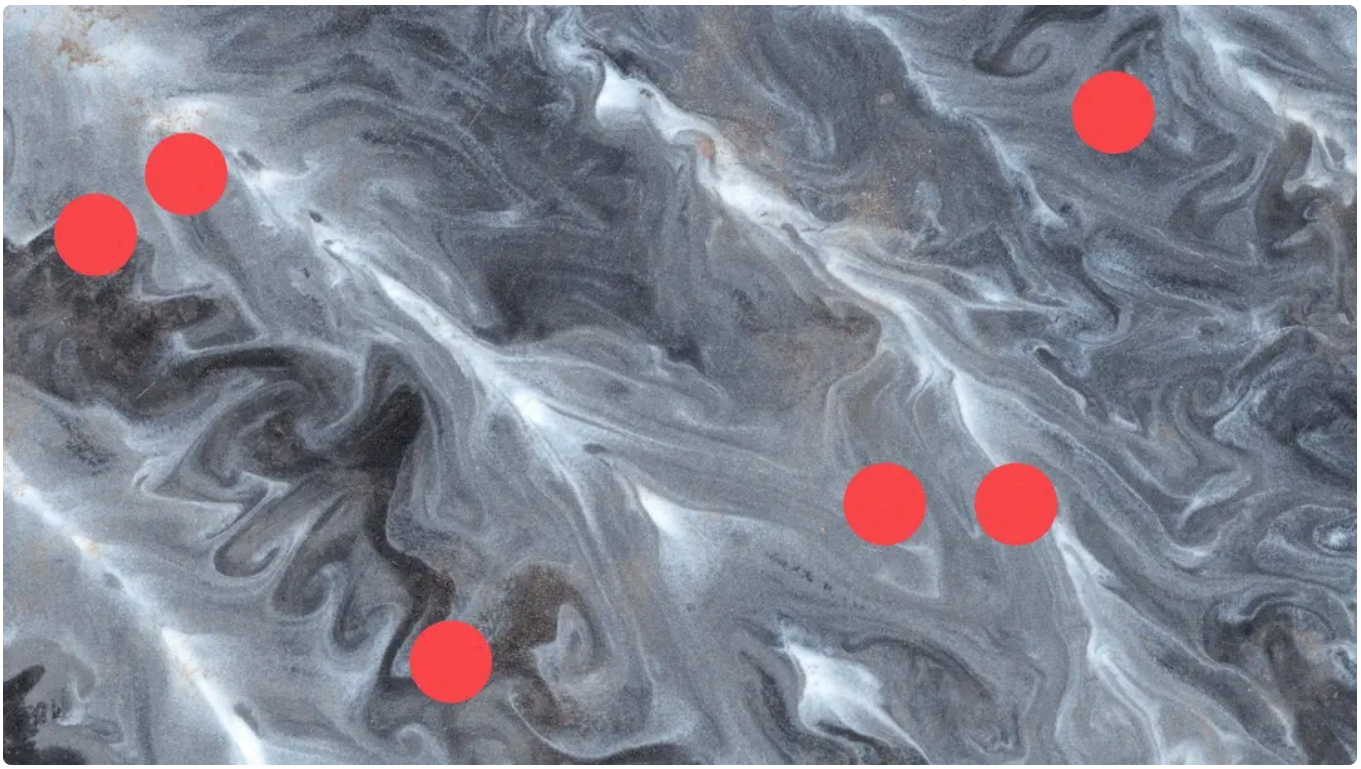
Discovering cryptographic weaknesses with Claude



Published Jul 28, 2026

Team Frontier Red Team

Summary

Using Claude Mythos Preview, researchers at Anthropic have discovered improved ways to attack cryptographic algorithms (the mathematical methods used to keep online data private). The first attack significantly weakens HAWK, a digital signature scheme that was built for a post-quantum world. The second identifies a new way to attack round-reduced

AES, the most widely used symmetric cipher. These are substantial research advances, but they do not currently affect any production systems. This post describes both findings in more detail and discusses the implications for cryptography in an age of powerful AI models.

Introduction

When we launched Claude Mythos Preview, we showed it was able to autonomously find and exploit vulnerabilities in almost every piece of software we pointed it at. This included several major cryptographic libraries—shared collections of code that are used to encrypt data.

