Robust and Fine-Grained Detection of AI Generated Texts

Anonymous ACL submission

Abstract

An ideal detection system for machine generated content is supposed to work well on any generator as many more advanced LLMs come into existence day by day. Existing systems often struggle with accurately identifying AI-generated content over shorter texts. Further, not all texts might be entirely authored by a human or LLM, hence we focused more over partial cases i.e human-LLM co-authored texts. Our paper introduces a set of models built for the task of token classification which are trained on an extensive collection of human-machine co-authored texts, which performed well over texts of unseen domains, unseen generators, texts by non-native speakers and those with adversarial inputs. We also introduce a new dataset of over 2.4M such texts mostly co-authored by several popular proprietary LLMs over 23 languages. We also present findings of our models' performance over each texts of each domain and generator. Additional findings include comparison of performance against each adversarial method, length of input texts and characteristics of generated texts compared to the original human authored texts.

1 Introduction

Recent advancements in large language models (LLMs) have significantly narrowed the gap between machine-generated and human-authored text. As LLMs continue to improve in fluency and coherence, the challenge of reliably detecting AI-generated content could become increasingly critical. This issue is particularly pressing in domains such as education and online media, where the authenticity of textual material is paramount. While early efforts such as the GLTR (Gehrmann et al., 2019a) provided valuable insights by leveraging statistical methods to differentiate between human and machine text, these methods often lag behind the rapid pace of LLM evolution. Likewise, initiatives aimed at mitigating neural fake news (Zellers et al., 2019a) have made significant strides in addressing the societal implications of AI-generated misinformation. However, as LLMs become more sophisticated, existing detection systems must be re-evaluated and enhanced to maintain their effectiveness. Further, Each domain comes with its version of the issue of detecting machine generated texts. For instance, proprietary LLMs with internet access and better knowledge cutoffs are more likely to be used in domains like academia. Similarly, bad actors might use an open source generators for the task of creating misinformation and deception through machine generated online content as such models can be hosted locally to not leave a trail and are more flexible in terms of not denying user requests. Hence, tailoring models and approaches for each specific domain/scenario might be better applicable for practical scenarios. We chose a token-classification approach to train a model for the task of distinguish writing styles within a text if more than one were found. This approach helped us achieve better performance over texts of unseen features (i.e domain, generator, adversarial inputs, non-native speakers' texts) as our models were trained to distinguish different styles within a text rather than classifying an input text as one of the two classes it was trained on. Further, we explored the findings and results upon testing our models over other benchmarks which consist of texts from unseen domain and generators. We also tested our models over benchmarks which consist of texts with various adversarial inputs and those written by non-native speakers. We feel our findings and datasets can aid in further research into mitigating the harms of AI generated texts.

Source	Dataset/Benchmark	Samples	Languages	Generators
(Lee et al., 2022)	CoAuthor	1,445	1	1
(Zhang et al., 2024)	MixSet	3,600	1	12
(Dugan et al., 2022)	RoFT	21,646	1	5
(Macko et al., 2024b)	MultiTude	4,070	11	8
(Artemova et al., 2025)	Beemo	19,600	1	10
Our Work	*placeholder*	2,447,221	23	12

Table 1: Comparison with other Human-LLM co-authored datasets & benchmarks

2 Related Works

A major portion of current research in detecting machine-generated content focuses on longer-form writing through binary classification. However, AI-generated misinformation is more likely to cause harm than its use in academia, making the distinction between AI and human-generated texts on social media platforms a critical challenge. Existing methods often struggle with accurately identifying AI-generated content over shorter texts. Moreover, binary classification approaches, which categorize texts as either human or AI-generated (Wang et al., 2024a), (Wang et al., 2024b), (Bhattacharjee et al., 2023), (Zellers et al., 2019b), (Macko et al., 2023), (Ghosal et al., 2023), (Dugan et al., 2024) are less practical in settings where texts could be co-authored by both humans and LLMs. In contrast, binary classification may be more effective for shorter texts commonly found on reviews and social media platforms (Macko et al., 2024a), (Ignat et al., 2024), where content typically consists of one or two sentences. Additionally, some detection works rely on detecting watermarks from AI-generated texts, (Chang et al., 2024), (Dathathri et al., 2024), (Sadasivan et al., 2024), (Zhao et al., 2023) but not all generators utilize watermarking limiting the applicability of such approaches. Few other approaches utilize statistical methods (Mitchell et al., 2023), (Kumarage et al., 2023), (Gehrmann et al., 2019b), (Hans et al., 2024), (Bao et al., 2023), but they can be prone to mis-classification against adversarial methods like rephrasing and humanizing. (Abassy et al., 2024) introduced a 4-way classification as entirely human authored, entirely llm authored, human-edited and llm-authored or llm-edited human-authored. An ideal detection system should be capable of identifying AI-generated content from any generator without depending on watermarking, especially since watermarking techniques may not be effective for shorter texts. Further an ideal detector should be robust against adversarial methods. To properly deal with co-authored text cases, a token classification approach to detect boundaries (Dugan et al., 2022), (Macko et al., 2024b) between machine authored and human authored portions might be more appropriate. Further in cases of AI usage in scenarios like academic cases, users are likely to use a proprietary LLM with better knowledge cutoffs than an open source LLM. Similarly, for AI misuse over social platforms, users are more likely to use a open-sourced model due to better flexibility and privacy. Hence, building models and benchmarks with a appropriate set of LLMs might be more applicable for practical scenarios. Many proprietary systems struggle at the task of fine-grained detection, further a large enough dataset to cover all POS-tag bi-grams of the text boundaries is required for such fine-grained detectors to work well (Kadiyala, 2024). Previous works in the similar direction include (Lee et al., 2022), (Zhang et al., 2024), (Dugan et al., 2023), (Macko et al., 2024b), (Liang et al., 2024) which utilize a dataset of limited size and limited number of generators or those less likely to be used, which might not be enough for a detector to work well on unseen domains and generators' texts. Further, the task of detection of such human-llm co-authored texts is a harder task compared to binary classification of texts based on authorship (Geng and Trotta, 2025) (Huang et al., 2025).

3 Dataset

Our dataset consists of around 2.45M samples. We used 12 different LLMs out of which 9 are popular proprietary LLMs : GPT-o1 (OpenAI, 2024), GPT-4o (etal., 2024), Gemini-1.5-Pro (DeepMind, 2024), Gemini-1.5-Flash, Claude-3.5-Sonnet (Anthropic, 2023), Claude-3.5-Haiku, Perplexity-Sonar-Large (Perplexity, 2023), Amazon-Nova-Pro (Intelligence, 2024), Amazon-Nova-Lite. We also included 3 open-source LLMs i.e Aya-23 (Aryabumi et al., 2024), Command-R-Plus (Cohere For AI, 2024), Mistral-large-2411 (Mistral AI, 2024) which produced outputs that are relatively difficult to distinguish from human written texts compared to other similar models in other benchmarks[1] as well as our own datasets. The samples range from 30 to 25K words in length with an average length of around 600 words. Table 1 provides an overview of comparison with other datasets and benchmarks for fine-grained detection of human-llm co-authored texts. Our dataset utilizes a better choice of generators which are more likely to be used in practical scenarios where as prior datasets are limited to smaller and limited number of models. Our dataset also compromises of over 100 times more samples, aiding other researchers in the field.

3.1 Dataset Distribution

The language distribution of the dataset and LLMs used can be seen in Figure 1. Each language-LLM pair has roughly 10000 samples. Among each set of the 10000 samples; training, development and test sets constitute 40%, 10%, 50% respectively. Additionally, among each set of 10000 samples, 10% were Completely human written, another 10% completely machine generated, and the other 80% were human-LLM co-authored i.e few portions of the text are machine generated and the rest are human written.

