

Public Cryptanalytic Effort in the Age of AI

A literature-based estimate of human research effort
and a scenario comparison with mathematical AI results

Working draft for discussion

9 October 2026 · Version 0.2

Abstract

How much public effort has been devoted to cryptanalysis, and how does that effort compare with the resources used in recent AI-assisted mathematical research? We assemble a multi-source literature corpus covering elliptic-curve discrete logarithms, lattice problems, and syndrome decoding, with factoring, isogenies, and multivariate cryptography as comparators. Fractional attribution yields approximately 465 elliptic-curve, 840 lattice, and 290 code-based research items in the principal view. A separate author-output model estimates approximately 594, 1,036, and 331 research-allocation years, respectively. At an explicit convention of 2,000 hours per research-year, these correspond to approximately 1.19 million, 2.07 million, and 0.66 million nominal person-hours. Neither publication counts nor these derived hours are direct time measurements.

We compare these estimates with literature-based effort estimates for selected mathematical problems and resource disclosures from OpenAI, Anthropic, and an independent evaluation of OpenAI models. Because historical scrutiny and the marginal cost of a breakthrough purchase different outputs, we do not fit an empirical conversion rate. Instead, assuming economic productivity multipliers of 1, 10, and 100 at \$250,000 per human research-year gives equivalent inference budgets of \$250,000, \$25,000, and \$2,500 per year. Under the middle scenario, the three principal cryptanalytic totals correspond to approximately \$14.85 million, \$25.91 million, and \$8.29 million. We also invert the comparison to express disclosed AI budgets in assumed human-year equivalents. These scenarios describe scale, not attack probabilities, security guarantees, or demonstrated replacement of human research. The draft identifies outstanding validation and accounting uncertainties.

Status. This paper synthesizes the repository as of 9 October 2026. Literature labels and person-time models remain provisionally machine-generated; this draft does not independently verify the mathematical claims discussed. No new rediscovery experiment is required for the adopted comparison.

1 Question, contribution, and interpretation

Security arguments for public-key cryptography partly rely on sustained attempts to defeat the underlying mathematical problems. Yet statements such as “decades of scrutiny” do not specify how many researchers were involved, which parameter regimes they studied, or whether their work concerned algorithms, implementation, structural properties, or concrete attack costs. Calendar age alone is a poor measure of effort.

Recent AI-assisted mathematics creates a second measurement problem. Laboratories report resources in tokens, inference dollars, elapsed time, or equivalent product usage. These disclosures are difficult to place beside a cryptographic community’s accumulated work. A useful comparison needs to distinguish three quantities:

1. **Historical human research effort on cryptanalysis**, estimated from the public literature and researchers’ publication portfolios.
2. **Historical human research effort around selected mathematical problems**, estimated by a similar paper-based procedure. This includes prior and partial work, not just the time required to prove a final theorem.
3. **Human-year equivalents assigned to AI resource use**, derived from an explicit productivity and price convention. These are hypothetical equivalents, not actual human years spent by OpenAI or Anthropic employees.

Actual human supervision, verification, and writing for an AI project constitute a fourth quantity. We report such information where available but do not infer it from elapsed model time. Similarly, a hundred agents running for an hour do not represent a hundred human hours.

This paper contributes a reproducible description of visible research effort and an assumption-based way to express its scale in units used in AI research. It does not attempt to estimate the probability that a specified AI budget will find a break. An unsuccessful line of cryptanalysis can provide useful evidence about a scheme while producing no publishable improvement; conversely, one new mathematical idea can invalidate substantial prior confidence. Research input and security output are not interchangeable.

2 Literature search and classification

2.1 Scope and sources

The primary classes are elliptic-curve discrete logarithms (ECDLP), structured/module/ring lattice problems, plain learning with errors (LWE), and syndrome decoding. Factoring and finite-field discrete logarithms, isogenies, and multivariate systems provide comparators. Generic lattice machinery is allocated between structured and plain lattices; the combined lattice class is the most stable headline unit. The corpus concerns mathematical assumptions relevant to public-key encryption and includes signature research when it analyzes shared mathematical machinery. It is not restricted to the exact deployed parameters of a particular standard.

