L173 Research proposal and report

D.J. Wischik

This course has been about taking interesting ML topics and translating them into the language of probability modelling, typically involving KL minimization. In the research report, you will do the same for an ML topic of your choice: you will take a vague topic and formulate it as a well-posed probability-modelling question. This assignment has two parts:

- research proposal, about 2/3 of a page (5% of total grade)
- research report, approximately 4,000 prose words (45% of total grade)

Your proposal should explain what your topic is, and state what you hope to achieve. I'm not expecting a probability model yet! I just want to see that you have spotted what there is to model, and the rough kind of model you are aiming for. Don't write a vacuous goal such as "I aim to explore ...". Instead, you might write something like "I conjecture that we can express this as a probability model for response ... along the lines of ... from session ..." **For the research report,** marks will be awarded for the formulation, i.e. for turning your conjecture into a precise probability model, not for building a working system. In general it need not involve any coding at all.

- Some of the topic suggestions are very open-ended PhD-level questions. You are expected to do 4,000 words' worth of interesting thinking, not to give a complete answer.
- If you find during your work that the formulation part is fairly straightforward, not enough to write 4,000 words about, then you may wish to include computational work — but such work needs to revolve around probability modelling, for example finding ways to test whether your proposed probability model actually fits the data, rather than simply presenting a classic ML leaderboard or ablation table. If you do computational work, small models and simple datasets are sufficient; there is no need to train at scale.

Your proposal is not binding. If your work on the report pulls you in a different direction, you are free to follow that direction, and you do not need to submit an updated proposal. You may however like to discuss your change in direction with the lecturer, to make sure your new direction is sound.

AI usage is permitted. State briefly in your report how you used it; one paragraph is sufficient.

Topic suggestions

You are free to propose your own topic. You might for example write a report that translates an ML paper of your choice, or a handful of papers, into the language of probability modelling. You might choose a topic that's related to your MPhil dissertation. Below is a list of further topic suggestions.

1. Give a rigorous account of t-SNE as an autoencoder, and sketch out a 't-SNE explorer' that can generate outputs given a point in the dimension-reduced space.
2. In the session on GANs, there are slides (not covered in lectures) that describe a probability-modelling interpretation of adversarial image generation. They explain how it can be seen as a constrained optimization problem, not unlike SFT. What would be the DPO equivalent?
3. Suggest a probability-modelling interpretation of adversarial text generation, for example designed to fool a classifier. The approach for generating adversarial images is based on making small tweaks to an image; for text, can we use the small tweaks from text diffusion?
4. What is the most hurtful thing you can say to an LLM? In other words, we want to optimize over tokens, to maximize the pain vector. [A lot of the translation was already done in lectures, so I'd expect to see coding for this report.]
5. Build a theory-of-mind model for an LLM that extracts several *independent* dimensions simultaneously, e.g. 'how frustrated the LLM is with the user' and 'how bored the LLM is with the topic'.
6. With AI agents, we can install *hooks* to review their responses and possibly reject them. Can a hook be baked into weights, as a critic?
7. Train a grouped conditional VAE on the FEMNIST dataset (`https://huggingface.co/datasets/flwrlabs/femnist`) which has images of letters grouped by writer. [I'd expect to see coding for this report.]

8. In the study into LLM experience of pain, the dataset was manually-labelled sentences. Suggest an approach based on contrastive learning instead. A good starting point is Yuksekgonul et al., "When and why vision-language models behave like bags-of-words" (`https://arxiv.org/abs/2210.01936`).

9. The text diffusion of Nie et al. was for a fixed-length output. Modify it to generate arbitrary-length outputs, based on world models.

10. When discussing world models, we considered the value of a function at one or more given target x-coordinates. In Oasis (the transformer-based version of Minecraft), we might think of the camera's viewpoint as the target location — but that's not actually how it works. Instead, it treats the user's motion controls as an input, and it doesn't calculate x-coordinates at all. Suggest how to model this.

11. Describe how to use a critic to train a world model.

12. Take a paper (or several papers) on uncertainty in language models and explain, in the language of this course, what it is measuring. A particularly good paper is by Kalai et al. (`https://arxiv.org/abs/2509.04664`) which discusses how an LLM would report its uncertainty.

13. One way to get a model to express its uncertainty is by providing an 'abstain' option. Explain how it should be scored, and what it achieves.

14. In the session on uncertainty, it was suggested that preference fine-tuning can achieve the same outcome as an uncertainty oracle. How?

15. In the session on uncertainty, it was suggested that uncertainty arising from data-in-context $\mathcal{O}$ and uncertainty arising from data-in-training $\mathcal{D}$ are essentially the same thing. How?