3.2 Dataset Creation

GPT-4o was used through a Azure OpenAI endpoint[2]. command-r-plus and aya-23 were used through cohere's API platform[3]. Rest of the models were used through open router's[4] API. The Rewritten samples were created by providing the generator LLM with the original text and a random prompt among writing an alternate version, a later update of what happened or a rephrased version of the same text. The samples which returned the exact text or a very similar text were once again regenerated. The partially machine generated texts were created by splitting the text at random locations and the generator was asked to finish the text. The split locations were chosen randomly from between the 30th word up to end of text. This was done to provide the LLM with enough context to better work towards text completion.

3.3 Original Data Source and Filtering

With a goal of training on one domain and testing on every other, we chose to train on old newspapers (HC-Corpora) as it has sufficient number of samples i.e 17.2M for 67 languages of the same domain. We then removed samples which originated after release of gpt-3 to avoid mislabelling of samples in our dataset. Further we sampled texts which were at least 3 sentences or 50 words long. For Chinese and Japanese, we sampled texts which were at least 100 characters long.

4 Our System

We have experimented with various multilingual transformer models (He et al., 2023), (Conneau, 2019), (Beltagy et al., 2020) with/without additional LSTM (Hochreiter, 1997) or CRF layers (Zheng et al., 2015) through a binary token-classification approach. We found that using additional CRF layer produced better results compared to other setups with the same model. All of the transformer models tested have produced nearly identical results over our test set. However, XLM-longformer gave better results over unseen domains and generators' texts, and was used in the end given the longer default context length of 16384. The token level predictions by the models were then mapped into word-level predictions. We use the model's predictions to separate text portions based on perceived authorship. Improving the performance required balancing pre and post boundary POS tag bi-grams to reduce error rates (Kadiyala, 2024).

5 Evaluation and Results

We evaluate the models at 3 levels of granularity : word level, sentence level and overall. For Chinese and Japanese, we performed evaluation at a character level instead of word-level. Each domain and user might have a different preference towards metrics and evaluation, hence we report 3 metrics at each level of granularity : accuracy, recall and precision. For word level mapping of predictions, in cases where part of a word i.e a few tokens are classified differently than others, we assigned the same label to the word as its first token. While mapping word level predictions to a sentence we

[1] https://raid-bench.xyz/
[2] https://azure.microsoft.com/en-us/products/ai-services/openai-service/
[3] https://dashboard.cohere.com/
[4] https://openrouter.ai/models

Model	o1	4o	Nova-Pro	Nova-Lite	Sonar-Large	Aya-23	Command R+	Large 2411	3.5-Sonnet	3.5-Haiku	1.5-Pro	1.5-Flash	Total
Arabic	10,000	9,997	10,000	9,989	—	9,985	9,985	9,955	10,000	9,995	9,974	10,000	109,880
Chinese	10,000	9,997	9,999	9,996	—	10,000	10,000	10,000	10,000	9,995	10,000	10,000	109,987
Czech	10,000	9,905	9,999	9,996	—	9,983	—	9,999	10,000	9,918	10,000	10,000	99,800
Dutch	10,000	9,969	10,000	9,962	—	10,002	—	10,000	10,000	9,884	10,000	10,000	99,817
English	10,000	9,998	10,000	9,994	9,961	9,978	9,989	10,000	10,000	9,997	9,998	10,000	119,915
French	10,000	9,972	9,999	9,977	9,990	9,982	9,993	10,000	10,000	9,935	10,000	10,000	119,848
German	10,000	9,983	10,000	9,995	10,000	9,995	9,993	9,998	10,000	9,969	10,000	10,000	119,933
Greek	10,000	9,997	9,947	9,992	—	9,940	—	10,000	10,000	9,974	10,000	10,000	99,851
Hebrew	10,000	9,998	10,000	9,999	—	9,982	—	10,000	10,000	9,924	10,000	10,000	99,902
Hindi	10,000	9,995	10,000	10,000	—	9,976	—	10,000	10,000	9,999	10,000	10,000	99,970
Indonesian	10,000	9,992	9,999	9,991	—	9,981	—	9,999	10,000	9,977	10,000	10,000	99,939
Italian	9,995	9,960	10,000	9,993	—	9,988	9,995	10,000	10,000	9,934	10,000	10,000	109,865
Japanese	9,989	9,962	10,000	9,999	—	10,000	10,000	9,999	10,000	9,907	10,000	10,000	109,856
Korean	10,000	9,986	9,998	9,996	—	9,869	9,898	9,997	10,000	9,956	9,997	10,000	109,697
Persian	9,999	9,996	9,998	9,999	—	9,998	—	10,000	10,000	9,991	10,000	10,000	99,981
Polish	10,000	9,978	9,998	9,993	—	9,954	—	10,000	10,000	9,925	10,000	10,000	99,848
Portuguese	9,999	9,982	9,998	9,991	9,939	9,993	9,993	9,996	10,000	9,893	10,000	10,000	119,784
Romanian	10,000	9,978	9,998	9,990	—	9,961	—	9,998	10,000	9,950	10,000	10,000	99,875
Russian	10,000	9,992	9,996	9,995	—	9,952	—	9,997	10,000	9,977	10,000	10,000	99,910
Spanish	10,000	9,975	9,999	9,997	9,933	9,980	9,978	9,997	10,000	9,932	10,000	10,000	119,791
Turkish	10,000	9,962	9,999	9,996	—	9,972	—	9,996	10,000	9,993	10,000	10,000	99,918
Ukrainian	10,000	9,993	9,995	9,998	—	9,988	—	9,997	10,000	9,954	10,000	10,000	99,925
Vietnamese	10,000	9,973	9,988	9,995	—	9,999	—	9,977	10,000	9,977	10,000	10,000	99,929
Total	229,982	229,541	229,910	229,833	49,823	229,458	99,824	229,925	230,000	228,956	229,969	230,000	2,447,221

Figure 1: Dataset distribution per each generator and language in our dataset

used majority voting, and in cases where consensus is not obtained, we assigned the same label as the first word. For evaluation over other benchmarks requiring binary classification of texts as human or machine written, we assign a human written label to the text if at least two thirds of the words get classified as human written. We also report several metrics, some of which can be seen in the below tables, rest can be found in Appendix D.

5.1 Seen Domains & Seen Generators

The results of our models over our dataset's test set can be seen in Table 2. The samples from both the data splits are of the same domain and originate from the same set of generators.

5.2 Unseen Domains & Unseen Generators

The models were tested twice over (Wang et al., 2024a): once by training on just 10000 samples of a single generator (Aya-23) and again later by training over our complete training data. The benchmark consists of 11,123 samples of peer reviews and student essays (Koike et al., 2024), the generators used were various versions of llama-2 and chat-gpt (earlier version of gpt-4). the samples would hence be from completely unseen domains and generators to our models. The results of both models can be seen in Table 3.