The acquisition combines the Cryptology ePrint Archive, DBLP, OpenAlex, and zbMATH. The ePrint harvest contains 27,249 live records; the DBLP dump contains about 8.7 million records, of which 375,911 survived an initial title/venue screen. OpenAlex phrase searches, abstract enrichment, and a citation snowball from 811 seeds extend coverage. A direct arXiv acquisition failed because of rate limits; arXiv entries indexed by the other sources remain represented. These acquisition counts describe different stages and must not be added as

distinct papers. The combined classification outputs contain 16,468 labeled rows, and the export contains 4,349 included or borderline records across all categories. The headline effort view is narrower.[14]

The study uses title and abstract screens followed by two classification passes. The second pass reviews included and borderline records plus an approximately 10% sample of exclusions. A later round adds 1,840 candidates found by a broader screen and 835 sampled lower-priority candidates. Version and title deduplication reduces repeated counting, but residual duplicates and incorrect metadata remain possible.

2.2 What counts as cryptanalytic work

The headline effort categories include attack algorithms, supporting implementations, structural cryptanalysis, cost estimates, computational records, negative results, and relevant failed or refuted claims. Surveys and general hardness reductions are tracked separately. The latter exclusion is a substantive convention: some reductions supply precisely the tools that later enable an attack. Case studies should therefore inspect whether this boundary omits important enabling work.

Fractional attribution assigns a paper’s weight across applicable problem classes. For example, generic lattice work is split equally between structured and plain lattices; it is counted once in the combined lattice view. A “touch” count instead credits a paper to every class it addresses. The two measures answer different questions. Authors’ name counts from paper metadata also differ from the deduplicated identities in the author map.

The primary view counts sampled lower-priority papers at their observed weight. A sensitivity view multiplies their contribution by ten to estimate the unsampled remainder. This extrapolation depends on a small number of positive labels and is not folded into the headline totals.

2.3 Quality evidence

Agreement between model passes is approximately $\kappa = 0.84\text{--}0.89$. Phase-1 exclusion audits report estimated miss rates around 0.5–1.1% within the relevant candidate pools. A separate benchmark uses the reference lists of 12 seed papers: 421 of 429 references were harvested, and review of remaining gaps suggests approximately 96–100% recall of in-scope references by class after the second round. These are checks on selected reference lists and classifier consistency, not guarantees of recall across the entire field. The second classifier could see the first label, so its agreement is not an independent human validation. Human spot-checks remain outstanding.[14]

3 Results from the public literature

Table 1 shows that the largest principal-class literature is the combined lattice class. Its fractional count is approximately 1.81 times ECDLP’s, while syndrome decoding is approximately 0.62 times ECDLP’s. Across the study’s source, attribution, and extrapolation views, the lattice/ECDLP ratio ranges roughly from 1.3 to 2.6 and the decoding/ECDLP ratio from 0.6 to 0.8. The spread mixes distinct definitions and sensitivity choices; it is not a confidence interval.

The historical pattern matters. Before 2006, ECDLP has more fractional items than lattices. In 2017–2026, lattices have approximately 3.29 times as many as ECDLP, and syndrome decoding slightly exceeds ECDLP. Both post-quantum leads appear larger in an ePrint-only view because older elliptic-curve and factoring research is poorly represented there. Adding older bibliographic sources changes the comparison materially. Publication volume also reflects incentives. Standardization competitions attract concentrated analysis, and lattice papers motivated by homomorphic encryption do not all transfer directly to ML-KEM parameters. A

Table 1: Main literature view: include = yes, effort categories A–F and H. Fractional counts allocate shared work; touch counts overlap. Era columns are fractional counts.

Class	Fractional	Touch	Before 2006	2006–16	2017–26
ECDLP	465.2	555	97.3	197.6	170.2
All lattices	840.0	862	55.0	225.1	559.9
MLWE / structured	477.4	803	32.5	133.2	311.6
Plain LWE	362.6	678	22.4	91.8	248.3
Syndrome decoding	289.5	325	24.7	80.1	184.7
Factoring / FF-DLP	849.1	940	212.5	343.0	293.7
Isogenies	80.4	83	0.0	2.2	78.2
Multivariate	174.7	190	13.0	67.7	94.0

mature consensus that generic methods are difficult to improve can reduce recent publication volume without implying neglect. Source coverage, parameter regime, and negative-result categories should accompany any headline ranking.

