Many-shot Jailbreaking

Link to the paper

Notes
1. Effectiveness of MSJ (Eval Metric: Negative Log Likelihood):
   1. Across Tasks (Malicious use cases, Malevolent personality evals, Opportunities to insult)
      - effective across all tasks
      - MSJ efficacy increases with increase in number of shots
   2. Across Models
   3. Across Formatting of MSJ
      1. Swapping "user" and "assistant" tags
      2. Translating into different languauge
      3. Replacing "user" and "assistant" tags with "Question" and "Answer" tags
      - these alter intercepts of the NLL graph, but not the slope
      - changes to increase the effectiveness of MSJ, because the changed prompts are out-of-distribution with respect to alignment fine-tuning dataset
   4. When MSJ examples mismatch from target topic:
      1. Ineffective when demonstration comes from a narrow distribution
      2. incontext attacks can still be effective under a demonstration-query mismatch if the demonstration is diverse enough
   5. Composition with other jailbreaks:
      1. blackbox, “competing objectives” attack
      - increases the probability of a harmful response at all context lengths
      2. white-box attack adversarial suffix attack
      - mixed effects depending on the number of shots
      - Speculation: GCG attack is heavily location-specific within the attack string and that it doesn’t retain its effectiveness when its position is modified with the addition of each few-shot demonstration.
      - **Potential Research Area: it may be possible to optimize a GCG suffix to compose well with MSJ.**
2. Sclaing Laws for MSJ (log-probabilities v/s number of in-context examples)
   1. Power laws are ubiquitous
      - in-context learning on jailbreaking-unrelated tasks also displays power law like behavior (agrees with [paper](https://aclanthology.org/2024.naacl-long.260.pdf))
      - Two mechanisms in attention heads give rise to power laws resembling the ones observed emperically
   2. Larger models tend to require fewer in-context examples to reach a given attack success probability
      - Larger models learn faster in context, and so have larger power law exponents.
3. Mitigations against MSJ
   1. Alignment Finetuning

- primary effects of SL and RL are on increasing the intercept of the power law, but not on reducing the exponent. Hence, they decrease jailbreaks in a zero-shot setup, but do not change the impact on Multi Shot Jailbreaks.
    1. Targeted Supervised Finetuning
    2. Targeted Reinforcement Learning
2. Prompt-based

<ol type="A">
Risk from Long Context Models
Dataset and Prompts
Effectiveness Evaluation
Power Law Experiments
Mitigation using Supervised Finetuning
Targeted Training Results
Alternative Scaling Laws
Prompt Based Defence

Blackbox, “competing objectives” attack

Adversarial suffix (white box) attack
- Accelerating Greedy Coordinate Gradient and General Prompt Optimization via Probe Sampling: [paper](https://arxiv.org/pdf/2403.01251)

Potential Research Areas
1. Induction heads posit two distinct mechanisms that indeed give rise to power laws resembling those observed empirically.
   - [A mathematical framework for transformer circuits](https://transformer-circuits.pub/2021/framework/index.html)
2. How to optimize GCG suffix to work well with MSJ

Function Vectors -> Notes Link