Token-Ensemble Text Generation: On Attacking the Automatic AI-Generated Text Detection

Anonymous ACL submission

Abstract

The robustness of AI-content detection models against sophisticated adversarial strategies, such as paraphrasing or word switching, is a rising concern in natural language generation (NLG) applications. This study proposes a novel token-ensemble generation strategy to challenge the robustness of current AI-content detection approaches by utilizing multiple sets of candidate generative large language models (LLMs). By randomly sampling token(s) from candidate language model sets, we find the token-ensemble approach significantly drops the performance of AI-content detection models. We evaluate the text quality produced under different token-ensemble settings based on annotations from hired human experts. We proposed a fine-tuned Llama2 model to distinguish the token-ensemble-generated text more accurately. Our findings underscore our proposed text generation approach's great potential in deceiving and improving detection models. This study's datasets, codes, and annotations are open-sourced [1].

1 Introduction

The pervasiveness of generative artificial intelligence (AI) has fundamentally reshaped information creation and dissemination approaches online. Powerful LLMs, like ChatGPT (OpenAI, 2022) and Llama 2 (Touvron et al., 2023), have accelerated this growth, blurring the lines between human-authored and machine-generated content (Sadasivan et al., 2023). While such technological advancement offers unprecedented opportunities and efficiency for natural language understanding and content creation (Gilardi et al., 2023; Qin et al., 2023), it comes with significant challenges and threats (Bang et al., 2023), particularly in misinformation dissemination (Huang et al., 2023), copyright violation (Karamolegkou et al., 2023), and decision trustworthiness (Choudhury and Shamszare, 2023). The capacity to accurately detect AI-generated content has become a crucial aspect of maintaining the integrity and reliability of information online.

The crucial role of detecting AI-generated content has spurred extensive research in this field. Existing works have primarily focused on developing AI content detectors, broadly categorized into supervised classifiers (Solaiman et al., 2019; Fagni et al., 2021; Mitrović et al., 2023) and zero-shot classifiers (Gehrmann et al., 2019; Mitchell et al., 2023; Su et al., 2023). These efforts are met with constant challenges in the form of adversarial methods, such as character substitution with homoglyphs, misspelling, paraphrasing, and word-switching (Wolff and Wolff, 2020; Sadasivan et al., 2023), which have proven effective to some degree. More recently, DetectGPT (Mitchell et al., 2023) and Fast-DetectGPT (Bao et al., 2023), which utilizes the concept of conditional probability curvature to highlight differences in word usage between large language models (LLMs) and humans in specific contexts, has achieved outstanding detection accuracy with significantly lowered the detection cost while also proving resilient against mainstream adversarial attacks.

To bring more insights into AI-generated text detection, we propose the token-ensemble text generation approach, which can be applied to attack neural text detectors. This approach manipulates the token selection process and alters the next-token probability distribution. Adversaries can effectively challenge the detection models' accuracy. Driven by the pressing need to explore and mitigate the vulnerabilities inherent in current AI-content detection methodologies (Sadasivan et al., 2023; Mitchell et al., 2023), we provide empirical evidence that our token-ensemble strategy could significantly affect the performance of AI-content detection models through exposing potential weak-

[1] https://anonymous.4open.science/r/token_ensemble-2462/

nesses in existing detection strategies. Our comprehensive performance benchmark and generation quality evaluation also introduce valuable insights for future research. By investigating the effects of token-ensemble generation on detection accuracies, we aim to shed light on the limitations of existing approaches and highlight the necessity for advanced detection technologies capable of countering sophisticated adversarial attacks.

The contributions of this paper are twofold. First, we provide empirical evidence of the significant impact our token-ensemble generation attack can have on the performance of AI-content detection models, indicating the potential weakness in current detection strategies. Secondly, we present comprehensive analyses and evaluations of our proposed approach, thereby providing insights for future research efforts to enhance the robustness of AI-content detection techniques against evolving adversarial attacks. The detection accuracy improvement depicts one effective pipeline to improve the performance of the existing AI-detection models, utilizing high-quality instance pairs generated through our token-ensemble approach.

We propose four research questions to thoroughly assess the effectiveness and limitations of our proposed token-ensemble strategy: (a) How significantly does the token-ensemble approach disrupt detection models? (b) How does the candidate language model selection influence the effectiveness of the token-ensemble approach? (c) How contextually coherent and fluent are the results of token-ensemble generation? (d) Can large language models benefit from the token-ensemble generation results on AI-generation detection?

2 Related Work

AI-Generated Content and Detection Models. Along with the advancements in content generation (Qin et al., 2023), efforts have been made to develop detection models capable of distinguishing human-written from AI-generated texts (Mitchell et al., 2023; Bao et al., 2023). Techniques leveraging word entropy analysis, machine learning classifiers, and next-token probability analysis have been explored (Kirchenbauer et al., 2023; Tang et al., 2023). For instance, Tang et al. (2023) have demonstrated using statistical methods and fine-tuned models to improve detection accuracy. However, these methods often struggle against straightforward manipulation strategies, such as paraphrasing attacks, highlighting the gap in the current detection capabilities (Sadasivan et al., 2023).

Adversarial Attacks on AI-Content Detection. The concept of adversarial attacks in AI content detection involves manipulating input textual information to deceive detection models into misclassifying AI-generated content as human-written (Sadasivan et al., 2023). Bao et al. (2023) and Mitchell et al. (2023) have shed light on the vulnerabilities of AI models to adversarial inputs, suggesting that even minor alterations can significantly impact model performance. Krishna et al. (2024) showcased the efficacy of paraphrasing attacks towards AI generation detection models and applicable retrieval-based solutions. Additional approaches like Liu et al. (2022) and Mao et al. (2024) also perform well, providing feasible solutions to detect AI-generated texts effectively.

3 Token-Ensemble Generation

Our proposed token-ensemble generation is a cultivated adversarial strategy designed to deceive AI-content detection models by exploiting their reliance on predicting the next-token distribution, as illustrated in Figure 1. When generating the next token based on the previous text, we randomly select one LLM from a pool of multiple LLMs and let it generate the next token(s). This process repeats until the text generation meets an ending condition. As a result, our token choices are created by shuffled probability distributions across candidate LLMs, creating a mixture of various token distribution predictions, unlike the general approach of only utilizing a single LLM to rephrase an existing AI-generated text.

