CS310 NLP project

12310513 娄毅彬　12310520 芮煜涵

Southern University of Science and Technology

Department of Computer Science and Engineering

July 24, 2025

Content

Project Background

Project Background

- 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**.

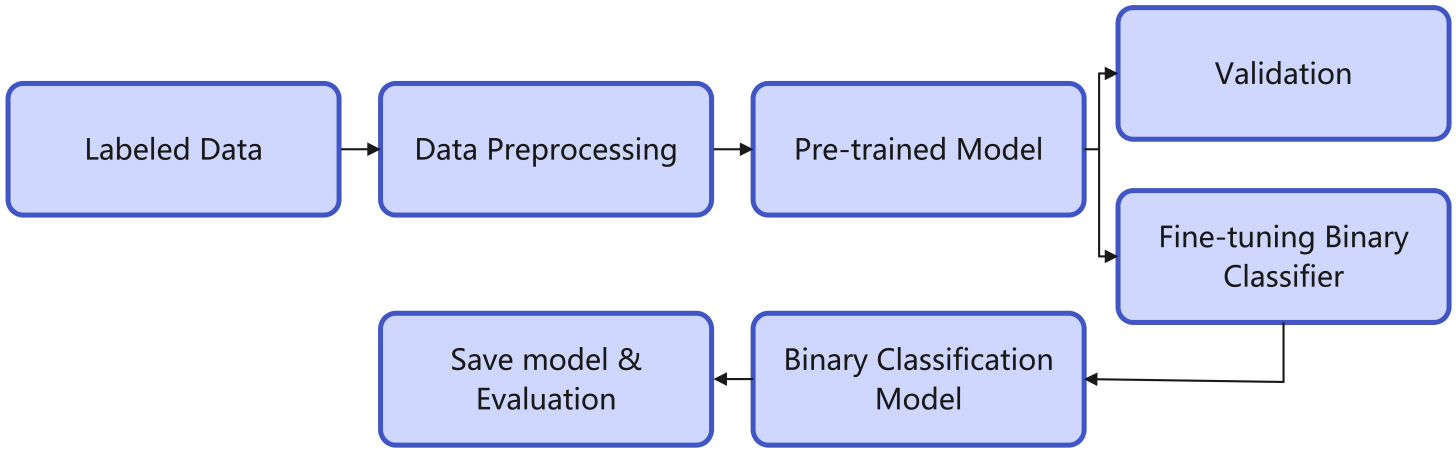
Supervised Learning Method

Supervised Learning Method: Fine-Tuning Steps

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

Using **multiple models** allows performance comparison and robustness analysis

Supervised Learning Method Framework



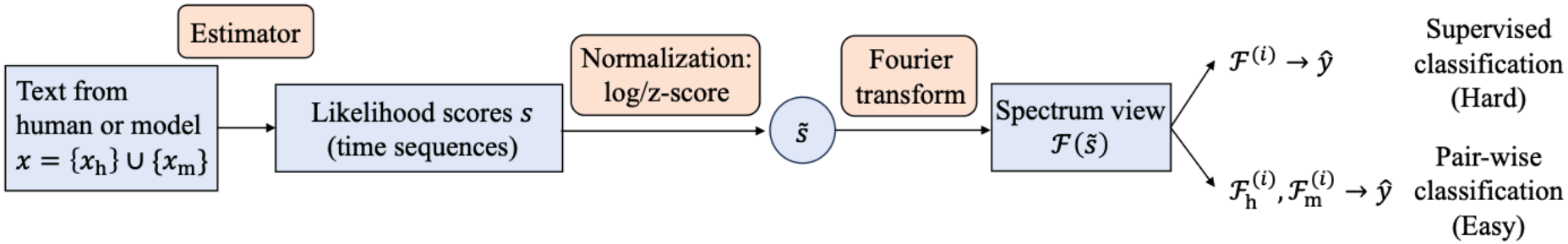
- **Binary classification model** at the core to distinguish human-written and machine-generated texts
- Process includes data preparation, preprocessing, pre-trained model loading, fine-tuning, validation, and model selection
- Validation metrics monitored during training to select the best model for final testing

Supervised Learning Method(Spectrum)

Supervised Learning Method(Spectrum) step

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:$\|X(\omega_k)\| = \sqrt{\text{Re}(X(\omega_k))^2 + \text{Im}(X(\omega_k))^2}$
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

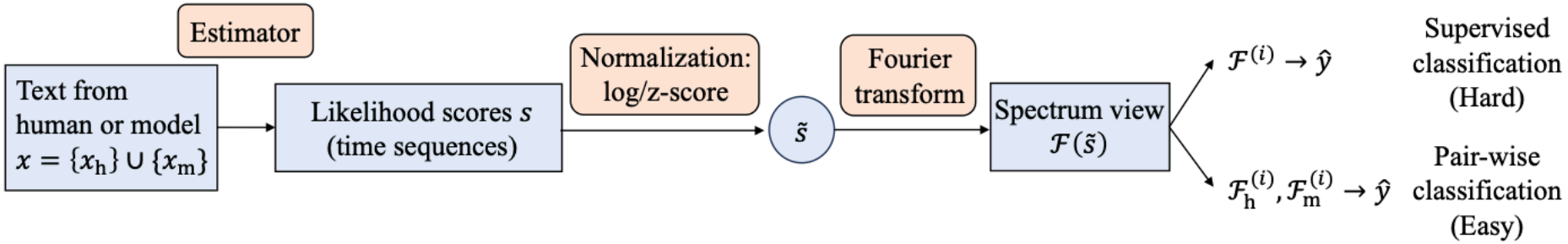
Supervised Learning Method(Spectrum) Framework



- **Augmented spectrum-based classifier** trained using frequency features derived from multiple circularized variants of the original likelihood sequence
- Process involves **circularization of the score sequence**, **Fourier transform** of each circularized version, and **averaging the resulting spectra** to form a robust feature representation
- Inspired by **circular convolution in signal processing**, this approach acts as a form of **data augmentation** to amplify weak periodicity and extract salient features for classification

Zero-shot Detection Methods

FourierGPT(Zero-shot) Framework



- **Heuristic spectrum-based classifier** designed to distinguish human-written and machine-generated texts using **frequency-domain features**
- Classification decision based on summed spectral power difference over selected low-frequency range with an empirical threshold

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

Data Preprocessing

Data Preprocessing for Supervised Method

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

Experiment Design

Supervised Learning Method Experiment

- Training parameters:
  - Epochs: 10
  - Learning rate Eng: 1×10^{-7}, Zh: 1×10^{-5}
  - Batch size: 16
  - Log interval: 50 steps
  - Optimizer: AdamW
- Models used:
  - English: bert-base-uncased, bert-base-multilingual-cased, roberta-base, xlm-roberta-base
  - Chinese: bert-base-chinese, xlm-roberta-base, roberta-base
- **Controlled variables** to ensure fair comparison:
  - Same training and validation splits used across all models
  - Consistent batch size, number of epochs, and optimizer settings per language
  - Evaluation metrics include accuracy, precision, recall, F1, and AUROC
- **Final step: Compare the five evaluation metrics across models and datasets**

Supervised Learning Method: Experimental Setup

- **In-Domain Evaluation:**
  - **Combine** data from all three domains (essay, reuter, wp) and **shuffle**
  - Split into train, validation, and test sets with ratio 8:1:1
  - Compute five evaluation metrics (accuracy, precision, recall, F1, AUROC) on test set
  - Report both **overall metrics** and metrics for **each individual domain**
- **Out-of-Domain (OOD) Evaluation:**
  - For each domain (A, B, C), **train and validate only** on data from **domain A**
  - Use **mixed-domain** data as the **test** set
  - Compute the five metrics on the test set overall and for each domain within test
  - **Discard metrics for domain A** in test, since training and validation data overlap
  - Repeat the same procedure by training/validating on domain B and C, respectively

FourierGPT supervised learning experiment

- Models used:
  - English: gpt2-xl, Mistral-7B-v0.1
  - Chinese: gpt2-chinese-cluecorpussmall, Wenzhong2.0-GPT2-3.5B-chinese
- **Controlled variables** to ensure fair comparison:
  - Consistent spectrum generation using circular shifts on NLL data with **fixed interpolation length**
  - Standard scaling, **fixed feature selection ($k = 120$)**, and **SVM** with **RBF kernel** and fixed hyperparameters
  - Comprehensive evaluation metrics including accuracy, precision, recall, F1 score, and AUROC;

