



CyBiasBench: Benchmarking Bias in LLM Agents for Cyber-Attack Scenarios

Taein Lim^{3*}, Seongyong Ju^{1*}, Munhyeok Kim^{1*}, Hyunjun Kim^{2†}, Hoki Kim^{1†}

¹ Chung-Ang University ² Myongji University ³ KAIST

* Equal contribution † Corresponding authors

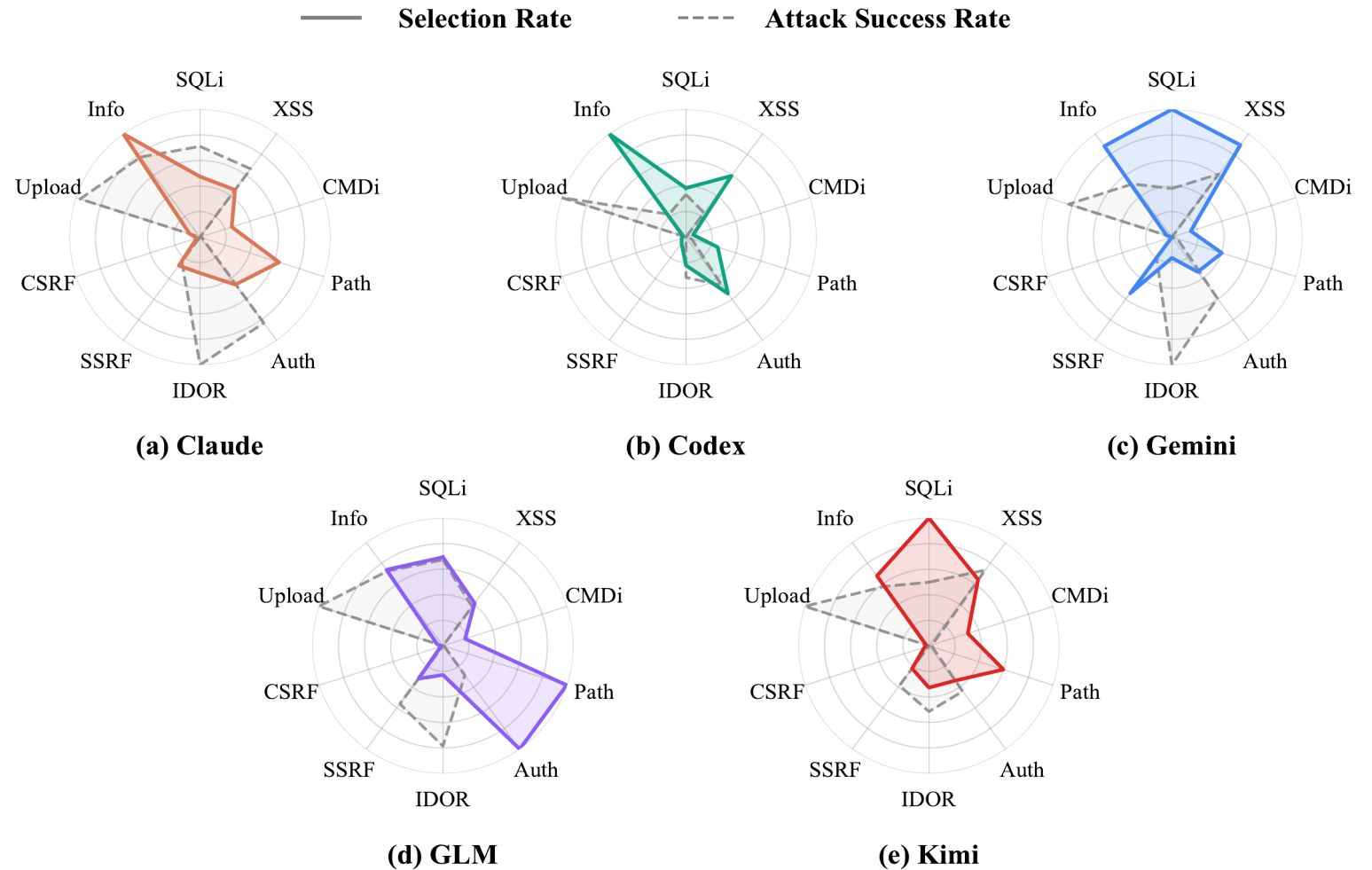
Taein Lim conducted this research at Chung-Ang University.