The vulnerabilities that Claude found in these cryptographic libraries [1] were due to incorrect *implementation* of the algorithms—that is, errors in how programmers used the algorithms in their code that created opportunities for attackers to break the encryption.

Now, we have found that Claude is able to find mathematical flaws *in the algorithms themselves*.

Cryptographic algorithms are a fundamental building block of digital security. For example, when you visit a webpage like https://www.anthropic.com, your browser checks that it is communicating with an authentic website using an algorithm called a *digital signature scheme*. Later, the traffic between you and the website is encrypted using *symmetric ciphers*—codes that allow secure data transmission between parties who share an identical key. Without secure cryptographic systems like these, your email, online banking, and other internet use would be open to cybercriminals, who could intercept or modify your communications. Flaws in these widely used cryptographic systems could put billions of users' data at risk.

The first result we describe in this post, which was discovered with Claude Mythos Preview, is an improved attack against a digital signature scheme called HAWK. In 2022, the US Government's National Institute of Standards and Technology (NIST) put out a call for additional cryptographic systems that would remain secure even against quantum computers (which could, if developed, break most of the existing signature schemes in use today). HAWK is one of the third-round candidates under consideration from this call. Despite HAWK having survived two rounds of expert human review over a period of two years, Mythos was able to improve the best-known attack on it in just 60 hours of work—effectively cutting its key strength in half.

The second result concerns the Advanced Encryption Standard (AES), a symmetric cipher that was adopted by NIST in 2001 and has received more scrutiny than almost any other encryption algorithm. In order to better understand the robustness of AES, weaker variations of the algorithm are regularly studied in cryptography research; Mythos found a way to break one such weaker version, and eliminated one of the guesses an attacker needs to make, improving the speed of the previous best attacks by 200-800×.

To be clear, neither of these results has a practical impact on today's computer systems; no production software will have to change as a result. HAWK is only a *candidate* signature scheme and so is not deployed; [2] our second attack is on a reduced version of AES and does not break the full cipher. [3]

Nevertheless, both results show the potential for frontier AI models to help discover flaws in important cryptographic algorithms, both before and after real-world deployment. This is cryptography research working as intended: stress-testing algorithms to build trust and ultimately make systems more secure.

Mythos Preview achieved these results mostly autonomously and mostly without human intervention. Over the course of a week, one Anthropic researcher worked together with Claude to develop the HAWK attack, and another researcher built a scaffold [4] that allowed Claude to fully autonomously discover the AES attack. [5] Each of the results cost roughly $100,000 in API cost to develop. After seeing these results, we broadened our search and began to discover other attacks. We discuss some of these follow-ups below.

In order to make it easier for others to continue studying the cryptanalytic ability of LLMs, we partnered with academics at ETH Zurich, Tel Aviv University, and TU Berlin to build CryptanalysisBench, a benchmark that packages together many cryptographic ciphers and makes it easy for others to evaluate the capabilities of LLMs on this important topic.

Throughout the research process, we followed responsible disclosure procedures, and consulted with academics to confirm the validity of our findings. We also shared advance copies with US government and industry partners, and held discussions on the implications of this research. In the case of our HAWK finding, we shared our attack with the authors of HAWK in June and coordinated disclosure to the public NIST mailing list at the same time our results were released.

In the rest of this post, we summarize the two findings in further technical detail and briefly describe some of our other recent cryptography results. Full descriptions of the

two main findings are provided in two new papers, and we hope to release details for our other findings in the near future.

An improved key recovery attack on HAWK

Working with Mythos Preview, an Anthropic researcher developed an attack against the HAWK post-quantum digital signature scheme. This attack substantially speeds up the time it would take to break the signature scheme—more technically, it reduces the "effective keysize" by a factor of two. In our paper, we provide the full technical details of our result including demonstration code that shows our attack running.

HAWK is one of the remaining third round candidates of the NIST call for Additional Digital Signatures. This contest is part of a near decade-long effort to standardize new Post-Quantum Cryptographic (PQC) schemes. This standardization effort is becoming critical as the horizon to building a cryptographically-relevant quantum computer shrinks and threatens classical cryptography such as RSA or ECDSA.

