Probabilistic Prompting of LLMs – Summary

Francesco Tinner, Wilker Aziz

October 21, 2024

Given some instruction c, and a response x, a LM can be queried to assign log-probability $\log P_{\text{LM}}(x|c)$ to x given c.[1] The instruction c can be used, for example, to provide the LM with a task description and/or few-shot examples.

We are interested in using the LM to characterise the probability $P(O = o|T = t, R = r, S = s)$ of filling with o the object slot of a given relation r in a given template t whose subject slot is already filled with s. For that, we regard the language model as a 'slot filler' that is constrained to filling the object slot with options from a predefined set Ω. This is, of course, a design choice we make in this research (as there could be many other ways to interact with an LM, beyond slot filling). Under this interpretation, $P(O = o|T = t, R = r, S = s)$ is obtained by renormalising the probability the LM assigns to the string $t(r, s, o)$ against all strings of the kind $t(r, s, o')$,[2] with o' in the set Ω of all possible relevant objects:

$$P(O = o|T = t, R = r, S = s) \triangleq \frac{\exp(g(t(r, s, o)))}{\sum_{o' \in \Omega} \exp(g(t(r, s, o')))} , \tag{1}$$

where $g(x) = \log P_{\text{LM}}(x|c)$ and we approximate Ω by $\mathcal{O}^+ \cup \mathcal{O}^-$.[3]

Multiple templates. We design a probe that assesses the model more comprehensively by testing the model's ability to fill object slots across a diverse set $\mathcal{T}(r)$ of templates for the relation r, which we obtain via automatic paraphrasing as described in §??? Because we have no strong reason to prefer one template over the other, we combine $K = |\mathcal{T}(r)|$ conditionals of the kind introduced in

[1] We simply condition on c and process the token-sequence x with the decoder, summing the log probabilities autoregressively assigned to the tokens in x.

[2] When we fill in a template t for r with subject s and object o, we obtain a string in English, we denote that string by $t(r, s, o)$.

[3] It is important that the instruction c is held constant in the definition of the conditional distribution. In effect, we are characterising the conditional distribution $P(O|C = c, T = t, R = r, S = s)$ specific to a given instruction c.

Eq (1) under a uniform prior over templates:

$$P(O=o|R=r,S=s) \triangleq \sum_{t \in \mathcal{T}(r)} \underbrace{P(T=t|R=r)}_{=1/K} P(O=o|T=t,R=r,S=s) \; . \quad (2)$$