

Uncovering LLM-Generated Code: A Zero-Shot Synthetic Code Detector via Code Rewriting

Tong Ye*

Zhejiang University
Hangzhou, China
tongye@zju.edu.cn

Yangkai Du*

Zhejiang University
Hangzhou, China
yangkaidu@zju.edu.cn

Tengfei Ma

Stony Brook University
New York, USA
tengfei.ma@stonybrook.edu

Lingfei Wu

Anytime.AI
New York, USA

lwu@anytime-ai.com

Xuhong Zhang

Zhejiang University
Hangzhou, China
zhangxuhong@zju.edu.cn

Shouling Ji

Zhejiang University
Hangzhou, China
sji@zju.edu.cn

Wenhai Wang

Zhejiang University
Hangzhou, China
zdzzlab@zju.edu.cn

ABSTRACT

Large Language Models (LLMs) have exhibited remarkable proficiency in generating code. However, the misuse of LLM-generated (synthetic) code has prompted concerns within both educational and industrial domains, highlighting the imperative need for the development of synthetic code detectors. Existing methods for detecting LLM-generated content are primarily tailored for general text and often struggle with code content due to the distinct grammatical structure of programming languages and massive "low-entropy" tokens. Building upon this, our work proposes a novel zero-shot synthetic code detector based on the similarity between the code and its rewritten variants. Our method relies on the intuition that the differences between the LLM-rewritten and original codes tend to be smaller when the original code is synthetic. We utilize self-supervised contrastive learning to train a code similarity model and assess our approach on two synthetic code detection benchmarks. Our results demonstrate a significant improvement over current synthetic content detectors designed for general texts, achieving a 20.5% increase in AUROC on the APPS benchmark and a 29.1% increase on the MBPP benchmark.

KEYWORDS

Synthetic Code Detector, Code Rewriting, Code Large Language Models

1 INTRODUCTION

Large Language models (LLMs) pre-trained on code have shown remarkable capability in understanding and generating code [5, 9, 22, 29]. Those code LLMs can function as professional coding assistants for programmers, offering intelligent code completion and document generation capabilities, such as Github Copilot¹. Furthermore, recent advancements in instruction tuning and AI alignment research [34, 47] have facilitated the development of general-purpose conversational LLMs [6, 31, 32, 41]. GPT-4, one

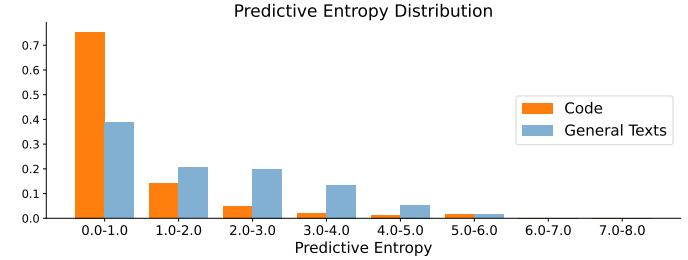


Figure 1: Token Entropy Distribution of Natural Texts and Code. We utilize 500 documents from XSum [28] dataset for token entropy estimation and 770 human-written code samples for code token entropy estimation.

of the most impressive conversational LLM released by OpenAI, can provide high-quality responses to general human requests in a zero-shot manner, including requests for generating code implementation for detailed coding specifications [4].

The breakthrough of LLM has dramatically improved the efficiency of the coding process and significantly lowered the barrier to entry for programming. However, it also raises concerns about the misuse of LLM-generated code. In programming education, students have already used LLMs to write solutions for coding assignments and exams [7, 20]. A recent study has shown that GPT-4 can achieve the average human-level performance in solving LeetCode problems [4]. This makes cheating with the use of LLMs in programming exams very attractive for students and also makes it difficult to assess students' programming abilities in education. Besides the educational domain, there is a significant demand for code security in real industrial applications, and LLM-generated code often contains security vulnerabilities [16]. An evaluation [35] revealed that in various security-relevant scenarios, 40% of Copilot-generated programs contain dangerous vulnerabilities. Therefore, if the code is synthetic, it requires an even more rigorous review.

*Both authors contributed equally to this research.

¹<https://copilot.github.com>

Building upon the aforementioned critical real-world scenarios, developing a synthetic code detector is crucial for addressing concerns regarding the misuse of LLM-generated code. Although there have been some efforts to detect LLM-generated pure text [11, 19, 27, 36, 51, 53], none of them have specifically focused on code content. The effectiveness of existing detection methods for general texts when applied to code content remains unknown. Unfortunately, our experiments, detailed in Section 5.1, reveal that even state-of-the-art detection methods, such as GLTR [11] and DetectGPT [27], exhibit a significant decline (approximately 32%) in detection performance when applied to the code domain.

To analyze the reasons for the failure, our meticulous experiments reveal that the primary issue lies in the core logic of existing state-of-the-art methods for text detection, which rely on the statistical log probability of tokens. LLMs assign higher log probabilities to the tokens they generate due to their high confidence in these self-generated tokens. Consequently, existing detection methods leverage this characteristic to differentiate between model-generated text and human-written text, achieving high detection accuracy. However, in the domain of code, this approach exposes its inherent weaknesses. The main reason is that, compared to natural language, code exhibits a more uniform grammatical structure. In specific programming languages and the given code context, the subsequent code token is often deterministic or has limited options. For example, in Python, functions always start with "def", and a ":" must follow the function signature "def *funciton_name()*". We refer to these tokens as "low-entropy" tokens. However, there exists a large number of such "low-entropy" tokens in code domain, greatly reducing the actual effective tokens for SOTA text detection methods to differentiate between human-written code and model-generated code, resulting in statistically insignificant differences between the two. We investigate the code token entropy distribution and the token entropy distribution (both estimated by Llama-13B) in Figure 1. The distribution of text token entropy is more dispersed than code, and more than 70% of code tokens have entropy smaller than 1.

Therefore, to address the detection challenge between human-written code and synthetic code, it is essential to abandon the current token log probabilities perspective. Adopting a more holistic, global view, on the other hand, opens up our thinking. Through a global and essential observation, we found that large language model, when tasked with rewriting synthetic code produced by either themselves or other LLMs, tend to exhibit remarkably similar coding patterns in their rewritten output compared to the original code. Conversely, when tasked with rewriting human-written code, the LLM-rewritten code diverges more significantly from the original code. As shown in Figure 2, the top left corner shows the original code (synthetic), which, after being rewritten by LLMs, yields the code in the top right corner. It's evident that the two are strikingly similar, with matching portions highlighted in orange. Conversely, the bottom left corner showcases human-written code. After undergoing the same LLM rewriting process, the LLM rewriting code in the bottom right corner shows a substantial number of differences compared to the original human-written code.

This observation inspired our hypothesis: the similarity between the original and rewritten code can be used as an indicator for detecting synthetic code. Furthermore, the generation of rewritten

code and the utilization of similarity as an indicator allows detection methods to be independent of statistically token likelihood estimation, thereby addressing the root cause of failure in current text detector for the code context. Based on this hypothesis, we propose a zero-shot synthetic code detector using Code Rewriting and Similarity Measurement.

Our method comprises two crucial steps: Code Rewriting and Similarity Measurement. When confronted with a piece of code, whether synthetic or human-written, we first employ a large language model to rewrite it using an appropriate prompt template. This process yields a pair consisting of the original code and its corresponding rewritten version. To accurately gauge the similarity between the original and rewritten code, we then develop a code similarity model that predicts a similarity score for the given code pair. Finally, using the cosine similarity metric, we estimate the expected similarity score by sampling m rewritten codes following the identical procedure and averaging their scores.

Due to the lack of synthetic code detection datasets, we construct two synthetic code detection datasets utilizing APPS [17] and MBPP [2], representing use cases in the framework of synthetic code exploitation. To confirm that our hypothesis and method are universally applicable to all types of code generation tools, we generate synthetic code using two open-sourced LLMs (CodeLlama [38], StarChat [22]) and two proprietary code generation APIs (GPT-3.5 [31], GPT-4 [32]) for detection. The experimental results demonstrate that our proposed method notably outperforms current SOTA methods. Specifically, our approach achieves a 20.5% improvement in detection performance (AUROC) on the APPS dataset and a 29.1% improvement on the MBPP dataset. Furthermore, from a practicality and minimal resource/permission requirement standpoint, our method only necessitates the capability to perform LLM inference or access to their APIs. This is in contrast to the previous detection methods that require knowledge of token log probabilities.

