

ChatLog: Carefully Evaluating the Evolution of ChatGPT Across Time

Anonymous EMNLP submission

Abstract

ChatGPT has achieved great success and can be considered to have acquired an infrastructural status. There are abundant works for evaluating ChatGPT on benchmarks. However, existing benchmarks encounter two challenges: (1) Disregard for periodical evaluation and (2) Lack of fine-grained features. In this paper, we construct ChatLog, an ever-updating dataset with large-scale records of diverse long-form ChatGPT responses for 21 NLP benchmarks from March, 2023 to now. We conduct a comprehensive performance evaluation to find that most capabilities of ChatGPT improve over time except for some abilities, and there exists a step-wise evolving pattern of ChatGPT. We further analyze the inherent characteristics of ChatGPT by extracting the knowledge and linguistic features. We find some stable features that stay unchanged and apply them on the detection of ChatGPT-generated texts to improve the robustness of cross-version detection. We will continuously maintain our project at GitHub to facilitate future research.

1 Introduction

ChatGPT¹ has achieved significant success (Wei et al., 2023; Liu et al., 2023a; Bubeck et al., 2023), solidifying its status as a foundational infrastructure. Consequently, there is a growing need for in-depth research and extensive applications based on ChatGPT’s capabilities (Laskar et al., 2023).

Whether it is to assess the capabilities of ChatGPT for building applications within a specific field or to gain insights for developing Large Language Model (LLM) from ChatGPT’s performance, benchmark testing is always crucial to provide a reference. Recently, there have been a large number of newly-constructed benchmarks such as AGIEval (Zhong et al., 2023), InstructEval (Chia et al., 2023) and HaluEval (Li et al., 2023a).

¹<https://openai.com/blog/chatgpt/>

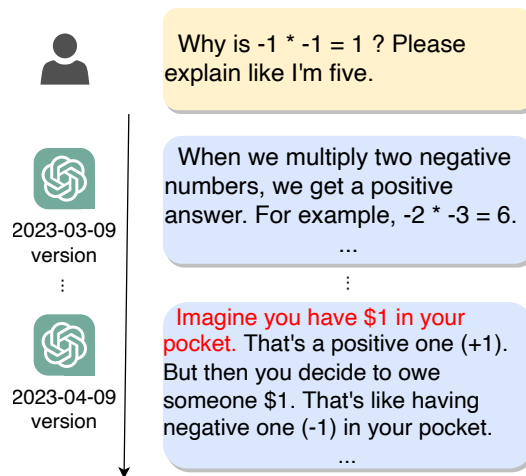


Figure 1: An example of different periods’ ChatGPT answers for the same question sampled from ELI5 dataset. Later version uses a **metaphor** for explanation.

However, the existing benchmarks encounter two challenges: (1) Lack of **periodic evaluation**: As Figure 1 shows, ChatGPT is continually evolving over time. This implies benchmark results only reflect a specific period, limiting their usefulness as ChatGPT operates without a public update schedule. (2) Absence of **fine-grained features**: ChatGPT is a generative language model which inherently inclines towards producing lengthy and diverse textual outputs (Guo et al., 2023). The linguistic characteristics exhibited in these outputs are also important part of ChatGPT’s capabilities. However, existing benchmarks often tend to summarize results using a limited number of overall metrics (Kocoń et al., 2023), typically overlooking the inherent features in outputs for evaluation.

For addressing these challenges, we have developed ChatLog, a dataset with (1) large-scale records of different ChatGPT versions that collect everyday responses from 2023-03-05 to now which is continuously updating, covering a wide range of NLP tasks (Wulczyn et al., 2017; Warstadt et al., 2019; Rajpurkar et al., 2016; Cobbe et al., 2021)

for automatic evaluation, and (2) long-form raw outputs of ChatGPT for fine-grained feature evaluation where each question is repeatedly tested three times to calculate the average feature score, thereby mitigating the influence of sampling randomness.

Based on the ChatLog dataset, we conduct an evaluation of ChatGPT from two perspectives: periodical evaluation and fine-grained features, leading to some interesting findings:

For periodical evaluation, we perform an automatic evaluation to observe how ChatGPT’s capabilities change over time. In tasks like text classification or sentiment analysis, particularly in multi-label emotion analysis, the later version’s ChatGPT has shown significant improvement. For example, ChatGPT’s F1 score on the GoEmotions dataset (Demszky et al., 2020), which requires identifying the emotions in a piece of text, increases from 24.90% in June to 28.88% in July. Besides, ChatGPT exhibits a surprising decline in other abilities that need inference on given texts, such as NLI (Wang et al., 2019b) and MRC (Rajpurkar et al., 2016), which identifies some weaknesses of ChatGPT that requiring further researches. Moreover, ChatGPT exhibits a step-wise pattern where it mainly changes after one day and stays stable in the next months. Each step change may correspond to a major update, which can help determine the effective scope of results on previous benchmarks.

For fine-grained features, we conduct feature extraction using off-the-shelf information extraction (Lu et al., 2022) and linguistic analysis (Lee et al., 2021) tools to obtain 10 knowledge features and 255 linguistic features. By calculating the variation of features, we find some features that remain consistent across time, we apply them on the ChatGPT detection task to help the detector learn more generalizable patterns. We use LightGBM (Ke et al., 2017) to combine 10 stable features with the RoBERTa-based (Liu et al., 2019) ChatGPT detector to improve 2.1 % accuracy of the base model on the test set consisting of generated responses from ChatGPT’s new versions. These findings yield valuable insights into features of ChatGPT and offer guidance for developers aimed at practical applications such as ChatGPT detection, ChatGPT style transfer, and LLM evaluation using ChatGPT, etc.

To summarize, our contribution is to construct the ChatLog dataset and conduct an extensive evaluation of temporal evolution and fine-grained features. ChatLog dataset will be continuously main-

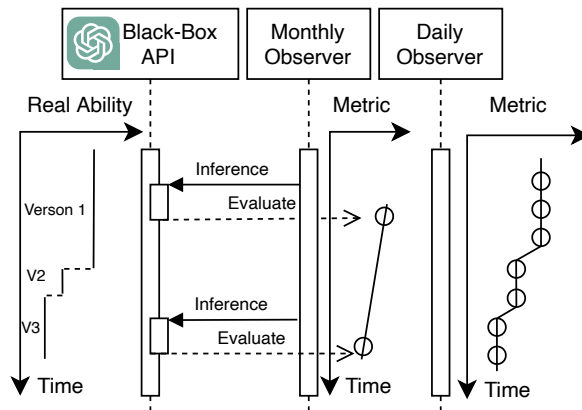


Figure 2: An illustration of periodical evaluation.

tained to (1) facilitate the analysis of the temporal patterns in ChatGPT, providing valuable insights for assessing the valid periods of benchmark results, and (2) serve as a resource for analyzing its fine-grained features or other potential attributes.

2 ChatLog

To investigate how ChatGPT is changing over time, we construct ChatLog, an ever-updating dataset that records ChatGPT’s responses for questions from 21 NLU and NLG benchmarks (Guo et al., 2023; Kocoń et al., 2023). In this section, we will introduce how we collect responses periodically.

2.1 Problem Definition

In our study, we approach the task of tracking changes in ChatGPT as a fitting problem. As depicted in Figure 2, we make the assumption that ChatGPT’s API may update over time due to many factors including the budget, the engineering constraints and the legal pressure, etc. Our work is not going to discuss why ChatGPT changes but focus on how it changes over time. Some works (Chen et al., 2023; Aiyappa et al., 2023) only use a limited number of fixed versions to study ChatGPT’s changes. In contrast, we perform both monthly and daily evaluation to capture the nuanced changes that occur over time.

Let P represent the language modeling probability of the changing black-box model, and t denote the time interval between every two evaluation with the same questions Q . After a time duration T , we can get a set of evaluation scores \mathcal{Y}_T for P during this period:

$$\mathcal{Y}_T = \{\bar{y}(x_i) | i = 1, \dots, \frac{T}{t}\} \quad (1)$$

Dataset	Task Type	Task	#Examples	Res.Len	Date Range	#Features
HC3-English	Generation	QA Dialogue	25,566 1,337	172.4 186.4	before 2023-01-18	49
Jack_of_all_trades	Classification	Pragmatic Semantic	18,777 19,953	2.5 7.3	before 2023-02-21	0
ChatLog-Daily (ours)	Generation	QA	$1,000 * n_d$	120.5	2023-03-05 to now	265
ChatLog-Monthly (ours)	Classification	Pragmatic Semantic	$18,777 * n_m$ $19,953 * n_m$	1.9 9.6	2023-03 to now	0

Table 1: Statistics of each dataset. n_d and n_m is the number of days and months from start. #Examples, Res.Len and #Features is respectively the number of collected responses, average response length, and number of extracted fine-grained features. Note that we are collecting data continuously and releasing them on GitHub periodically.

$$\bar{y}(x_i) = \frac{\sum_1^N y(x_i)}{N} \quad (2)$$

$$x_i = P_i(Q) \quad (3)$$

where x_i and P_i denote the responses and corresponding version’s API of the i -th period, y indicates the evaluation metric. In order to eliminate the sampling randomness of P_i , we repeatedly send Q for N times on the same i -th period so that we can get the average evaluation score $\bar{y}(x_i)$ to represent P_i ’s ability. Therefore, we can leverage each P_i ’s score to fit the changes of P over time.

2.2 Data Collection

Data Source. To compare with previous versions’ ChatGPT in open-source datasets, we select two typical datasets, including:

- **HC3-english** (Guo et al., 2023). A dataset consisting of approximately 26k questions with their corresponding human expert responses and ChatGPT responses, covering domains such as computer science, open-domain, finance, medicine, law, and psychology.

- **Jack_of_all_trades** (Kocoń et al., 2023) A dataset to evaluate ChatGPT’s performance on 25 diverse analytical NLP tasks on existing public NLP datasets, such as sentiment analysis, emotion recognition, and recognizing textual entailment, etc. We select 20 of these English-based tasks and datasets that provide full evaluation scripts.

Questions for ChatLog-Monthly. As shown in Table 1, ChatLog-Monthly is a dataset of 38,730 question-answer pairs collected every month. There have been 7 months’ data recorded and evaluated until the paper submission. To comprehensively evaluate ChatGPT monthly, we collect all questions from Jack_of_all_trades dataset which actually is composed of existing benchmarks.

Category	Source Dataset
Sentiment Analysis	ColBERT (Annamoradnejad and Zoghi, 2020) Tweeveal (Barbieri et al., 2020) GoEmotions (Demszky et al., 2020)
Text Classification	CoLa (Warstadt et al., 2019) Wikipedia Talk (Wulczyn et al., 2017) Wic (Pilehvar and Camacho-Collados, 2019) SMS Spam (Hidalgo et al., 2012) Sarcasmania (Siddiqui, 2019)
Natural Language Inference	SuperGLUE (Wang et al., 2019a), GLUE (Wang et al., 2019b)
Reading Comprehension	SQuAD v2.0 (Rajpurkar et al., 2018) RACE (Xu et al., 2017)
Mathematical Reasoning	GSM8K (Cobbe et al., 2021)

Table 2: Source datasets of ChatLog-Monthly.

As shown in Table 2, the questions cover a wide range of NLP tasks, including sentiment analysis (Annamoradnejad and Zoghi, 2020; Barbieri et al., 2020; Demszky et al., 2020), text classification (Warstadt et al., 2019; Wulczyn et al., 2017; Pilehvar and Camacho-Collados, 2019; Hidalgo et al., 2012; Siddiqui, 2019), natural language inference (Wang et al., 2019a,b), reading comprehension (Rajpurkar et al., 2018; Xu et al., 2017) and reasoning (Cobbe et al., 2021). Datasets’ details are listed in Appendix A.

Questions for ChatLog-Daily. To monitor ChatGPT’s daily changes, we sample 1000 questions from the HC3 dataset. These 1000 questions have been repeatedly sent to ChatGPT from March 5 to now. There have been 207 days’ data recorded and evaluated until the paper submission. The original dataset of these 1000 questions is ELI5 (Fan et al., 2019), a long-form question-answering dataset composed of threads from the Reddit forum "Explain Like I’m five". The questions are open-ended and the answers are supposed to be comprehensible by a five-year-old, which means the answers should

Task Name	Task Category	ChatGPT March[%]	ChatGPT April[%]	ChatGPT May[%]	ChatGPT June[%]	ChatGPT July[%]	ChatGPT August[%]	ChatGPT Sept.[%]
CoBERT	Sentiment	74.94	73.00 ~	73.30 ~	72.40 ~	75.24 ~	73.01 ~	74.84 ~
TweetEmoji	Sentiment	16.14	18.24 ↑	17.74 ~	18.05 ~	15.86 ↓	17.40 ~	17.10 ~
TweetSent	Sentiment	66.26	66.20 ~	66.20 ~	65.65 ~	61.05 ~	61.72 ~	62.82 ~
TweetStance	Sentiment	52.02	54.23 ~	52.98 ~	53.44 ~	48.96 ~	51.33 ~	51.23 ~
GoEmo	Sentiment	27.29	29.45 ~	27.53 ~	24.90 ~	28.88 ↑	28.31 ~	25.10 ↓
GoEmoPer0	Sentiment	23.97	23.64 ~	24.51 ~	21.56 ↓	23.81 ↑	22.28 ~	22.41 ~
GoEmoPer1	Sentiment	22.47	22.32 ~	23.85 ~	23.09 ~	22.68 ~	21.62 ~	22.87 ~
GoEmoPer2	Sentiment	21.29	22.77 ~	23.26 ~	23.29 ~	19.65 ↓	23.70 ↑	23.32 ~
GoEmoPer3	Sentiment	24.69	22.16 ↓	24.26 ~	23.19 ~	21.53 ~	23.39 ~	25.07 ~
CoLa	Classify	82.07	80.25 ~	80.06 ~	79.96 ~	79.67 ~	80.44 ~	79.87 ~
Aggression	Classify	75.90	76.12 ~	76.01 ~	75.91 ~	71.78 ~	71.06 ~	70.28 ~
AggressionPer	Classify	72.97	72.80 ~	72.47 ~	72.17 ~	69.33 ~	69.87 ~	69.85 ~
WordContext	Classify	57.99	62.23 ~	62.54 ~	61.60 ~	62.23 ~	59.72 ~	57.52 ~
Spam	Classify	82.81	84.29 ~	84.29 ~	82.41 ~	74.80 ~	72.28 ~	75.04 ~
Sarcasm	Classify	46.34	48.50 ~	48.70 ~	48.40 ~	45.46 ~	46.19 ~	44.97 ~
TextEntail	NLI	82.31	80.34 ~	80.05 ~	80.34 ~	79.28 ~	80.72 ~	79.98 ~
WNLI	NLI	71.83	74.65 ~	74.65 ~	76.06 ~	71.83 ~	69.01 ~	67.61 ~
SQuAD	MRC	58.07	48.20 ↓	48.63 ~	48.99 ~	46.24 ~	45.57 ~	43.75 ~
ReAging	MRC	75.90	76.77 ~	76.36 ~	75.87 ~	73.74 ~	72.02 ~	72.54 ~
GSM8K	Reasoning	78.00	77.70 ~	78.10 ~	78.00 ~	73.40 ~	74.30 ~	73.70 ~

Table 3: ChatGPT’s quantitative performance on each task of ChatLog-Monthly from March to September, where *ChatGPT Sept.* refers to ChatGPT’s performance in September. The ↑ or ↓ symbol refers to the performance increases or decreases over 10% compared with the previous month. The ~ means the change is smaller than 10%.

be simple and easy to understand. By evaluating ChatGPT on these questions, we can assess its text generation and instruction following ability.

