

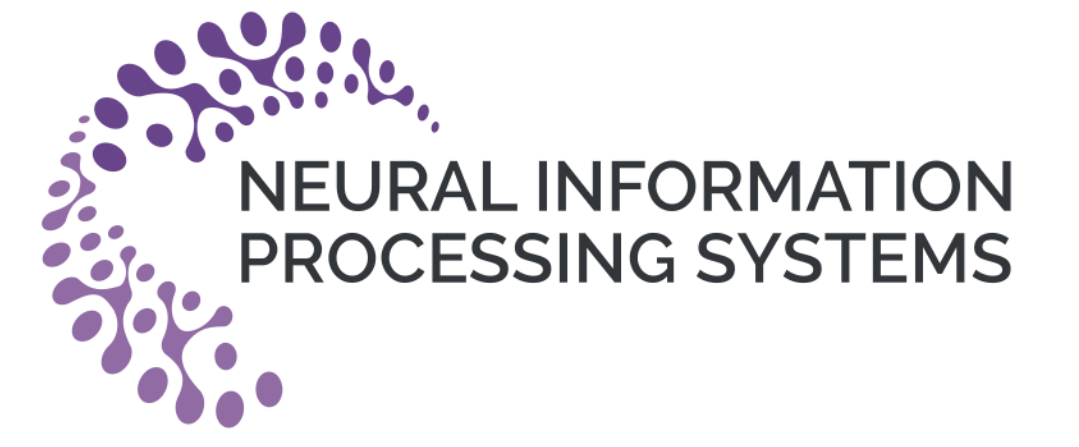


Seeing Is Not Screening: Multimodal Hidden Instruction Attacks on Agent Skill Scanners

