Building an LLM identifiability challenge

Guy Shpigler and **Nir Mezamer** and **Yuval Mizrahi** and **Yuval Mor**
School of Computer Science, Tel-Aviv University
School of Electrical Engineering, Tel-Aviv University
{shpigler,nirmezamer,yuvalmizrahi,yuvalmor3}@mail.tau.ac.il

Abstract

Natural Language Processing (NLP) has made remarkable progress in recent years across various language-related tasks, and the use of large language models (LLMs) for text generation, summarization, and writing is becoming increasingly widespread. However, these advancements have also introduced a new challenge, as distinction between human-generated and machine-generated text becomes less clear. With LLMs becoming more sophisticated, they can produce text that closely resembles human-written content in terms of style, syntax, and coherence. This blurring of lines poses significant implications for various applications where textual authenticity is crucial, such as in journalism, social media, and online encyclopedia (Wikipedia) where they all serve as primary sources of knowledge for people worldwide. Our research project is designed to help answering this fundamental question:
Given a text, was it generated by a human or by a large language model (LLM)?
To address this question, we first design a pipeline to gather diverse human-generated texts in specific domains (Wikipedia, Reddit posts and comments, newspaper articles). By selecting a wide array of sources, we ensure that our dataset captures a broad spectrum of writing styles, topics, and formats. This diversity is crucial for creating a robust challenge. Based on these human-texts, we then design a pipeline to collect LLM-based texts mimicking the same diversity and domains as the natural ones. This involves using the large language model Gemini[1] to generate texts that replicate the styles and contexts of the human texts, and can be modified to use other LLMs. Our goal is to create a dataset where the machine-generated texts are as indistinguishable as possible from the human-generated ones at first glance. Subsequently, the dataset will form the foundational training infrastructure for machine learning models.

[1] https://ai.google.dev/

1 Introduction

To address the classification question, we aim to build a robust infrastructure for creating a Neural Network model that, given a text input, can determine whether it was generated by a human or by an LLM. The model's performance will largely depend on the quality of the dataset we build, which will be used for the model's learning process. The main challenge of our project was to create two text corpora: one of texts generated by humans and the other of texts generated by LLMs. Both corpora need to be representative, diverse, and balanced in size. We started by building the corpus of human-generated texts. To meet the requirements mentioned above, we chose to extract texts from various sources such as Wikipedia, Reddit (a social network) and newspapers. To create this text corpus, we wrote python scripts that use open APIs such as Wikipedia API, Reddit API, and Newspaper3k API. From these, we extracted Wikipedia pages, posts and comments from Reddit, and published newspaper articles from the BBC, which together form the corpus of human-generated texts. The topics in each domain were selected using a random mechanism to ensure as much diversity as possible. Simultaneously, to create the dataset of texts generated by LLMs, we used the Gemini API. We hand-crafted several prompts for each domain to achieve the most appropriate one to use when generating texts. Using the labeled data, we trained a classifier model to distinguish between human-generated and LLM-generated texts. We used a pre-trained BERT model and fine-tuned it on our dataset.

2 Engines

3 Document Body

3.1 Footnotes

Footnotes are inserted with the `\footnote` command.[2]

3.2 Tables and figures

Do not override the default caption sizes.

3.3 Hyperlinks

Users of older versions of LaTeX may encounter the following error during compilation:

```
\pdfendlink ended up in different
nesting level than \pdfstartlink.
```

This happens when pdfLaTeX is used and a citation splits across a page boundary. The best way to fix this is to upgrade LaTeX to 2018-12-01 or later.

3.4 Citations

Table **??** shows the syntax supported by the style files. We encourage you to use the natbib styles. You can use the command `\citet` (cite in text) to get "author (year)" citations, like this citation to a paper by **?**. You can use the command `\citep` (cite in parentheses) to get "(author, year)" citations (**?**). You can use the command `\citealp` (alternative cite without parentheses) to get "author, year" citations, which is useful for using citations within parentheses (e.g. **?**).