2.3 Response Process

OpenAI policy. We use the API called *gpt-3.5-turbo*, which is the model that powers ChatGPT and is optimized for conversational formats. According to the documentation of OpenAI API, *gpt-3.5-turbo* will be updated with their latest model iteration. OpenAI provides a clear pricing structure and usage policy for its API. Pricing of API calls is based on tokens. For *gpt-3.5-turbo*, it costs \$0.002 per 1000 tokens. Furthermore, OpenAI has implemented policies to prevent abusive use of the API, including malicious requests or attempts to overload the server. As OpenAI continues to train ChatGPT on new data, we can observe how its capabilities evolve alongside the ongoing analytical progress in our research.

Prompt and parameter. We aim to carefully address the issue of controlling variables, specifically focusing on prompts and parameters. To maintain evaluation consistency with prior research, we use the same prompt and keeping the prompts and parameters fixed. For instance, we include the prompt "explain like I’m five" after each question on ELI5 (Fan et al., 2019), following the approach taken by HC3. Additionally, we utilize

the default parameters provided by OpenAI during API calls. Although the default decoding parameter (temperature = 1) introduces randomness in responses, we mitigate this by sending the same set of 1000 questions to ChatGPT three times each day. Consequently, we can calculate evaluation or feature scores by averaging the three generated results. To further ensure the stability of our results, we conduct a statistical test in Section 3.3.

3 Periodical Evaluation

To investigate ChatGPT’s behavior across time, we evaluate its performance on various benchmarks using automatic metrics.

3.1 Evaluation Settings

We quantify ChatGPT’s performance on classification tasks using two classification metrics, accuracy and macro F1 score. Accuracy measures the proportion of correctly classified instance out of the total number of instances, which provides a simple measure of classification. The F1 score is a commonly used metric that combines precision and recall metrics into a single score. The macro F1 score is calculated as the unweighted average of the F1 score for each class. It is used to evaluate the overall performance across all classes equally. For most tasks of ChatLog-Monthly, we use F1 as the metric except four tasks: CoLa, Word-

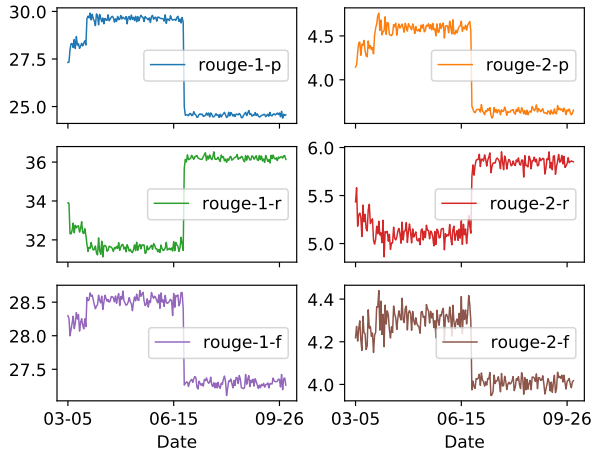


Figure 3: Trend of ChatGPT’s rouge scores and linguistic features on ChatLog-Daily dataset using 3 trials’ average score (i.e. $N = 3$) every day since Mar 5, 2023.

Context, WNLI and GSM8K, which utilize accuracy to measure. For evaluation on the generation task (Fan et al., 2019) of ChatLog-Daily, we choose the widely-used rouge (Lin, 2004) metric to measure the token-level overlap with the reference.

3.2 Evaluation Results

We report ChatGPT’s understanding performance on ChatLog-Monthly and its generation performance on ChatLog-Daily.

Understanding Performance. In Table 3, we exhibit ChatGPT’s performance after March and have the following observations:

(1) **Progressive abilities:** In tasks that require text classification or sentiment analysis, such as offensive language detection and multi-label emotion analysis, the later version’s ChatGPT has shown significant improvement. For example, ChatGPT’s F1 score on the GoEmotions dataset (Demszky et al., 2020), which requires identifying the emotions in a piece of text, increases from 24.90% in June to 28.88% in July. But the performance on GoEmo drops to 25.10% in September. We hypothesize that the fluctuation may be attributed to the sampling randomness as other tasks have shown stable results.

(2) **Regressive abilities:** In tasks that rely on limited given texts, such as Natural Language Inference (NLI) and Machine Reading Comprehension (MRC), ChatGPT shows a decline in performance on these benchmarks including TextEntail, WNLI, and SQuAD. For example, ChatGPT’s F1 score on the SQuAD dataset, which requires an understanding of the given Wikipedia passage, decreased

from 58.07% in March to 48.20% in April. This may be caused by the version update in March that changes the parameter of ChatGPT helping it gain some new abilities while forgetting some other aspects, leading to the drop in the MRC task.

(3) **Step-wise pattern:** There are 2 notable changes are observed, one is from March to April, the other is from June to July. While minimal variations are recorded in April and May. This suggests that ChatGPT might have been updated with a new version during March and June while stay stable in April and May. These results reveal the complexities of evaluating ChatGPT as new versions may get different evaluation results over time. Our records can help determine the effective scope of evaluation results on benchmarks, satisfying the need for tracking of ChatGPT’s performance.

Generation Performance. In addition to ChatLog-Monthly which mainly evaluates the NLU abilities, we also evaluate ChatGPT’s generation ability on the ChatLog-Daily dataset with ELI5 (Fan et al., 2019) questions for the long-form QA task using the rouge (Lin, 2004) metric. rouge-1 and rouge-2 indicate the unigram and bigram overlap between the reference answer and the generated answer. And suffix p,r,f represents precision, recall, and f1 respectively. To rule out the possibility that the changes in ChatGPT are just from randomness and confirm the stability of our evaluation results, we conduct 3 trials for each data everyday. We have the following findings:

(1) As shown in Figure 3, we can see that ChatGPT’s generated answers are getting more concise as the rouge-1-precision score increases. While the declining performance of ChatGPT on the rouge-1-recall metric indicates that ChatGPT produces less redundant text to get higher readability.

(2) However, only using reference-guided metrics may not evaluate the quality of generated texts comprehensively (Ke et al., 2022). For example, the case in Figure 1 shows that ChatGPT can use the metaphor for answering questions, this semantic-level metaphoric capability can not be reflected by token-level rouge score. This limitation inspires our further evaluation in Section 4.

3.3 Statistical Test

In addition to the visual inspection for the changes over time, we conduct a statistical test that measures the degree of variation that can be attributed to version change rather than sampling randomness.

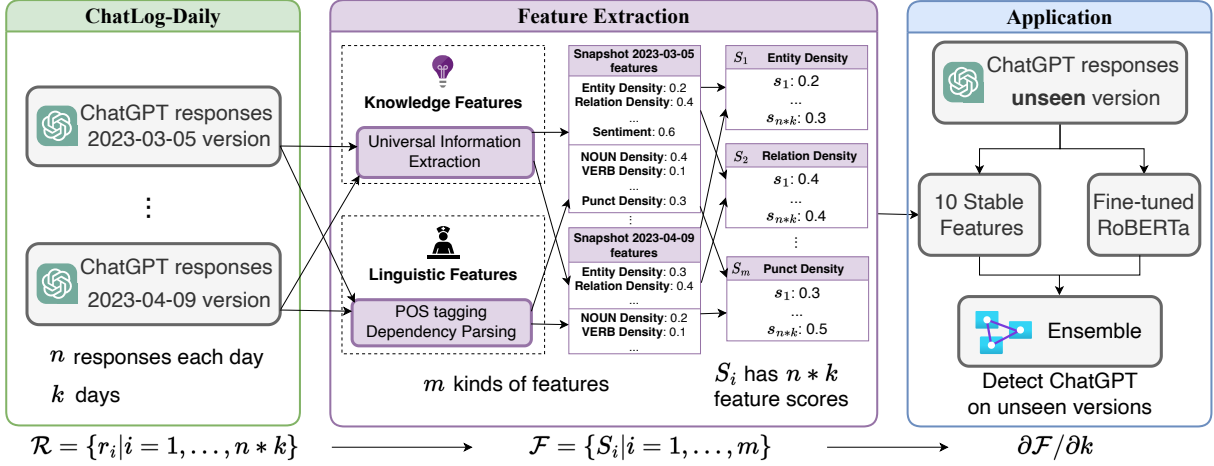


Figure 4: The framework of our feature extraction and application process. We first extract knowledge and linguistic features on the long-form answers from ChatLog-Daily dataset. Then we conduct an analysis on these features to find unchanged stable features over time and apply these features on the ChatGPT-generated text detection task.

T	$ \mathcal{Y} $	$\bar{\mathcal{Y}}$	S	Z
03-05 to 03-23	18	0.2816	0.0035	—
03-24 to 05-09	46	0.2965	0.0020	16.98
05-10 to 06-27	46	0.2960	0.0019	-1.006
06-28 to 08-14	48	0.2457	0.0015	-142.0
08-15 to 10-02	49	0.2457	0.0016	0.1629

Table 4: Result of the statistical test, where Z is the z -score of the current (T_1) and the previous (T_2) period.

Let T_1 and T_2 denote two successive periods, if the model’s version is not changed from T_2 to T_1 , then their average performance μ_1, μ_2 should be equal. Therefore, we can prove the significance of version change by testing the following null hypothesis,

$$H_0 : \mu_1 = \mu_2 \quad (4)$$

Let $\bar{\mathcal{Y}}$ and S denote the average and standard deviation of evaluation scores \mathcal{Y} on each period. We can calculate the z -score by:

$$Z = \frac{\bar{\mathcal{Y}}_1 - \bar{\mathcal{Y}}_2}{\sqrt{\frac{S_1}{|\mathcal{Y}_1|} + \frac{S_2}{|\mathcal{Y}_2|}}} \quad (5)$$

As shown in Table 4, we calculate four different periods’ z -score of rouge-1-p with the corresponding previous period. We find that (1) There are two significant version changes, one is in March 24 with a z -score of 16.98, another is in June 28 with a z -score of -142. Both of them are larger than the critical value 2.58 of z -score at the 99% confidence level. This is consistent with the fact that OpenAI updates their gpt-3.5-turbo API in two weeks after

a version change. (2) Between the changes, the model’s performance stays on a stable merit. For each stable version after March 24, we split their records into two evenly long parts, and the z -score of each part is -1.006 and 0.1629 respectively, which are smaller than the critical value 1.96 of z -score at the 95% confidence level. Therefore, we believe that ChatGPT are stable on these days.

4 Fine-grained Features

To analyze the underlying characteristics of ChatGPT beyond evaluation performance, we perform a comprehensive extraction of knowledge and linguistic features on the ChatLog-Daily dataset to discover ChatGPT generated texts’ unchanged features over time. We also explore the application of these features on AI detection.

4.1 Feature Extraction Settings

Figure 4 shows our pipeline and each sub-process is introduced in the following sections.

Extraction Process. We first take n source queries \mathcal{Q} as input then send the query to LLM API each day across k days to get $n * k$ responses \mathcal{R} . Then we apply off-the-shelf information and linguistic feature extraction tools (Lu et al., 2022; Lee et al., 2021) on the raw responses \mathcal{R} to extract m sentiment, knowledge and linguistic features for each r in \mathcal{R} . Therefore, we can get a set of m snapshot feature score collections $\mathcal{F} = \{S_i | i = 1, \dots, m\}$ for the next feature analysis stage, where each snapshot S_i contains $n * k$ scores.

Type	Tool	Branch	#Features
Knowledge	UIE	NE	5
Knowledge	UIE	OP	2
Knowledge	CogIE	RE	1
Knowledge	CogIE	FP	2
Linguistic	LingFeat	AdSem	48
Linguistic	LingFeat	Disco	28
Linguistic	LingFeat	Synta	109
Linguistic	LingFeat	LxSem	56
Linguistic	LingFeat	ShaTr	14

Table 5: Statistics of feature types, extractive tools, branches and the number of features under each branch.

Feature Category. As shown in Table 5, we use UIE (Lu et al., 2022) and CogIE (Jin et al., 2021) tools to extract knowledge information such as named entities (NE), opinions(OP), and entity relation extraction (RE), as well as more abstract knowledge like frame-semantic parsing (FP) of semantic frames and arguments. In addition to knowledge features, we also utilize the LingFeat (Lee et al., 2021) tool to obtain linguistic features such as Advanced Semantic (AdSem), Discourse (Disco), Syntactic (Synta), Lexico Semantic (LxSem) and Shallow Traditional Features (ShaTr). Overall, our extraction covers a wide range of features, with syntax and semantics being the most prominent.

4.2 Detection Application Settings

We further explore the application of extracted ChatGPT features on our ChatLog-Daily dataset. We show a preliminary use case: improving the robustness of the ChatGPT detection model, which aims to classify whether a paragraph of text is generated by ChatGPT or humans. The evaluation details and baselines are as follows:

Dataset. The portion of training, valid and test set is 9:1:10. The training and the valid set consists of 1000 ChatGPT and 1000 human answers sampled from the ELI5 part of HC3 (Guo et al., 2023) collected before January 18, 2023. For the test set, we sample 1000 ChatGPT responses from our ChatLog-Daily dataset, which consists of ChatGPT’s responses to ELI5 (Fan et al., 2019) questions for each day from March 5 to June 10. We also sample 1000 human answers for the test set to make the human and ChatGPT labels balanced.

Evaluation. We use accuracy to evaluate the performance of detection classifiers. We train models on HC3 corpus (Guo et al., 2023) with ChatGPT data gathered before January 18, 2023, and test them on our new dataset collected after March 5

to evaluate the robustness of unseen versions’ data. And please refer to Appendix C.1 for further detailed implementation information.

Baselines. Following Guo et al. (2023), we select 3 typical detection baselines for fine-tuning on our dataset, which are:

- **RoBERTa** (Liu et al., 2019) is a pre-trained Transformer-based neural network that can be fine-tuned to handle classification tasks.

- **PPL** (Jelinek et al., 1977) is a metric to model the probability of a word sequence on a distribution. We adapt it for the detection task.

- **GLTR** (Gehrmann et al., 2019) utilizes the exceedingly powerful Test-2 evaluation metric, which is constructed from the number of tokens occupying the top ranks contained within the language model’s predicted probability distributions. Subsequently, logistic regression techniques were applied and a model was trained to classify input texts.

4.3 Feature Analysis

To effectively detect ChatGPT, we conduct an analysis to discover those **stable features** that stay unchanged over time by minimizing the variation coefficient of extracted features. As we get a set of m snapshot feature score collections $\mathcal{F} = \{S_i | i = 1, \dots, m\}$ for the analysis, where each snopshot S_i contains $n * k$ scores. We can calculate the variation coefficient c_h for each feature score set S_h :

$$\mu_h = \frac{\sum_{i=1}^n \sum_{j=1}^k s_{ij}}{n * k} \quad (6)$$

$$\sigma_h = \frac{\sum_{i=1}^n \sum_{j=1}^k (s_{ij} - \frac{\sum_{l=1}^k s_{il})^2}{k}}{n * k} \quad (7)$$

$$c_h = \frac{\sigma_h}{\mu_h} \quad (8)$$

where $h \in \{1, \dots, m\}$ indicates the feature’s index. Specifically, we calculate the variation coefficient on ChatLog-Daily, after filtering those features that always are zero, the rest features with minimum variation coefficient are selected. Finally, we select the top 10 stable features to combine with RoBERTa for ChatGPT detection.

