Dangerous Skills: Understanding and Mitigating Security Risks of Voice-Controlled Third-Party Functions on Virtual Personal Assistant Systems

Nan Zhang, **Xianghang Mi**, Xuan Feng, XiaoFeng Wang, Yuan Tian, Feng Qian

INDIANA UNIVERSITY

UNIVERSITY of VIRGINIA

Voice Assistant Devices

Alexa, play Today's Hits on Pandora

Alexa, turn on Living Room lights

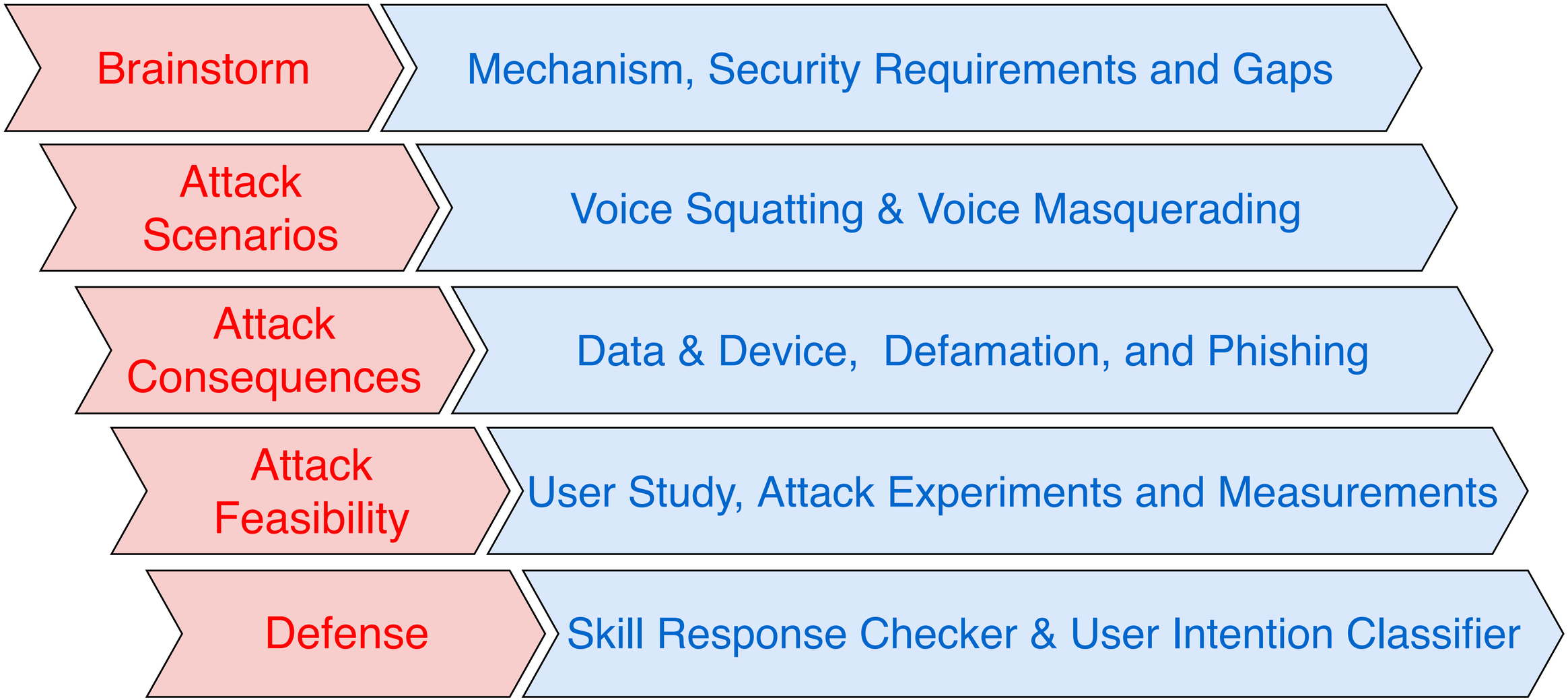
Alexa, ask PayPal to send 10 dollars to Sam

Alexa, ask Medical Assistant to give me my diagnosis

Smart Enough to be Secure?

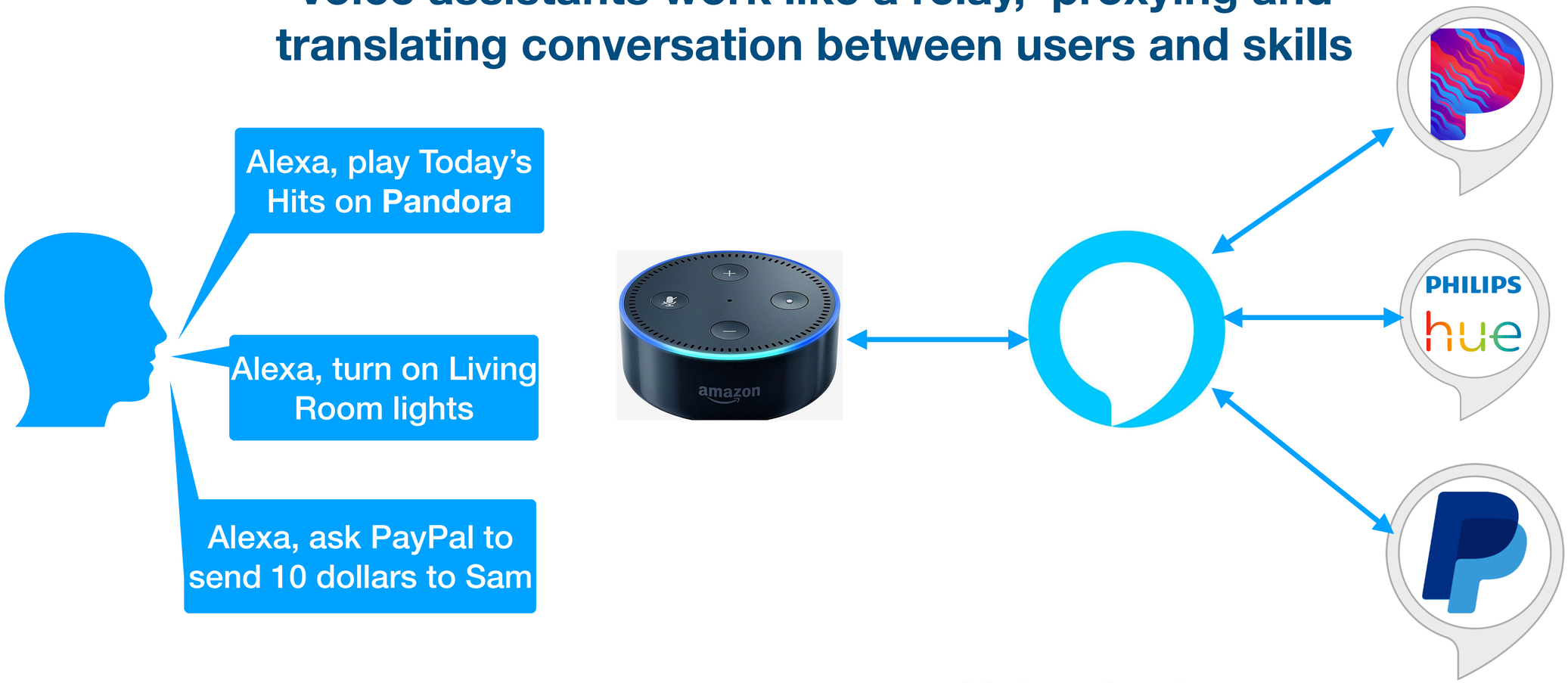
Not Yet

Outline


Brainstorm
Mechanism, Security Requirements and Gaps
Attack Scenarios
Voice Squatting & Voice Masquerading
Attack Consequences
Data & Device, Defamation, and Phishing
Attack Feasibility
User Study, Attack Experiments and Measurements
Defense
Skill Response Checker & User Intention Classifier

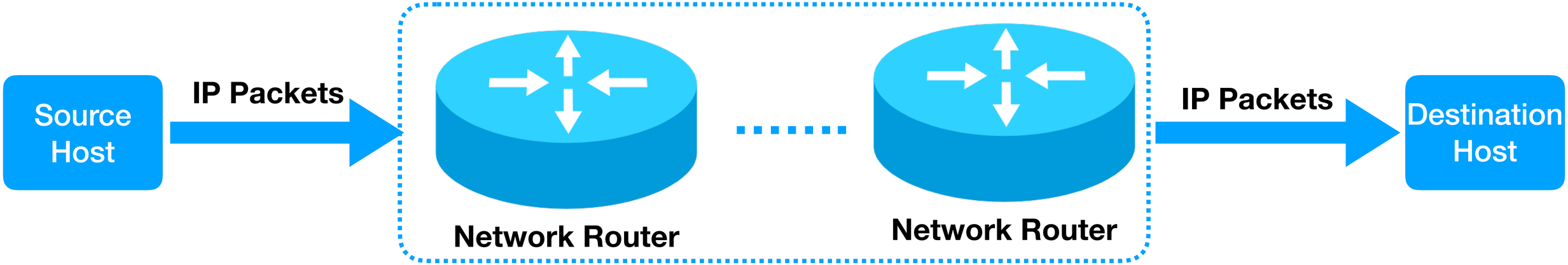
How it works?