To facilitate further research, we have made our dataset, code, and trained code similarity model checkpoint publicly available². Overall, the main contributions can be outlined as follows:

- We identify a performance gap when applying synthetic content detectors designed for general texts to the code domain. Through detailed analysis, we uncovered the reasons for this discrepancy.
- We propose a novel detection approach that utilizes code rewriting and similarity measurement, effectively addressing the unique challenges of synthetic code detection.
- Our proposed zero-shot synthetic code detector is applicable to both open-sourced code LLMs and closed-source LLMs like ChatGPT/GPT-4, which solely offer APIs.
- Extensive experiments and analysis demonstrate the effectiveness of our method, revealing significant improvements in accuracy and robustness compared to existing methods.

2 RELATED WORK

2.1 LLMs for Code Generation

With the rapid rise in the popularity of LLMs, an increasing number of studies have begun focusing on Code LLMs aimed at achieving

²<https://anonymous.4open.science/r/code-detection-6B35>

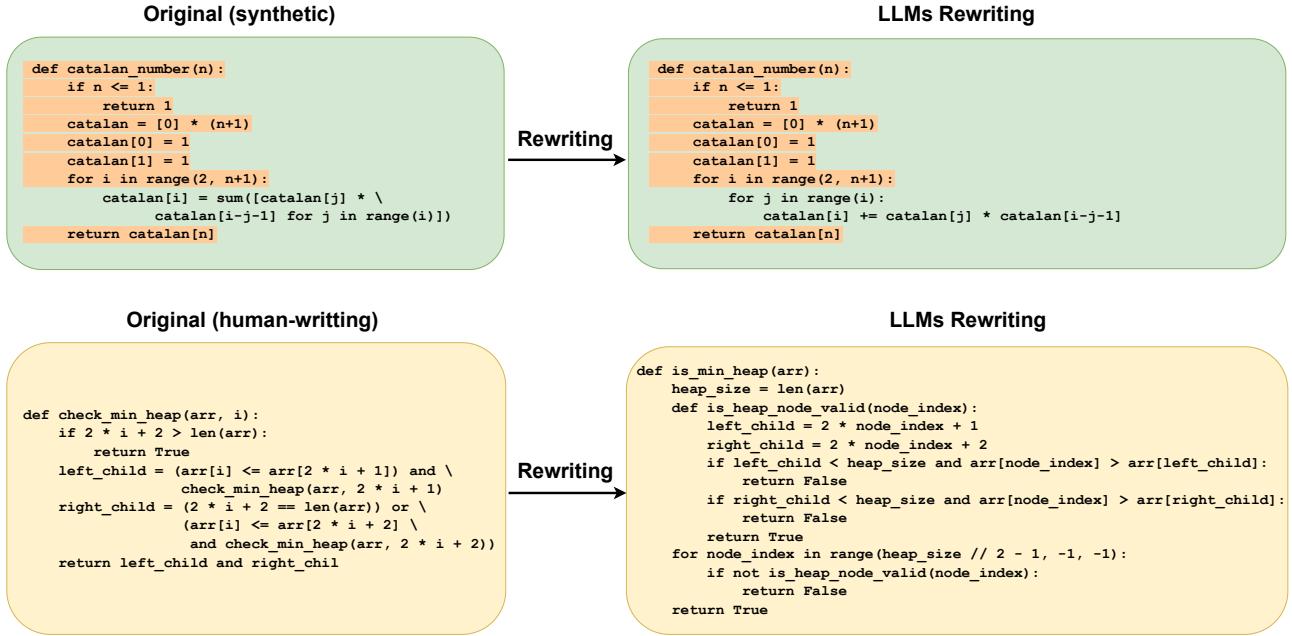


Figure 2: Large Language Models can only make minor modifications to the synthetic code. We use the StarChat model [22] to rewrite the original code and observe that the similarity between the rewritten and original code is much higher when the original code is generated from GPT-3.5-Turbo.

automatic software engineering. Those Code LLMs are continually pre-trained on large-scale code corpus from an initial pre-trained model on text with auto-denoising [46] or causal language modeling task [5, 9, 22, 29]. A crucial capability of Code LLMs is code generation. Code generation is a left-to-right decoding process that utilizes functional requirements or code context as a prompt. Pioneering works such as CodeX [5], AlphaCode [24], CodeGeex [52], SantaCoder [1], and GPT-4 [33] have all demonstrated powerful code generation capabilities. Recent advances in general purpose LLMs [6, 31, 32, 41] further enhanced their understanding of human instructions, resulting in the generation of code that better meets the requirements and exhibits higher quality.

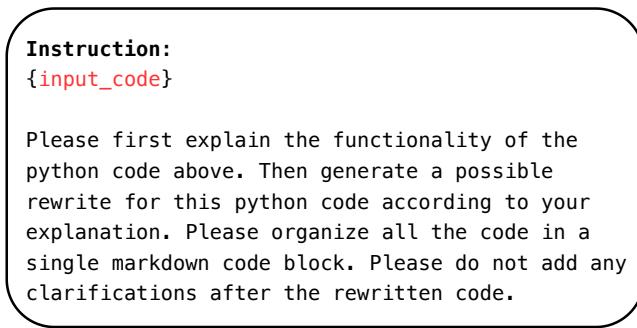
As code large language models continue to mature, more and more people from various industries are integrating these models into their daily lives. According to a community survey led by StackOverflow [39], 44% of experienced developers and 55% of beginners already used AI coding assistant and most of them use two tools, ChatGPT (83%) and Github Copilot (56%). Therefore, in certain areas such as education and industrial settings, automated detection of whether code is written by humans or generated by models has become increasingly important.

2.2 Detection of Synthetic Text

Detecting AI-generated (synthetic) text has been studied before the emergence of LLM. The main works on synthetic text detection follow two lines of research. One research line formulates detection as a binary classification problem by collecting synthetic texts from generative models and training a supervised model based

on pre-trained transformers [19, 51, 53] or other neural models [3, 43]. Another research line detects synthetic text by designing zero-shot metrics. These metrics measure the relationship between a given text and the text distribution of generative models. Gehrmann et al. [11] claim that synthetic texts are sampled from the head of generative models’ distribution, so the average log probability score under the generative model \mathcal{G} of a given text, i.e., $\frac{1}{L} \sum_{i=1}^L \log(P_{\mathcal{G}}(x_i|x_1, \dots, x_{i-1}))$, can be a simple and effective zero-shot metric (called GLTR). Su et al. [40] proposed two metrics, LRR and NPR, where the former combines log-rand and log-likelihood to better magnify the differences between human-written and machine-generated text, while the latter is primarily based on the idea that the log-rank of machine-generated texts should be more sensitive to smaller perturbations. Mitchell et al. [27] further improved the GLTR metrics by proposing that synthetic texts tend to occupy negative curvature regions of the model’s log probability function, i.e., the local maximum of the generative model’s distribution. They use the average token probability disparity between the given text and perturbed texts to detect whether the given text is located at the negative curvature regions of $\log p(x)$. Overall, all of these methods converge on the same idea: LLMs tend to generate tokens with higher confidence. For tokens generated by LLMs themselves, there will be assigned a higher log probability. Therefore, effective differentiation can be achieved by statistically analyzing token log probabilities.

In addition to detection methods, some approaches have explored adding watermarks to the generated text during the generation process behind LLM services [21, 50]. Through watermark extraction,

**Figure 3: Prompt Template for Code Rewriting.**

it becomes feasible to discern whether a given text originates from a model. However, it's worth noting that the inclusion of a watermarking algorithm could potentially compromise the quality of the generated text. Consequently, it cannot be presumed that all users or service providers will be inclined to produce watermarked text.

3 METHODOLOGY

3.1 Problem Definition

We focus on the zero-shot synthetic code detection problem. The synthetic code detection problem involves determining whether a code snippet x is generated from a generative LLM or an API service (such as ChatGPT). The term "zero-shot" refers to the scenario where we do not have access to a dataset with labeled synthetic code and human-written code for model training. Notably, all previous zero-shot synthetic text detection methods, including GLTR and DetectGPT, consider the "white-box" setting, where log probability score of the generative model is accessible. However, in the synthetic code detection problem, we argue that the "white-box" assumption is too strong because in many cases, some commercial models only provide the generated content and do not provide the log probability score. Therefore, we do not assume access to the log probability score and study the "black-box" setting, where only generated content is available.

