Configuration Manual

MSc Research Project
MS in Data Analytics

Shweta Yadav
Student ID: x21209251

School of Computing
National College of Ireland

Supervisor: Prasanth Nayak

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Shweta Yadav
Student ID:	x21209251
Programme:	MS in Data Analytics
Year:	2023
Module:	MSc Research Project
Supervisor:	Prasanth Nayak
Submission Due Date:	14/08/2023
Project Title:	Configuration Manual
Word Count:	898
Page Count:	6

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Shweta Yadav
Date:	18th September 2023

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Configuration Manual

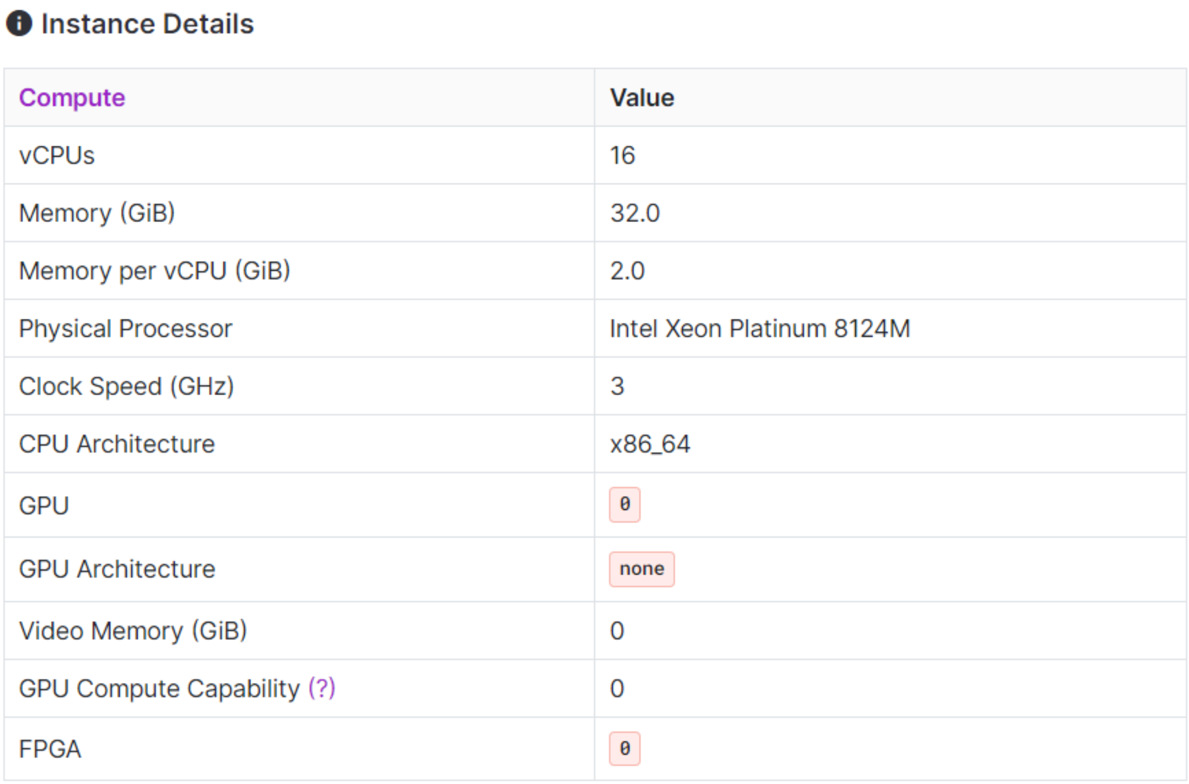
Shweta Yadav
x21209251

1 Introduction

This configuration manual lists all hardware and software requirements to replicate the results of the research. A step-by-step guide taken from data acquisition to model implementation are described in this manual.

2 Hardware and Software Configurations

The model used for this study is OpenAI's davinci, which is a large language model, and fine-tuning such large pre-trained LLMs is a computation-intensive task. Selection of appropriate resources helps in the efficient execution of code. The entire framework from data gathering to model evaluation is done on the AWS EC2 instance with ubuntu AMI. Figure 1 shows the aws EC2 instance details used for the experiment.



