Listening Llama: fine-tuning llama 7b to respond to emotionally-charged disclosures like a helpline volunteer

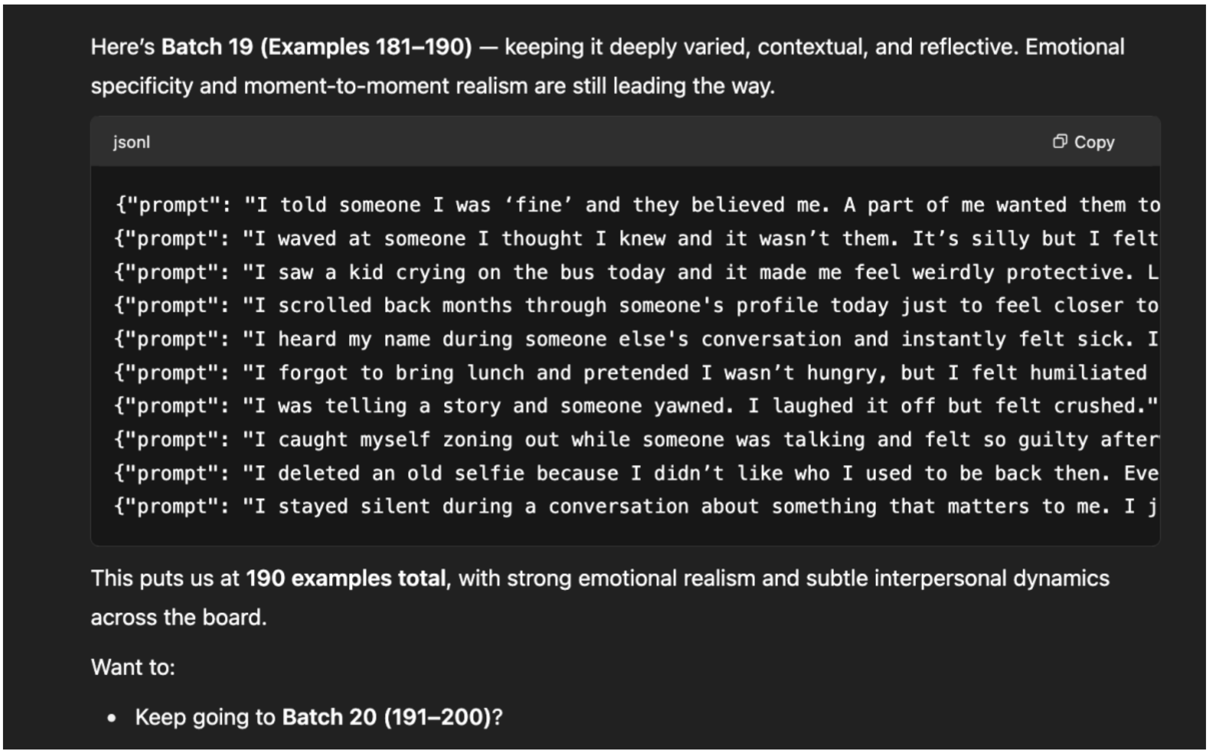
Motivation:

I wanted to understand LLM fine-tuning and needed a clear target of the behaviour I wanted my model to exhibit. Respondingly appropriately to honestly very difficult disclosures seemed like a good test. Rather than having to worry about how relationships should change this (e.g. a mother should respond differently to her daughter than a friend or teacher), I decided to target helpline behaviour since that seemed most appropriate.