We carefully set the completion criterion during the token-ensemble process (i.e., the whole text length reaches around 200 tokens after attaching the latest generated token(s)). In Bao et al. (2023), the setting for AI text generation is controlled by the total token length, more than 50 and less than 200 tokens, where the initial prompt includes the first 30 tokens of human-written text. However, as the average token lengths of data instances across all three datasets are around 170 tokens, we set our completion criteria as 'stop generating next token(s) when the content token length is greater than 170.' We also explored another completion criterion and found that it did not significantly change the performance results, as shown in Appendix D.

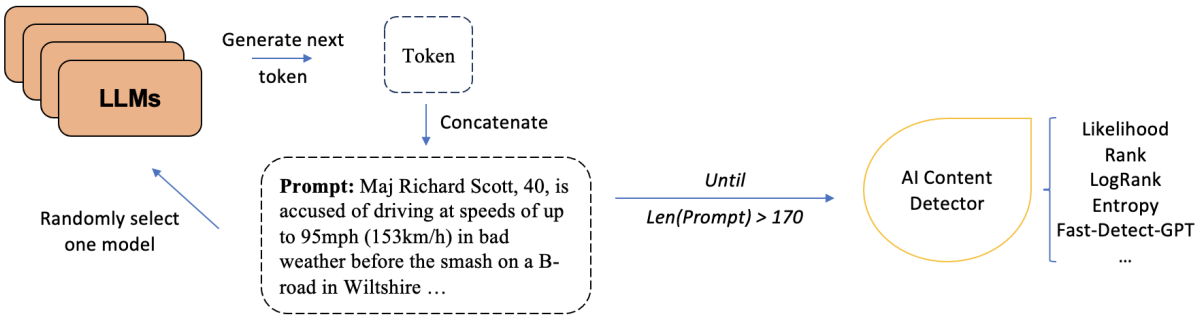

Figure 1: Pipeline illustration of token-ensemble generation attack.

4 Experiments

4.1 Dataset

To prepare the datasets for evaluating our approach, we start with three human-written text datasets from different domains. These include the XSum dataset (Narayan et al., 2018), which features news articles, the sQuAD dataset (Rajpurkar et al., 2016) based on Wikipedia documents, and the WritingPrompts dataset (Fan et al., 2018) containing story scripts. Collecting human-written texts from diverse public sources, these datasets serve as a comprehensive benchmark for detecting AI-generated content. All datasets are open-sourced, and our use complies with their intended purposes.

As in Bao et al. (2023), we select 300 instances from sQuAD and 500 instances each from the XSum and WritingPrompts datasets. This selection ensures a diverse representation of content types. We also confirm that the datasets do not include offensive content or sensitive information, such as individuals' names or other unique identifiers, ensuring ethical compliance in our research.

We further augment these datasets by incorporating AI-generated texts by various LLMs, like GPT-2 (Radford et al., 2019), as they are used to produce AI-generated texts in Bao et al. (2023). AI-generated texts serve as a baseline for evaluating our token-ensemble generation attack, establishing a comprehensive benchmark for distinguishing between human-written and AI-generated content.

4.2 Candidate Models for Token-ensemble

Our token-ensemble generation strategy concatenates the next token generated by a random candidate language model chosen from the designated sets. To make a fair comparison and illustrate the technical potential of our proposed token-ensemble approach, we collect eight language models and put them in two sets: one is smaller and relatively well explored, and another one is slightly bigger (no more than 10 billion parameters) and more advanced. In that case, we select the GPT-2 (gpt2-xl, with 1.5 billion parameters) (Radford et al., 2019), OPT (opt-2.7b) (Zhang et al., 2022), GPT-Neo (gpt-neo-2.7B) (Black et al., 2021), and GPT-J (gpt-j-6B) (Wang and Komatsuzaki, 2021) as the *classic LLMs* set; we select the Llama2 (llama-2-7b) (Touvron et al., 2023), Phi-2 (microsoft/phi-2, with 2.7 billion parameters) (GenAI, 2023), Mistral (mistralai/Mistral-7B-v0.3) (Infra, 2024), and Gemma (google/gemma-7b) (Google, 2024) as the *advanced LLMs* set.

4.3 Experimental Settings

We chose those eight open-sourced models, which are between 1 and 10 billion parameters, to make our proposed token-ensemble approach efficiently implemented on small GPU servers. A 40GB of GPU memory is enough to execute our proposed approach in both settings. If you want to optimize the generation speed using our proposed alternative options, 100GB of GPU memory would be enough to reach the average speed of 10 seconds per instance (around 170 tokens generated). The above selections are made to encompass a wide range of generative capabilities and styles, ensuring the robustness and generalizability of our findings. The detailed inference settings are listed in Appendix A.

We comprehensively examine the performance of our attack by testing different token lengths from 1 to 5 and a random number between them. Additionally, we include the setting of sentence-level ensembles, where each LLM generates an entire sentence instead of a single token. For this, we ask the model to generate the following 50 tokens given the previous text and get the first sentence from those generated 50 tokens. The motivation for dynamically selecting tokens is to create a shuffled distribution derived from the collective outputs

of all the LLMs, making AI-generated texts free from the signature of a certain language model while exploiting the LLMs' capability to generate human-like content. The example of prompt setting is listed in the Appendix B.

4.4 AI-Generated Text Detection

We use multiple AI-generated text detection models. For the traditional statistical methods, we adopt likelihood (average log probabilities), rank (average token ranks arranged by descending order on probabilities), LogRank (average log value of ranks), and entropy (average token entropy of the generated content) (Gehrmann et al., 2019; Solaiman et al., 2019; Ippolito et al., 2020). In addition to those methods, we use Fast-DetectGPT (Bao et al., 2023), which has achieved a higher speed and better AUROC score than DetectGPT (Mitchell et al., 2023). Fast-DetectGPT remains robust even after the paraphrasing attack (Sadasivan et al., 2023), ensuring its capability as a strong baseline for our proposed token-ensemble generation attack. We note that the detection approaches mentioned are all open-sourced.