Zero-shot detection method experiment

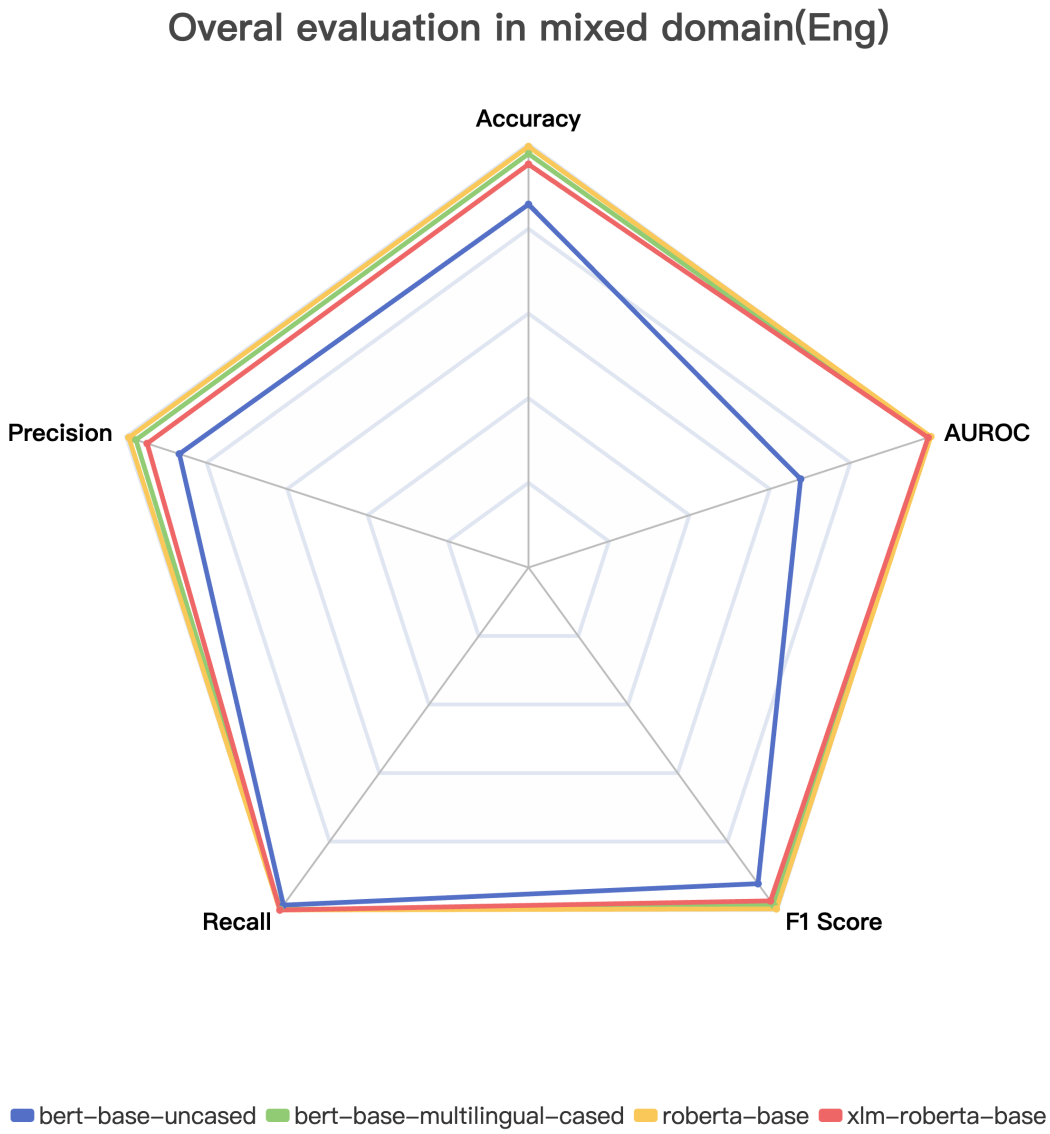
- Models used:
  - English: gpt2-xl, Mistral-7B-v0.1
  - Chinese: gpt2-chinese-cluecorpussmall, Wenzhong2.0-GPT2-3.5B-chinese
- **Controlled variables** to ensure fair comparison:
  - Same procedure to generate spectrum across all models
  - Same k range and threshold $\epsilon = 0$
  - Evaluation metrics include accuracy, and plot AUROC curve

Zero-shot Detection Method: Experimental Setup

- A sample is classified as model-generated if $\sum_{i=1}^{k} P_{\textbf{model}}(\omega_i) - \sum_{i=1}^{k} P_{\textbf{human}}(\omega_i) > \varepsilon$; otherwise, it is classified as human-written.
- The optimal $k \in [1, 50]$ is selected by evaluating accuracy, precision, recall, and F1 score for each k, and choosing the one with the **highest accuracy**.
- Given the **selected** k, multiple power thresholds are tested to compute true positive rate (TPR) and false positive rate (FPR) pairs, forming the ROC curve and enabling the calculation of AUROC.

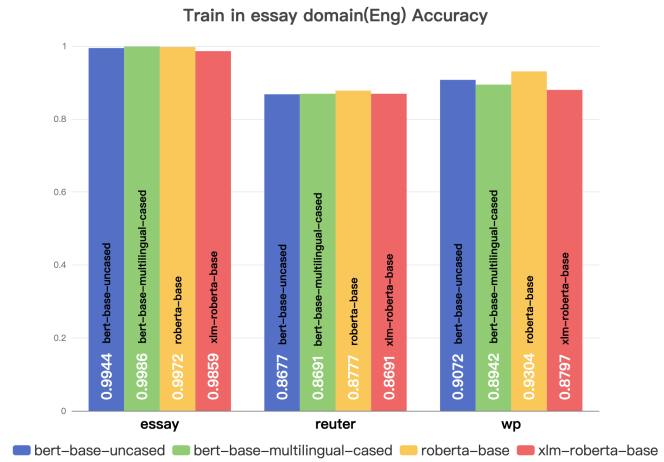
Result

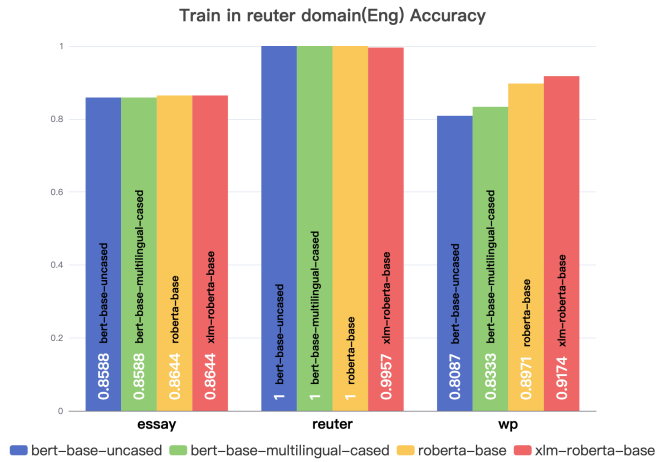
Supervised Learning Method(Eng):in domain

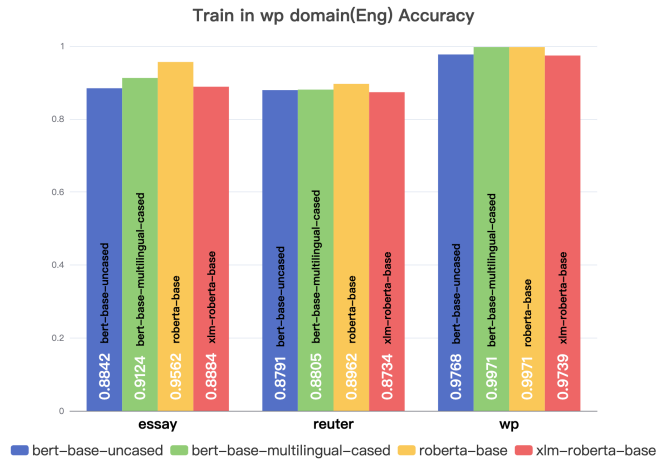


In the English mixed domain, **roberta-base and xlm-roberta-base** outperform bert-base-uncased and bert-base-multilingual-cased in accuracy, F1 score, and AUROC.

Supervised Learning Method (Eng): Out of Domain

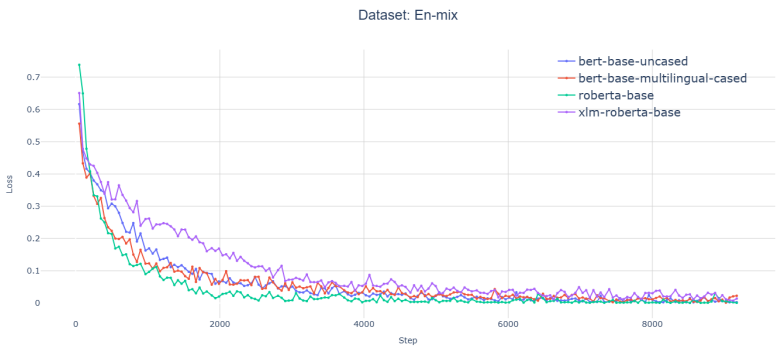




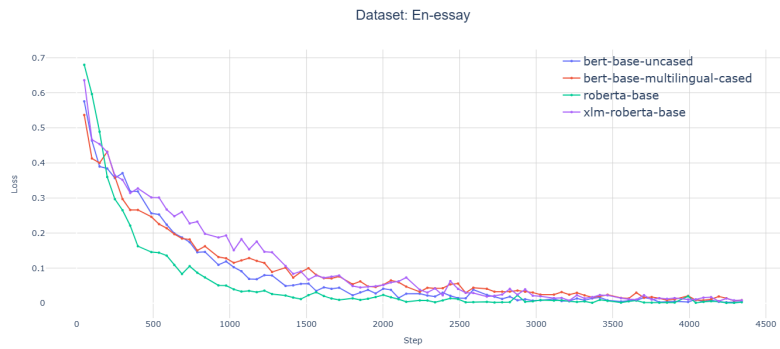


- Perform well within their **own training domain**, achieving accuracy generally **above 0.88**
- Accuracy significantly decreases when **tested on other domains**, often dropping **below 0.80**
- Among the four models, **roberta-base** shows relatively better generalization across different domains.

