GPA2Bench: Abstract-Writing

[On the Detectability of ChatGPT Content: Benchmarking, Methodology, and Evaluation through the Lens of Academic Writing.](https://arxiv.org/abs/2306.05524) [Github: CheckGPT-v2](https://github.com/liuzey/CheckGPT-v2)

Three main focuses/tasks:

1. **Composing**: The author gives a title to ChatGPT and asks it to write the complete abstract from scratch.
2. **Completing**: Part of abstract is provided, GPT asked to complete with w words.
3. **Polishing**: Entire abstract provided to GPT and its asked to revise.

Three different fields:

1. CS for technical/engineering writing,
2. physics for science writing,
3. humanities and social science for liberal arts writing

Four different style prompts:

* Zero-shot
  + Composing: Here is the title of an academic research paper. Please write a paper abstract about it: {input}.
  + Completing: Here is the first half of the abstract of an academic research paper. Please complete its second half with approximate {X} words: {input}.
  + Polishing: Here is the abstract of an academic research paper. Please rewrite it for clarity: {input}
* Context Integrated: “for specific discipline”
  + Composing: Write an abstract of a research paper in {discipline} with first-person, clear, and academic language about "{title}"
* Persona: “you are an expert”
  + Composing: I want you to act as an academic paper writer. You are familiar with the topics in {discipline}. You will be responsible for writing a paper abstract. Your task is to generate an abstract for a paper with a given title. Please only include the written abstract in your answer. Here is the title of the paper: "{input}"
* Detailed requirements/instructions
  + Composing: Please act as an expert paper writer and write the abstract section of a paper from the perspective of a paper reviewer to make it fluent and elegant. Please only include the written abstract in your answer. Here are the specific requirements: 1.Enable readers to grasp the main points or essence of the paper quickly. 2. Allow readers to understand the important information, analysis, and arguments throughout the entire paper. 3. Help readers remember the key points of the paper. 4. Please clearly state the innovative aspects of your research in the abstract, emphasizing your contributions. 5. Use concise and clear language to describe your findings and results, making it easier for reviewers to understand the paper. Here is the title of the paper: "{input}

GPT Dataset:

* Gpt-3.5-turbo
* 50,000 samples for each prompt (12) , task (3), and discipline (3).
  + 600,000 samples for each discipline.
* If we only look at the composing task (no human input involved), 600,000

Human Dataset:

* Size: 150,000 samples, 50000 in each discipline
* Collection Method:
  + CS: top-tier conference papers and arXiv
  + Physics: arXiv
  + Humanities: Springer’s SSRN
* CS/Physics paper all published prior to 2019 (release of GPT-3) to ensure no AI was used to assist in writing the paper.

Composing Only:

|  |  |  |  |
| --- | --- | --- | --- |
|  | gpt | human | gpt breakdown |
| CS | 200,000 | 50,000 | 50,000 per prompt, 4 prompts |
| HSS | 200,000 | 50,000 | 50,000 per prompt, 4 prompts |
| PHX | 200,000 | 50,000 | 50,000 per prompt, 4 prompts |

Files to look at:

GPABench2/\_/gpt\_task1\_prompt1.json

GPABench2/\_/gpt\_task1\_prompt2.json

GPABench2/\_/gpt\_task1\_prompt3.json

GPABench2/\_/gpt\_task1\_prompt4.json

GPABench2/\_/ground.json

MGTBench: Essay, WP, Reuters

[MGTBench: Benchmarking tMachine-Generated Text Detection](https://arxiv.org/abs/2303.14822). [GitHub](https://github.com/xinleihe/MGTBench)

Intended to be a modular framework designed to benchmark detection methods.

Human Datasets

* from [Ghostbuster: Detecting Text Ghostwritten by Large Language Models](https://arxiv.org/abs/2305.15047):
* 3 subsets:
  + Student Essays:
    - 1000 essays available on IvyPanda, encompassing high school/university and several disciplines.
    - ChatGPT-turbo used to generate a prompt corresponding to essay.
  + Creative Writing
    - Collected from r/WritingPrompts, a subreddit for creative writing.
    - To avoid contamination, top 50 posters in Oct 2022 and 100 post by each user.
    - 1000 samples
  + News (Reuters):
    - 50/50 author identification dataset (Reuters Corpus Vol 1 RCV1)
      * Top 50 authors with text labeled with at least one subtopic of CCAT (corporate/industrial) so that genre/topic is not a distinguishing factor. So similar topics
      * A writing process with black text

        AI-generated content may be incorrect.5000 texts (100 per author): og paper split in 50

AI Datasets:

* ChatGPT-turbo, ChatGLM, Dolly, GPT4ALL, StableLM, Claude
* 1000 per essay, news, wp
* Something is wrong with this. Datasets are weird. Somehow very used?