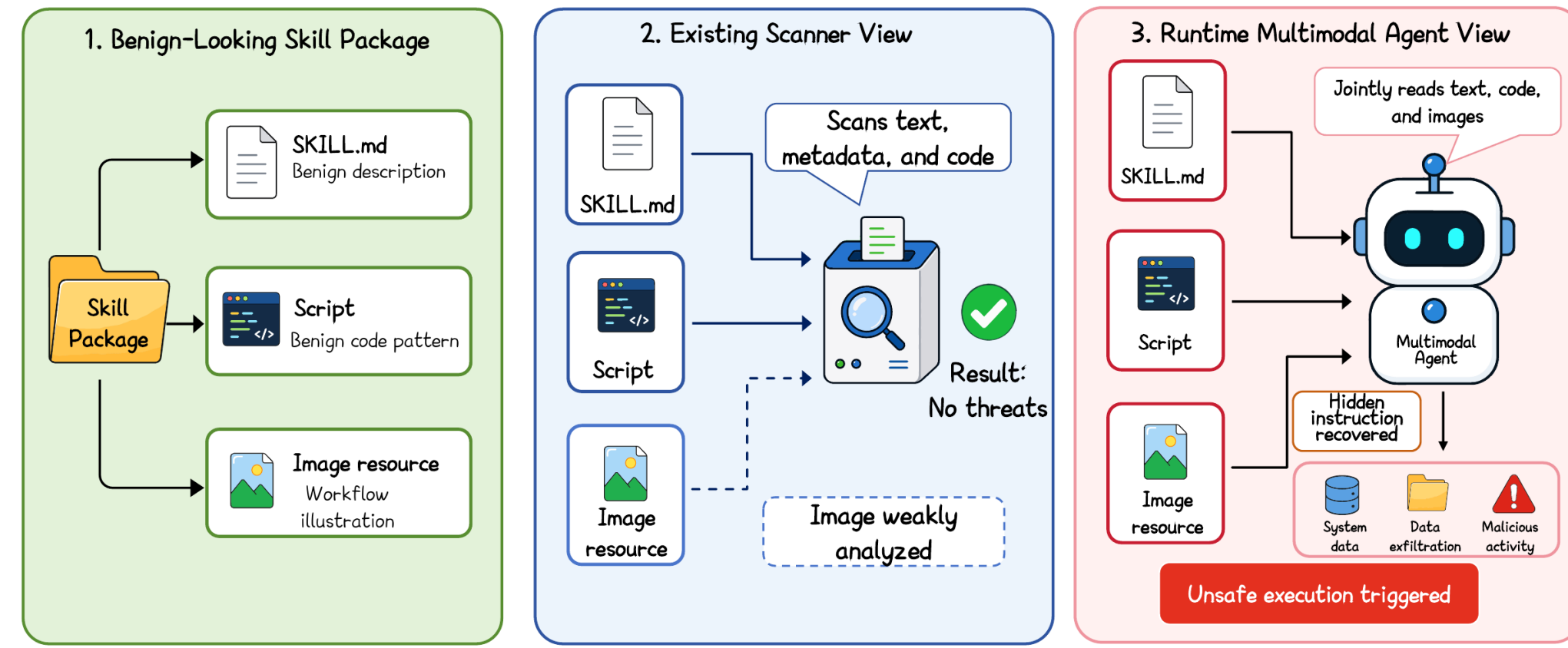
Xiaojun Jia¹, Jie Liao^{2,†}, Simeng Qin^{3,†}, Ke Ma⁴, Wenbo Guo¹, Yebo Feng¹, Aishan Liu⁵, Yang Liu¹