4.5 AI-Generation Detection Evaluation

We select the detection accuracy in the area under the receiver operating characteristic (AUROC) as the evaluation metric to illustrate the performance of AI-generation detection approaches, which is feasible to showcase the detectors' performance on the whole spectrum of the thresholds (Bao et al., 2023). To interpret the AUROC scores, an AUROC score of 0.5 indicates a random level detection capability, and an AUROC score of 1.0 indicates a perfect level detection capability. The effectiveness of the token-ensemble attack against the detection models is quantified by comparing the AUROC score of detecting text generated by previously mentioned LLMs with detecting text generated by the token-ensemble method.

4.6 Token-Ensemble Generation Quality Evaluation

We randomly selected five instances for each dataset to construct a comprehensive analysis of the generation result of our proposed token-ensemble generation approach. We then hired three well-trained human experts (one male and two female from the institution) to annotate the quality of token-ensemble generation results in seven experimental settings and the baseline generated by the GPT-2 model. In the field of dialogue system and natural language generation evaluations, the coherence (Cervone et al., 2018; Ye et al., 2021) and fluency (Martindale and Carpuat, 2018; Kann et al., 2018) have long been perceived as key metrics to evaluate generation results.

5 RQ1: How Significantly Does the Token-Ensemble Approach Disrupt Detection Models?

Datasets	GPT-2	OPT	Neo	GPT-J
XSum	0.9922	0.9806	0.9881	0.9771
sQuAD	0.9990	0.9949	0.9956	0.9854
Writing	0.9982	0.9972	0.9981	0.9974
Avg.	0.9965	0.9909	0.9939	0.9866
Bao et al. (2023)	*0.9967*	*0.9908*	*0.9940*	*0.9866*

Table 1: Replication of the detection AUROC scores of Fast-DetectGPT on single language model generation with the same settings as Bao et al. (2023). The final row lists the original results from Bao et al. (2023)

We build the baseline by replicating Fast-DetectGPT's performance in detecting AI-generated content. We use each of the four LLMs (GPT-2, OPT, GPT-Neo, and GPT-J) to generate the AI content and use Fast-DetectGPT to detect them using the same settings. We successfully achieve the Fast-DetectGPT AUROC scores in our experimental settings as shown in Table 1. For each dataset, the average AUROC scores across the four models serve as the baseline in Table 2 when applying the classic LLMs set.

As shown in Table 2, our experiments demonstrate a notable reduction in the performance across nearly all traditional AI content detection methods (such as likelihood, rank, and LogRank) and the SoTA detection model, Fast-DetectGPT, when tasked with identifying texts generated through token-ensemble methods. This ensemble attack method performs better when it generates fewer tokens at each step before concatenating them to complete the text generation process. The significant drops in performance across different detection approaches underscore the effectiveness of the token-ensemble generation in exploiting the inherent weaknesses of current mainstream detection techniques. Surprisingly, compared to other metrics, the token-ensemble generation method increases the AUROC score for

Datasets	Detection Method	Baseline	TL=1	TL=2	TL=3	TL=4	TL=5	Rand.	Sent.
XSum	Likelihood	0.7837	0.3147	**0.2015**	0.2466	0.2664	0.2923	0.2295	0.4492
	Rank	0.8068	0.3787	**0.3625**	0.4018	0.4246	0.4520	0.3892	0.5797
	LogRank	0.8117	0.3877	**0.2827**	0.3307	0.3572	0.3792	0.3165	0.5191
	Entropy	**0.5300**	0.6983	0.8068	0.8209	0.8269	0.8106	0.8177	0.7435
	Fast-DetectGPT	0.9845	0.4573	**0.4431**	0.5653	0.6288	0.6406	0.5245	0.8062
sQuAD	Likelihood	0.7573	**0.2602**	0.2626	0.3237	0.3757	0.3783	0.2986	0.5493
	Rank	0.7836	**0.3684**	0.4212	0.4586	0.5216	0.4806	0.4474	0.6224
	LogRank	0.8090	**0.3657**	0.3877	0.4535	0.4955	0.4961	0.4232	0.6357
	Entropy	**0.5617**	0.7721	0.8149	0.8049	0.7801	0.7724	0.8025	0.7152
	Fast-DetectGPT	0.9937	**0.5068**	0.6035	0.7002	0.7565	0.7541	0.6810	0.9062
Writing	Likelihood	0.8905	0.7131	0.6780	0.6965	0.7027	0.7075	**0.6727**	0.7866
	Rank	0.8186	0.6542	**0.6458**	0.6702	0.6671	0.6725	0.6527	0.7145
	LogRank	0.9158	0.7728	**0.7490**	0.7650	0.7683	0.7705	0.7426	0.8305
	Entropy	**0.3752**	0.4357	0.5449	0.5857	0.5969	0.5934	0.5981	0.5135
	Fast-DetectGPT	0.9977	**0.7817**	0.8718	0.9253	0.9321	0.9364	0.8996	0.9581

Table 2: All AI content detection metric AUROC scores for the XSum, sQuAD, and Writing datasets, reported in various token-ensemble generation settings. Baseline scores come from the average score for each dataset in Table 1. Compared with the baseline AUROC score at each row, we highlighted the most deviated AUROC score in bold. TL is token length. Rand means that the token number is random between 1 and 5. Sent is the sentence-ensemble. Here, the candidate models are GPT-2, OPT, GPT-Neo, and GPT-J

all three datasets when evaluated using the entropy metric. This suggests that our approach makes the generated content more easily distinguishable from the perspective of token entropy distribution.

Furthermore, when using the Fast-DetectGPT method, the one or two-token-ensemble attack setting performs the best, decreasing the AUROC score from 0.9845 to 0.4431 for XSum, from 0.9937 to 0.5068 for sQuAD, and from 0.9977 to 0.7817 for WritingPrompts. Generally, the attack effectiveness decreases as the token number increases during ensemble generation. At the same time, for the entropy metric, the token-ensemble attack makes it more accurate to distinguish between human-written and AI-generated content, from 0.5300 to 0.8269 for XSum, from 0.5617 to 0.8149 for sQuAD, and from 0.3752 to 0.5981 for Writing.