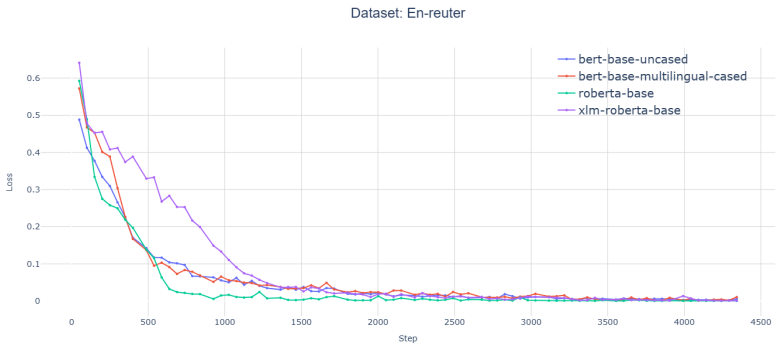
Supervised Learning Method: Loss Curve (Eng)



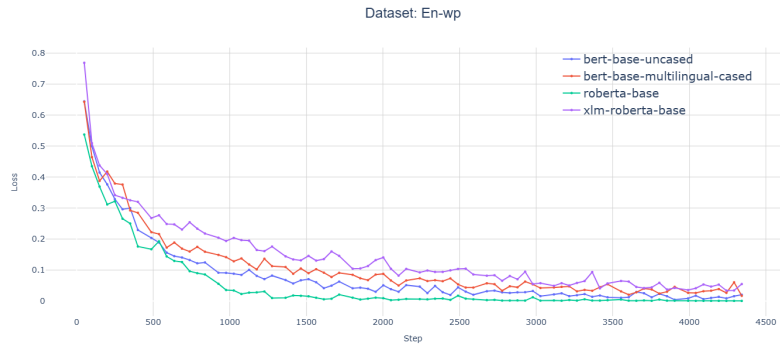
(a) Eng-mix



(b) Eng-essay



(c) Eng-reuter



(d) Eng-wp

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.

Supervised Learning Method: Metrics Comparison (Eng)

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

Supervised Learning Method: Metrics Comparison (Eng)

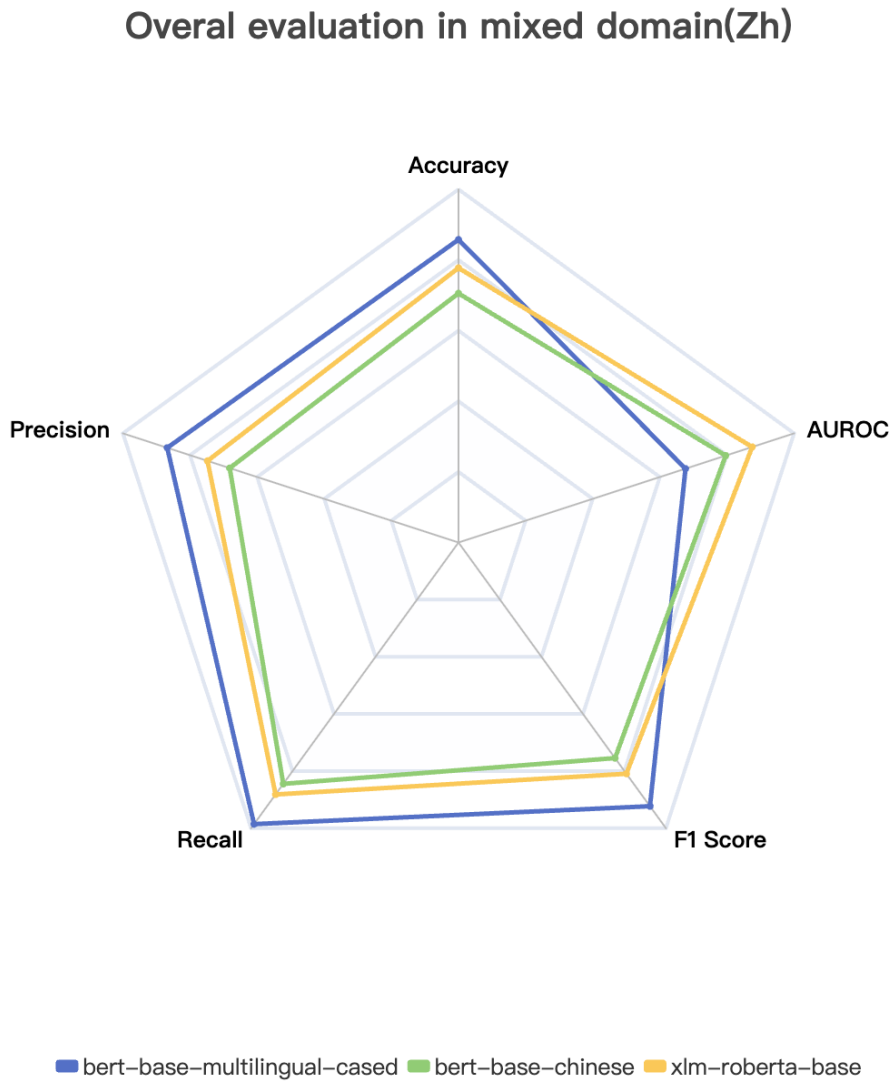
FT domain	mixed	essay	reuter	wp
Accuracy	0.9943 / 0.9887 / 0.9972 / 0.9971	0.9353 / 0.9972 / 0.8777 / 0.9304	0.9205 / 0.8644 / 1.0000 / 0.8971	0.9495 / 0.9562 / 0.8962 / 0.9971
Precision	0.9934 / 0.9871 / 0.9967 / 0.9967	0.9340 / 0.9967 / 0.8787 / 0.9330	0.9160 / 0.8644 / 1.0000 / 0.8948	0.9482 / 0.9532 / 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: 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

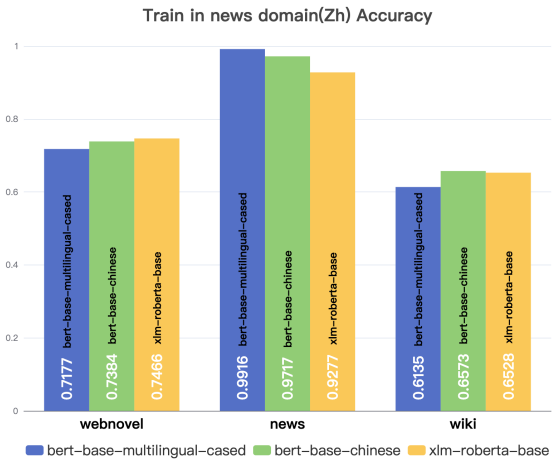
Table: xlm-roberta-base (Eng) overall/essay/reuter/wp

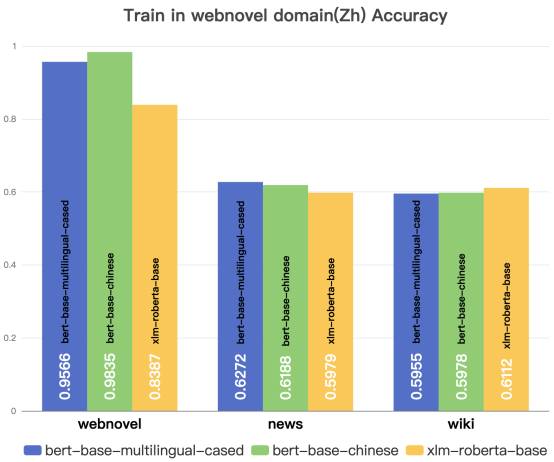
Supervised Learning Method(Zh):in domain

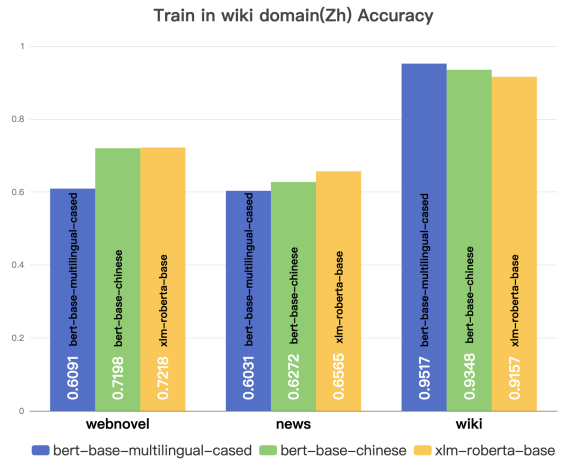


In the Chinese mixed domain, **bert-base-multilingual-cased** achieves the highest recall and precision, while **xlm-roberta-base** leads in AUROC and F1 score, outperforming bert-base-chinese across most metrics.

