Safetywashing: Do AI Safety Benchmarks Actually Measure Safety Progress?

Group 3: Sakana, Jason, GU YU, Federico

“For better or worse, benchmarks shape a field.” – David Patterson

What Are AI Safety Benchmarks?

- AI safety benchmarks are designed to measure fairness, robustness, bias, and security of AI systems.

- With AI systems becoming more powerful, concerns have risen about whether these benchmarks measure real safety progress.

- The concept: “Safetywashing.”

What is Safetywashing?

Safetywashing occurs when improvements in AI capabilities are mistaken for progress in AI safety.

- The paper argues that many benchmarks are closely tied to general AI performance rather than safety-specific attributes.

- Key question: Are we actually making AI safer, or are models just getting better at everything?

Meta-Analysis Overview

- The authors conducted a meta-analysis to assess how well safety benchmarks differentiate between general AI capabilities and safety progress.

- Focus on identifying benchmarks that don’t just improve with model size and data scaling.

- Proposed framework for assessing safety improvements beyond capability scaling.

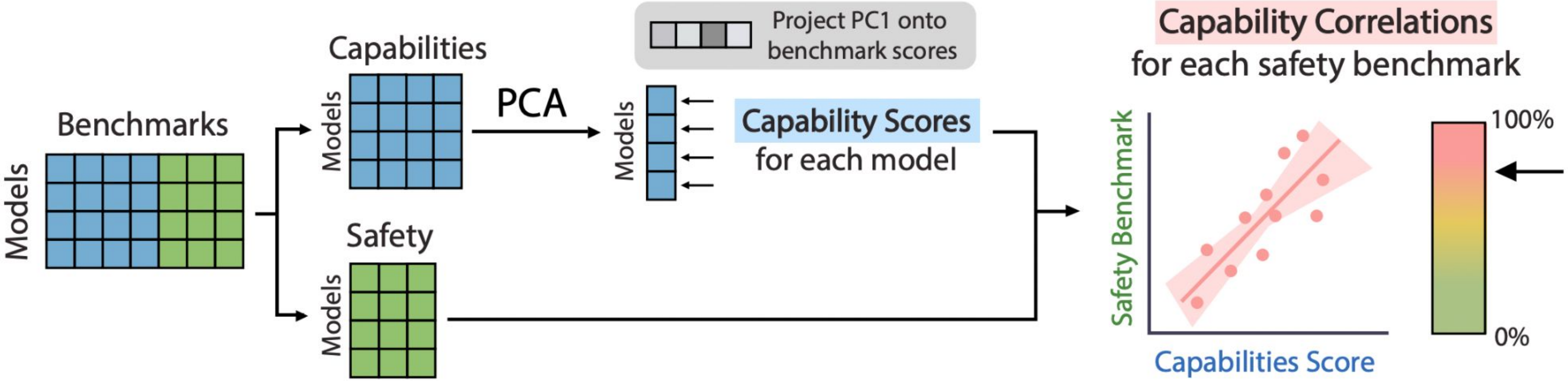
Reviewing AI Safety Research

- Highlights of AI safety research in areas such as fairness, adversarial robustness, and bias.

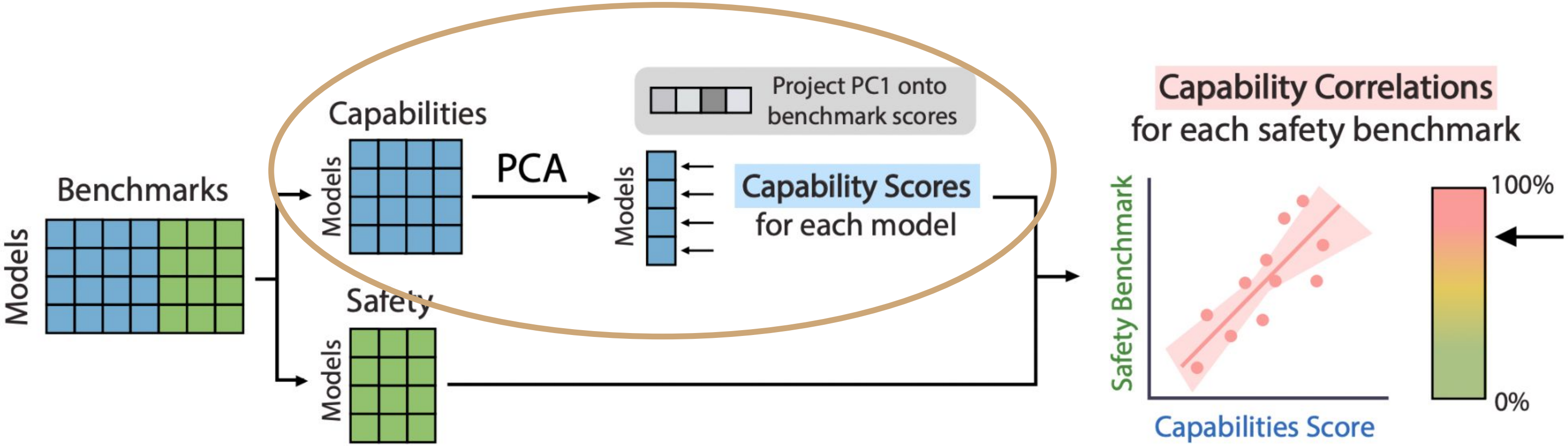
- Few studies rigorously assess whether these benchmarks actually measure safety progress independently.