3.2 Overview

Our method is grounded in the intuition that every programmer has a unique coding style and tends to solve coding problems following their own routines and habits. Similarly, generative models can be likened to senior programmers with their distinctive coding style. This method is inspired by the code reviewing and refactoring process. During these stages, the original author of the code typically does not participate. Managers assign these tasks to other programmers to prevent the original authors from being stuck in their habitual routines, which might hinder them from making significant improvements. In a similar vein, LLMs tend to exhibit consistent code writing patterns due to inherent biases in their training data. Therefore, LLMs can be viewed as programmers with unique coding styles and habitual routines shaped by the common biases present in their training data.

Therefore, based on consistent code writing patterns, we hypothesize that when generative LLMs are tasked to rewrite synthetic

Algorithm 1: Zero-shot Synthetic Code Detection

Input:

x : code snippet;

\mathcal{G} : generative model;

\mathcal{M} : similarity model;

m : number of rewriting;

ϵ : threshold;

Output:

true: x is generated by \mathcal{G} ;

false: x is not generated by \mathcal{G} ;

```

1  $x'_i \sim \mathcal{G}(\cdot | x)$ ,  $i \in [1 \dots m]$ 
2 // generate  $m$  rewriting of  $x$ .
3  $score \leftarrow \frac{1}{m} \sum_i \frac{\mathcal{M}(x) \cdot \mathcal{M}(x'_i)^T}{\|\mathcal{M}(x)\|_2 \times \|\mathcal{M}(x'_i)\|_2}$ 
4 // estimate the similarity score.
5 if  $score > \epsilon$  then
6   | return true      //  $x$  is generated by  $\mathcal{G}$ 
7 end
8 else
9 | return false     //  $x$  is not generated by  $\mathcal{G}$ 
10 end

```

code, the differences between the rewritten and the original code tend to be smaller compared to when the original code is human-written. Refer to Figure 2 for an intuitive understanding. Building on this hypothesis, we propose a zero-shot synthetic code detection method using Code Rewriting and Similarity Measurement. The overall design principle is as follows: for a given code snippet x , we process code rewriting multiple times using a generative LLM \mathcal{G} . Then, we use a code similarity model \mathcal{M} to measure the average similarity between rewritten code and original code x . We train a code similarity model \mathcal{M} using self-supervised contrastive learning [10] on unlabeled code to measure the code similarity effectively. Consequently, we use the average similarity score as the detection metric. We summarize the proposed approach in Algorithm 1. Next, the two parts of Code Rewriting and Similarity Measurement will be explained in detail.

3.3 Code Rewriting

Given a code snippet x , we generate a rewriting of x utilizing the chain of thought prompting method [48]. We prompt the generative model \mathcal{G} with the original code snippet x and instruct the model to initially generate an analysis and explanation of x , followed by a potential rewrite x' . The intermediate code analysis can help the model to understand the original code and generate a valid rewrite according to the analysis. It should be noted that x is normalized by omitting all comments and empty lines because comments are very easy to manipulate while maintaining their functionality. They can either be a source of noise or be utilized to evade detection.

To extract the code from the model's response, we require the model to return the code in Markdown format, and we subsequently remove the in-line comments during post-processing. The exact prompt template is depicted in Figure 3.

3.4 Similarity Measurement

To effectively measure the similarity between the rewritten code and the original code, we require a code similarity model \mathcal{M} , which can predict a similarity score S for the input code pair. In the domain of AI for software engineering, code similarity learning is an active area of research. Many studies address this problem by focusing on code representation learning. This involves generating dense semantic representations by leveraging various code structures [8, 15, 23, 44, 45, 49] and using vector distances to quantify similarity. We use GraphCodeBERT [15] as the base similarity model since its model checkpoint is off-the-shelf, widely used, and has been pre-trained on large-scale code.

To enhance the function-level representation of GraphCodeBERT further from a pre-trained transformer, we employ self-supervised contrastive learning. Following [10], we adapt the unsupervised SimCSE method to the code domain, using standard dropout as the data augmentation method for contrastive learning. For an input code snippet x , we use the last-layer hidden states of [CLS] token as code representation. During the unsupervised training stage, we add an MLP layer to obtain the final representation \mathbf{h} of x .

Formally, for a batch of input code snippets $\{x_i\}_{i=1}^N$, we feed the batch to the GraphCodeBERT twice and get two embeddings, $\{\mathbf{h}_i\}_{i=1}^N$ and $\{\mathbf{h}'_i\}_{i=1}^N$, with different dropout masks applied, the training objective of SimCSE is:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^N \log \frac{e^{sim(\mathbf{h}_i, \mathbf{h}'_i)/\tau}}{\sum_{j=1}^N e^{sim(\mathbf{h}_i, \mathbf{h}'_j)/\tau}} \quad (1)$$

Where τ is a temperature hyperparameter and is set to 0.1. The $sim(\mathbf{h}_i, \mathbf{h}'_i)$ is the cosine similarity $\frac{\mathbf{h}_i \cdot \mathbf{h}'_i}{\|\mathbf{h}_i\| \|\mathbf{h}'_i\|}$.

The MLP layer is dropped after the unsupervised contrastive learning stage, and we only use the last-layer hidden states of the [CLS] token as code snippet representation following [10]. For a given code snippet x and its rewriting x' sampled from \mathcal{G} as introduced in Section 3.3, we obtain the final representations of x and x' by feeding them into the similarity model \mathcal{M} . It is worth noting that our framework is not dependent on a specific similarity model; \mathcal{M} can be any implementation that accurately model code similarity, such as OpenAI’s text embedding services [13].

We use the cosine similarity function and estimate the expectation of the similarity score by sampling m rewritten code following the same procedure, as shown in Equation (2):

$$score(x) = \mathbb{E}_{x' \sim \mathcal{G}(\cdot | x)} sim(\mathcal{M}(x), \mathcal{M}(x')) \quad (2)$$

A larger m can lead to a more accurate estimation of expectation but at the cost of generating more rewritten code. Excitingly, our experiments found that only using 4 rewrites can achieve outstanding detection performance.

4 EXPERIMENT SETTING

4.1 Benchmarks

Considering the absence of existing benchmarks for evaluating synthetic code detectors, we build two synthetic code detection benchmarks in Python using APPS [17] and MBPP [2]. The choice of Python as the programming language for constructing our benchmark stems from several key considerations. Firstly, a significant

portion of training data for prevalent code generation models like CodeGen [29], Incoder [9], and StarCoder [22] is constituted by Python code. These models exhibit notably superior code generation capabilities in Python than in other languages. Consequently, the challenge and significance of detecting code generated by LLMs are most pronounced within the context of Python. Secondly, popular datasets used for code generation evaluation, such as HumanEval [5], MBPP [2], and APPS [17], primarily provide Python test cases. We adopted Python as our chosen language to align with this prevailing practice. Furthermore, Python ranks among the most widely used programming languages, adding practical significance to validating our detection performance. Specifically, for the generation of synthetic code, we utilize the natural language description of each sample as a prompt and employ four widely-used code generation tools (CodeLlama [38], StarChat [42], GPT-3.5 [31] and GPT-4 [32]).

Choice of Generation Tools. We select two open-sourced Code LLMs, CodeLlama-13B-Instruct [38] and StarChat-Alpha [42], along with two proprietary generation APIs, GPT-3.5-turbo and GPT-4 [32]. According to the survey [39] conducted by StackOverflow, GPT-3.5-turbo is the most extensively used tool in daily programming. GPT-4 represents the state-of-the-art in code generation tools to date. Consequently, we have chosen these two models to typify proprietary generation APIs. CodeLlama [38] and StarChat [42] are two potent and popular open-sourced code generation tools with millions of downloads on HuggingFace³⁴. We opted for the popular 13B and 15.5B versions due to resource constraints.

Constructing the APPS benchmark. APPS [17] is a benchmark for code generation. Each sample in APPS is a coding problem taken from coding contest websites, e.g., CodeForces, LeetCode, and HackerRank, and contains problem descriptions and submitted solutions for the problem. The original APPS Benchmark comprises 5,000 test data and 5,000 training data instances. In our construction, we exclusively utilized the test data. The rationale for excluding the training data stems from concerns that the training data might have been incorporated into the training data of code LLMs. If the training data are used as prompts and those prompts are present in the LLM’s training data, it could potentially result in code generation that perfectly matches human-written code, leading to inaccuracies in evaluating the detection outcomes. Within the pool of 5,000 test data instances, we identified and removed prompts with web hyperlinks from the Markdown format of coding challenges, as such links could affect code quality and are rarely used by developers in practice. This reduced the dataset to 3,846 instances. We then randomly selected 1,540 instances (40%) from this refined set and used 770 of them for synthetic code generation.

