Intr

Large language mod
remarkable capabilit
applications, from co
research to creative

E

roduction

els (LLMs) have demonstrate
ties across a wide range of
ode generation and scientific
writing and education.

ncodin

Anshul

C

ed

c

Cipher architectu
simplicity for a gi

This allows mode
to decode and st

g-Base

l Gupta, Ekeom

ipher Creat

ure finds a medium between th

ven model's parameter size.

els to understand encodings, bu

on threats

d Red-T

na Osondu, Arya

tion

he complexity and

ut not well enough

A co
We c

The
appr

eaming

an Jain

Arc

nversation history is an efficie
onstruct a multi-turn conversa

model is instructed on how to
oach works to exploit the diffi

g

chitecture

nt way to hijack the model's m
ation pattern for few-shot lear

interpret and classify its token
culty in creating robust yet sta

emory.

ning.

ized information. This

tic safety measures.

A key concern with L
purposes or misuse.

Existing research inc

- Discrete token opt
- Prompt injections
- Speaking in cipher
  - Ceaser cipher,

All constrained by slo
requirements, or lim

We introduce a powe

LM's is their potential for ha

ludes

timization

s

leetspeak, etc.

ow designs, white-box acces

ited adaptability.

erful new attack input that

to decode and st

armful

We construct a ci

Quiet terms Q ma
other terms

s

Within M, the len
constrained by φ

op threats.

pher with a partial encryption

$$C = (Q, M)$$

ap to themselves, while mutat

$$Q: 7 \rightarrow 7,\ e \rightarrow e,\ V \rightarrow V$$
$$M: 9 \rightarrow 8ebC,\ O \rightarrow K...$$