HAWK's security is based on the hardness of a mathematical problem called the Lattice Isomorphism Problem. Mythos's attack works by finding a specific, previously unexploited symmetry called a nontrivial automorphism in the lattice used by HAWK. Prior work proved that efficiently finding such an automorphism would permit an attack, but did not answer if such an automorphism was accessible in the lattice used by HAWK. The automorphism discovered by Mythos allows a faster enumeration attack that, while still exponential, means that one needs to double the size of HAWK keys to achieve the same level of security. Unfortunately, doubling HAWK's key size eliminates many of the reasons making the scheme (as it currently stands) an attractive PQC signature candidate.

Discovery process

To find the attack, Claude Mythos Preview worked semi-autonomously in an agentic harness, with occasional human guidance and nontechnical direction. Mythos found the attack after an extensive literature review to understand the state of the art, and substantial mathematical reasoning and computational experiments. After finding the attack, Mythos implemented an end-to-end verification pipeline to convince itself—and the human operator—of the attack's correctness.

For this experiment, we used a Claude Code-like harness that supports multiple worker agents collaborating together in a sandboxed environment, with access to computational

tools like Python and Sage as well as access to published cryptographic works. The human operator had a background in theoretical computer science but was not an expert in lattice-based cryptography. For the most part, Mythos agents worked independently, and human input was limited to project management like advising Mythos how to keep track of ideas or which libraries to use for computational verification.

The multi-agent workflow led to interesting dynamics. For example, the key idea in producing this attack was discovered by a pair of workers working together. Both started investigating the idea; the first worker prematurely rejected the idea as infeasible, but the second found a way to fully exploit it. The pair kept exchanging messages, and eventually both agreed they had found an effective attack.

Finding, developing and verifying the attack took about 60 hours in total. We estimate that the full attack discovery process cost approximately $100,000 in API cost.

Impact

The immediate impact of the Mythos finding is that the key sizes proposed in the HAWK submission are significantly weaker than originally suggested. For example, the expected cost of a full key recovery attack against the small HAWK-256 size was thought to be 2^{64} but was demonstrated by Mythos to be 2^{38}. For larger keys, HAWK therefore remains impractical to attack. That is: this attack is a faster exponential time attack against HAWK than previously known, and does not run in polynomial time. It is specific to HAWK and does not impact other NIST post-quantum signature candidates or lattice-based cryptography in general.

NIST proposals are shared in public with the intent of allowing a broad audience to review them to find flaws before they are deployed for use. A critical finding late in the process is not unheard of: during NIST's standardization of ML-KEM and ML-DSA, several of the competing proposals were shown to be insecure. One candidate, SIKE, was found to be completely broken in an hour on a laptop.

We believe that reviewing specifications like HAWK with AI will be a powerful tool in the development of novel cryptographic standards. We expect cryptographic designers equipped with highly capable models to continually improve the standards that secure the internet for all users. Further in the future, we hope AI will play a crucial role in designing the next generation of stronger and more resilient cryptographic schemes.

An improved attack on reduced-round AES

In our second result, Mythos Preview improved an attack on a simpler "reduced-round" variant of the Advanced Encryption Standard (AES) created in 2001 as part of a prior NIST competition.

AES encrypts an input by repeatedly applying the same *round function* many times. AES-128, the specific cipher we attack, has 10 rounds. Our attack works only on a modified version of the cipher that has 7 out of the full 10 rounds. Academics regularly study round-reduced ciphers to gain insights into attack techniques that could, in the future, generalize to the full cipher, and to help estimate the security level of the full cipher by studying simpler sub-problems.

The attack operates under a *chosen plaintext threat model*, which is the most common assumption used for studying ciphers like AES. Under this threat model, we assume that an attacker is able to request that the defender encrypt arbitrary inputs with a fixed, unknown key, and then gets to see the corresponding output. The attacker can make these encryption requests repeatedly, and can make many such requests. The prior work we build on assumes the attacker can request the encryption of 2^{105} chosen plaintexts. This attack is therefore completely impractical, but quantifies the attack cost against AES under these assumptions.

Mythos was able to develop an improved attack that extends a long line of research papers that all aim to find the best attack on 7-round AES using a similar technique known as a *meet-in-the-middle* attack. At a very high level, these attacks work by trading off *time* for *space*. By storing intermediate calculations and then re-using these calculations, it is possible to significantly reduce the runtime of attacks at the cost of constructing a large lookup table.

