CS310 NLP project

12310513 Lou Yibin 12310520 Rui Yuhan

I. Introduction

- LLMs are increasingly used to generate text in various domains.
- **Distinguishing** machine-generated text from human-written text has become a critical task.
- Two Major detection methods: **supervised learning and zero-shot detection.**
- Supervised learning performs well on specific domains but has **limited generalization**.
- Zero-shot detection offers **more generalizability across domains**.

This project aims to **implement and compare** these two approaches to evaluate their **cross-domain detection performance**. The detailed implement can be found at **https://github.com/Nahuyiur/CS310-NLP**.

II. Methodology

A. Supervised Learning Method

We adopt a supervised learning approach by fine-tuning pre-trained Transformer models.

1) Select **pre-trained Transformer models** for **fine-tuning**, e.g., BERT variants and RoBERTa
2) **Tokenize** labeled dataset texts using model-specific tokenizers, set max sequence length
3) Construct **data loaders** for training, validation, and testing splits
4) Initialize model with **classification head** for binary output
5) Train models end-to-end with AdamW optimizer, monitor training loss
6) Validate after each epoch, save best-performing model by validation F1 score
7) **Evaluate** final model on test set using metrics: accuracy, precision, recall, F1, AUROC

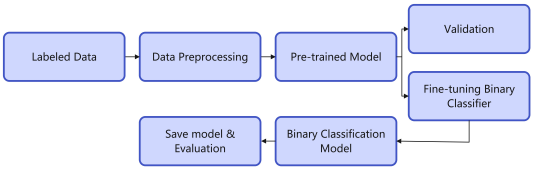


Figure 1: Supervised Learning Method Framework

At the core of our method is **a binary classification model** designed to distinguish between **human-written and machine-generated texts**. The process comprises data preparation, preprocessing, loading of pre-trained models, fine-tuning, validation, and model selection. During training, **validation metrics** are continuously monitored to identify and retain the best-performing model for final evaluation.

B. Supervised Learning Method(Spectrum)

We propose a supervised learning method based on spectral representations of language model outputs.

1) Compute the **negative log-likelihood (NLL)** scores from a language model, e.g., gpt2-xl and Mistral-7B-v0.1
2) Apply **z-score normalization** to the sequence to standardize it
3) Perform a **DFT** on the normalized sequence to transform it from the **time domain** to the **frequency domain**
4) compute magnitude:
$\|X(\omega_k)\| = \sqrt{\mathrm{Re}(X(\omega_k))^2 + \mathrm{Im}(X(\omega_k))^2}$
5) Use the spectrum magnitude sequence as input features for classification.
6) Implement **Augmented spectrum classifier** by averaging Fourier spectra of circularized likelihood sequences to enhance weak periodic patterns for classification

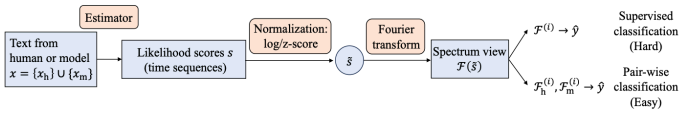


Figure 2: Supervised Learning Method(Spectrum) Framework

An augmented spectrum-based classifier is trained using frequency features obtained from multiple circularized variants of the original likelihood sequence. **This involves circularizing the sequence, applying the Fourier transform to each variant, and averaging the resulting spectra to construct a robust representation**. Inspired by circular convolution in signal processing, this method serves as a form of data augmentation to enhance weak periodic signals and improve classification performance.

C. Zero-shot Detection Methods

A **heuristic spectrum-based classifier** is proposed to distinguish human-written from machine-generated texts by leveraging **frequency-domain features**. The classification decision is based on the summed spectral power difference over a selected low-frequency range, and an empirical threshold ε, according to the following criterion:

$$\left| \sum_{k=1}^{\delta} \|X_{\text{Human}}(\omega_k)\| - \sum_{k=1}^{\delta} \|X_{\text{Model}}(\omega_k)\| \right| > \varepsilon$$

III. Data Preprocessing

English Data:

- Source: Ghostbuster dataset (https://github.com/vivek3141/ghostbuster-data)
- Domains: **essay, reuter, wp**
- Each domain contains 6 types of LLM-generated texts and one human-written (HM) type
- **CSV with fields** —text, label **(1 for LLM, 0 for HM)**, domain
- Created mixed-domain and domain-specific CSV files
- Split into train, validation, test sets with ratio **8:1:1**
- Data cleaning applied to handle corrupted or missing entries

Chinese Data:

- Domains: **news, webnovel, wiki**
- Human-written data and Qwen2-72b generated data
- **Same preprocessing pipeline** as English data
- Input format well structured; tested both **concatenated input-output** and **input only**
- Found better results **without concatenation** of input

Data Preprocessing for Zero-shot Method

- **Same datasets** as supervised method used for consistency
- For each domain, data split into **two separate text files**: one for human-written (HM), one for LLM-generated (LLM) texts
- Each line in txts corresponds to **one data entry** from the original dataset
- Related metadata such as topics and prompts stored in separate files, **aligned by line number** across HM and LLM files
- Missing or anomalous values identified and cleaned
- For English data, each domain contains 6 LLM samples and 1 HM sample per topic, duplicate HM for 6 times to **balance sample counts**

Processed Data Overview

CSV Files:

- **English Data:**
  - eng_essay.csv, eng_reuter.csv, eng_wp.csv
  - eng_hm_essay.csv, eng_hm_reuter.csv, eng_hm_wp.csv
  - eng_llm_essay.csv, eng_llm_reuter.csv, eng_llm_wp.csv
  - eng_mix (mixed domain)
- **Chinese Data:**
  - zh_domain
  - zh_mix (mixed domain)

TXT Files:

- **English Data:**
  - eng_essay_hm.txt, eng_essay_llm.txt
  - eng_reuter_hm.txt, eng_reuter_llm.txt
  - eng_wp_hm.txt, eng_wp_llm.txt

- **Chinese Data:(similar)**
  - zh_news_hm.txt, zh_news_llm.txt
  - zh_webnovel_hm.txt, zh_webnovel_llm.txt
  - zh_wiki_hm.txt, zh_wiki_llm.txt

IV. Experiments

A. Experiment Design

1) Supervised Learning Method: We conduct a supervised learning experiment using several pre-trained Transformer models for both English and Chinese. **The training setup includes 10 epochs, a batch size of 16, the AdamW optimizer, and log intervals of 50 steps.** The learning rates are set to 1×10^{-7} **for English and** 1×10^{-5} **for Chinese.**

The English models include `bert-base-uncased`, `bert-base-multilingual-cased`, `roberta-base`, **and** `xlm-roberta-base`, while the Chinese models include `bert-base-chinese`, `xlm-roberta-base`, **and** `roberta-base`. To ensure fair comparison, we control for training/validation splits, batch size, number of epochs, and optimizer settings within each language group.