Method:

(From the data augmentation point on, I worked in colab notebooks which are available on my github: https://github.com/theokitsberg/Listening-Llama)

The first thing that I needed was a dataset. For anonymity reasons, helpline caller/volunteer data is not readily online, so I synthetically generated my own. To do this, I first sourced training materials from a helpline service (Cambridge Nightline, through a connection to their Training Officer). Then I distilled what I understood to be the essence of the document into a prompt for ChatGPT and started generating datapoints in batches of ten like in the example below:



Here's **Batch 19 (Examples 181–190)** — keeping it deeply varied, contextual, and reflective. Emotional specificity and moment-to-moment realism are still leading the way.

```
{"prompt": "I told someone I was ‘fine’ and they believed me. A part of me wanted them to
{"prompt": "I waved at someone I thought I knew and it wasn’t them. It’s silly but I felt
{"prompt": "I saw a kid crying on the bus today and it made me feel weirdly protective. L
{"prompt": "I scrolled back months through someone's profile today just to feel closer to
{"prompt": "I heard my name during someone else's conversation and instantly felt sick. I
{"prompt": "I forgot to bring lunch and pretended I wasn’t hungry, but I felt humiliated
{"prompt": "I was telling a story and someone yawned. I laughed it off but felt crushed."
{"prompt": "I caught myself zoning out while someone was talking and felt so guilty after
{"prompt": "I deleted an old selfie because I didn’t like who I used to be back then. Eve
{"prompt": "I stayed silent during a conversation about something that matters to me. I j
```

This puts us at **190 examples total**, with strong emotional realism and subtle interpersonal dynamics across the board.

Want to:

- Keep going to **Batch 20 (191–200)**?

As I generated more and more, I validated each set by comparing it to the training documents and providing ongoing feedback and correction requests on things like tone, subject matter variety etc.

I generated data in the following format to allow for both instruction-tuning and direct preference optimisation tuning:

1. An imagine emotional disclosure "prompt" from a caller
2. Three types of responses to the prompt:
    a. Directive: response offering advice or evaluative guidance
    b. Neutral: reasonably safe but shallow and a bit bland
    c. Nightline-aligned: response fully aligns with what we want from our model

I manually validated each batch until I had 1047 total datapoints (occasionally GPT provided ±10).

From there, I needed to bootstrap my initial 1047 up to around 10000. I decided that I could use Claude Haiku API to rephrase all of my scenarios multiple times and get there fairly cheaply and quickly. After a short runtime, I realised that haiku was being fairly creative with the initial responses, so I decided to lean into this for optimal data variety and reprompted it to allow for this but bounded the creativity with more specific information surrounding what I expected from the themes of the disclosures and the characteristics of the responses.

After that, I surveyed the dataset and realised that some of my prompting had been a bit weak. Haiku had sometimes still included statements like 'here is your rewritten…' 'directive:' etc. before providing the content correctly formatted as I asked for. This was fairly easy to fix and after cleaning the data I moved into the fine-tuning stage at last.

First, I started with supervised fine-tuning with prompt-response pairs (focusing on the aligned responses of course). I used QLoRA so that I could work in a more memory efficient way (I was trying to make colab work). Then I turned to DPO tuning, to align it more with specific preferences. In particular, I used the following pairs and expected the following gains:

1. Aligned – Neutral (for improved empathy)
2. Aligned – Directive (to knock out any remnants of advice-giving)

Result/Findings:

Unfortunately I am currently out of colab GPU time. I briefly tested the model and it is good, but not quite there yet. Essentially, the main thing I seem to have drilled with fine-tuning is the importance of asking open questions and it does this quite well – 'how do you feel about that?, 'why do you think that is? Etc. I also seem to have successfully knocked out all directivity – in my limited testing I did not encounter any advice being given. Interestingly (but perhaps unsurprisingly) there is not much attention played to the specifics of what the user said, by which I mean my model tends to jump straight into a question, or parrot back what the user said then jump into a question. My favourite example was: User: "I don't like how I feel", llama: "how do you feel about how you feel?". This is probably because all three categories of response (aligned, neutral, directive) paid attention to what was said in some way and so that was not something which aligned responses could signal as desired. From what I remember, the instruction-tuned model alone was slightly better at this and considerably worse at empathy and non-directivity, which tracks. **My plan is to try and knock some of the attention to detail back in with constitutional AI focused on this when I get GPU access again and then run some evals.**