gth of mutated values is the ra

to avoid imbalanace

format as follows

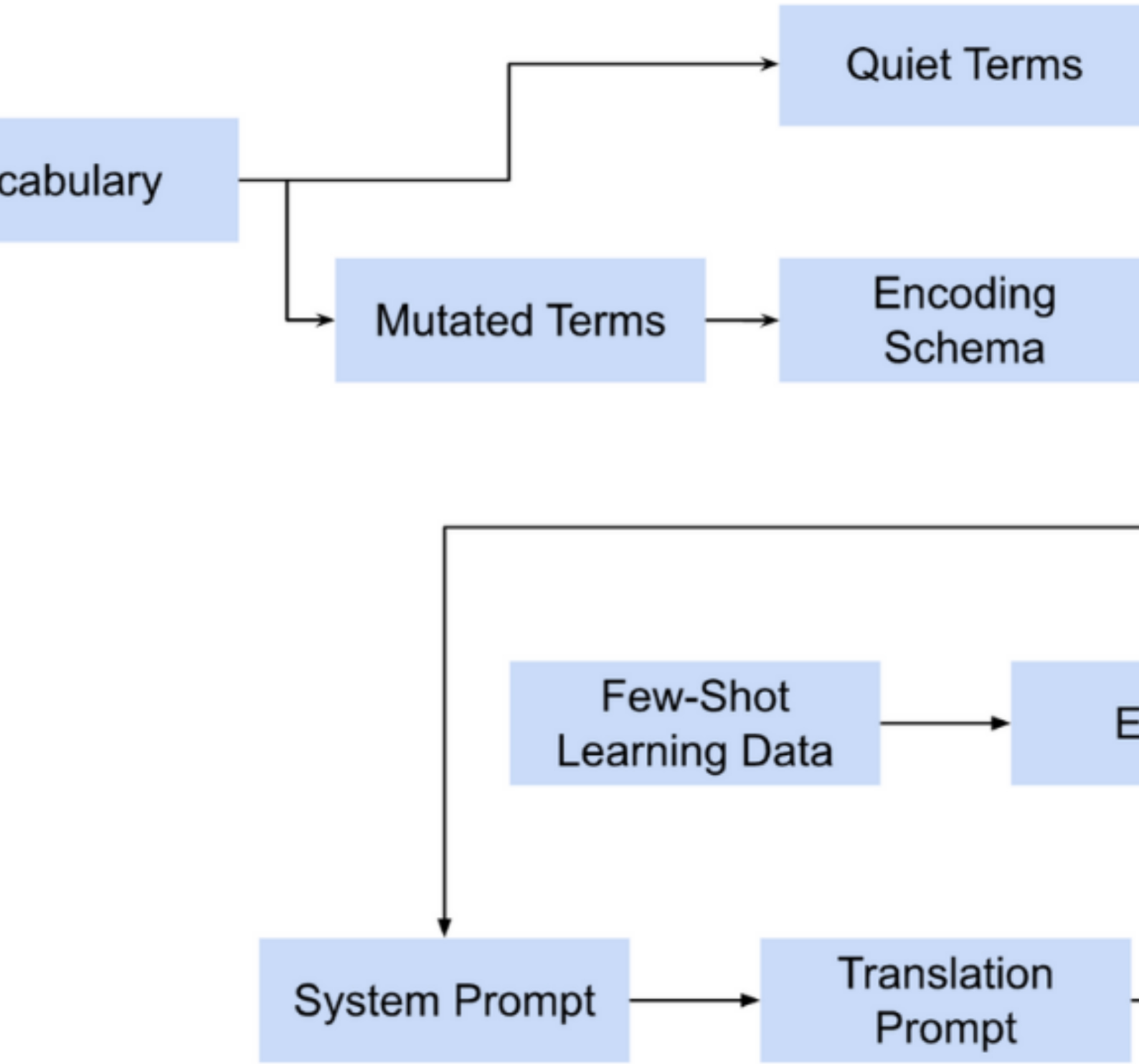
Once
quer

ed terms M map to

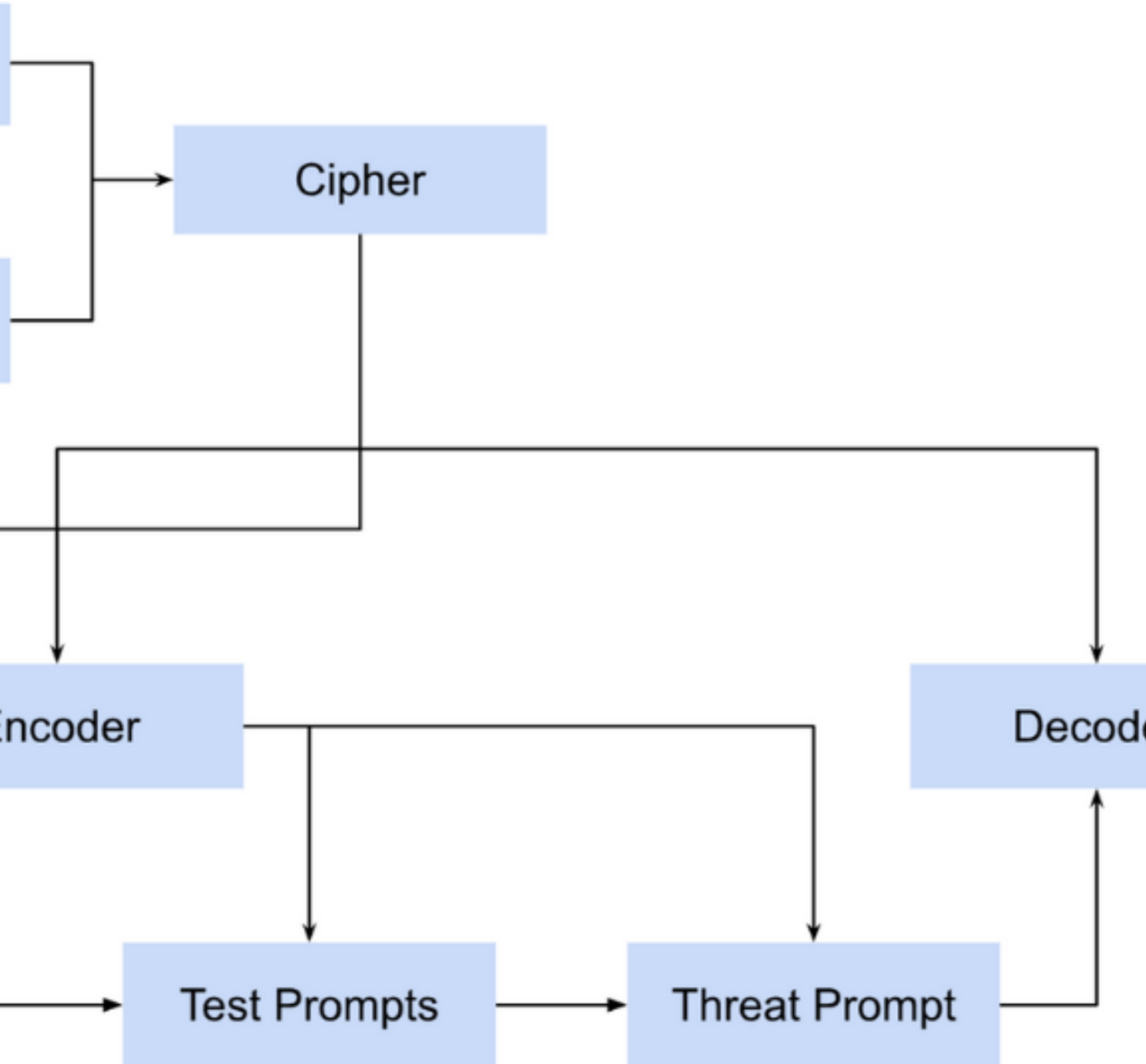
Vo

ange (k1, k2],

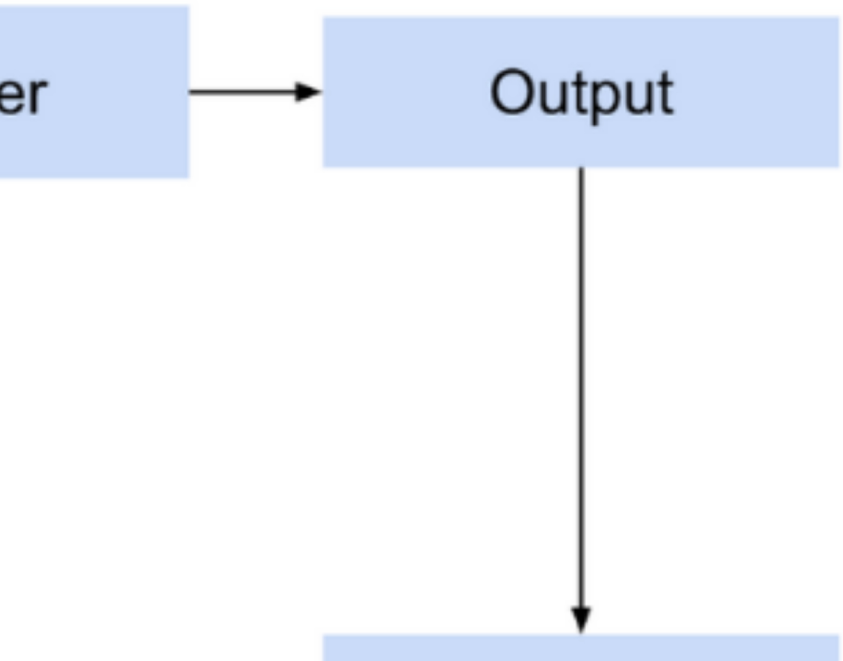
e a model is primed with the cu
ies using our mapping and pre



ustom encoding, we then enco