Model performance is evaluated using **five metrics: accuracy, precision, recall, F1, and AUROC**. The final step involves comparing these metrics across all models and datasets.

We conduct both **in-domain and out-of-domain (OOD)** evaluations. For in-domain evaluation, data from the three domains (essay, reuter, wp) are combined and shuffled, then **split into training, validation, and test sets in an 8:1:1 ratio.** We compute five evaluation metrics—accuracy, precision, recall, F1, and AUROC—on the test set, reporting both overall and per-domain results.

For OOD evaluation, models are trained and validated on a single domain and tested on mixed-domain data. **Metrics are computed overall and per domain**, excluding results from the training domain to avoid data leakage. This process is repeated for each domain as the training source.

2) FourierGPT supervised learning experiment: The models used in our experiments include **GPT2 variants for both English (gpt2-xl, Mistral-7B-v0.1)** and **Chinese (gpt2-chinese-cluecorpussmall, Wenzhong2.0-GPT2-3.5B-chinese).**

To ensure fair comparison, we maintain consistent spectrum generation by applying circular shifts on NLL data with a fixed interpolation length, followed by standard scaling and fixed feature selection ($k = 120$).

Classification is performed using an SVM with an RBF kernel and fixed hyperparameters. Evaluation metrics include accuracy, precision, recall, F1 score, and AUROC.

3) Zero-shot detection method experiment: The zero-shot detection experiments utilize **GPT2 variants for English (`gpt2-xl`, Mistral-7B-v0.1) and Chinese (`gpt2-chinese-cluecorpussmall`, Wenzhong2.0-GPT2-3.5B-chinese)**. To ensure fair comparison, the same spectrum generation procedure is applied across all models, with a consistent k range and threshold $\varepsilon = 0$. **Evaluation metrics include accuracy and the AUROC curve.**

In the detection method, a sample is classified as model-generated if

$$\sum_{i=1}^{k} P_{\text{model}}(\omega_i) - \sum_{i=1}^{k} P_{\text{human}}(\omega_i) > \varepsilon,$$

and as human-written otherwise. The optimal $k \in [1, 50]$ is selected by evaluating accuracy, precision, recall, and F1 for each k, choosing the one with the highest accuracy.

With the selected k, multiple power thresholds are tested to compute true positive and false positive rates, which form the ROC curve and enable calculation of the AUROC metric.

V. Result

A. Supervised Learning Method

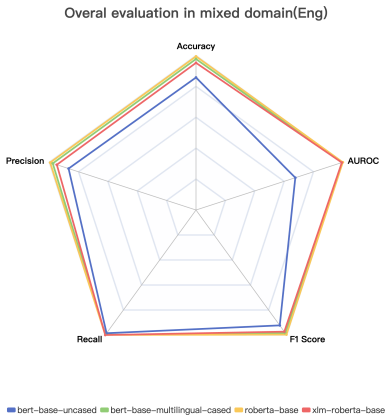


Figure 3: Overal evaluation in mixed domain(Eng)

1) English Dataset: In the English mixed domain, **roberta-base and xlm-roberta-base** outperform bert-base-uncased and bert-base-multilingual-cased in accuracy, F1 score, and AUROC.

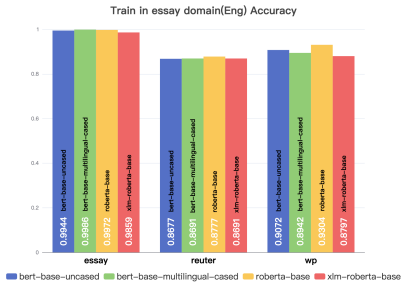


Figure 4: Accuracy in essay domain(Eng)

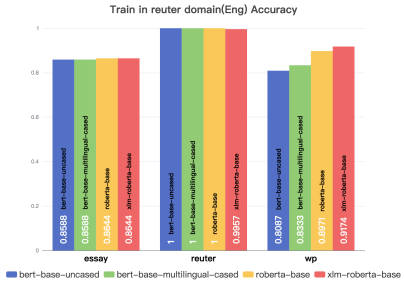


Figure 5: Accuracy in reuter domain(Eng)

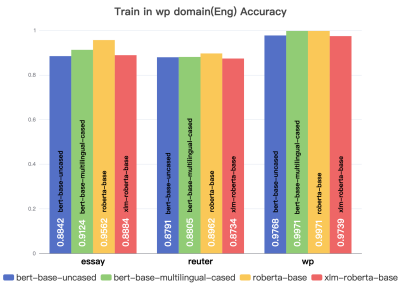
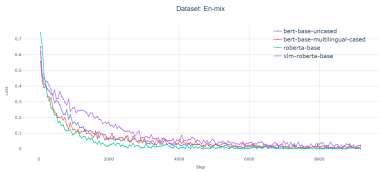
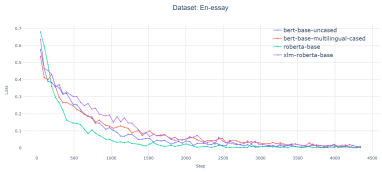


Figure 6: Accuracy in wp domain(Eng)

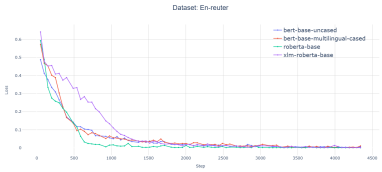
The models perform well within their **own training domain**, achieving accuracy generally **above 0.88**. However, accuracy significantly decreases when **tested on other domains**, often dropping **below 0.80**. Among the four models, **roberta-base** demonstrates relatively better generalization across different domains.