Instance Details

Compute	Value
vCPUs	16
Memory (GiB)	32.0
Memory per vCPU (GiB)	2.0
Physical Processor	Intel Xeon Platinum 8124M
Clock Speed (GHz)	3
CPU Architecture	x86_64
GPU	0
GPU Architecture	none
Video Memory (GiB)	0
GPU Compute Capability (?)	0
FPGA	0

Figure 1: EC2 Instance Details

For the experiment 'Python' is used as the programming language. Table 1 details the libraries used and their respective versions.

Library	**Version**
openai	0.27.8
bs4	0.0.1
pandas	2.0.3
numpy	1.25.0
sacrebleu	2.3.1
nltk	3.8.1

Table 1: Python libraries and versions

3 Dataset acquisition

TICO-19 dataset Anastasopoulos et al. (2020) is publicly available for advancing the research and enhancing machine translation for Covid-19-related information. The dataset contains domain-specific terminologies and translated sentences that can be used for evaluating the model output. Terminologies are available in sgm files and are to be extracted by hierarchically selecting the correct tag and attribute. Once terms are extracted, these are fed to a generative AI model to generate a parallel corpus(source and target language statements) which finally acts as the input for the research. Generating parallel corpus is an iterative process and repeated depending on the model's translation accuracy.

4 Project development

Only the most crucial, necessary steps have been discussed in this section. The implementation of the study is divided into four subsections. The code is available on github[1] The folder structure followed for executing the code is as follows:

1. /home/ubuntu/thesis/data
2. /home/ubuntu/thesis/data/input/
3. /home/ubuntu/thesis/data/output/
4. /home/ubuntu/thesis/script/
5. /home/ubuntu/thesis/fineTune/

4.1 Generating training data

1. Install the libraries mentioned in Table 1 using command *pip install library name*
2. Sign up at OpenAI [2] for accessing the API and generating the API key. This will be a paid service. Update the key in the generateText.py script available at location */home/ubuntu/thesis/script*.

---

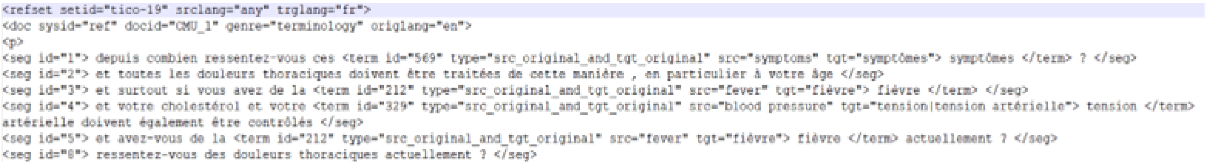
[1] `https://github.com/shweta-0511/fineTuningDavinci/tree/master`

[2] `https://openai.com/blog/chatgpt`

3. Download the input file test.en-fr.fr.sgm from TICO-19[3]. Figure 2 shows the sample records from the sgm file for terminology extraction. Upload the file to location */home/ubuntu/thesis/data/input/*.

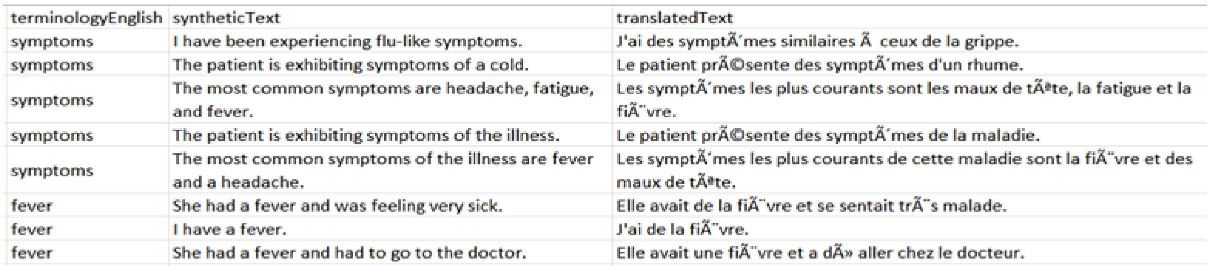
4. Execute script generateText.py, to extract terms and call API for synthetic text and its translation. Update line 62 in the script with chosen source language and line 101 with the chosen target language. Use command *python3 /home/ubuntu/thesis/script/generateTex*
Figure 3 shows the generated parallel corpus store in a csv file.

