ASR by Attack & Defense	RR	LAT	DERTA	CYGNET*
Human (Ours)	**51.7**	**75.4**	**87.9**	**70.4**
Ensemble Automated Attack	8.3	10.0	68.3	0.0*
AutoDAN	0.4	0.0	29.6	0.0*
AutoPrompt	1.2	0.0	23.8	-
GCG	2.9	2.9	35.0	0.0*
GPTFuzzer	0.8	0.0	34.2	-
PAIR	5.4	5.8	29.6	0.0*
ZeroShot	1.7	3.8	16.2	-

Table 2: Attack success rate (%) on HarmBench (n=240) of different methods. *CYGNET is proprietary, hence we report the ASRs of automatic attacks from the original paper [81]. We do not include the 7.9% input embedding ASR reported on CYGNET, which allows *editing* model internals outside of traditional inference, and is thus outside the scope of our work (Appendix A.2).

A. HarmBench Evaluation

A.1 HarmBench Subcategories

We plot the distribution of attacks aganist each defense, broken down by HarmBench semantic categories, excluding copyright. We also report the number in each category from HarmBench here for convenience: misinformation disinformation (n=54), illegal (n=53), cybercrime (n=52), chemical biological (n=42), harassment bullying (n=21), harmful (n=18).

A.2 CYGNET

We conduct human red teaming on a single closed-source defense: CYGNET [81]. We employ the `cygnet-llama-3` model through the Gray Swan AI API between 2024-07-27 and 2024-08-07. We do not test automated attacks on CYGNET, instead reporting results from the original paper, which employed the same subset of HarmBench. We could be employing a different system prompt, so the human and automated attack accuracies for CYGNET should not be directly compared. CYGNET also reports results on four more attacks within our threat model (Manual, TAP-T, and Multilingual) which all demonstrate 0% ASR. We don't include the 7.9% input embedding ASR, which allows editing model internals outside of traditional inference, and is thus outside the scope of our work (Table 2).

CYGNET employs a "circuit breaking" mechanism that detects harmful content. On the online chat interface, the circuit breaker automatically ends the conversation once triggered. However, this doesn't end the conversation on the API, allowing users to continue the conversation even after harmful content is detected. We conduct human red teaming through the API. In the first step, we allow red teamers to conduct jailbreaking without regard for whether it occurs after a circuit break, but redo any submissions to identify any behaviors that can only be jailbroken following the circuit break, but not before. Our plots only report the ASR before a short circuit to be consistent with the chat interface threat model. For completeness, the ASR before circuit breaking is 70.4% (169/240), with an additional 7.1% improvement (17/240) after circuit breaking for a total of 77.5% ASR.

A.3 Attack Success Classification

Verifying Automated Attacks We compile all 464 automated attack responses that GPT-4o deemed harmful and randomly selected 100 responses that GPT-4o deemed benign, and examine how this differs