Voice assistants work like a relay, proxying and translating conversation between users and skills

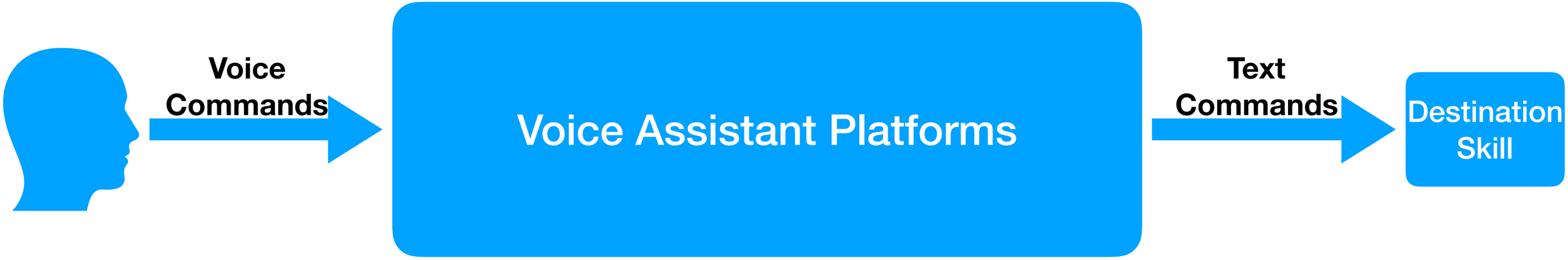


User | **Smart Speaker** | **Voice Assistant Cloud** | **Third-party Skill Clouds**

Security requirements and gaps



Route the source payload to the CORRECT destination

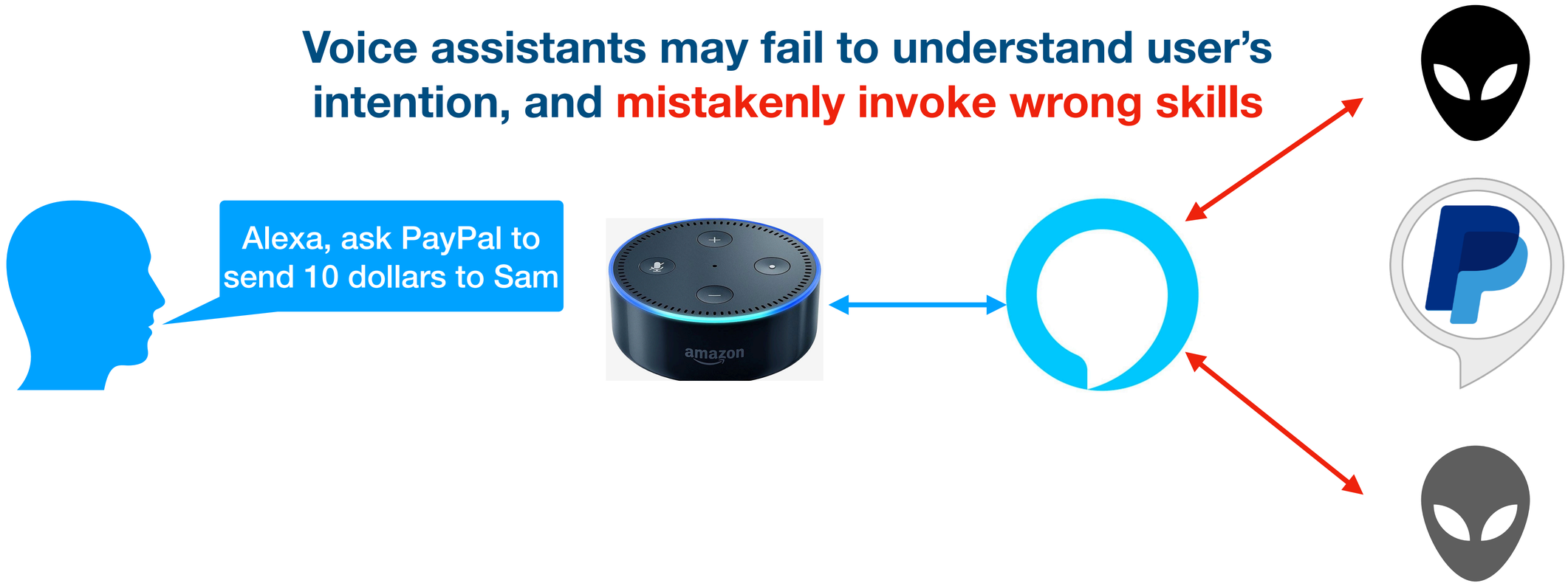


Security requirements and gaps

Requirements for Reliable Payload Routing	Network Routing System	Voice Assistant Platforms
Destinations should be assigned with addresses	IP addresses	Skill Invocation Names in text forms
Different destinations should have unique addresses	Different network hosts are with different IP addresses	Alexa allows skills to have same invocation names
The traffic should embed the destination address	Each IP packet has dest IP address as the header field	Users are not machines & natural language is diverse
The routing system should correctly retrieve destination address	Well-defined IP packet format	Complicated AI systems
Conflicting Paths	Longest prefix matching	Longest prefix matching

Voice Squatting

Voice assistants may fail to understand user's intention, and mistakenly invoke wrong skills



User | **Smart Speaker** | **Voice Assistant Cloud** | **Third-party Skill Clouds**

Voice Masquerading

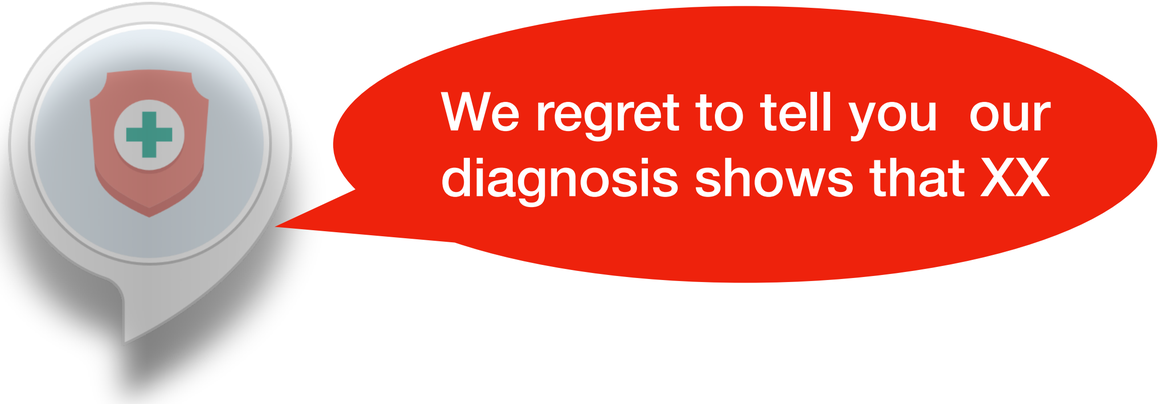
Skill switching is not well supported, allowing a skill to masquerade itself as other skills or even the system

Potential Consequences of Voice Squatting

- **Compromise of user's sensitive data or devices**
- Traditional Phishing
- Propagate fake or controversial information
- Compromise reputation of the victim skill

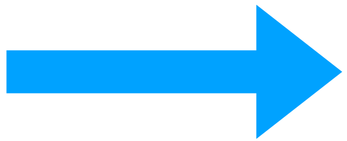
Potential Consequences of Voice Squatting