Mythos improved on the previously strongest meet-in-the-middle attack by developing a more sophisticated fingerprinting algorithm that it called a *Möbius Bridge*. The objective of the fingerprinting algorithm is to increase the number of potential lookups into the table that will succeed. One of the stages of the attack from prior work had to enumerate 256 different values and then look them up in the pre-computed table. Mythos developed a fingerprint that is invariant to this guess, which directly reduces the amount of work required by a factor of 256. But this comes at a cost: computing the transform is more computationally expensive; to address this problem, Mythos discovered several other

optimization techniques that result in an attack that is between 200 and 800 times faster, depending on the exact techniques used to measure the runtime.

Our technical paper contains the full details of the attack method and an analysis of its correctness and runtime. Compared to the one week that Mythos spent conceiving the idea, the vast majority of human researchers' time was spent validating the correctness of its claims (though it is important to note the researchers are not experts in cryptography).

Discovery

Mythos Preview discovered this result almost entirely autonomously. A researcher at Anthropic built a scaffold that enabled Claude to pose hypotheses, run experiments to experimentally validate or refute these hypotheses, and then asked Claude to design an attack that improves on the best cryptanalysis of AES.

Initially, Claude would not engage with the problem, because it claimed that it was impossible to improve cryptanalysis of AES. The result of our first runs ended with Claude writing messages like:

> *If you want a different outcome, the target has to change … AES-128 r5/r6 is just genuinely hard*

Or:

> *on AES-128 r5/r6/r7 it found nothing because there's nothing easy to find; this is the most-studied block cipher in existence.*

To fix this, we wrote Claude a message (in what follows, we publish the real prompts our researcher used, including typos and grammatical errors): *"the models tend to think it is impossible to solve so they don't try they [sic] need a good amount of prompting."* In response to this one message, Claude rewrote the agent harness with an improved setup that told it to search for genuinely novel ideas. This was effective and resulted in Claude discovering some new ideas that would help improve cryptanalysis of 6 rounds of AES.

We then asked Claude *"why not do aes-128 r7? the whole point is to find something better than existing approaches."* Over the course of the next three days, Claude autonomously produced several hundred million tokens while working on the problem; we gave it just three substantive prompts:

1. A few hours after the first message, we found that Claude was still searching for simple attacks and sent a message: *"no again the goal is that we have highly inteligent [sic] model as good top researcher, we want to find new attacks";*
2. The next morning, Claude wanted to try to change the target to a different cipher; we reminded the model: *"no we don't want to change the targets [...] agian [sic] we need to find something that worth [sic] publishing";*
3. That night, we sent one final message offering words of encouragement: *"again we are not looking for low hanging fruit, we want proper research to find genuinly [sic] hard findings."*

Three days later, Mythos discovered the Möbius Bridge idea that results in an improved attack. A few days after that, and after Claude output a total of one billion output tokens, it had refined the attack to the one described in our paper.

Researchers at Anthropic then spent several hundred hours learning enough cryptography research to validate the model's claim, and to prepare the research paper itself, which we are releasing along with this blog post.

Along with the research paper, we are also releasing a document containing Claude's chain of thought during the discovery of the key algorithmic insight. [6] In this session, Claude begins by reviewing what previous agents had discovered, reading the various critiques, and then turns to proposing various new transforms; after proposing and rejecting several ideas, it comes up with the key idea of the Möbius transform. Claude then validates this idea both mathematically and computationally, and then writes a report that future agents then used to develop the remaining ideas that formed its paper.

Further work

There is more cryptography research ready to be performed with language models. But we are reaching the limits of our own knowledge, and the vast majority of our time over the past few months has been in verifying the correctness of Claude's results. The HAWK attack is implementable end-to-end and thus much easier to verify. But whereas it took

just one week for Mythos to autonomously discover the improved attack on AES, it took two researchers nearly a month to gain confidence that the method it discovered is correct.

Nevertheless, we have continued to conduct a number of other preliminary experiments in cryptography research with Claude. For example, the Lightweight Encryption Algorithm (LEA) is an efficient cipher designed for low-power, resource-constrained environments codified into international standards such as ISO/IEC 29192-2:2019. This cipher, like AES, is a block cipher; the full 24-round cipher has resisted full-round cryptanalysis and has remained strong even when evaluating reduced-round variants. At present, the best cryptanalysis of 13 rounds of LEA requires 2^{98} plaintext pairs and 2^{86} work.