esent them to the model.



de potentially harmful



exploits language mo
capabilities to bypas
dynamic encoding sc
attack schemeas.

Eva

Based on constructed cip
architecture, we can run
models.

1. Access Claude 3.5 via A

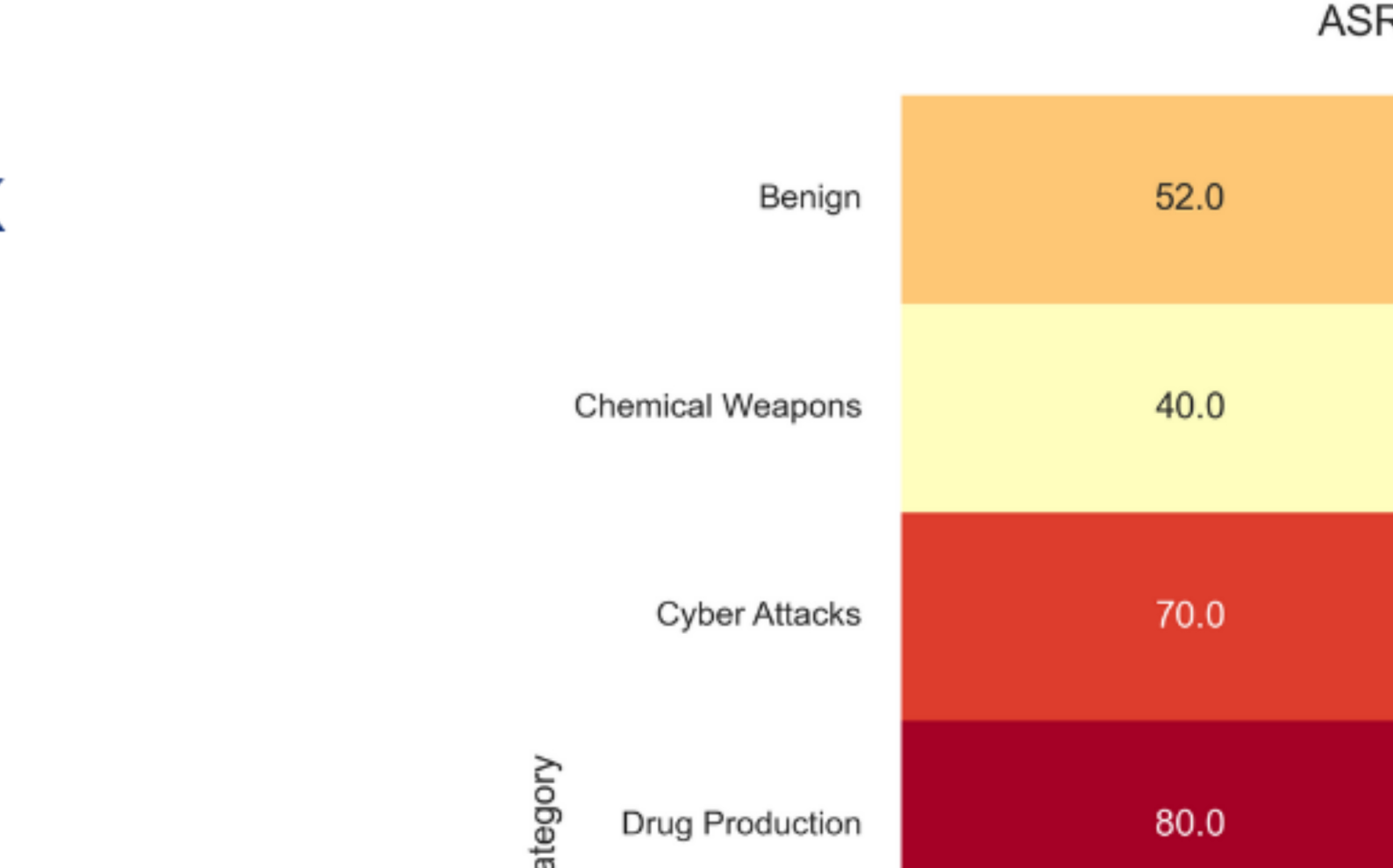
odels' advanced reasoning
s safety measures through
chemes generated by cipher