4.4 Application of Stable Features

Effectiveness of Features on ChatGPT Detection To demonstrate the application of the derived features, we employ them in the context of

Model	Test Accuracy (Unseen Dates)	Test True Negative Rate	Test False Positive Rate	Test False Negative Rate	Test True Positive Rate
RoBERTa (Liu et al., 2019)	96.2 \pm 0.1	50.0 \pm 0.0	0.0 \pm 0.0	3.8 \pm 0.1	46.2 \pm 0.1
PPL (Guo et al., 2023)	86.7 \pm 0.5	49.6 \pm 0.1	0.4 \pm 0.1	13.0 \pm 0.4	37.0 \pm 0.4
GLTR (Gehrmann et al., 2019)	77.6 \pm 0.4	49.6 \pm 0.1	0.3 \pm 0.1	22.1 \pm 0.4	27.9 \pm 0.4
RoBERTa + random 10 features	95.3 \pm 0.3	49.0 \pm 1.2	0.9 \pm 1.2	3.7 \pm 3.6	46.3 \pm 3.6
RoBERTa + stable 10 features	98.1 \pm 0.1	49.2 \pm 0.8	0.8 \pm 0.8	1.1 \pm 1.2	48.9 \pm 1.2

Table 6: Accuracies (%) on the test set of baselines and our detectors, along with 5 random trials’ standard deviations.

the ChatGPT detection task. While baseline detectors (Guo et al., 2023) are trained on ChatGPT corpus collected from history, the features learned from previous corpus may not stay consistent in future versions. Specifically, we use either randomly selected 10 features or the most stable 10 features to combine with RoBERTa. The combined 11 features are utilized by LightGBM (Ke et al., 2017) for final label prediction. With results in Table 6, we have three observations:

(1) **Performance drops on new versions:** The accuracy of RoBERTa, gltr, and ppl classifiers decreases rapidly from the valid set with old data to the test set with newly collected corpus, which shows the difference between ordinary ChatGPT texts and up-to-date texts, and demonstrates that different periods’ ChatGPT may have different text styles. This is consistent with the fact that OpenAI updates their gpt-3.5-turbo API periodically.

(2) **Fine-tuned classifiers are accurate:** Among all classifiers, gltr and ppl actually depend on the white-box hypothesis to predict (Gehrmann et al., 2019), which have more interpretability but also encounter larger performance drop on the test dataset than the fine-tuned RoBERTa binary classifier, which achieves over 96% accuracy. This is because the fine-tuned RoBERTa classifier has more parameters than gltr and ppl, which can help it to better adapt to the unseen versions.

(3) **Stable features enhance the model:** The incorporation of our 10 most stable features significantly boosts the performance of the RoBERTa detector on the test dataset, whereas the utilization of random features actually hurt the performance, which proves that our stable features can effectively improve the detector’s robustness on unseen versions’ ChatGPT texts. This is because the stable features are more likely to be consistent across different versions, which can help the detector to learn more generalizable patterns.

5 Related Work

After the success of LLMs (Glaese et al., 2022; OpenAI, 2023), many efforts have been devoted to evaluating the capabilities of LLMs, such as natural language understanding (Laskar et al., 2023; Das et al., 2023; Li et al., 2023a), reasoning (Frieder et al., 2023; Liu et al., 2023a; Bian et al., 2023), and generation abilities (Bubeck et al., 2023; Zhou et al., 2023; Gao et al., 2023). Some find that recent ChatGPT and GPT4 perform better than previous models (Zhong et al., 2023; Chia et al., 2023), and there may exist an evolutionary tree of language models (Yang et al., 2023). Inspired by the concept of evolution (Ogundare and Araya, 2023), we study whether ChatGPT can evolve across its versions.

Due to the strong generation ability, LLMs have significantly impacted various fields, such as education (Khalil and Er, 2023). Students may use ChatGPT for finishing assignments and exams (Haensch et al., 2023; Herbold et al., 2023). Consequently, there is a need for effective detection methods (Kirchenbauer et al., 2023; Mitchell et al., 2023; Krishna et al., 2023) to address this problem (Tang et al., 2023). We hope our extracted features may assist LLM detection application (Liu et al., 2023b; Li et al., 2023b; Yu et al., 2023).

6 Conclusion

This study introduces ChatLog, a dynamic dataset that evaluates ChatGPT’s evolution across various benchmarks. The evaluation encompasses 21 tasks, offering a comprehensive assessment of ChatGPT’s evolving natural language understanding and generation capabilities over time. Additionally, we extract features from ChatGPT’s daily long-form responses to investigate its intrinsic characteristics, which can significantly enhance the detection application for ChatGPT. We hope that this research will contribute to a greater understanding of ChatGPT’s evolution over time and will have practical implications for real-world applications.

553 Limitations

554 Although we have conducted extensive experi- 599
555 ments, there are still some limitations for our work: 600
556 (1) The LLM responses in our paper are all col- 601
557 lected from ChatGPT. And all the analytical experi- 602
558 ments are based on ChatGPT responses. Whether 603
559 the observed phenomena widely exist on other 604
560 LLMs is an open question and needs more explo- 605
561 rations. (2) The datasets employed in our experi- 606
562 ments are entirely in English, imposing limitations 607
563 on the linguistic features considered in our evalua- 608
564 tion. We chose English just because of its exten- 609
565 sive resources, and we intuitively believe that the 610
566 phenomena observed in experiments are not depen- 611
567 dent on the English language. Therefore, we look 612
568 forward to future studies that include experiments 613
569 using more diverse languages. (3) The application 614
570 mentioned in our paper just covers the ChatGPT de- 615
571 tection task. The potential usages of our work are 616
572 still under exploration. Besides, in the ChatGPT de- 617
573 tection experiment, the extracted stable features are 618
574 only applied on RoBERTa base model. Whether 619
575 these features are useful for other models is unclear 620
576 and more experiments are needed to verify it. 621

577 Ethics Statement

578 In this section, we will discuss the ethical consider- 622
579 ation for our work. 623

580 **Licenses.** For open-accessible datasets used in 618
581 our work, we have checked their licenses. The 619
582 HC3 (Guo et al., 2023) dataset is shared under 620
583 the CC-BY-SA license and the JAT dataset (Ko- 621
584 coń et al., 2023) is released under the MIT li- 622
585 cense. The Licenses for the feature extraction 623
586 tools are also available. UIE (Lu et al., 2022) is 624
587 released under the Creative Commons Attribution- 625
588 NonCommercial-ShareAlike 4.0 International Pub- 626
589 lic License for Noncommercial use only, Co- 627
590 gIE (Jin et al., 2021) is shared under the Apache-2.0 628
591 license, and LingFeat (Lee et al., 2021) is shared 629
592 under the CC-BY-SA-4.0 license., 630

593 **Ethics Considerations for ChatGPT** ChatGPT 631
594 complies with the AI ethical guidelines set by the 632
595 European Union². These guidelines place em- 633
596 phasis on various ethical aspects, including hu- 634
597 man oversight, technical robustness, safety, pri- 635
598 vacy, data governance, transparency, diversity, non- 636

²<https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>

discrimination, societal and environmental well- 599
being, and accountability. They highlight the im- 600
portance of empowering humans, ensuring the 601
safety and accuracy of AI systems, promoting sus- 602
tainability, and establishing accountability mech- 603
anisms for potential negative consequences. This 604
encourages our work to record ChatGPT’s daily 605
responses that may help policy makers conduct reg- 606
ulations for generative AI systems like ChatGPT. 607

References 608

- Rachith Aiyappa, Jisun An, Haewoon Kwak, and Yong- 609
Yeol Ahn. 2023. [Can we trust the evaluation on ChatGPT?](#) *arXiv preprint arXiv:2303.12767*. 610
611
- Issa Annamoradnejad and Gohar Zoghi. 2020. [Colbert: Using bert sentence embedding for humor detection.](#) *arXiv preprint arXiv:2004.12765*. 612
613
614
- Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 615
1998. [The Berkeley FrameNet project.](#) In *Proceedings of COLING*. 616
617
- Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. [TweetEval: Unified benchmark and comparative evaluation for tweet classification.](#) In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650. 618
619
620
621
622
623
- Ning Bian, Xianpei Han, Le Sun, Hongyu Lin, Yaojie Lu, and Ben He. 2023. [Chatgpt is a knowledgeable but inexperienced solver: An investigation of commonsense problem in large language models.](#) *arXiv preprint arXiv:2303.16421*. 624
625
626
627
628
- Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. [Sparks of artificial general intelligence: Early experiments with gpt-4.](#) *arXiv preprint arXiv:2303.12712*. 629
630
631
632
633
634
- Lingjiao Chen, Matei Zaharia, and James Zou. 2023. [How is chatgpt’s behavior changing over time?](#) *arXiv preprint arXiv:2307.09009*. 635
636
637
- Yew Ken Chia, Pengfei Hong, Lidong Bing, and Soujanya Poria. 2023. [INSTRUCTEVAL: Towards holistic evaluation of instruction-tuned large language models.](#) *arXiv preprint arXiv:2306.04757*. 638
639
640
641
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. [Training verifiers to solve math word problems.](#) *arXiv preprint arXiv:2110.14168*. 642
643
644
645
646
- Israel Cohen, Yiteng Huang, Jingdong Chen, Jacob Benesty, Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. 2009. [Pearson correlation coefficient. Noise reduction in speech processing](#), pages 1–4. 647
648
649
650
651

652	Mithun Das, Saurabh Kumar Pandey, and Animesh Mukherjee. 2023. Evaluating chatgpt’s performance for multilingual and emoji-based hate speech detection . <i>arXiv preprint arXiv:2305.13276</i> .	Zhuoran Jin, Yubo Chen, Dianbo Sui, Chenhao Wang, Zhipeng Xue, and Jun Zhao. 2021. CogIE: An information extraction toolkit for bridging texts and CogNet . In <i>Proceedings of ACL</i> , pages 92–98.	705
653			706
654			707
655			708
656	Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions . In <i>Proceedings of ACL</i> , pages 4040–4054.	Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. LightGBM: A highly efficient gradient boosting decision tree . In <i>Proceedings of NeurIPS</i> , pages 3146–3154.	709
657			710
658			711
659			712
660	Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long form question answering . In <i>Proceedings of ACL</i> , pages 3558–3567.	Pei Ke, Hao Zhou, Yankai Lin, Peng Li, Jie Zhou, Xiaoyan Zhu, and Minlie Huang. 2022. CTRL Eval: An unsupervised reference-free metric for evaluating controlled text generation . In <i>Proceedings of ACL</i> , pages 2306–2319.	714
661			715
662			716
663			717
664	Simon Frieder, Luca Pinchetti, Ryan-Rhys Griffiths, Tommaso Salvatori, Thomas Lukasiewicz, Philipp Christian Petersen, Alexis Chevalier, and Julius Berner. 2023. Mathematical capabilities of chatgpt . <i>arXiv preprint arXiv:2301.13867</i> .	Mohammad Khalil and Erkan Er. 2023. Will chatgpt get you caught? rethinking of plagiarism detection . <i>arXiv preprint arXiv:2302.04335</i> .	719
665			720
666			721
667			721
668			721
669	Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling large language models to generate text with citations . <i>arXiv preprint arXiv:2305.14627</i> .	John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. 2023. A watermark for large language models . <i>arXiv preprint arXiv:2301.10226</i> .	722
670			723
671			724
672			725
673	Sebastian Gehrmann, Hendrik Strobelt, and Alexander Rush. 2019. GLTR: Statistical detection and visualization of generated text . In <i>Proceedings of ACL</i> , pages 111–116.	Jan Kocoń, Igor Cichecki, Oliwier Kaszyca, Mateusz Kochanek, Dominika Szydło, Joanna Baran, Julita Bielaniec, Marcin Gruza, Arkadiusz Janz, Kamil Kanclerz, et al. 2023. Chatgpt: Jack of all trades, master of none . <i>arXiv preprint arXiv:2302.10724</i> .	726
674			727
675			728
676	Amelia Glaese, Nat McAleese, Maja Trębacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, et al. 2022. Improving alignment of dialogue agents via targeted human judgements . <i>arXiv preprint arXiv:2209.14375</i> .	Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. 2023. Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense . <i>arXiv preprint arXiv:2303.13408</i> .	731
677			732
678			733
679			734
680			735
681			735
682	Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How close is chatgpt to human experts? comparison corpus, evaluation, and detection . <i>arXiv preprint arXiv:2301.07597</i> .	Md Tahmid Rahman Laskar, M Saiful Bari, Mizanur Rahman, Md Amran Hossen Bhuiyan, Shafiq Joty, and Jimmy Xiangji Huang. 2023. A systematic study and comprehensive evaluation of chatgpt on benchmark datasets . <i>arXiv preprint arXiv:2305.18486</i> .	736
683			737
684			738
685			739
686			740
687	Anna-Carolina Haensch, Sarah Ball, Markus Herklotz, and Frauke Kreuter. 2023. Seeing chatgpt through students’ eyes: An analysis of tiktok data . <i>arXiv preprint arXiv:2303.05349</i> .	Bruce W. Lee, Yoo Sung Jang, and Jason Lee. 2021. Pushing on text readability assessment: A transformer meets handcrafted linguistic features . In <i>Proceedings of EMNLP</i> , pages 10669–10686.	741
688			742
689			743
690			744
691	Steffen Herbold, Annette Hautli-Janisz, Ute Heuer, Zlata Kikteva, and Alexander Trautsch. 2023. AI, write an essay for me: A large-scale comparison of human-written versus chatgpt-generated essays . <i>arXiv preprint arXiv:2304.14276</i> .	Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023a. HaluEval: A large-scale hallucination evaluation benchmark for large language models .	745
692			746
693			747
694			748
695			748
696	José María Gómez Hidalgo, Tiago A Almeida, and Akebo Yamakami. 2012. On the validity of a new sms spam collection . In <i>2012 11th International Conference on Machine Learning and Applications</i> , volume 2, pages 240–245. IEEE.	Linyang Li, Pengyu Wang, Ke Ren, Tianxiang Sun, and Xipeng Qiu. 2023b. Origin tracing and detecting of LLMs . <i>arXiv preprint arXiv:2304.14072</i> .	749
697			750
698			751
699			752
700			753
701	Fred Jelinek, Robert L Mercer, Lalit R Bahl, and James K Baker. 1977. Perplexity—a measure of the difficulty of speech recognition tasks . <i>The Journal of the Acoustical Society of America</i> , 62(S1):S63–S63.	Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries . In <i>Text Summarization Branches Out</i> , pages 74–81.	752
702			753
703			754
704			754
		Hanmeng Liu, Ruoxi Ning, Zhiyang Teng, Jian Liu, Qiji Zhou, and Yue Zhang. 2023a. Evaluating the logical reasoning ability of chatgpt and gpt-4 . <i>arXiv preprint arXiv:2304.03439</i> .	755
			756
			757
			758