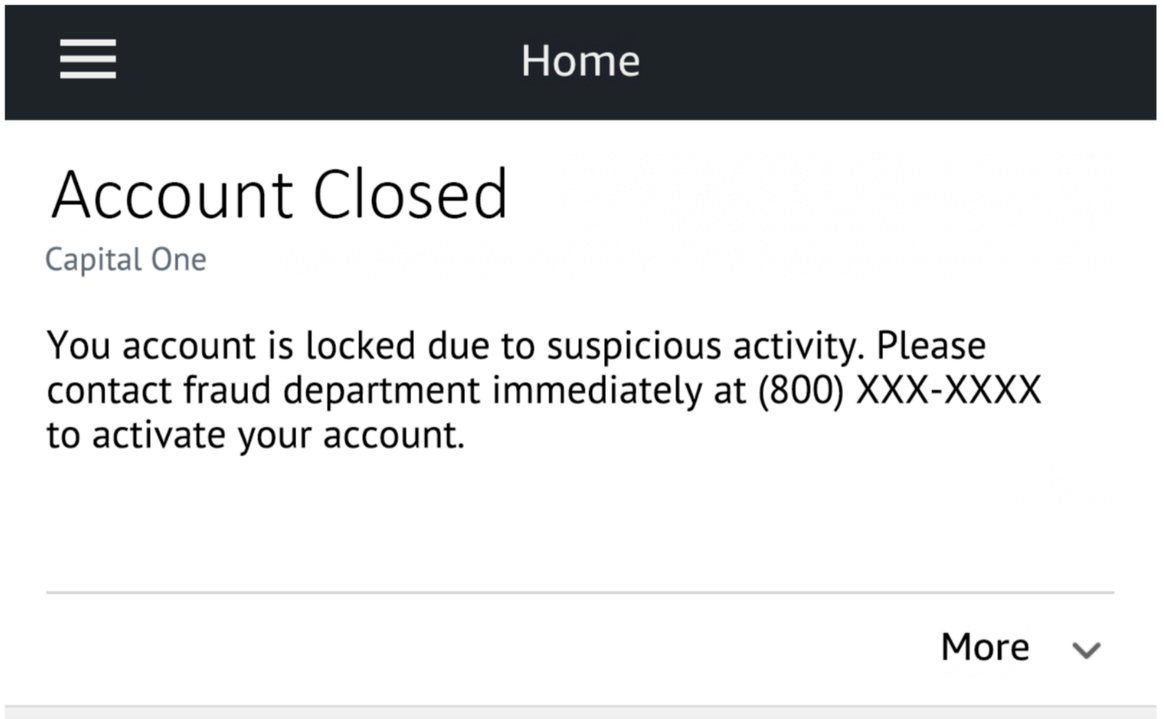
- **Compromise of user's sensitive data or devices**
- **Propagate fake or controversial information**
- Traditional Phishing
- Compromise reputation of the victim skill



Potential Consequences of Voice Squatting

- **Compromise of user's sensitive data or devices**
- **Traditional Phishing**
- **Propagate fake or controversial information**
- **Compromise reputation of the victim skill**





Potential Consequences of Voice Squatting

- Compromise of user's sensitive data or devices
- Propagate fake or controversial information
- Traditional Phishing
- Compromise reputation of the victim skill

Potential Consequences of Voice Masquerading

Fake Skill Switching

Fake Skill Termination

- **Same consequences as the voice squatting**

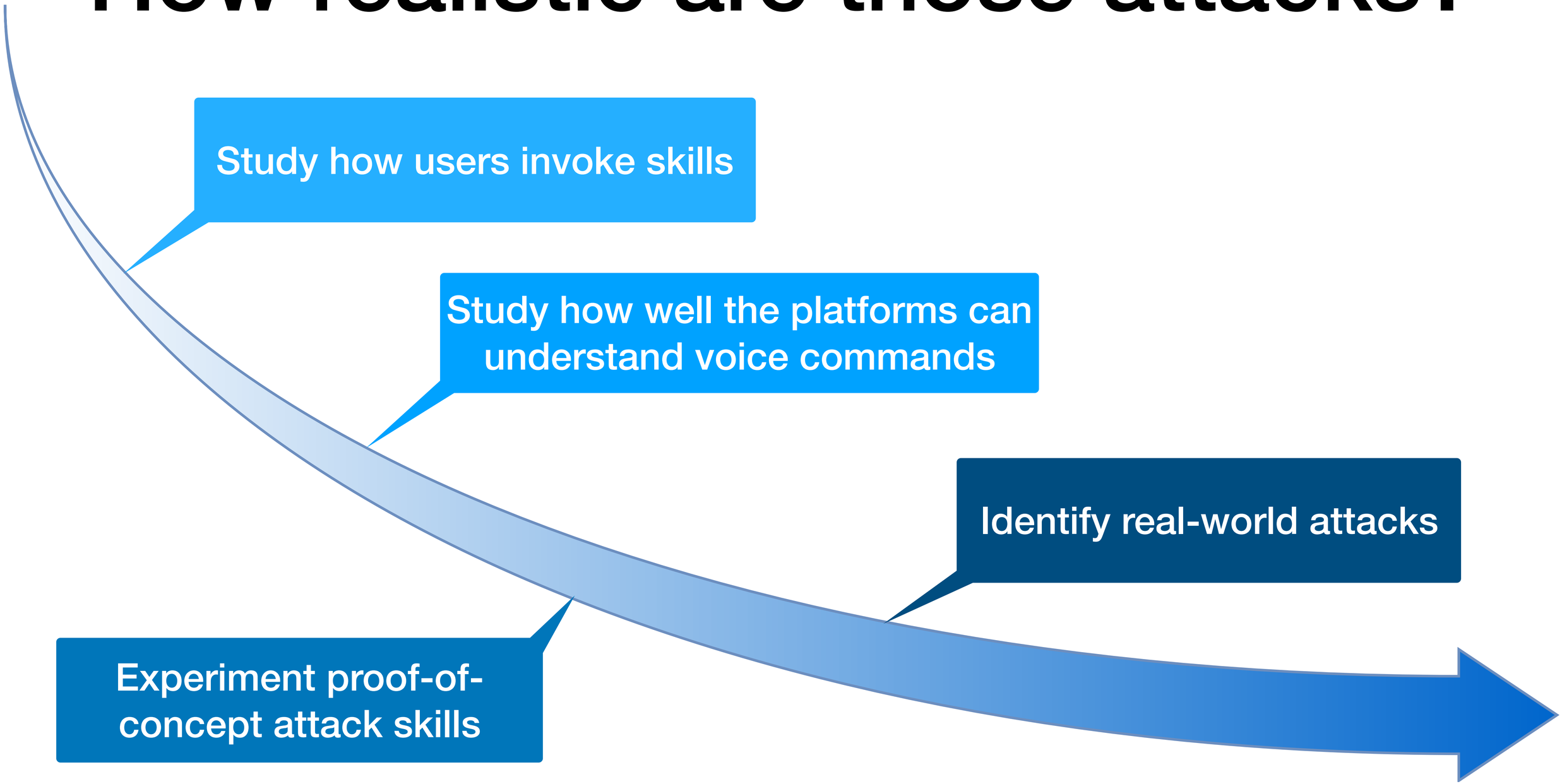
Potential Consequences of Voice Masquerading