For those 770 problems, we generate synthetic code using the chain of thought prompting [48] to improve the generation quality using four generation tools. The exact prompt used is shown in Figure 4. The generation temperature is set to 0.7, and top_p is set to 0.95. We removed all the comments and empty lines for the generated code. For the remaining 770 problems (1540 minus 770),

³<https://huggingface.co/codellama>

⁴<https://huggingface.co/HuggingFaceH4/starchat-alpha>

we randomly sample a solution from all valid solutions as human-written code. The detailed dataset statistics are listed in Table 1.

Generator	CodeLlama	StarChat	GPT-3.5	GPT-4
# Sample	770/770	770/770	770/770	100/100
# Char	630/458	630/456	630/420	440/401
# Line	19.1/15.0	19.1/14.6	19.1/14.7	18.7/13.6

Table 1: APPS Benchmark Statistics. We count the averaged number of chars and lines for human-written/synthetic code for all versions of benchmark generated by four different code generation tools.

Constructing the MBPP benchmark. MBPP [2] is another benchmark for evaluating AI code generation. It contains 1,000 crowd-sourced Python programming problems designed to be solvable by entry-level programmers, covering programming fundamentals, standard library functionality, etc. Compared to APPS, MBPP’s problems are more straightforward and frequently encountered in daily programming practice. We follow the same construction process as APPS. The detailed dataset statistics are listed in Table 2.

Generator	CodeLlama	StarChat	GPT-3.5	GPT-4
# Sample	233/233	233/233	233/233	100/100
# Char	256/192	256/251	256/265	267/238
# Line	10.0/6.6	10.0/8.2	10.0/9.2	10.4/8.4

Table 2: MBPP Benchmark Statistics.

All intermediate scripts used to construct the benchmark, as well as the final two benchmarks, are publicly available.

4.2 Similarity Model Training Details

We continue to train a code similarity model by exploiting unsupervised SimCSE from initial GraphCodeBERT as introduced in Section 3.4. In the SimCSE training stage, we collect 160k code snippets from CodeSearchNet [18] Python subset as training data. The GraphCodeBERT model trained by SimCSE is fixed for all subsequent experiments once training is complete. We use the Adam optimizer and set the maximum learning rate to $1e - 4$ with linear decay. We use 4 NVIDIA 3090s with batch size 16 on each GPU and train the model for 5 epochs. In addition to the GraphCodeBERT model, we also experimented with the UniXcoder [14] as another similarity model by following the same SimCSE training process.

4.3 Detector LLMs

Our approach necessitates a rewriting LLM, \mathcal{G} , to function as the detector LLM. Additionally, other zero-shot detectors for text, such as GLTR and DetectGPT, require a scoring LLM for likelihood estimation, as introduced in Section 2.2. We examine CodeLlama, StarChat-Alpha, and GPT-3.5 as detector LLMs for our method while using CodeLlama and StarChat-Alpha for baseline comparison. We employ nucleus sampling for code rewriting, setting the top_p to 0.95 and the temperature to 0.8.

4.4 Metrics

We utilize the Area Under the Receiver Operating Characteristic (AUROC) curve as the evaluation metric for all the detectors on our benchmarks, following DetectGPT [27]. AUROC measures a classification model’s performance across various thresholds, providing a comprehensive evaluation of the method’s overall performance.

4.5 Baselines

We consider the following zero-shot detection methods utilizing a surrogate model to approximate the true distribution of the generative model: $\log p(x)$ [11], **LogRank**, **Rank**, **Entropy** [27], **LRR**, **NPR** [40] and **DetectGPT** [27]. To adapt these detection methods to code content and effectively approximate the true distribution of the generative code LLM, the surrogate score model should be replaced with LLMs trained on large-scale code content. In this study, we examine two open-sourced code LLMs: StarChat-Alpha and CodeLlama. We also select two supervised detectors and compare them to the zero-shot detectors: **GPTZero** [12], a leading AI content detection services, is trained on millions of synthetic texts sampled from various generative models, including ChatGPT, GPT-4 and Bard. **OpenAI-Detector** [30], an open-sourced detector, is trained on texts sampled from GPT-2 and is based on Roberta-large [26].

$\log p(x)$. This method calculates the average log probability of each token. Code snippets with higher $\log p(x)$ scores are more likely to be synthetic [11].

LogRank and Rank. These metrics use the averaged (log-) rank of tokens in the predicted distribution as a detection metric. Code snippets with lower ranks tend to be sampled from AI models.

Entropy. We use the averaged predictive entropy of each token as another baseline following Mitchell et al. [27]. Synthetic code will be more "in-distribution" for the generative model, leading to more confident predictions.

LRR and NPR. These two metrics, introduced by Su et al. [40], are used in our study with the identical calculation formula as in the original paper without any modifications, ensuring fairness.

$$\text{LRR} = \left| \frac{\frac{1}{t} \sum_{i=1}^t \log p_\theta(x_i | x_{<i})}{\frac{1}{t} \sum_{i=1}^t \log r_\theta(x_i | x_{<i})} \right| \quad (3)$$

where $r_\theta(x_i | x_{<i}) \geq 1$ is the rank of token x_i conditioned on the previous tokens.

$$\text{NPR} = \frac{\frac{1}{n} \sum_{p=1}^n \log r_\theta(\tilde{x}_p)}{\log r_\theta(x)} \quad (4)$$

where small perturbations are applied on the target text x to produce the perturbed text \tilde{x}_p .

DetectGPT. DetectGPT [27] uses the average token probability disparity between the given text and perturbed texts to detect whether the given text is located at the probability curvature of the generative model. To better adapt DetectGPT for the code domain, we made minor modifications. Specifically, we replace the original perturbation model, T5-large [37] with CodeT5-large [46]. All other settings are the same as in the original implementation.

GPTZero. It is a proprietary detection API [12]. We request the provided API with the input code following their official documentation. We use the "completely_generated_prob" field returned by their API as the detection score for AUROC calculation. "completely_generated_prob" refers to the probability that the entire code snippet is generated by an AI model.

OpenAI-Detector. The finetuned Roberta-large model checkpoint provided by the authors is used for detection [30]. This model is a binary classifier, and we use the "*log_softmax*" score of the final output layer as the detection score for AUROC calculation.

5 EXPERIMENT RESULTS

5.1 Main Results

We list the AUROC score of all zero-shot detectors on the APPS and MBPP benchmarks in Table 3. Among the seven baselines, LogRank, LRR, and NPR are the most compelling methods. However, the detection performance of baseline zero-shot detectors, including log $p(x)$, DetecGPT, LRR, and NPR drop significantly on the code benchmarks compared to their reported performance on detecting general texts [27, 40]. This observation holds true regardless of the generation tools and detector LLMs used. The main reason, as mentioned earlier, lies in the fact that these SOTA zero-shot text detectors are based on the log probability of tokens statistically. In the code domain, code exhibits a more uniform grammatical structure, often with a relatively fixed space for the next code tokens. As a result, there are many "low-entropy" tokens in code, which significantly reduces the number of effective tokens that SOTA text detection methods can use to differentiate between human-written code and synthetic code, as illustrated in Figure 1. Furthermore, We observe that the previous baselines demonstrate improved performance when the Detector LLM and the Generation LLM are identical, and exhibit a decline in performance when they do not match.

As shown in Table 3, our proposed methods significantly outperform previous methods by all four prevalent code generation tools, with a 20.5% improvement on the APPS benchmark and 29.1% improvement on the MBPP benchmark. This improvement is primarily due to our proposed methods of detecting synthetic code from a more holistic perspective instead of relying on token-wise scores. The holistic code rewriting and similarity measurement approach effectively avoids the inherent shortcomings of previous methods when applied to the code domain, while also leveraging the LLM's consistent coding style and the tendency of LLMs to output the most confident and likelihood code during generation.

It is worth noting that our method can achieve overwhelming detection performance with only 2 rewriting for estimating the similarity expectation. Furthermore, using more rewriting ($m = 4$ and $m = 8$) can lead to better performance.