759	Yikang Liu, Ziyin Zhang, Wanyang Zhang, Shisen Yue, Xiaojing Zhao, Xinyuan Cheng, Yiwen Zhang, and Hai Hu. 2023b. ArguGPT: evaluating, understanding and identifying argumentative essays generated by gpt models. <i>arXiv preprint arXiv:2304.07666</i> .	814
760		815
761		816
762		817
763		
764	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. <i>arXiv preprint arXiv:1907.11692</i> .	818
765		819
766		820
767		821
768		822
769	Yaojie Lu, Qing Liu, Dai Dai, Xinyan Xiao, Hongyu Lin, Xianpei Han, Le Sun, and Hua Wu. 2022. Unified structure generation for universal information extraction. In <i>Proceedings of ACL</i> , pages 5755–5772.	823
770		824
771		825
772		826
773	Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature. <i>arXiv preprint arXiv:2301.11305</i> .	827
774		828
775		829
776		830
777		831
778		832
779	Oluwatosin Ogundare and Gustavo Quiros Araya. 2023. Comparative analysis of chatgpt and the evolution of language models. <i>arXiv preprint arXiv:2304.02468</i> .	833
780		834
781	OpenAI. 2023. Gpt-4 technical report., <i>OpenAI</i> .	835
782		836
783	Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. In <i>Proceedings of NAACL-HLT</i> , pages 1267–1273.	837
784		
785		
786	Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don’t know: Unanswerable questions for SQuAD. In <i>Proceedings of ACL</i> , pages 784–789.	838
787		839
788		840
789	Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In <i>Proceedings of EMNLP</i> , pages 2383–2392.	841
790		842
791		
792		
793	Paulo Shakarian, Abhinav Koyyalamudi, Noel Ngu, and Lakshminivihari Mareedu. 2023. An independent evaluation of chatgpt on mathematical word problems (mwp). <i>arXiv preprint arXiv:2302.13814</i> .	843
794		844
795		845
796		846
797	R. Siddiqui. 2019. Sarcasmania: Sarcasm exposed! http://www.kaggle.com/rmsharks4/sarcasmania-dataset .	847
798		
799		
800	Ruixiang Tang, Yu-Neng Chuang, and Xia Hu. 2023. The science of detecting llm-generated texts. <i>arXiv preprint arXiv:2303.07205</i> .	848
801		849
802		850
803	Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019a. Superglue: A sticker benchmark for general-purpose language understanding systems. In <i>Proceedings of NeurIPS</i> , pages 3261–3275.	851
804		852
805		853
806		854
807		855
808		856
809	Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019b. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In <i>Proceedings of ICLR</i> .	857
810		858
811		859
812		860
813		861
	Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. <i>Transactions of the Association for Computational Linguistics</i> , 7:625–641.	862
		863
		864
		865
		866
	Xiang Wei, Xingyu Cui, Ning Cheng, Xiaobin Wang, Xin Zhang, Shen Huang, Pengjun Xie, Jinan Xu, Yufeng Chen, Meishan Zhang, et al. 2023. Zero-shot information extraction via chatting with chatgpt. <i>arXiv preprint arXiv:2302.10205</i> .	
	Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In <i>Proceedings of the 26th International Conference on World Wide Web, WWW 2017, Perth, Australia, April 3-7, 2017</i> , pages 1391–1399.	
	Yichong Xu, Jingjing Liu, Jianfeng Gao, Yelong Shen, and Xiaodong Liu. 2017. Towards human-level machine reading comprehension: Reasoning and inference with multiple strategies. <i>arXiv preprint arXiv:1711.04964</i> .	
	Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Bing Yin, and Xia Hu. 2023. Harnessing the power of LLMs in practice: A survey on chatgpt and beyond. <i>arXiv preprint arXiv:2304.13712</i> .	
	Xiao Yu, Yuang Qi, Kejiang Chen, Guoqiang Chen, Xi Yang, Pengyuan Zhu, Weiming Zhang, and Nenghai Yu. 2023. Gpt paternity test: Gpt generated text detection with gpt genetic inheritance. <i>arXiv preprint arXiv:2305.12519</i> .	
	Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2023. AGIEval: A human-centric benchmark for evaluating foundation models. <i>arXiv preprint arXiv:2304.06364</i> .	
	Wangchunshu Zhou, Yuchen Eleanor Jiang, Peng Cui, Tiannan Wang, Zhenxin Xiao, Yifan Hou, Ryan Cotterell, and Mrinmaya Sachan. 2023. RecurrentGPT: Interactive generation of (arbitrarily) long text. <i>arXiv preprint arXiv:2305.13304</i> .	
	A Evaluation Details	853
	A.1 Source Benchmark Statistics	854
	To compare ChatGPT’s performance with its previous version and SOTA, we tested the performance of ChatGPT with the same prompts and questions in prior research (Guo et al., 2023; Kocoń et al., 2023; Shakarian et al., 2023), the corresponding tasks or datasets are as follows:	855
		856
		857
		858
		859
		860
	• Aggression and AggressionPer. The Aggression and AggressionPer tasks, which are based on the Wikipedia Talk Labels dataset (Wulczyn et al., 2017), are offensive language detection tasks. These tasks are presented in the form of binary classification to classify text as either aggressive or	861
		862
		863
		864
		865
		866

not. Specifically, AggressionPer uses the personalized variant by providing user annotations with prompts.

- **CoLa.** The CoLa task, which is based on The Corpus of Linguistic Acceptability (Warstadt et al., 2019), is a linguistic acceptability task that requires classifying input text as either grammatically correct or not.

- **ColBERT.** The ColBERT task, which is based on the ColBERT dataset (Annamoradnejad and Zoghi, 2020), is a humor recognition task. This task is binary classification, which is to classify text as either humor or not humor.

- **Sarcasm** The Sarcasm task, which is based on the Sarcasmania dataset (Siddiqui, 2019), is sarcasm classification task. This task is presented in the form of binary classification, with the task being to classify text as either sarcastic or non-sarcastic.

- **Spam.** The Spam task, which is based on the SMS Spam Collection v.1 (Hidalgo et al., 2012), is a spam detection task. This task is presented in the form of classifying text as either spam or not.

- **WordContext.** The WordContext task, which is based on the Wic dataset (Pilehvar and Camacho-Collados, 2019), is a word sense disambiguation task. This task is in the form of binary pair classification to test whether the highlighted word in two different contexts expresses the same meaning.

- **TextEntail.** The TextEntail task, which is based on one of the SuperGLUE benchmark (Wang et al., 2019a), is recognizing textual entailment task. This task is binary sentence pair classification, with the task being to test whether the two sentences are "entailed" or "not_entailed".

- **WNLI.** The WNLI task, which is based on the SuperGLUE Winograd NLI dataset from the GLUE benchmark (Wang et al., 2019b), is a natural language inference task. This task is binary sentence pair classification, which is to test whether the ambiguous pronoun in the second sentence is entailed with the pronoun in the first sentence.

- **SQuAD.** The SQuAD task is based on SQuAD v2.0 (Rajpurkar et al., 2018), an extractive question-answering dataset. The language model should answer the question and infer whether the answer can be found in the given context by giving its position.

- **GSM8K.** The GSM8K task is based on the GSM8K (Cobbe et al., 2021), a maths word problem dataset. This task is framed as mathematical

reasoning. Chain-of-thought reasoning prompting is used for this task.

- **GoEmo and GoEmoPer.** The GoEmo and GoEmoPer tasks, which are based on the GoEmotions dataset (Demszky et al., 2020), are sentiment analysis tasks. These tasks are framed as multi-label classification, with the tasks being to classify the emotions in a piece of text as emotion labels in 27 possible emotions plus neutral. Specifically, GoEmoPer tasks use individual annotator annotations with prompts.

- **TweetEmoji, TweetSent, TweetStance.** The TweetEmoji task, TweetSent task, and TweetStance task, which are based on the Tweeteval dataset (Barbieri et al., 2020), are emoji prediction task, sentiment analysis task and stance detection task respectively. These tasks are presented in the form of multi-class classification. The TweetEmoji task focuses on emoji prediction for a given tweet, the TweetSent task aims to classify the emotion of a given tweet as negative, neutral, or positive, and the TweetStance task focuses on detecting the stance of a piece of text as none, against or favor.

- **ReAding.** The ReAding task is based on the RACE dataset (Xu et al., 2017), a reading comprehension dataset. This task is framed as multiple choice QA to choose the most likely option of four statements.

A.2 Evaluation Stability

To rule out the possibility that the changes in ChatGPT are just from randomness and confirm the stability of our evaluation results, we conduct 3 trails for each data everyday (sending the same question to OpenAI api for 3 times with the same hyper-parameters and prompts). We compare the difference of evaluating and analyzing ChatGPT by using results of 1 query and 3 repeated queries' average score for each data in Figure 5 and 6. We observed that, after taking the average of three results, the fluctuations in the feature and evaluation scores of ChatGPT decrease. Specifically, in the stable range after April, the curve of only one trail's result exhibits relatively large fluctuations, but it doesn't reach the level of difference observed between March and April due to version changes. These phenomena indicate that the observed temporal variations are indeed primarily attributed to version changes rather than randomness.

We also calculate the Pearson correlation coefficient using results of 1 query and 3 repeated

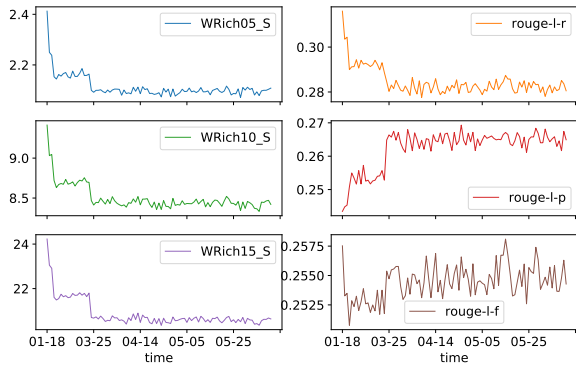


Figure 5: Trend of ChatGPT’s rouge scores and linguistic features on ChatLog-Daily dataset using **only 1 trail**’s score every day from March 5 to June 10, 2023. Note that the initial data at January 18 comes from HC3 (Guo et al., 2023) dataset. WRich_S denotes Semantic Richness coverage on several topics extracted from Wikipedia.

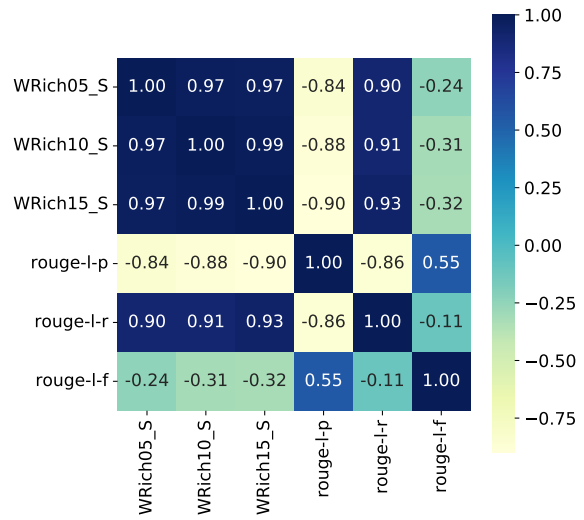


Figure 7: **Pearson correlation coefficient** between the three rouge metrics rouge-l-{precision, recall, f1} and three semantic features WRich{05,10,15}_S, on the ChatLog-Daily dataset using **only 1 trail**’s score. WRich_S denotes Semantic Richness coverage on several topics extracted from Wikipedia.

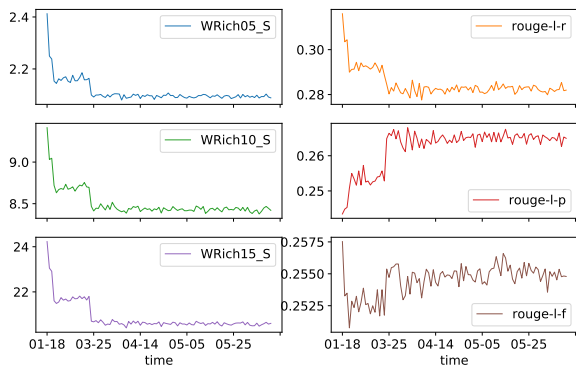


Figure 6: Trend of ChatGPT’s rouge scores and linguistic features on ChatLog-Daily dataset using **3 trails’ average score** every day from March 5 to June 10, 2023. Note that the initial data at January 18 comes from HC3 (Guo et al., 2023) dataset. WRich_S denotes Semantic Richness coverage on several topics extracted from Wikipedia.

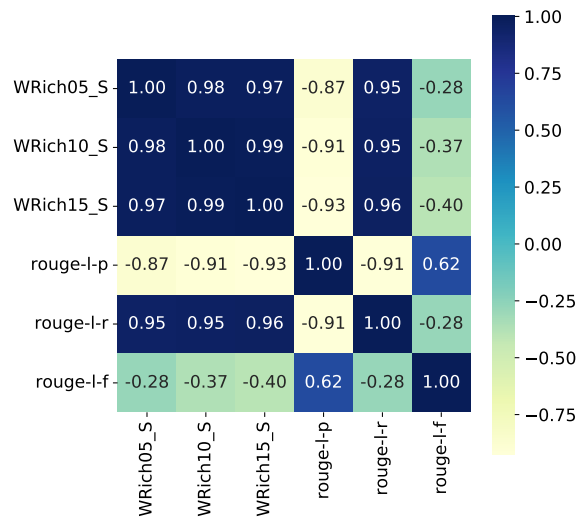


Figure 8: **Pearson correlation coefficient** between the three rouge metrics rouge-l-{precision, recall, f1} and three semantic features WRich{05,10,15}_S, on the ChatLog-Daily dataset using **3 trails’ average score**. WRich_S denotes Semantic Richness coverage on several topics extracted from Wikipedia.

968 queries' average score for each data in Figure 7
969 and 8. The results indicate a stronger correlation
970 between the features and rouge scores of ChatGPT
971 when using the average value of three responses.
972 This suggests that employing the averaging ap-
973 proach can help ChatGPT's features regress to a
974 reasonable range, thus establishing a stable rela-
975 tionship with the rouge metric.

976 **A.3 Case Study**

977 This section contains sampled examples from
978 the ChatLog-Monthly dataset for some evaluation
979 tasks. ChatGPT's performance on these tasks is not
980 perfect. When the text is short or contains little in-
981 formation, it may have difficulty making judgments.
982 In addition, the prompt is not always valid, so Chat-
983 GPT may give out-of-bounds results or its output
984 may be in a format that requires post-processing.
985 The following examples from Table 7 to Table 13
986 show some of the above situations, including ab-
987 solutely correct (content and format), correct but
988 wrong formatted, unable to answer, incorrect, out
989 of range, etc.

QUESTION: Which one of the attributes: "aggressive", "non-aggressive" describes a given text? Text: Lets see...no final warning, and talk-page stalking. Nice.

MODEL ANSWER (INCORRECT): expected: non-aggressive, answer: aggressive

QUESTION: Which one of the attributes: "aggressive", "non-aggressive" describes a given text? Text: Personal Attack Check the history of that IP and you will see that it is not at all a personal attack. Please also remember that you can't blank talk pages that don't belong to you.

MODEL ANSWER (INCORRECT): expected: non-aggressive, answer: aggressive

QUESTION: Knowing that a certain person described each text with a given attribute: Text: 'Benji dude. I wouldnt bother these lot on here are clearly deluded. If MSNBC/BBC or CNN are not reliable sources because they go back to an interview then this is one of the reasons wikipedia is critizised for, and Portia keeps reverting edits made and class them as vandalism yet coming back with "I understand your passion". You cannot find any retraction by Stiers or his people if you google it, if were really such an issue for Stiers he would have retracted it immediatly and it hasnt publically, sooo silly.'; Attribute: "non-aggressive" Text: Xman52 is vandalizing the Hopkins page.; Attribute: "non-aggressive" Text: RE: the gods template The text is different, looks better then the plain ass text used on everything else. And once again, who gives a shit? I use it on a few pages, what does it matter? ; Attribute: "non-aggressive" write which one of the attributes: "aggressive", "non-aggressive" would this person describe a given text? Write your answer in the form of a Python list containing the appropriate attribute. Text: HA! That is ridiculous. I've seen his posts. He actually believes what he says and that the WTC was taken down by fire and damage. What is he a mailboy?

