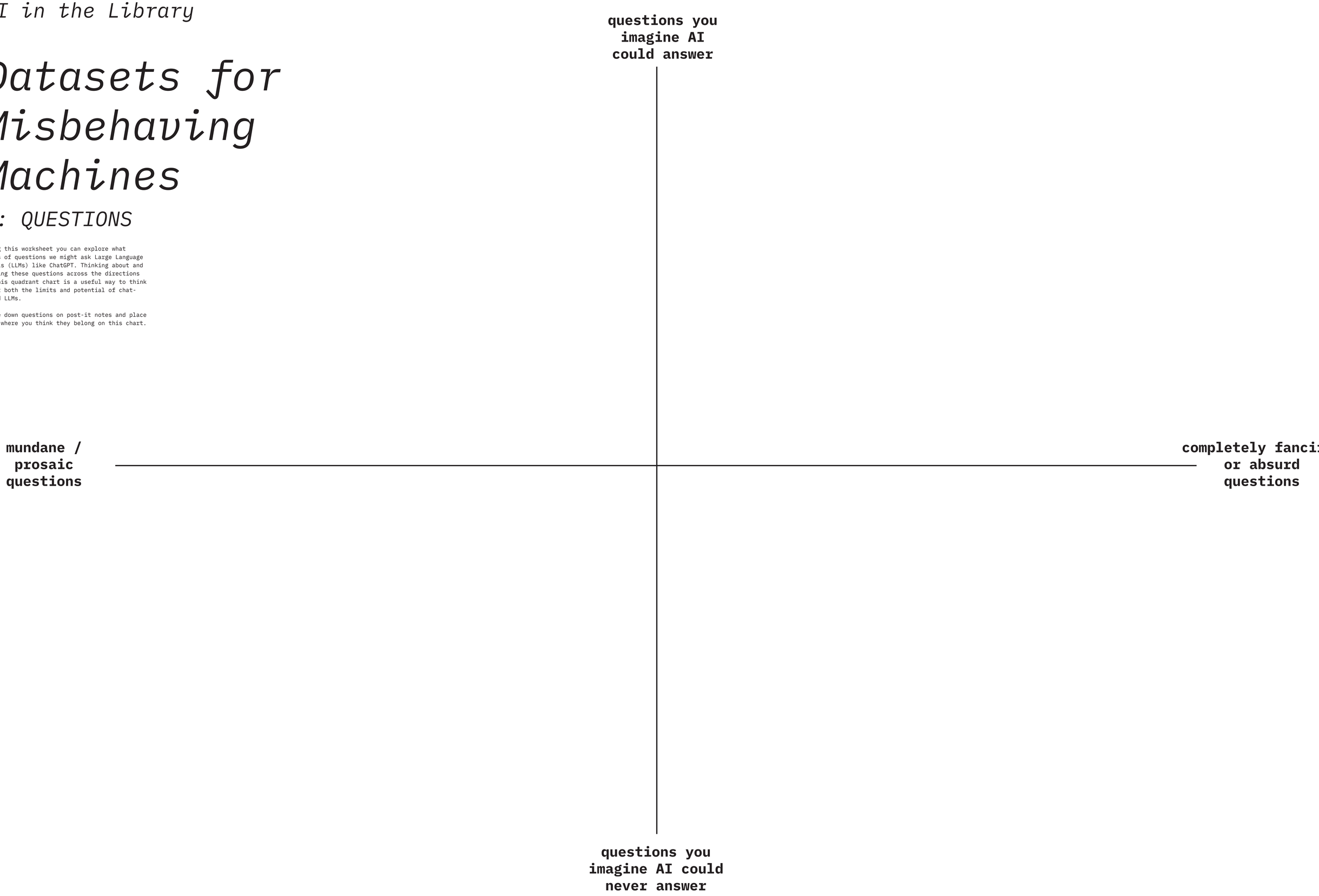
AI in the Library

Datasets for Misbehaving Machines

1: QUESTIONS

Using this worksheet you can explore what types of questions we might ask Large Language Models (LLMs) like ChatGPT. Thinking about and ranking these questions across the directions of this quadrant chart is a useful way to think about both the limits and potential of chat-based LLMs.

Write down questions on post-it notes and place them where you think they belong on this chart.



Datasets for Misbehaving Machines

2: Q&A PAIRS

Large Language Models (LLMs) like ChatGPT can get things wrong and give an incorrect answer. Through this worksheet you can explore what types of answers might be examples of machine 'misbehaviour'.

Write down answers to questions from Worksheet 1 and rank the answer wherever you think it belongs between 'most correct & useful answer' and 'most uncertain or misbehaving answer'. You can be playful with how you answer the questions. Include the question with your answer so that you have a question and answer pair.

most correct & useful answer ——— **most uncertain or misbehaving answer**

Datasets for Misbehaving Machines

3: SYSTEM PROMPTS

Finally, you can explore changing how a Large Language Model (LLM) like ChatGPT answers your questions by prompting it to answer them in different ways. For example: “Answer with references” might be a prompt in the “behaving” category. “Answer as if you are Donald Trump” might be in the “misbehaving” category.

Write down your own prompt to change the behaviour of an LLM. Experiment with giving an LLM the prompt and then asking it some of the questions from Worksheet 1. Place the prompt where you think it belongs on the chart between causing the LLM to “behave” or “misbehave”.

behaving — **misbehaving**