5.3 Unseen Domains & Unseen Generators & Non-Native Speakers

The models were tested by training on just 10k samples each from Aya-23 for English and Arabic Separately. The benchmark's samples for Arabic were from (Alfaifi, 2013) and (Zaghouani et al., 2024). The samples for English consist of ETS and IELTS student essays sampled from non-native speakers (Chowdhury et al., 2025). Our models were used for inference directly over these texts and the strings of predicted tokens were then used to for binary classification based on how frequently the perceived authorship changed from human to

Language ↓	Partial cases	Unchanged cases	Rewritten cases	Overall
Arabic	97.16	90.69	97.55	96.44
Chinese*	93.13	76.28	91.40	86.58
Czech	96.23	79.63	93.84	94.98
Dutch	96.83	77.60	94.13	95.31
English	97.32	90.23	97.68	96.02
French	96.89	74.46	96.52	94.91
German	96.64	76.54	95.92	95.28
Greek	96.25	82.21	92.08	94.37
Hebrew	96.52	80.56	95.34	95.70
Hindi	97.08	92.60	97.24	96.34
Indonesian	97.20	84.92	97.19	96.64
Italian	96.44	80.69	96.84	95.38
Japanese*	92.74	83.80	92.81	86.13
Korean	97.29	84.13	94.74	95.77
Persian	96.60	88.61	96.19	94.36
Polish	96.63	88.52	92.75	95.94
Portuguese	96.46	88.51	90.29	94.89
Romanian	97.59	78.06	95.15	96.10
Russian	96.64	79.98	95.58	94.02
Spanish	96.38	71.60	96.69	94.47
Turkish	95.74	83.00	94.48	93.62
Ukrainian	95.74	74.03	96.57	93.53
Vietnamese	94.41	77.99	96.65	89.67
Average	**96.26**	**81.94**	**95.11**	**94.19**

Table 2: Word-Level Accuracy (.2f) of the models on the test dataset for each case

* Character level evaluations were done instead for Japanese and Chinese

Metrics →	Accuracy	Precision	Recall	F1
Initial Model	86.51	91.61	87.46	89.49
Final Model	86.00	87.16	92.25	89.63

Table 3: Word level Metrics over Mgtd-bench (.2f) through our models (zero-shot, unseen domains, unseen generators)

Metrics →	Accuracy	Precision	Recall	F1
Arabic	95.9	96.1	94.5	95.2
English	99.1	98.7	99.3	99.0
Arabic-Best	96.1	96.1	95.0	95.5
English-Best	99.3	99.0	99.2	99.1

Table 4: Overall Metrics over ETS essays (.1f) through our detectors (zero-shot, unseen generators, unseen domain) VS best submissions (fine-tuned on same generators anbd domain)

LLM and vice-versa i.e the number of changes and whether the longest string consists of ones or zeroes. The metrics obtained for each language can be seen in Table 4.

5.4 Unseen Domains & Partially Seen Generators & Adversarial Inputs

We have also tested over raid-bench (Dugan et al., 2025) which consists of texts from 11 generators and 8 domains. among them roughly 10% would be from a seen domain (news articles) while the rest are unseen by our models. The dataset's texts were also created using various sampling strategies (greedy, random, etc.). The texts were also modified to have adversarial methods including homoglyphs, mis-spellings, alternative spellings, article deletion etc. Among the 11 generators used, Gpt-4 is one which is similar to the generator whose outputs our model has been trained on (Gpt-4o). However, both of them have different linguistic and stylistic features, similar to how Gpt-4 is different from Gpt-3. We have tested our model's performance once again upon being trained on our own full training data. Additionally, we have also performed an error analysis to find out what domains, models, attack strategies and decoding strategies effected the model's performance and to what extent. This can be seen in Figure 2, Figure 3, Figure 4 and Figure 5. The texts were classified as machine generated if at least one third of the tokens within the model's context length were classified as machine generated. The F1 score obtained with the initial model trained on a single generator was 0.63 and the F1 score grew to 0.79 upon being trained on our full dataset. Evaluation was done directly without performing any preprocessing of the texts and neither were our models trained on texts with any of the adversarial methods.

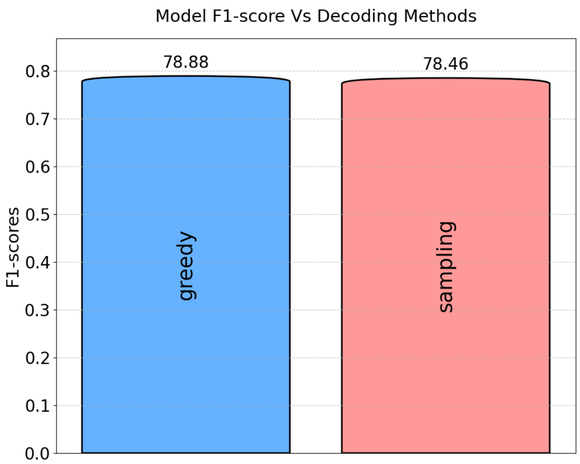

Figure 2: F1 scores VS text sampling method used : Sampling strategy did not effect detection capability

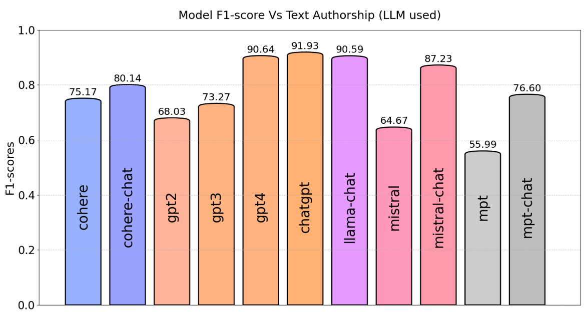

Figure 3: F1 scores VS the generator's texts : instruct/chat variants were harder to detect.

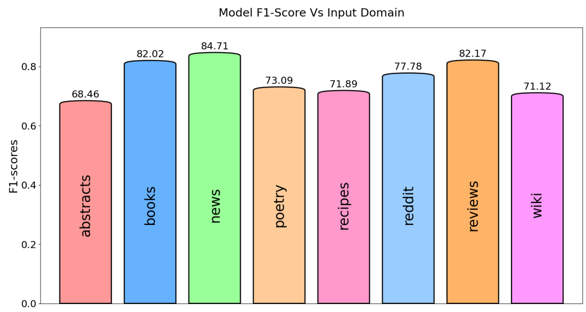

Figure 4: F1 scores VS each domain's texts : news i.e the only training data domain was easier to detect.

5.5 Comparisons with proprietary systems

While many proprietary systems claim to have excellent results, they often struggle with unseen domains and generators. Further many systems like ZeroGPT[5] and GPTZero[6] do not provide fine-grained predictions through their API. This would require manually evaluating the results through the UI by counting correct and incorrectly classified word counts. Hence, these systems were not used for comparisons for fine-grained scores. Several users have tried this manfully over a subset (Kadiyala, 2024) only to find a large gap in performance. For binary classifications as well, they have a minimum length threshold to even work, which would become incompatible over a large number of samples of the benchmark.

6 Other Observations

The sentences inside which text authorship switches from human to LLM or vice versa were found to be relatively shorter that the original text portions which they replaced. LLMs may be likely to finish the current sentence earlier than usual to move on to the next sentence in text completion scenarios. The mean length of the original portion and the replaced portions of those sentences for each language and generator can be seen in Table 5 and Table 6 respectively. This observation was consistent across all languages and generators with a 20-30% reduction and a larger reduction in Hindi. For Chinese and Japanese too, we did observe a 20-30% reduction in character count when comparing the original and replaced portions of the sentence after the text boundary. Although there is a good variation in this feature across languages, the mean and medians observed for each language were similar for all the LLMs. This is further elaborated in Appendix A.