MODEL ANSWER (CORRECT): expected: aggressive, answer: aggressive

Table 7: Aggression and AggressionPer task examples.

QUESTION: Describe the grammatic acceptability of the given text. Choose your answer from provided list and map your answer with following unacceptable: 0, acceptable: 1 and return an integer as a result. Do not explain yourself. Input data: Text: Michelle kept the desk with the papers. Possible acceptability: unacceptable, acceptable

MODEL ANSWER (INCORRECT): expected: 0, answer: 1

QUESTION: Describe the grammatic acceptability of the given text. Choose your answer from provided list and map your answer with following unacceptable: 0, acceptable: 1 and return an integer as a result. Do not explain yourself. Input data: Text: Most of the fruit is rotten. Possible acceptability: unacceptable, acceptable

MODEL ANSWER (CORRECT): expected: 1, answer: 1

QUESTION: Answer should be only python list containing string either "spam" or "not spam". Content of the SMS: IMPORTANT MESSAGE. This is a final contact attempt. You have important messages waiting out our customer claims dept. Expires 13/4/04. Call 08717507382 NOW!.

MODEL ANSWER (CORRECT): expected: spam, answer: spam

QUESTION: Answer should be only python list containing string either "spam" or "not spam". Content of the SMS: I don't know u and u don't know me. Send CHAT to 86688 now and let's find each other! Only 150p/Msg revd. HG/Suite342/2Lands/Row/W1J6HL LDN. 18 years or over..

MODEL ANSWER (INCORRECT): expected: spam, answer: not spam

QUESTION: Which one of the attributes: "funny", "not funny" describes a given text? Write your answer in the form of a Python list containing the appropriate attribute. Text: But is it too late to try?

MODEL ANSWER (NONE): expected: not funny, answer: not enough information is given to determine whether the text is funny or not funny. both attributes could be possible. therefore, the python list containing the appropriate attribute cannot be provided.

QUESTION: Which one of the attributes: "funny", "not funny" describes a given text? Write your answer in the form of a Python list containing the appropriate attribute. Text: Eleven days until the National Convention! Are you as excited as we are? #FillYourBucket #cantwait

MODEL ANSWER (CORRECT): expected: funny, answer: ["funny"]

Table 8: CoLa, ColBERT, Sarcasm and Spam task examples.

QUESTION: Write the reasoning and highlight the answer at the end in format: 'Answer: '. The question is: Carlos is planting a lemon tree. The tree will cost \$90 to plant. Each year it will grow 7 lemons, which he can sell for \$1.5 each. It costs \$3 a year to water and feed the tree. How many years will it take before he starts earning money on the lemon tree?

MODEL ANSWER (INCORRECT AND WRONG UNDERSTANDING): expected: 13, answer: to calculate when carlos will start earning money on the lemon tree, we need to compare the income to the expenses.

each year, carlos will earn 7 lemons x \$1.5/lemon = \$10.5 in revenue. his yearly costs are \$3 to water/feed the tree.

so each year, his profit will be \$10.5 - \$3 = \$7.5.

we need to figure out how many years it will take to earn back the initial \$90 investment.

$\$90 / \$7.5 = 12$

answer: it will take 12 years before carlos starts earning money on the lemon tree.

QUESTION: Which one of the attributes: "funny", "not funny" describes a given text? Write your answer in the form of a Python list containing the appropriate attribute. Text: Eleven days until the National Convention! Are you as excited as we are? #FillYourBucket #cantwait

MODEL ANSWER (CORRECT): expected: funny, answer: ["funny"]

QUESTION: Write the reasoning and highlight the answer at the end in format: 'Answer: '. The question is: Melanie is a door-to-door saleswoman. She sold a third of her vacuum cleaners at the green house, 2 more to the red house, and half of what was left at the orange house. If Melanie has 5 vacuum cleaners left, how many did she start with?

MODEL ANSWER (INCORRECT AND WRONG CALCULATION): expected: 18, answer: to solve this problem, we need to work backwards from the end. melanie has 5 vacuum cleaners left after selling a third at the green house, 2 more to the red house, and half of what was left at the orange house.

let's start with the half sold at the orange house. if half of what was left after selling to the green and red houses is 5, then the total left after selling to the green and red houses must be 10.

to find out how many she started with, we need to work backwards again. we know that she sold a third at the green house and 2 more to the red house, which means that the total sold at those two houses is:

$(1/3)x + 2$

where x is the total number of vacuum cleaners she started with.

so we can set up an equation:

$x - ((1/3)x + 2) - ((1/2)(x - (1/3)x - 2)) = 5$

simplifying:

$x - (1/3)x - 2 - (1/2)x + (1/6)x + 1 = 5$

multiplying both sides by 6 to eliminate the fractions:

$6x - 2x - 24 - 3x + x + 6 = 30$

simplifying:

$2x - 18 = 30$

$2x = 48$

$x = 24$

answer: melanie started with 24 vacuum cleaners.

QUESTION: Write the reasoning and highlight the answer at the end in format: 'Answer: '. The question is: Eliza's rate per hour for the first 40 hours she works each week is \$10. She also receives an overtime pay of 1.2 times her regular hourly rate. If Eliza worked for 45 hours this week, how much are her earnings for this week?

MODEL ANSWER (CORRECT): eliza's earnings for the first 40 hours:

40 hours x \$10/hour = \$400

eliza's overtime earnings for the additional 5 hours:

5 hours x (\$10/hour x 1.2) = \$60

total earnings for eliza this week:

$\$400 + \$60 = \$460$

answer: \$460

Table 9: MathQA task examples.

QUESTION: Are the given texts expressing the same sense of the noun board: “Room and board .” and “He nailed boards across the windows .”? Return your answer as a letter: "T" if the sense is the same or "F" if it's not the same.

MODEL ANSWER (CORRECT): expected: F, answer: F

QUESTION: Are the given texts expressing the same sense of the verb circulate: “Circulate a rumor .” and “This letter is being circulated among the faculty .”? Return your answer as a letter: "T" if the sense is the same or "F" if it's not the same.

MODEL ANSWER (INCORRECT): expected: F, answer: T

QUESTION: Having the sentence "Always before, Larry had helped Dad with his work. But he could not help him now, for Dad said that his boss at the railroad company would not want anyone but him to work in the office." tell me if the following sentence "He could not help Larry now." is true or false? Answer a number "0" if false or "1" if true.

MODEL ANSWER (CORRECT BUT WRONG FORMAT): expected: 0, answer: False. The sentence should actually say "Dad could not help Larry now."

QUESTION: Having the sentence "There is a pillar between me and the stage, and I can't see it." tell me if the following sentence "I can't see around the stage." is true or false? Answer a number "0" if false or "1" if true.

MODEL ANSWER (INCORRECT): expected: 1, answer: 0

QUESTION: Having premise "Nokia, Texas Instruments and other leading makers of mobile phones have formally complained to Brussels that Qualcomm, the US mobile chipmaker, has unfairly used its patents on 3G technologies." judge if the following hypothesis "Texas Instruments produces mobile phones." are logically connected with the premise? Answer "entailment" if yes, or "not_entailment" if no.

MODEL ANSWER (CORRECT): expected: entailment, answer: entailment

Table 10: WordContext, textEntail and WNLI task examples.

QUESTION: The context is 'In 2006, a toxic waste spill off the coast of Côte d'Ivoire, from a European ship, prompted the Commission to look into legislation against toxic waste. Environment Commissioner Stavros Dimas stated that "Such highly toxic waste should never have left the European Union". With countries such as Spain not even having a crime against shipping toxic waste, Franco Frattini, the Justice, Freedom and Security Commissioner, proposed with Dimas to create criminal sentences for "ecological crimes". The competence for the Union to do this was contested in 2005 at the Court of Justice resulting in a victory for the Commission. That ruling set a precedent that the Commission, on a supranational basis, may legislate in criminal law – something never done before. So far, the only other proposal has been the draft intellectual property rights directive. Motions were tabled in the European Parliament against that legislation on the basis that criminal law should not be an EU competence, but was rejected at vote. However, in October 2007, the Court of Justice ruled that the Commission could not propose what the criminal sanctions could be, only that there must be some.'. Give the result in JSON format with keys 'text' and 'answer_start' with values: Answer to the question 'In which year did a toxic waste spill from a European ship prompt the Commission to look into legislation against waste?' in 1 words, the number of character at which the answer appears in the context. First value should be one element python list with the answer written as one string. Second value should be one element python list.

MODEL ANSWER (CORRECT): expected: {'text': ['2006', '2006', '2006'], 'answer_start': [3, 3, 3]}, answer: {"text": ["2006"], "answer_start": [3]}

QUESTION: The context is 'Today, roller skating is easy and fun. But a long time ago, it wasn't easy at all. Before 1750, the idea of skating didn't exist. That changed because of a man named Joseph Merlin. Merlin's work was making musical instruments. In his spare time he liked to play the violin. Joseph Merlin was a man of ideas and dreams. People called him a dreamer. One day Merlin received an invitation to attend a fancy dress ball. He was very pleased and a little excited. As the day of the party came near, Merlin began to think how to make a grand entrance at the party. He had an idea. He thought he would get a lot of attention if he could skate into the room. Merlin tried different ways to make himself roll. Finally, he decided to put two wheels under each shoe. These were the first roller skates. Merlin was very proud of his invention and dreamed of arriving at the party on wheels while playing the violin. On the night of the party Merlin rolled into the room playing his violin. Everyone was astonished to see him. There was just one problem. Merlin had no way to stop his roller skates. He rolled on and on. Suddenly, he ran into a huge mirror that was hanging on the wall. Down fell the mirror, breaking to pieces. Nobody forgot Merlin's grand entrance for a long time!'. Give correct answer to the question: 'People thought Merlin was a dreamer because he _.' from A, B, C, D where A, B, C, D correspond to each item in python list ['often gave others surprises', 'was a gifted musician', 'was full of imagination', 'invented the roller skates']. The answer must be only one letter.

MODEL ANSWER (CORRECT): expected: C, answer: C

QUESTION: The context is 'Children have their own rules in playing games. They seldom need a judge and rarely trouble to keep scores. They don't care much about who wins or loses, and it doesn't seem to worry them if the game is not finished. Yet, they like games that depend a lot on luck, so that their personal abilities cannot be directly compared. They also enjoy games that move in stages, in which each stage, the choosing of leaders, the picking-up of sides, or the determining of which side shall start, is almost a game in itself. Grown-ups can hardly find children's games exciting, and they often feel puzzled at why their kids play such simple games again and again. However, it is found that a child plays games for very important reasons. He can be a good player without having to think whether he is a popular person, and he can find himself being a useful partner to someone of whom he is ordinarily afraid. He becomes a leader when it comes to his turn. He can be confident, too, in particular games, that it is his place to give orders, to pretend to be dead, to throw a ball actually at someone, or to kiss someone he caught. It appears to us that when children play a game they imagine a situation under their control. Everyone knows the rules, and more important, everyone plays according to the rules. Those rules may be childish, but they make sure that every child has a chance to win.'. Give correct answer to the question: 'To become a leader in a game the child has to _.' from A, B, C, D where A, B, C, D correspond to each item in python list ['play well', 'be popular among his playmates', 'be confident in himself', 'wait for his turn']. The answer must be only one letter.

MODEL ANSWER (INCORRECT AND WRONG FORMAT): expected: D, answer: C (be confident in himself)

Table 11: SQuAD and ReAding task examples.

QUESTION: From the given list of all emotions, choose the ones that the input text arouses in most people reading it. Write your answer in the form of a Python list containing exactly 1 selected most matching emotion. List of all emotions: admiration, amusement, anger, annoyance, approval, caring, confusion, curiosity, desire, disappointment, disapproval, disgust, embarrassment, excitement, fear, gratitude, grief, joy, love, nervousness, optimism, pride, realization, relief, remorse, sadness, surprise, neutral. Input text: Well, there's cubs and otters too.

MODEL ANSWER (CORRECT): expected: neutral, answer: ["neutral"]

QUESTION: From the given list of all emotions, choose the ones that the input text arouses in most people reading it. Write your answer in the form of a Python list containing exactly 1 selected most matching emotion. List of all emotions: admiration, amusement, anger, annoyance, approval, caring, confusion, curiosity, desire, disappointment, disapproval, disgust, embarrassment, excitement, fear, gratitude, grief, joy, love, nervousness, optimism, pride, realization, relief, remorse, sadness, surprise, neutral. Input text: I would try rubbing alcohol, high percentage. Looks greasy to me.

MODEL ANSWER (INCORRECT): expected: optimism, answer: ["disgust"]

QUESTION: From the given list of all emotions, choose the ones that the input text arouses in most people reading it. List of all emotions: admiration, amusement, anger, annoyance, approval, caring, confusion, curiosity, desire, disappointment, disapproval, disgust, embarrassment, excitement, fear, gratitude, grief, joy, love, nervousness, optimism, pride, realization, relief, remorse, sadness, surprise, neutral. Text: UN report says population before war: 300,000 [NAME] population before war: 2,000,000. Write your answer in the form of a Python list containing exactly 1 unique selected most matching emotion. Do not explain yourself.

MODEL ANSWER (INCORRECT AND OUT OF RANGE): expected: neutral, answer: ["disbelief"]

QUESTION: Knowing that a certain person described known texts with provided emotions: Known text: Honestly if I didn't have this many favorite characters not yet in the game, I would probably waste my 7 votes on Merlinus.; Emotion: love. Known text: The Japanese and the British can do it well, but they probably get like 10x the funding per capita.; Emotion: pride With what emotions would this person describe the unknown text? List of all emotions: admiration, amusement, anger, annoyance, approval, caring, confusion, curiosity, desire, disappointment, disapproval, disgust, embarrassment, excitement, fear, gratitude, grief, joy, love, nervousness, optimism, pride, realization, relief, remorse, sadness, surprise, neutral. Unknown text: the most romantic thing that can possibly happen to me. Write your answer in the form of a Python list containing exactly 1 unique selected most matching emotion. Do not explain yourself.

MODEL ANSWER (CORRECT): expected: love, answer: ["love"]

Table 12: GoEmo and GoEmoPer task examples.

QUESTION: Describe the sentiment of the given text. Choose your answer from provided list and map your answer with following negative: 0, neutral: 1, positive: 2 and return an integer as a result. Do not explain yourself. Input data: Text: An interesting security vulnerability - albeit not for the everyday car thief Possible sentiment: negative, neutral, positive

MODEL ANSWER (CORRECT): expected: 1, answer: 1

QUESTION: Describe the sentiment of the given text. Choose your answer from provided list and map your answer with following negative: 0, neutral: 1, positive: 2 and return an integer as a result. Do not explain yourself. Input data: Text: Can someone tell me where I can get a good red eyeshadow from?? (Not animal testing brands pls x) Possible sentiment: negative, neutral, positive

MODEL ANSWER (CORRECT BUT WRONG FORMAT): expected: 1, answer: Neutral: 1

QUESTION: Assuming that you have to describe the stance from the given list which is evoked after reading the text about Hillary Clinton by majority of people, which one would you select? Map your answer with following none: 0, against: 1, favor: 2. Do not explain yourself. Input data:

Text: President Jimmy Carter: There's no doubt Hillary will get the nomination. And when she does I'll be happy to support her #SemST

Possible stances: none, against, favor

MODEL ANSWER (CORRECT): expected: 2, answer: 2

Table 13: TweetEmoji, TweetSent and TweetStance task examples.