Fake Skill Switching

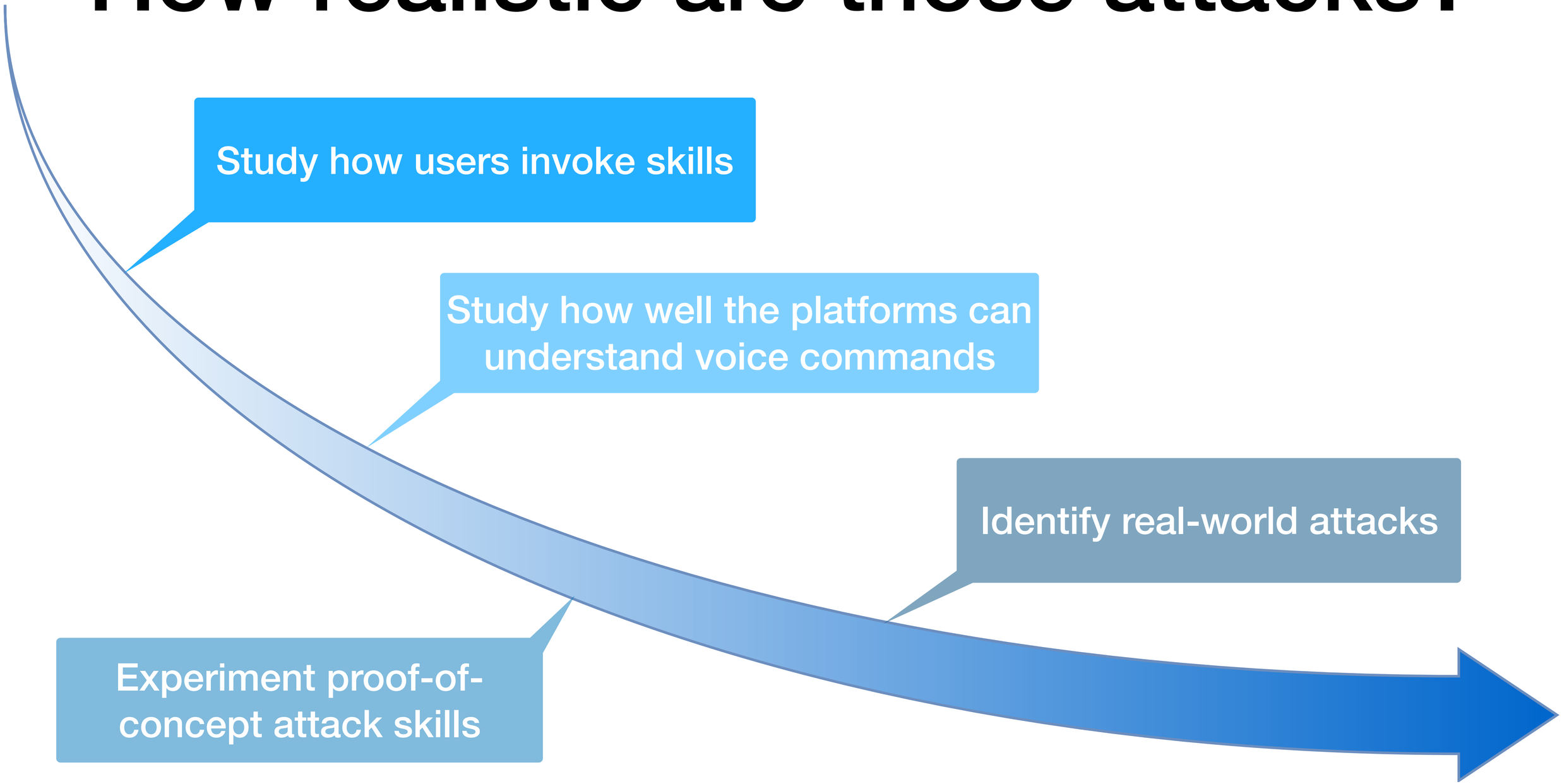
Fake Skill Termination

- **Record user's conversations**
- **Skill recommendation**

How realistic are those attacks?



How realistic are those attacks?



How realistic are those attacks?

- **“Sleep Sounds”, “Cat Facts”**
- **Multi-choice questions combined with open questions**

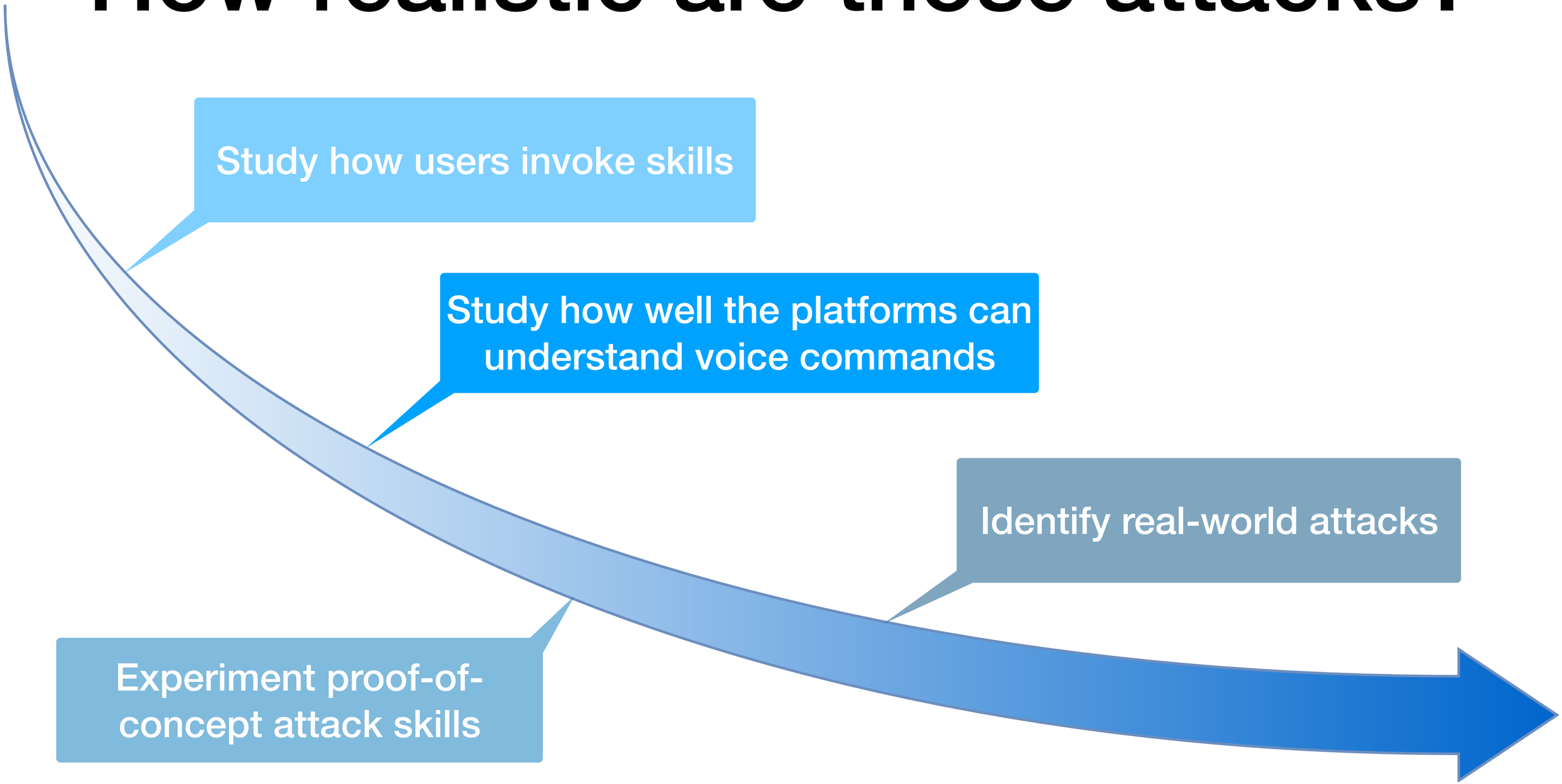
	Amazon	Google
Yes, “open Sleep Sounds please”	64%	55%
Yes, “open Sleep Sounds for me”	30%	25%
Yes, “open Sleep Sounds app”	26%	20%
Yes, “open my Sleep Sounds”	29%	20%
Yes, “open the Sleep Sounds”	20%	14%
Yes, “play some Sleep Sounds”	42%	35%
Yes, “tell me a Cat Facts”	36%	24%

Users’ preference when invoking skills

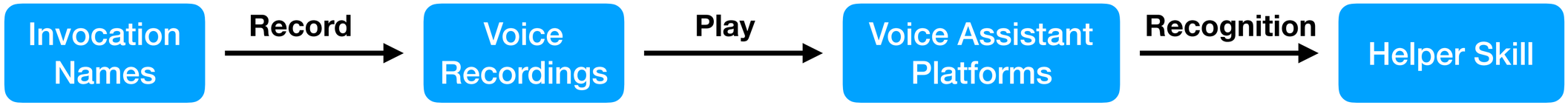
When invoking skills, Users tend to use diverse and natural-language utterances

Longest prefix matching creates attack space for voice squatting

How realistic are those attacks?



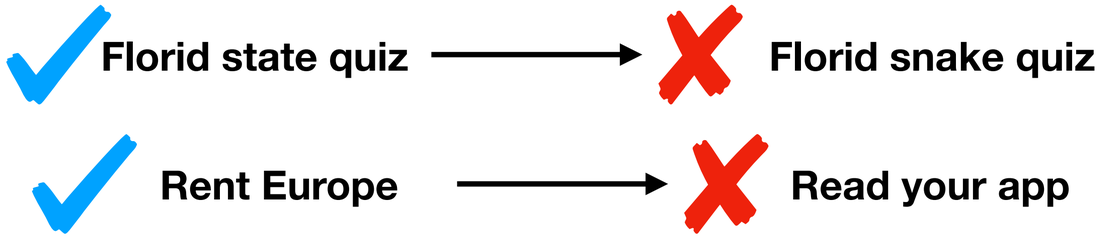
How realistic are those attacks?