[5]https://github.com/zerogpt-net/zerogpt-api
[6]https://gptzero.stoplight.io/docs/gptzero-api/

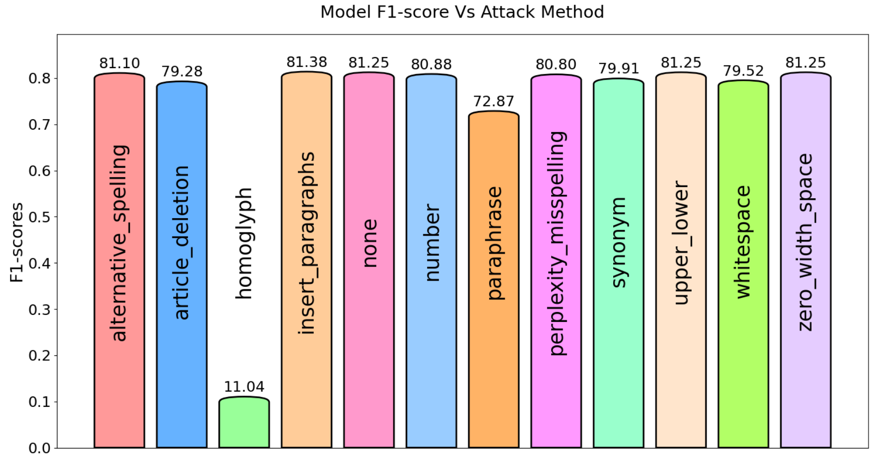

Figure 5: F1 scores VS adversarial method used in the input texts : homo-glyphs are the only real issue, and paraphrases to a small extent, while the rest can be handled through pre-processing.

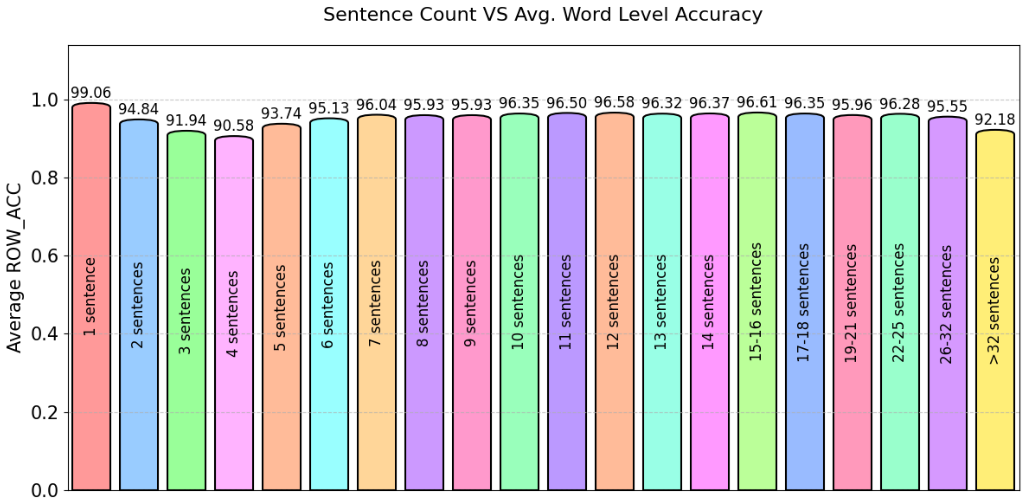

Figure 6: Accuracy VS length of input texts (sentence count)

7 Conclusion

Despite not being trained on the domains or generators, the models built through our approach performed well over several benchmarks as seen in subsection 5.3 and subsection 5.4 over inputs which were from non-native speakers and consist of adversarial methods. Further, one case where many proprietary systems struggle is when the inputs were too short, which our models were able to overcome as seen in Figure 6, which demonstrates our models' accuracy over our test set compared to input text's sentence count. Table 7 displays our model's performance over English subset of our dataset for each generator. A similar trend from subsection 5.4 was observed with models which are likely less instruction-tuned / not instruction-tuned tend to produce texts which are harder to distinguish than their alternatives.

7.1 Scalability and scope for extension

The original dataset used to train our current models as mentioned in section 3 consists of samples over 60 languages which would cover 70% of the world population's primary language, and all of the languages are supported by existing multilingual transformer models making the process of scaling the work to more languages easier. Despite not being trained on the generators or domains' texts, our models were able to perform well on several benchmarks. Even reaching a F1 score of 0.79 against adversarial inputs while they were neither trained over them nor pre-processed. Similarly, creation and usage of such large datasets of other domains along with ours might result in robust and better models. We couldn't explore the relation between instruction tuning sample size of LLMs and detectability of their texts due to the proprietary

Language	Length of Original part	Length of generated part
Arabic	17	13
Czech	11	8
Dutch	12	10
English	15	11
French	14	11
German	12	9
Greek	15	12
Hebrew	11	9
Hindi	26	12
Indonesian	11	8
Italian	15	14
Korean	9	7
Persian	17	15
Polish	10	7
Portuguese	15	11
Romanian	14	11
Russian	11	9
Spanish	15	12
Turkish	10	8
Ukrainian	11	8
Vietnamese	18	14
Average	13.8	10.4

Table 5: Median length (words) of original & newly generated parts of the sentences - Language wise : Models tend to finish off current sentence after authorship switch quickly before continuing with the rest of the text.

Generator	Length of original part	Length of generated part
Amazon-Nova-Pro	14	10
Amazon-Nova-Lite	12	10
Aya-23-35B	11	10
Claude-3.5-Haiku	18	10
Claude-3.5-Sonnet	16	10
Command-R-Plus	16	10
GPT-4o	12	10
GPT-o1	11	9
Gemini-1.5-Pro	15	10
Gemini-1.5-Flash	9	10
Mistral-Large-2411	11	10
Perplexity-Sonar-large	15	11
Average	13.3	10

Table 6: Median length (words) of original & newly generated parts of the sentences : Generator wise

Generator	Accuracy
Amazon-Nova-Pro	94.90
Amazon-Nova-Lite	95.26
Aya-23-35B	91.75
Claude-3.5-Haiku	96.07
Claude-3.5-Sonnet	95.97
Command-R-Plus	93.92
GPT-4o	91.78
GPT-o1	96.61
Gemini-1.5-Flash	92.34
Gemini-1.5-pro	93.38
Mistral-Large-2411	93.47
Perplexity-Sonar-large	94.91
Average	94.31

Table 7: Word level accuracy (.2f) of our models over our dataset (English)

* excluding Chinese and Japanese

nature of most of the generators we used, but a similar study using open-data models could uncover more insights.

7.2 Scope for Improvement

As seen in Figure 5, almost none of the adversarial methods affected the models built through our approach other than paraphrasing and homo-glyphs. However homo-glyphs can be pre-processed by mapping them to the actual character they were imitating in the text. This would require a large collection of homo-glyph to character mapping set to use for pre-processing. Further, paraphrased samples of various number of iterations being included in the training dataset might lead to further improvements. It is also worth exploring how detectable are texts in cases where multiple generators contribute a portion each in a human authored text. Other missing adversarial methods that are likely to be used in practical scenarios include usage of proprietary systems that 'humanize' a given text in an attempt to evade detection.

7.3 Ideal Usage

The models were built primarily for a human-in-the-loop use cases where the model would try to flag most of the likely machine-generated portions while the flagged content can be validated either through an ensemble of models or a human and hence a tilt towards higher recall can be observed in the metrics as seen in Table 13.

Limitations

Just like any other detector or classifier, no detector can guarantee a 100% accuracy and hence the models are not meant to be used directly for decision making but are meant to be used in a human-in-the-loop scenarios. Furthermore, the experiments carried out did not include cases of multiple LLMs co-authoring a portion each of the same text.

