**CS310 NLP Group Project Guideline**

**Spring 2025**

(this draft will be constantly updated)

**Goal Description**

To effectively detect large language models (LLMs)-generated texts, especially to distinguish them from real human-written ones, is becoming a more and more important task. The task can be approached with two technical paths: 1) **supervised learning**-based detection; 2) likelihood metrics-based **zero-shot** detection. The former is similar to building a text classification model for tasks such as sentiment analysis etc., which can be done by fine-tuning a transformer encoder-based model (e.g., BERT) on an annotated dataset with binary labels (e.g., “0” for human-written and “1” for LLM-generated). The main advantage of this approach is that a supervised learning model can perform well provided with sufficient amount of data, and will be useful for a focused task-domain (e.g., news, fictions etc.). The limitation is also obvious – it is not a generic method, which means a detection model trained on one type of text data may fail on others, that is, relatively poor out-of-domain (OOD) performance. The latter approach, likelihood-based zero-shot detection, is a more generic solution – the detection algorithm/pipeline developed for one text-domain/languages/LLM can also work well on others, that is, better overall OOD performance.

The goal of this project is twofold: First, implement a series of supervised learning-based detection models, and test their performances under the OOD condition. Second, pick one of the zero-shot detection methods, and test it on the same setting. Compare the performances of the two methods, and discuss your findings.

**Datasets**

* Ghostbuster English data: A collection of LLM-generated texts together with human ground truths developed by Verma et al. [1]  
  <https://github.com/vivek3141/ghostbuster-data>
* Chinese data: A collection texts generated by Qwen-2 on three domains: News articles Wikipedia documents, and Web novels. Also shipped with human ground truths.   
  (the data will be uploaded to the course website)

**Zero-shot Detection Methods**

You can choose from one of the published works as follows:

* Fast-DetectGPT [2]: A method based on the probability curvatures texts. It is an improvement over DetectGPT[3],   
  Repo: <https://github.com/baoguangsheng/fast-detect-gpt>
* FourierGPT[4]: A method based on the spectral representations of text likelihood.   
  Repo: <https://github.com/CLCS-SUSTech/FourierGPT>
* GPT-who [5]: A method based on the psycholinguistic features.   
  Repo: <https://github.com/saranya-venkatraman/gpt-who>

It is okay if there are other methods you would like to use, but make sure to justify your choice.

**Notes on Supervised Method**

A binary classification model for distinguishing human vs. generated texts is sufficient. You do not need to train a model for multi-class detection, e.g., Claude vs. GPT-4 vs. GPT-3.5 etc.

You can start with training English-specific models. For optional experiment, you can test how multi-lingual models work on the task.

**General Expected Results**

Your results should roughly fall into this table, with a combination of datasets and methods. The expected results include accuracy, precision, recall, F-1, AUROC etc.

The detection tasks can be at a finer degree, for example, the performance on News, Wikipedia, Webnovel in Chinese separately. This is optional, and should be done only if you have time.

|  |  |  |
| --- | --- | --- |
|  | English | Chinese |
| Supervised | AA | BB |
| Zero-shot | CC | DD |

**References**

[1] V. Verma, E. Fleisig, N. Tomlin, and D. Klein, "Ghostbuster: Detecting text ghostwritten by large language models," *arXiv preprint arXiv:2305.15047,* 2023.

[2] G. Bao, Y. Zhao, Z. Teng, L. Yang, and Y. Zhang, "Fast-detectgpt: Efficient zero-shot detection of machine-generated text via conditional probability curvature," *arXiv preprint arXiv:2310.05130,* 2023.

[3] E. Mitchell, Y. Lee, A. Khazatsky, C. D. Manning, and C. Finn, "Detectgpt: Zero-shot machine-generated text detection using probability curvature," in *International Conference on Machine Learning*, 2023: PMLR, pp. 24950-24962.

[4] Y. Xu, Y. Wang, H. An, Z. Liu, and Y. Li, "Detecting Subtle Differences between Human and Model Languages Using Spectrum of Relative Likelihood," Miami, Florida, USA, November 2024: Association for Computational Linguistics, in Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pp. 10108-10121.

[5] S. Venkatraman, A. Uchendu, and D. Lee, "GPT-who: An Information Density-based Machine-Generated Text Detector," Mexico City, Mexico, June 2024: Association for Computational Linguistics, in Findings of the Association for Computational Linguistics: NAACL 2024, pp. 103-115.