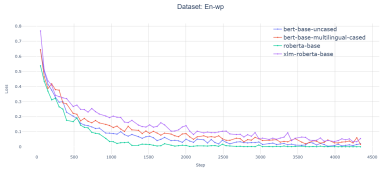
((a)) Eng-mix



((b)) Eng-essay



((c)) Eng-reuter



((d)) Eng-wp

Figure 7: Supervised Learning Method: Loss Curve (Eng)

Among the four models, **roberta-base** exhibits the fastest loss decline across all datasets. Eventually,

all models **converge and stabilize** with the loss values near **0.005 to 0.001**, indicating effective training and good convergence behavior.

In the supervised learning method, we compare model performance across English datasets using multiple evaluation metrics.

Table I: bert-base-uncased (Eng) overall/essay/reuter/wp

FT domain	mixed	essay	reuter	wp
Accuracy	0.8577 / 0.8460 / 0.8578 / 0.8696	0.9234 / 0.9944 / 0.8677 / 0.9072	0.8896 / 0.8588 / 1.0000 / 0.8087	0.9129 / 0.8842 / 0.8791 / 0.9768
Precision	0.8679 / 0.8640 / 0.8590 / 0.8813	0.9291 / 0.9935 / 0.8889 / 0.9091	0.9421 / 0.8657 / 1.0000 / 0.9797	0.9107 / 0.8829 / 0.8811 / 0.9742
Recall	0.9857 / 0.9755 / 0.9983 / 0.9834	0.9868 / 1.0000 / 0.9669 / 0.9934	0.9297 / 0.9902 / 1.0000 / 0.7980	0.9973 / 0.9984 / 0.9934 / 1.0000
F1 Score	0.9231 / 0.9163 / 0.9234 / 0.9296	0.9571 / 0.9967 / 0.9262 / 0.9494	0.9358 / 0.9238 / 1.0000 / 0.8796	0.9520 / 0.9371 / 0.9339 / 0.9869
AUROC	0.6764 / 0.6967 / 0.5726 / 0.8464	0.9526 / 0.9999 / 0.9178 / 0.9474	0.8558 / 0.8016 / 1.0000 / 0.9022	0.9435 / 0.8843 / 0.8828 / 1.0000

Table II: bert-base-multilingual (Eng) overall/essay/reuter/wp

FT domain	mixed	essay	reuter	wp
Accuracy	0.9772 / 0.9774 / 0.9844 / 0.9696	0.9210 / 0.9986 / 0.8691 / 0.8942	0.8977 / 0.8588 / 1.0000 / 0.8333	0.9296 / 0.9124 / 0.8805 / 0.9971
Precision	0.9758 / 0.9760 / 0.9837 / 0.9679	0.9432 / 0.9984 / 0.9051 / 0.9275	0.9412 / 0.8636 / 1.0000 / 0.9785	0.9553 / 0.9270 / 0.9437 / 0.9967
Recall	0.9984 / 0.9984 / 0.9983 / 0.9983	0.9670 / 1.0000 / 0.9470 / 0.9536	0.9407 / 0.9935 / 1.0000 / 0.8278	0.9637 / 0.9755 / 0.9156 / 1.0000
F1 Score	0.9870 / 0.9871 / 0.9910 / 0.9829	0.9550 / 0.9992 / 0.9256 / 0.9404	0.9409 / 0.9240 / 1.0000 / 0.8969	0.9595 / 0.9506 / 0.9294 / 0.9983
AUROC	0.9938 / 0.9942 / 0.9948 / 0.9948	0.8951 / 1.0000 / 0.8050 / 0.8857	0.8161 / 0.6701 / 1.0000 / 0.9163	0.9573 / 0.8957 / 0.9276 / 0.9999

Table III: roberta-base (Eng) overall/essay/reuter/wp

FT domain	mixed	essay	reuter	wp
Accuracy	0.9943 / 0.9887 / 0.9972 / 0.9971	0.9353 / 0.9972 / 0.8777 / 0.9304	0.9205 / 0.8644 / 1.0000 / 0.8971	0.9495 / 0.9562 / 0.8962 / 0.9971
Precision	0.9934 / 0.9871 / 0.9967 / 0.9967	0.9340 / 0.9967 / 0.8787 / 0.9330	0.9160 / 0.8644 / 1.0000 / 0.8948	0.9482 / 0.9532 / 0.8992 / 0.9967
Recall	1.0000 / 1.0000 / 1.0000 / 1.0000	0.9956 / 1.0000 / 0.9950 / 0.9917	1.0000 / 1.0000 / 1.0000 / 1.0000	0.9962 / 0.9984 / 0.9901 / 1.0000
F1 Score	0.9967 / 0.9935 / 0.9983 / 0.9983	0.9638 / 0.9984 / 0.9332 / 0.9615	0.9561 / 0.9273 / 1.0000 / 0.9445	0.9716 / 0.9753 / 0.9425 / 0.9983
AUROC	1.0000 / 1.0000 / 1.0000 / 1.0000	0.9797 / 1.0000 / 0.9662 / 0.9696	0.9024 / 0.9039 / 1.0000 / 0.9316	0.9826 / 0.9799 / 0.9392 / 1.0000

Table IV: xlm-roberta-base (Eng) overall/essay/reuter/wp

FT domain	mixed	essay	reuter	wp
Accuracy	0.8777 / 0.8686 / 0.8649 / 0.9000	0.9119 / 0.9859 / 0.8691 / 0.8797	0.9257 / 0.8644 / 0.9957 / 0.9174	0.9115 / 0.8884 / 0.8734 / 0.9739
Precision	0.8766 / 0.8691 / 0.8641 / 0.8975	0.9123 / 0.9839 / 0.8678 / 0.8917	0.9249 / 0.8644 / 0.9951 / 0.9253	0.9093 / 0.8913 / 0.8716 / 0.9711
Recall	0.9995 / 0.9984 / 1.0000 / 1.0000	0.9940 / 1.0000 / 1.0000 / 0.9818	0.9951 / 1.0000 / 1.0000 / 0.9851	0.9973 / 0.9918 / 1.0000 / 1.0000
F1 Score	0.9340 / 0.9293 / 0.9271 / 0.9460	0.9514 / 0.9919 / 0.9292 / 0.9346	0.9587 / 0.9273 / 0.9975 / 0.9543	0.9513 / 0.9389 / 0.9314 / 0.9853
AUROC	0.9618 / 0.9502 / 0.9653 / 0.9800	0.8359 / 0.9999 / 0.6537 / 0.7428	0.8115 / 0.5012 / 1.0000 / 0.9369	0.9291 / 0.8743 / 0.8691 / 0.9978