5.2 Effectiveness of Detector LLM

We observe that using GPT-3.5-turbo as the detector LLM for rewriting achieves optimal performance in detecting code generated by StarChat, GPT-3.5-turbo, and GPT-4, except for CodeLlama. Optimal performance in detecting code generated by CodeLlama is

attained when using CodeLlama self as the LLM detector. The primary reason for this is that among the four models (CodeLlama, StarChat, GPT-3.5-turbo, and GPT-4), StarChat is fine-tuned on an instruction dataset distilled from GPT-3.5-turbo and GPT-4. Therefore, these three models may share a closer distribution. In contrast, CodeLlama is instruction-tuned on a dataset constructed by self-instructing Llama-2 [41]. We suspect that the suboptimal performance in detecting CodeLlama-generated code using GPT-3.5-turbo is due to the differences in the instruction datasets' distributions.

Regarding the optimal use of GPT-3.5-turbo over StarChat for detecting StarChat-generated code, our observation of the rewritten code generated by both GPT-3.5-turbo and StarChat revealed that StarChat tends to oversimplify the rewritten code in complex problems, while GPT-3.5-turbo provides more accurate and complete rewritten code.

5.3 Comparison to Supervised Detector

In Table 3, the first two rows represent two supervised detectors, GPTZero and OpenAI-Detector. We compare them with the zero-shot detectors and find that GPTZero and OpenAI-Detector fail to provide useful detection for code content, performing similarly to random guessing. Despite being trained on millions of labeled samples, these supervised detectors may only capture a small fraction of code content. This observation suggests that supervised detectors are prone to overfitting the training distribution and struggle to generalize to new domains without adaptive tuning. Furthermore, our proposed zero-shot detection methods demonstrate promising generalization capability when evaluated on different code distributions and outperform supervised detectors.

5.4 Ablation Study

We conducted an ablation experiment to analyze the contributions of two primary components in our design, i.e., Code Rewriting and Similarity Measurement. We considered two ablation settings: first, replacing the Code Rewriting with in-fill perturbation following DetectGPT [27] while retaining the Similarity Measurement (Sim); second, retaining Code Rewriting (CR) and replacing the Similarity Measurement between original and rewritten code with token-wise score difference (i.e., LogProb, LogRank, Rank and Entropy). The results are presented in Table 4.

The ablation results indicate that the "CR + Sim" method significantly enhances the detection performance. Replacing either Code Rewriting or Similarity Measurement leads to a noticeable decrease in performance, underscoring the importance of these two components.

6 MODEL ANALYSIS

We conduct comprehensive experiments to evaluate our proposed methods across diverse settings and scenarios. These experiments encompass a range of factors, including the choices of the Similarity Model, the impact of generation prompts, decoding strategy, generalizability to different programming languages, detection of revised synthetic code, consideration of code correctness, and assessment of code length impact. Furthermore, we perform a sensitivity analysis for the choice of m to understand our methods' effectiveness

Dataset		APPS					MBPP					
Methods		CodeLlama	StarChat	GPT-3.5	GPT-4	Avg.		CodeLlama	StarChat	GPT-3.5	GPT-4	Avg.
GPTZero		52.71	56.25	53.68	58.24	55.22		59.53	60.82	57.29	61.55	59.80
OpenAI		56.32	50.08	48.48	55.81	52.67		48.81	47.40	43.31	46.44	46.49
Using CodeLlama as Detector LLM												
log $p(x)$		66.14	59.40	64.58	59.27	62.35		50.70	53.84	63.05	53.35	55.24
LogRank		69.79	61.54	67.31	62.89	65.38		60.76	58.56	68.05	58.91	61.57
Rank		52.17	48.63	50.77	48.04	49.90		25.99	35.75	42.03	36.33	35.03
Entropy		58.91	54.71	61.49	55.87	57.75		37.22	44.90	50.88	43.40	44.10
DetectGPT		61.28	57.71	62.06	53.41	59.85		56.28	53.18	66.56	63.84	59.96
LRR		67.15	62.16	67.82	60.06	64.30		53.25	56.32	64.29	54.62	57.12
NPR		65.49	60.08	66.53	58.62	62.68		54.37	55.10	68.85	64.96	60.82
Ours $m = 2$		80.78	72.91	73.12	68.19	73.75		77.90	68.88	76.36	75.02	74.54
Ours $m = 4$		85.42	76.53	77.70	74.29	78.49		82.91	71.50	79.83	77.71	77.99
Ours $m = 8$		87.77	78.13	80.23	74.51	80.16		86.21	75.70	83.58	81.75	81.81
Using StarChat as Detector LLM												
log $p(x)$		66.41	65.27	65.54	62.18	64.85		55.81	64.86	69.91	60.17	62.69
LogRank		66.95	65.81	66.74	64.25	65.93		58.69	65.31	69.56	59.55	63.28
Rank		53.85	48.24	50.37	49.77	50.56		37.24	44.48	47.97	46.47	44.04
Entropy		56.55	55.43	59.60	55.30	56.72		39.03	48.72	55.35	47.22	47.33
DetectGPT		60.92	58.23	61.52	58.62	61.26		54.41	55.74	66.49	65.02	60.42
LRR		66.55	68.91	68.45	65.88	67.45		56.80	66.74	69.87	60.16	63.39
NPR		64.47	63.60	66.43	65.00	64.88		54.20	60.00	70.43	66.17	62.70
Ours $m = 2$		81.93	77.23	72.46	72.89	76.13		79.68	73.79	79.28	69.24	75.50
Ours $m = 4$		85.51	79.24	74.58	77.35	79.17		80.61	76.44	81.05	74.67	78.19
Ours $m = 8$		87.24	81.35	76.28	77.84	80.68		83.67	79.00	83.17	78.04	80.97
Using GPT-3.5-Turbo as Detector LLM												
Ours $m = 2$		77.84	81.67	79.02	79.04	79.39		66.21	77.29	83.05	83.45	77.50
Ours $m = 4$		78.21	82.22	82.12	78.69	80.31		67.00	78.87	85.39	82.23	78.37
Ours $m = 8$		78.47	82.48	83.25	80.87	81.27		67.66	79.23	86.23	84.00	79.28

Table 3: Main Results. The first two rows list two benchmarks and their corresponding four Generation Tools. The following sections detail the AUROC scores of our methods, with variations in m , alongside seven other zero-shot detectors, using three Detector LLMs. CodeLlama abbreviates CodeLlama-13B-Instruct, StarChat represents StarChat-Alpha-15B, and GPT-3.5 refers to GPT-3.5-turbo. Red indicates the best performance, while blue indicates the best performance among all seven baselines.

Dataset		APPS			
Methods		CodeLlama	StarChat	GPT-3.5	GPT-4
CR + Sim		87.77	76.96	80.23	74.51
CR + LogProb		57.29	46.70	50.58	55.13
CR + LogRank		75.28	64.76	75.85	67.69
CR + Rank		62.34	55.42	60.59	59.34
CR + Entropy		60.39	59.42	68.20	59.35
Perturb + Sim		74.55	71.55	74.23	53.43

Table 4: Ablation Study. All results are reported when using CodeLlama as the Detector LLM.

better. For all experiments in this section, we use GPT-3.5-turbo as the generation tool.

6.1 Choice of Similarity Model

In the Methodology section, we exploit self-supervised contrastive learning [10] to train a better code similarity model based on GraphCodeBERT (GCB-SimCSE). To investigate the impact of the Similarity Model, we experiment with three other variants of code similarity models: UniXcoder [14] trained by SimCSE (Unix-SimCSE), the original GraphCodeBERT model with average pooling (GCB-avg) and OpenAI’s text embedding service (Text-ada-002) [13]. The results on the APPS benchmark (generated by GPT-3.5-turbo) are listed in Table 5. The Unix-SimCSE attains the highest performance, yet using the other three alternatives still surpasses the previous zero-shot detectors. The experimental results indicate that our method is not dependent on one specific similarity model. Moreover, as UniXcoder is an enhanced version of GraphCodeBERT, the superior performance of Unix-SimCSE over our original GCB-SimCSE implies that the detection performance can be further improved by utilizing a more strong and robust similarity model.

Method	CodeLlama	StarChat	GPT-3.5
Unix-SimCSE	81.21	77.84	86.37
GCB-SimCSE	80.23	76.28	83.25
GCB-avg	72.79	70.95	72.22
Text-ada-002	70.18	69.29	76.48

Table 5: Test with different similarity models. We use CodeLlama, StarChat and GPT-3.5-turbo as Detector LLMs.

