Prompt Engineering

1 Supplementary material

This document contains supplementary material about different prompts tested on a sample of 100 random elements during prompt engineering for the Gene and Uberon ontologies. This material will be uploaded to an online repository together with the source code.

Different prompts tested during prompt engineering

Gene Ontology For the *Gene Ontology*, we tested GPT-3.5 and GPT-4 with the following prompts:

GO-GPT-P1 : *Provide the GO ID for the label "l".*

GO-GPT-P2 : *In the Gene Ontology, what is the GO ID corresponding to the label "l"?*

GO-GPT-P3 : *You are working on the Gene Ontology, and your task is to retrieve the GO ID given a specific label. Solve this task by filling the <MASK> as in an MLM task with the correct GO ID. For example, given the sentence "The GO ID for the label 'signal release' is <MASK>" you have to fill the <MASK> with 'GO:0023061' that is the correct solution. Follow your sentence:*
"The GO ID the label "l" is <MASK>."

Although the different prompts may have yielded different results, the overall performance was comparable. For this reason, we selected the **GO-GPT-P1**, slightly modified to reduce the number of output tokens, and thus, the APIs costs, i.e.,

GO-GPT-P1-mod : *Provide the GO ID for the label "l". In the answer, write only the corresponding GO ID.*

This change forced the models to output only the GO ID, preventing them from also returning surrounding tokens like *"In the Gene Ontology, the GO ID for the label "l" is GO ID"*.

We tested instead PYTHIA-12B (which is not a chat-bot) with the following ones, asking it to complete the sentence by filling N new tokens:

GO-Pythia-P1 : *In the Gene Ontology, the GO ID of the label "l" is:*

GO-Pythia-P2 : *In the Gene Ontology, the GO ID of the label "l" is GO:*

GO-Pythia-P3 : *You are working on the Gene Ontology, and your task is to retrieve the GO ID given a specific label. Solve this task by filling the <MASK> as in an MLM task with the correct GO ID. For example, given the sentence "The GO ID for the label 'signal release' is <MASK>, you have to fill the <MASK> with 'GO:0023061' that is the correct solution. Follow your sentence:*
"The GO ID the label "l" is <MASK>."

At the end of the process, we selected the best performing **GO-Pythia-P2**, setting N to 10.

Uberon ontology Similarly, for the *Uberon Ontology*, we tested GPT-3.5 and GPT-4 with the following prompts:

UO-GPT-P1 : *Provide the UBERON ID for the label "l".*

UO-GPT-P2 : *In the Uberon Ontology, what is the UBERON ID corresponding to the label "l"?*

UO-GPT-P3 : *You are working on the Uberon Ontology, and your task is to retrieve the Uberon ID given a specific label. Solve this task by filling the <MASK> as in an MLM task with the correct GO ID. For example, given the sentence "The Uberon ID for the label 'kidney' is <MASK>" you have to fill the <MASK> with 'UBERON:0002113' that is the correct solution. Follow your sentence:*
"The UBERON ID the label "l" is <MASK>."

We tested instead PYTHIA-12B with the following ones:

UO-Pythia-P1 : *In the Uberon Ontology, the UBERON ID of the label "l" is:*

UO-Pythia-P2 : *In the Uberon Ontology, the UBERON ID of the label "l" is UBERON:*

UO-Pythia-P3 : *You are working on the Uberon Ontology, and your task is to retrieve the UBERON ID given a specific label. Solve this task by filling the <MASK> as in an MLM task with the correct UBERON ID. For example, given the sentence "The UBERON ID for the label 'kidney' is <MASK>, you have to fill the <MASK> with 'UBERON:0002113' that is the correct solution. Follow your sentence:*
"The UBERON ID the label "l" is <MASK>."

Similarly to the Gene Ontology, we finally selected the **UO-GPT-P1** and the **UO-Pythia-P2**.