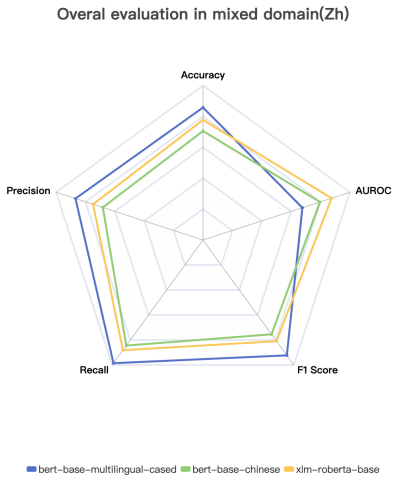


Figure 8: Overal evaluation in mixed domain(Zh)

2) Chinese Dataset: In the Chinese mixed domain, **bert-base-multilingual-cased** achieves the highest recall and precision, while **xlm-roberta-base** leads in AUROC and F1 score, outperforming bert-base-chinese across most metrics.

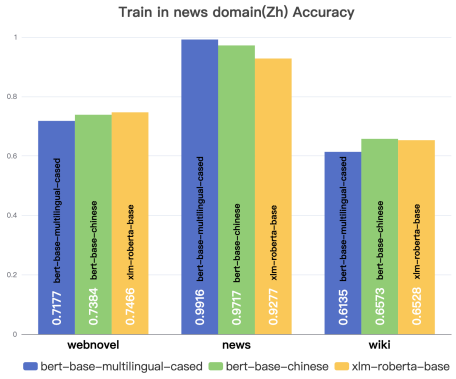


Figure 9: Accuracy in news domain(Zh)

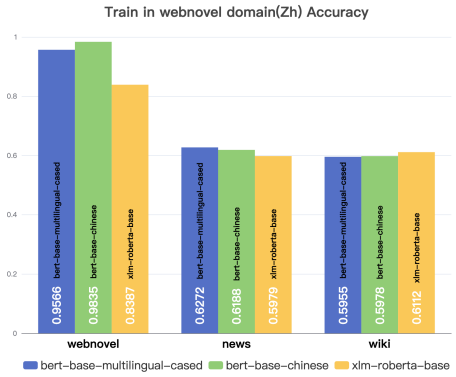


Figure 10: Accuracy in webnovel domain(Zh)

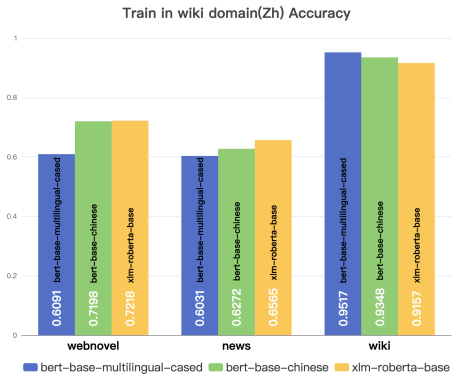
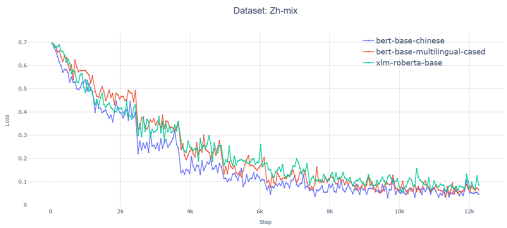
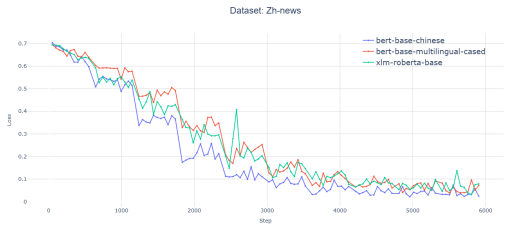


Figure 11: Accuracy in wiki domain(Zh)

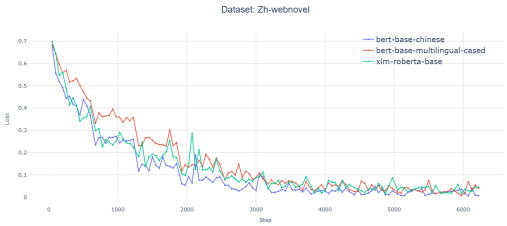
The models demonstrate strong performance within their **own training domain**, achieving accuracy mostly **above 0.83**. However, accuracy notably drops when **tested on other domains**, often falling between **0.63 and 0.72**. Among the three models, **bert-base-chinese** exhibits relatively better generalization across different domains.



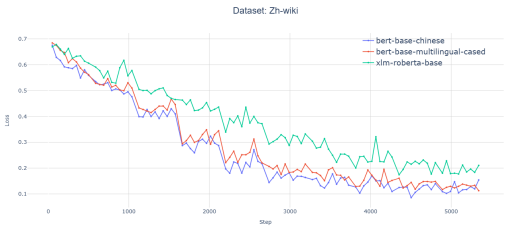
((a)) Zh-mix



((b)) Zh-news



((c)) Zh-webnovel



((d)) Zh-wiki

Figure 12: Supervised Learning Method: Loss Curve (Zh)

Among the three models, **bert-base-chinese** shows the fastest loss decline. All models stabilize with loss values around **0.05**, demonstrating stable and effective training.

In the supervised learning method, we compare model performance across Chinese datasets using multiple evaluation metrics.