References

Mervat Abassy, Kareem Elozeiri, Alexander Aziz, Minh Ngoc Ta, Raj Vardhan Tomar, Bimarsha Adhikari, Saad El Dine Ahmed, Yuxia Wang, Osama Mohammed Afzal, Zhuohan Xie, Jonibek Mansurov, Ekaterina Artemova, Vladislav Mikhailov, Rui Xing, Jiahui Geng, Hasan Iqbal, Zain Muhammad Mujahid, Tarek Mahmoud, Akim Tsvigun, Alham Fikri Aji, Artem Shelmanov, Nizar Habash, Iryna Gurevych, and Preslav Nakov. 2024. Llm-detectaive: a tool for fine-grained machine-generated text detection.

AYG Alfaifi. 2013. Arabic learner corpus v1: A new resource for arabic language research. In *Second Workshop on Arabic Corpus Linguistics*.

Anthropic. 2023. Model card: Claude 3. Technical report, Anthropic. Accessed: 2024-04-27.

Ekaterina Artemova, Jason Lucas, Saranya Venkatraman, Jooyoung Lee, Sergei Tilga, Adaku Uchendu, and Vladislav Mikhailov. 2025. Beemo: Benchmark of expert-edited machine-generated outputs.

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Jon Ander Campos, Yi Chern Tan, Kelly Marchisio, Max Bartolo, Sebastian Ruder, Acyr Locatelli, Julia Kreutzer, Nick Frosst, Aidan Gomez, Phil Blunsom, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. 2024. Aya 23: Open weight releases to further multilingual progress.

Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. 2023. Fast-detectgpt: Efficient zero-shot detection of machine-generated text via conditional probability curvature. *arXiv preprint arXiv:2310.05130*.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Amrita Bhattacharjee, Tharindu Kumarage, Raha Moraffah, and Huan Liu. 2023. Conda: Contrastive domain adaptation for ai-generated text detection.

Yapei Chang, Kalpesh Krishna, Amir Houmansadr, John Wieting, and Mohit Iyyer. 2024. Postmark: A robust blackbox watermark for large language models. *arXiv preprint arXiv:2406.14517*.

Yutian Chen, Hao Kang, Vivian Zhai, Liangze Li, Rita Singh, and Bhiksha Raj. 2023. Gpt-sentinel: Distinguishing human and chatgpt generated content. *arXiv preprint arXiv:2305.07969*.

Shammur Absar Chowdhury, Hind Almerekhi, Mucahid Kutlu, Kaan Efe Keleş, Fatema Ahmad, Tasnim Mohiuddin, George Mikros, and Firoj Alam. 2025. GenAI content detection task 2: AI vs. human – academic essay authenticity challenge. In *Proceedings of the 1stWorkshop on GenAI Content Detection (GenAIDetect)*, pages 323–333, Abu Dhabi, UAE. International Conference on Computational Linguistics.

Cohere For AI. 2024. c4ai-command-r-plus-08-2024 (revision dfda5ab).

A Conneau. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Sumanth Dathathri, Abigail See, Sumedh Ghaisas, Po-Sen Huang, Rob McAdam, Johannes Welbl, Vandana Bachani, Alex Kaskasoli, Robert Stanforth, Tatiana Matejovicova, et al. 2024. Scalable watermarking for identifying large language model outputs. *Nature*, 634(8035):818–823.

DeepMind. 2024. Gemini v1.5 report. `https://storage.googleapis.com/deepmind-media/gemini/gemini_v1_5_report.pdf`. Accessed: 2025-02-08.

Liam Dugan, Alyssa Hwang, Filip Trhlik, Josh Magnus Ludan, Andrew Zhu, Hainiu Xu, Daphne Ippolito, and Chris Callison-Burch. 2024. Raid: A shared benchmark for robust evaluation of machine-generated text detectors. *arXiv preprint arXiv:2405.07940*.

Liam Dugan, Daphne Ippolito, Arun Kirubarajan, Sherry Shi, and Chris Callison-Burch. 2022. Real or fake text?: Investigating human ability to detect boundaries between human-written and machine-generated text.

Liam Dugan, Daphne Ippolito, Arun Kirubarajan, Sherry Shi, and Chris Callison-Burch. 2023. Real or fake text?: Investigating human ability to detect boundaries between human-written and machine-generated text. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(11):12763–12771.

Liam Dugan, Andrew Zhu, Firoj Alam, Preslav Nakov, Marianna Apidianaki, and Chris Callison-Burch. 2025. GenAI content detection task 3: Cross-domain machine generated text detection challenge. In *Proceedings of the 1stWorkshop on GenAI Content Detection (GenAIDetect)*, pages 377–388, Abu Dhabi, UAE. International Conference on Computational Linguistics.

OpenAI etal. 2024. Gpt-4 technical report.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander Rush. 2019a. GLTR: Statistical detection and visualization of generated text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 111–116, Florence, Italy. Association for Computational Linguistics.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander M Rush. 2019b. Gltr: Statistical detection and visualization of generated text. *arXiv preprint arXiv:1906.04043*.

Mingmeng Geng and Roberto Trotta. 2025. Human-llm coevolution: Evidence from academic writing.

Soumya Suvra Ghosal, Souradip Chakraborty, Jonas Geiping, Furong Huang, Dinesh Manocha, and Amrit Singh Bedi. 2023. Towards possibilities & impossibilities of ai-generated text detection: A survey. *arXiv preprint arXiv:2310.15264*.

Abhimanyu Hans, Avi Schwarzschild, Valeriia Cherepanova, Hamid Kazemi, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2024. Spotting llms with binoculars: Zero-shot detection of machine-generated text.

HC-Corpora. Old newspapers. https://www.kaggle.com/datasets/alvations/old-newspapers.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTav3: Improving deBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

S Hochreiter. 1997. Long short-term memory. *Neural Computation MIT-Press*.

Baixiang Huang, Canyu Chen, and Kai Shu. 2025. Authorship attribution in the era of llms: Problems, methodologies, and challenges. *SIGKDD Explor. Newsl.*, 26(2):21–43.

Oana Ignat, Xiaomeng Xu, and Rada Mihalcea. 2024. Maide-up: Multilingual deception detection of gpt-generated hotel reviews.

Amazon Artificial General Intelligence. 2024. The amazon nova family of models: Technical report and model card. *Amazon Technical Reports*.

Ram Mohan Rao Kadiyala. 2024. RKadiyala at SemEval-2024 task 8: Black-box word-level text boundary detection in partially machine generated texts. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 511–519, Mexico City, Mexico. Association for Computational Linguistics.

Ryuto Koike, Masahiro Kaneko, and Naoaki Okazaki. 2024. Outfox: Llm-generated essay detection through in-context learning with adversarially generated examples. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 21258–21266.

Tharindu Kumarage, Joshua Garland, Amrita Bhattacharjee, Kirill Trapeznikov, Scott Ruston, and Huan Liu. 2023. Stylometric detection of ai-generated text in twitter timelines. *arXiv preprint arXiv:2303.03697*.

Mina Lee, Percy Liang, and Qian Yang. 2022. Coauthor: Designing a human-ai collaborative writing dataset for exploring language model capabilities. In *Proceedings of the 2022 CHI conference on human factors in computing systems*, pages 1–19.

Weixin Liang, Zachary Izzo, Yaohui Zhang, Haley Lepp, Hancheng Cao, Xuandong Zhao, Lingjiao Chen, Haotian Ye, Sheng Liu, Zhi Huang, et al. 2024. Monitoring ai-modified content at scale: A case study on the impact of chatgpt on ai conference peer reviews. *arXiv preprint arXiv:2403.07183*.

Dominik Macko, Jakub Kopal, Robert Moro, and Ivan Srba. 2024a. Multisocial: Multilingual benchmark of machine-generated text detection of social-media texts.