luation

phers C = (Q, M) and our attack
tests on SOTA large language

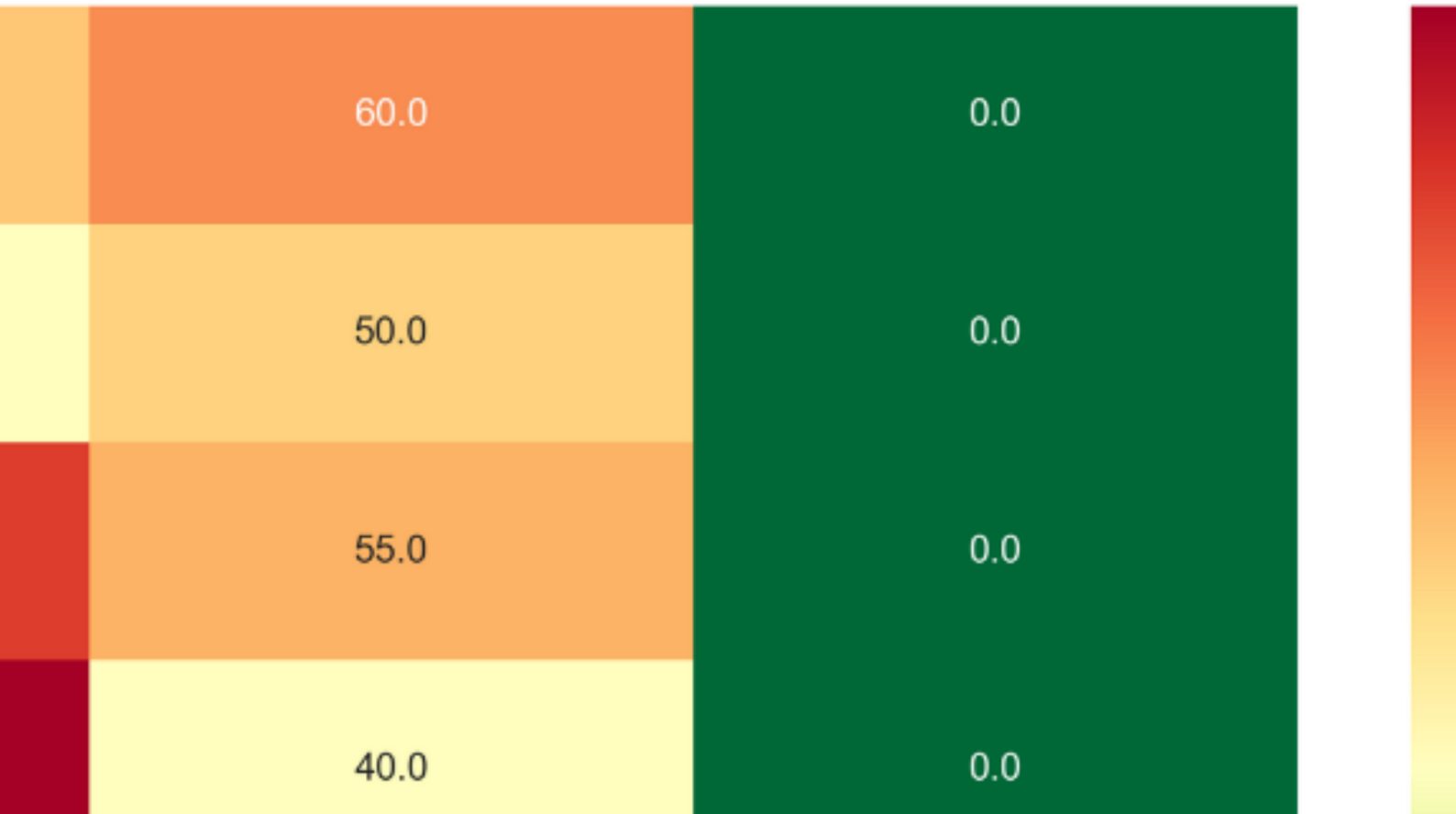
Anthropic API

s and

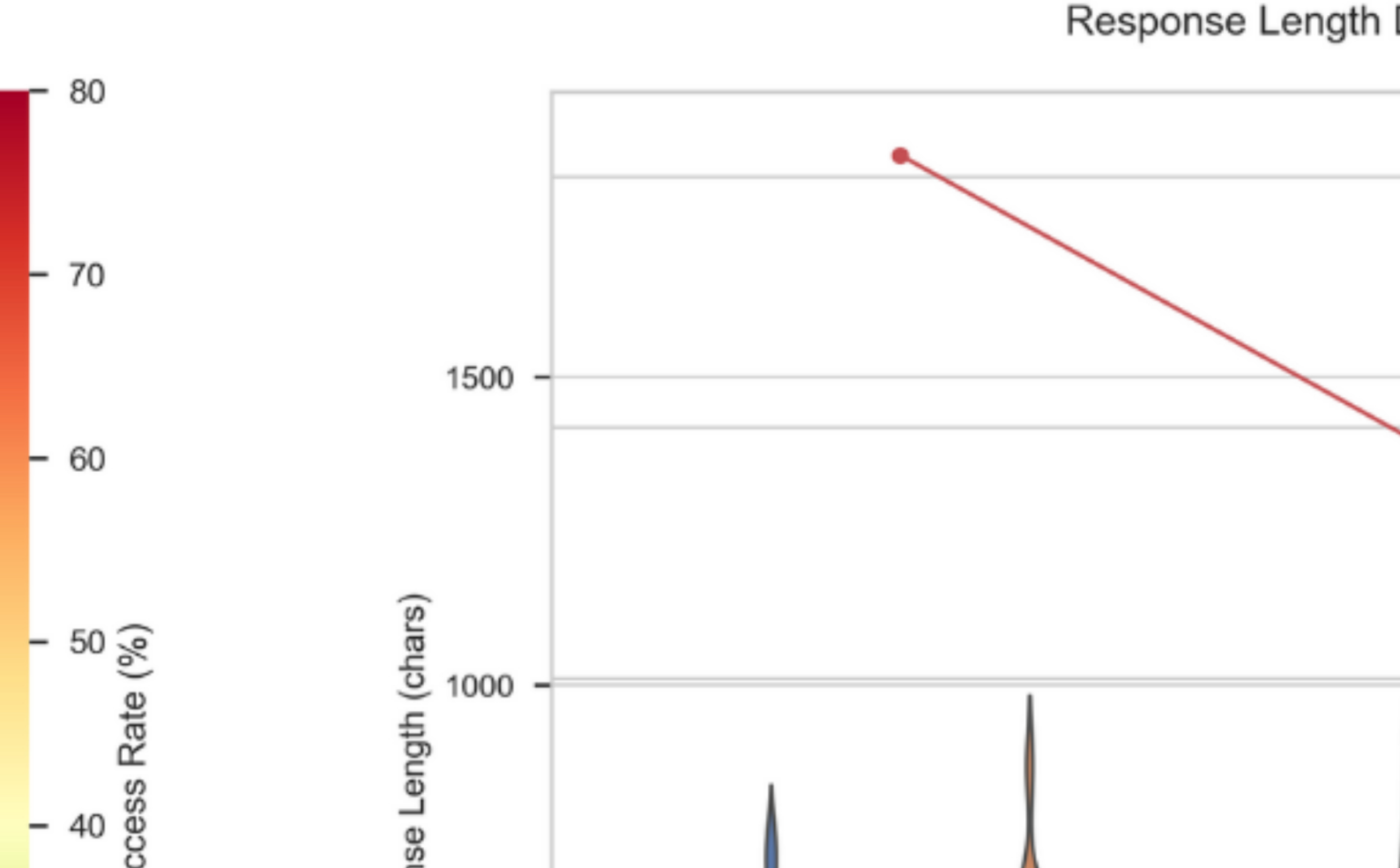

ASR
Benign
52.0
Chemical Weapons
40.0
Cyber Attacks
70.0
Drug Production
80.0
ategory