Table V: bert-base-multilingual-cased (Zh) overall/webnovel/news/wiki

FT domain	mixed	webnovel	news	wiki
Accuracy	0.7055 / 0.7880 / 0.6440 / 0.6820	0.7304 / 0.9566 / 0.6272 / 0.5955	0.7155 / 0.6091 / 0.6031 / 0.9517	0.6820 / 0.6091 / 0.6031 / 0.9517
Precision	0.6818 / 0.7238 / 0.6289 / 0.6998	0.6722 / 0.9190 / 0.5997 / 0.5952	0.6924 / 0.5681 / 0.6009 / 0.9602	0.6998 / 0.5681 / 0.6009 / 0.9602
Recall	0.8447 / 0.9056 / 0.8226 / 0.8125	0.9701 / 0.9979 / 0.9201 / 0.9943	0.8441 / 0.7876 / 0.7778 / 0.9583	0.8125 / 0.7876 / 0.7778 / 0.9583
F1 Score	0.7546 / 0.8046 / 0.7128 / 0.7520	0.7941 / 0.9568 / 0.7262 / 0.7447	0.7608 / 0.6601 / 0.6780 / 0.9592	0.7520 / 0.6601 / 0.6780 / 0.9592
AUROC	0.7961 / 0.8879 / 0.6976 / 0.7649	0.7528 / 0.9988 / 0.6629 / 0.5219	0.7881 / 0.6637 / 0.6390 / 0.9936	0.7649 / 0.6637 / 0.6390 / 0.9936

Table VI: bert-base-chinese (Zh) overall/webnovel/news/wiki

FT domain	mixed	webnovel	news	wiki
Accuracy	0.7219 / 0.7942 / 0.6754 / 0.6933	0.7376 / 0.9835 / 0.6188 / 0.5978	0.7920 / 0.7384 / 0.9717 / 0.6573	0.7564 / 0.7198 / 0.6272 / 0.9348
Precision	0.6893 / 0.7236 / 0.6452 / 0.7060	0.6741 / 0.9668 / 0.5869 / 0.5977	0.7511 / 0.6814 / 0.9500 / 0.6573	0.7251 / 0.7010 / 0.6008 / 0.9181
Recall	0.8759 / 0.9270 / 0.8791 / 0.8277	0.9841 / 1.0000 / 0.9805 / 0.9848	0.9151 / 0.8584 / 1.0000 / 0.8826	0.8786 / 0.7296 / 0.9123 / 0.9773
F1 Score	0.7715 / 0.8128 / 0.7442 / 0.7620	0.8014 / 0.9831 / 0.7343 / 0.7439	0.8250 / 0.7767 / 0.9744 / 0.6869	0.7945 / 0.7150 / 0.7457 / 0.9468
AUROC	0.8271 / 0.9125 / 0.7516 / 0.7920	0.7549 / 0.9999 / 0.7053 / 0.5985	0.8396 / 0.7989 / 0.9950 / 0.7081	0.8328 / 0.7817 / 0.6977 / 0.9927

Table VII: xlm-roberta-base (Zh) overall/webnovel/news/wiki

FT domain	mixed	webnovel	news	wiki
Accuracy	0.7774 / 0.8676 / 0.7309 / 0.7292	0.6849 / 0.8387 / 0.5979 / 0.6112	0.7784 / 0.7466 / 0.9277 / 0.6528	0.7610 / 0.7218 / 0.6565 / 0.9157
Precision	0.7479 / 0.8062 / 0.6975 / 0.7487	0.6324 / 0.7566 / 0.5734 / 0.6056	0.7656 / 0.6751 / 0.8868 / 0.7386	0.7158 / 0.6518 / 0.6190 / 0.9450
Recall	0.8819 / 0.9549 / 0.8811 / 0.8182	0.9841 / 0.9807 / 0.9825 / 0.9886	0.8454 / 0.9142 / 0.9922 / 0.9376	0.9190 / 0.9077 / 0.9376 / 0.9110
F1 Score	0.8094 / 0.8743 / 0.7786 / 0.7819	0.7700 / 0.8542 / 0.7241 / 0.7511	0.8035 / 0.7767 / 0.9365 / 0.6869	0.8048 / 0.7587 / 0.7457 / 0.9277
AUROC	0.8751 / 0.9531 / 0.8099 / 0.8259	0.7549 / 0.9704 / 0.7112 / 0.5307	0.8685 / 0.8667 / 0.9910 / 0.7081	0.8304 / 0.8204 / 0.7144 / 0.9831

B. Supervised Learning Method(Spectrum)

Figure 13: Overal evaluation in mixed domain(Eng)

1) English Dataset: **In the English mixed domain**, the metrics of the two models are almost identical.

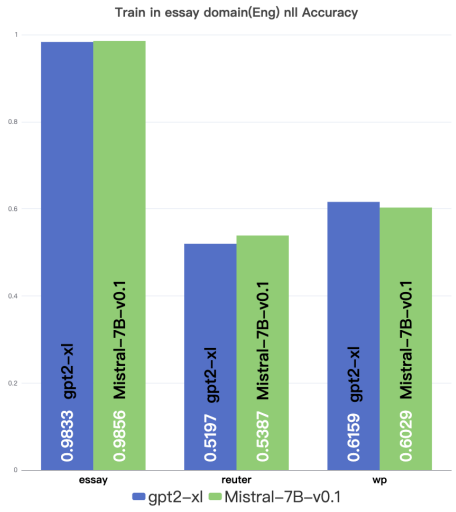


Figure 14: Accuracy in essay domain(Eng)

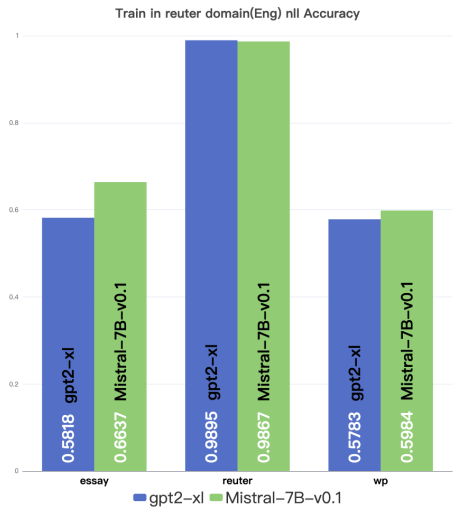


Figure 15: Accuracy in Reuter domain(Eng)

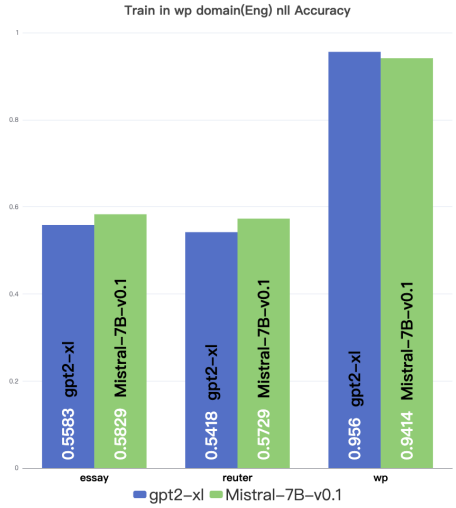


Figure 16: Accuracy in wp domain(Eng)

For **out of domain test**, the models perform well within their **own training domain**, achieving accuracy **above 0.95**. However, accuracy significantly decreases when **tested on other domains**, often dropping **below 0.6**. Among the two models, **Mistral-7B-v0.1** demonstrates relatively better generalization across different domains.