Dominik Macko, Robert Moro, Adaku Uchendu, Jason Lucas, Michiharu Yamashita, Matúš Pikuliak, Ivan Srba, Thai Le, Dongwon Lee, Jakub Simko, and Maria Bielikova. 2023. Multitude: Large-scale multilingual machine-generated text detection benchmark. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, page 9960–9987. Association for Computational Linguistics.

Dominik Macko, Robert Moro, Adaku Uchendu, Ivan Srba, Jason Samuel Lucas, Michiharu Yamashita, Nafis Irtiza Tripto, Dongwon Lee, Jakub Simko, and Maria Bielikova. 2024b. Authorship obfuscation in multilingual machine-generated text detection. *arXiv preprint arXiv:2401.07867*.

Mistral AI. 2024. Mistral large 2407. https://mistral.ai/en/news/mistral-large-2407. Accessed: 2025-02-08.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature. In *International Conference on Machine Learning*, pages 24950–24962. PMLR.

OpenAI. 2024. Openai system card. https://cdn.openai.com/o1-system-card.pdf. Accessed: 2025-02-08.

Perplexity. 2023. Sonar. https://sonar.perplexity.ai/. Accessed: 2025-02-08.

Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. 2024. Can ai-generated text be reliably detected?

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Osama Mohammed Afzal, Tarek Mahmoud, Giovanni Puccetti, Thomas Arnold, Chenxi Whitehouse, Alham Fikri Aji, Nizar Habash, Iryna Gurevych, and Preslav Nakov. 2024a. Semeval-2024 task 8: Multidomain, multimodel and multilingual machine-generated text detection.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Osama Mohanned Afzal, Tarek Mahmoud, Giovanni Puccetti, Thomas Arnold, Alham Fikri Aji, Nizar Habash, Iryna Gurevych, and Preslav Nakov. 2024b. M4gt-bench: Evaluation benchmark for black-box machine-generated text detection.

Wajdi Zaghouani, Abdelhamid Ahmed, Xiao Zhang, and Lameya Rezk. 2024. Qcaw 1.0: Building a qatari corpus of student argumentative writing. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13382–13394.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019a. Defending against neural fake news. *Advances in neural information processing systems*, 32.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019b. Defending against neural fake news. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Qihui Zhang, Chujie Gao, Dongping Chen, Yue Huang, Yixin Huang, Zhenyang Sun, Shilin Zhang, Weiye Li, Zhengyan Fu, Yao Wan, et al. 2024. Llm-as-a-coauthor: Can mixed human-written and machine-generated text be detected? In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 409–436.

Xuandong Zhao, Prabhanjan Ananth, Lei Li, and Yu-Xiang Wang. 2023. Provable robust watermarking for ai-generated text. *arXiv preprint arXiv:2306.17439*.

Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip HS Torr. 2015. Conditional random fields as recurrent neural networks. In *Proceedings of the IEEE international conference on computer vision*, pages 1529–1537.

A Pre- and Post- Boundary Comparisons

The mean and median word counts of the text portions in a sentence after the text authorship shifts from human to LLM can be seen in Table 11 and Table 12 in comparison to the texts they replace.

B Dataset Creation

The max_new_tokens value specified to the generator during creation of partial cases was randomized between 80% to 200% of the length of the portion that is being replaced. The prompts used for creation of the partial samples and rewritten samples can be seen in Table 8 and Table 9 respectively.

continue this text in Language directly :
complete this text in Language, respond directly :

Table 8: Prompts used in dataset creation : Partial cases

Rewrite this in Language a different way :
Generate an alternative version of this in Language :
Generate a later update to this in Language :
Generate a previous version of this in Language ;

Table 9: Prompts used in dataset creation : Rewritten cases

Hyperparameter	Value
Seed (Training)	1024
Seed (Shuffling)	1024
Number of Epochs	5
Per Device Batch Size (Train)	12
Per Device Batch Size (Eval)	30
Context Length	16384
Learning Rate	5e-5
Weight Decay	0
Dropout (CRF Layer)	0.075

Table 10: Training Hyper-parameters used

C Reproducibility

We used multilingual longformer [7] with an additional CRF layer. The hyper-parameters used for training the models can be seen in Table 10. We built a separate model for each language, the training was done over A100 SXM over 10h each.

D Other Metrics

The metrics over each type of text for each language and LLM separately can be seen in Table 14, Table 15, Table 16, Table 17, Table 18, Table 19, Table 20, Table 21, Table 22, Table 23, Table 24, Table 25.

[7] https://huggingface.co/hyperonym/xlm-roberta-longformer-base-16384

Language ↓	**Mean length of old text portion**	**Mean length of new text portion**	**Median length of Old text portion**	**Median length of New text portion**
Arabic	18.73	16.25	17	13
Czech	12.02	9.52	11	8
Dutch	13.73	13.39	12	10
English	16.04	14.59	15	11
French	15.50	13.16	14	11
German	13.03	10.89	12	9
Greek	16.74	14.87	15	12
Hebrew	12.64	10.65	11	9
Hindi	40.56	15.42	26	12
Indonesian	12.44	9.56	11	8
Italian	17.54	16.39	15	14
Korean	9.85	8.08	9	7
Persian	18.83	19.88	17	15
Polish	11.42	8.84	10	7
Portuguese	16.52	13.29	15	11
Romanian	16.30	13.50	14	11
Russian	12.27	10.63	11	9
Spanish	17.18	14.81	15	12
Turkish	11.81	9.74	10	8
Ukrainian	12.04	10.39	11	8
Vietnamese	20.06	18.01	18	14
Average	**16.19**	**12.95**	**13.76**	**10.43**

Table 11: Comparison of replaced and generated text portion lengths (word count) : Language wise

Generator↓	**Mean length of old text portion**	**Mean length of new text portion**	**Median length of Old text portion**	**Median length of New text portion**
Amazon-Nova-Pro	16.02	13.27	12	10
Amazon-Nova-Lite	18.12	12.87	14	10
Aya-23-35B	13.70	12.87	11	10
Claude-3.5-Haiku	20.19	13.13	18	10
Claude-3.5-Sonnet	17.32	12.98	16	10
Command-R-Plus	16.92	13.28	16	10
GPT-4o	13.49	12.84	12	10
GPT-o1	14.83	12.36	11	9
Gemini-1.5-Flash	19.68	13.44	15	10
Gemini-1.5-pro	12.50	13.48	9	10
Mistral-Large-2411	12.85	12.85	11	10
Perplexity-Sonar-large	17.13	13.45	15	11
Average	**16.06**	**13.07**	**13.33**	**10**

Table 12: Comparison of replaced and generated text portion lengths (word count) : Generator wise

E License

The xlm-longformer model we used was available with an mit license, we are releasing the models and datasets through CC BY-NC 4.0[8] which permits usage for research purposes.