Supervised Learning Method (Zh): Out of Domain

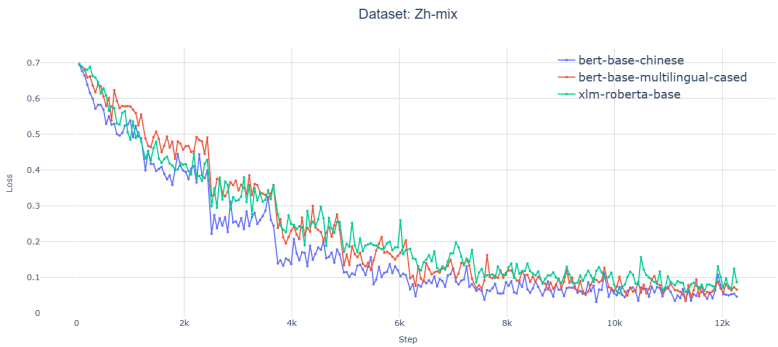




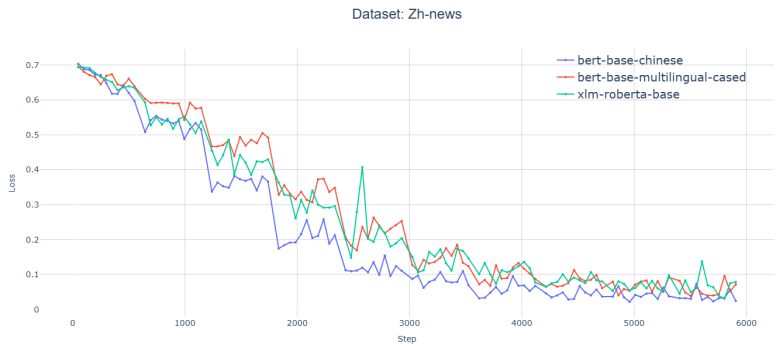


- Performance is strong within their **own training domain**, with accuracy mostly **above 0.83**
- Accuracy notably drops when **tested on other domains**, often falling in **0.63 to 0.72**
- Among the three models, **bert-base-chinese** shows relatively better generalization across different domains.

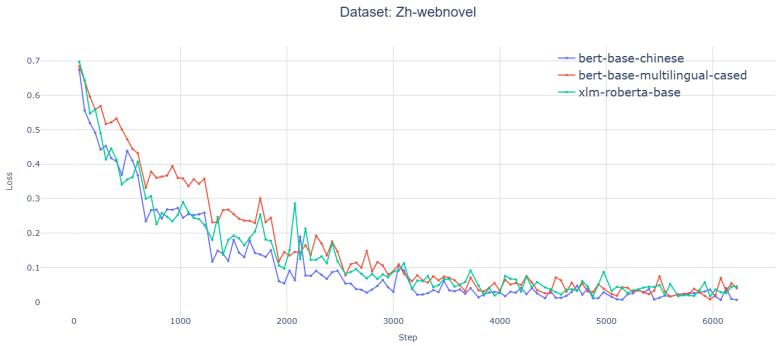
Supervised Learning Method: Loss Curve (Zh)



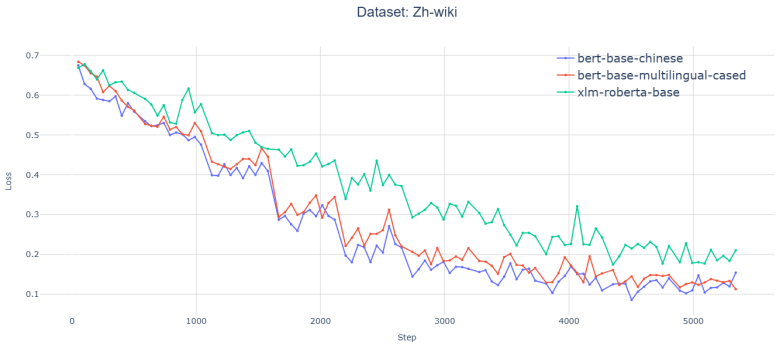
(a) Zh-mix



(b) Zh-news



(c) Zh-webnovel



(d) Zh-wiki

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.

Supervised Learning Method: Metrics Comparison (Zh)

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

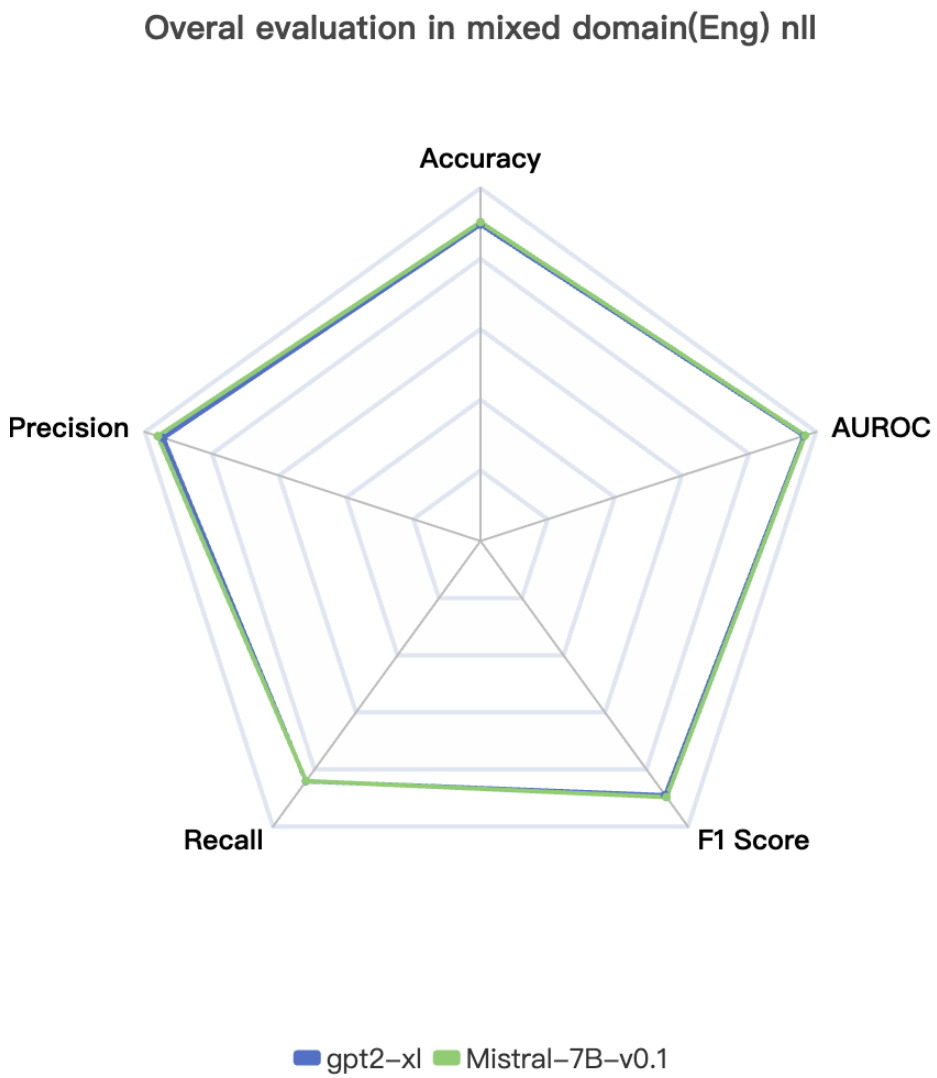
FT domain	mixed	webnovel	news	wiki
Accuracy	0.7219 / 0.7942 / 0.6754 / 0.6933	0.7376 / 0.9835 / 0.6188 / 0.5978	0.7920 / 0.7384 / 0.9717 / 0.6573	0.7564 / 0.7198 / 0.6272 / 0.9348
Precision	0.6893 / 0.7236 / 0.6452 / 0.7060	0.6741 / 0.9668 / 0.5869 / 0.5977	0.7511 / 0.6814 / 0.9500 / 0.6573	0.7251 / 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: bert-base-chinese (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

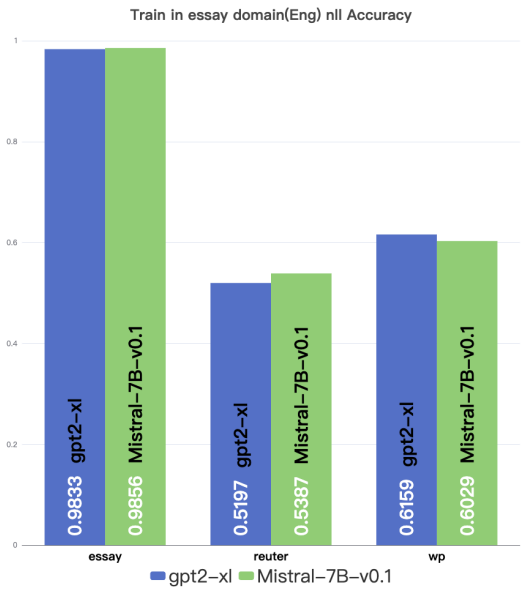
Table: xlm-roberta-base (Zh) overall/webnovel/news/wiki

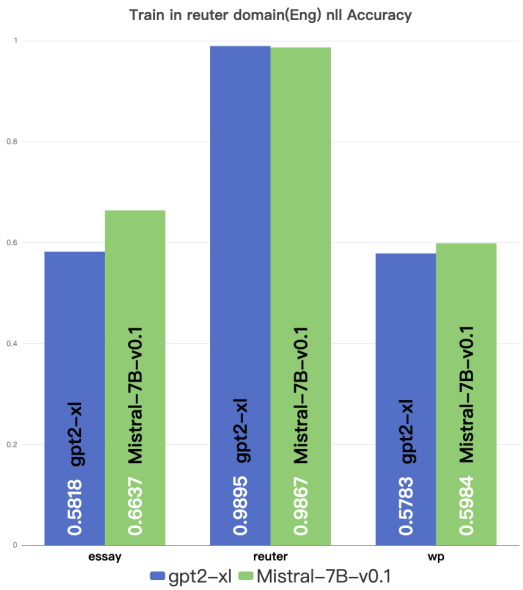
Supervised Learning Method(Spectrum)(Eng): In domain

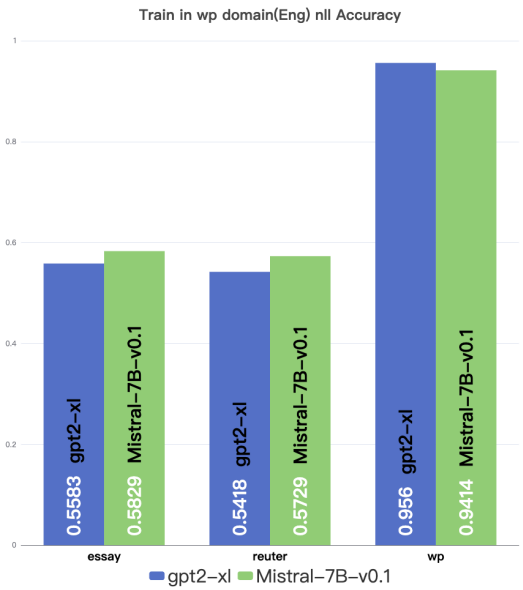


In the English mixed domain, the metrics of the two models are almost identical.