WritingPrompts: Creative Writing (2018, pre-chatgpt)

[Hierarchical Neural Story Generation](https://arxiv.org/abs/1805.04833)

Investigates AI’s story generation capabilities: creative systems that can built coherent and fluent passages about a text.

Human Dataset:

* Collected from r/WritingPrompts, a subreddit for creative writing.
* Each prompt can have multiple story responses
* Diverse topic, length detail
  + At least 30 words
  + Avoiding general profanity and inappropriate content
  + Inspired by prompt (but not every single req)
* Scraped 3 years of prompts
  + Cleaned out bot responses, deleted responses, special announcements, mod comments, and stories <30 words
* For their experiment, limit length to 1000 words and limit vocab size for prompts and stories to words appearing more than 10 times each.

Uses the prompts on Language models (NOT LLMS) – can’t use probably

Seems to have set the precedent for creative writing datasets, cause MGTBench, Verma, and other papers have used this.

MAGE: Multi-domain, Multi-model

[MAGE: Machine-generated Text Detection in the Wild.](https://arxiv.org/abs/2305.13242) [GitHub](https://github.com/yafuly/MAGE). [Hugging Face](https://huggingface.co/datasets/yaful/MAGE)

Human Dataset:

* Opinion Statements: 804 opinion statements from r/ChangeMyView, 1000 reviews from Yelp
* News: 1000 articles from XSum, 777 (such a weird number) articles from TLDR\_news
* Question Answering: 1000 answers from ELI5 dataset
* Story Generation: 1000 prompted stories from r/WritingPrompts (WP), 1000 stores from ROCStories Corpora.
* Commonsense reasoning: 1000 sentence sets for reasoning from HellaSwag
* Knowledge Illustration: 1000 Wikipedia paragraphs from SQuAD contexts
* Scientific Writing: 1000 abstracts of scientific articles from SciXGen

AI Dataset:

* 27 LLMs
  + OpenAI GPT: text-davinci-002 (legacy), text-davinci-003, gpt-turbo-3.5
  + LLaMA: 6b/13B/30B/65B
  + GLM-130B
  + FLAN-T5: small, base, large, xl, xxl
  + OPT: 125M, 350M, 1.3B, 2.7B, 6.7B, 13B, 30B, iml1.3B, iml-30B
  + BigScience: T0-3B, T0-11B, BLOOM-7B1,
  + EleutherAI: GPT-J-6B, GPT-NeoX-20B
* Three types of prompts:
  + Continuation: continue generation based on provided 30 words
    - Prompt: “I spend my summer as a representative of the college…”
  + Topical: generate text based on a topic (ex. argument, story topic)
    - Prompt: “Generate a counter-argument to refute the following opinion: HandwritingCursive is an important skill…”
  + Specified: topical prompt but with info about text source (ex. BBC)
    - Directed more towards OpenAI.
    - Prompt: “generate a counter argument to refute the following Reddit post: HandwritingCursive….”

Construction example:

* Interesting methodology
  + Sample 1000 human sources from yelp
  + Prompt 27 models, receive 26,235 machine-gen texts
  + For data balance, sample data again from yelp and get a total of 37706 samples (for test and validation). They found no impact on model performance (they talk about it more extensively).
  + Default is continuation prompt
  + Clean dataset and filter out too long or too short text

A table with numbers and a few words

AI-generated content may be incorrect.

Topical/Specified only:

* Only text-davinci-002, text-davinci-003, and gpt-turbo-3.5 (all legacy/deprecated)
* No yelp subset, only cmv, eli5, tldr, xsum, wp, roct, hswag
* 3852

For some reason ran into issue with hugging face and error (maybe cause of datasets and/or python version, even though it seems to be an older issue), loaded datasets with dask?

RAID: Robust AI Detection, ACL-Long

[RAID: A Share Benchmark For Robust Evaluation of Machine-Generated Text Detectors](https://aclanthology.org/2024.acl-long.674/) [GitHub](https://github.com/liamdugan/raid?tab=readme-ov-file)

Few detectors are evaluated on shared benchmark datasets, and when they are they lack model variability, sample strategy, adversarial attacks. RAID addresses this.

A screenshot of a survey

AI-generated content may be incorrect.