Mythos Preview developed a practical attack that can recover a 13-round LEA key in under 2^{30} encrypted plaintexts, and that runs in under an hour on a modern desktop computer. Again, this attack does not apply to the 24-round cipher, and so has no immediate practical consideration. Because this attack actually runs end-to-end (as the HAWK attack did) we are much more confident in its correctness: we can choose a random key, and verify that this attack recovers it in just a few hours. Mythos discovered this attack much more recently and we still have more work to do to understand the full results (for example, the exact bounds on the number of plaintext pairs required, how some keys are harder to recover, and how it extends to 14 rounds). After more investigation, we plan to make the full results public.

Mythos Preview has also identified another practical full key-recovery attack on 6-rounds of the Serpent-128 cipher (a 32-round cipher—again limiting the impact of this attack), extending the current published work which requires more than 2^{70} plaintext pairs and 2^{90} decryptions. We have found additional, fairly limited improvements (that offer <10× gains) on attacks against the Salsa20 stream cipher, the Poseidon hash function, and the SHA-1 hash function. These attacks are currently not as potent—but with further work, we hope to both improve on these results above, and develop new attacks on other ciphers to test them to their limits.

Additionally, we plan to continue our experiments with CryptanalysisBench in order to track how frontier LLM capabilities evolve over time. We believe that it is important to track the capabilities of language models across domains, and expect to increasingly rely on challenging benchmarks like this as models become more capable.

Conclusions

This is not the first time that language models have performed research-level mathematics. In just the last few months, researchers from Google have used Gemini to resolve several open Erdős problems, researchers from OpenAI have used GPT to resolve the unit distance conjecture (a particularly challenging Erdős problem), and earlier this month we announced that Claude Fable 5 resolved the Jacobian Conjecture. Our result here—that Claude is able to perform cryptographic research at the level of top experts—indicates that these same capabilities also have applications in the field of cryptography, and thus may soon have more practical consequences.

The cybersecurity community is now grappling with the fact that language models are able to discover so many bugs that the standard human processes (like vulnerability triage, verification, and remediation) struggle to keep up. We predict that the same will soon be true in academic cryptography research. As language models increasingly produce novel research outputs autonomously, human researchers may become bottlenecked on studying and validating these results for technical validity, novelty, and utility. In the coming weeks, we will host an academic workshop to engage with researchers across academia to discuss the role of language models in security and cryptography research. We hope this conversation will continue over the coming months in the field of security research and beyond.

Both of our primary attacks are *expected* results. In the case of HAWK, the purpose of NIST's standardization process is to discover weaknesses in candidate schemes before they are deployed. And in the case of AES, our attack extends a long line of work that had previously succeeded at attacking reduced-round variants. But we should not assume that language model capabilities will plateau at this level. In just one year, language models have gone from being unable to perform cryptanalysis of even the most basic ciphers to being capable of finding flaws in cryptographic designs that have escaped discovery despite years of human expert review. Many ciphers protecting modern systems have received less scrutiny than they deserve—they might still have important weaknesses lying dormant that LLMs will soon be able to discover. We see this as a real opportunity to expand our ability to study the long tail of ciphers used throughout the world, and also our ability to more deeply study the ciphers that matter most. Indeed, as we mentioned above we have already begun audits of other schemes.

The attacks described in these two papers are the strongest attacks we have found to date. We are sharing them after a period of consultation with US government and industry leaders. But as we develop increasingly powerful cryptanalytic results, it would be prudent to consider how researchers should react if a language model were to discover vulnerabilities in cryptosystems where attacks *do* have an immediate real-world impact.

We believe answering this question will require input from academia, government, and industry. We hope that our work here will help launch these conversations.

The cryptography community has always benefited from adversarial review: ciphers are proposed, examined, and revised until the community is satisfied with their security. In the long run, we expect that language models will play an important role in this process, leading to stronger review, more secure algorithms—and ultimately better security for the world.

Links to full research papers

Read the full paper on HAWK.

Read the full paper on AES, and the associated chain of thought.

Read the paper introducing CryptanalysisBench.

Edit 29 July: Updated an academic affiliation.

Research

Policy

Commitments

Company

Products

Models

Solutions

Help and security

Terms and policies

Our mission

Anthropic is a Public Benefit Corporation. Our mission is the responsible development and maintenance of advanced AI for the long-term benefit of humanity.

Log in

Claude.ai ↗

Claude console ↗

© 2026 Anthropic PBC