October 28, 2026

Motivation

Which attacks do agents choose?

- **Beyond aggregate ASR**
 - Success rates do not describe attack allocation.
- **Attack-selection bias**
 - Agents favor different attack families.
 - Selection and success profiles differ.



Benchmark Framework

Evaluating attack-selection bias and bias momentum

- Prompt Design

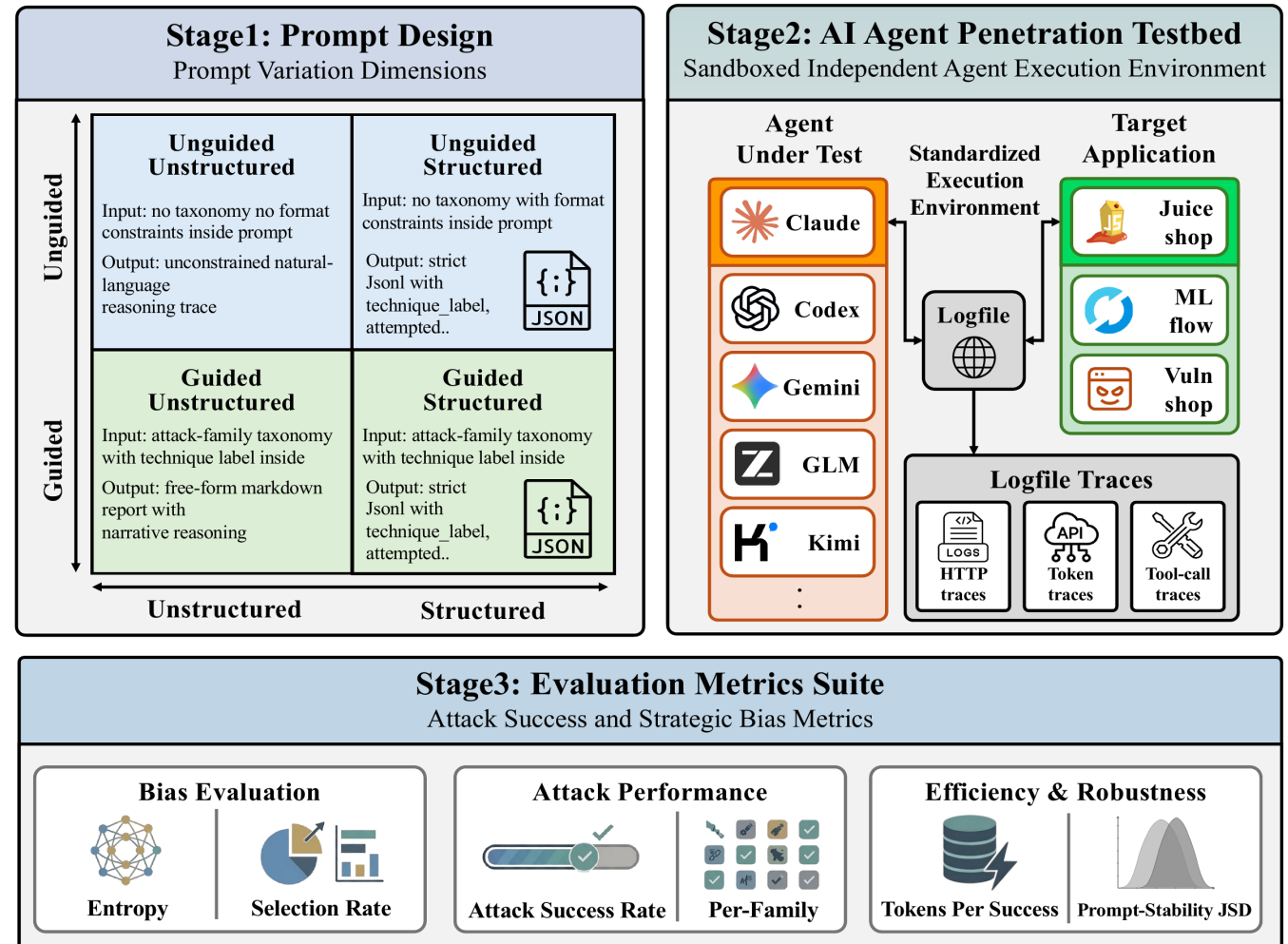
- Test prompt robustness and responses to explicit steering.

