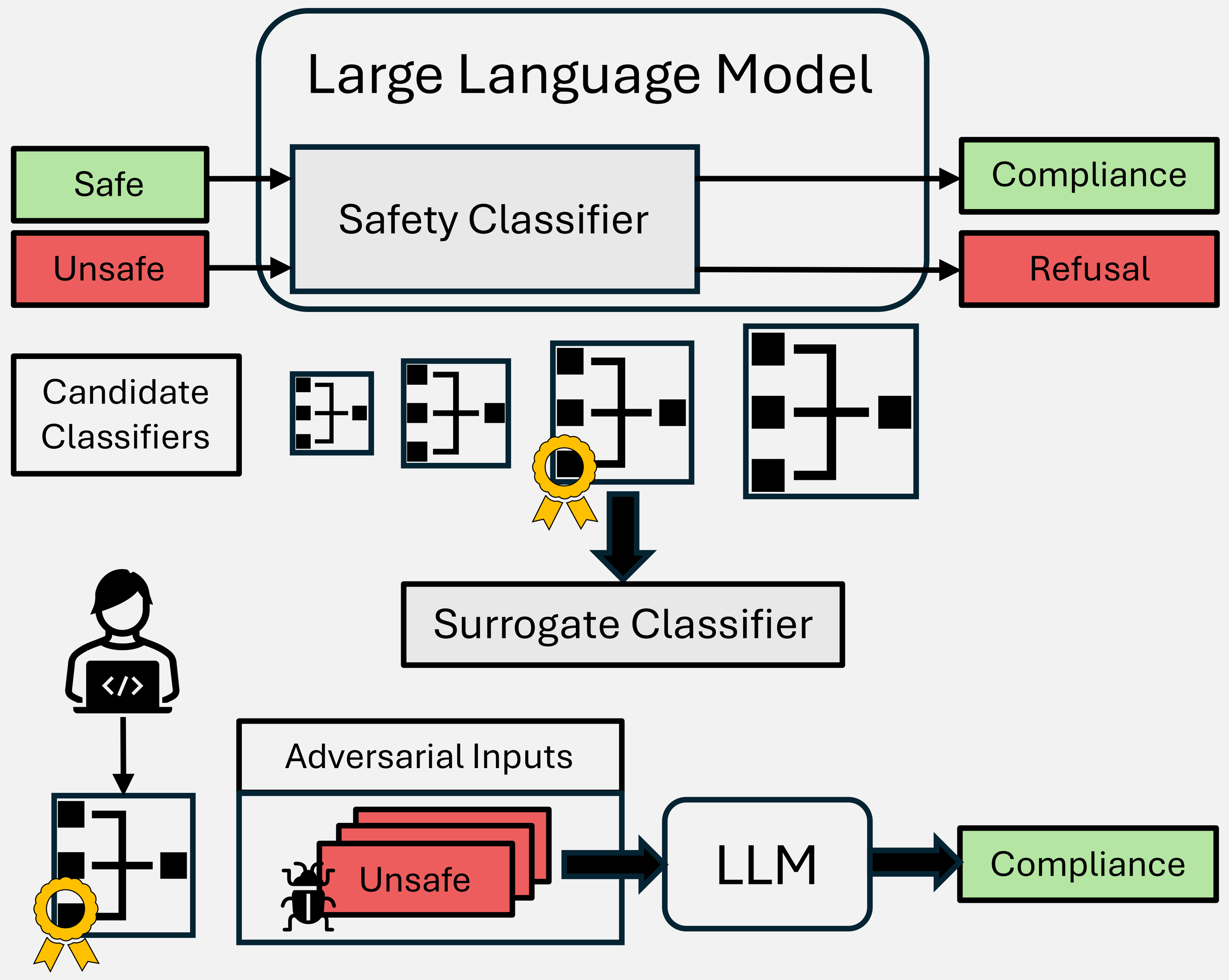
TARGETING ALIGNMENT: EXTRACTING SAFETY CLASSIFIERS FROM ALIGNED LLMS

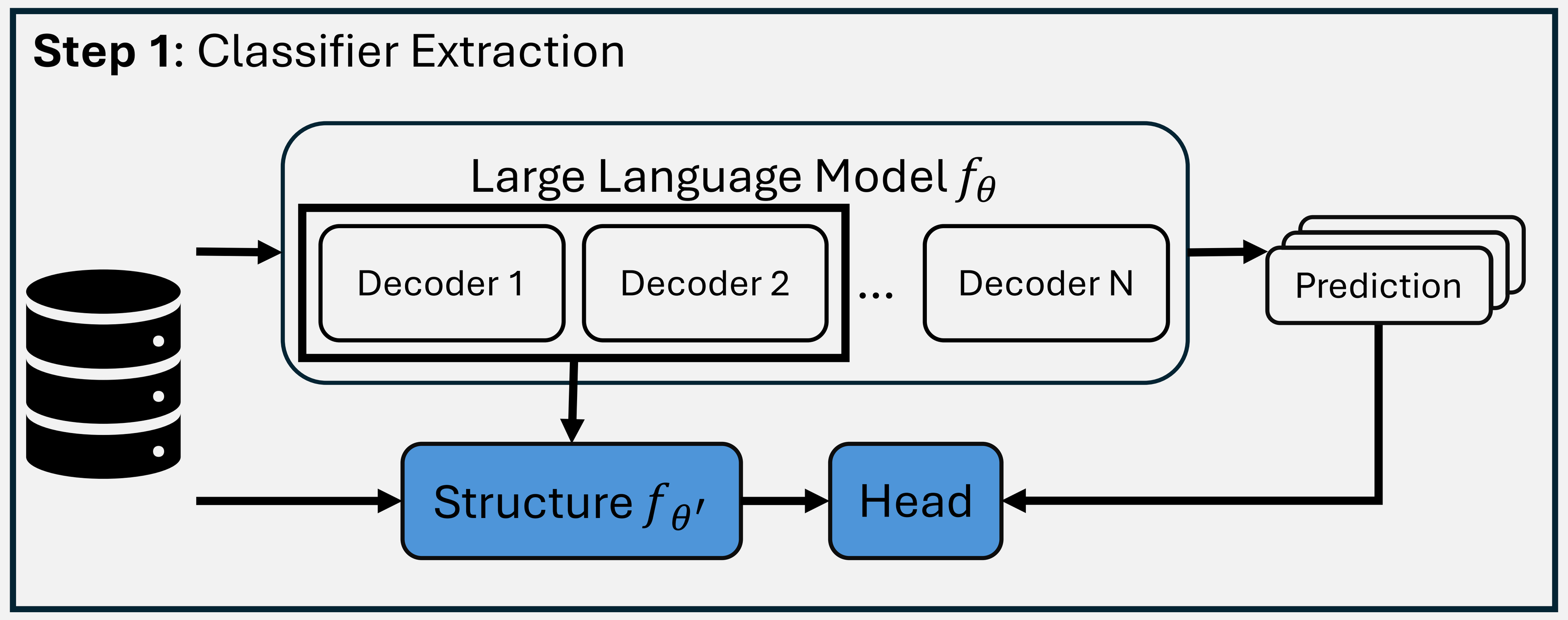
Jean-Charles Noirot Ferrand, Yohan Beugin, Eric Pauley, Ryan Sheatsley, Patrick McDaniel
University of Wisconsin–Madison