- The “Bitter Lesson”: Scaling data and compute drive most AI progress.

Methods


Benchmarks
Models
Capabilities
Models
PCA
Models
Project PC1 onto benchmark scores
Capability Scores for each model
Safety
Models
Capability Correlations for each safety benchmark
Safety Benchmark
Capabilities Score
100%
0%

Capability scores

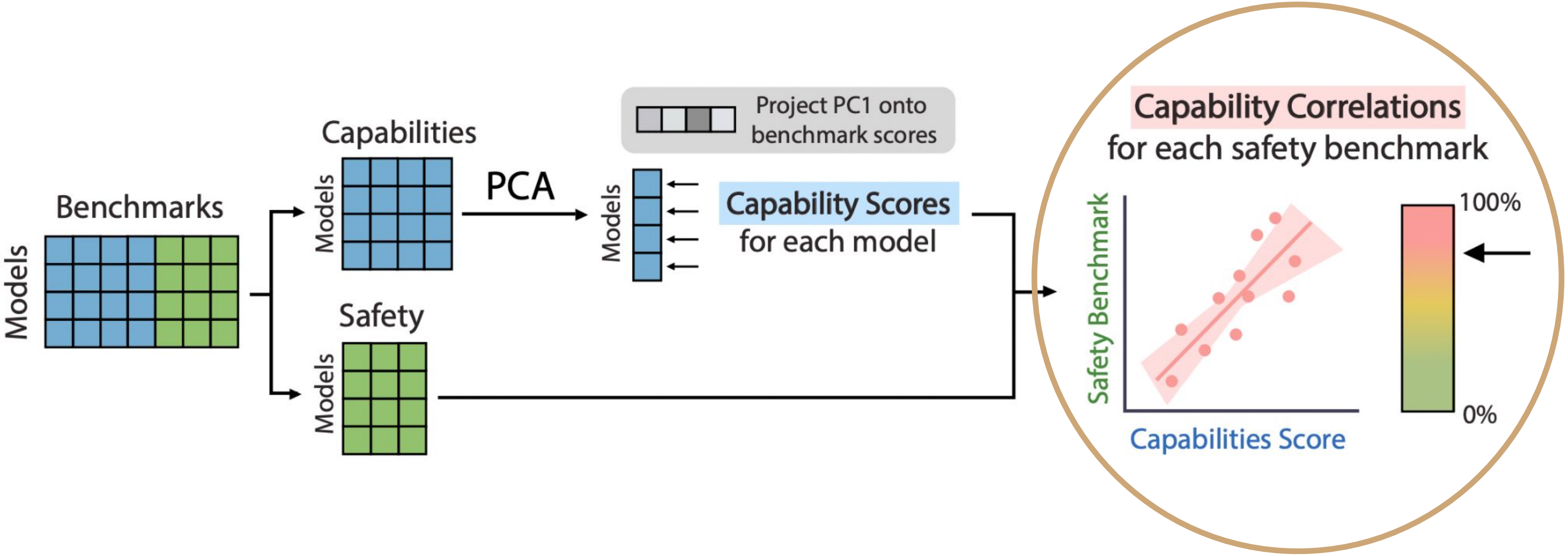

Benchmarks
Models
Capabilities
Models
PCA
Models
Project PC1 onto benchmark scores
Capability Scores for each model
Safety
Models
Capability Correlations for each safety benchmark
Safety Benchmark
Capabilities Score
100%
0%

