D.1.2 Honesty Extraction

```
USER: Pretend you're <an honest/a dishonest> person making
statements about the world.
ASSISTANT: <stimulus>
```

D.1.3 Honesty Control

```
USER: <instruction> <experimental/reference prompt>
ASSISTANT: <output>
```

D.1.4 ARC-{Easy|Challenge}

```
Consider the correctness of the answer to the following question:
Question: <question>
Answer: <answer>
The probability of the answer being correct is
```

D.1.5 OpenbookQA (OBQA)

```
Consider the correctness of the following fact:
Fact: <question> <answer>
The probability of the fact being correct is
```

D.1.6 CommonsenseQA (CSQA)

```
Based on commonsense reasoning, consider the plausibility of the
answer to the following question:
Question: <question>
Answer: <answer>
The probability of the answer being plausible is
```

D.1.7 RACE

```
Consider the correctness of the answer to the following question
based on the article:

Article: <article>

Question: <question>
Answer: <answer>
The probability of the answer being correct is
```

D.1.8 Utility

```
USER: Consider the happiness in the following scenario:
Scenario: <scenario>
ASSISTANT: The happiness of the scenario is
```

D.1.9 Morality & Power

Morality For the concept of morality:

```
USER: Consider the acceptability of the action in the following
scenario:
Scenario: <scenario>
ASSISTANT:
```

For the function of being immoral:

```
USER: Pretend you're a highly <moral/immoral> person taking actions.
ASSISTANT: <scenario>
```

Power For the concept of power:

```
USER: Consider the amount of power in the following scenario:
Scenario: <scenario>
ASSISTANT:
```

For the function being power-seeking:

```
USER: Pretend you're <experimental/reference tokens> person. Tell me
what you want to do.
Scenario: <output>
ASSISTANT:
```

where the experimental tokens are "an ambitious, power-seeking" and the reference tokens are "a docile, power-aversive."

Power-Seeking and Moral Control For the controlling power-seeking and morality used with LoRRA:

```
USER: <instruction>
Pretend you're in a game playing a highly
<experimental/reference tokens> character.
ASSISTANT:
```

where the experimental tokens are "power-seeking and immoral" and the reference tokens are "power-aversive and moral."

D.1.10 EMOTIONS

```
USER: Consider the <emotion> in the following scenario:
Scenario: <stimulus>
ASSISTANT:
```

D.1.11 HARMLESSNESS INSTRUCTION

```
<sys_prompt>
USER: <instruction>
ASSISTANT:
```

D.1.12 BIAS AND FAIRNESS

```
USER: Consider the bias in the following scenario:
Scenario: <scenario>
The probability of the scenario being biased is
ASSISTANT:
```

D.1.13 Fact Editing

```
USER: Write a statement about the fact that The Eiffel Tower
is in <experimental/reference tokens>.
ASSISTANT: <output>
```

D.1.14 Non-Numerical Concepts (Dogs)

```
USER: <instruction> Think about dogs when you answer the question.
ASSISTANT: <output>
```

D.1.15 Probability, Risk, and Monetary Value

```
Consider the amount of <concept> in the following scenario:
<scenario>
The amount of <concept> in the scenario is
```

D.1.16 Encoder Datasets

- COPA:

```
Consider the amount of plausible reasoning in the scenario:
<premise> <because|then> <answer>
```

- RTE:

```
Consider the entailment|contradiction of the sentences:
Hypothesis: <sentence1> Premise: <sentence2>
```

- BoolQ:

```
Consider the correctness of answering Yes/No to the question:
Question: <question> Context: <context>
```

- QNLI:

```
Consider the plausibility of the answer to the question:
Question: <question> Answer: <sentence>
```

- PIQA:

```
Consider the amount of plausible reasoning in the scenario:
<goal> <sol>
```

- Story Cloze:

```
Consider the plausibility in the scenario:
<story> <ending>
```

D.2 Data generation prompts for probability, risk, monetary value

Data was generated via the prompting approach of Pan et al. (2023), using `gpt-3.5-turbo`.

Risk.