Supervised Learning Method(Spectrum)(Eng):OOD

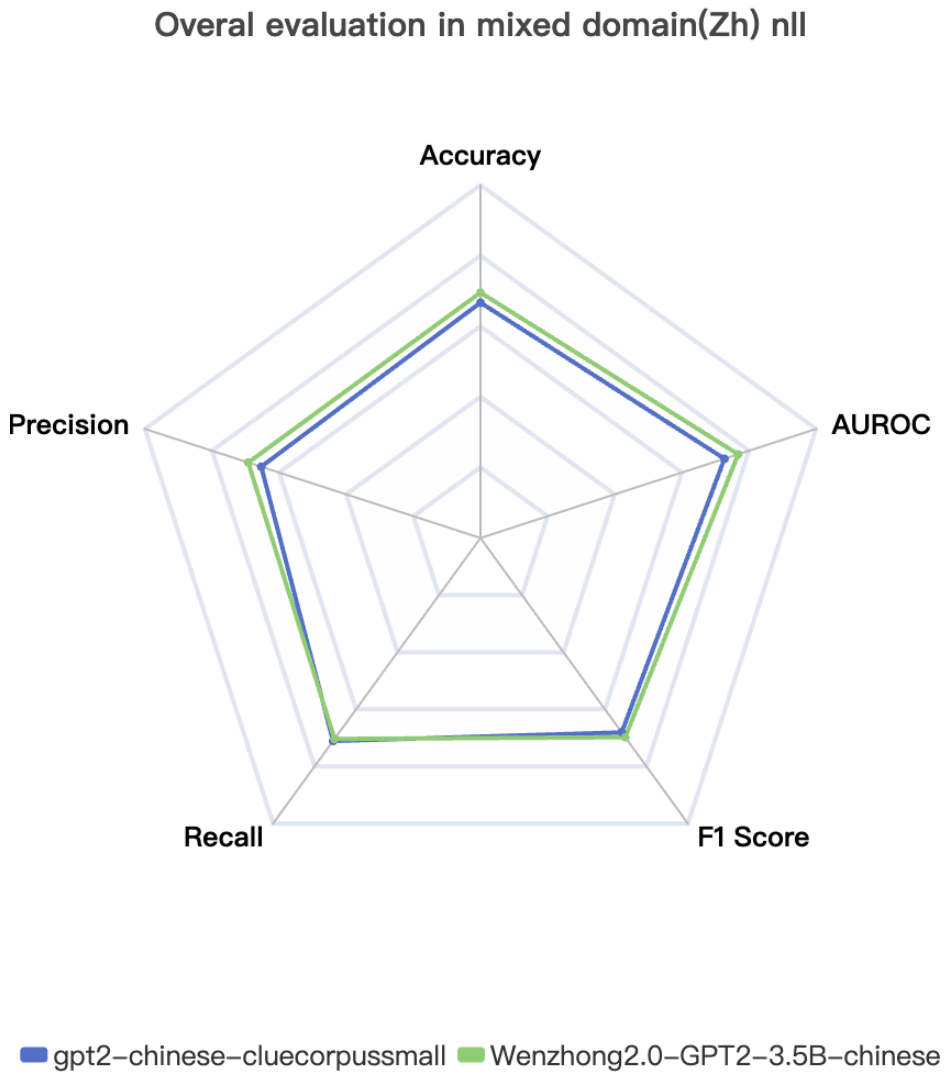






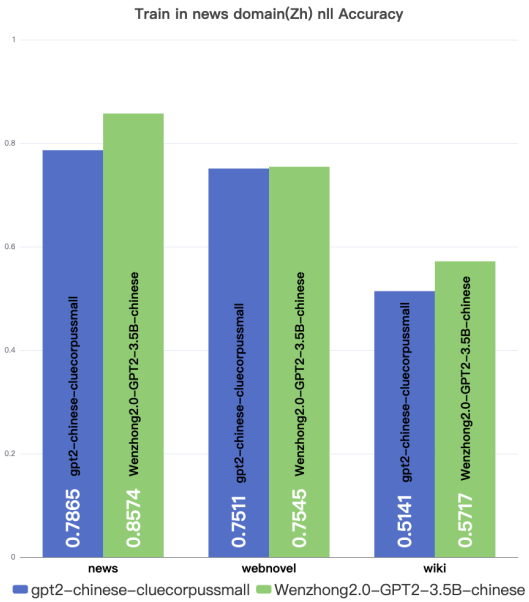
- Perform well within their **own training domain**, **above 0.95**
- Accuracy significantly decreases when **tested on other domains**, often dropping **below 0.6**
- Among the two models, **Mistral-7B-v0.1** shows relatively better generalization across different domains.

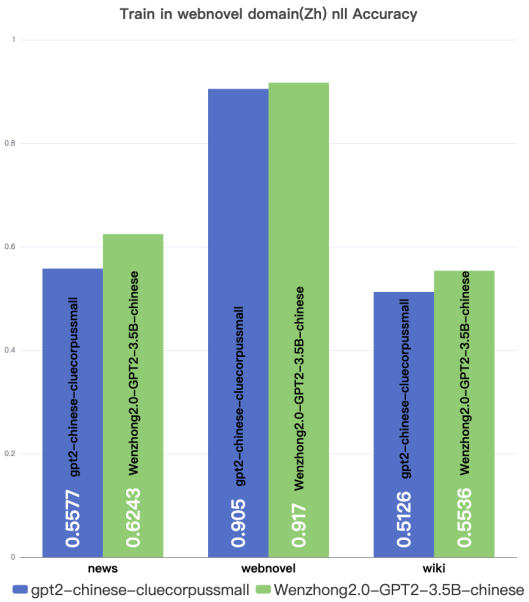
Supervised Learning Method(Spectrum)(Zh):In domain

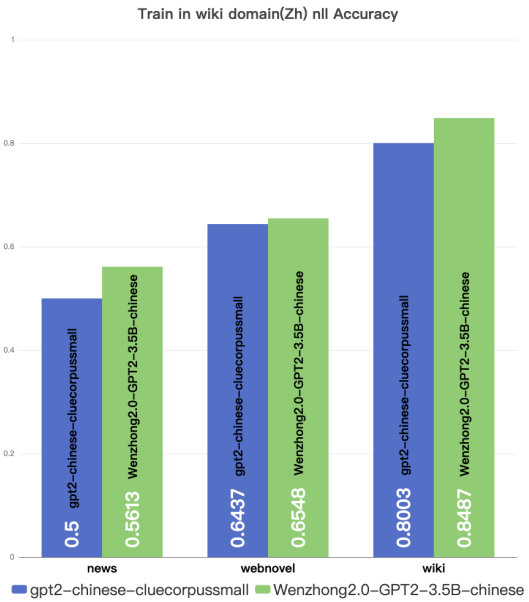


In the Chinese mixed domain, **Wenzhong2.0-GPT2-3.5B-chinese** performs better across most metrics except recall.

Supervised Learning Method(Spectrum)(Zh):OOD

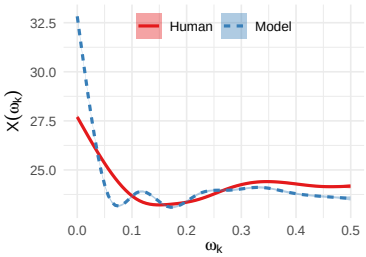






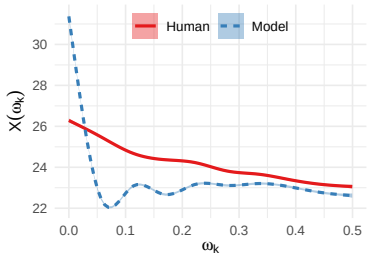
- Perform well within their **own training domain**, **above 0.8**
- Accuracy significantly decreases when **tested on other domains**, often dropping **below 0.6**
- Among the two models, **Wenzhong2.0-GPT2-3.5B-chinese** shows relatively better generalization across different domains.