Dataset	Code Contest C++			
	Methods	CodeLlama	StarChat	GPT-3.5
$\log p(x)$	67.82	61.92	73.55	69.40
LogRank	59.35	57.12	66.68	62.43
Rank	54.67	52.19	60.29	54.21
Entropy	43.83	39.52	53.74	55.50
DetectGPT	62.99	60.45	75.47	64.05
Ours $m = 8$	89.87	83.42	90.82	88.85

Table 6: Detection Results on C++. All results are reported when using CodeLlama as the Detector LLM.

6.2 Impact of Generation Prompts

It is widely acknowledged that LLM outputs can be notably influenced by prompts. Therefore, the prompts used to generate synthetic code also have a substantial impact on the generated output. To assess this impact, we modified the prompts used in our benchmark for synthetic code generation. Specifically, we compared the effectiveness of two prompting approaches: chain of thought prompts and direct prompts. The latter directly asks the model to generate a solution without preliminary analysis. The difference between these two prompts is illustrated in Figure 4 and the results under two different prompts are presented in Figure 5.

The results in Figure 5 indicate that our proposed methods are more robust than previous zero-shot detectors when the generation prompts are varied across both benchmarks.

6.3 Impact of Decoding Strategy

Adjusting the temperature parameter in LLMs balances output diversity and accuracy, with lower temperatures yielding more deterministic and consistent results, while higher temperatures produce more varied and creative outputs. To investigate this, we conducted an experiment on the APPS benchmark where we varied the generator temperature across the range of [0.2, 0.4, 0.8] while keeping the rewriting temperature constant at 0.8. The results are shown in Figure 6. This temperature range is chosen based on the findings from Codex paper [5], which identified $T = 0.2$ and $T = 0.8$ were optimal for pass@1 and pass@100 rates, respectively. The results in Figure 6 demonstrate that our method exhibits superior consistency across different temperatures.

6.4 Generalizability to Different Programming Languages

To assess the generalizability of our method to different programming languages, we construct an additional C++ benchmark using the Code Contest dataset [25]. We generate synthetic code using the same four-generation tools. The detection performance on Code Contest C++ is presented in Table 6. Our method still achieves notable improvement compared to other zero-shot baselines on C++ benchmarks. Moreover, we observe that synthetic C++ code is easier to detect for both our method and other baselines when compared to Python, suggesting a more significant distribution gap between synthetic and human-written C++ code.

6.5 Detecting Revised Synthetic Code

In real-world scenarios, humans may revise synthetic code before using it. Naturally, this raises the question of whether zero-shot synthetic code detectors remain effective when a human user makes minor revisions to the generated code. We consider the most straightforward code modification: identifier renaming. This does not affect the functionality of the original code and even does not require the user to understand the code's logic. To mimic identifier renaming, we extract all the identifier names and randomly replace 10%, 20%, and 50% fraction with "*var_i*". Then we evaluate all the zero-shot detectors on the revised synthetic code to observe how performance changes with increasing fractions of replaced identifiers. The results are plotted in Figure 7.

From Figure 7, it's evident that the performance of all detectors declines as the replacement fraction increases, degrading to random guess when 50% of the identifiers are replaced. However, our method consistently outperforms all other detectors across all replacement fraction levels. This phenomenon can be attributed to the behavior of GPT-3.5-turbo and StarChat when rewriting the revised synthetic code. They tend to restore "*var_i*" to variable names that contain code semantics, resulting in a lower similarity between the rewritten code and the revised synthetic code compared to the similarity between the rewritten code and the original synthetic code. We consider detection in this or more complex adversarial scenarios as a focus for future work.

6.6 Impact of Code Correctness

We consider that the distribution of correct code solutions likely differs from incorrect ones since the correct synthetic code is much closer to human-written code and is more difficult to detect. To explore the impact of code correctness, we separately present the detection AUROC for both correct and incorrect codes on the MBPP benchmark when using CodeLlama as Detector LLM in Figure 8.

In Figure 8, we observe that both with our method and other baselines, detecting correct code is more challenging than detecting incorrect code. Nonetheless, our method consistently outperforms other baselines when it comes to correct code.

6.7 Impact of Code Length

We also investigate how the length of the code snippet impacts the detection performance of the zero-shot synthetic code detectors. We divide the samples in the APPS benchmark into five groups

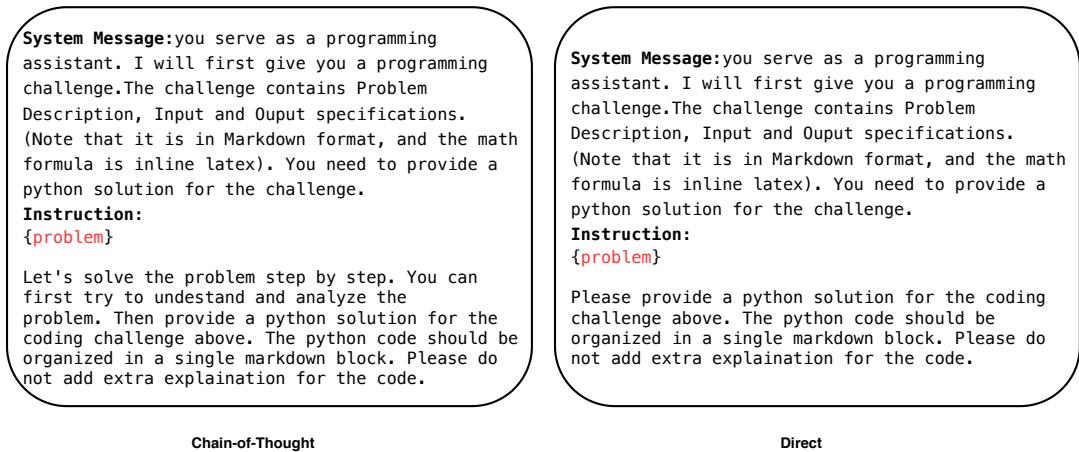


Figure 4: Prompt templates for synthetic code generation. We use chain-of-thought prompting for reporting main results in Table 3. We use direct prompting as part of model analysis for the impact of generation prompts.

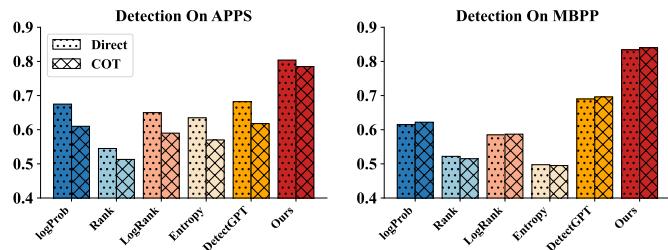


Figure 5: Impact of generation prompts.

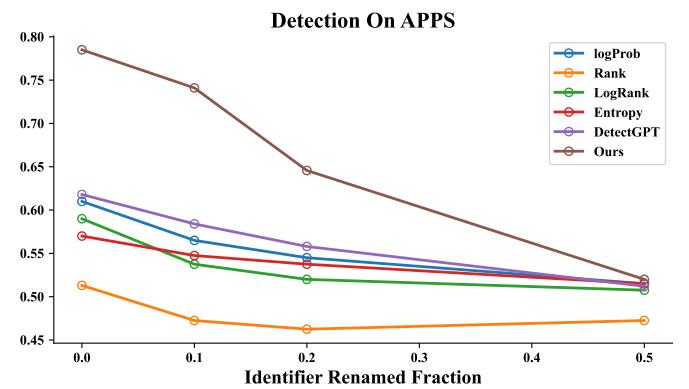


Figure 7: Detecting Revised Synthetic Code.

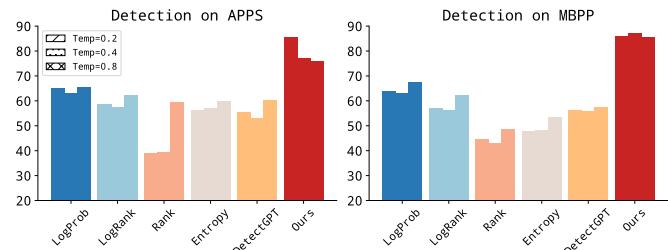


Figure 6: Impact of Decoding Strategy.

according to the code string length. Each group has an equal number of synthetic and human-written code snippets. The results are plotted in Figure 9.