4 From publications to research-years and person-hours

4.1 Two distinct effort proxies

The repository provides two approaches. The first assigns total effort priors to weighted papers. For a problem P , with attribution weight w_i , category k_i , inclusion factor b_i (one for included work and one-half for borderline work), and assumed effort μ_{k_i} ,

$$H_P^{\text{papers}} = \sum_i b_i w_i \mu_{k_i}. \quad (1)$$

In the existing problem-calibration implementation, categories A/C/F/H receive 1–2 person-years per paper; D receives 0.5–1.5, E receives 2–10, and S receives 0.5–1. G is excluded. The midpoint sum supplies the displayed central values. These are total years across authors, not years per author. These priors differ in some categories from the general project’s original plan and author-map fallback; this draft preserves the implemented definitions rather than silently homogenizing them.

The second approach allocates each researcher’s output across classes within a moving five-year window. It uses deduplicated DBLP and zbMATH publications, identity matching, an output-rate floor, and adjustments near the observation boundary. Let a_{pcy} denote researcher p ’s publication share allocated to class c in year y . The corresponding research-allocation total is schematically

$$H_c^{\text{share}} = \sum_{p,y} a_{pcy} + H_c^{\text{fallback}}. \quad (2)$$

Researchers without usable career publication histories enter through a per-paper fallback. The author map identifies approximately 3,240 people across classes. It reports both the allocation total and a second total incorporating assumed research-time fractions by career stage.[14]

Equation (2) is a proxy for research allocation, not a time-use survey. It assumes that output share reflects effort share. The apparent agreement between paper-based and author-based estimates is useful but not wholly independent: both use the same classified literature, and part of the author estimate directly reuses paper priors.

4.2 The hours convention

For a legible comparison, we adopt

$$h_0 = 2,000 \text{ hours per research-year}, \quad L_c^{\text{nominal}} = h_0 H_c^{\text{share}}. \quad (3)$$

The 2,000-hour figure is an accounting convention, not a measured annual schedule. Calling the product *nominal person-hours* prevents it from being confused with logged work. Using 1,500 or 2,200 hours instead scales all reported hours by 0.75 or 1.10, respectively, and does not change the year-based dollar scenarios.

The repository’s career-adjusted alternative multiplies publication shares by estimated research-time fractions. Its nominal hours are lower because faculty and other researchers also perform nonresearch work. Table 2 presents both conventions. The adopted AI-budget scenarios retain the unadjusted research-allocation totals for consistency with the existing method. Switching to the adjusted totals would lower the budgets proportionally; it is a change of human-effort definition, not evidence about AI productivity.

Table 2: Estimated human effort through the 2026 observation window. Hours are in thousands, using 2,000 hours/year. The last column uses the repository’s career-stage research fractions. Values are provisional proxies.

Class	Allocation years	Nominal hours (k)	Adjusted hours (k)
ECDLP	593.8	1,187.6	877.8
All lattices	1,036.4	2,072.8	1,517.0
MLWE / structured	583.6	1,167.2	852.2
Plain LWE	452.8	905.6	664.8
Syndrome decoding	331.4	662.8	497.4
Factoring / FF-DLP	1,111.1	2,222.2	1,695.0
Isogenies	116.6	233.2	170.4
Multivariate	210.2	420.4	299.4