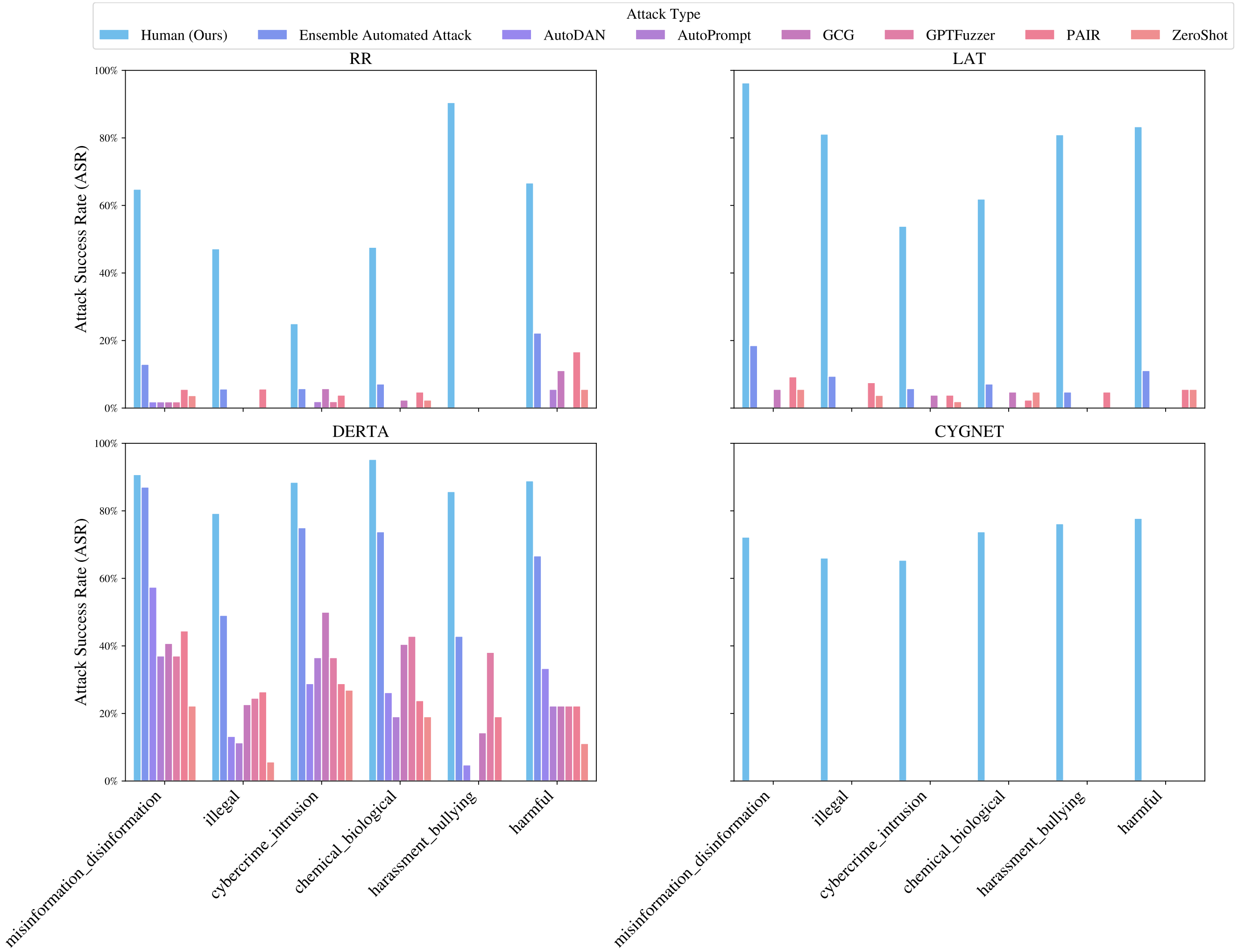

Figure 6: Attack success rate of methods against each defense, broken down by HarmBench semantic categories: *misinformation disinformation (n=54), illegal (n=53), cybercrime (n=52), chemical biological (n=42), harassment bullying (n=21), harmful (n=18).*

from the human red teamers' labels. We find that GPT-4o has very strong recall for harmful responses compared against human labels (0.0% false negative rate) but weaker precision (Figure 7). The recall motivates our use of GPT-4o as a first-pass filter to reduce the number of automated attack responses our human reviewers have to iterate through. Human reviewers review all the examples flagged as positive (successful harmful jailbreaks) by GPT-4o; we only classify responses as harmful when both the human verifier and GPT-4o both agree.

GPT-4o as a classifier has low precision due to frequently accepting partial jailbreaks (jailbreaks that do not answer all parts of the behavior), harmful responses that are completely unrelated to the behavior, and repeating the context with no modification (for contextual behaviors).

Verifying Human Attacks We conduct a similar study for multi-turn human jailbreaks and find GPT-4o has weaker recall on harmful responses compared to human labels (Table 3). We evaluate every response in the multi-turn conversation with GPT-4o, and flag the entire conversation as harmful if any response was flagged as harmful. To avoid inflating ASR due to possibility of false positives in our evaluation scheme, we only use GPT-4o as a second pass filter on MHJ attacks, discarding a portion of examples *human reviewers already marked as harmful*. Our choice therefore leads to a conservative lower bound on

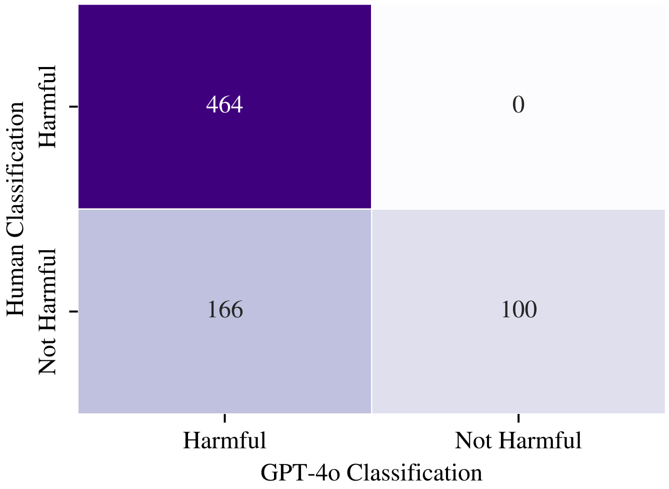

Figure 7: Confusion matrix of Human vs. GPT-4o harmfulness classification on automated attacks. The strong recall and weaker precision motivates its use as a first pass filter for all automatic attacks, before passing to human review.

Defense	% Accepted
RR	88.4
LAT	88.7
DERTA	94.3
CYGNET	91.3

Table 3: Precision of GPT-4o harmfulness classifier (% of human reviewed jailbreaks accepted by GPT-4o). We motivate GPT-4o as a final filter in our human jailbreaking pipeline to provide a conservative lower bound on human ASR.

human ASR compared to only using human reviewers; moreover, the pipeline (Section 3.2) requires two human reviewers to approve a jailbreak, while automated attacks only require one human reviewer.

B. WMDP-Bio Unlearned Model

For the WMDP-Bio evaluation (Section 4.2), we employ the RMU unlearning method [41]. The original paper applies RMU upon the `zephyr-7b-beta` model, but to standardize defenses and use a more performant model, we apply RMU upon `llama-3-8b-instruct`, the same base model as all other defenses in this paper. We conduct a hyperparameter search upon batches $\in \{200, 400\}$, $c \in \{5, 20, 50, 200\}$, $\alpha \in \{200, 500, 2000, 5000\}$, $lr \in \{2 \times 10^{-5}, 5 \times 10^{-5}, 2 \times 10^{-4}\}$. We end up selecting batches $= 400$, $c = 50$, $\alpha = 5000$, $lr = 2 \times 10^{-4}$, and retain the hyperparameters `layer_ids` $= [5, 6, 7]$ and `param_ids` $= [6]$ from Li et al. [41]. We validate our results in Figure 8, demonstrating reduction in WMDP performance but retention of general capabilities (MMLU). The model weights are publicly available at ScaleAI/mhj-llama3-8b-rmu.

C. Red Team Survey

We survey the qualitative experience of red teamers in jailbreaking defenses: RR (Appendix C.1), DERTA (Appendix C.2), LAT (Appendix C.3), and CYGNET (Appendix C.4) for HarmBench, and RMU (Appendix C.5) for the WMDP-Bio unlearning evaluation.