In general, the performance of all detectors improves as the code length falls within the 0-1000 character range. The $\log p(x)$, LogRank, and Entropy methods achieve optimal detection performance on code snippets between 600-1000 while exhibiting inferior performance on code lengths smaller than 200. DetectGPT's performance surpasses ours when the code exceeds 1000 characters. However, DetectGPT is inferior to the other baselines for code lengths less than 600, the range where the majority (70%) of code snippets lie. Our method displays strong performance across all

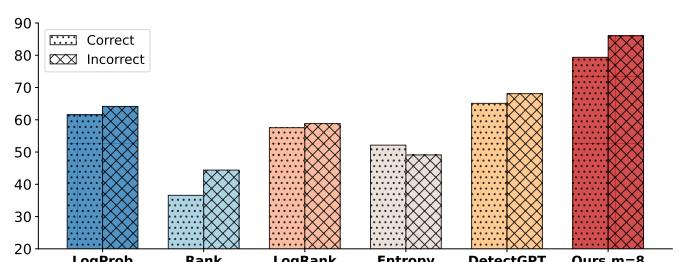


Figure 8: Impact of code correctness.

ranges and is more robust to changes in code length when compared to the other baselines.

6.8 Choice of m

In our primary experiment, we set the number of rewriting m in the range of [2, 4, 8] due to limited computational resources. However, increasing m can reduce randomness in code sampling and enhance

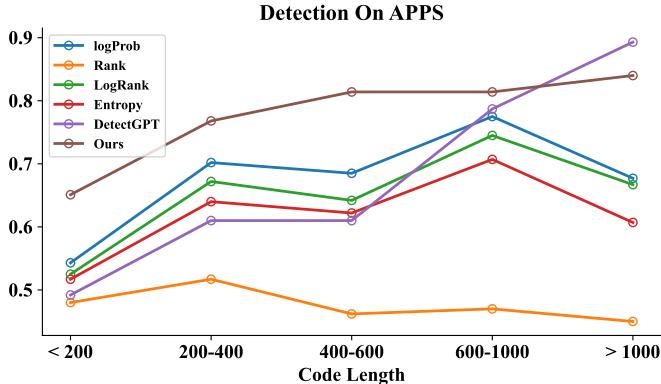
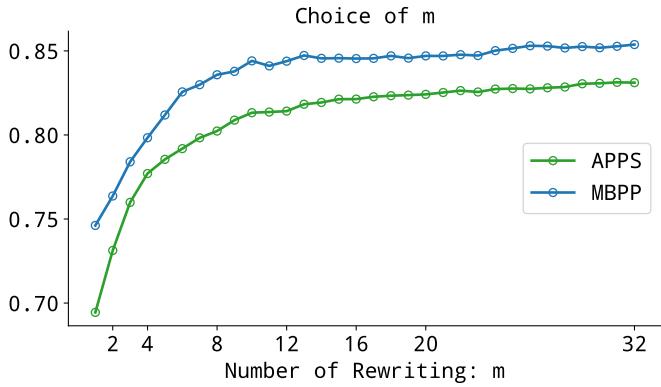


Figure 9: Impact of Code Length.

the accuracy of expectation estimation. To investigate this, we conduct experiments by setting the maximum value of m to 32 on the APPS and MBPP benchmarks (generated by GPT-3.5-turbo) and plotted the detection AUROC against changes of m in Figure 10. The result indicates that the detection performance increases with m with slight fluctuation and saturates around 20 - 32 rewrites.

Figure 10: Choice of m . CodeLlama is used as Detector LLM.

7 CONCLUSION

In this paper, we identified the gap when applying synthetic content detectors designed for general texts to code content, and proposed a novel zero-shot synthetic code detector via code rewriting and similarity measurement. We utilize the similarity between rewritten and original code as an indicator for detecting synthetic code. Our proposed method addresses the unique challenges posed by the massive "low-entropy" tokens in code domain. Through exhaustive experiments, we uncovered several findings regarding synthetic code detection. The proposed method demonstrates superior performance and robustness in detecting synthetic code compared to other zero-shot baselines.

REFERENCES

- [1] Loubna Ben Allal, Raymond Li, Denis Kocetkov, Chenghao Mou, Christopher Akiki, Carlos Munoz Ferrandis, Niklas Muennighoff, Mayank Mishra, Alex Gu, Manan Dey, et al. 2023. SantaCoder: don't reach for the stars! *arXiv preprint arXiv:2301.03988* (2023).
- [2] Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. 2021. Program Synthesis with Large Language Models. *arXiv:2108.07732 [cs.PL]*
- [3] Anton Bakhtin, Sam Gross, Myle Ott, Yuntian Deng, Marc'Aurelio Ranzato, and Arthur Szlam. 2019. Real or Fake? Learning to Discriminate Machine from Human Generated Text. *arXiv:1906.03351 [cs.LG]*
- [4] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4. *arXiv:2303.12712 [cs.CL]*
- [5] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374* (2021).
- [6] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeonjae Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. PaLM: Scaling Language Modeling with Pathways. *arXiv:2204.02311 [cs.CL]*
- [7] Paul Denny, James Prather, Brett A Becker, James Finnie-Ansley, Arto Hellas, Juho Leinonen, Andrew Luxton-Reilly, Brent N Reeves, Eddie Antonio Santos, and Sami Sarsa. 2024. Computing education in the era of generative AI. *Commun. ACM* 67, 2 (2024), 56–67.
- [8] Zhangyu Feng, Daya Guo, Duyu Tang, Nan Duan, Xiaocheng Feng, Ming Gong, Linjun Shou, Bing Qin, Ting Liu, Dixin Jiang, and Ming Zhou. 2020. CodeBERT: A Pre-Trained Model for Programming and Natural Languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 1536–1547. <https://doi.org/10.18653/v1/2020.findings-emnlp.139>
- [9] Daniel Fried, Armen Aghajanyan, Jessy Lin, Sida Wang, Eric Wallace, Freda Shi, Ruiqi Zhong, Wen-tau Yih, Luke Zettlemoyer, and Mike Lewis. 2022. Incoder: A generative model for code infilling and synthesis. *arXiv preprint arXiv:2204.05999* (2022).
- [10] Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821* (2021).
- [11] Sebastian Gehrmann, Hendrik Strobelt, and Alexander M Rush. 2019. Gltr: Statistical detection and visualization of generated text. *arXiv preprint arXiv:1906.04043* (2019).
- [12] GPTZero. 2023. *The Global Standard for AI Detection Humans Deserve the Truth*. <https://gptzero.me/>
- [13] Ryan Greene, Sanders, Lilian Weng, and Arvind Neelakantan. 2022. *New and improved embedding model*. <https://openai.com/blog/new-and-improved-embedding-model>
- [14] Daya Guo, Shuai Lu, Nan Duan, Yanlin Wang, Ming Zhou, and Jian Yin. 2022. UniXcoder: Unified Cross-Modal Pre-training for Code Representation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, Dublin, Ireland, 7212–7225. <https://doi.org/10.18653/v1/2022.acl-long.499>
- [15] Daya Guo, Shuo Ren, Shuai Lu, Zhangyu Feng, Duyu Tang, Shujie LIU, Long Zhou, Nan Duan, Alexey Svyatkovskiy, Shengyu Fu, Michele Tufano, Shao Kun Deng, Colin Clement, Dawn Drain, Neel Sundaresan, Jian Yin, Dixin Jiang, and Ming Zhou. 2021. GraphCodeBERT: Pre-training Code Representations with Data Flow. In *International Conference on Learning Representations*. <https://openreview.net/forum?id=jLoC4ez43PZ>
- [16] Jingxuan He and Martin Vechev. 2023. Large language models for code: Security hardening and adversarial testing. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*. 1865–1879.
- [17] Dan Hendrycks, Steven Basart, Saurav Kadavath, Mantas Mazeika, Akul Arora, Ethan Guo, Collin Burns, Samir Puranik, Horace He, Dawn Song, and Jacob Steinhardt. 2021. Measuring Coding Challenge Competence With APPS. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*. <https://openreview.net/forum?id=sD93GOzH3i5>
- [18] Hamel Husain, Ho-Hsiang Wu, Tiferet Gazit, Miltiadis Allamanis, and Marc Brockschmidt. 2020. CodeSearchNet Challenge: Evaluating the State of Semantic

