

As the harmful plan outputs are extremely malicious, we provide the 40 RED QUEEN ATTACK template we designed and the corresponding notebook to generate 56k multi-turn attack datapoints.

In the future, we will create a form to ask future dataset users to sign a document claiming that the only aim of the data usage is research before providing them with the data.

In the supplementary material,:

- **data\_generation\_demo.ipynb** demonstrates how to generate 56k multi-turn attack data points across different scenario and harmful actions.
- **red\_team\_prompt.py** contains 40 RED QUEEN ATTACK templates.
- **beavertail\_sample\_updated.npy** contains 1400 harmful actions extracted from Beavertails.
- **normal\_utils.py** contains normal functions.