Figure 17: Overal evaluation in mixed domain(Zh)

2) Chinese Dataset: **In the Chinese mixed domain, Wenzhong2.0-GPT2-3.5B-chinese** performs better across most metrics except recall.

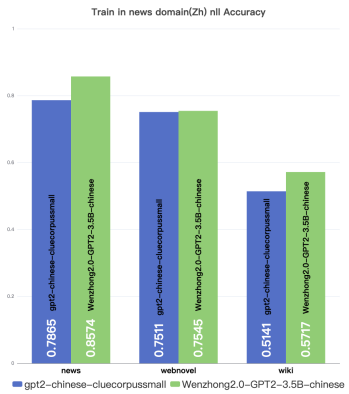


Figure 18: Accuracy in news domain(Zh)

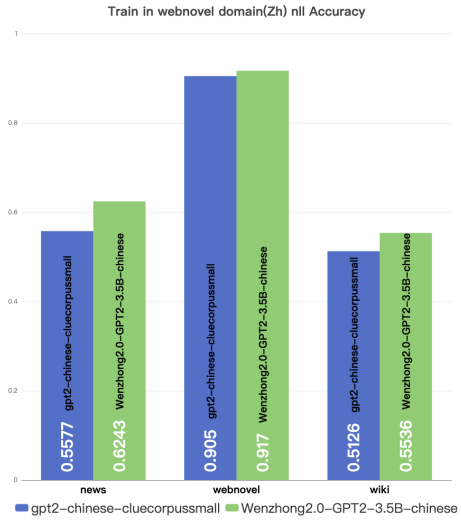


Figure 19: Accuracy in Webnovel domain(Zh)

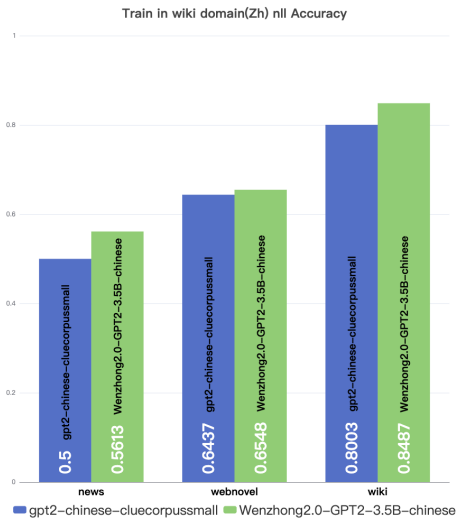


Figure 20: Accuracy in wiki domain(Zh)

For **out of domain test**, the models perform well within their **own training domain**, achieving accuracy **above 0.8**. However, accuracy significantly decreases when **tested on other domains**, often dropping **below 0.6**. Among the two models, **Wenzhong2.0-GPT2-3.5B-chinese** exhibits relatively better generalization across different domains.

C. Zero-shot Detection

We analyze the zero-shot detection performance using **aggregated power spectra across different datasets.**

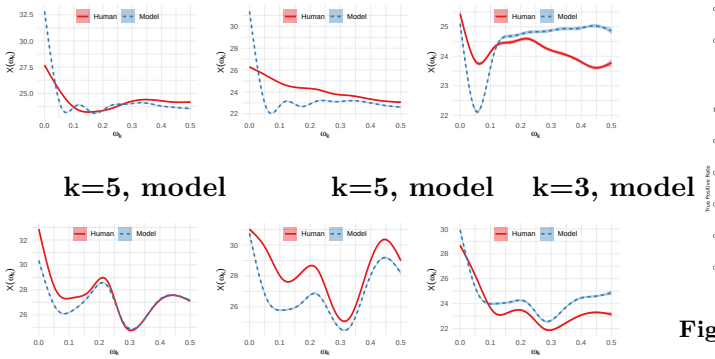


Figure 21: Aggregated power spectra for zero-shot detection across different datasets.

For English texts, the model can distinguish between human-written and LLM-generated texts using only **a small number of low-frequency** features. Notably, for English texts, LLM-generated texts exhibit **higher** spectral power than human texts, whereas for Chinese texts, LLM-generated texts show **lower** spectral power.

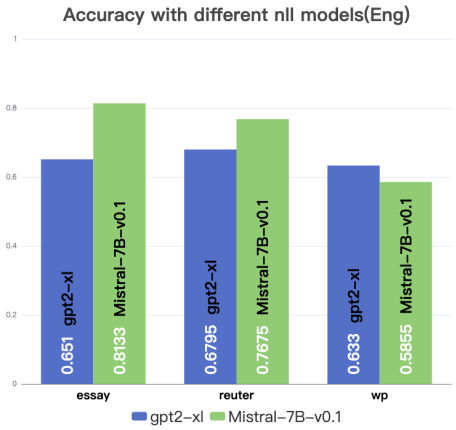


Figure 22: Accuracy with different nll models(Eng)

1) English Dataset: **Mistral-7B-v0.1** outperforms GPT2-XL in both the Essay and Reuter domains, achieving accuracy scores **above 0.75**. However, both models **struggle** on the wp domain, with accuracies of **0.63** and **0.59** respectively. Additionally, **Mistral-7B-v0.1** generates **better negative log-likelihood (NLL)** scores.

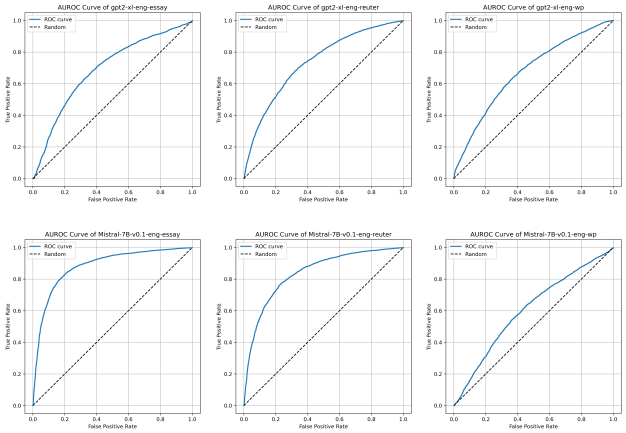


Figure 23: AUROC curves of GPT2-XL and Mistral-7B-v0.1 across English essay, reuter, and wp domains.