[8] https://creativecommons.org/licenses/by-nc/4.0/deed.en

Language ↓	Accuracy	Precision	Recall	F1-score
Arabic	96.44	92.50	97.17	94.78
Chinese*	86.58	87.03	86.46	86.75
Czech	94.98	94.57	97.96	96.23
Dutch	95.31	93.34	97.97	95.60
English	96.02	92.34	98.44	95.29
French	94.91	93.64	98.42	95.97
German	95.28	94.87	98.38	96.59
Greek	94.37	93.69	96.51	95.08
Hebrew	95.70	95.32	97.94	96.61
Hindi	96.34	89.72	96.66	93.06
Indonesian	96.64	95.61	98.29	96.93
Italian	95.38	95.04	97.58	96.29
Japanese*	86.13	85.64	94.17	89.70
Korean	95.77	95.29	97.69	96.48
Persian	94.36	84.45	96.88	90.24
Polish	95.94	96.76	97.19	96.97
Portuguese	94.89	91.92	96.07	93.95
Romanian	96.10	95.81	98.53	97.15
Russian	94.02	86.67	97.29	91.67
Spanish	94.47	90.02	98.14	93.90
Turkish	93.62	88.56	97.17	92.66
Ukrainian	93.53	86.58	97.93	91.90
Vietnamese	89.67	77.23	97.44	86.17
Average	**94.19**	**91.16**	**96.97**	**93.91**

Table 13: Word-level Metrics of our models over each language : our test set

* Character level evaluations were done instead for Japanese and Chinese

Language ↓	Partial cases	Unchanged cases	Rewritten cases	Overall
Arabic	97.56	88.10	98.09	97.17
Chinese*	93.70	75.35	91.60	87.00
Czech	96.63	80.10	94.36	95.20
Dutch	95.21	78.00	92.10	95.23
English	97.77	89.61	98.87	96.60
French	97.34	72.85	97.14	95.28
German	96.92	75.73	95.29	95.58
Greek	95.65	81.80	82.85	92.96
Hebrew	97.35	70.89	96.26	95.73
Hindi	96.65	92.82	96.67	96.59
Indonesian	97.27	85.73	95.76	96.60
Italian	96.88	80.88	94.70	95.65
Japanese*	97.48	88.57	93.94	96.85
Korean	97.68	84.15	93.73	95.39
Persian	96.91	89.45	93.22	94.05
Polish	96.96	87.52	92.32	95.98
Portuguese	95.28	94.15	96.32	95.28
Romanian	96.53	76.55	96.64	96.23
Russian	96.46	79.10	94.45	94.03
Spanish	96.92	71.75	96.97	94.97
Turkish	95.55	82.68	98.17	92.65
Ukrainian	95.39	73.81	95.48	93.90
Vietnamese	94.46	76.17	97.14	88.74

Table 14: Case wise accuracies over all languages for each generator : amazon-nova-pro

Language ↓	Partial cases	Unchanged cases	Rewritten cases	Overall
Arabic	96.56	90.93	95.62	95.88
Chinese*	93.20	76.99	93.57	87.77
Czech	97.81	79.40	94.58	96.43
Dutch	97.07	78.10	92.85	95.19
English	98.11	89.25	98.73	96.80
French	97.59	76.28	97.79	96.07
German	98.01	76.34	95.52	96.76
Greek	96.00	79.78	88.01	93.66
Hebrew	98.05	83.84	94.35	96.78
Hindi	96.49	91.59	95.30	95.38
Indonesian	97.99	85.03	97.18	97.06
Italian	96.95	80.81	95.45	95.54
Japanese*	98.07	76.50	93.02	92.78
Korean	98.20	82.49	95.42	95.68
Persian	97.40	88.48	94.92	95.31
Polish	97.55	89.32	93.83	96.63
Portuguese	92.67	87.92	95.09	94.35
Romanian	97.97	76.44	93.76	96.20
Russian	97.22	81.47	96.30	95.10
Spanish	97.49	71.55	97.15	94.98
Turkish	96.63	83.99	90.84	93.87
Ukrainian	96.74	74.80	99.91	94.24
Vietnamese	95.28	78.87	97.03	89.76

Table 15: Case wise accuracies over all languages for each generator : amazon-nova-lite

Language ↓	Partial cases	Unchanged cases	Rewritten cases	Overall
Arabic	95.14	92.22	97.76	96.05
Chinese*	85.65	75.66	89.10	82.40
Czech	89.75	79.02	87.94	89.24
Dutch	92.45	79.97	92.99	93.12
English	93.01	90.49	96.96	93.52
French	93.28	73.12	95.14	92.72
German	89.89	77.28	92.53	90.06
Greek	92.04	80.69	91.83	91.75
Hebrew	96.71	82.75	91.54	95.32
Hindi	94.18	93.62	92.96	94.88
Indonesian	90.91	83.55	95.28	92.85
Italian	89.21	75.43	88.87	88.87
Japanese*	75.56	78.10	91.21	75.64
Korean	95.04	85.46	92.92	94.14
Persian	93.81	87.28	95.29	92.98
Polish	90.40	86.41	89.15	90.81
Portuguese	92.69	91.17	90.96	92.69
Romanian	93.65	78.15	95.16	93.17
Russian	93.00	79.77	92.12	92.20
Spanish	91.30	72.87	93.17	91.88
Turkish	90.19	82.59	98.19	90.77
Ukrainian	87.77	73.69	97.57	90.69
Vietnamese	87.70	76.08	96.83	88.54

Table 16: Case wise accuracies over all languages for each generator : Aya-23

Language ↓	Partial cases	Unchanged cases	Rewritten cases	Overall
Arabic	98.82	91.63	95.75	97.18
Chinese*	86.51	75.93	86.18	87.75
Czech	99.80	80.78	91.56	97.55
Dutch	99.49	77.57	86.16	96.41
English	99.37	90.98	98.32	97.76
French	99.63	74.86	90.15	96.30
German	99.79	77.23	94.62	97.37
Greek	99.90	87.33	82.84	97.46
Hebrew	98.94	83.48	82.61	96.27
Hindi	98.72	92.35	95.48	97.23
Indonesian	99.55	88.19	94.11	98.05
Italian	99.97	81.43	93.49	97.48
Japanese*	98.33	87.97	91.08	97.02
Korean	99.40	84.22	94.45	96.93
Persian	97.99	89.54	90.00	94.54
Polish	99.60	88.75	85.69	97.62
Portuguese	99.17	90.82	82.19	96.52
Romanian	99.93	78.70	92.11	97.11
Russian	99.26	80.44	92.17	95.36
Spanish	99.31	71.65	93.24	95.91
Turkish	98.60	81.65	92.86	94.62
Ukrainian	99.27	73.52	91.46	93.99
Vietnamese	98.42	77.05	92.41	91.85

Table 17: Case wise accuracies over all languages for each generator : Claude-3.5-Haiku

Language ↓	Partial cases	Unchanged cases	Rewritten cases	Overall
Arabic	98.63	91.82	100.00	96.58
Chinese*	92.64	78.36	95.14	88.43
Czech	99.30	77.22	99.88	97.62
Dutch	99.30	77.53	99.52	97.30
English	99.35	90.02	99.76	98.03
French	99.53	73.66	99.97	97.06
German	99.52	76.06	99.69	97.33
Greek	99.00	80.60	99.62	95.83
Hebrew	97.68	82.69	99.88	96.46
Hindi	99.12	92.51	99.88	97.63
Indonesian	99.66	84.55	100.00	98.43
Italian	99.69	81.13	99.99	98.13
Japanese*	98.59	87.36	99.64	98.04
Korean	98.77	83.49	99.87	97.15
Persian	98.35	87.92	99.97	96.01
Polish	99.00	90.30	99.26	98.20
Portuguese	98.74	89.65	83.74	96.38
Romanian	99.18	80.36	99.71	97.46
Russian	99.33	80.55	99.93	94.55
Spanish	99.06	71.68	99.92	96.65
Turkish	98.45	83.13	99.96	95.32
Ukrainian	99.06	74.14	99.87	95.62
Vietnamese	98.10	77.40	99.92	88.87

Table 18: Case wise accuracies over all languages for each generator : Claude-3.5-Sonnet