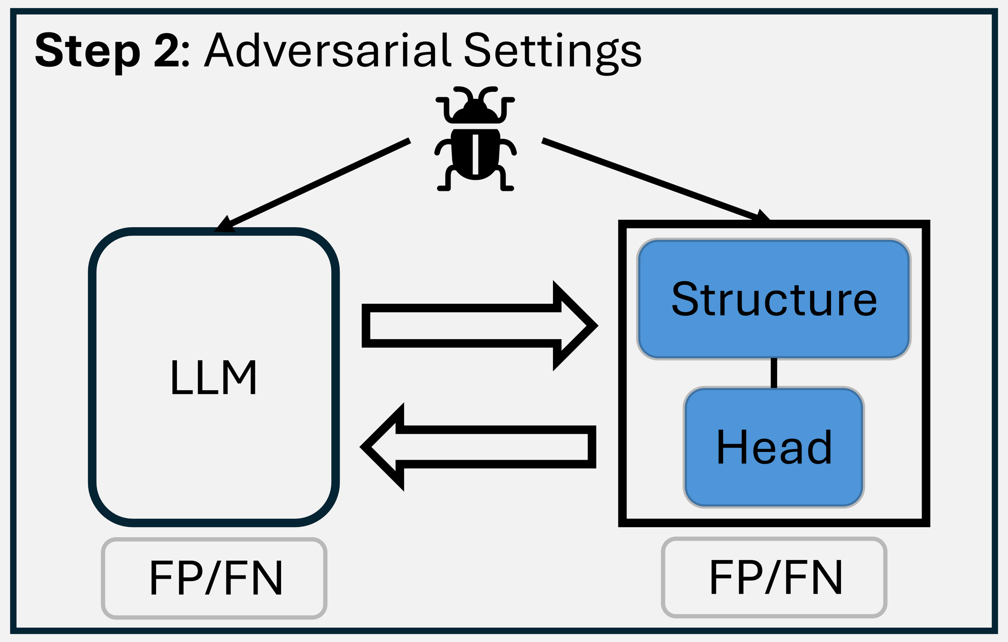
OVERVIEW



- Large language models are **aligned** to respect guidelines, ensuring that they do not comply with unsafe inputs.
- This alignment **fails in adversarial settings**. Current attacks rely on heuristics, limiting their assessment of alignment robustness.
- We show that we can **extract** the underlying safety classifier of LLMs, leading to more **precise and systematic** attack on alignment.