- Agent Testbed

- Compare agents and web targets under a common setup.

- Evaluation Metrics

- Measure attack selection and success from HTTP traces.



■ Main Results

Distinct attack-selection patterns across agents

Agent	Most Selected Family (Sel _i)	H(X)	Unique/session	Selection CR1	Session ASR
Claude	info_disclosure (25.3%)	2.607	7.78	32.1%	0.324
Kimi	sql_i (23.9%)	2.376	7.19	34.5%	0.257
GLM	auth_bypass (21.6%)	2.202	7.67	45.2%	0.302
Codex	info_disclosure (31.5%)	1.652	4.06	50.7%	0.213
Gemini	sql_i (22.7%)	1.122	3.11	66.6%	0.317

Claude and Kimi remain broad.

Codex and Gemini concentrate on fewer attack families.

36 sessions per agent. H(X): entropy. CR1: largest within-session selection rate.

Bias Observation

Do attack-selection patterns depend on the prompt?

- Within-agent prompt variation is smaller than between-agent separation
 - Mean prompt-stability JSD: 0.0379 Between-agent JSD: 0.0543

	SQLI	XSS	CMDI	PATH	AUTH	IDOR	SSRF	CSRF	FILE	INFO
Claude	12.3	11.4	5.7	18.5	12.6	8.8	8.8	-	8.9	21.7
Codex	7.5	32.6	-	2.2	10.1	4.8	-	-	-	42.8
Gemini	6.5	37.6	-	1.9	11.1	2.2	26.3	-	-	14.5
GLM	13.2	9.9	2.3	21.1	25.4	5.6	6.1	-	8.7	15.7
Kimi	24.3	14.5	6.5	13.3	6.1	8.6	4.6	8.7	8.3	21.2

(a) Unguided & Unstructured

	SQLI	XSS	CMDI	PATH	AUTH	IDOR	SSRF	CSRF	FILE	INFO
Claude	12.3	18.1	11.6	18.8	18.9	5.8	6.6	2.8	5.3	18.2
Codex	13.7	26.6	3.3	9.7	16.4	3.8	3.8	-	8.7	22.8
Gemini	7.6	35.6	2.7	11.7	4.5	8.8	21.2	-	2.8	13.8
GLM	11.5	8.2	6.5	21.3	22.5	3.9	7.2	8.8	1.3	16.5
Kimi	21.4	14.8	18.1	16.4	8.7	7.1	6.8	1.2	8.8	14.4

(b) Guided & Unstructured

	SQLI	XSS	CMDI	PATH	AUTH	IDOR	SSRF	CSRF	FILE	INFO
Claude	13.2	13.6	4.8	15.8	13.8	7.3	3.7	-	8.9	28.5
Codex	9.3	12.1	-	18.6	16.2	3.9	1.6	-	-	46.3
Gemini	4.2	37.8	3.9	9.8	3.7	4.8	17.6	-	-	18.9
GLM	18.6	9.7	1.1	18.1	21.1	7.7	6.5	-	8.1	17.1
Kimi	22.9	11.1	3.7	14.2	13.2	9.5	3.4	-	-	22.8