B Full Feature Analysis

In this section, we first introduce the feature details of our feature extraction process to complete the full category in Table 5. Then we will show full knowledge and linguistic feature analysis results of representative feature for each category. In section 4.3, we just show 3 features (WRich_S) in AdSem category, actually there are 9 categories of total 265 features.

B.1 Feature Meaning

We list representative features of each category in the following Tables. Table 14 to 18 introduce the linguistic feature categories. As those linguistic features are extracted by the LingFeat³ tool, Lee et al. (2021) has introduced the entire 255 linguistic features in their paper, please refer to their paper’s appendix for full linguistic feature information. On the other hand, we also extract 10 knowledge features using UIE (Lu et al., 2022) and CogIE (Jin et al., 2021) tools. Features in Table 19 and 20 are named entities and opinions extracted by UIE⁴. And features in Table 21 and Table 22 are entity relations and semantic frames extracted by CogIE⁵.

³<https://github.com/brucewlee/lingfeat>

⁴<https://github.com/universal-ie/UIE>

⁵<https://github.com/jinzhuror/CogIE>

<i>idx</i>	Code	Definition
1	WRich05_S	Richness, 50 topics extracted from Wikipedia Dump
2	WClar05_S	Clarity, 50 topics extracted from Wikipedia Dump
3	WNois05_S	Noise, 50 topics extracted from Wikipedia Dump
4	WTopc05_S	# of topics, 50 topics extracted from Wikipedia Dump
5	WRich10_S	Richness, 100 topics extracted from Wikipedia Dump
6	WClar10_S	Clarity, 100 topics extracted from Wikipedia Dump
7	WNois10_S	Noise, 100 topics extracted from Wikipedia Dump
8	WTopc10_S	# of topics, 100 topics extracted from Wikipedia Dump
9	WRich15_S	Richness, 150 topics extracted from Wikipedia Dump
10	WClar15_S	Clarity, 150 topics extracted from Wikipedia Dump
...
20	BTopc05_S	# of topics, 50 topics extracted from WeeBit Corpus
21	BRich10_S	Richness, 100 topics extracted from WeeBit Corpus
22	BClar10_S	Clarity, 100 topics extracted from WeeBit Corpus
23	BNois10_S	Noise, 100 topics extracted from WeeBit Corpus
24	BTopc10_S	# of topics, 100 topics extracted from WeeBit Corpus
25	BRich15_S	Richness, 150 topics extracted from WeeBit Corpus
26	BClar15_S	Clarity, 150 topics extracted from WeeBit Corpus
27	BNois15_S	Noise, 150 topics extracted from WeeBit Corpus
28	BTopc15_S	# of topics, 150 topics extracted from WeeBit Corpus
29	BRich20_S	Richness, 200 topics extracted from WeeBit Corpus
30	BClar20_S	Clarity, 200 topics extracted from WeeBit Corpus
32	BTopc20_S	# of topics, 200 topics extracted from WeeBit Corpus
33	ORich05_S	Richness, 50 topics extracted from OneStop Corpus
34	OClar05_S	Clarity, 50 topics extracted from OneStop Corpus
35	ONois05_S	Noise, 50 topics extracted from OneStop Corpus
36	OTopc05_S	# of topics, 50 topics extracted from OneStop Corpus
...
48	OTopc20_S	# of topics, 200 topics extracted from OneStop Corpus

Table 14: Advanced Semantic(AdSem) feature category, including Wikipedia Knowledge Features (W), WeeBit Knowledge Features (B), OneStop Knowledge Features (O).

<i>idx</i>	Code	Definition
49	to_EntiM_C	total number of Entities Mentions
50	as_EntiM_C	average number of Entities Mentions per sentence
51	at_EntiM_C	average number of Entities Mentions per token (word)
52	to_UEnti_C	total number of unique Entities
53	as_UEnti_C	average number of unique Entities per sentence
54	at_UEnti_C	average number of unique Entities per token (word)
55	ra_SSToT_C	ratio of SS transitions : total, count from Entity Grid
56	ra_SOToT_C	ratio of SO transitions : total, count from Entity Grid
57	ra_SXToT_C	ratio of SX transitions : total, count from Entity Grid
58	ra_SNTToT_C	ratio of SN transitions : total, count from Entity Grid
...
72	LoCohPW_S	Local Coherence for PW score from Entity Grid
73	LoCohPU_S	Local Coherence for PU score from Entity Grid
74	LoCoDPA_S	Local Coherence dist. for PA score from Entity Grid
75	LoCoDPW_S	Local Coherence dist. for PW score from Entity Grid
76	LoCoDPU_S	Local Coherence dist. for PU score from Entity Grid

Table 15: Discourse (Disco) feature category, including Entity Density Features (EnDF). Entity Grid Features (EnGF).

<i>idx</i>	Code	Definition
77	to_NoPhr_C	total count of Noun phrases
78	as_NoPhr_C	average count of Noun phrases per sentence
79	at_NoPhr_C	average count of Noun phrases per token
80	ra_NoVeP_C	ratio of Noun phrases : Verb phrases count
...
124	ra_AvAjP_C	ratio of Adv phrases : Adj phrases count
125	to_TreeH_C	total parsed Tree Height of all sentences
126	as_TreeH_C	average parsed Tree Height per sentence
127	at_TreeH_C	average parsed Tree Height per token
128	to_FTree_C	total length of Flattened parsed Trees
129	as_FTree_C	average length of Flattened parsed Trees per sentence
130	at_FTree_C	average length of Flattened parsed Trees per token
131	to_NoTag_C	total count of Noun tags
132	as_NoTag_C	average count of Noun tags per sentence
133	at_NoTag_C	average count of Noun tags per token
...
185	ra_CoFuW_C	ratio of Content words to Function words

Table 16: Syntactic (Synta) feature category, including Phrasal Features (PhrF), Tree Structural Features (TrSF), Part-of-Speech Features (POSF)

<i>idx</i>	Code	Definition
186	SimpNoV_S	unique Nouns/total Nouns #Noun Variation
187	SquaNoV_S	(unique Nouns**2)/total Nouns #Squared Noun Variation
188	CorrNoV_S	unique Nouns/sqrt(2*total Nouns) #Corrected Noun Var
...
197	CorrAvV_S	unique Adv/sqrt(2*total Adv) #Corrected Adv Var
198	SimpTTR_S	unique tokens/total tokens #TTR
199	CorrTTR_S	unique/sqrt(2*total) #Corrected TTR
200	BiLoTTR_S	log(unique)/log(total) #Bi-Logarithmic TTR
201	UberTTR_S	(log(unique)) ² /log(total/unique) #Uber
202	MTLDTTR_S	#Measure of Textual Lexical Diversity (TTR, 0.72)
203	to_AAKuW_C	total AoA (Age of Acquisition) of words, Kuperman
204	as_AAKuW_C	average AoA of words per sentence, Kuperman
205	at_AAKuW_C	average AoA of words per token, Kuperman
...
217	at_AACoL_C	average AoA of lem, Cortese and K norm per token
218	to_SbFrQ_C	total SubtlexUS FREQcount value
219	as_SbFrQ_C	average SubtlexUS FREQcount value per sentence
220	at_SbFrQ_C	average SubtlexUS FREQcount value per token
...
241	at_SbLIC_C	average SubtlexUS Lg10CD value per token

Table 17: Lexico Semantic (LxSem) feature category, including Variation Ratio Features (VarF), Type Token Ratio Features (TTRF), Psycholinguistic Features (PsyF) and Word Familiarity Features (WorF).

<i>idx</i>	Code	Definition
242	TokSenM_S	total count of tokens x total count of sentence
243	TokSenS_S	sqrt(total count of tokens x total count of sentence)
244	TokSenL_S	log(total count of tokens)/log(total count of sent)
245	as-Token_C	average count of tokens per sentence
246	as_Sylla_C	average count of syllables per sentence
247	at_Sylla_C	average count of syllables per token
248	as_Charac_C	average count of characters per sentence
249	at_Charac_C	average count of characters per token
250	SmogInd_S	Smog Index
251	ColeLia_S	Coleman Liau Readability Score
252	Gunning_S	Gunning Fog Count Score (New, US Navy Report)
253	AutoRea_S	Automated Readability Idx (New, US Navy Report)
254	FleschG_S	Flesch Kincaid Grade Level (New, US Navy Report)
255	LinseaW_S	Linsear Write Formula Score

Table 18: Shallow Feature (ShaF) category.

<i>idx</i>	Code	Definition
256	Person	total count of Person entities
257	Organization	total count of Organization entities
258	Location	total count of Location entities
259	Work	total count of Work entities
260	Time	total count of Time entities

Table 19: Named entity (NE) feature category.

<i>idx</i>	Code	Definition
261	Opinion	total count of Opinion expression words
262	Sentiment	A probability for negative, neutral or positive sentiment

Table 20: Opinion (OP) feature category.

<i>idx</i>	Code	Definition
263	Relation	total count of entity relations

Table 21: Relation Extraction (RE) feature category. Actually there are 500 kinds of relation classes in CogIE tool that we use.

<i>idx</i>	Code	Definition
264	Lexical Unit	total count of evoking words
265	Frame Element	total count of semantic role words

Table 22: Frame-Semantic Parsing (FP) feature category. Actually there are 800 kinds of Lexical Units and 1900 Frame Elements in CogIE tool that we use. Lexical Unit is a word that plays the role of evoking the corresponding frame. Frame Element indicates different semantic roles associated with the frame.

1013 B.2 Knowledge Feature Analysis

1014 In this section, we will demonstrate the knowl- 1063
1015 edge features on our ChatLog-Daily dataset to ob- 1064
1016 serve ChatGPT’s patterns on knowledge. As knowl- 1065
1017 edge features are mostly extracted mentions in sen- 1066
1018 tences that can be linked with entity, relation or 1067
1019 frame in the pre-defined knowledge schemas such 1068
1020 as Wikipedia and FrameNet (Baker et al., 1998), 1069
1021 these knowledge are commonly string type rather 1070
1022 than numerical type. Therefore, it is hard to com- 1071
1023 prehensively describe these string features with our 1072
1024 defined numerical calculation. In this work, we 1073
1025 limit the research scope by just using the count of 1074
1026 these string type features to summarize, for exam-
1027 ple, using the count of relations in one document
1028 to summarize the relation category features. We
1029 leave this to future works for the exploration of
1030 more elaborate designs on knowledge features.

1031 **Entity and Relation Analysis** From the results
1032 presented in Figure 9, we can observe some in-
1033 teresting patterns in ChatGPT’s entity and relation
1034 features over time. Firstly, we note that there is an
1035 overall decrease in the frequency of named entity
1036 features in ChatGPT’s answers, with the average
1037 frequency value for Organization feature decreas-
1038 ing from around 0.8 in January to around 0.4 in
1039 April. Additionally, the most common instance’s
1040 frequency also shows an downward trend, suggest-
1041 ing that ChatGPT’s ability to incorporate concise
1042 named entities in its responses is improving over
1043 time.

1044 Similarly, we can observe an decrease in the fre-
1045 quency of the relation feature, with the average and
1046 maximum frequency value also showing an down-
1047 ward trend across the three dates presented. This
1048 result suggests that ChatGPT has become more ef-
1049 fective and efficient in generating responses that
1050 contain relations between entities over time.

1051 **Opinion Analysis** Figure 10 showcases the fre-
1052 quencies of opinion words and sentiment features
1053 extracted by ChatGPT. These features play a cru-
1054 cial role in understanding the intention that can
1055 make a significant difference in the text’s overall
1056 meaning.

1057 From the results presented in Figure 10, we can
1058 see that the frequency of opinion words in Chat-
1059 GPT’s responses has decreased gradually over time.
1060 Specifically, the average frequency values for these
1061 features have decreased from 1.2 to 0.8. This sug-
1062 gests that ChatGPT has become more objective in

generating responses that express less opinions. 1063

1064 However, we can also see a unchanged value in
1065 the frequency of sentiment classification feature
1066 that indicating positive, negative, and neutral senti-
1067 ment are stable in ChatGPT’s responses over time.
1068 As shown in Table 23, the most frequent sentiment
1069 is negative, maybe because that the questions are
1070 collected from Reddit, which are controversial and
1071 subjective. Actually there are almost the 40% re-
1072 sults are positive and neutral, which demonstrates
1073 a function of ChatGPT to generate responses that
1074 convey the correct sentiment.

1075 **Frame Analysis** From the results presented in
1076 Figure 11, we observe an overall downward trend
1077 in the frequency of semantic frame words over time.
1078 This suggests that ChatGPT is becoming better at
1079 generating responses that contain semantic frames
1080 or frames used to represent entities and events. Sim-
1081 ilarly, we can observe an decreasing trend in the
1082 frequency of argument role features such as Agent,
1083 Patient, and Theme in ChatGPT’s responses over
1084 time. This indicates an improvement in ChatGPT’s
1085 ability to recognize the argument roles of entities
1086 and events in generating responses that convey the
1087 meaning correctly.

1088 Overall, the results presented in Figure 9, 10
1089 and 11 demonstrate that ChatGPT’s patterns on
1090 knowledge are showing improvement over time.
1091 The decreasing frequency values of named entity,
1092 relation, opinion and frame features suggest that
1093 ChatGPT has become more proficient in generating
1094 responses that contain less complex knowledge to
1095 fit the instruction "Explain like I am five". Nonethe-
1096 less, this study is limited in its scope of long-form
1097 QA task on ELI5 (Fan et al., 2019) dataset, and fur-
1098 ther research is required to explore more elaborate
1099 knowledge analysis for more tasks and to evaluate
1100 the generalizability of our findings.

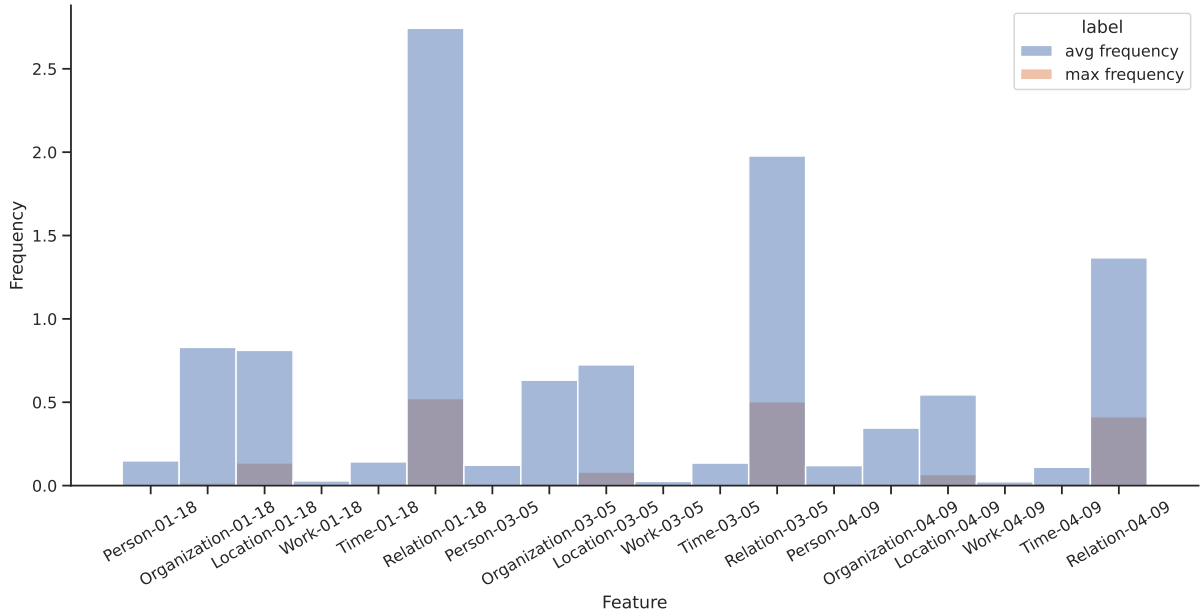


Figure 9: The extracted named entity and relation features’ frequencies across time, where avg frequency denotes the average number of the feature in one ChatGPT answer and max frequency denotes the most common instance’s frequency. Note that we demonstrate three dates’ results in this figure, i.e. 2023-01-18, 2023-03-05 and 2023-04-09, which come from HC3 and ChatLog-Daily dataset respectively.