Human Dataset:

* 2000 documents for each of 4 areas, 8 target domains
  + Factual Knowledge
    - News: BBC articles with associated titles, five categories (sport, tech, entertainment, politics, business
    - Wikipedia:
  + Generalization and Reasoning:
    - Abstracts: abstracts + titles from arXiv, post-2023 to avoid memorization
    - Recipes: recipes + dish names, requires common sense reasoning?
  + Creative and Conversational Skills
    - Reddit: reddit post + title, first-person / informal style
    - Poetry: poemhunter.com, title and genre, randomly spread over genre/topic
  + Knowledge of Specific Media
    - Books: plot-centric summaries of books with titles, first-person narrative style
    - Reviews: not 2000 but max amount, IMDb, review + movie name
* Public pre-2022 datasets
* A close-up of a list

  AI-generated content may be incorrect.Corresponding generation prompt like “Write a recipe…” , both continuation and chat-style

AI Dataset:

* One output per prompt across all 11 models

A screenshot of a computer

AI-generated content may be incorrect.Models:

M4:

[M4: Multi-Generator, Mult-Domain, and Multi-Lingual Black-Box Machine-Generated Text Detection](https://arxiv.org/abs/2305.14902). [Github](https://github.com/mbzuai-nlp/M4)

To achieve a more general machine-generated text detection, by encompassing multiple languages, various LLMs and diverse domains.

A table with numbers and text

AI-generated content may be incorrect.

Total Human: total size of dataset I think. Samples 3000

1147K total, 102k English , 42k others

Human Dataset:

* English Dataset:
  + Wikipedia (March 22) Ver, 3000 at random with min 1000 characters
  + WikiHow, 3000 at random with min 1000 characters
  + Reddit (ELI5), 3 categories (ExplainLikeImGive, AskScience, AskHistorians), 1000 QA pairs each with top user rating, at least 1000 characters, questions with titles that then end in question mark
  + arXiv: 3000 at random with min length 1000 chars
  + PeerRead: 586 papers with multiple reviews, 5798 total
* Also have Chinese, Russian, Arabic, Urdu, Indonesian, and Bulgarian datasets

AI Dataset:

* All Multilingual: GPT-4, ChatGPT, GPT3.5 (Text-davinci-003), Cohere, Dolly-v2, BLOOMz 176B
* Prompted to write:
  + article given title (Wikipedia), length 250 words at least
  + abstract given paper title (ArXiv)
  + peer review given title and abstract (PeerRead): 4 prompts for each paper (2 title only, 2 title and abstract), 2344 total
  + an answer given question (Reddit)
  + other prompts given other lang datasets
* Different style prompts (2-8)
  + A table with numbers and letters

    AI-generated content may be incorrect.Different numbers due to the fact that some prompts did not generate diversity across the different models/outputted wrong or weird formatting, were problematic for other reasons.

CHEAT: ChatGPT written abstract test

[CHEAT: A large-scale Dataset for Detecting ChatGPT WrittEn AbsTracts](https://arxiv.org/abs/2304.12008). [GitHub](https://github.com/botianzhe/CHEAT):

Focuses specifically on scientific writing in the form of research paper abtracts.

Human Dataset:

* 15395 Human Abstracts, average length 163.9 words
* IEEE Xplore
* All Computer Science papers: NLP, Cimputer Vision, Machine Learning

AI Dataset:

* 35,504 ChatGPT (gpt-3.5-turbo) written abstracts
* 3 tasks:
  + Generation (15395): “Generate a 200-word abstract of the paper in English based on the title and keywords; your answer only needs to include the generated paragraph.", followed by the title and keywords in the human-written data.
  + Polish (15395): "Polish the following paragraphs in English, your answer just needs to include the polished text.", followed by the human-written abstract.
  + Mix (15395, not new): decomposing polished abstracts and human-written text into semantics then constructing random mask and mixing the polished abstract with the human abstract.

HC3: Human ChatGPT Comparison Corpus (HC3)

[How Close is ChatGPT to Human Experts? Comparison Corpus, Evaluation, and Detection.](https://arxiv.org/abs/2301.07597) [GitHub](https://github.com/Hello-SimpleAI/chatgpt-comparison-detection)

Human Dataset:

* English: 24322 questions, 58546 human answers
* Different Domains: open-domain, computer science, finance, medicine, law, and psychology
* Sources: ELI5, WikiQA (for openQA), FiQA (finance), Medical Dialog Dataset, Crawled Wikipedia for cs qs
* Both English and Chinese

AI Dataset:

* GPT 3.5 series, only preview model available on website at time
* Thread refreshed
* 26903 ChatGPT answers

HC3 Plus:

[HC3 Plus: A Semantic-Invariant Human ChatGPT Comparison Corpus](https://arxiv.org/abs/2309.02731). GitHub

Updated HC3 dataset that adds Summarization, Translation and Paraphrasing tasks

Human Dataset:

* CNN/DailyMail: unique news articles. Article and Highlights
* Xsum: BBC articles, wide range of domain, article+one sentence summary
* Paraphrasing dataset: original HC3 questions

AI dataset:

* Generated using GPT-3.5-Turbo-0301
* Don’t say numbers. Also don’t differentiate between tasks, don’t know how I would filter out translation tasks, or would I include them.