6 RQ2: How Does the Candidate Language Model Selection Influence the Effectiveness of the Token-Ensemble?

The effectiveness of our proposed token-ensemble approach, as one adversarial attack towards AI-generation detection models, could be significantly influenced by the candidate language models used for generating tokens. This section explores how variations in selecting these models could affect the ability to deceive AI content detection systems. The experiment results using the advanced LLMs set are listed in Table 3. When testing on the Fast-DetectGPT method, the one or two-token-ensemble attack setting performs even better than in Table 2, decreasing the AUROC score from 0.9845 to 0.2191 for XSum, from 0.9937 to 0.2051 for sQuAD, and from 0.9977 to 0.5734 for WritingPrompts. The attack effectiveness decreases as the token number increases during ensemble generation. Similarly, at the same time, for the entropy metric, the token-ensemble attack makes it more accurate to distinguish between human-written and AI-generated content, from 0.5300 to 0.9339 for XSum, from 0.5617 to 0.9095 for sQuAD, and from 0.3752 to 0.6515 for Writing.

Comparing the AUROC scores among various settings and datasets from Table 2 and Table 3, we find that the general score distributions are the same. At the same time, the advanced LLMs set is more successful at deceiving the AI-generation detection approaches except the entropy method. The detection methods we investigated struggle more with text generated from the token-ensemble approach using advanced LLMs set as candidates, except for the entropy method. Surprisingly, the entropy detection approach performs much better for all our token-ensemble generation settings. For XSum and sQuAD datasets, the entropy detection

Datasets	Detection Method	Baseline	TL=1	TL=2	TL=3	TL=4	TL=5	Rand.	Sent.
XSum	Likelihood	0.7837	0.0977	**0.0485**	0.0809	0.0970	0.1331	0.0640	0.3023
	Rank	0.8068	**0.1980**	0.1998	0.2416	0.2720	0.2955	0.2138	0.4642
	LogRank	0.8117	0.1334	**0.0777**	0.1217	0.1417	0.1835	0.0973	0.3389
	Entropy	**0.5300**	0.8780	0.9339	0.9047	0.8842	0.8628	0.9237	0.7495
	Fast-DetectGPT	0.9845	0.2191	**0.2139**	0.2890	0.2985	0.3400	0.2576	0.5382
sQuAD	Likelihood	0.7573	**0.0638**	0.0960	0.1618	0.1965	0.2195	0.1465	0.4800
	Rank	0.7836	**0.1883**	0.2387	0.3009	0.3168	0.3333	0.2722	0.5286
	LogRank	0.8090	**0.0985**	0.1507	0.2365	0.2699	0.2954	0.2168	0.5360
	Entropy	**0.5617**	0.9095	0.8785	0.8249	0.8122	0.7891	0.8345	0.6290
	Fast-DetectGPT	0.9937	**0.2051**	0.2742	0.3656	0.4458	0.4531	0.3531	0.7046
Writing	Likelihood	0.8905	**0.4223**	0.4828	0.5295	0.5801	0.5781	0.4975	0.7126
	Rank	0.8186	0.5154	**0.5327**	0.5695	0.5972	0.5900	0.5390	0.6641
	LogRank	0.9158	0.4933	**0.5501**	0.5965	0.6414	0.6382	0.5658	0.7487
	Entropy	**0.3752**	0.6338	0.6515	0.6400	0.5947	0.6051	0.6388	0.5102
	Fast-DetectGPT	0.9977	**0.5734**	0.7134	0.7727	0.7916	0.7939	0.7170	0.8846

Table 3: All AI content detection metric AUROC scores for the XSum, sQuAD, and Writing datasets, reported in various token-ensemble generation settings. The difference is that the candidate models used here are Llama2, Phi-2, Mistral, and Gemma.

method works the best among all five detection methods; for the Writing dataset, the entropy approach performs very well except when we ensemble at the sentence level. The variation in performance is likely due to the different syntactic and semantic patterns these models introduce. Statistical analysis further supports the assumption that the more advanced the models in the ensemble, the lower the likelihood of accurate detection, especially for the current SOTA detection model like Fast-DetectGPT.

The above findings suggest that candidate model selection strategy could also play a critical role in the success of adversarial attacks aimed at evading the current AI content detectors. One complementary strategy is that the detection systems should better incorporate a variety of detection models, including the traditional ones like entropy-based detection methods, to avoid -adversarial attacks based on ensembling several generation results from multiple LLMs. Our findings showcased the strategic advantage of employing diverse detection methods for generating adversarial content, as it introduces a level of complexity that current detection models may not be fully equipped to handle.

7 RQ3: How Contextually Coherent and Fluent are the Results of the Token-Ensemble Generation?

One essential aspect of deploying adversarial strategies like the token-ensemble approach is ensuring that the generated text not only deceives detection systems but also retains good coherence and contextual fluency comparable to the level of human-written text. This section examines the text quality produced by our token-ensemble approach to assess its capability and limitations.

We employed two linguistic quality metrics, i.e., coherence and fluency, from human annotations to evaluate the quality of the generated text and its resemblance to the human-written level. We collected annotations from three human experts regarding *coherence* (the contextual information of the given short text should logically make sense) and *fluency* (the given short text should read naturally, mimicking the style and syntax of human natural language) to quantify the token-ensemble generation results from a scale of 1 to 7 (1 means very bad, 7 means very good quality). We randomly select five instances for each dataset and report the average score of five instances in different generation settings, as listed in Table 4.

Among the baseline generation from the GPT-2 model and token-ensemble generation in seven settings for each instance, we then asked the

Datasets	Candidate LLMs	Baseline	TL=1	TL=2	TL=3	TL=4	TL=5	Rand.	Sent.
XSum	Classic LLMs	**4.3/4.8**	3.9/3.4	3.4/3.6	4.0/4.2	2.9/3.3	3.7/4.2	3.5/3.4	3.4/4.0
	Advanced LLMs	**4.3/4.7**	2.0/1.5	3.9/3.4	3.3/3.1	3.7/3.6	4.2/3.7	3.6/3.7	3.9/4.3
sQuAD	Classic LLMs	5.2/**5.7**	4.0/3.8	5.1/5.3	4.1/4.9	4.1/4.6	4.3/4.9	**5.4**/5.3	4.7/5.6
	Advanced LLMs	**5.3/5.7**	3.1/3.5	3.5/3.9	4.3/4.9	4.7/5.0	3.7/4.3	4.8/4.8	4.7/5.6
Writing	Classic LLMs	**3.4/4.1**	2.8/3.2	2.4/2.9	2.9/2.9	**3.4**/3.9	2.8/2.7	2.8/2.5	2.8/3.4
	Advanced LLMs	**4.3/4.0**	3.4/3.4	4.2/**4.0**	3.4/3.1	3.3/3.9	2.2/2.5	3.3/2.8	3.2/3.2

Table 4: The average coherence/fluency scores for sampled XSum, sQuAD, and Writing sub-datasets, annotated by three hired human experts and recorded in various token-ensemble generation settings. The difference is that the candidate models used here are Llama2, Phi-2, Mistral, and Gemma.