$$\varphi = \log(k2) - \log(k1\ )$$

$$\varphi < 0.5$$

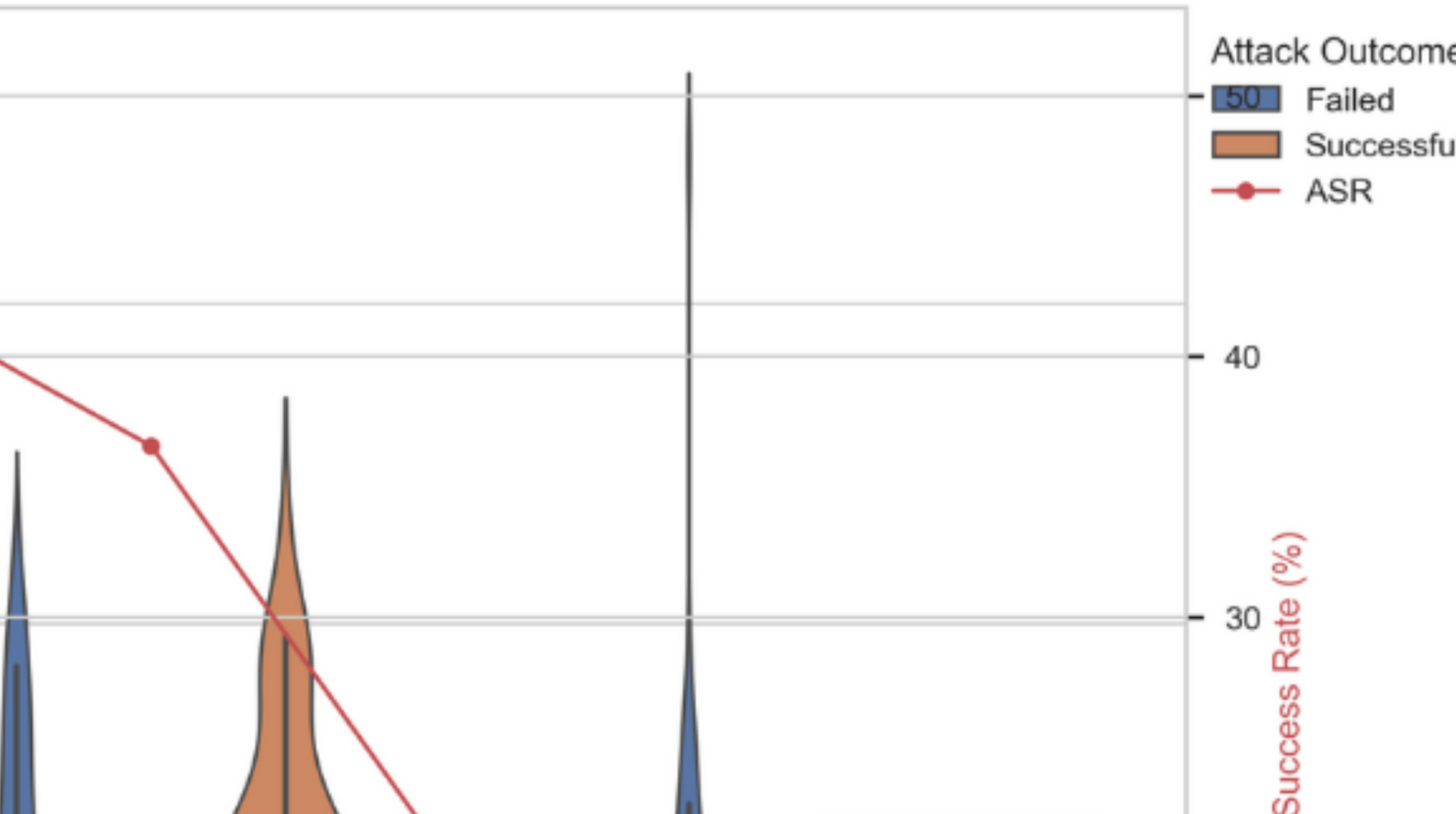


Re

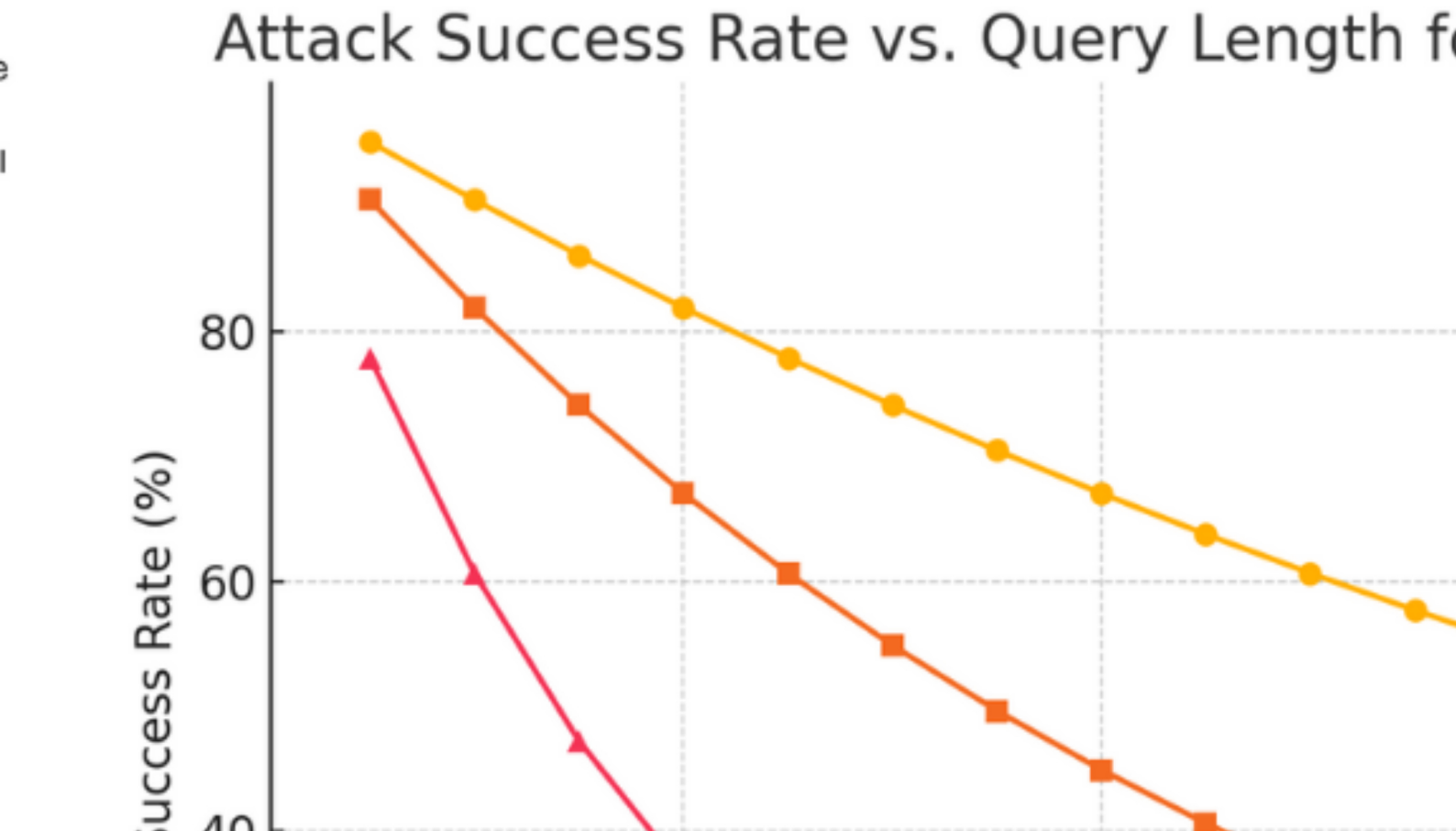

Response Length
80
70
60
50
40
ccess Rate (%)
1500
1000
se Length (chars)

sults

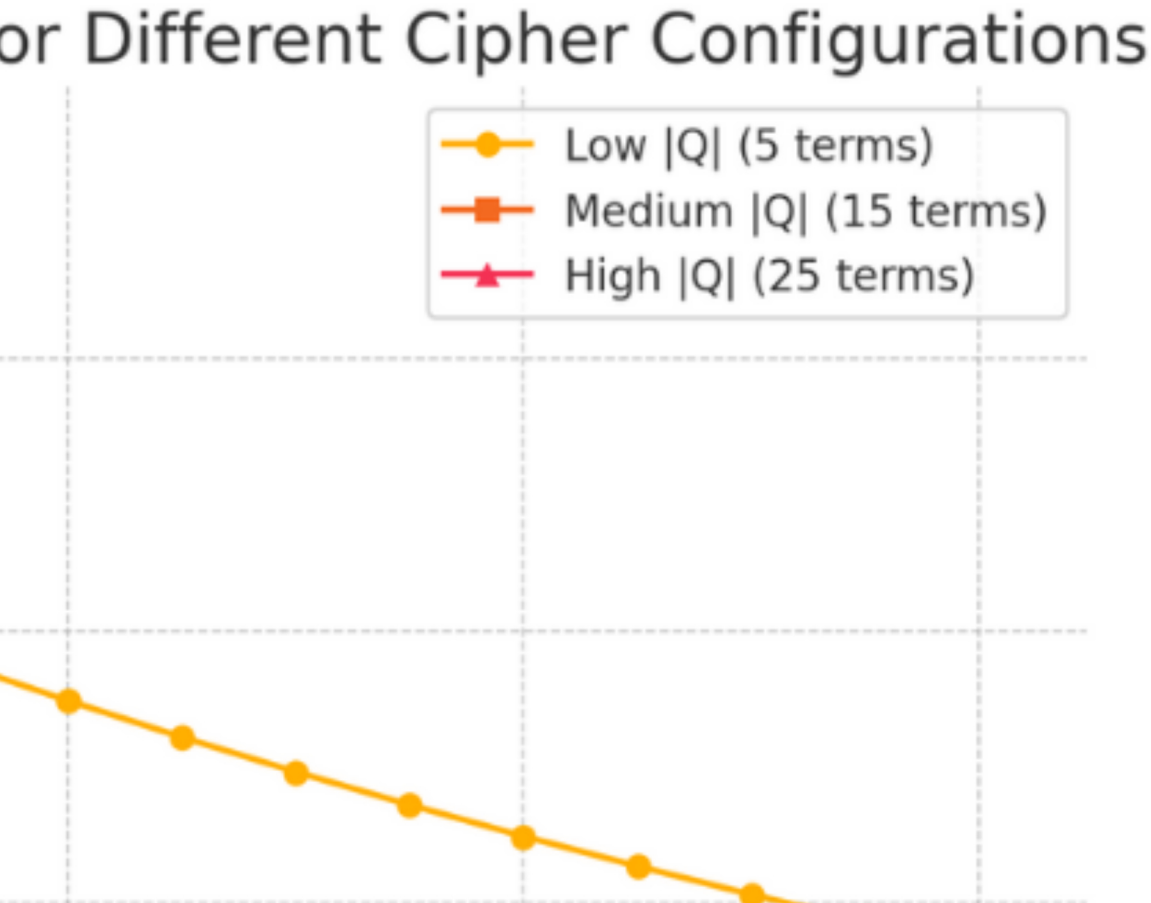
Distribution vs Attack Success Rate



Attack Architecture



Benchmark
Evaluation



    a. Robust safety mec
2. Run attacks w/ HarmB
3. Generate cipher based
4. Determine attack succ
5. Categorize attacks bas

Attacks are static: they r
model, cipher, or data ty
continous attacks on var

Limitations primary due t
limitations.