Language ↓	Partial cases	Unchanged cases	Rewritten cases	Overall
Arabic	87.12	85.34	86	82
Chinese*	88.90	86.45	89	84
English	92.45	90.12	91	88
French	89.78	87.21	90	85
German	90.23	88.05	89	86
Italian	89.12	87.00	89	86
Japanese*	87.77	85.88	88	83
Korean	88.56	86.34	87	85
Portuguese	90.12	88.34	89	87
Spanish	90.45	88.12	89	87

Table 19: Case wise accuracies over all languages for each generator : Command-R-Plus

Language ↓	Partial cases	Unchanged cases	Rewritten cases	Overall
Arabic	95.74	91.26	96.31	95.01
Chinese*	92.63	77.87	92.51	86.36
Czech	93.30	80.96	91.67	91.87
Dutch	94.14	74.85	90.84	92.58
English	94.94	90.02	92.84	94.01
French	92.87	75.18	93.90	90.89
German	93.67	75.82	93.99	91.97
Greek	94.22	81.18	95.67	92.11
Hebrew	92.60	82.51	95.10	91.85
Hindi	96.56	92.44	96.95	96.16
Indonesian	95.08	84.88	95.88	94.70
Italian	93.35	79.95	92.72	92.72
Japanese*	93.98	88.44	94.19	93.84
Korean	94.44	84.84	93.09	93.61
Persian	94.83	88.32	94.68	93.34
Polish	94.53	89.51	89.36	93.73
Portuguese	95.50	88.58	85.07	93.93
Romanian	94.59	77.44	92.73	93.26
Russian	92.90	80.17	97.34	92.61
Spanish	93.54	69.64	91.87	92.13
Turkish	92.83	83.80	88.09	91.37
Ukrainian	91.69	74.93	96.81	90.33
Vietnamese	92.10	77.39	93.32	88.25

Table 20: Case wise accuracies over all languages for each generator : GPT-4o

Language ↓	Partial cases	Unchanged cases	Rewritten cases	Overall
Arabic	99.10	90.46	97.08	97.92
Chinese*	95.08	76.01	86.18	87.30
Czech	98.84	80.76	86.30	97.05
Dutch	98.80	77.47	89.03	97.12
English	99.07	88.92	94.25	96.91
French	98.77	76.17	91.74	96.53
German	98.92	76.66	89.69	97.12
Greek	98.87	81.60	85.68	97.05
Hebrew	98.97	83.94	97.33	98.10
Hindi	99.10	92.12	97.42	97.33
Indonesian	98.96	84.98	99.00	98.45
Italian	97.04	80.93	99.10	96.16
Japanese*	90.78	73.63	85.24	78.08
Korean	99.16	83.18	83.13	97.09
Persian	98.72	87.01	94.43	95.50
Polish	99.04	90.29	86.92	97.86
Portuguese	98.65	88.47	83.50	95.18
Romanian	98.77	76.50	98.37	97.57
Russian	98.98	78.13	87.02	95.23
Spanish	98.80	71.70	92.79	96.12
Turkish	98.94	82.37	87.33	96.55
Ukrainian	99.05	75.07	93.34	96.45
Vietnamese	98.29	78.87	91.75	92.24

Table 21: Case wise accuracies over all languages for each generator : GPT-o1

Language ↓	Partial cases	Unchanged cases	Rewritten cases	Overall
Arabic	93.90	88.40	98.59	94.86
Chinese*	89.29	75.98	94.92	85.09
Czech	91.43	78.82	97.15	92.04
Dutch	95.08	78.10	91.91	94.86
English	96.57	91.03	97.34	94.64
French	95.92	76.69	98.72	94.04
German	96.22	78.67	98.80	95.15
Greek	92.04	84.32	98.56	93.52
Hebrew	91.90	69.05	98.50	92.51
Hindi	95.55	93.52	98.72	95.66
Indonesian	96.84	83.45	98.55	95.92
Italian	94.70	76.47	97.03	94.89
Japanese*	84.59	87.53	96.26	87.86
Korean	94.54	85.07	99.41	94.66
Persian	95.68	89.47	96.54	94.66
Polish	93.51	88.07	97.34	94.60
Portuguese	95.22	88.75	94.90	94.28
Romanian	97.31	75.73	97.53	96.30
Russian	94.96	78.53	99.40	92.82
Spanish	95.30	74.56	99.47	94.09
Turkish	94.82	82.44	96.67	92.51
Ukrainian	89.37	73.96	98.39	91.68
Vietnamese	91.06	80.25	99.45	87.71

Table 22: Case wise accuracies over all languages for each generator : Gemini-1.5-Pro

Language ↓	Partial cases	Unchanged cases	Rewritten cases	Overall
Arabic	98.55	89.55	98.00	95.88
Chinese*	89.94	75.81	94.03	87.75
Czech	98.45	80.42	99.78	96.67
Dutch	98.22	78.00	99.39	95.86
English	97.92	90.02	99.39	95.26
French	98.10	73.66	99.94	95.59
German	98.71	74.75	99.58	96.74
Greek	98.96	85.16	98.90	98.09
Hebrew	97.64	82.90	99.71	96.82
Hindi	98.17	93.06	99.62	97.32
Indonesian	99.11	85.64	99.72	97.65
Italian	98.25	83.62	99.89	98.28
Japanese*	95.71	88.47	97.35	95.88
Korean	98.36	84.40	99.36	97.10
Persian	96.67	88.45	95.52	93.87
Polish	99.01	87.67	99.27	97.84
Portuguese	96.72	88.12	90.12	95.52
Romanian	99.84	77.30	99.75	98.49
Russian	97.62	81.30	99.07	94.29
Spanish	97.50	68.98	99.91	94.37
Turkish	97.69	84.99	99.36	95.47
Ukrainian	97.46	75.89	99.67	93.70
Vietnamese	96.29	79.35	99.81	91.12

Table 23: Case wise accuracies over all languages for each generator : Gemini-1.5-Flash

Language ↓	Partial cases	Unchanged cases	Rewritten cases	Overall
Arabic	96.53	92.00	99.40	95.68
Chinese*	95.05	76.30	97.63	88.07
Czech	96.99	78.78	94.73	94.98
Dutch	94.46	78.10	91.01	94.46
English	96.74	90.32	99.80	95.91
French	97.06	74.65	97.22	94.35
German	96.69	76.14	98.43	94.77
Greek	95.78	79.75	96.39	92.69
Hebrew	95.40	83.36	97.30	93.86
Hindi	96.25	92.00	99.35	95.29
Indonesian	96.67	83.15	96.32	95.55
Italian	97.34	82.59	99.30	96.01
Japanese*	97.05	87.46	94.90	96.12
Korean	96.65	82.64	97.39	95.02
Persian	94.75	89.34	98.55	92.34
Polish	96.69	87.13	93.69	95.36
Portuguese	96.37	88.20	93.68	94.69
Romanian	97.22	77.49	97.42	95.45
Russian	96.64	80.58	97.90	93.09
Spanish	95.54	69.92	98.82	92.76
Turkish	91.99	80.61	98.47	91.99
Ukrainian	96.21	74.58	97.16	93.20
Vietnamese	92.42	78.44	98.60	88.25

Table 24: Case wise accuracies over all languages for each generator : Mistral-Large-2411

Language ↓	Partial cases	Unchanged cases	Rewritten cases	Overall
English	97.10	91.08	99.71	96.64
French	95.53	72.58	99.49	93.94
German	94.98	77.21	99.50	94.29
Portuguese	92.66	89.06	98.17	94.03
Spanish	94.79	72.31	99.80	93.99

Table 25: Case wise accuracies over all languages for each generator : Perplexity-Sonar-Large