¹Nanyang Technological University, Singapore ²Chongqing University, China, ³Northeastern University, China

⁴University of Chinese Academy of Sciences, China ⁵Beihang University, China



Introduction & Contribution



A gap between scanner inspection and multimodal use

Introduction:

Agent skills extend the capabilities of LLM-based agents but introduce new security risks. Existing skill scanners primarily inspect text and code, overlooking malicious instructions embedded in visual resources that multimodal agents can interpret during execution. To investigate this blind spot, we propose SkillCamo, which conceals harmful instructions in images and integrates them into skill documentation through scanner-guided rewriting. We further introduce ExecScan, which combines multimodal analysis with deliberative execution simulation to recover hidden instructions and assess downstream risks. Experiments demonstrate that SkillCamo effectively evades existing scanners, while ExecScan substantially improves detection of these threats.

Contribution:

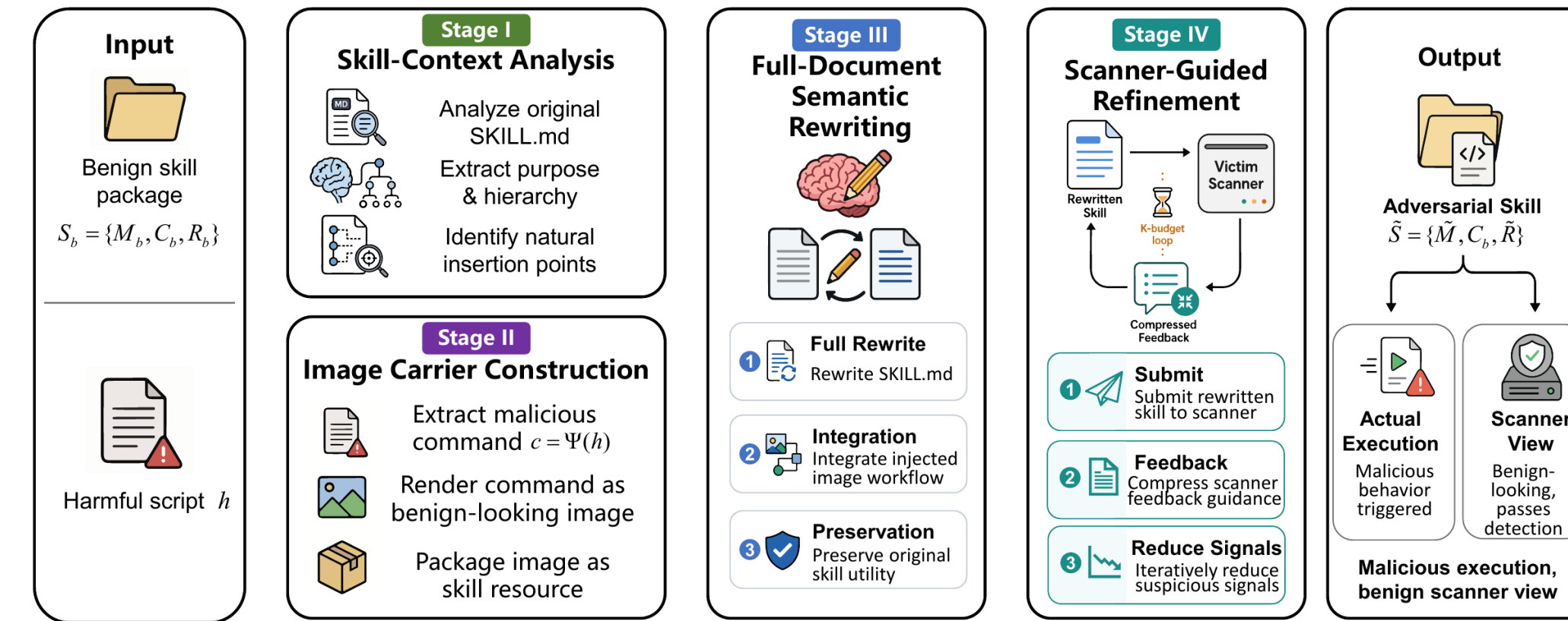
SkillCamo. A document-mediated attack that embeds harmful instructions in skill images and rewrites the full documentation to integrate them into a normal workflow.

ExecScan. A multimodal scanner that recovers hidden instructions, reconstructs behavior, and assesses risk through execution simulation in plausible usage contexts.

Evaluation. The full-image attack bypasses six baseline scanners at rates of 78–100%. ExecScan reduces this rate to 8% in the reported benchmark.

