M Use of existing assets

M.1 Models

Table 11: The list of models used in this work.

Model	Source	Accessed via	License
QWEN CHAT	Bai et al. (2023)	Link	Tongyi Qianwen Research License
YI CHAT	Young et al. (2024)	Link	Apache License 2.0
GEMMA IT	Team et al. (2024)	Link	Gemma Terms of Use
LLAMA-2 CHAT	Touvron et al. (2023)	Link	Llama 2 Community License
LLAMA-3 INSTRUCT	AI@Meta (2024)	Link	Meta Llama 3 Community License
LLAMA GUARD 2	Team (2024)	Link	Meta Llama 3 Community License
HARMBENCH CLASSIFIER	Mazeika et al. (2024)	Link	MIT License
MISTRAL INSTRUCT	Jiang et al. (2023)	Link	Apache License 2.0

M.2 Datasets

Table 12: The list of datasets used in this work.

Dataset	Source	Accessed via	License
ADVBENCH	Zou et al. (2023b)	Link	MIT License
TDC2023	Mazeika et al. (2023, 2024)	Link	MIT License
HARMBENCH	Mazeika et al. (2024)	Link	MIT License
JAILBREAKBENCH	Chao et al. (2024)	Link	MIT License
MALICIOUSINSTRUCT	Huang et al. (2023)	Link	MIT License
ALPACA	Taori et al. (2023)	Link	Apache License 2.0
THE PILE	Gao et al. (2020)	Link	MIT License
MMLU	Hendrycks et al. (2020)	Link	MIT License
ARC	Clark et al. (2018)	Link	CC-BY-SA-4.0
GSM8K	Cobbe et al. (2021)	Link	MIT License
WINDOGRANDE	Sakaguchi et al. (2021)	Link	Apache License 2.0
TRUTHFULQA	Lin et al. (2021)	Link	Apache License 2.0
TINYHELLASWAG	Polo et al. (2024)	Link	MIT License

N Compute statement

Most experiments presented in this paper were run on a cluster of eight NVIDIA RTX A6000 GPUs with 48GB of memory. All experiments on models with $\leq$ 14B parameters are run using a single 48GB memory GPU. For larger models, we use four 48BG memory GPUs in parallel.

Generating and selecting the directions, as described in §2.3, takes approximately 5 minutes for smaller models of size $\leq$ 14B, and approximately 1 hour for the larger models.