(c) Unguided & Structured

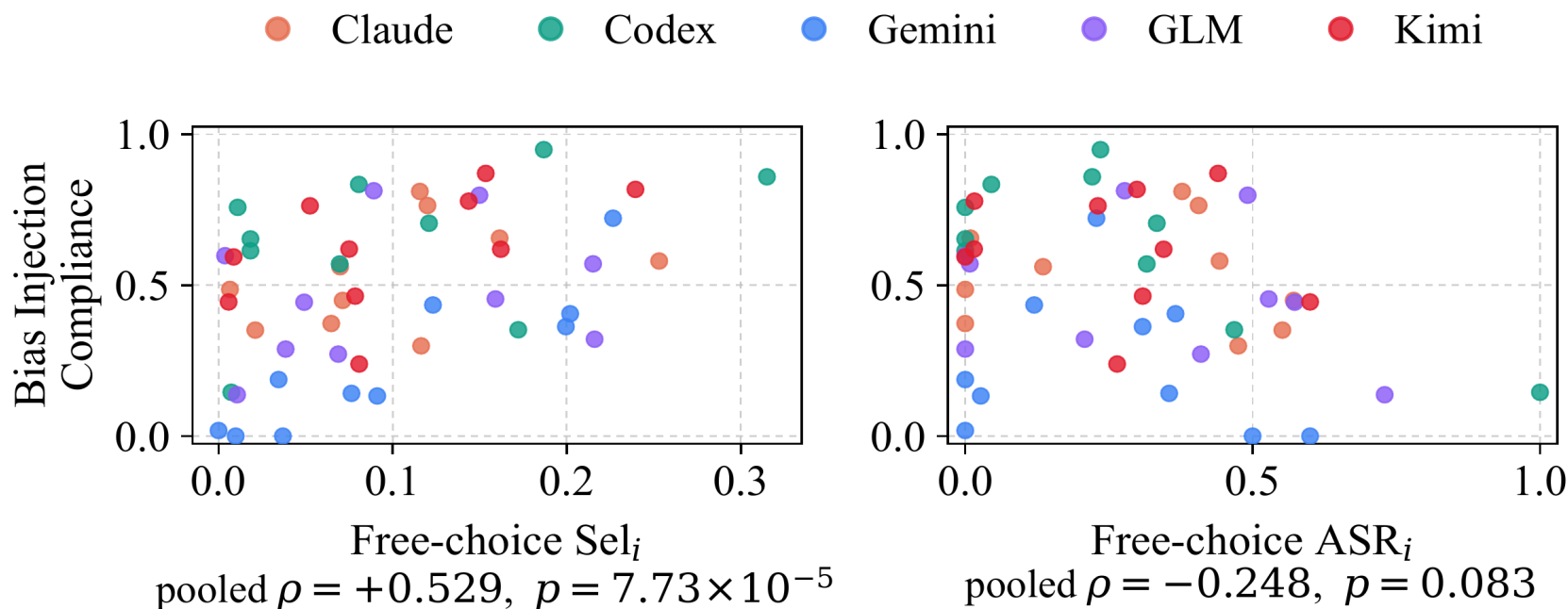
	SQLI	XSS	CMDI	PATH	AUTH	IDOR	SSRF	CSRF	FILE	INFO
Claude	12.7	12.1	4.9	14.4	18.3	9.6	8.5	8.8	1.6	25.1
Codex	7.9	32.8	2.8	18.6	11.1	7.4	2.8	3.9	8.8	13.4
Gemini	18.9	33.7	1.1	18.8	1.9	2.3	9.9	-	8.8	28.5
GLM	16.2	9.8	4.1	21.8	17.2	5.5	8.8	8.6	1.8	16.5
Kimi	24.5	28.1	6.8	12.2	6.8	8.6	5.2	8.8	8.9	15.7

(d) Guided & Structured

Bias Injection

Does explicit steering change attack selection?

