Advanced Methods in Text Analytics
Exercise 8: Large Language Models - Part 2
Solutions

Daniel Ruffinelli

FSS 2025

1 Making Predictions with LLMs

(a) TODO: num layers, hidden sizes and vocab sizes of each of the models.

(b) TODO: what is BatchEncoding object, what is input_ids (token ids), what is attention mask, how many tokens are used to encode your prompt? len(input_ids)

(c) TODO: what is CausalLMOutputWithPast, what is logits, what is past_key_values. Logits are of size (batch size, input sequence length, vocab size)

(d) TODO: See code, explain params in batch decode, tokens make sense, difference between models? This function is useful for sampling, e.g. greedy, top k.

2 Prompting

(a) TODO: difference gpt vs llama, output of gpt is always new line, likely wants to create a sentence, so greedy on first token does not work, we would have to decode more next tokens as well, answer Paris contained in top 10 tokens for gpt, but how to use it? Top k sampling? It's a factual question, we want a deterministic answer. If we try a more ICL-like prompt, then the top token is the correct answer, but this too is brittle because of tokenization issues, e.g. Rome is often tokenized with a space, so we need to add the space to the prompt, and this is not consistent across models or examples. Try to out!

(b) TODO: ICL prompt has instruction, demonstration and question.

(c) TODO: are all tasks clear with zero shot? Which ones and which ones not? Do both models understand the tasks? What tasks become clear with demonstrations?

(d) TODO: can any task be solved with an added natural instruction?

(e) TODO: go from asking natural question to ICL in zero shot, n shot and then with instructions. I think the difference is: instruct model can do natural question (e.g. ask a capital city), and it can also do zero shot when a natural language instruction is included (e.g. do translate to french). They should be able to also generate responses in a given mood, but we do that in the next task.

3 Generating Longer Responses

(a) See code.

(b) TODO: before we saw greedy not good for llama and gpt sometimes, but now with more tokens decoded? Were they trying to talk?

(c) TODO: try different sampling methods, see code in RAW notebook.

(d) TODO: Try different lengths: short means incomplete sentences, long means model continues conversation by itself. Also try different system prompts. Do all models follow the instructions?