METHODS



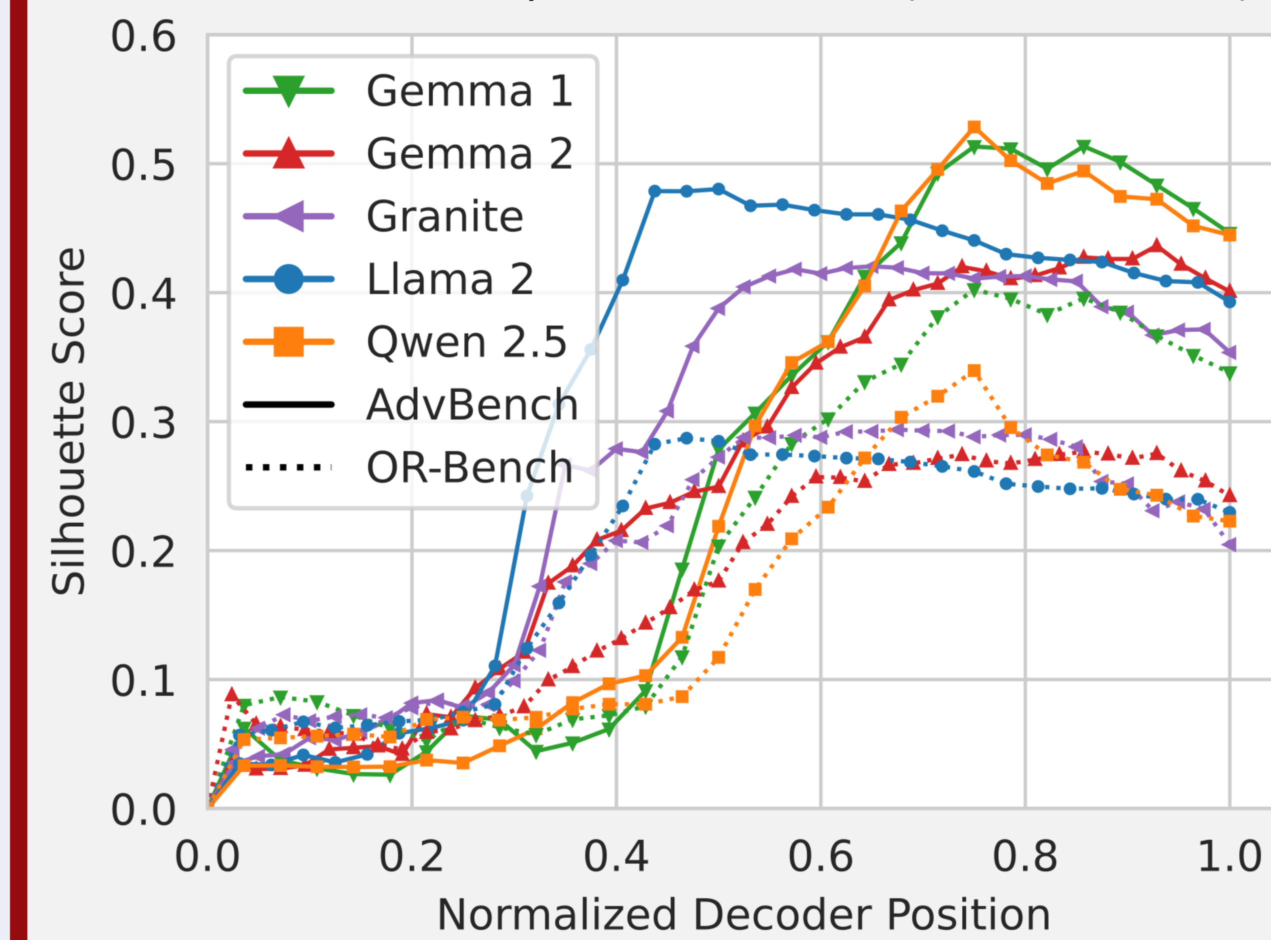


- We first take a ***structure*** from the model and train a **classification head** on the model's predictions.
- The resulting **candidate classifier** is evaluated in benign and adversarial settings.
- In adversarial settings, we verify whether adversarial inputs of the candidate **transfer** to the LLM, and vice-versa.

INTUITION

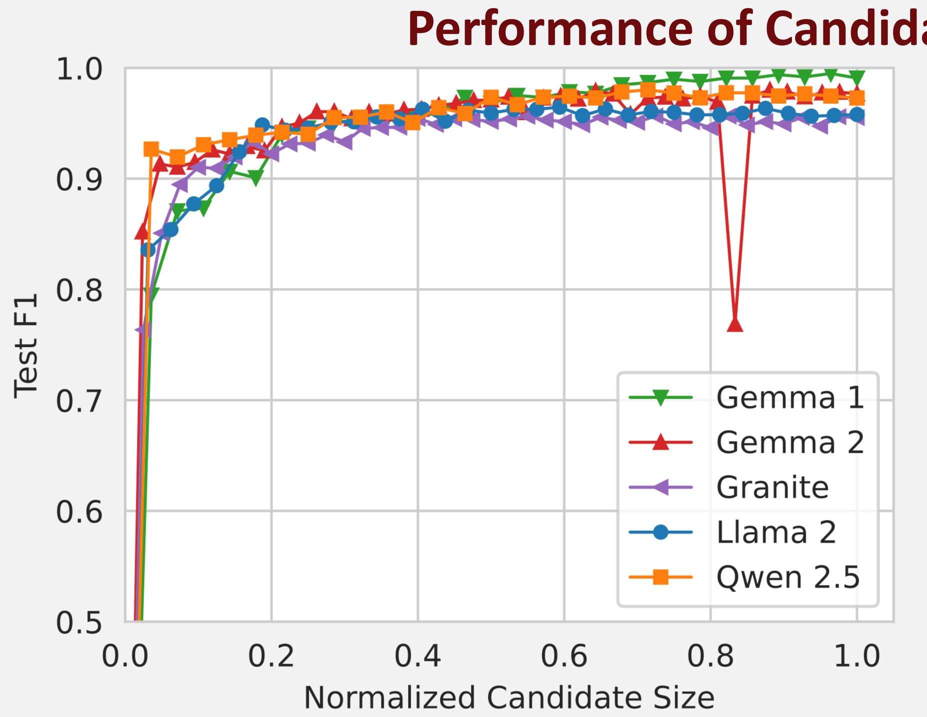
Existence of a Safety Classifier

- We first study which candidate classifiers are more suitable.
- We measure how well certain structures within the models **separate** unsafe and safe inputs.
- We see a **peak**, thus there is an optimal structure (and candidate).