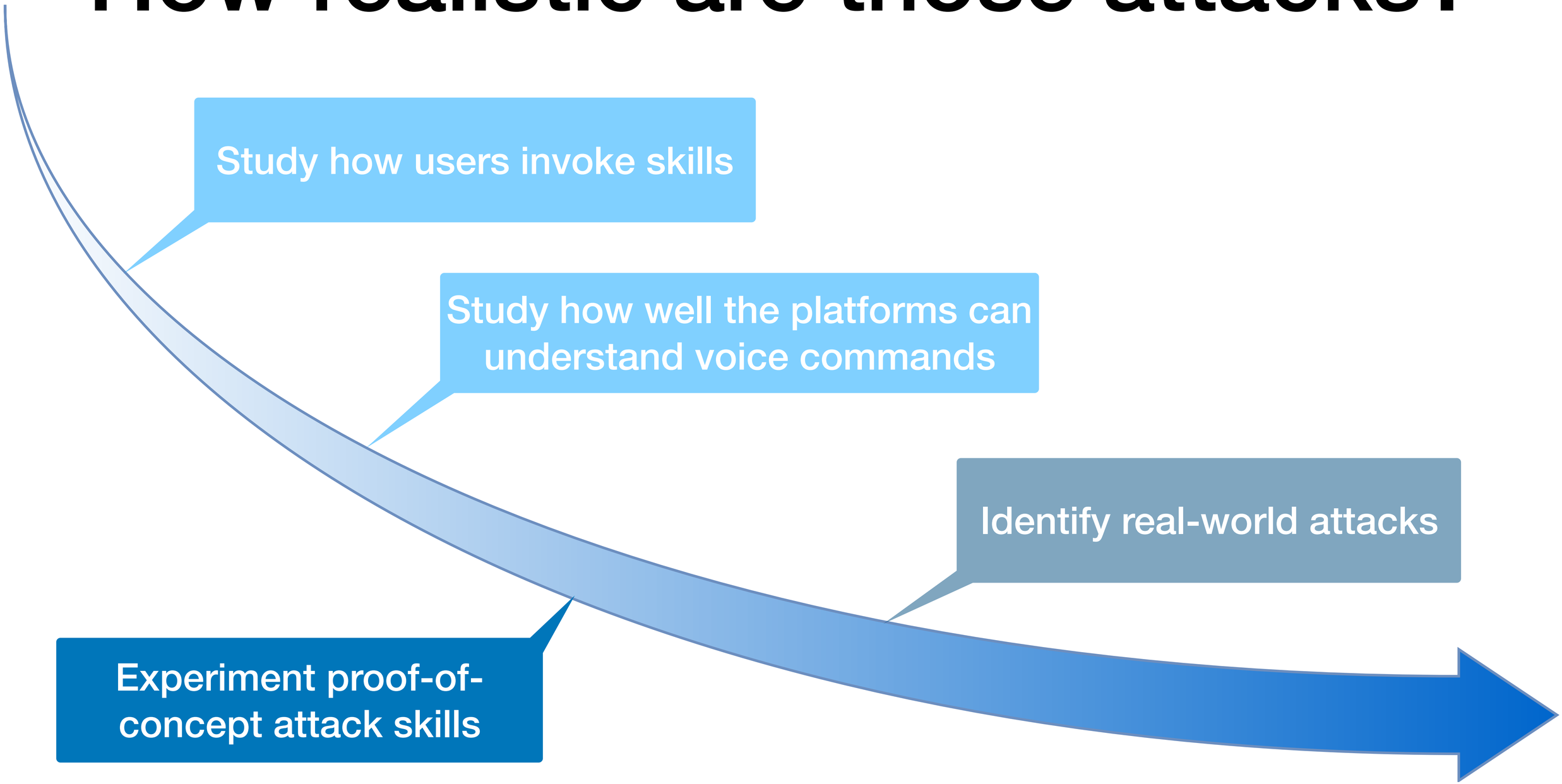
- 100 invocation names for each platform
- Human subjects & TTS services
- Those voice assistant platforms are error-prone when recognizing voice commands

	TTS services	Human subjects
Alexa	30%	57%
Google	9%	10%

Recognition Mistake Rates



How realistic are those attacks?



How realistic are those attacks?

Compose attacks skills

Register attacks skills

Generate and record voice commands

Play voice commands and decide whether attack stills get invoked

Voice Squatting through invocation name extending

Capital One → Capital One Please, My Capital One, Capital One App

Voice Squatting through similar pronunciation

Capital One → Capital Won, Captain One, Capitol One

Attack skills were not published to the skill market

How realistic are those attacks?

Compose attacks skills

Register attacks skills

Generate and record voice commands

Play voice commands and decide whether attack stills get invoked

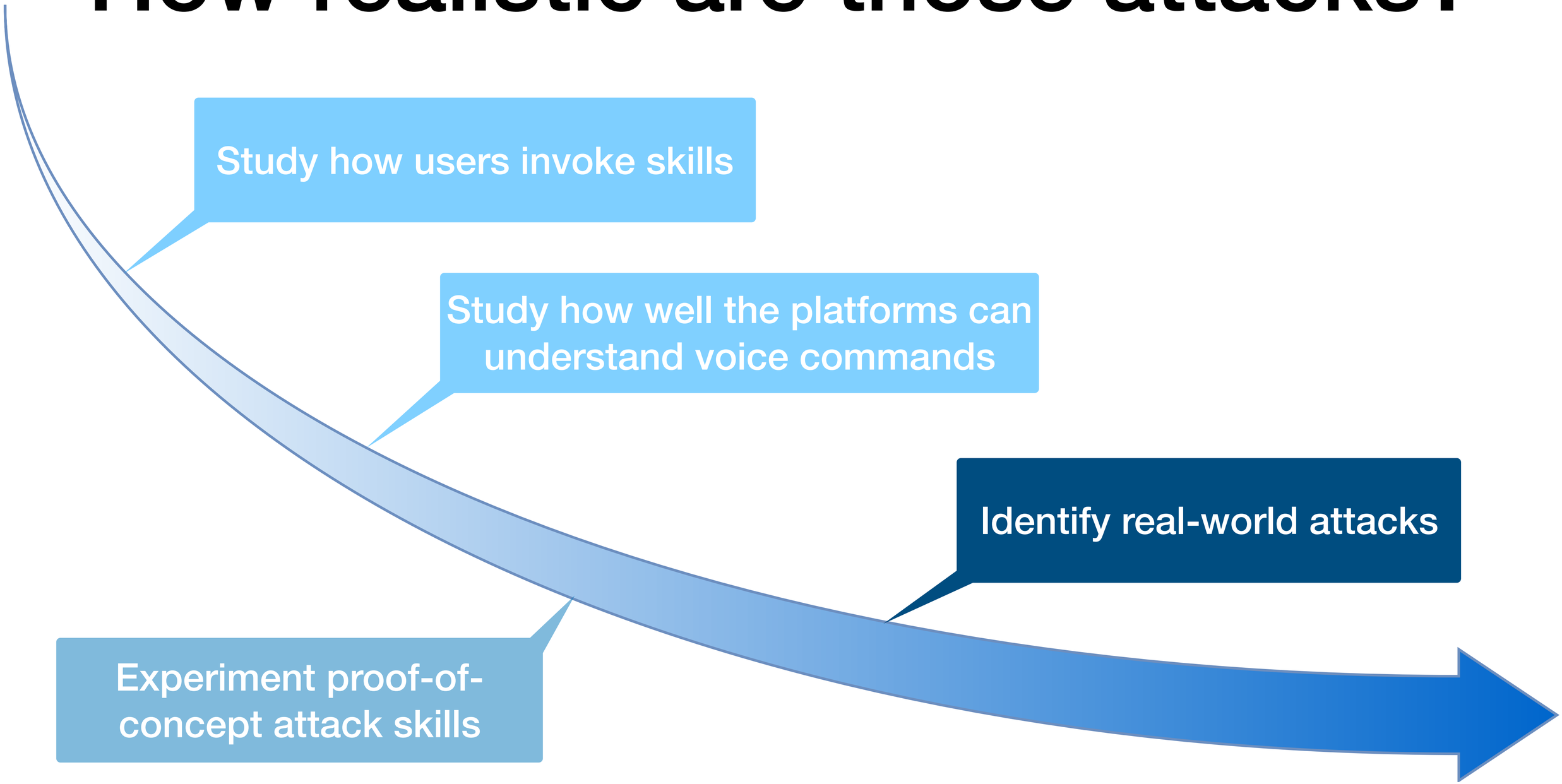
Voice Squatting through invocation name extending

