FYP Plan Proposal

If downsized several small LLMs together is impossible to be more cost efficient than a buying tokens for a larger LLM, should we use ‘cooperation between small LLMs’ itself as in the Introduction to my Report?

I would like to start to Write Report around next Feb to March, how much should we push to that stage? I would like to already have our small LLMs deployed by that time so there is a nice period for me to get the data in the final outcome.

~~Acquisition of fine tuning material: I’ll talk to Alexandra about what we already have, about what that Computer Science MSC student done, but ideally just finetuning a 1B model to do a simple task (like give correct and several incorrect) might need 1000+ fine tuning data.~~

Time line for 2 months

Main Parts:

1. Prompt gpt-4o-mini structured as what we expect Llama3 as it to be (1 to more structure etc.) and fix the output issue of Llama3 using correct structure of Chain. Week 11
2. Achieve the Multi-Agent cooperation with the chain build, including 1 to many list
   1. comparison agent (from many to many split down to 1 to many),
   2. reasoning agent (for use 1 line to give reason)
      1. reasoning agent give parts (the thing mentioned in the meeting that for 1+1 = 2 , when the answer is -2 instead of just giving wrong, tell the sign is incorrect)
   3. parsers between
   4. other optimisation agents (grammar checker etc.)
3. Deployment, the input interface integration etc. work, I really want this to happen before the end of the project so there are nice outcome I can show.

I think maybe 2 can be done in 8 weeks, 1 agent by another and meanwhile I can focus on the data acquisition. Idk if this is fair