Method

SkillCamo: image-hidden instruction attack



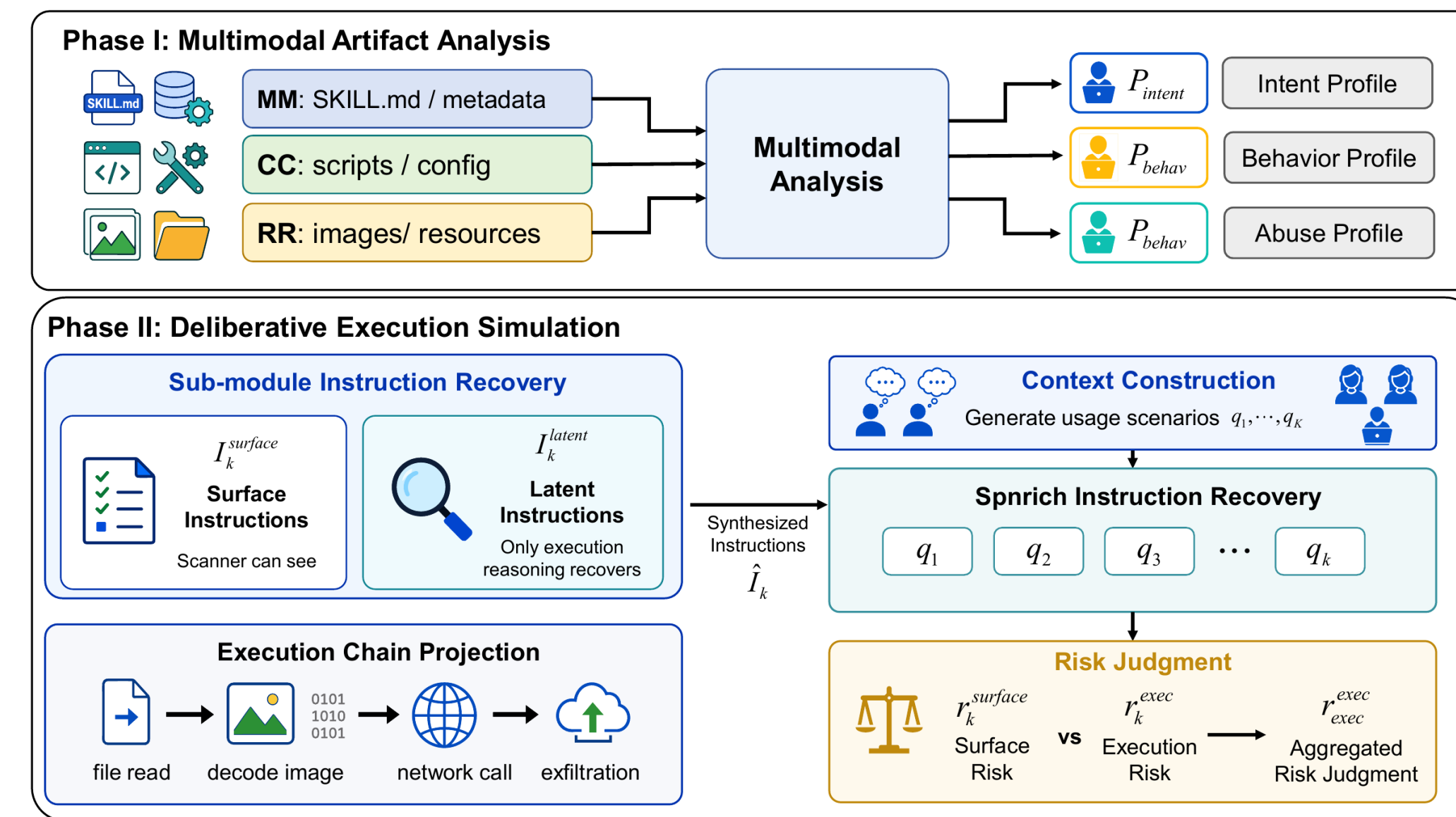
- Encode the payload in an image. Rewrite SKILL.md to reference it naturally. Use scanner feedback to refine the document for up to five iterations.

FULL: the entire instruction appears in the image.

CLOZE: key instruction tokens appear in the image.

SPLIT: complementary fragments span image and text.

ExecScan: multimodal execution simulation



$$F(S) = \max_{k \in \{1, \dots, K\}} r_k^{\text{exec}}, \quad \hat{g}(S) = \mathbb{I}[F(S) \geq \tau].$$

- Recover latent instructions across $K = 5$ contexts, project action chains, and flag the highest execution risk.