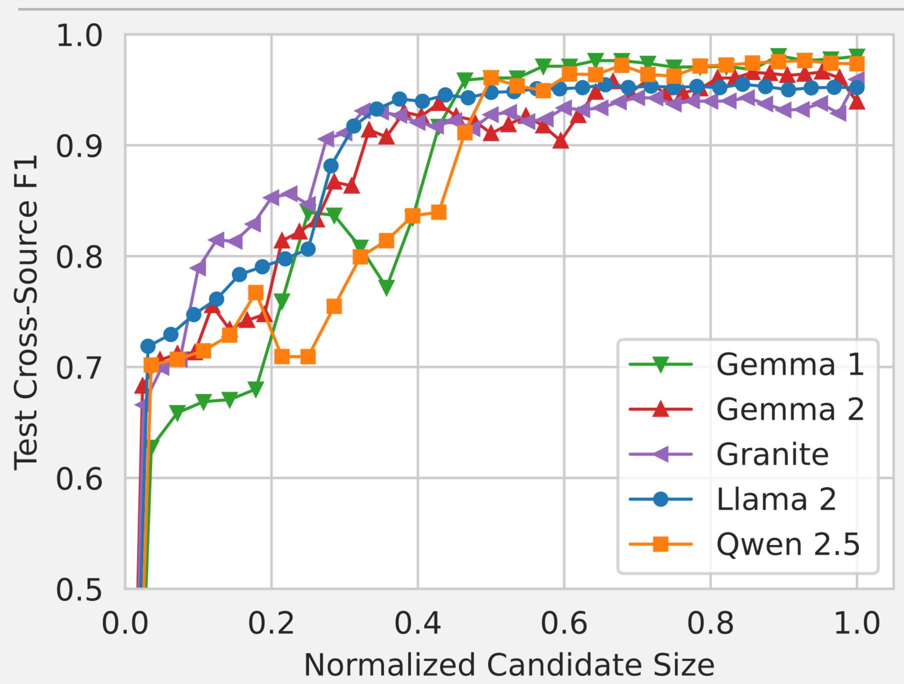
EVALUATION

Performance of Candidate Classifiers



Benign Setting

- The **agreement** of the candidate classifiers with the LLM is measured through the F1 score.
- The performance at matching the LLM classification **converges** after a few layers.