Zero-shot Detection: Aggregated Power Spectrum



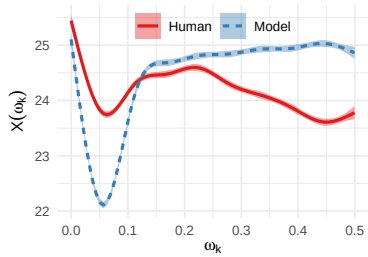
k=5,

higher=model



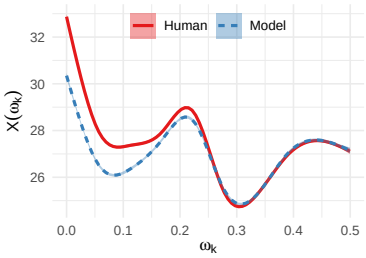
k=5,

higher=model



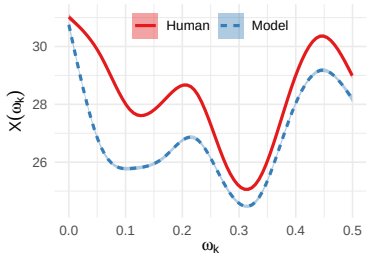
k=3,

higher=model



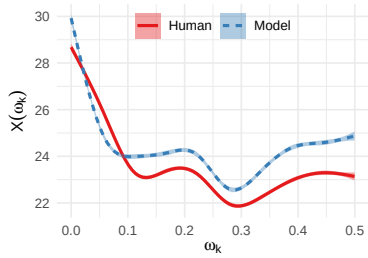
k=46,

higher=human



k=50,

higher=human

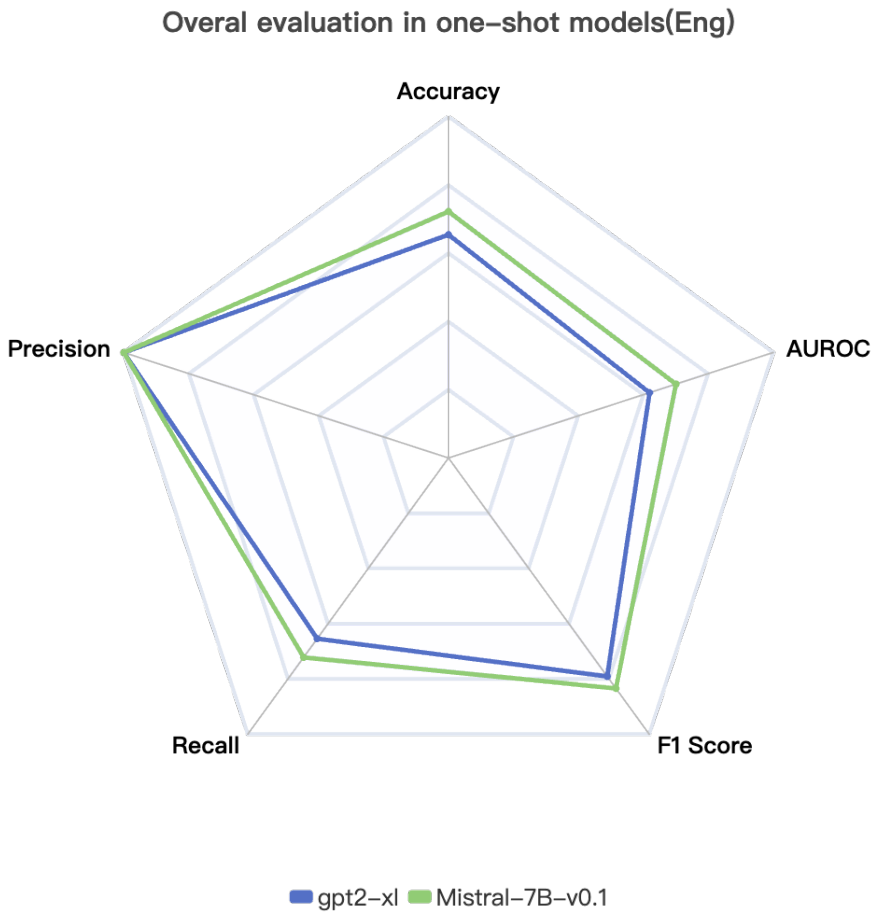


k=49,

higher=human

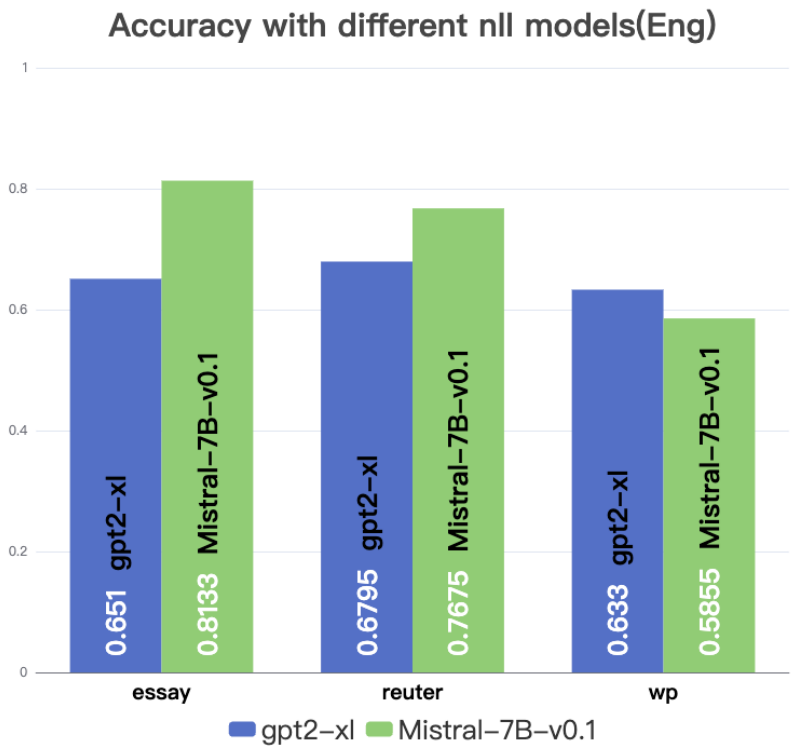
- For English texts, model can distinguish between hm and llm texts using only **a small number of low-frequency** features.
- For **English** texts, the llm text has **higher** power than the hm text while for **Chinese**, where llm text has **lower** power..

Zero-shot Detection (Eng)



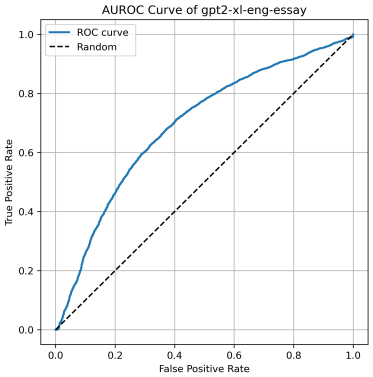
Mistral-7B-v0.1 slightly outperforms gpt2-xl across most metrics in one-shot detection for English texts.

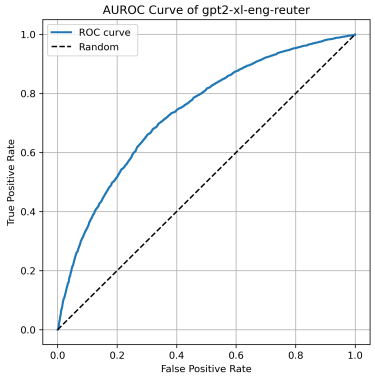
Zero-shot Detection: Accuracy (Eng)

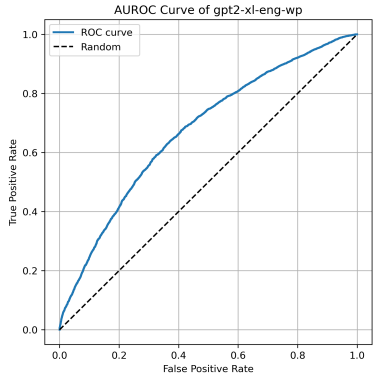


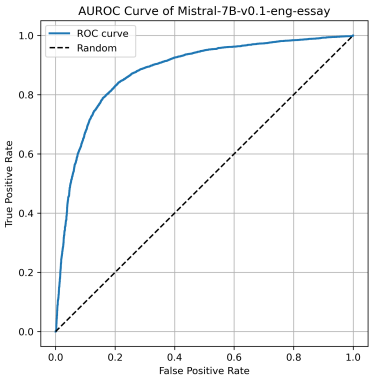
- **Mistral-7B-v0.1** Outperforms GPT2-XL in Essay and Reuter Domains(**both above 0.75**)
- Both Models **Struggle** on the wp Domain(**0.63 and 0.59**)
- **Mistral-7B-v0.1** generate **better nll**

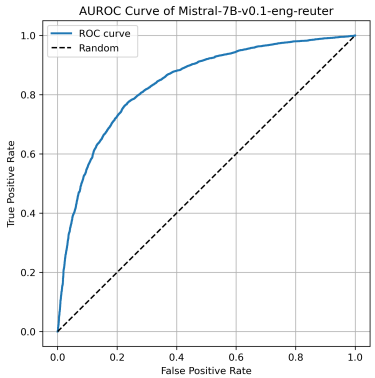
Zero-shot Detection: AUROC curve (Eng)

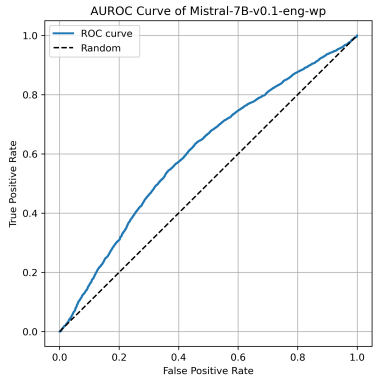












- On the essay and reuter, the ROC curve of **Mistral-7B-v0.1** is closer to the top-left corner, indicating stronger ability to **distinguish positive and negative samples**.
- The ROC curves on the **wp** are closer to the diagonal line, suggesting a higher chance of random guessing.