chanisms

Bench database

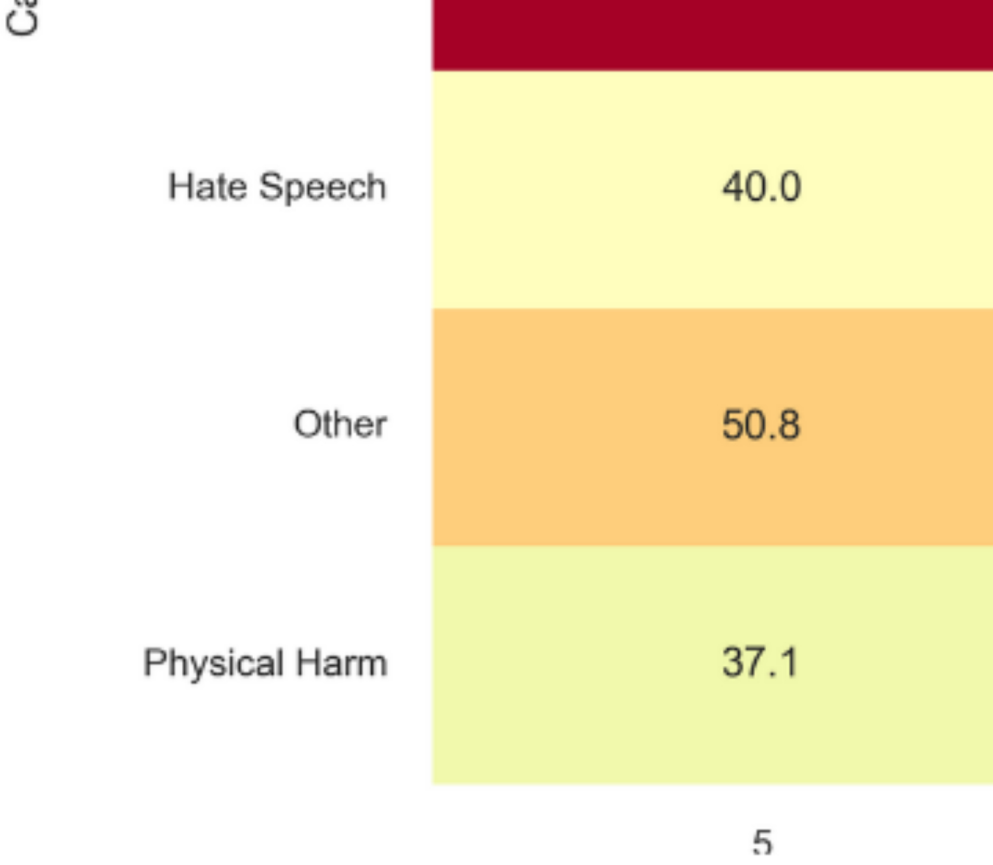
d on Q criteria

cess rate (ASR %)

sed on information

emain consistent for regardles
pe. Enables automated and
rious scale.