Datasets	Fine-tune Status	Baseline	TL=1	TL=2	TL=3	TL=4	TL=5	Rand.	Sent.
XSum	Not fine-tuned	0	0.4	0.4	0.2	0.4	0.4	0.4	**0.6**
	After fine-tuning	0.2	0.6	0.2	0.6	0.6	0.6	**0.8**	0.2
sQuAD	Not fine-tuned	**0.6**	0.4	0	**0.6**	0	**0.6**	0.2	**0.6**
	After fine-tuning	**0.8**	0.6	0.4	0.6	**0.8**	0.6	**0.8**	0.4
Writing	Not fine-tuned	0.2	0	**0.4**	0	**0.4**	0.2	**0.4**	0.2
	After fine-tuning	0.4	0.6	0.2	0.8	0.6	**1.0**	0.6	0.4

Table 5: The accuracy of utilizing the Llama2 (llama-2-13b-chat) for the AI-generation detection task on the sampled XSum, sQuAD, and Writing sub-datasets before and after the specific fine-tuning process based on the selected high-quality machine-generated instances.

human experts to select the one they think has the highest probability of being written by a real human. Note that we disclose neither the source of the given short text nor our experimental settings before annotation. The annotation process takes 14 working hours, and we reimburse 280 US dollars. The detailed settings of human annotation collection and our instruction to the human experts are listed in Appendix C.

We can observe a high relevance between the coherence and fluency metric scores; the fluency score is mostly higher than the coherence score, indicating our proposed token ensemble approach is generating more fluent and less coherent text. As the baseline approach (generated by a single GPT-2 model) mostly receives the best coherence and fluency scores, our proposed token-ensemble approach remains of relatively good quality, especially when the token number increases during the ensemble process. It is also surprising to find that utilizing the advanced LLMs sets as the candidate LLMs does not ensure the generation quality improvement, though more difficult to be distinguished by the SOTA detection methods as shown in Table 2 and Table 3.

Both single language model generation and token-ensemble approach generate a medium text level regarding the rating range from 1 to 7. While some token-ensemble configurations produced text with high linguistic quality scores (like sentence-level and random token length ensemble), others displayed noticeable discrepancies in fluency and coherence. Annotations for most human-like generation for each instance (selecting from baseline and seven token-ensemble settings) are attached in Appendix C.

8 RQ4: Can LLMs Benefit from the Token-Ensemble Generation Results on AI-Generation Detection?

A critical challenge in AI-generated content detection is improving the accuracy and robustness of the detection model. In this section, we investigate whether fine-tuning the large language models using high-quality instance pairs generated from our token-ensemble approach can improve their ability to detect our generated high-quality texts through token-ensemble settings.

We implemented and fine-tuned the Llama2 model with 13 billion parameters (llama-2-13b-chat), utilizing the LoRA (Hu et al., 2021) during Fine-tuning for higher speed and lower GPU

memory consumption. The LoRA technique allows efficient fine-tuning of LLMs by adjusting only a small subset of model parameters. We filter the high-quality generation results from the 240 instances based on the quality annotation collected in Section 7. Only the instances with coherence and fluency scores equal to or bigger than 5 (from the original rating scale of 1 to 7) would be selected for the fine-tuning dataset. For this experiment, the instance pairs comprised outputs from token-ensemble configurations that were manually curated to ensure high quality in terms of coherence and contextual relevance. The detailed prompt and fine-tuning settings are attached in Appendix E.

The results in Table 5 showcase the significant improvement in five instances for each dataset and AI-generation settings (single GPT-2 generation and seven token ensemble scenarios) in detection accuracy post-fine-tuning in most cases (except the XSum and Writing when TL =2, and XSum and sQuAD when adopting sentence ensemble). The Llama2 model here, once fine-tuned by the specific downstream datasets, demonstrated its improved sensitivity towards subtle cues and patterns brought by our proposed token-ensemble approach, where the mainstream detection models do not distinguish well. The improved accuracy scores among various datasets and generation settings indicate the promising benefits of the detection method built upon the language models, which can effectively and efficiently learn the implicit nuance introduced by our proposed token-ensemble attack.

9 Robustness Analysis

Several trade-offs emerge while implementing our proposed token-ensemble approach, particularly regarding generation speed versus resource consumption and the balance between deception capability and the quality of the generated text. Those trade-offs highlight the complexities in designing effective and efficient adversarial AI strategies.

The efficiency of generating text through the token-ensemble method can vary significantly based on the computational resources consumed. For instance, generating 100 tokens that take up much memory space on a strong CPU server may take approximately 30 minutes. In contrast, the same process on an A100 GPU with 100GB of memory can produce 170 tokens in just 10 seconds. Our findings reveal another important trade-off between the strength of the deception achieved by the token-ensemble method and the coherence and fluency of the generated text: certain configurations (e.g., TL=1) of the token-ensemble approach were extremely effective in deceiving state-of-the-art LLM-based detection methods (i.e., Fast-DetectGPT); however, they often resulted in generated text that lacked coherence and fluency. One feasible solution is to find a balance where the deception is successful (e.g., a random guess success rate) when the quality of the generated text is comparable to other language model generation results (e.g., the filtered results from the GPT-2 model collected in Bao et al. (2023)).