Type	Feature	Most Common Instances of 3 snapshots		
		2023-01-18	2023-03-05	2023-04-09
Entity and Relation	Person	John Smith	Jesus	Jesus
	Organization	ISIS	Apple	Apple
	Location	United States	Earth	Earth
	Work	The Mona Lisa	work	NY Times
	Time	night	night	school hours
	Relation	subclass of	shares border with	shares border with
Opinion	Opinion	hope	want	want
	Sentiment	negative	negative	negative
Frame	Frame	Causation	Causation	Causation
	Argument	entity	entity	entity

Table 23: Detailed instance allocation of each knowledge feature. Note that we demonstrate three snapshots’ most common instances respectively in this table, i.e. 2023-01-18, 2023-03-05 and 2023-04-09, which come from HC3 and ChatLog-Daily dataset.

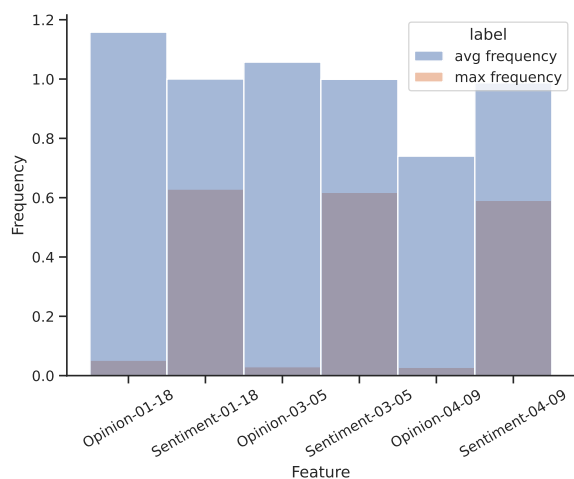


Figure 10: The extracted opinion and sentiment features' frequencies across time, where avg frequency denotes the average number of the feature in one ChatGPT answer and max frequency denotes the most common instance's frequency. Note that we demonstrate three dates' results in this figure, i.e. 2023-01-18, 2023-03-05 and 2023-04-09, which come from HC3 and ChatLog-Daily dataset respectively.

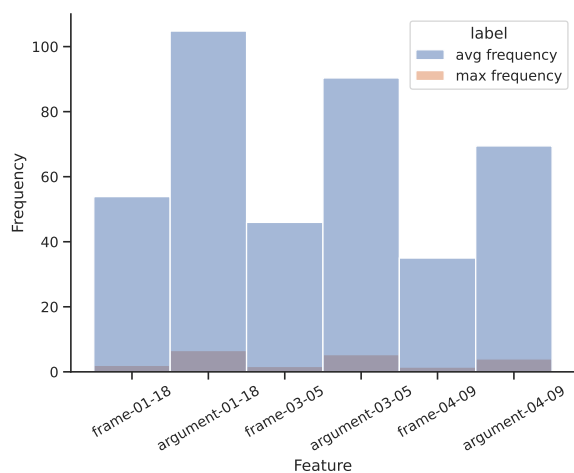


Figure 11: The extracted frame and argument features' frequencies across time, where avg frequency denotes the average number of the feature in one ChatGPT answer and max frequency denotes the most common instance's frequency. Note that we demonstrate three dates' results in this figure, i.e. 2023-01-18, 2023-03-05 and 2023-04-09, which come from HC3 and ChatLog-Daily dataset respectively.

B.3 Linguistic Feature Analysis

Further Trend Analysis We show the linguistic features' trends from Figure 12 to Figure 17. There are 5 feature categories including Advanced Semantic (AdSem), Discourse (Disco), Syntactic (Synta), Lexico Semantic (LxSem) and Syntactic (ShaF).

One interesting trend observed in the Advanced Semantic feature category from Figure 12 is the increase in the mean value of Clarity (BClar) features over time. This suggests that ChatGPT has become more effective in making its expression more clear in responses over time. Similarly, the number of topic (BTopc20_S) features decreases across time, which suggests that ChatGPT has become more accurate in generating fluent responses without loading its writing with fancy topics.

In the Discourse category from Figure 13, we observed a simultaneous drop in the number of entities (Enti_C) and Local Coherence from Entity Grid (LoCoh_S) scores, indicating that ChatGPT is utilizing more concrete and clear words rather than specific entities in its responses over time, which makes itself easier understood.

In the Syntactic feature category from Figure 14, we observed a decrease in the length of Flattened parsed Trees (FTree_C) and Local Coherence from count of Noun Tags (NoTag_C) scores, which indicates that ChatGPT has become more effective in structuring sentences so that the distance between dependent words in the flattened parsed trees is minimized. This trend suggests that ChatGPT is generating more coherent and organized sentences over time.

In terms of the Lexico Semantic (LxSem) feature category from Figure 15 and 16, we observed a decrease in the number of Noun Variation (NOV_S) and the Age of Acquisition of words by Kuperman (AAKuW_C), indicating that ChatGPT is utilizing more frequently occurring words in its responses over time. This trend is consistent with an increasing ability to generate focused and accurate responses.

Finally, in the ShaF feature category from Figure 17, we observed that the total count of tokens (Tok_S) and readability scores decreased over time. This suggests that ChatGPT has become more concise in generating responses with less words, resulting in a more precise ability to convey meaning.

Taken as a whole, these findings suggest that ChatGPT has been developed and improved over

time with an increasing ability to generate coherent, fluent, and concise responses. Analysis of linguistic features provides insights into the specific aspects of understanding ChatGPT's evolving patterns across time.

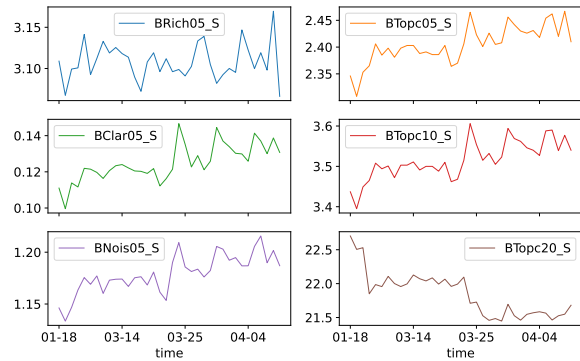


Figure 12: Trend of ChatGPT's Advanced Semantic (AdSem) linguistic features including Richness (BRich), Clarity (BClar), Noise (BNois) and number of topics (BTopc) scores on ChatLog-Daily dataset every day from March 5 to April 9, 2023. Note that the data at start point of January 18 comes from HC3 dataset.

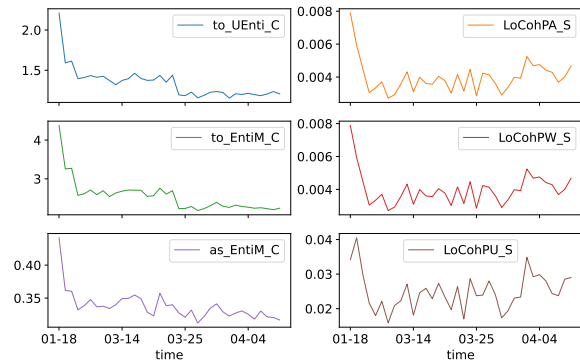


Figure 13: Trend of ChatGPT's Discourse (Disco) linguistic features including number of entities (Enti_C) and Local Coherence from Entity Grid (LoCoh_S) scores on ChatLog-Daily dataset every day from March 5 to April 9, 2023. Note that the data at start point of January 18 comes from open-access HC3 dataset.

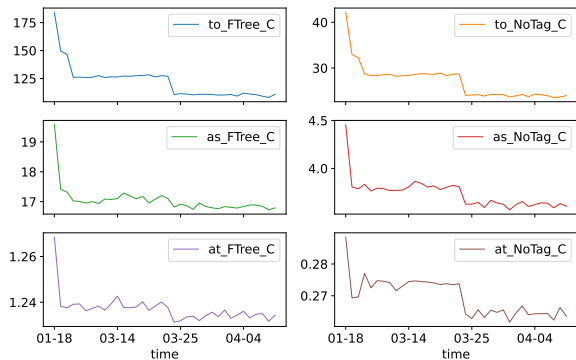


Figure 14: Trend of ChatGPT’s Syntactic (Synta) linguistic features including length of Flattened parsed Trees (FTree_C) and Local Coherence from count of Noun Tags (NoTag_C) scores on ChatLog-Daily dataset every day from March 5 to April 9, 2023. Note that the data at start point of January 18 comes from open-access HC3 dataset.

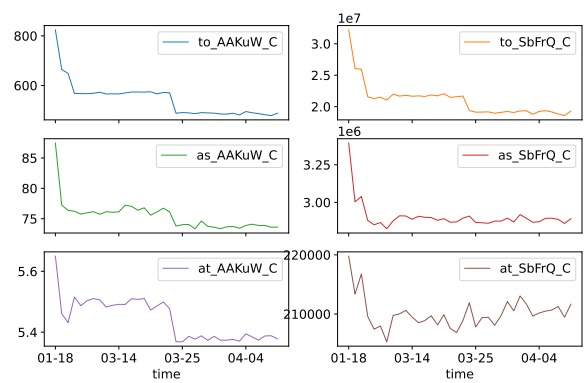


Figure 16: Trend of ChatGPT’s Lexico Semantic (LxSem) linguistic features including the number of Age of Acquisition of words by Kuperman (AAKuW_C) and the SubtlexUS FREQcount value (SbFrQ_C) scores on ChatLog-Daily dataset every day from March 5 to April 9, 2023. Note that the data at start point of January 18 comes from HC3 dataset.

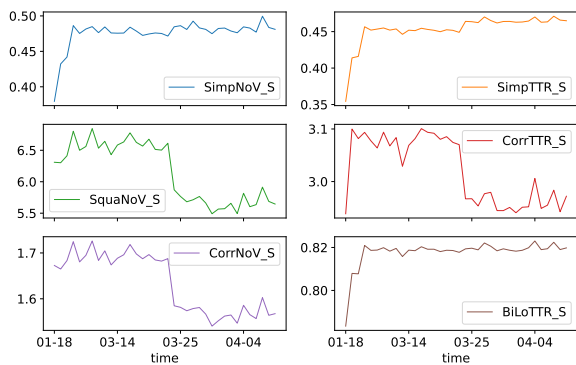


Figure 15: Trend of ChatGPT’s Lexico Semantic (LxSem) linguistic features including the number of Noun Variation (NOV_S) and the number of TTR (TTR_S) scores on ChatLog-Daily dataset every day from March 5 to April 9, 2023. Note that the data at start point of January 18 comes from HC3 dataset.

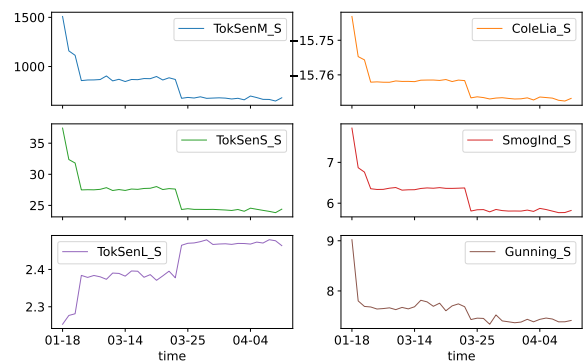


Figure 17: Trend of ChatGPT’s Syntactic (ShaF) linguistic features including total count of tokens (Tok_S) and readability scores on ChatLog-Daily dataset every day from March 5 to April 9, 2023. Note that the data at start point of January 18 comes from HC3 dataset.

1157 **Further Correlation Analysis** To comprehen-
1158 sively verify the correlation between extracted fea-
1159 tures and performance scores, we calculate the
1160 Pearson correlation coefficient (Cohen et al., 2009)
1161 between the nine rouge metrics rouge- $\{1,2,1\}$ -
1162 $\{precision, recall, f1\}$ and three semantic features
1163 WRich $\{05,10,15\}$ _S. The normalized results of
1164 ChatLog-Daily are shown in Figure 18. We can
1165 see that the WRich_S semantic richness feature
1166 scores positively correlate with the rouge- $\{1,2,1\}$ -
1167 r performance scores, which can be concluded from
1168 the correlation scores in the upper left part of the
1169 matrix.

1170 **Further Variation Analysis** In Table 25. we cal-
1171 culate the variation coefficient on ChatLog-Daily
1172 with data from March 5, 2023, to June 10, 2023,
1173 and show features with minimum variation coef-
1174 ficient. These stable features may not change sig-
1175 nificantly over time, which can be used for down-
1176 stream tasks. However, if we change the time span
1177 of the ChatLog-Daily, will these selected features
1178 stay minimum variation on new data? To answer
1179 this question, we calculate the variation coefficient
1180 on ChatLog-Daily with data from March 5, 2023, to
1181 March 28, 2023, and show features with minimum
1182 variation coefficient in Table 24. In fact, the 10
1183 most stable features on new time spans in Table 24
1184 are identical to the 10 features of longer time span
1185 in Table 25. Only the rank of the ra_ONToT_C
1186 and at_VeTag_C feature exchanges with each other
1187 while this doesn't affect the set of the 10 most sta-
1188 ble features. This phenomenon shows that even
1189 after changing the time span, our discovered fea-
1190 tures still have the minimum variation coefficient.