5. Upon successful execution the script will generate a csv file at the location
/home/ubuntu/thesis/data/output/



```
<refset setid="tico-19" srclang="any" trglang="fr">
<doc sysid="ref" docid="CMU_1" genre="terminology" origlang="en">
<p>
<seg id="1"> depuis combien ressentez-vous ces <term id="569" type="src_original_and_tgt_original" src="symptoms" tgt="symptômes"> symptômes </term> ? </seg>
<seg id="2"> et toutes les douleurs thoraciques doivent être traitées de cette manière , en particulier à votre âge </seg>
<seg id="3"> et surtout si vous avez de la <term id="212" type="src_original_and_tgt_original" src="fever" tgt="fièvre"> fièvre </term> </seg>
<seg id="4"> et votre cholestérol et votre <term id="329" type="src_original_and_tgt_original" src="blood pressure" tgt="tension|tension artérielle"> tension </term>
artérielle doivent également être contrôlés </seg>
<seg id="5"> et avez-vous de la <term id="212" type="src_original_and_tgt_original" src="fever" tgt="fièvre"> fièvre </term> actuellement ? </seg>
<seg id="8"> ressentez-vous des douleurs thoraciques actuellement ? </seg>
```

Figure 2: Sample data in TICO-19



terminologyEnglish	syntheticText	translatedText
symptoms	I have been experiencing flu-like symptoms.	J'ai des symptÃ´mes similaires Ã ceux de la grippe.
symptoms	The patient is exhibiting symptoms of a cold.	Le patient prÃ©sente des symptÃ´mes d'un rhume.
symptoms	The most common symptoms are headache, fatigue, and fever.	Les symptÃ´mes les plus courants sont les maux de tÃªte, la fatigue et la fiÃ¨vre.
symptoms	The patient is exhibiting symptoms of the illness.	Le patient prÃ©sente des symptÃ´mes de la maladie.
symptoms	The most common symptoms of the illness are fever and a headache.	Les symptÃ´mes les plus courants de cette maladie sont la fiÃ¨vre et des maux de tÃªte.
fever	She had a fever and was feeling very sick.	Elle avait de la fiÃ¨vre et se sentait trÃ¨s malade.
fever	I have a fever.	J'ai de la fiÃ¨vre.
fever	She had a fever and had to go to the doctor.	Elle avait une fiÃ¨vre et a dÃ» aller chez le docteur.

Figure 3: Parallel Corpus

4.2 Fine-Tuning the model

Once data is generated. The next step is to transform the data into the correct format for fine-tuning the model.

1. Execute script fineTuneData.py to convert csv to JSON. Use command
python3 /home/ubuntu/thesis/script/fineTuneData.py,

2. Upon successful execution the script will generate a JSON file at the location. Figure 4 shows the sample data from the transformed file.
/home/ubuntu/thesis/fineTune/data/

3. Set OpenAI key using command *export OPENAI_API_KEY=”your API key”*

4. Execute command to start data transformation.
openai tools fine_tunes.prepare_data -f /home/ubuntu/thesis/fineTune/data/data.json

5. Read the prompt and enter **Y or n** to accept or reject the transformations.

---

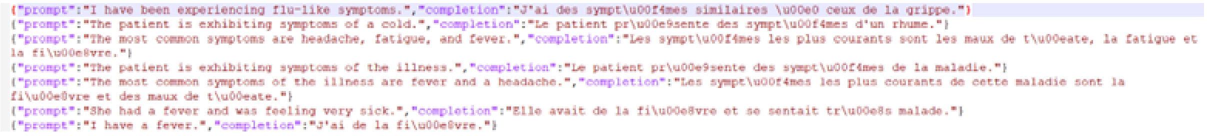
[3] `https://www.statmt.org/wmt21/terminology-task.html`

6. Upon successful completion, the jsonL file is written at the location.
   /home/ubuntu/thesis/fineTune/data/.
   Figure 5 shows the sample data from jsonL file. This will be the input for fine-tuning the model.