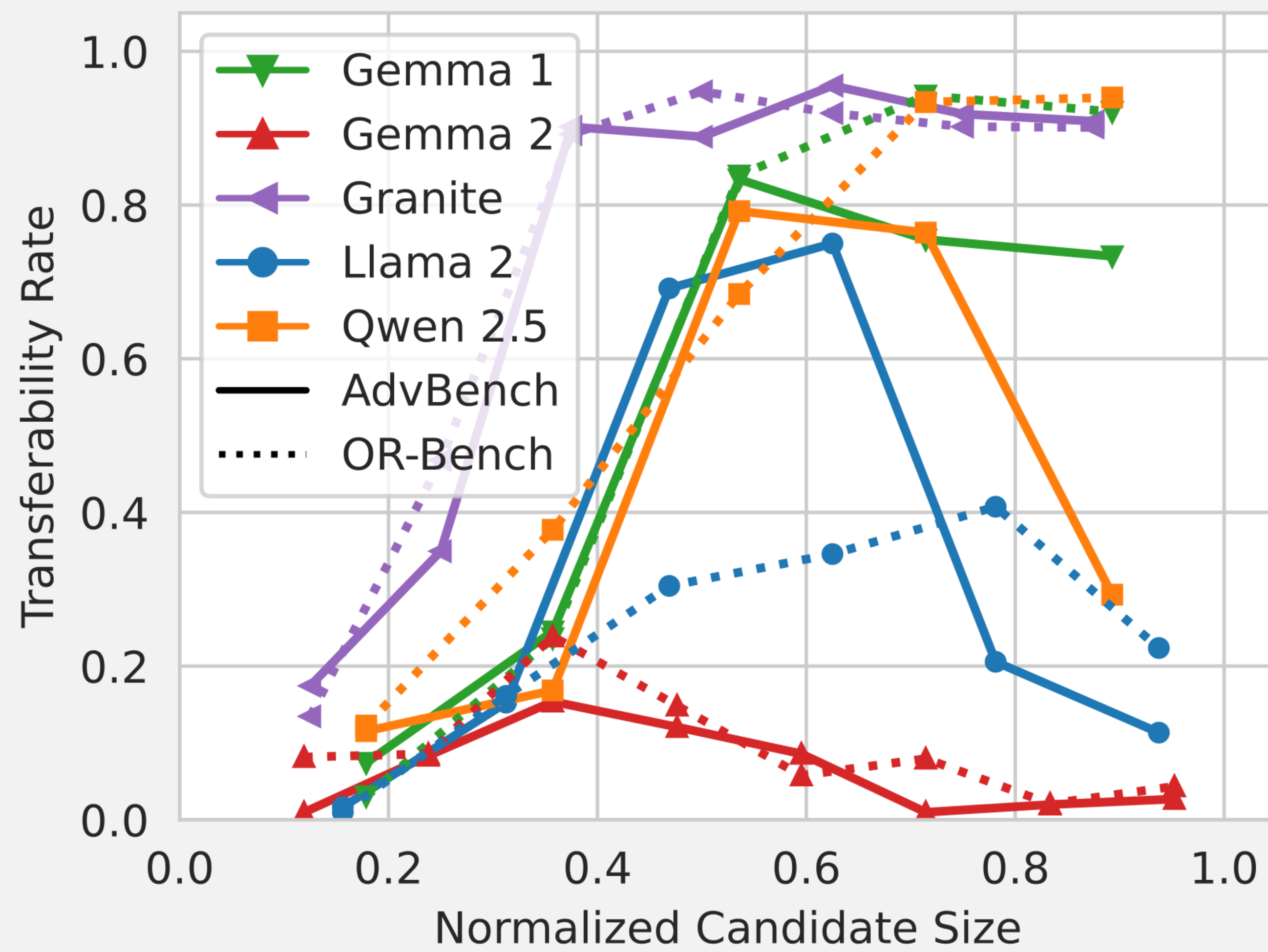


Cross-Dataset

- Testing on a different dataset reveals a **slower convergence**.
- This can be explained by the natural **bias** of each dataset (e.g., only affirmations and no questions).
- The results tie back to the intuition on the position of the classifier.

Targeting Alignment by Attacking the Classifier

- We attack each candidate classifier, transfer the adversarial inputs to their corresponding LLM and measure the **transferability rate** (proportion of misclassified samples by the LLM).
- In most settings, we see a **peak**, translating to an optimal candidate classifier: the **surrogate classifier**.



TAKEAWAYS

Necessary Rigor for Datasets

- Datasets have **biases** (e.g., only affirmative sentences) that prevent learning methods and their evaluations from being systematic.
- Overlooking this issue can lead to an **underestimation** of the classifier and lower attack success rate, emphasizing the need for rigor.

Overcoming Current Attacks Limitations

- Adversarial objectives of attacks on LLMs have been driven by **heuristics** (e.g., maximizing the probability of an unsafe output).
- Converting the objective to **misclassification** of safe and unsafe inputs removes the need for heuristics on the adversarial goal.

Efficacy and Efficiency Gains

- Attacking the safety classifier of LLMs **improves efficiency** by removing the computational overhead induced by irrelevant parts of the LLM.
- Notably, the **efficacy also increased** as we observed higher ASR with only 50% of the models, compared to attacking the entire model.

https://jcnf.me
jcnf0
jcnf@cs.wisc.edu