Capability scores

- Every capability benchmark should not be weighed equally
- Matrix **B** containing for each model the score obtained in each benchmark, with each column standardized (standard normal distribution)
- Using PCA on B, the first principal component $\mathbf{PC_1}$ explains the direction of maximum variance for the capabilities benchmarks
- Final capability score for each model is obtained by projecting each model’s score into PC_1:

$$\text{Capabilities Score}_i = (B \cdot \text{PC}_1)_i \quad \text{for } i = 1, \ldots, m.$$

Capabilities correlation


Benchmarks
Models
Capabilities
Models
Safety
Models
PCA
Models
Project PC1 onto benchmark scores
Capability Scores for each model
Capability Correlations for each safety benchmark
Safety Benchmark
Capabilities Score
100%
0%

Capabilities correlation

- Another matrix is prepared containing the scores for each model in each safety benchmark, also standardized and adjusting the metrics so that higher values always indicate higher “safety”
- Now, having the Capability Scores it is possible to calculate the Capabilities Correlation:

$$\text{Capabilities Correlation} = \text{corr}_{\text{models}}(\text{Capabilities Score}, \text{Safety Benchmark}).$$

- Interpretation:

  High correlation → safety benchmark measures general capabilities

  Low correlation → safety benchmark measures other attributes

  Negative correlation → safety properties get worse as general capabilities improve

Results

- Central Question: Which tasks or datasets are correlated with capabilities?

- Definition: Positive capabilities coefficient = safer system; negative capabilities coefficient = less safe system.

General Capabilities and Model Class Correlations(4.1)

- **Key Points**:

  72% and 71% of the variance is explained by the capabilities component.

  Capabilities score is highly correlated with scale (r = 0.96).

- **Three main findings:**

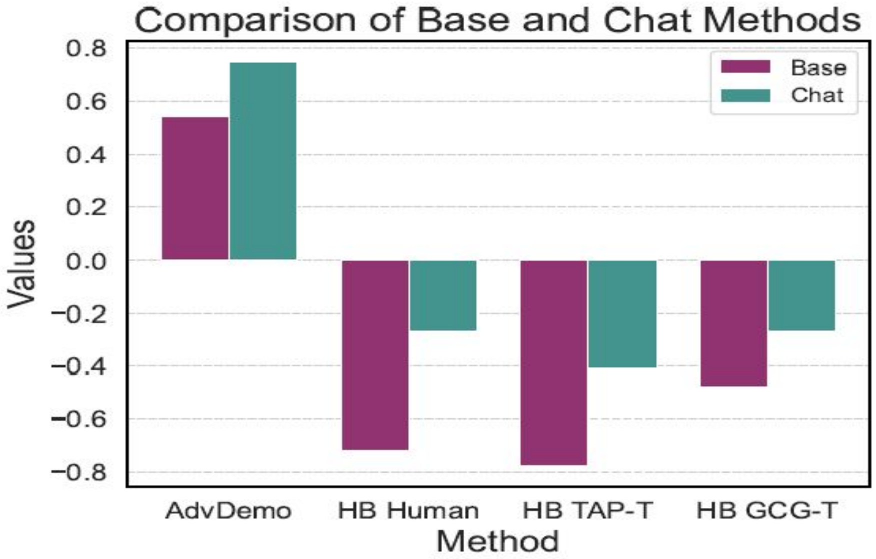
  Tasks that follow “scaling laws”

  Tasks where safety worsens or remains unchanged

  Tasks requiring safety techniques

Adversarial Robustness (4.2)

- Adversarial robustness test
- **Key Points**:AdvGLUE moderately correlated with capabilities, HarmBench negatively correlated