7. Execute command to start fine-tuning.
   openai api fine_tunes.create -t /home/ubuntu/thesis/fineTune/data/data_prepared.jsonl -m davinci

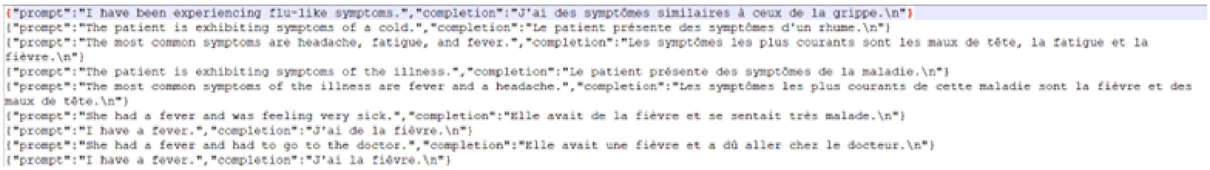
8. Execute command to resume fine-tuning in case execution is interrupted.
   openai api fine_tunes.follow -i fine-tune id returned by last command

9. Execute command to get the status of fine-tuning process.
   openai api fine_tunes.get -i fine-tune id
   Wait and rerun the command until the status changes to completed and the fine-tuned model name is returned. Nomenclature followed *davinci:ft-personal-YYYY-mm-dd-hh-mm-ss*



```
{"prompt":"I have been experiencing flu-like symptoms.","completion":"J'ai des sympt\u00f4mes similaires \u00e0 ceux de la grippe."}
{"prompt":"The patient is exhibiting symptoms of a cold.","completion":"Le patient pr\u00e9sente des sympt\u00f4mes d'un rhume."}
{"prompt":"The most common symptoms are headache, fatigue, and fever.","completion":"Les sympt\u00f4mes les plus courants sont les maux de t\u00eate, la fatigue et
la fi\u00e8vre."}
{"prompt":"The patient is exhibiting symptoms of the illness.","completion":"Le patient pr\u00e9sente des sympt\u00f4mes de la maladie."}
{"prompt":"The most common symptoms of the illness are fever and a headache.","completion":"Les sympt\u00f4mes les plus courants de cette maladie sont la
fi\u00e8vre et des maux de t\u00eate."}
{"prompt":"She had a fever and was feeling very sick.","completion":"Elle avait de la fi\u00e8vre et se sentait tr\u00e8s malade."}
{"prompt":"I have a fever.","completion":"J'ai de la fi\u00e8vre."}
```

Figure 4: JSON data for fine-tuning



```
{"prompt":"I have been experiencing flu-like symptoms.","completion":"J'ai des symptômes similaires à ceux de la grippe.\n"}
{"prompt":"The patient is exhibiting symptoms of a cold.","completion":"Le patient présente des symptômes d'un rhume.\n"}
{"prompt":"The most common symptoms are headache, fatigue, and fever.","completion":"Les symptômes les plus courants sont les maux de tête, la fatigue et la
fièvre.\n"}
{"prompt":"The patient is exhibiting symptoms of the illness.","completion":"Le patient présente des symptômes de la maladie.\n"}
{"prompt":"The most common symptoms of the illness are fever and a headache.","completion":"Les symptômes les plus courants de cette maladie sont la fièvre et des
maux de tête.\n"}
{"prompt":"She had a fever and was feeling very sick.","completion":"Elle avait de la fièvre et se sentait très malade.\n"}
{"prompt":"I have a fever.","completion":"J'ai de la fièvre.\n"}
{"prompt":"She had a fever and had to go to the doctor.","completion":"Elle avait une fièvre et a dû aller chez le docteur.\n"}
{"prompt":"I have a fever.","completion":"J'ai la fièvre.\n"}
```

Figure 5: JsonL data for fine-tuning

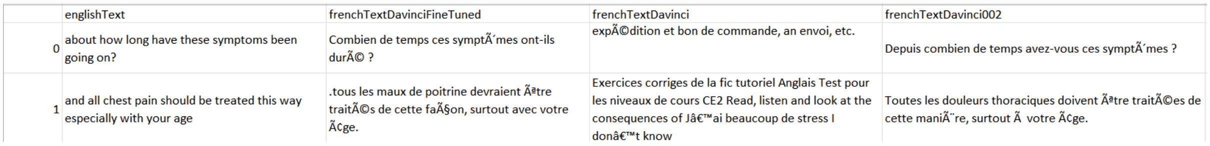
4.3 Generating evaluation data

Once the model is fine-tuned. The next step is to download and upload test data and generate translations using the models for comparison.