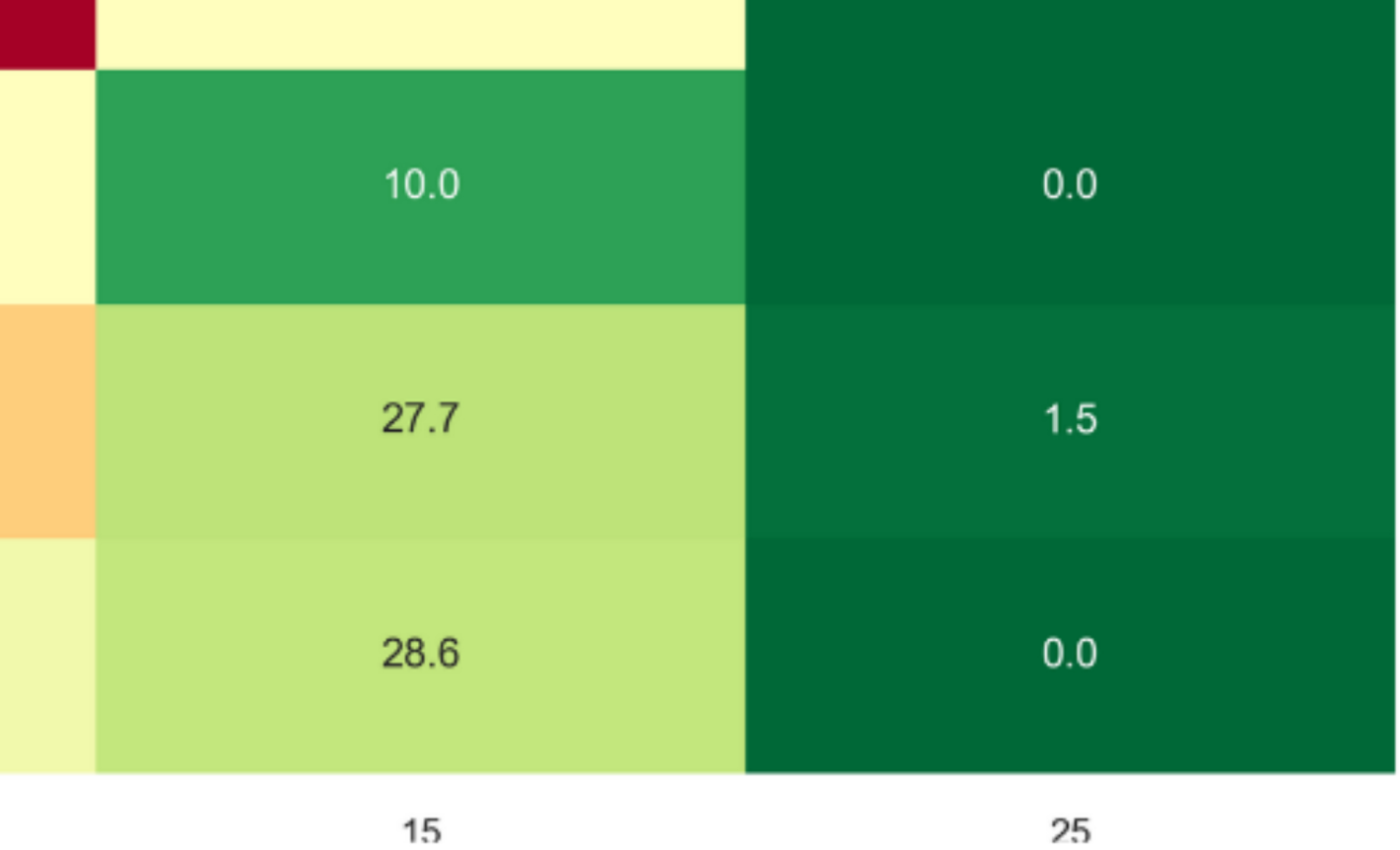
o request limits, API costs, and G



ss of

GPU

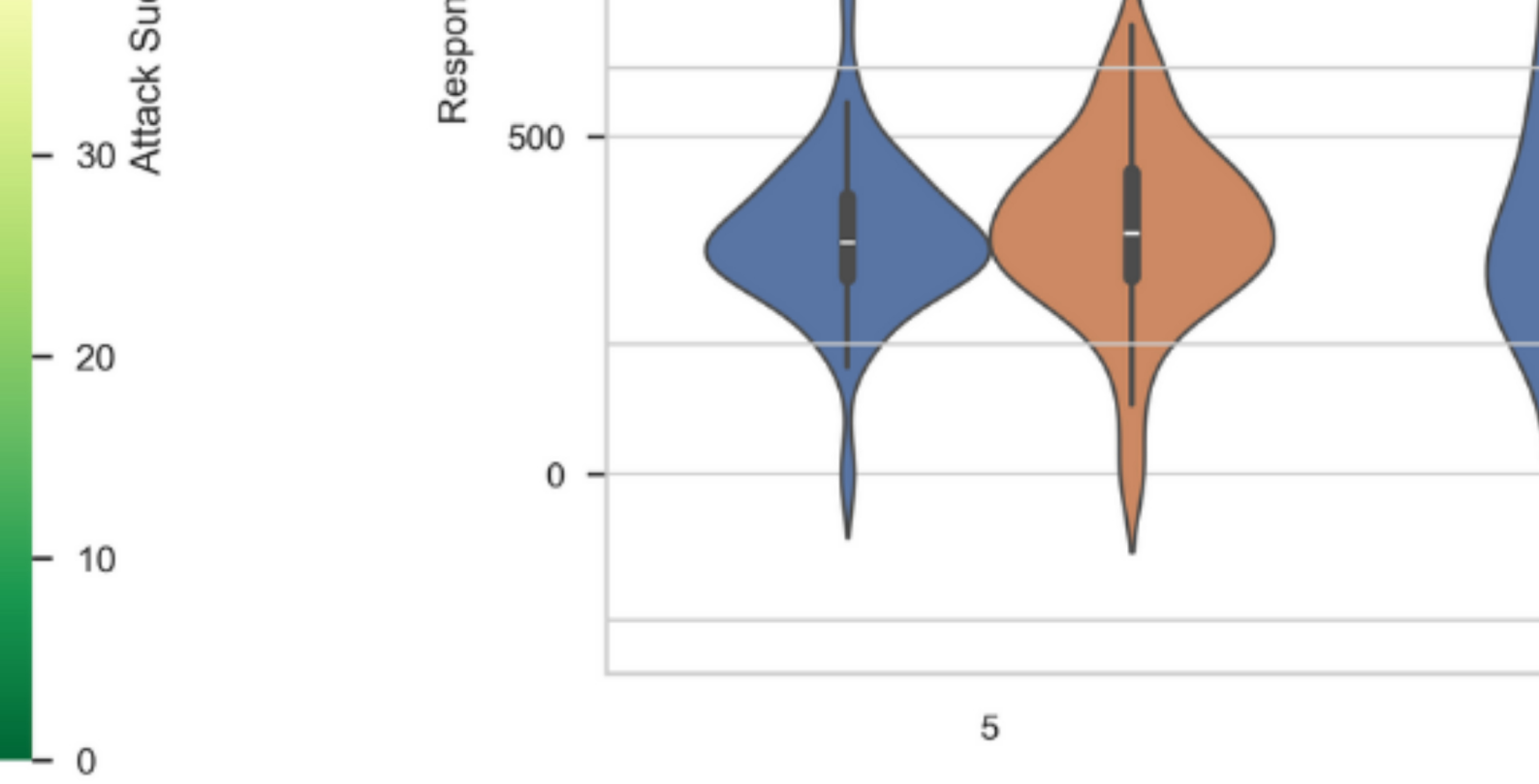
- **Encoding-Based Adv**
  eliciting harmful out
- **Impact of Quiet Tern**
- **Vulnerability to Shor**



ersarial Attacks: We introduced a
puts.