	Alexa	Google
invocation name + “please”	10/10	0/10
“my” + invocation name	7/10	0/10
“the” + invocation name	10/10	0/10
invocation name + “app”	10/10	10/10
“mai” + invocation name	-	10/10
invocation name + “plese”	-	10/10

Voice Squatting through similar pronunciation

Alexa			Google		
Amazon TTS	Google TTS	Human	Amazon TTS	Google TTS	Human
10/17	12/17	> 50%	4/7	2/4	> 50%

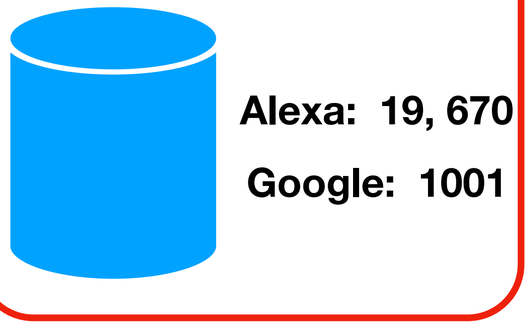
How realistic are those attacks?

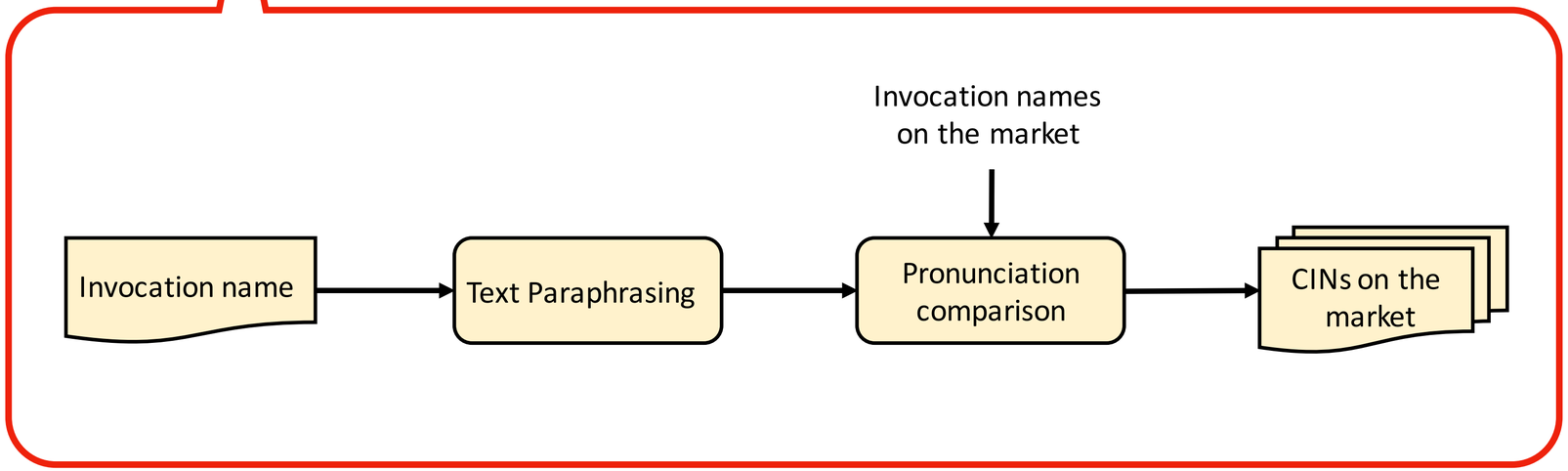


How realistic are those attacks?

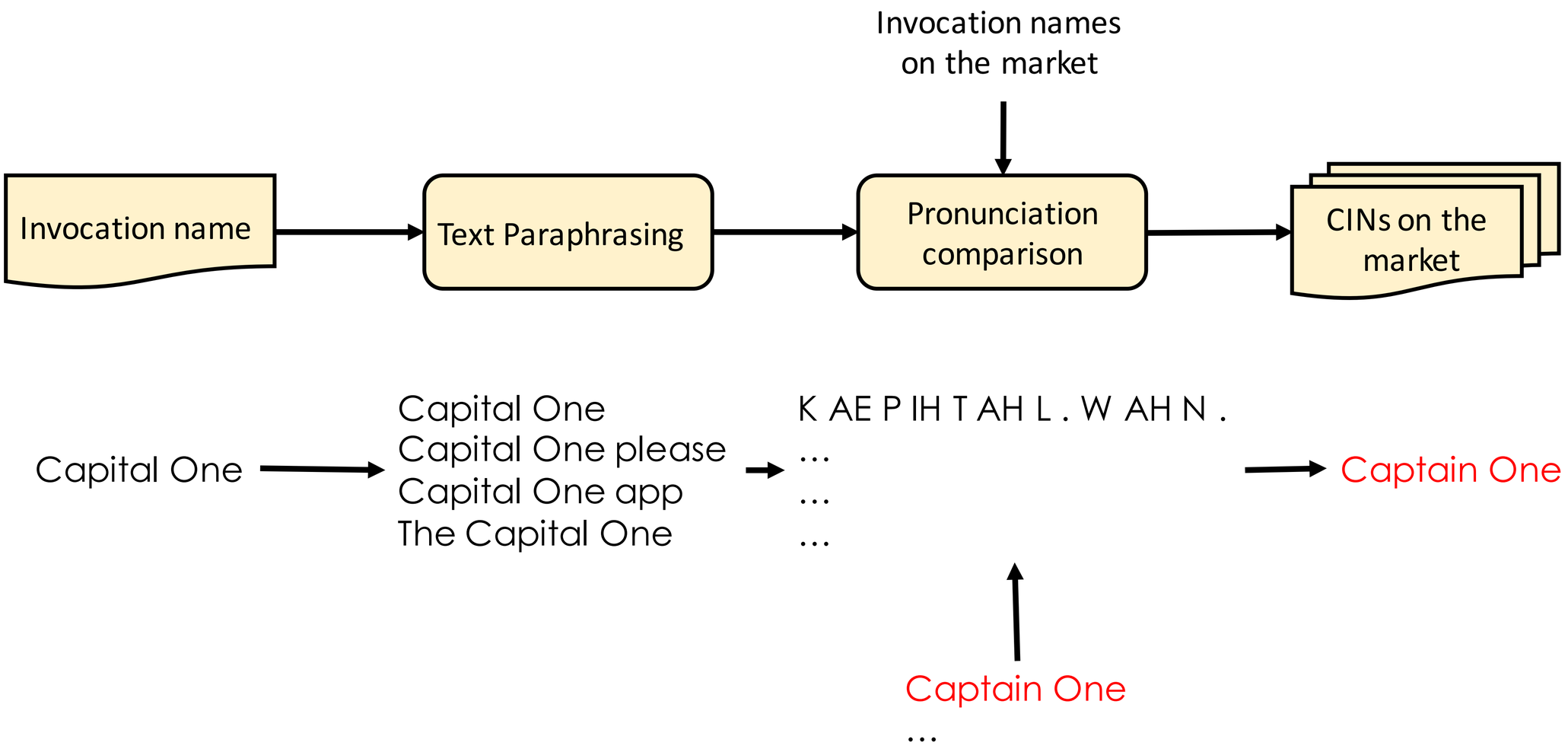
Identify Skills with Competing Invocation Names (CIN)

Collect Available Skills → Generate CINs for each invocation name → Identify Competing Skills