1. Download and upload blind_test.en-fr.en.sgm at location
   /home/ubuntu/thesis/data/input

2. Execute script generateTestData.py using command
   python3 /home/ubuntu/thesis/script/generateTestData.py.

3. Upon successful completion the script will generate csv file containing the source language sentence and translated output from the three models at the location.
   /home/ubuntu/thesis/data/output
   Figure 6 shows the sample data from the evaluation data generated.



	englishText	frenchTextDavinciFineTuned	frenchTextDavinci	frenchTextDavinci002
0	about how long have these symptoms been going on?	Combien de temps ces symptÃ´mes ont-ils durÃ© ?	expÃ©dition et bon de commande, an envoi, etc.	Depuis combien de temps avez-vous ces symptÃ´mes ?
1	and all chest pain should be treated this way especially with your age	.tous les maux de poitrine devraient Ãªtre traitÃ©s de cette faÃ§on, surtout avec votre Ã¢ge.	Exercices corriges de la fic tutoriel Anglais Test pour les niveaux de cours CE2 Read, listen and look at the consequences of Jâ€™ai beaucoup de stress I donâ€™t know	Toutes les douleurs thoraciques doivent Ãªtre traitÃ©es de cette maniÃ¨re, surtout Ã votre Ã¢ge.

Figure 6: Evaluation Data

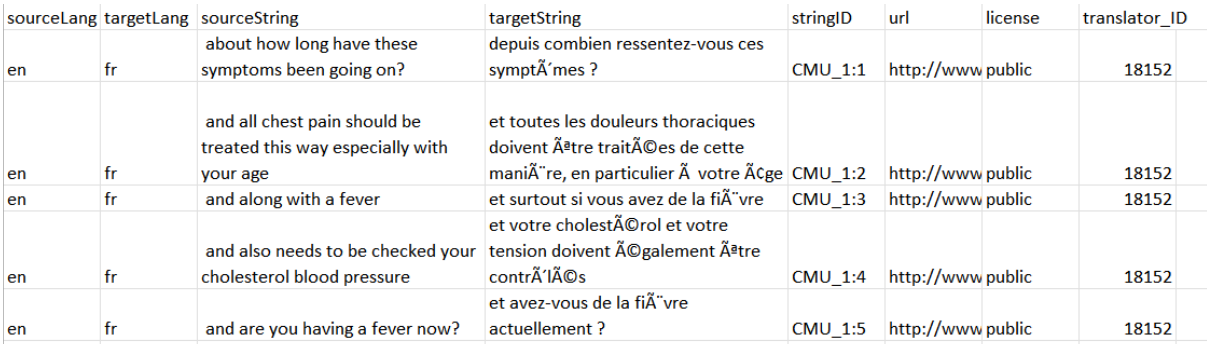
4.4 Evaluating the model output

The last step is to evaluate the translation quality.

1. Download and upload test.en-fr.tsv file at the location. */home/ubuntu/thesis/data/input.*
   Figure 7 shows the sample records from the file.

2. Execute script evaluateModel.py using command *python3 /home/ubuntu/thesis/script/evaluateModel.py.*

3. The script will return BLEU score of the three models.



sourceLang	targetLang	sourceString	targetString	stringID	url	license	translator_ID
en	fr	about how long have these symptoms been going on?	depuis combien ressentez-vous ces symptÃ´mes ?	CMU_1:1	http://www	public	18152
en	fr	and all chest pain should be treated this way especially with your age	et toutes les douleurs thoraciques doivent Ãªtre traitÃ©es de cette maniÃ¨re, en particulier Ã votre Ã¢ge	CMU_1:2	http://www	public	18152
en	fr	and along with a fever	et surtout si vous avez de la fiÃ¨vre	CMU_1:3	http://www	public	18152
en	fr	and also needs to be checked your cholesterol blood pressure	et votre cholestÃ©rol et votre tension doivent Ã©galement Ãªtre contrÃ´lÃ©s	CMU_1:4	http://www	public	18152
en	fr	and are you having a fever now?	et avez-vous de la fiÃ¨vre actuellement ?	CMU_1:5	http://www	public	18152

Figure 7: Ground Truth Data

References

Anastasopoulos, A., Cattelan, A., Dou, Z.-Y., Federico, M., Federman, C., Genzel, D., Guzmán, F., Hu, J., Hughes, M., Koehn, P. et al. (2020), ‘Tico-19: the translation initiative for covid-19’, *arXiv preprint arXiv:2007.01788* .