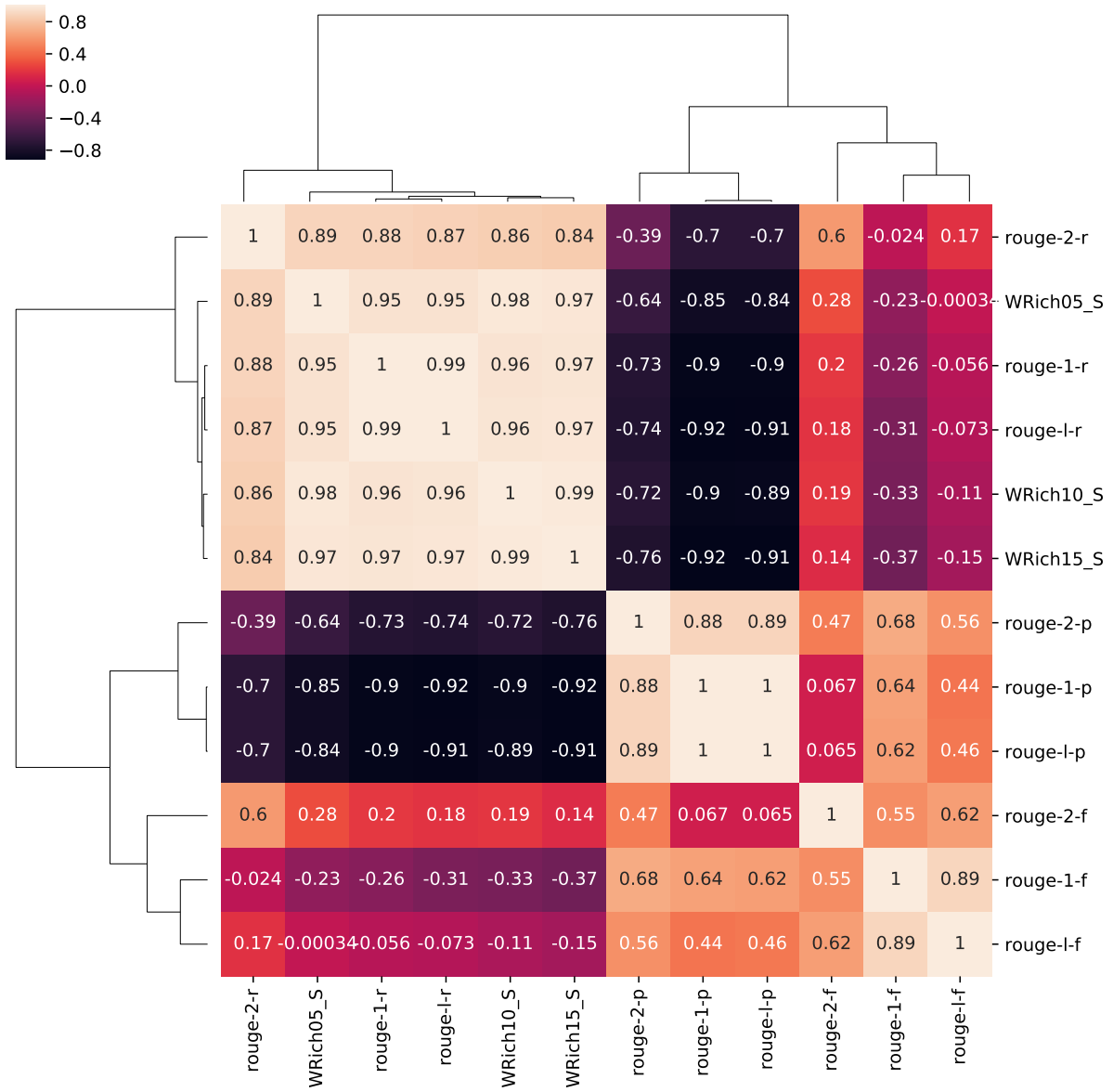


Figure 18: **Pearson correlation coefficient** between the nine rouge metrics rouge- $\{1,2,1\}$ - $\{precision, recall, f1\}$ and three semantic features WRich $\{05,10,15\}$ _S, on the ChatLog-Daily dataset after normalization. WRich_S denotes Semantic Richness coverage on several topics extracted from Wikipedia.

Branch	Feature	Feature Description	$ \mu $	$ \sigma $	$\frac{ \sigma }{ \mu }$
ShaTr	ColeLia_S	Coleman Liau Readability Score	15.76	1.7e-4	1.1e-5
Disco	ra_NNTo_C	Ratio of nn transitions to total	0.97	2.1e-4	2.2e-4
AdSem	BClar20_S	Semantic Clarity from 200 topics	0.80	7.9e-4	9.9e-4
AdSem	BClar15_S	Semantic Clarity from 150 topics	0.77	9.8e-4	1.3e-3
LxSem	BiLoTTR_S	Type Token Ratio Features	0.82	1.8e-3	2.2e-3
Synta	at_FTree_C	Average length of flattened Trees per token	1.24	2.9e-3	2.4e-3
Synta	at_ContW_C	Average count of Content words per token	0.64	1.9e-3	2.9e-3
LxSem	at_SbL1C_C	Average SubtlexUS Lg10CD value per token	3.36	1.3e-3	3.8e-3
Disco	ra_ONToT_C	Ratio of on transitions to total	2.0e-3	9.0e-6	4.5e-3
Synta	at_VeTag_C	average count of Verb POS tags per token	0.18	8.3e-4	4.6e-3

Table 24: Top 10 stable linguistic features from 2023-01-18 to 2023-03-28. $|\mu|$, $|\sigma|$, $\frac{|\sigma|}{|\mu|}$ is respectively the average, variation and variation coefficient of feature scores.

Branch	Feature	Feature Description	$ \mu $	$ \sigma $	$\frac{ \sigma }{ \mu }$
ShaTr	ColeLia_S	Coleman Liau Readability Score	15.76	1.6e-4	1.0e-5
Disco	ra_NNTo_C	Ratio of nn transitions to total	0.97	2.5e-4	2.5e-4
AdSem	BClar20_S	Semantic Clarity from 200 topics	0.80	8.3e-4	1.0e-3
AdSem	BClar15_S	Semantic Clarity from 150 topics	0.77	9.8e-4	1.3e-3
LxSem	BiLoTTR_S	Type Token Ratio Features	0.82	1.9e-3	2.3e-3
Synta	at_FTree_C	Average length of flattened Trees per token	1.24	3.0e-3	2.4e-3
Synta	at_ContW_C	Average count of Content words per token	0.63	1.9e-3	3.0e-3
LxSem	at_SbL1C_C	Average SubtlexUS Lg10CD value per token	3.36	1.3e-3	3.8e-3
Synta	at_VeTag_C	average count of Verb POS tags per token	0.18	8.6e-4	4.8e-3
Disco	ra_ONToT_C	Ratio of on transitions to total	2.2e-3	1.1e-5	5.2e-3

Table 25: Top 10 most stable linguistic features from 2023-01-18 to 2023-04-09. $|\mu|$, $|\sigma|$, $\frac{|\sigma|}{|\mu|}$ is respectively the average, variation and variation coefficient of feature scores.

C Details of Detectors

In this section, we first introduce the details of our training and evaluation of the ChatGPT detection task. Then we will show the baselines’ settings, hyper-parameters, and the full detection results on the everyday corpus of ChatLog-Daily.

Module	Parameter	Value
Feature Extraction	n	1000
	k	98
	m	265
Application of LightGBM	boosting type	gdbt
	learning rate	0.05
	num_leaves	31
	feature fraction	0.9
	bagging fraction	0.8
	bagging freq	5
	verbose	0
	boost round	50
early stop rounds	10	

Table 26: Hyper-parameters for each module.

C.1 Implementation Details

In our study, we utilized a binary classification approach to investigate whether they can detect ChatGPT responses over time, to maintain robustness on potential changes in ChatGPT’s new versions. To do so, we designed a detector based on the RoBERTa (Liu et al., 2019) pre-trained language model, which has demonstrated outstanding performance in various NLP tasks.

Our detector baseline is trained on a sample of HC3 (Guo et al., 2023) dataset comprising human responses and ChatGPT responses, which are used to fine-tune the RoBERTa model. During fine-tuning, the RoBERTa model is trained to predict whether a given response is human or ChatGPT given question and answer (RoBERTa-qa⁶) or only answer (RoBERTa-single⁷). Specifically, we directly use the trained checkpoints from huggingface.

While the RoBERTa pre-trained language model has been shown to be particularly effective in various NLP applications, it should be noted that a major disadvantage of using a pre-trained model as

⁶<https://huggingface.co/spaces/Hello-SimpleAI/chatgpt-detector>

⁷<https://huggingface.co/spaces/Hello-SimpleAI/chatgpt-detector-single>

a detector is the lack of interpretability. That is, the model is essentially a black box, making it difficult to understand the specific features that it uses to classify responses. To solve this potential problem and enhance the robustness of RoBERTa, we use an ensemble model LightGBM (Ke et al., 2017) to train a logistic regression model on the output probability of RoBERTa-single and extra 10 stable features discovered during feature extraction.

The hyper-parameters of our experiment’s each module are summarized in Table 26. For the RoBERTa-based baselines, we run their inference process on each response of our ChatLog-Daily dataset to collect the predicted classification probability using a single NVIDIA GeForce RTX 3090 GPU. Then the probability is combined with linguistic features to train and test the LightGBM regression model using CPU.

C.2 Full Detection Results

In fact, the test dataset in Section 4.4 is sampled from ChatLog-Daily and is blended with human responses to make the label balanced. Here we will show all the classification results for each day’s full corpus without balanced human response labels.

With classification results in Figure 19 and Figure 20, we have following observations:

(1) The accuracy of RoBERTa-single, gltr and ppl classifiers decreases rapidly from 2023-01-18 corpus of HC3 dataset to newly collected corpus in March and April of ChatLog-Daily, which shows the difference between ordinary ChatGPT texts in January and up-to-date texts in March and April. Besides, the texts’ classification accuracy also has continuous marginally fluctuations during March and April, which demonstrates that there may exist timely updating of ChatGPT as mentioned by OpenAI.

(2) Notice that the RoBERTa-qa classifier encodes both question and answer to classify whether the answer is written by ChatGPT. As our everyday asking questions stay the same with the training set, it may cheat on the question information to get high accuracy. If we remove the question information (i.e. RoBERTa-single), its performance drops to 90%.

(3) Among classifiers, gltr and ppl depend on linguistic features, which have the larger performance drop than fine-tuned classifiers based on RoBERTa, which demonstrates the robustness of the pre-trained language model based classifiers.

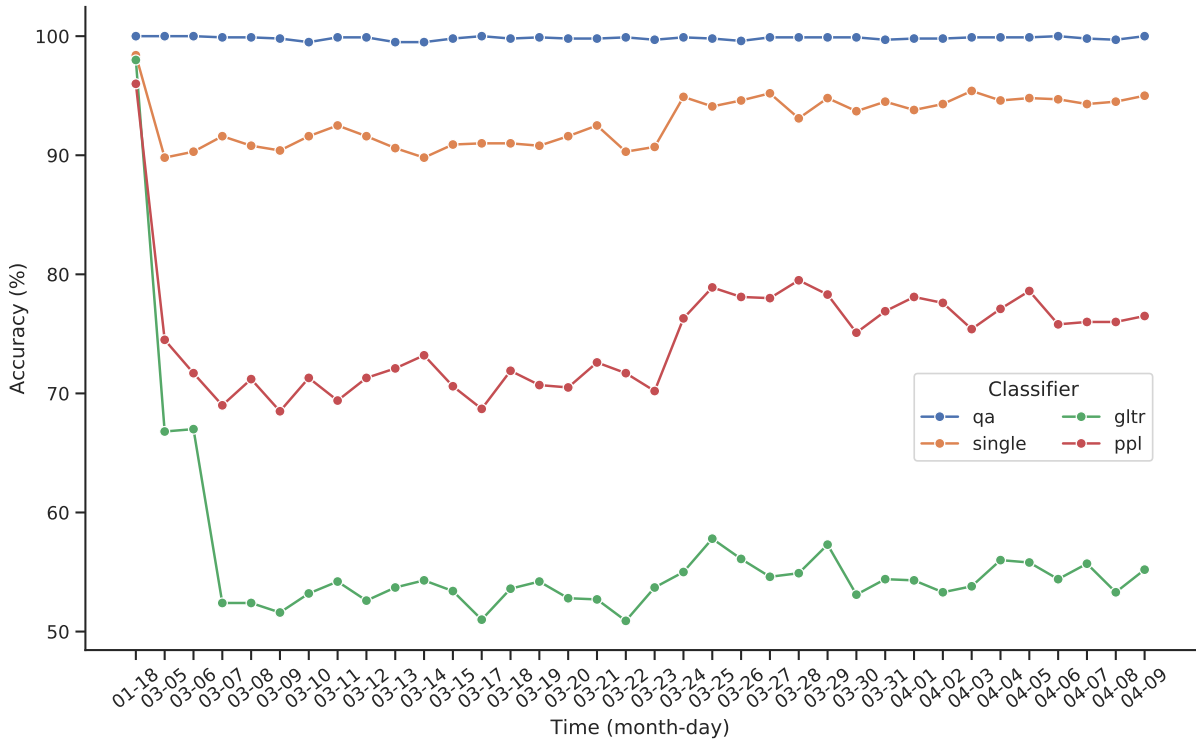


Figure 19: Accuracy of classifiers trained on HC3-01-18 corpus to predict the same 1000 questions' answers collected during full following days in March and April, where each day's data are unbalanced 1000 ChatGPT responses without human responses.

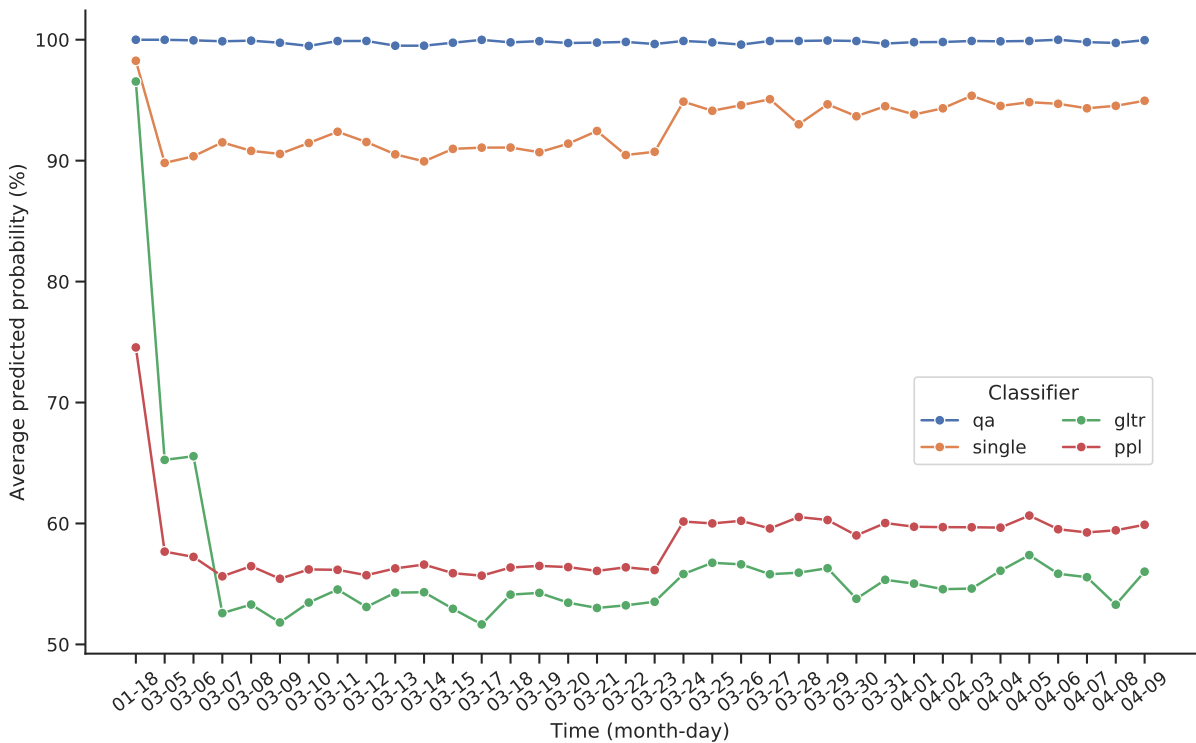


Figure 20: Output ChatGPT classification probability (confidence) of classifiers trained on HC3-01-18 corpus to predict the same 1000 questions' answers collected during the following days in March and April, where each day's data are unbalanced 1000 ChatGPT responses without human responses.