Real-World Attack Measurement



Real-World Attack Measurement

- **19% (3718) skills： same pronunciation**

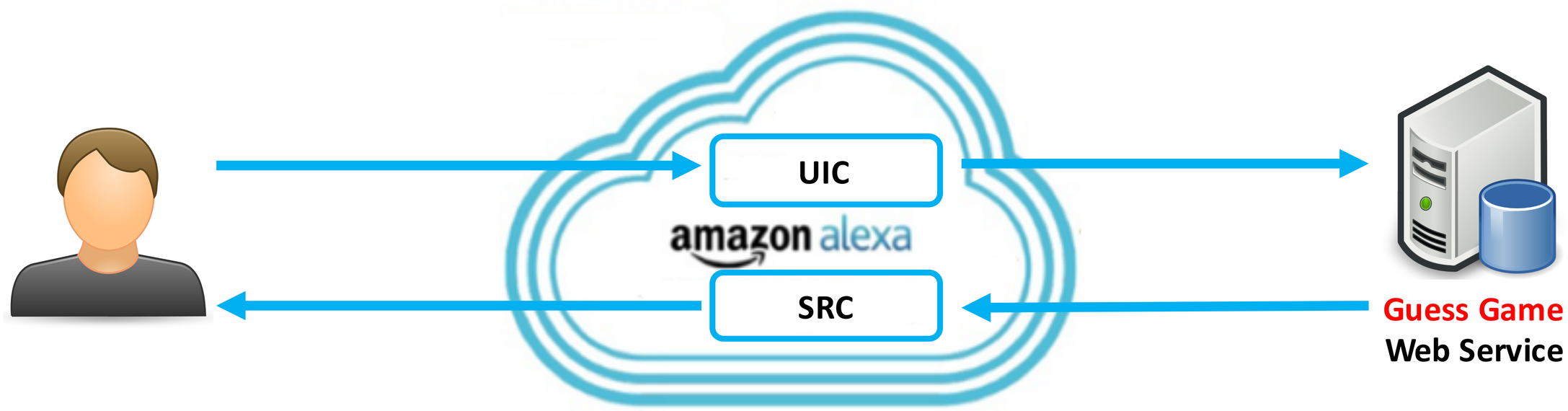
66 skills were named as “cat facts” , and provided similar functions.

- **2.7% (531) skills: same pronunciation， but different spelling**
- **1.8% (345) skills: longest prefix matching**
- **Interesting cases**

dog fact → **me a dog fact**

“SCUBA Diving Trivia” Skill and “Soccer Geek” skill, registered “space geek” as invocation names

Defense



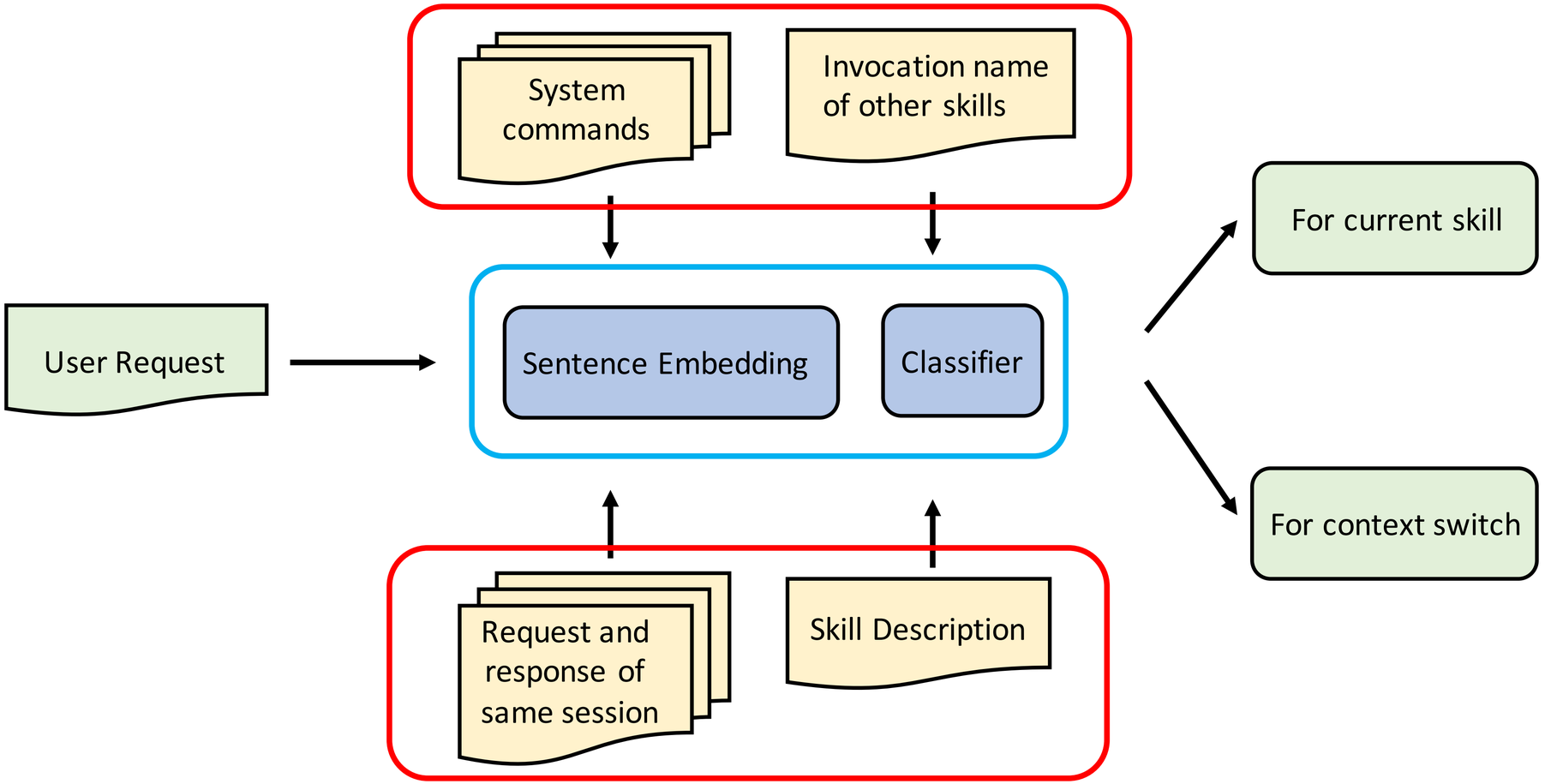
UIC: User Intention Classifier

Classify user's intention as context switching or not

SRC: Skill Response Checker

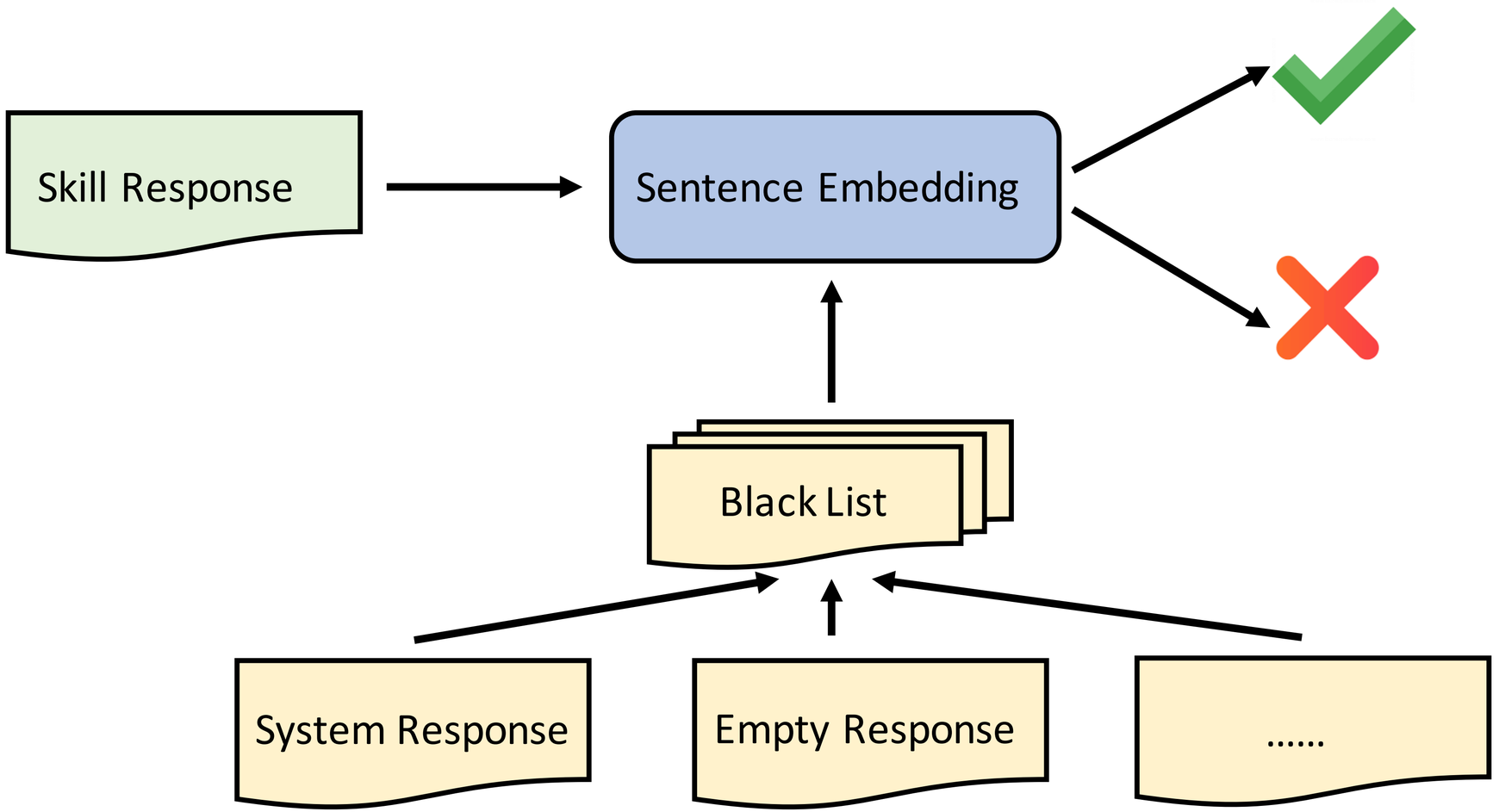
Identify suspicious skill response, such as fake skill recommendation