10 Conclusion

This study presents a novel attack strategy for AI content detection using a token-ensemble approach, effectively challenging current detection models by leveraging multiple, including relatively smaller, mainstream LLMs. This strategy, inspired by the success of ensemble methods in machine learning for boosting predictive performance, proves that a coordinated attack using multiple, weaker models can more effectively bypass advanced AI content detection systems than a singular, more powerful model. Our findings highlight a significant advancement in the arms race between creating and detecting AI-generated content, providing fresh insights into improving detection capabilities.

By manipulating candidate selection from a diverse array of large language models, we demonstrated a substantial impact on the detection methods' ability to identify AI-generated content accurately. Our further investigations illustrate surprising findings that: (1) the token-ensemble approach could generate texts that are much harder for LLM-based detection methods to distinguish while maintaining comparable coherence and fluency quality compared with the GPT-2 model direct generation results; (2) different candidate LLMs would lead to very different qualities of generation results and the advancement of candidate models would not guarantee a better quality; (3) LLM-based AI-generation detection approach could benefit from simple fine-tuning to achieve better understanding on the implicit nuance introduced by our proposed token-ensemble approach. Future research could consider building large language model based detection systems specifically fine-tuned through specially designed datasets derived from adversarial attack tactics to improve the detection performance.

Ethical Statements

The author's Institutional Review Board has approved the experiment design of collecting human annotations, and the approval number will be disclosed in the camera-ready version. We provide all annotators with information on mental consultant hotlines and clinics, considering that the LLMs generation results might contain uncomfortable information like social bias. We suggest the annotators stop or quit the annotation process anytime if they feel necessary. The annotators are reimbursed based on their recorded working hours at a rate above the average salary requirement in the US.

Regarding the ethical concerns associated with AI content generation and detection, addressing the various dimensions of risk, fairness, privacy, and security issues is imperative. We want to outline the potential ethical considerations of our work, underscoring the drawbacks of misuse and possible negative consequences.

Our research, primarily technical explorations, opens trails to potentially harmful applications. The token-ensemble text generation method, which was practical and straightforward to deploy, could be easily adopted to deceive the current AI content detection services, which would raise concerns regarding the spread of disinformation or the creation of fake user profiles. Such risks highlight the importance of developing robust detection mechanisms to identify and mitigate adversarial attacks. Mitigation strategies might include the development of more sophisticated detection algorithms, implementing ethical guidelines for AI-generated content, or promoting transparency in AI deployments.

Regarding fairness, deploying technologies that leverage LLMs' deception capability could inadvertently amplify the misuse of LLMs on their inherent biases toward historically marginalized groups or minority groups. Our research methodology and applications should be carefully scrutinized to avoid bias issues, ensuring that the development and deployment of generative AI models in our experimental settings do not exacerbate social inequalities. Exploring adversarial attacks in AI-content detection applications could also involve privacy and security considerations. The inadequate practices could inadvertently facilitate malicious activities without scrutinizing the content.

Limitations

The findings of our study highlight a significant vulnerability in current AI-content detection models when faced with sophisticated adversarial attack strategies. Our proposed method has proven effective in degrading the performance of SOTA detection approaches. The ability of ensemble-generated texts to deceive detection underscores the complexity of distinguishing between human and AI-generated content, which the evolving capabilities of generative AI technologies would magnify. As AI technologies advance, the potential for misuse in spreading misinformation or reinforcing social bias through generated deceptive content could keep increasing. Though the special fine-tuned LLM could perform better towards specific tasks and datasets, that pipeline may not work for other scenarios. Further investigations on the efficacy and robustness of fine-tuning LLM are expected.

While our study provides valuable insights, it is important to acknowledge the limitations of our work. The scope of our experiments was constrained by the selection of finite LLM candidates and finite detection methods from a wide range of options. We selected the most mainstream LLMs and detection methods, while a more comprehensive benchmark would be better to illustrate the capability and limitations of our proposed token-ensemble generation approach. Moreover, the datasets used in our experiments may not fully capture the diversity of human and AI-generated content encountered in real-world scenarios. Including more varied and nuanced datasets could improve the analysis of the effectiveness of our proposed token-ensemble attack. Finally, exploring alternative adversarial strategies and their countermeasures can provide a broader perspective on the race between AI content generation and detection.

References

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

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*.

Sid Black, Gao Leo, Phil Wang, Connor Leahy, and Stella Biderman. 2021. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow. If you use this software, please cite it using these metadata.

Alessandra Cervone, Evgeny Stepanov, and Giuseppe Riccardi. 2018. Coherence models for dialogue. *arXiv preprint arXiv:1806.08044*.

Avishek Choudhury and Hamid Shamszare. 2023. Investigating the impact of user trust on the adoption and use of chatgpt: Survey analysis. *Journal of Medical Internet Research*, 25:e47184.

Tiziano Fagni, Fabrizio Falchi, Margherita Gambini, Antonio Martella, and Maurizio Tesconi. 2021. Tweepfake: About detecting deepfake tweets. *Plos one*, 16(5):e0251415.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

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

Microsoft GenAI. 2023. Phi-2: The surprising power of small language models. Microsoft Research Blog.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd-workers for text-annotation tasks. *arXiv preprint arXiv:2303.15056*.

Google. 2024. Gemma: Introducing new state-of-the-art open models. Google Developers Blog.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Fan Huang, Haewoon Kwak, and Jisun An. 2023. Is chatgpt better than human annotators? potential and limitations of chatgpt in explaining implicit hate speech. In *Companion Proceedings of the ACM Web Conference 2023*, WWW '23 Companion, page 294–297, New York, NY, USA. Association for Computing Machinery.

Fan Huang, Haewoon Kwak, Kunwoo Park, and Jisun An. 2024. Chatgpt rates natural language explanation quality like humans: But on which scales? *arXiv preprint arXiv:2403.17368*.

Deep Infra. 2024. Huggingface: Mistral-7b-v0.3.

Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. 2020. Automatic detection of generated text is easiest when humans are fooled. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1808–1822, Online. Association for Computational Linguistics.

Katharina Kann, Sascha Rothe, and Katja Filippova. 2018. Sentence-level fluency evaluation: References help, but can be spared! In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 313–323, Brussels, Belgium. Association for Computational Linguistics.

Antonia Karamolegkou, Jiaang Li, Li Zhou, and Anders Søgaard. 2023. Copyright violations and large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7403–7412, Singapore. Association for Computational Linguistics.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. 2023. A watermark for large language models. *arXiv preprint arXiv:2301.10226*.

Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. 2024. Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense. *Advances in Neural Information Processing Systems*, 36.

Xiaoming Liu, Zhaohan Zhang, Yichen Wang, Hang Pu, Yu Lan, and Chao Shen. 2022. Coco: Coherence-enhanced machine-generated text detection under data limitation with contrastive learning. *arXiv preprint arXiv:2212.10341*.

Chengzhi Mao, Carl Vondrick, Hao Wang, and Junfeng Yang. 2024. Raidar: generative ai detection via rewriting. *arXiv preprint arXiv:2401.12970*.

Marianna J Martindale and Marine Carpuat. 2018. Fluency over adequacy: A pilot study in measuring user trust in imperfect mt. *arXiv preprint arXiv:1802.06041*.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature. *arXiv preprint arXiv:2301.11305*.

Sandra Mitrović, Davide Andreoletti, and Omran Ayoub. 2023. Chatgpt or human? detect and explain. explaining decisions of machine learning model for detecting short chatgpt-generated text. *arXiv preprint arXiv:2301.13852*.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! Topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium.

OpenAI. 2022. Introducing chatgpt. OpenAI Blog.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is chatgpt a general-purpose natural language processing task solver? *arXiv preprint arXiv:2302.06476*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. 2023. Can ai-generated text be reliably detected? *arXiv preprint arXiv:2303.11156*.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, et al. 2019. Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203*.

Jinyan Su, Terry Yue Zhuo, Di Wang, and Preslav Nakov. 2023. Detectllm: Leveraging log rank information for zero-shot detection of machine-generated text. *arXiv preprint arXiv:2306.05540*.

Ruixiang Tang, Yu-Neng Chuang, and Xia Hu. 2023. The science of detecting llm-generated texts. *arXiv preprint arXiv:2303.07205*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. https://github.com/kingoflolz/mesh-transformer-jax.

Max Wolff and Stuart Wolff. 2020. Attacking neural text detectors. *arXiv preprint arXiv:2002.11768*.

Zheng Ye, Liucun Lu, Lishan Huang, Liang Lin, and Xiaodan Liang. 2021. Towards quantifiable dialogue coherence evaluation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2718–2729, Online. Association for Computational Linguistics.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. Opt: Open pre-trained transformer language models.

A Generation Settings

Our token-ensemble generation approach does not necessarily require GPU resources. As we tested on the CPU server, the one-token ensemble generation setting would need around 30 minutes to generate 100 tokens without specific speed optimization. However, using the A100 GPU server to accelerate the generation speed would only take approximately 10 seconds to generate 170 tokens in all token-ensemble settings, which would take up to 100GB of the GPU memory usage two A100 GPU 80GB GPU cards. We completed the experiments using the GPU server provided by Nvidia, under the support of the grant [anonymized]. (Detailed grant information will be revealed in the camera-ready version.)

B Token-Ensemble Prompt Setting

We prompt the randomly selected generation LLM with the first 30 tokens of the human-written original sentence. For example, the prompt of the first instance in the XSum Dataset is:

> Maj Richard Scott, 40, is accused of driving at speeds of up to 95mph (153km/h) in bad weather before the smash on a B-road in Wiltshire

C Token-Ensemble Generation Quality Annotation

In addition to the human annotations, we also tried to ask the ChatGPT (version 3.5 and 4) to provide the annotation regarding the ChatGPT's capability to understand the numerical scale and evaluate text quality similar to human performance (Huang et al., 2024). We use the exact instructions we give to human experts as the prompt of the ChatGPT input. However, after the manual inspection of the ChatGPT annotations, we find only ChatGPT-4 could understand the task and give out annotations in the format we requested. Still, further inspection showcased that around 20% of the annotation scores are significantly contrary to human expert annotations. Thus, we do not include annotations from language models in this work.

The most human-like generation results, voted by three human annotators for each instance, are listed in Table 7. The instructions we provide to human experts are listed in Table 8.

D Robustness Test

In our original settings, to fully replicate settings in Bao et al. (2023), we adopt the completion criteria of exceeding 170 tokens in our token-ensemble generation attack since the average length of human-written content in Bao et al. (2023) is around 170. We attached the robustness test results for our token-ensemble generation attack under a different completion criterion of exceeding 100 tokens ended by a period or exceeding 150 tokens for each instance, as shown in Table 6. The new completion criterion resulted in a more significant AUROC score drop in our token-ensemble attack. Thus, we believe that our efforts to make the AI-generated and human-written text similar in length, around 170, is required for a fairer comparison.

E Llama2 QA and Fine-tune Setting

We prompt the Llama2 model with the prompt design below to collect AI-generation text detection classification results:

> Please answer whether the given short text is generated by Artificial Intelligence models but not written from real human. Please answer by Yes, No or Uncertain. And then explain why in shortly in one or two sentences.
>
> The short text is: *Generation results.*

To avoid overfitting the language model in the fine-tuning process, we created the dataset with 37 annotated high-quality AI-generated texts with 37 human-written texts with the same source distribution (28 from sQuAD, 7 from XSum, and 2 from Writing). We adopt the template below to utilize the human-written and high-quality AI-generated texts:

> ### Question: Please answer whether the given short text is generated by Artificial Intelligence models but not written from real human. Please answer by Yes, No or Uncertain. And then explain why shortly in one or two sentences.
>
> The short text is:*instance text*.
>
> ### Answer:*instance label*. Yes means the short text is more likely to be generated by AI models but not written by real human. No means the contrary.