Zero-shot Detection: Metrics Comparison (Eng)

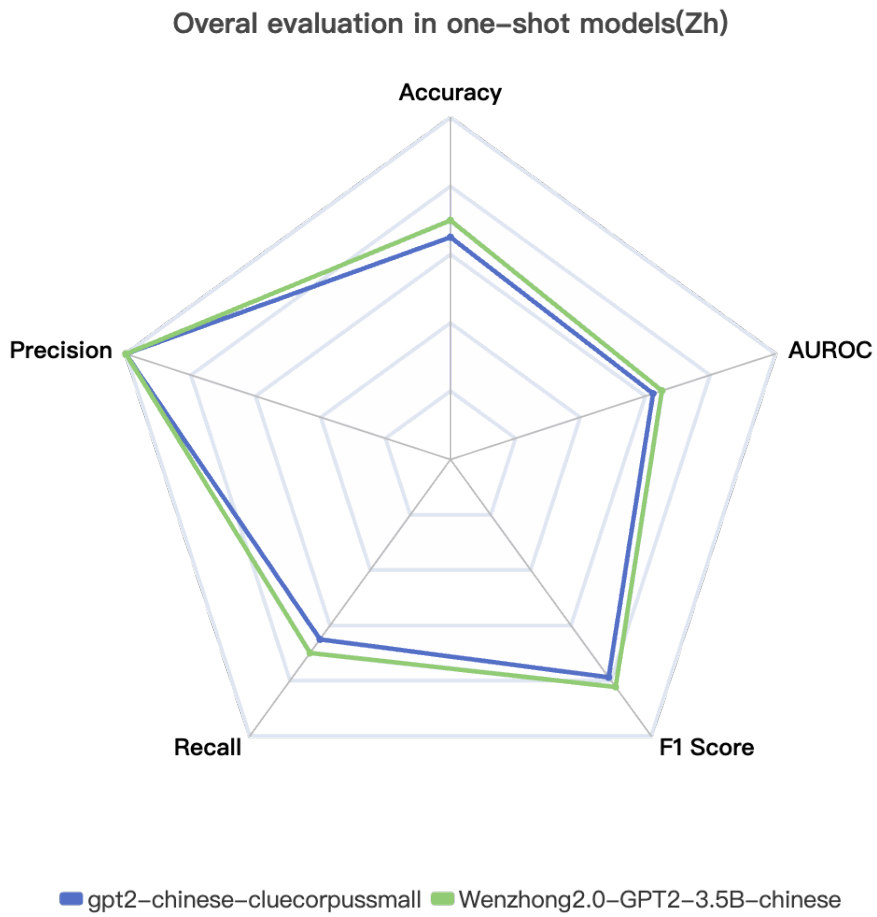
FT domain	**essay**	**reuter**	**wp**
Accuracy	0.6510	0.6795	0.6330
AUROC	0.6899	0.7348	0.6732

Table: gpt2-xl-Eng

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

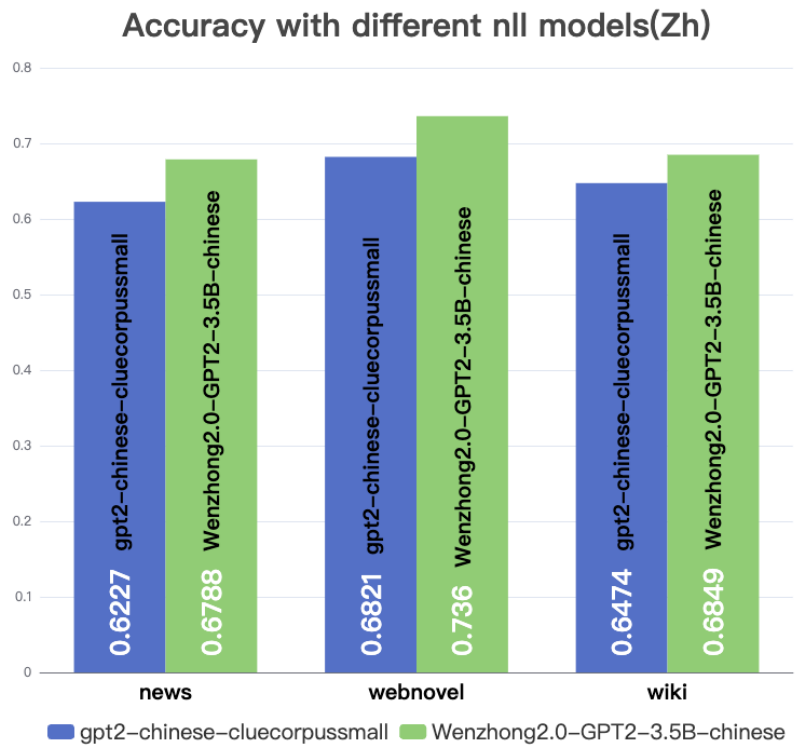
Table: Mistral-7B-v0.1-Eng

Zero-shot Detection (Zh)



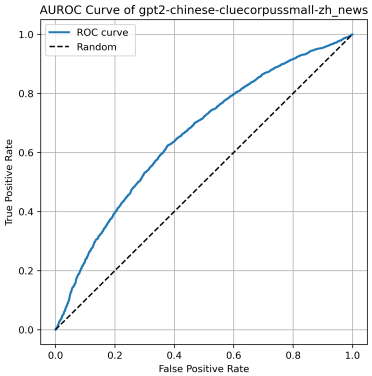
Wenzhong2.0-GPT2-3.5B-chinese slightly outperforms gpt2-chinese-cluecorpussmall across most metrics in one-shot detection for Chinese texts.

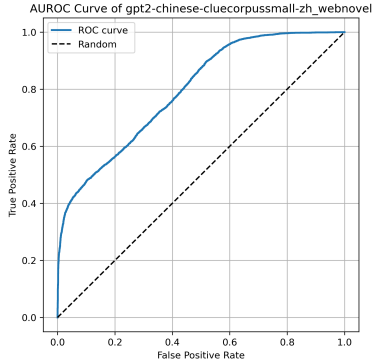
Zero-shot Detection: Accuracy (Zh)

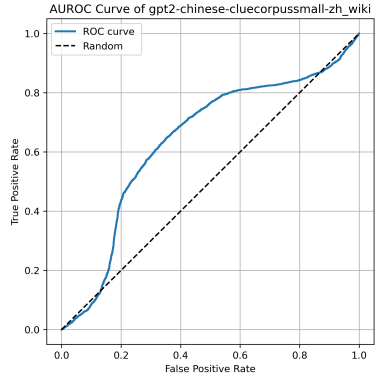


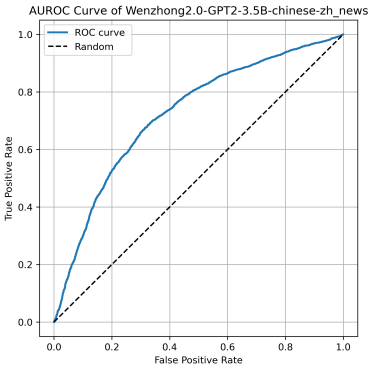
- **Wenzhong2.0-GPT2-3.5B** outperforms gpt2-chinese-cluecorpussmall across all domains
- Both models achieve the highest accuracy and AUROC in the **webnovel** domain(**0.73 and 0.82**)
- Detecting differences in **news** texts is more challenging(**0.64 and 0.68**)

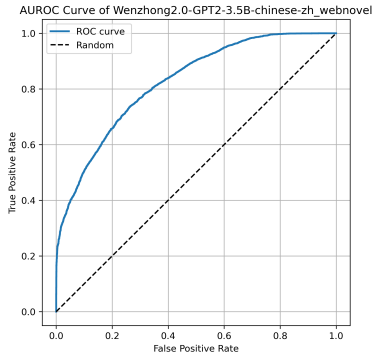
Zero-shot Detection: AUROC curve (Zh)

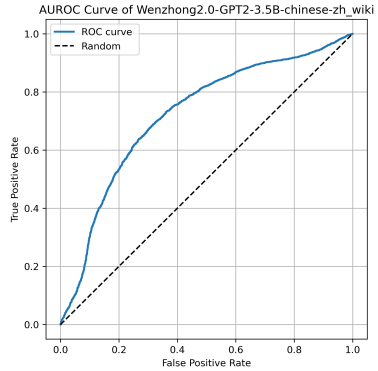












- The ROC curves for g**pt2-chinese-cluecorpussmall** show good **early rise** in news and webnovel datasets but remain flat and close to the diagonal in the wiki dataset, indicating weaker discrimination.
- The **Wenzhong2.0-GPT2-3.5B** has steeper, more left-top concentrated curves in news and webnovel datasets, demonstrating stronger classification performance.