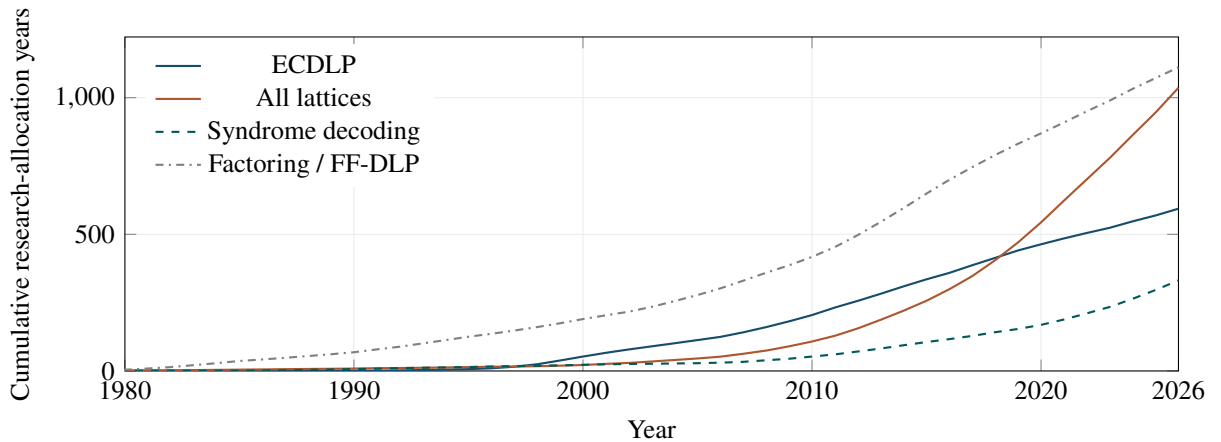


Figure 1: Author-output model of cumulative public effort. Attribution spreads publications over moving windows, so curves are smoothed estimates rather than observed project timelines. The 2026 observation window is incomplete.

4.3 Uncertainty in the human estimates

The main uncertainties are shared assumptions, not just sampling noise. If effort per paper is systematically twice the assumed level, an entire paper-based estimate doubles. Independently drawing one effort value for each paper makes the sum appear increasingly precise without resolving that common uncertainty. For example, the HAWK yes-only estimate of 35.4 years has a reported Monte Carlo interval of 33.2–37.6 under its independent priors. That interval omits uncertainty in the priors, missing literature, attribution, and unpublished work.

We therefore treat the central estimates as explicit scenarios and retain alternate definitions. The author model’s publication-rate floor, name resolution, and treatment of unknown career stages deserve particular scrutiny. A paper can represent a brief observation, several researchers’ multiyear project, or a computational campaign with substantial engineering. Neither model measures unreported failed projects, private industry work, or classified work. Those omissions are not known scalar corrections.

5 Mathematical AI results and their human context

5.1 Historical literature around selected problems

The problem-specific search uses curated seeds, citation expansion, and a rubric distinguishing work on a problem from work that merely assumes it. It covers HAWK/module-LIP, the Unique Games and 2-to-1 Games lines, and 22 selected Erdős problems. Table 3 shows a subset relevant to the OpenAI and Anthropic comparison; the full Erdős sample is in Appendix B.

Table 3: Paper-based historical effort around selected problems, including borderline work at half weight. Ranges sum the endpoints of the implemented effort priors; they are sensitivity ranges, not confidence intervals. Hours use the same 2,000-hour convention.

Problem line	Estimated years	Nominal hours (k)	Prior-endpoint years
Unique Games line	85.8	171.6	55.2–116.4
2-to-1 Games line	20.3	40.6	13.5–27.1
HAWK / module-LIP	38.4	76.8	25.6–51.2
Erdős #1	22.1	44.1	14.7–29.4
Erdős #74	3.0	6.0	2.0–4.0
Erdős #126	4.9	9.8	3.3–6.5
Erdős #548	9.8	19.5	6.5–13.0
Erdős #571	14.4	28.7	9.6–19.2

These estimates are neither time spent by the labs nor counterfactual time that a human team would need for the new result. They measure published work surrounding the problem under the selected scope. A complete audit of historical cutoffs, AI-assisted papers, duplicates, and problem variants remains necessary before describing every item as pre-discovery human labor. Classification based on abstracts can miss AI involvement. Unique Games and the 2-to-1 line overlap and must not be added. Their central values do not imply that the recent results replaced 86 or 20 human years.

The broader mathematical sample is selected, not representative of mathematics. Several small samples have weak agreement between classification passes. Neither the age of a conjecture nor its fame determines how much active effort it received. A failed proof search may leave no paper; a paper with many downstream applications may contain little effort directed at the original question.

5.2 What the laboratories disclose

Table 4 separates disclosed units from the conversions introduced later. We use primary laboratory accounts and the independent FrontierMath Erdős evaluation. We have not revalidated the mathematical claims or their proof artifacts.