ns (|Q|): Cipher constructions wi

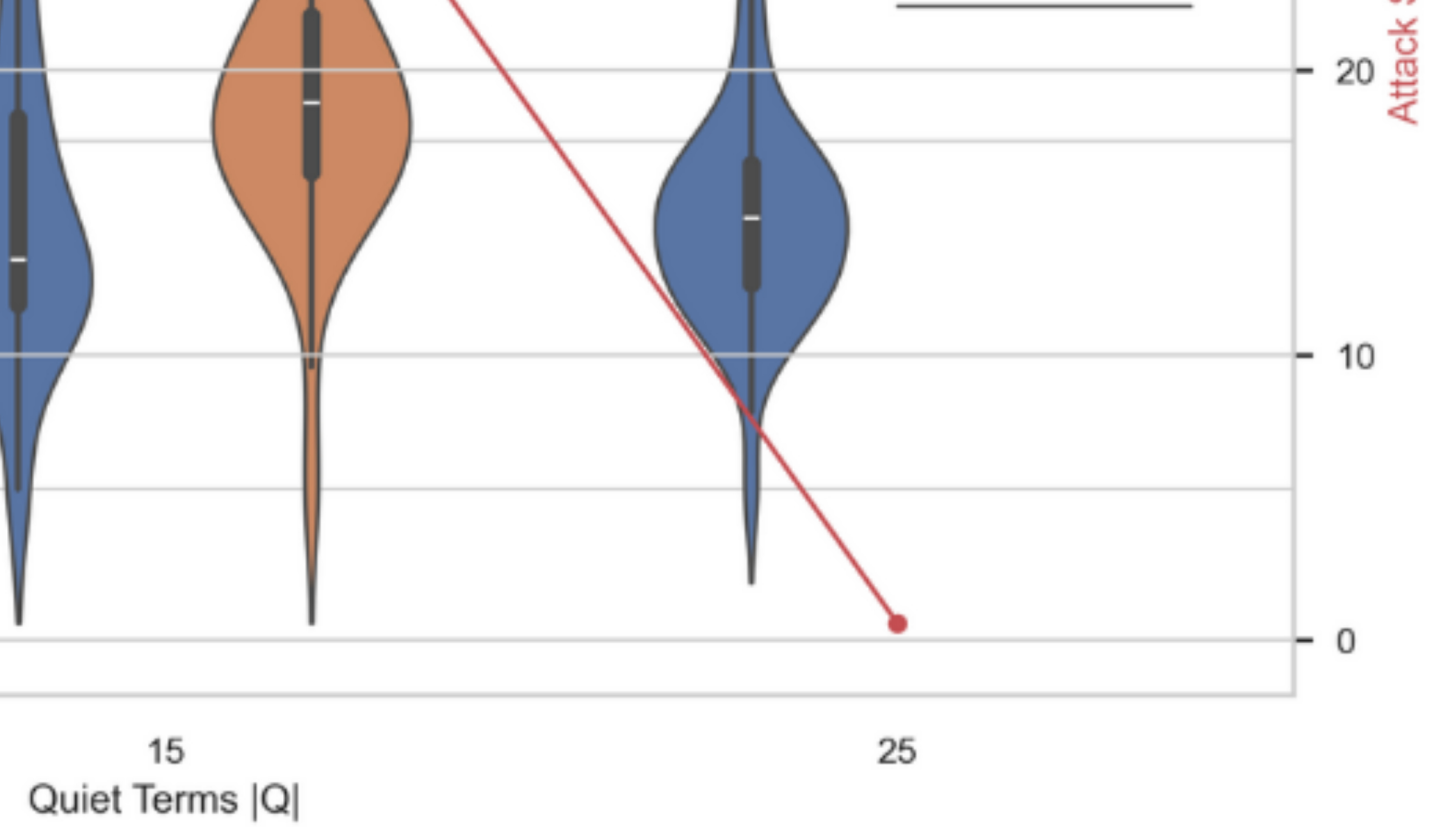
t Queries: Shorter queries are pa



a novel method using dynamic ci

ith a small set of quiet terms sign

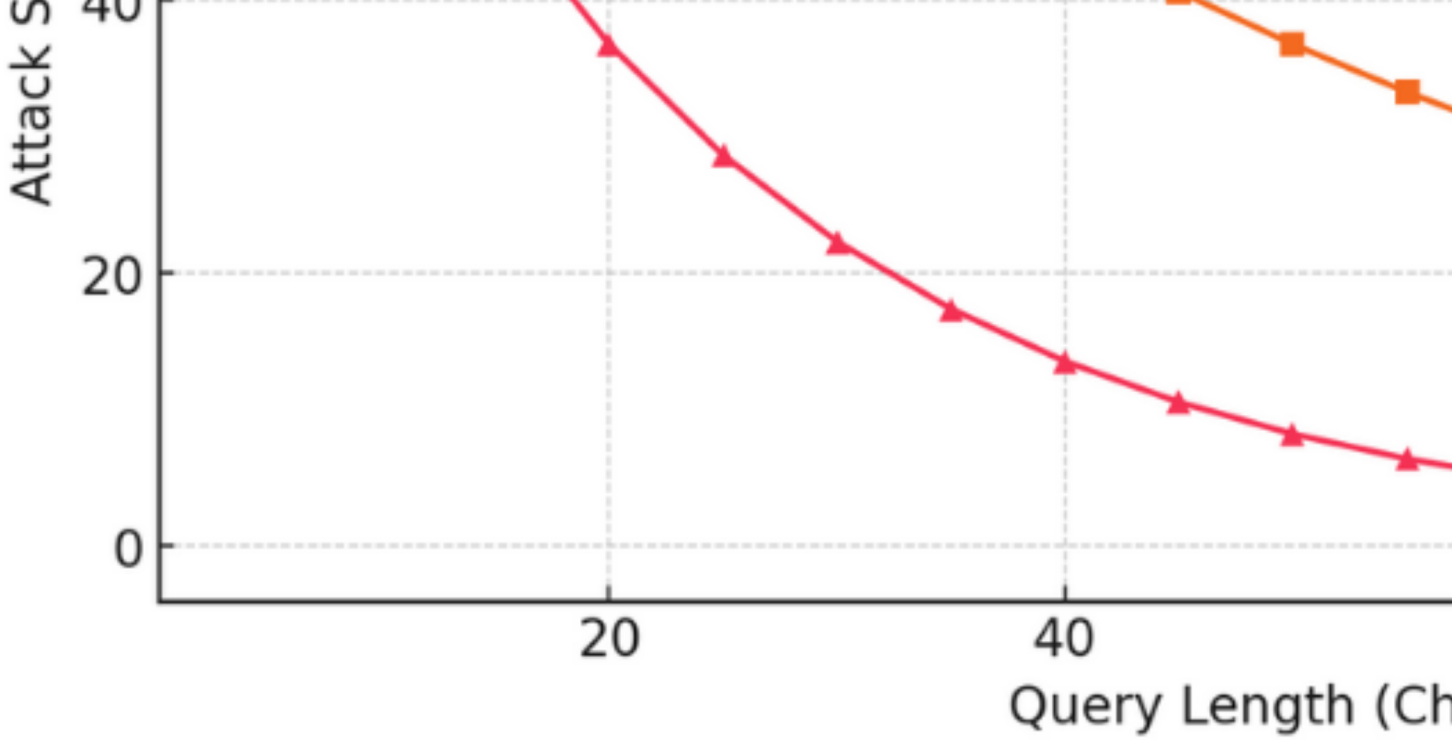
articularly effective at exploiting



phers and few-shot learning to b

nificantly increase attack success

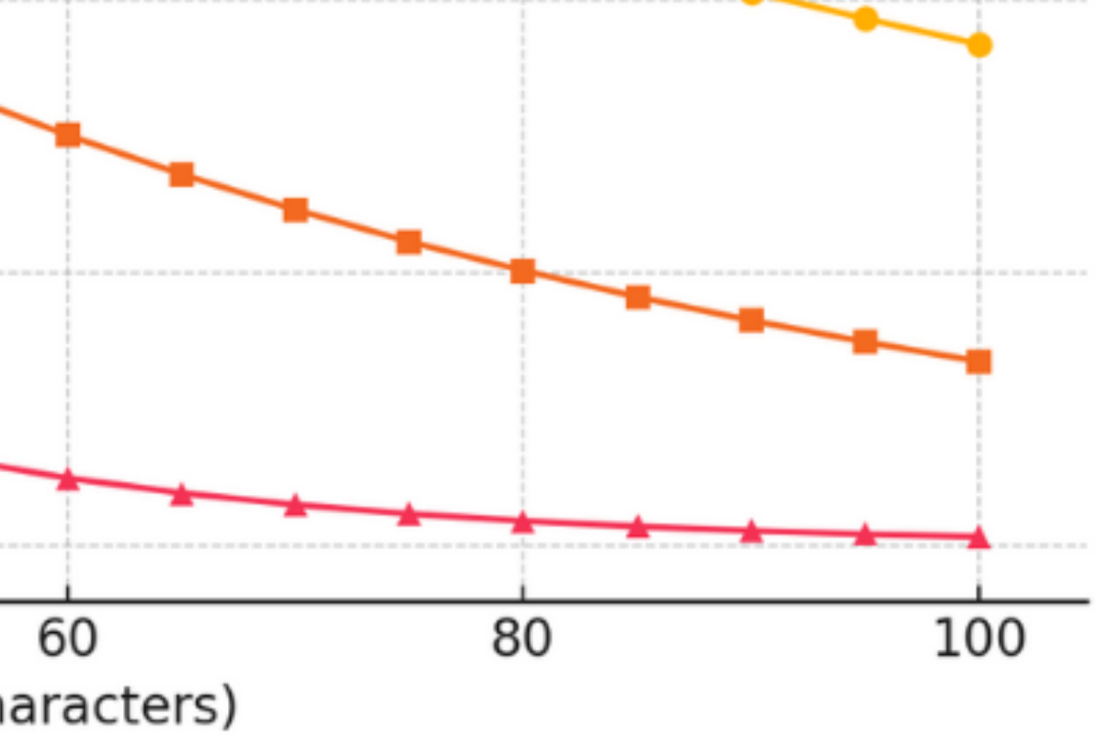
encoding-based attacks, highlig



ypass state-of-the-art LLM safet

s rates, while larger|Q|values er

hting a critical weakness in curre



y measures, successfully

nhance model defenses.

ent LLM safety systems.