README

Final Project for 50.040 Natural Language Processing

Mohammad Saif Zia | Rose Evangeline Anne Dagman Destor | Joel Lim | Tania Koh Tze Ern

Default project brief

Refer to ***Project brief.pdf***

Directory structure

```
Root/
  config.py
  utils.py
  optimizer_test.npy
  Task 1 and 2.ipynb
  Task 3 per language.ipynb
  Task 3 all language.ipynb
  pretrain.txt
  XNLI-1.0/
      xnli.dev.jsonl
      xnli.dev.tsv
      xnli.test.jsonl
      xnli.test.tsv
  XNLI-MT-1.0/
      multinli/
          multinli.train.en.tsv
          multinli.train.es.tsv
          ...
          multinli.train.vi.tsv
          multinli.train.zh.tsv
      xnli/
          xnli.dev.en.jsonl
          xnli.dev.en.tsv
          xnli.test.en.jsonl
          xnli.test.en.tsv
  README.md
```

Reproducibility

1. Git clone this project `git clone https://github.com/NLP-Project-2025/Main-tasks.git`
2. Install dependencies `pip install -r requirements.txt`
3. For inference, download our weights from Google Drive and store the corresponding weights (`.pt`) in folder `best_model`. > Note: You may need to change the directories accordingly especially for the dataloader and model file paths.

4. For training and testing, download XNLI dataset (monolingual) from

- https://dl.fbaipublicfiles.com/XNLI/XNLI-1.0.zip and save under `XNLI-1.0`.
- https://dl.fbaipublicfiles.com/XNLI/XNLI-MT-1.0.zip and save under `XNLI-MT-1.0`.

   You can also download our custom dataset (code-switched) in HuggingFace. See repo for implementation.

5. How to run:

- `Task1_2.ipynb` → Custom GPT-2 implementation + English NLI fine-tuning
- `Task3_per_lang.ipynb` → Zero-shot multilingual evaluation + Per-language fine-tuning
- `Task3_all_lang.ipynb` → All-language fine-tuning
- `Attention_visualisation.ipynb` → For attention layer analyses on negation, temporal order and lexical overlap
- Extended task is found in this repo.
- Our code-switched generation is found in this repo.

Results Summary

Language	Tokenizer Fertility ↓	Zero-shot (EN-tuned) Acc. (%)	Per-language Acc. (%)	Multilingual (all-lang) Acc. (%)
English	1.11	83.23	83.23	83.31
French	1.75	43.83	—	—
Spanish	1.83	46.47	70.16	72.46
Swedish	2.08	39.92	—	—
German	2.10	40.28	—	—
Turkish	2.73	40.50	—	—
Vietnamese	3.62	39.22	65.67	68.34
Chinese (zh)	3.77	37.94	65.47	66.10
Arabic	4.70	37.19	—	—
Hindi	5.12	34.55	—	—
Urdu	5.16	34.93	—	—
Bulgarian	5.53	37.72	—	—
Russian	5.90	37.52	—	—
Greek	6.16	35.79	—	—
Thai	9.48	34.85	—	—

Results from Task 3. Languages in bold were selected for further fine-tuning. All models were trained using the AdamW optimizer with the following hyperparameters:

- Epochs: 5 (the checkpoint with the highest test accuracy was chosen, regardless of epoch)
- Batch size: 8
- Learning rate: 2×10
- Weight decay: 0.01
- Warmup ratio: 0.1

Overall (en+vi+zh+es)
- **Multilingual mixed-test accuracy**: **72.57%** (20040 samples)

To-do: add Full report link