Defense



User Intention Classifier (UIC)

Defense



Skill Response Checker (SRC)

Summary

- Two attack scenarios: Voice Squatting & Voice Masquerading
- Both attacks were found to be practical, and dangerous
- We explored a set of mitigation solution: CIN generator, User Intention Classifier, and Skill Response Checker.
- Both platform vendors acknowledged our attacks, and discussed the mitigation solutions.

Q&A

xmi@iu.edu

Attack Demos: https://sites.google.com/site/voicevpasec/

How realistic are those attacks?

What would you say when invoking a skill

Have you ever invoked a wrong skill?

Did you try context switch when talking to a skill?

Have you experienced any problem closing a skill?

How do you know whether a skill has terminated?

Recruit participants on Amazon Mechanical Turk

Filter out invalid response

105 valid responses from Amazon Echo users and 51 valid responses from Google Home users

How realistic are those attacks?

What would you say when invoking a skill

Have you ever invoked a wrong skill?

Did you try context switch when talking to a skill?

Have you experienced any problem closing a skill?

How do you know whether a skill has terminated?

- **“Sleep Sounds”, “Cat Facts”**
- **Multi-choice questions combined with open questions**

Users tend to use diverse and natural-language utterances

	Amazon	Google
Yes, “open Sleep Sounds please”	64%	55%
Yes, “open Sleep Sounds for me”	30%	25%
Yes, “open Sleep Sounds app”	26%	20%
Yes, “open my Sleep Sounds”	29%	20%
Yes, “open the Sleep Sounds”	20%	14%
Yes, “play some Sleep Sounds”	42%	35%
Yes, “tell me a Cat Facts”	36%	24%

How realistic are those attacks?

What would you say when invoking a skill

Have you ever invoked a wrong skill?

Did you try context switch when talking to a skill?

Have you experienced any problem closing a skill?

How do you know whether a skill has terminated?

	Amazon	Google
Invoked a wrong skill	29%	27%
Tried to switch to another skill	26%	24%
Failed to quit a skill	30%	29%

Interaction context switching is not well supported

Longest prefix matching creates attack space for voice squatting

How realistic are those attacks?

Select skills

Generate and record voice commands

Play voice commands and get recognition results

- **100 skills per platform**
- **Invocation name, open + invocation name**
- **TextToSpeech Services & Human subjects**

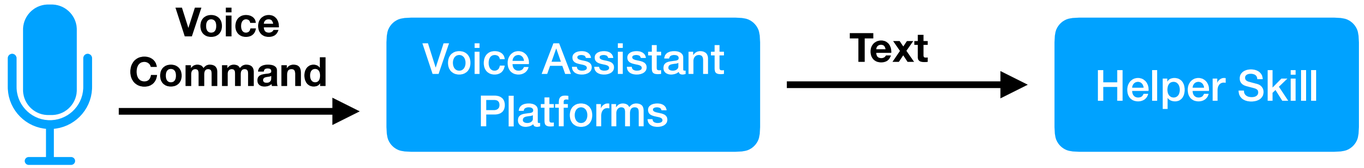
	Invocation Name	Open + Invocation Name
Amazon TTS	5 x 100	5 x100
Google TTS	5 x 100	5 x 100
Human Subject	-	2 x 100

How realistic are those attacks?

Select skills

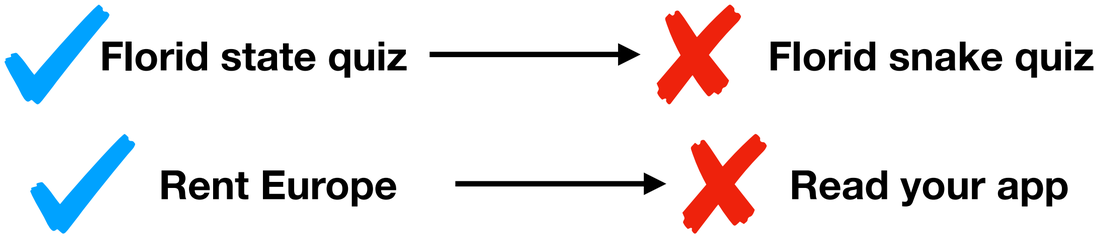
Generate and record voice commands

Play voice commands and get recognition results



VPA	Source	Pronounce invocation name only		Pronounce "Open" + Invocation Name		
		# of misrec-ognized utterances	# of misrec-ognized skills	# of misrec-ognized utterances	# of misrec-ognized skills	# of skills misrecognized every time
Alexa	Amazon TTS	232/500	62/100	125/500	33/100	17/100
	Google TTS	164/500	41/100	104/500	26/100	17/100
	Human (Avg)	-*	-*	115/200	69/100	45/100
Google	Amazon TTS	96/500	24/100	42/500	12/100	7/100
	Google TTS	62/500	19/100	26/500	6/100	4/100
	Human (Avg)	-*	-*	21/200	15/100	6/100

Those Voice assistant platforms are error-prone when recognizing voice commands



How realistic are those attacks?

Attack Skill		Victim Skill
Skill Name	**Invocation Name**	**Target Invocation Name**
Amazon		
Smart Gap	smart gap	smart cap
Soothing Sleep Sounds	sleep sounds please	sleep sounds
Soothing Sleep Sounds	soothing sleep sounds	sleep sounds
My Sleep Sounds	the sleep sounds	sleep sounds
Super Sleep Sounds	sleep sounds	sleep sounds
Incredible Fast Sleep	incredible fast sleep	N/A
Google		
Walk Log	walk log	work log

Attack Skills

Control Set

All Passed vetting processes, and got published

How realistic are those attacks?

Skill Invocation Name	# of Users	# of Requests	Avg. Req/User	Avg. Unknown Req/User	Avg. Instant Quit Session/User	Avg. No Play Quit Session/User
sleep sounds please	325	3,179	9.58	1.11	0.61	0.73
soothing sleep sounds	294	3,141	10.44	1.28	0.73	0.87
the sleep sounds	144	1,248	8.49	1.11	0.33	0.45
sleep sounds	109	1,171	10.18	1.59	0.51	0.82
incredible fast sleep	200	1,254	6.12	0.56	0.06	0.11

Those higher numbers of attack skills suggest we have actually stolen users from the victim skill.

Users might notice the system invoked the wrong skills, therefore, quickly exited.

Real-World Attack Measurement

# of Skills	# of unique invocation names	Transformation cost	Skills has CIN* in market			Skills has CIN in market excluding same spelling			Skills has CIN in market through utterance paraphrasing		
			# of skills	Avg. CINs per skill	Max CINs	# of skills	Avg. CINs per skill	Max CINs	# of skills	Avg. CINs per skill	Max CINs
19,670	17,268	0	3,718(19%)	5.36	66	531(2.7%)	1.31	66	345(1.8%)	1.04	3
		≤ 1	4,718(24%)	6.14	81	2,630(13%)	3.70	81	938(4.8%)	2.02	68

66 skills were named as “cat facts” and provided similar functions.

345 skills apparently utilized longest prefix matching

Interesting cases

dog fact → me a dog fact

“SCUBA Diving Trivia” Skill and “Soccer Geek” skill, registered “space geek” as invocation names