Zero-shot Detection: Metrics Comparison (Zh)

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

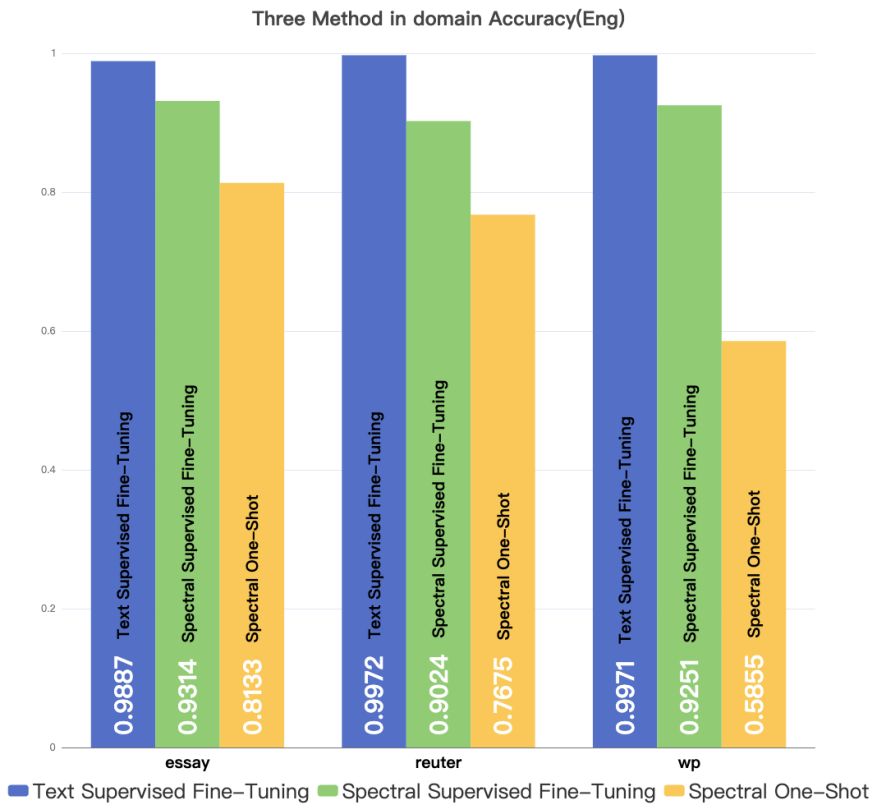
Table: gpt2-chinese-cluecorpussmall-Zh

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

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

Result analysis

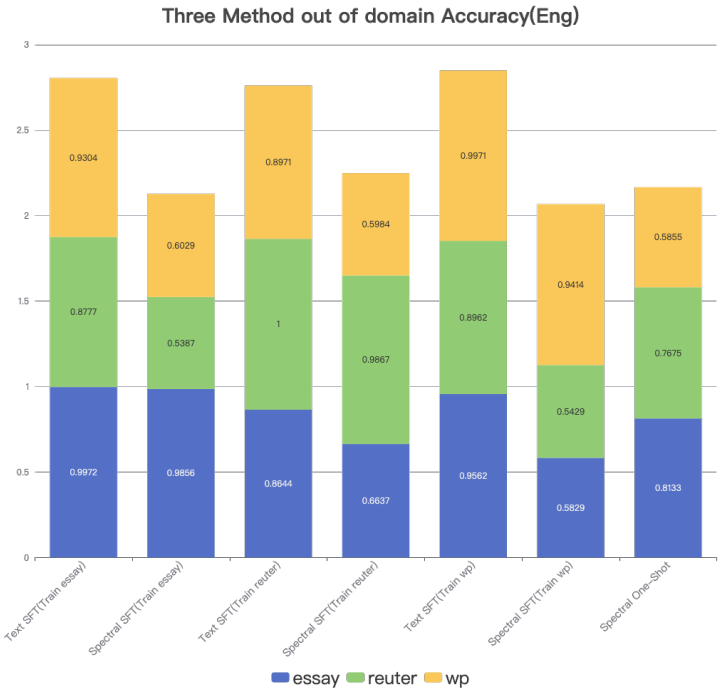
Method comparison(English in domain)



Text Supervised Fine-Tuning achieves the highest accuracy across all domains.

Zero-Shot shows the lowest accuracy and **high domain sensitivity**, especially in the wp domain(**0.58**).

Method comparison(English out of domain)

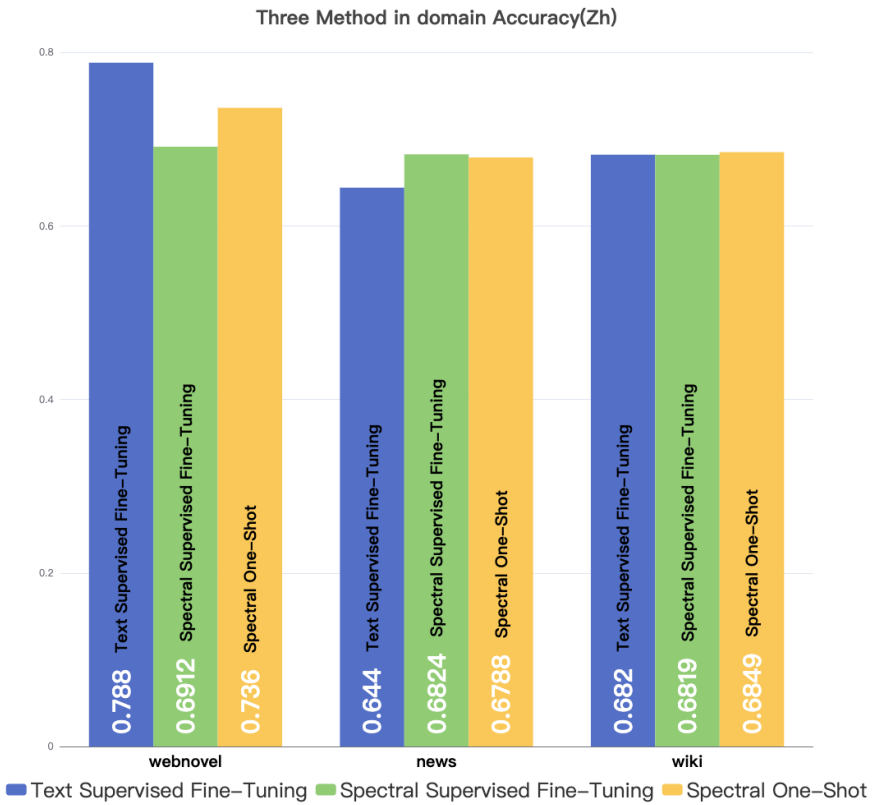


Text Supervised Fine-Tuning demonstrates **strong generalization across domains**(above 0.85)

S**pectral Supervised Fine-Tuning** performs poorly in cross-domain(below **0.6**)

Spectral Zero-Shot performs well on **essay(0.81) and reuter(0.77)**

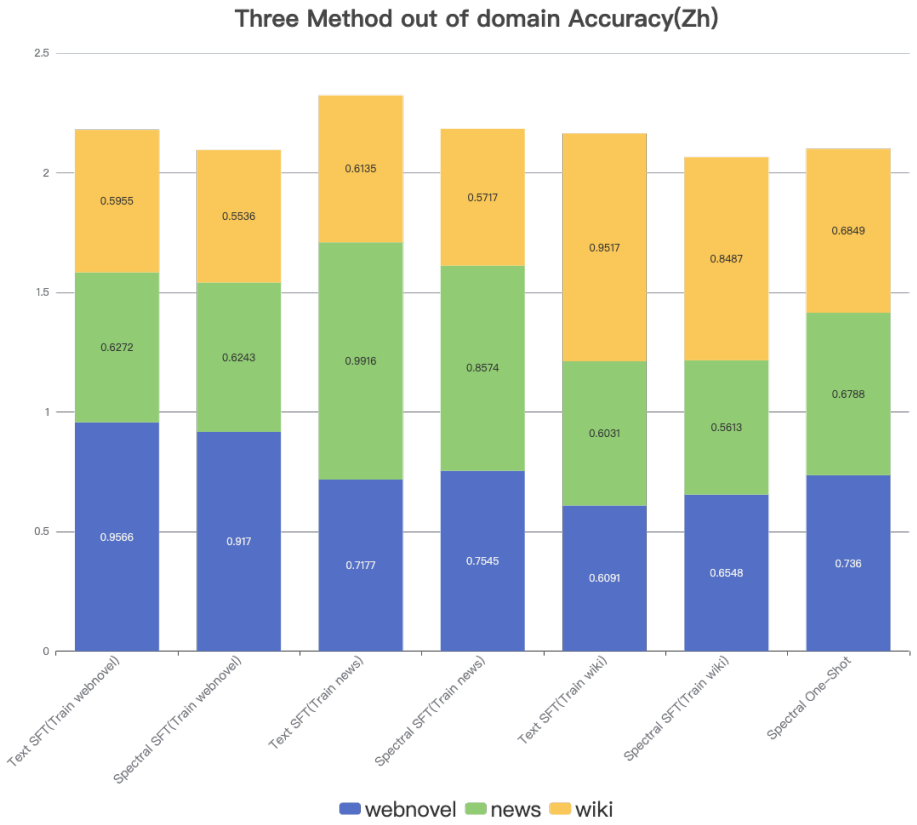
Method comparison(Zh in domain)



Text Supervised Fine-Tuning outperforms all other methods in the **webnovel** domain(**0.788**)

Two FourierGPT methods perform competitively in the **news and wiki** domains(around **0.68**)

Method comparison(Zh out of domain)



Two **supervised learning** methods performs poorly(below **0.65**)

Zero-shot detection demonstrates **strong generalization across domains**(above **0.65**)

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.

Reasoning

1. **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.

2. **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.

Conclusion

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.

Future Work

- 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.