- Code Search. arXiv:1909.09436 [cs.LG]
- [19] 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*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 1808–1822. <https://doi.org/10.18653/v1/2020.acl-main.164>
- [20] Majeed Kazemitaabar, Xinying Hou, Austin Henley, Barbara Jane Ericson, David Weinrop, and Tovi Grossman. 2023. How novices use LLM-based code generators to solve CS1 coding tasks in a self-paced learning environment. In *Proceedings of the 23rd Koli Calling International Conference on Computing Education Research*. 1–12.
- [21] John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. 2023. A Watermark for Large Language Models. arXiv:2301.10226 [cs.LG]
- [22] Raymond Li, Loubna Ben Allal, Yangtian Zi, Niklas Muennighoff, Denis Kocetkov, Chenghao Mou, Marc Marone, Christopher Akiki, Jia Li, Jenny Chim, et al. 2023. Starcoder: may the source be with you! *arXiv preprint arXiv:2305.06161* (2023).
- [23] Xiaonan Li, Daya Guo, Yeyun Gong, Yun Lin, Yelong Shen, Xipeng Qiu, Daxin Jiang, Weizhu Chen, and Nan Duan. 2022. Soft-Labeled Contrastive Pre-Training for Function-Level Code Representation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 118–129. <https://doi.org/10.18653/v1/2022.findings-emnlp.9>
- [24] Yujia Li, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, Rémi Leblond, Tom Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, et al. 2022. Competition-level code generation with alphacode. *Science* 378, 6624 (Dec. 2022), 1092–1097.
- [25] Yujia Li, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, Rémi Leblond, Tom Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, Thomas Hubert, Peter Choy, Cyprien de Masson d’Autume, Igor Babuschkin, Xinyun Chen, Po-Sen Huang, Johannes Welbl, Sven Gowal, Alexey Cherapanov, James Molloy, Daniel J. Mankowitz, Esme Sutherland Robson, Pushmeet Kohli, Nando de Freitas, Koray Kavukcuoglu, and Oriol Vinyals. 2022. Competition-level code generation with AlphaCode. *Science* 378, 6624 (Dec. 2022), 1092–1097. <https://doi.org/10.1126/science.abq1158>
- [26] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692 [cs.CL]
- [27] Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. 2023. DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature. arXiv:2301.11305 [cs.CL]
- [28] 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*, Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun’ichi Tsujii (Eds.). Association for Computational Linguistics, Brussels, Belgium, 1797–1807. <https://doi.org/10.18653/v1/D18-1206>
- [29] Erik Nijkamp, Bo Pang, Hiroaki Hayashi, Lifu Tu, Huan Wang, Yingbo Zhou, Silvio Savarese, and Caiming Xiong. 2022. Codegen: An open large language model for code with multi-turn program synthesis. *arXiv preprint arXiv:2203.13474* (2022).
- [30] OpenAI. 2019. *gpt-2-output-dataset*. <https://github.com/openai/gpt-2-output-dataset>
- [31] OpenAI. 2022. *Introducing ChatGPT*. <https://openai.com/blog/chatgpt>
- [32] OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL]
- [33] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Berglund, Irwan Bello, Jake Berdine, Gabriel Bernaddet-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yuwei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschle, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhen Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chah Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makajnu, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O’Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emry Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belutte Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael Pokorny, Michelle Pokrass, Witchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Rousset, Nick Ryder, Mario Saltarelli, Ted Sanders, Shihani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cérion Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL]
- [34] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. arXiv:2203.02155 [cs.CL]
- [35] Hammond Pearce, Baleegh Ahmad, Benjamin Tan, Brendan Dolan-Gavitt, and Ramesh Karri. 2022. Asleep at the keyboard? assessing the security of github copilot’s code contributions. In *2022 IEEE Symposium on Security and Privacy (SP)*. IEEE, 754–768.
- [36] Jianpeng Pu, Zain Sarwar, Sifat Muhammad Abdullah, Abdullah Rehman, Yoonjin Kim, Parantapa Bhattacharya, Momin Javed, and Bimal Viswanath. 2023. Deepfake text detection: Limitations and opportunities. In *2023 IEEE Symposium on Security and Privacy (SP)*. IEEE, 1613–1630.
- [37] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.* 21, 1, Article 140 (jan 2020), 67 pages.
- [38] Baptiste Rozière, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoting Ellen Tan, Yossi Adi, Jingyu Liu, Tal Remez, Jérémie Rapin, Artyom Kozhevnikov, Ivan Evtimov, Joanna Bitton, Manish Bhatt, Cristian Canton Ferrer, Aaron Grafftaiori, Wenhan Xiong, Alexandre Défossez, Jade Copet, Faisal Azhar, Hugo Touvron, Louis Martin, Nicolas Usunier, Thomas Scialom, and Gabriel Synnaeve. 2023. Code Llama: Open Foundation Models for Code. arXiv:2308.12950 [cs.CL]
- [39] StackOverflow. 2023. *Developer sentiment around AI/ML*. <https://stackoverflow.com/labs/developer-sentiment-ai-ml/>
- [40] Jinyan Su, Terry Yu Zhuo, Di Wang, and Preslav Nakov. 2023. DetectImlm: Leveraging log rank information for zero-shot detection of machine-generated text. *arXiv preprint arXiv:2306.05540* (2023).
- [41] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esioubi, Jude Fernandes, Jeremy Fu, Wenjin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madiari Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jinya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poultion, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2:

- Open Foundation and Fine-Tuned Chat Models. arXiv:2307.09288 [cs.CL]
- [42] Lewis Tunstall, Nathan Lambert, Nazneen Rajani, Edward Beeching, Teven Le Scao, Leandro von Werra, Sheon Han, Philipp Schmid, and Alexander Rush. 2023. Creating a Coding Assistant with StarCoder. *Hugging Face Blog* (2023). <https://huggingface.co/blog/starchat>.
- [43] Adaku Uchendu, Thai Le, Kai Shu, and Dongwon Lee. 2020. Authorship attribution for neural text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 8384–8395.
- [44] Xin Wang, Yasheng Wang, Fei Mi, Pingyi Zhou, Yao Wan, Xiao Liu, Li Li, Hao Wu, Jin Liu, and Xin Jiang. 2021. SynCoBERT: Syntax-Guided Multi-Modal Contrastive Pre-Training for Code Representation. arXiv:2108.04556 [cs.CL]
- [45] Xin Wang, Yasheng Wang, Yao Wan, Jiawei Wang, Pingyi Zhou, Li Li, Hao Wu, and Jin Liu. 2022. CODE-MVP: Learning to Represent Source Code from Multiple Views with Contrastive Pre-Training. In *Findings of the Association for Computational Linguistics: NAACL 2022*, Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (Eds.). Association for Computational Linguistics, Seattle, United States, 1066–1077. <https://doi.org/10.18653/v1/2022.findings-naacl.80>
- [46] Yue Wang, Weishi Wang, Shafiq Joty, and Steven CH Hoi. 2021. Codet5: Identifier-aware unified pre-trained encoder-decoder models for code understanding and generation. *arXiv preprint arXiv:2109.00859* (2021).
- [47] Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. Finetuned Language Models are Zero-Shot Learners. In *International Conference on Learning Representations*. <https://openreview.net/forum?id=gEZrGCozdqR>
- [48] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brianichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), Vol. 35. Curran Associates, Inc., 24824–24837. https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abc4-Paper-Conference.pdf
- [49] Fangke Ye, Shengtian Zhou, Anand Venkat, Ryan Marcus, Nesime Tatbul, Jesmin Jahan Tithi, Niranjan Hasabnis, Paul Petersen, Timothy Mattson, Tim Kraska, Pradeep Dubey, Vivek Sarkar, and Justin Gottschlich. 2021. MISIM: A Neural Code Semantics Similarity System Using the Context-Aware Semantics Structure. arXiv:2006.05265 [cs.LG]
- [50] KiYoon Yoo, Wonhyuk Ahn, Jiho Jang, and Nojun Kwak. 2023. Robust Multi-bit Natural Language Watermarking through Invariant Features. arXiv:2305.01904 [cs.CL]
- [51] Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. *Advances in neural information processing systems* 32 (2019).
- [52] Qinkai Zheng, Xiao Xia, Xu Zou, Yuxiao Dong, Shan Wang, Yufei Xue, Zihan Wang, Lei Shen, Andi Wang, Yang Li, et al. 2023. Codegeex: A pre-trained model for code generation with multilingual evaluations on humaneval-x. *arXiv preprint arXiv:2303.17568* (2023).
- [53] Wanjuan Zhong, Duyu Tang, Zenan Xu, Ruize Wang, Nan Duan, Ming Zhou, Jiahai Wang, and Jian Yin. 2020. Neural Deepfake Detection with Factual Structure of Text. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 2461–2470. <https://doi.org/10.18653/v1/2020.emnlp-main.193>