3.5 References

The LaTeX and BibTeX style files provided roughly follow the American Psychological Association format. If your own bib file is named `custom.bib`, then placing the following before any appendices in your LaTeX file will generate the references section for you:

```
\bibliographystyle{acl_natbib}
\bibliography{custom}
```

You can obtain the complete ACL Anthology as a BibTeX file from https://aclweb.org/anthology/anthology.bib.gz. To include both the Anthology and your own .bib file, use the following instead of the above.

```
\bibliographystyle{acl_natbib}
\bibliography{anthology,custom}
```

Please see Section 4 for information on preparing BibTeX files.

[2]This is a footnote.

3.6 Appendices

Use `\appendix` before any appendix section to switch the section numbering over to letters. See Appendix A for an example.

4 BibTeX Files

Unicode cannot be used in BibTeX entries, and some ways of typing special characters can disrupt BibTeX's alphabetization. The recommended way of typing special characters is shown in Table **??**.

Please ensure that BibTeX records contain DOIs or URLs when possible, and for all the ACL materials that you reference. Use the `doi` field for DOIs and the `url` field for URLs. If a BibTeX entry has a URL or DOI field, the paper title in the references section will appear as a hyperlink to the paper, using the hyperref LaTeX package.

Limitations

ACL 2023 requires all submissions to have a section titled "Limitations", for discussing the limitations of the paper as a complement to the discussion of strengths in the main text. This section should occur after the conclusion, but before the references. It will not count towards the page limit. The discussion of limitations is mandatory. Papers without a limitation section will be desk-rejected without review.

While we are open to different types of limitations, just mentioning that a set of results have been shown for English only probably does not reflect what we expect. Mentioning that the method works mostly for languages with limited morphology, like English, is a much better alternative. In addition, limitations such as low scalability to long text, the requirement of large GPU resources, or other things that inspire crucial further investigation are welcome.

Ethics Statement

Scientific work published at ACL 2023 must comply with the ACL Ethics Policy.[3] We encourage all authors to include an explicit ethics statement on the broader impact of the work, or other ethical considerations after the conclusion but before the references. The ethics statement will not count toward the page limit (8 pages for long, 4 pages for short papers).

[3]https://www.aclweb.org/portal/content/acl-code-ethics

Acknowledgements

This document has been adapted by Jordan Boyd-Graber, Naoaki Okazaki, Anna Rogers from the style files used for earlier ACL, EMNLP and NAACL proceedings, including those for EACL 2023 by Isabelle Augenstein and Andreas Vlachos, EMNLP 2022 by Yue Zhang, Ryan Cotterell and Lea Frermann, ACL 2020 by Steven Bethard, Ryan Cotterell and Rui Yan, ACL 2019 by Douwe Kiela and Ivan Vulić, NAACL 2019 by Stephanie Lukin and Alla Roskovskaya, ACL 2018 by Shay Cohen, Kevin Gimpel, and Wei Lu, NAACL 2018 by Margaret Mitchell and Stephanie Lukin, BibTEX suggestions for (NA)ACL 2017/2018 from Jason Eisner, ACL 2017 by Dan Gildea and Min-Yen Kan, NAACL 2017 by Margaret Mitchell, ACL 2012 by Maggie Li and Michael White, ACL 2010 by Jing-Shin Chang and Philipp Koehn, ACL 2008 by Johanna D. Moore, Simone Teufel, James Allan, and Sadaoki Furui, ACL 2005 by Hwee Tou Ng and Kemal Oflazer, ACL 2002 by Eugene Charniak and Dekang Lin, and earlier ACL and EACL formats written by several people, including John Chen, Henry S. Thompson and Donald Walker. Additional elements were taken from the formatting instructions of the *International Joint Conference on Artificial Intelligence* and the *Conference on Computer Vision and Pattern Recognition*.

A Example Appendix

This is a section in the appendix.