Datasets	Detection Method	Baseline	TL=1	TL=2	TL=3	TL=4	TL=5	Rand.	Sent.
XSum	Likelihood	0.7837	0.2176	0.1543	0.1988	0.2383	0.2537	**0.1883**	0.4087
	Rank	0.8068	**0.3073**	0.3548	0.3970	0.4108	0.4144	0.3844	0.5610
	LogRank	0.8117	0.2719	**0.2254**	0.2811	0.3221	0.3314	0.2574	0.4725
	Entropy	0.5300	0.7623	0.8467	0.8602	0.8438	0.8258	0.8489	0.7523
	Fast-DetectGPT	0.9845	**0.3187**	0.4053	0.5363	0.5833	0.5897	0.4971	0.7559
sQuAD	Likelihood	0.7573	**0.1812**	0.2722	0.2949	0.3208	0.3775	0.2684	0.5305
	Rank	0.7836	**0.3117**	0.3825	0.4294	0.4648	0.4908	0.4089	0.5975
	LogRank	0.8090	**0.2523**	0.3341	0.4101	0.4294	0.4906	0.3829	0.6053
	Entropy	0.5617	0.8262	0.8306	0.7945	0.8017	0.7707	0.8107	0.7181
	Fast-DetectGPT	0.9937	**0.3986**	0.5583	0.6401	0.6905	0.7433	0.6104	0.8974
Writing	Likelihood	0.8905	**0.5637**	0.6017	0.6315	0.6612	0.6786	0.6287	0.7361
	Rank	0.8186	**0.5737**	0.6140	0.6328	0.6389	0.6449	0.6402	0.6873
	LogRank	0.9158	**0.6263**	0.6748	0.7065	0.7269	0.7418	0.7030	0.7843
	Entropy	0.3752	0.5694	0.6177	0.6477	0.6301	0.6118	0.6221	0.5740
	Fast-DetectGPT	0.9977	**0.7168**	0.8378	0.8965	0.9105	0.9106	0.8680	0.9430

Table 6: All AI content detection metric AUROC scores for the XSum, sQuAD, and Writing datasets, reported in various token-ensemble generation settings. Baseline scores come from the average score for each dataset in Table 1. Compared with the baseline AURCO score at each row, we highlighted the most deviated AUROC score in bold. TL is token length. Rand means that the token number is random between 1 and 5. Sent is the sentence-ensemble.

Datasets	Candidate LLMs	Baseline	TL=1	TL=2	TL=3	TL=4	TL=5	Rand.	Sent.
XSum	Classic LLMs	2	3	1	4	1	1	2	1
	Advanced LLMs	4	-	1	1	2	3	-	4
sQuAD	Classic LLMs	3	-	3	1	-	1	5	2
	Advanced LLMs	4	1	-	-	3	2	2	3
Writing	Classic LLMs	6	-	3	-	5	-	-	1
	Advanced LLMs	4	1	1	2	4	-	2	1

Table 7: The vote counts the most human-like generated text from human expert annotations for five instances from each of the three datasets.

As for the hyper-parameter setting for the fine-tuning process, we fine-tuned the Llama2 model at one GPU server containing eight A100 80GB GPUs. We set the lora_alpha value as 16 and the lora_dropout as 0.1 for LoRA; we set the optimizer as pages_adamw_32bit, learning rate as 0.0002, and weight decay as 0.001 for 5 epochs.

Instructions for each instance, you should:

1. Score 1 7 (1 means very bad, 7 means very good) on the coherence score and fluence score for the column of 'a', 'b', 'c', 'd', 'e', 'f', 'g', 'h'.
* Coherence: The contextual information of the given short text (around 200 words) should logically make sense, i.e., maintaining topic consistency and logical sequence. * Fluency: The text should read naturally, mimicking the style and syntax of human natural language.
2. From the column of 'a', 'b', 'c', 'd', 'e', 'f', 'g', 'h', select the best one that you think has the highest probability that writes by a real human and put the number in the column of 'best' (Normally, the best one should be the one that possesses the highest coherence and fluency scores you annotated in the previous step).

Do note that:
1. We have 6 files and 5 instances each. Completing all labeling, for one instance, should take less than 6 mins. But feel free to take more time if necessary.
2. All texts selected are only a slice of 200 tokens of their original source, so please do not consider the potential incomplete sentence at the end of the text as one of your scoring criteria.

Examples:
1. [As Muslim institutions of higher learning , the madrasa had the legal designation of waqf . In central and eastern Islamic lands , the view that the madrasa , as a religious trust for pious educational endeavors , is the institutional and social prec ursory for the mosque (Dar ul -Kh air -) is prevalent (Ibd ah 198 5 ; N. A hmad 198 7) . The madrasa served as a place of religious instruction and a center for m ak tab (schools) . It provided lodging , board , and medical care for the instructors and m ustaf a (students) . The madrasa was both a place of instruction for religious m at terial and a living environment . The madrasa therefore served two purposes : the disse mination and perpet uation of the teachings of the Islamic faith and the pro p aga tion of Islamic culture and heritage . The mad rasa was the first step in the path to higher education] [Coherence: 4, Fluency: 5]
2. [Boston has a continental climate with some maritime influence , and using the -3 C (27 F) coldest month (January) isotherm , the city lies within USDA hardiness zone 5 b , with an average annual minimum temperature of around -2 .2 C (27 F) . Bostons climate is compar atively warm for the latitude due to its location within the Northeastern United States . Boston is often identified as a coastal city , and experiences regular and strong effects of maritime climate . However , since Boston is far from the most eastern coastline of the state, its climate has little maritime influence . The effects of the ocean can be seen in the average rainfall rate (around 4 3 inches or 1 09 centim et ers of snow per year) , a rain shadow that prevents heavy rainfall from accumulating in the summer and a generally war mer average annual temperate , but the city is rarely affected by extreme cold ,] [Coherence: 6, Fluency: 7]
3. [South Korea : The event was held in Seoul , which hosted the 1988 Summer Olympics , on April 27 . Intended torchbearers Choi Seung-kook and Park Won-sun boycotted the games . To demonstrate their disapproval, many South Koreans wore black rib ons to mourn the massacre. http : //en . w ikinews . org . 30 Nov . 2 0 1 6 . Korea , South - 19 88 Summer Olympics . This website is not to be accredited with any additional information about this event . The images were taken and used in this website by the users . http : // www . z iman . co . kr . h o d oj 69 0 . 46 . 01 0 . ↑ h t t p : //www.al jazeera . co m / program /ins igh t a l/ asia pacif is m / 2 0 1 7/ 0 5 / 01 3 / 8 4 6 5 492 . 27 0162 2 < eos >.] [Coherence: 1, Fluency: 2]

Table 8: Instruction information we give to our hired human experts to annotate the quality scores for given short texts and select the most human-like generation result from eight candidates for each instance.