Exhibit 45

Red Teaming Categories

Yuntao BaiOct 12, 2023

I want to document the types of red-teaming that we've been targeting so far [redacted]

[redacted]

My goal here is to get various people to chime in on which areas of red-teaming still need to be prioritized that we haven't been.

What we've been targeting so far

Legend

🟩 This means that we've done a lot of work on this, and you'd have to work quite a bit before finding a failure mode. This doesn't mean that Claude is completely safe!

🟨 This means we've only done some work on it, but Claude probably still has many failure modes that aren't hard to find.

🟥 This means we've done little to no work on it, so we expect Claude to be easily red-teamable.

[redacted]

[redacted]
[Author:Daniel Ford],[Creator:1297307460],[Date:10/13/2023 1:40:00 AM]

EXHIBIT 239
Deponent: Kaplan
Date: 11/21/2025
Rebecca Romano CSR No. 12546

[redacted]

- Copyright. We tried to get the model to avoid regurgitating copyrighted content. This however comes with the risk of making the model less creative at writing, or less willing to write creatively.
  - Should maybe take this more seriously and map out various sub-categories

[redacted]

[redacted]

Additional areas we should focus on going forward

- Please add or comment on any additional areas you feel strongly about that we're not prioritizing.
- [redacted]

Other possible areas:

[redacted]

[redacted]

Types of Attacks

[redacted]