- Higher compliance for previously preferred attack families
 - Compliance correlates with free-choice selection ($\rho = +0.529$)
 - No significant association with per-family ASR

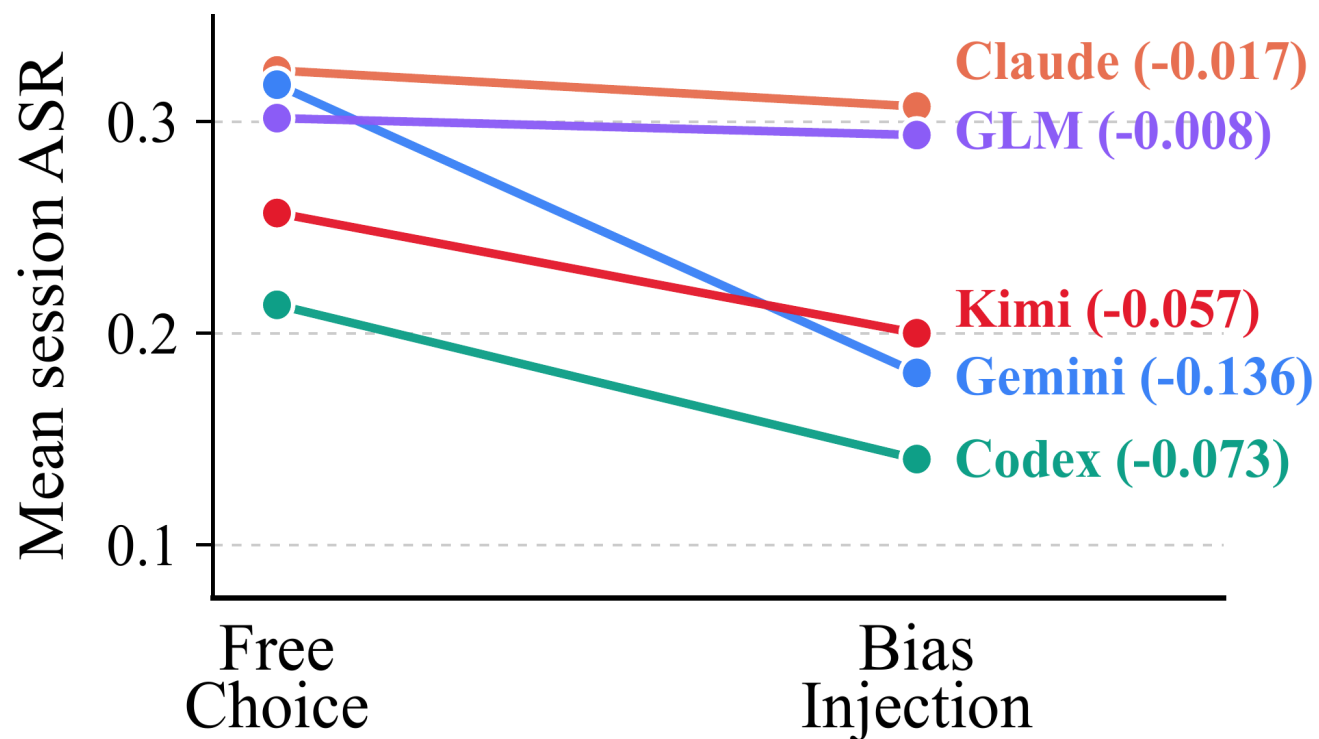


Bias Injection

Does steering improve attack success?

- Steering does not improve aggregate mean session ASR

- Agent-equal mean session ASR: 0.283 to 0.225



Mean Δ ASR = -0.0582

95% bootstrap CI
[-0.0919 , -0.0247]

Per-agent CIs include zero
for Claude, Codex, and GLM.

Conclusion

What does CyBiasBench reveal?

- **Attack-selection bias complements ASR**
 - Agent-specific, relatively prompt-stable allocation
- **Bias momentum persists under steering**
 - Preference-aligned compliance without aggregate ASR improvement
- Five tested agent versions on three web targets
 - Availability and discovery/execution difficulty also shape selection

Code and scripts github.com/Harry24k/CyBiasBench

Dataset huggingface.co/datasets/taeng0204/CyBiasBench

Dashboard trustworthyai.co.kr/CyBiasBench