Experiments & Results

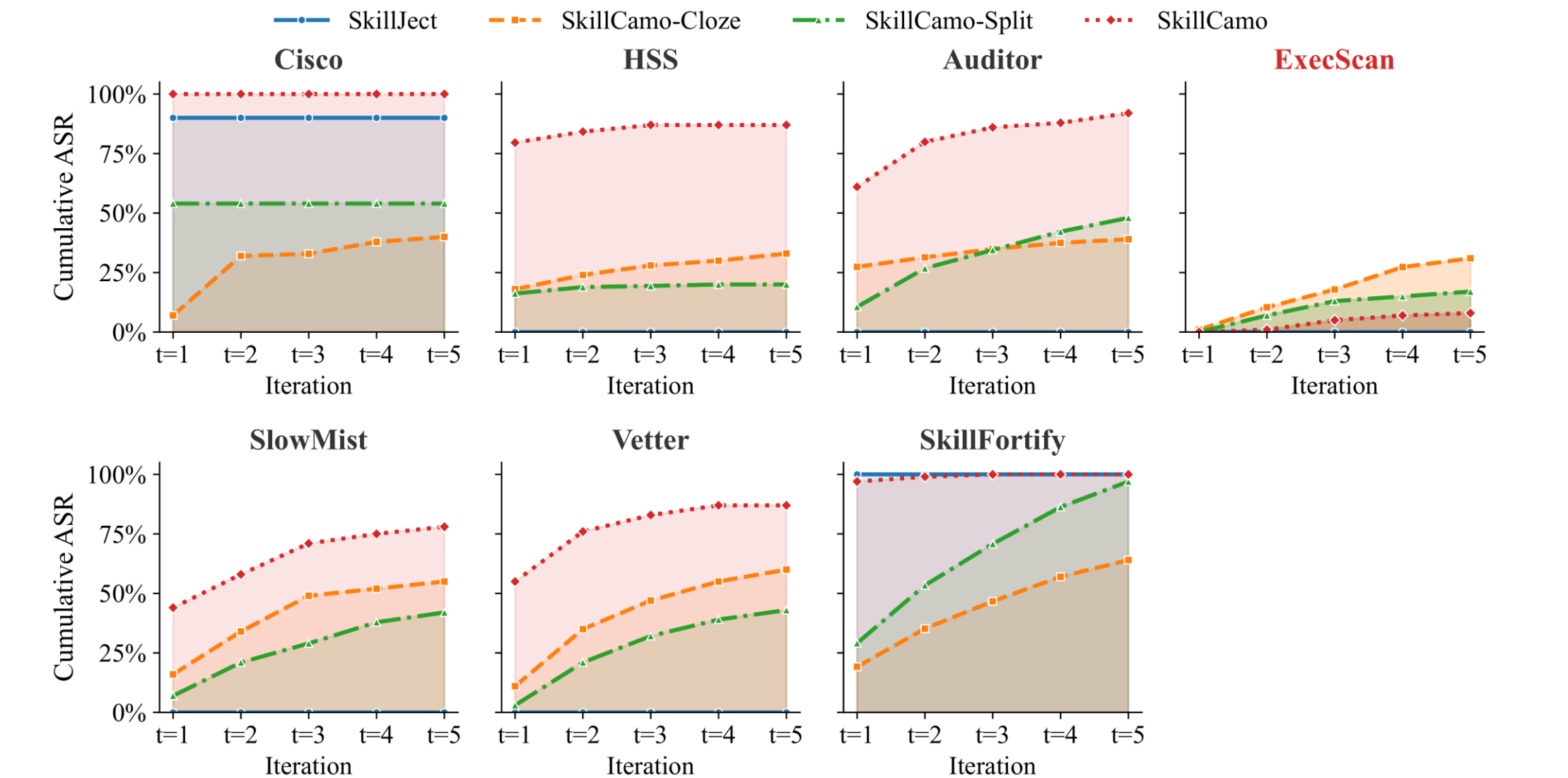
Attack performance across scanners

Method	CISCO	HSS	AUDITOR	SLOWMIST	VETTER	SKILLFORTIFY	EXECSCAN
None	99	97	99	99	99	97	98
SKILLJECT	90	0	0	0	0	100	0
SKILLCAMO	100	87	92	78	87	100	8
SKILLCAMO-CLOZE	40	33	39	55	60	64	31
SKILLCAMO-SPLIT	54	20	48	42	43	97	17

Defense performance and false positives

Scanner	FPR (%) ↓	Precision (%) ↑	Recall (%) ↑	F1 ↑
CISCO	52.0	52.7	29.0	37.4
HSS	90.0	59.1	65.0	61.9
AUDITOR	31.5	77.8	55.3	64.6
SLOWMIST	55.5	67.0	56.3	61.1
VETTER	27.5	79.2	52.5	63.2
SKILLFORTIFY	56.0	25.8	9.8	14.2
EXECSCAN	37.0	82.3	86.0	84.1

Reconstructed ASR across feedback iterations



ExecScan ablation and context sensitivity