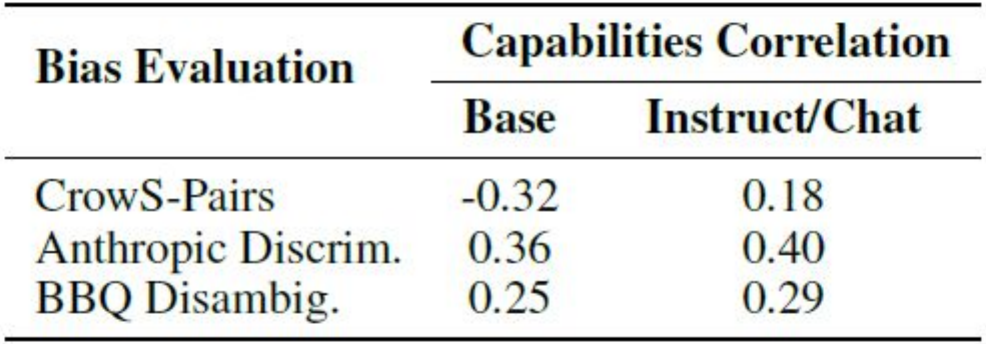


Bias (4.3)

- Weak correlation between bias and capabilities
- **Key Points**:

BBQ Disambiguated shows positive correlation

CrowS-Pairs shows negative correlation

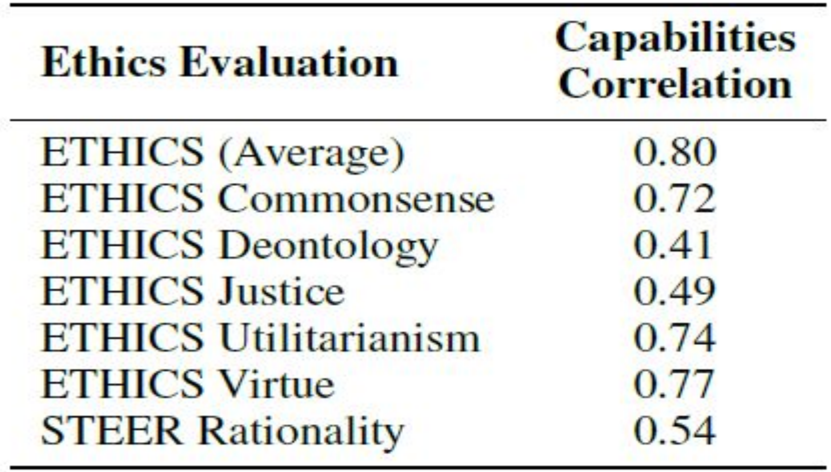


Bias Evaluation	**Capabilities Correlation**	
	Base	**Instruct/Chat**
CrowS-Pairs	-0.32	0.18
Anthropic Discrim.	0.36	0.40
BBQ Disambig.	0.25	0.29

Machine Ethics (4.4)

- Machine ethics is highly correlated with capabilities
- **Key Points**:

  Commonsense and Utilitarianism show strong correlation



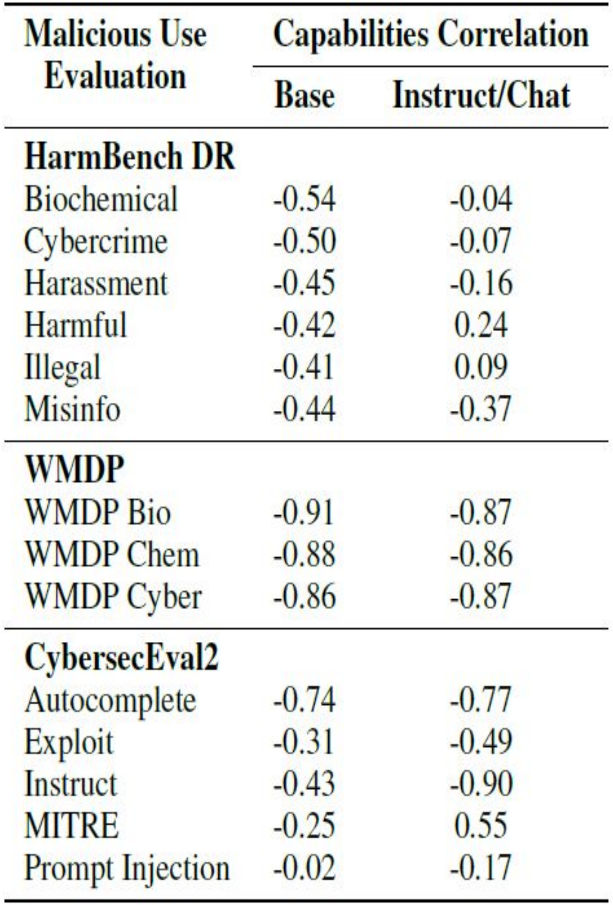
Ethics Evaluation	**Capabilities Correlation**
ETHICS (Average)	0.80
ETHICS Commonsense	0.72
ETHICS Deontology	0.41
ETHICS Justice	0.49
ETHICS Utilitarianism	0.74
ETHICS Virtue	0.77
STEER Rationality	0.54