The ROC curves for **Mistral-7B-v0.1** on the essay and reuter datasets are closer to the top-left corner, indicating a stronger ability to distinguish between positive and negative samples. Conversely, the ROC curves on the **wp** dataset approach the diagonal line, suggesting performance close to random guessing

and highlighting the increased difficulty in classification within this domain.

In the Zero-shot detection method, we compare model performance across English datasets using multiple evaluation metrics.

Table VIII: gpt2-xl-Eng

FT domain	essay	reuter	wp
Accuracy	0.6510	0.6795	0.6330
AUROC	0.6899	0.7348	0.6732

Table IX: Mistral-7B-v0.1-Eng

FT domain	essay	reuter	wp
Accuracy	0.8133	0.7675	0.5855
AUROC	0.8798	0.8375	0.6044

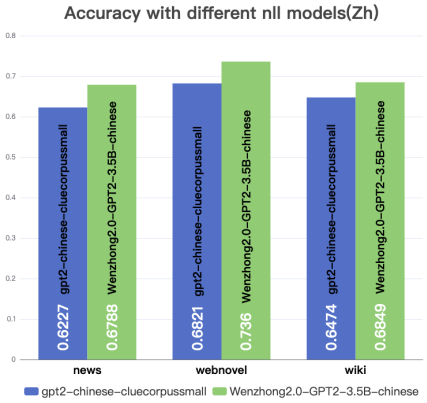


Figure 24: Accuracy with different nll models(Zh)

2) Chinese Dataset: **Wenzhong2.0-GPT2-3.5B** consistently outperforms **gpt2-chinese-cluecorpussmall** across all evaluated domains. Both models achieve their highest accuracy and AUROC in the **webnovel** domain (0.73 and 0.82, respectively), while detecting differences in **news** texts proves more challenging, with corresponding scores of 0.64 and 0.68.

The ROC curves for **gpt2-chinese-cluecorpussmall** exhibit a pronounced early rise in the news and webnovel datasets but remain flat and close to the diagonal in the wiki dataset, indicating weaker discrimination ability. In contrast,

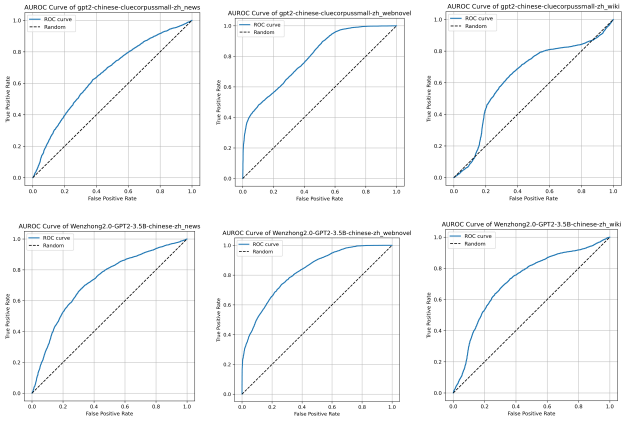


Figure 25: Accuracy across Chinese domains for GPT2-chinese-cluecorpussmall (top row) and Wenzhong2.0-GPT2-3.5B (bottom row) in news, webnovel, and wiki datasets.

Wenzhong2.0-GPT2-3.5B demonstrates steeper and more top-left concentrated ROC curves on the news and webnovel datasets, reflecting stronger classification performance.

In the Zero-shot detection method, we compare model performance across Chinese datasets using multiple evaluation metrics.

Table X: gpt2-chinese-cluecorpussmall-Zh

FT domain	news	webnovel	wiki
Accuracy	0.6227	0.6821	0.6474
AUROC	0.6561	0.7933	0.6381

Table XI: Wenzhong2.0-GPT2-3.5B-chinese-Zh

FT domain	news	webnovel	wiki
Accuracy	0.6788	0.7360	0.6849
AUROC	0.7232	0.8240	0.7216

VI. Result analysis

A. In-Domain Accuracy on English Datasets

To evaluate different detection approaches under in-domain settings, we compare the performance of **Text Supervised Fine-Tuning, Spectral Supervised Fine-Tuning, and Spectral Zero-Shot methods on English datasets.**

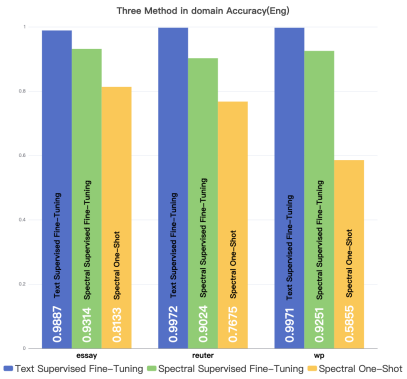


Figure 26: In-domain accuracy comparison across English datasets.

Text Supervised Fine-Tuning achieves the highest accuracy across all domains, demonstrating the effectiveness of direct supervision. In contrast, the **Zero-Shot method performs the worst,** with notably high sensitivity to domain variation, especially in the wp domain where its accuracy drops to 0.58.

B. Out-of-Domain Accuracy on English Datasets

We next examine how each **method generalizes** to unseen domains in the English setting.

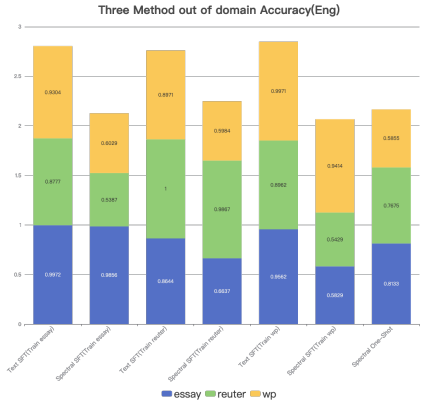


Figure 27: Cross-domain (out-of-domain) accuracy comparison on English datasets.

