**BSc Project Log Book**

**Student Name: Amirali Famili**

**Project Name: Emotion Interpretability in LLMs**

**Internal Supervisor: Dr. Rodrigo Souza Wilkens**

**External Supervisor: (if any) No External Supervisor**

**Log of Activities**

**Date**

**Key activities**

**Outcomes**

**Comments/Action plans**

**02/10/2024**

**Going through project ideas**

**Interest in emotion learning in LLMs**

**Need more time to decide on the scope of the project**

**03/10/2024 - 08/10/2024**

**Ideas on the project topic, scope and planning for the meeting**

**Designed detailed diagram of the architecture of the model**

**Project goals are still unclear and needs supervision**

**12/10/2024**

**Went through Ideas with supervisor.**

**Decided major aspects of the project.**

**Still need to decide topic (title) of the project**

**23/10/2024 - 25/10/2024**

**Research on the unanswered questions for the project topic.**

**Narrowed down the topics in my interest**

**Should discuss the topics with supervisor**

**25/10/2024 - 30/10/2024**

**Discussing ideas with supervisor regarding topic, goals, etc…**

**Decided on the topic of Emotion learning in LLMs**

**Need to research the explainability techniques and model(QWEN)**

**31/10/2024**

**Research and think on the project goals**

**Wrote down and shared the clear project goals with supervisor.**

**Need further research on the topic.**

**1/11/2024 - 3/11/2024**

**Research on QWEN and Explainability techniques, red articles**

**Obtained a clear vision on the scope of the project.**

**Need more detailed research on QWEN architecture, literature**

**04/11/2024 - 08/11/2024**

**Worked on the literature review and further research**

**Made a draft and sent it to supervisor.**

**Literature review needs major improvements (based on feedback)**

**09/11/2024 - 13/11/2024**

**Resolved the issues of the literature review mentioned by the supervisor, improved the architecture and content of the literature review.**

**Made another draft and shared it with my supervisor**

**The improvements mentioned by the supervisor helped a lot and made me correct the architecture of the literature review and its contents and scope.**

**15/11/2024 - 19/11/2024**

**Red articles, in depth research for the literature review, key improvements of the literature review, added transformer architecture.**

**Made the final adjustments to the literature review and made it ready for submission.**

**The abstract still needs work, as well as the conclusion part, the project roadmap and specification needs to be reviewed and corrected.**

**28/11/2024 - 07/12/2024**

**Defining project repository, file structure, downloaded necessary resources, and researched on different code implementations.**

**Gathered knowledge on probing techniques, as well as touch and transformer libraries. Managed to interact with hugging face hub in order to download default models on local computer.**

**I should pre-process the datasets, I have downloaded and take action on how to probe simple models.**

**Due to lack of memory I couldn’t download QWEN2-7b, instead I plan to start working on smaller models like BERT to accomplish the probing task.**

**10/12/2024 - 14/12/2024**

**Wrote the code for pre-processing the datasets (GoEmotion and ISEAR), made relevant plots for demonstration in Jupyter notebook, and a separate python file for pre-processing and calling the datasets into code.**

**Managed to finish the work on the datasets, which made my work easier by just importing the file and calling the relevant function for retrieval of the dataset as a data frame.**

**The ISEAR dataset in particular had several issues, I downloaded it from this GitHub repository https://github.com/JULIELab/ISEAR, some parts of it required manual corrections and some was done with python.**

**16/12/2024 - 18/12/2024**

**Wrote a code for probing the BERT model and managed to get some results from this code on where does emotion learning happens in BERT based on the data from GoEmotion and ISEAR datasets.**

**Got a clear understanding on how does a classifier probe works, how to work with default models locally and access their hidden states and attention heads.**

**Each time I run the code it could take 2-3 hours, with only the first 1000 samples. This process involves: giving one batch of samples and their labels to the model, only calculate the forward pass (since we are only working on the default model and weights are frozen), extract the hidden states of the model for each sample, give the hidden states of each of the n layers to each of the n probes as inputs along with the same label passed to the model. After training each probe with 3 epochs, I validate the probes by getting each of the n probes to classify the 1000 samples and try to predict the label (probes with highest accuracies are considered the layers where emotion learning is happening). All this was done in a single loop! (With No Programming Norm)**

**22/12/2024 - 30/12/2024**

**Wrote a code for downloading different QWEN models into my local computer and attempting to pass a single line of string to the model to validate their response and extract the hidden states for that string.**

**I ran into various errors and issues just tying to use different models of the QWEN series (\*\*\*). My main issue was that I didn’t have a Nvidia GPU to take advantage of CUDA, to resolve it I had to offload the hole model on CPU (So far I only managed to get it to work for QWEN2-0.5B model, which is considered the most lightweight of all the QWEN series).**

**I have made a detailed explanation regarding the issues I ran into and how I managed to resolve them in this hackmd notebook:**

**https://hackmd.io/@JvKZuse-TeuIH4jY23k2TA/HkFWZ25Lyl**

**I am struggling with my available storage space, my Mac has 128Gb of space almost 100Gb of which are file systems. Moreover, Despite the large size of the QWEN model series them selves (shards), since the number of layer (as well as neurones per layer) increases when working with big models with 7 billion parameters, the size of the extracted hidden states also increases (extracted per layer), which leads to memory issues.**

**Even ignoring all the memory issues, since I don’t have a Nvidia GPU, so I can’t properly run QWEN models (since it’s recommended by it’s developers to use CUDA), and my kernel always crashes when trying to load the model shards, because I have to offload the entire model on CPU/RAM which rans out of memory.**

**08/1/2025**

**Shared what I have done so far with my supervisor.**

**They also suggested that I should work on a smaller model like QWEN2-0.5B.**

**I think at this stage it’s best to write a generalised code which works with every given model name and given dataset. Doing so would make my work easier once I got the computational power needed for working on larger QWEN models.**

**Date**

**Key activities**

**Outcomes**

**Comments/Action plans**

**Date**

**Key activities**

**Outcomes**

**Comments/Action plans**

**Date**

**Key activities**

**Outcomes**

**Comments/Action plans**

**Date**

**Key activities**

**Outcomes**

**Comments/Action plans**

**Date**

**Key activities**

**Outcomes**

**Comments/Action plans**

**Date**

**Key activities**

**Outcomes**

**Comments/Action plans**

**Date**

**Key activities**

**Outcomes**

**Comments/Action plans**

**Date**

**Key activities**

**Outcomes**

**Comments/Action plans**

**Date**

**Key activities**

**Outcomes**

**Comments/Action plans**

**Date**

**Key activities**

**Outcomes**

**Comments/Action plans**

**Date**

**Key activities**

**Outcomes**

**Comments/Action plans**

**Date**

**Key activities**

**Outcomes**

**Comments/Action plans**

**Date**

**Key activities**

**Outcomes**

**Comments/Action plans**