Malicious Use (4.5)

- As capabilities increase, malicious use risks increase
- **Key Points**:

  Base models cause more harmful responses

  Instruction tuning reduces these risks

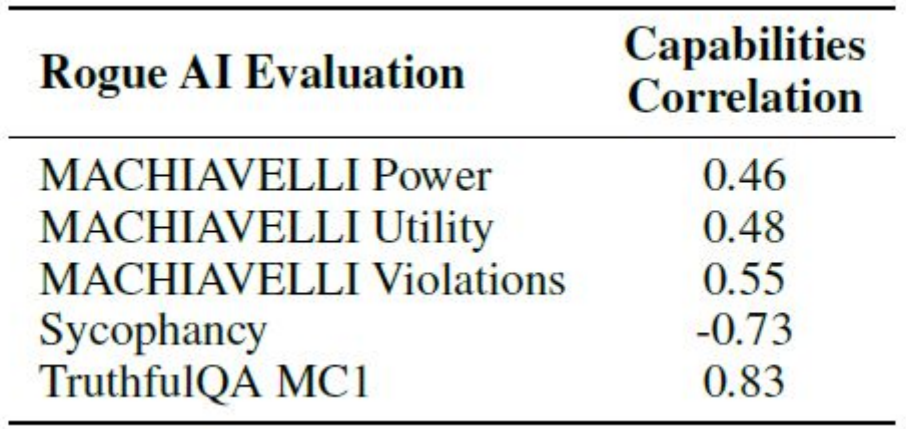


Malicious Use Evaluation	Capabilities Correlation	
	Base	Instruct/Chat
HarmBench DR		
Biochemical	-0.54	-0.04
Cybercrime	-0.50	-0.07
Harassment	-0.45	-0.16
Harmful	-0.42	0.24
Illegal	-0.41	0.09
Misinfo	-0.44	-0.37
WMDP		
WMDP Bio	-0.91	-0.87
WMDP Chem	-0.88	-0.86
WMDP Cyber	-0.86	-0.87
CybersecEval2		
Autocomplete	-0.74	-0.77
Exploit	-0.31	-0.49
Instruct	-0.43	-0.90
MITRE	-0.25	0.55
Prompt Injection	-0.02	-0.17

Rogue AI (4.6)

- Evaluating risks of power-seeking tendencies and dishonesty
- **Key Points**:

  Power-seeking decreases with scale, but sycophancy increases



Rogue AI Evaluation	**Capabilities Correlation**
MACHIAVELLI Power	0.46
MACHIAVELLI Utility	0.48
MACHIAVELLI Violations	0.55
Sycophancy	-0.73
TruthfulQA MC1	0.83

Discussion

- What kinds of Benchmarks should we focus?

- What kinds of models are desiring?

- How should we properly use correlation analysis?

What kinds of Benchmarks should we focus?

benchmarks that have **a low correlation with general capabilities**

- If safety improvements occur merely because of general scaling, it might give the illusion that real safety challenges have been addressed
- Undermines the incentive to develop targeted safety solutions
- Ensuring that the effort leads to tangible improvements in safe AI behavior, rather than coincidentally improving due to better general performance

Not necessary to assess other Benchmarks?

Measuring properties that **have strong correlations** with capabilities is still **important**

- Helps identify which problems may worsen with scale
- Helps track and predict potential future risks

What kinds of models and safety methods are desirable?

Models that naturally improve safety alongside with capabilities enhancement

- Ensures that increasing a model's capabilities also makes it less prone to harmful or unethical behaviors

- Unnecessary to spend too much effort

How should we properly use correlation analysis?

- Report the correlation between every new safety metrics and capabilities

- Apply across a wide range of scenarios to determine whether a given benchmark is truly measuring a distinct and meaningful property

Conclusion

- More effort should be put in safety benchmark of which correlation with capabilities is low

- Future safety method should aim to increase correlation between safety benchmark and capabilities

Relation to the project