Train on	Essay	Reuter	Wp
Text SFT	0.9972	0.8644	0.9562
Spectrum SFT	0.9856	0.6637	0.5829
One-shot	0.8133		

Table XII: Cross-domain accuracy when tested on Essay (English datasets).

Train on	Essay	Reuter	Wp
Text SFT	0.8777	1.0000	0.8962
Spectrum SFT	0.5387	0.9867	0.5429
One-shot	0.7675		

Table XIII: Cross-domain accuracy when tested on Reuter (English datasets).

Train on	Essay	Reuter	Wp
Text SFT	0.9304	0.8971	0.9971
Spectrum SFT	0.6029	0.5984	0.9414
One-shot	0.5855		

Table XIV: Cross-domain accuracy when tested on Wp (English datasets).

Text Supervised Fine-Tuning exhibits strong cross-domain generalization, maintaining accuracy above 0.85. However, **Spectral Supervised Fine-Tuning performs poorly in this setting**, with accuracy dropping below 0.6. Interestingly, t**he Spectral Zero-Shot method shows promising results on the essay (0.81) and reuter (0.77) domains**, highlighting its potential in specific scenarios.

C. In-Domain Accuracy on Chinese Datasets

We perform a similar evaluation for **Chinese datasets**, measuring in-domain performance of the three methods.

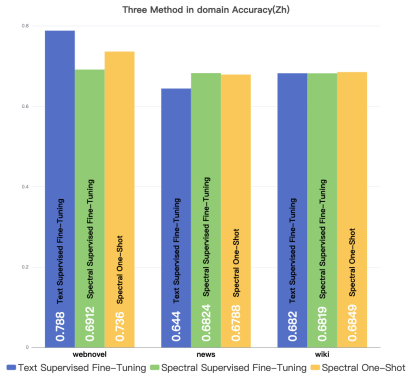


Figure 28: In-domain accuracy comparison across Chinese datasets.

Text Supervised Fine-Tuning leads in the webnovel domain with an accuracy of 0.788. However, the **two FourierGPT-based methods also perform competitively**, particularly in the news and

wiki domains, where their accuracies remain around 0.68.

D. Out-of-Domain Accuracy on Chinese Datasets

We further assess the **cross-domain generalization ability** of each method on Chinese datasets.

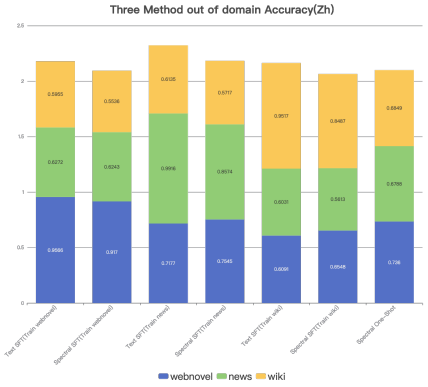


Figure 29: Cross-domain (out-of-domain) accuracy comparison on Chinese datasets.

Train on	Webnovel	News	Wiki
Text SFT	0.9566	0.7177	0.6091
Spectrum SFT	0.9170	0.7545	0.6548
One-shot	0.6849		

Table XV: Cross-domain accuracy when tested on Webnovel (Chinese datasets).

Train on	Webnovel	News	Wiki
Text SFT	0.6272	0.9916	0.6031
Spectrum SFT	0.6243	0.8574	0.5613
One-shot	0.6788		

Table XVI: Cross-domain accuracy when tested on News (Chinese datasets).

Train on	Webnovel	News	Wiki
Text SFT	0.5955	0.6153	0.9517
Spectrum SFT	0.5536	0.5717	0.8487
One-shot	0.7360		

Table XVII: Cross-domain accuracy when tested on Wiki (Chinese datasets).

Both supervised learning methods struggle in the out-of-domain scenario, with accuracy below 0.65. In contrast, **the Zero-Shot detection method demonstrates relatively strong generalization**, maintaining performance above 0.65 across all domains.

E. Reasoning

1) **Which method performs best in in-domain scenarios, and why?**

Answer:

Text Supervised **Fine-Tuning consistently achieves the highest accuracy in in-domain tasks** because it enables the model to learn detailed domain-specific features and patterns, whereas **Zero-Shot** methods lack access to such labeled examples during training.

2) **Which method performs best in out-of-domain (OOD) scenarios, and why?**

Answer:

Zero-Shot (One-Shot) methods show better generalization in out-of-domain settings, especially in the **Chinese datasets**, because they do not rely heavily on domain-specific training data. Instead, they utilize more generalizable spectral features or pretrained knowledge, making them more adaptable to unseen domains despite lower in-domain accuracy.

3) **Why do the One-Shot results differ from those reported in essay?**

Answer:

Our English dataset is generated by **gpt3.5-turbo**, which uses **Reinforcement Learning from Human Feedback (RLHF)** to improve the quality and safety of its generated content by leveraging human feedback. It also undergoes specialized fine-tuning to **enhance reasoning, maintain memory and logical consistency** in multi-turn conversations. So it's harder to discriminate it from human than GPT 3.5 which is used in the paper.

4) **Why do supervised methods achieve lower accuracy in Chinese compared to English?**

Answer:

The lower accuracy in Chinese supervised learning is mainly due to the increased **complexity and diversity of the Chinese language**, such as rich characters and more ambiguous context.

VII. Conclusion

A. Summary

- This project compares **supervised learning** and **zero-shot detection** methods for cross-domain text generation detection.
- **Supervised learning** achieves the best performance **in-domain**, especially with **text supervised fine-tuning**.
- **Zero-shot detection** demonstrates stronger **out-of-domain generalization**, particularly on Chinese datasets.
- The **spectrum-based methods** provide effective feature representations to enhance detection robustness.

B. Future

- Explore **more advanced pretrained models**, including larger multilingual models, for detection tasks.
- Investigate integrating **multi-modal information** (e.g., speech, images) to improve detection accuracy.
- Develop **more efficient spectral analysis and feature augmentation techniques** to boost zero-shot detection performance.
- Expand datasets in **size and diversity**, focusing on real-world and varied text sources.

References

[1] Xu, Y., Wang, Y., An, H., Liu, Z., & Li, Y. (2024). *Detecting subtle differences between human and model languages using spectrum of relative likelihood*. arXiv preprint arXiv:2406.19874.