Table 4: Selected AI research disclosures as checked on 9 October 2026. Different rows have different accounting boundaries and must not be summed.

Result or evaluation	Disclosed resource figure	Accounting boundary
Anthropic HAWK	About \$100,000 API cost; about 60 hours	Discovery process; token count and active human hours unspecified.[1]
Anthropic reduced-round AES	About \$100,000 API cost; about one billion output tokens	Also reports substantial human validation; prior session allocation is incomplete.[1]
OpenAI ten advances	About \$2,000 at Sol API rates	Reference-price valuation of solution tokens across ten results; not a measured all-in project bill.[18]
OpenAI October collection	Roughly three hours of Pro-equivalent thinking per result	Release-wide average; no per-family dollar or token budget.[16]
OpenAI Navier–Stokes claim	About 130 billion output tokens; 88 hours plus 17 hours formalization	About 300 billion output tokens across all attempted problems; input tokens and total dollar bill unspecified.[17]
FrontierMath Erdős / OpenAI model	\$218 for #74; \$247 for #126	Successful benchmark attempts, not costs of the whole search.[11]

The October OpenAI catalogue contains 719 manuscripts in 372 families from approximately 4,000 posed problems; its README says verification is incomplete. Those aggregation units are not interchangeable. The collection includes the problem lines used in our human-effort comparison, but the overall average is insufficient to assign an inference bill to either line.[19]

FrontierMath’s default evaluation scored two resolutions among 68 conjectures for a prerelease OpenAI model. Additional, non-systematic attempts reached five conjectures. The paper notes that metering at another model’s prices caused attempts to exceed the intended actual-dollar budget. Its discussion’s roughly \$10,000 expected cost per resolution is therefore a budget/success-rate heuristic, not a reconciled campaign ledger. We do not infer missing trial costs or treat successful-attempt prices as a complete denominator.[11]

5.3 HAWK as a nearby comparison, not an exchange-rate estimate

HAWK is useful because the human and AI-assisted results address closely related mathematical advances. The human trace-zero approach reduces key recovery to exact-SVP calls in dimension $3n/4 + 1$; the Anthropic result reaches $n/2 + 1$. These are related but different deliverables.[15, 20] The AI work also uses existing mathematical tools, so its discovery cost cannot be interpreted as the cost of recreating the full preceding literature.

The human group’s active effort is undisclosed. The repository’s earlier 0.3–1 person-year suggestion was a hypothetical range, not an observation. At \$250,000 per year, \$100,000 equals 0.4 years before supervision and other expenses. That is a break-even threshold, not evidence of cost parity. Anthropic describes its figure as the full HAWK discovery process; whether earlier unsuccessful campaigns are included is unknown. It also reports sharing the attack with the HAWK team in June, so July publication dates do not establish either discovery durations or independence.[1]

A separate historical normalization assigns the reported \$100,000 to the estimated 35.4 years of HAWK literature effort, or 38.4 years including borderline work:

$$D_X^{\text{HAWK}} = \$100,000 \frac{H_X^{\text{papers}}}{H_{\text{HAWK}}^{\text{papers}}}. \quad (4)$$

This produces approximately \$2,825 or \$2,604 per historical paper-based year. It assumes this historical-effort-to-breakthrough-cost ratio transfers across problems. A shared multiplicative effort prior cancels if the same paper estimator is used on both sides; coverage and attribution differences do not. We do not apply this normalization directly to the author-output totals, which use a different estimator.

6 Assumed productivity and equivalent inference budgets

6.1 Forward conversion: human effort to an AI budget

We adopt a loaded human cost convention $c = \$250,000$ per research-year and an assumed productivity multiplier $s \in \{1, 10, 100\}$. Conditional on access to existing literature, assume AI-assisted research supplies comparable research effort at s times the human productivity per inference dollar. Define

$$D_c(s) = \frac{cH_c^{\text{share}}}{s}. \quad (5)$$

These scenarios correspond to \$250,000, \$25,000, and \$2,500 per human research-year. They are not percentiles, confidence limits, or fitted estimates; no one scenario is preferred. Values outside this range remain possible. Applying the same multiplier across classes preserves their human-effort ranking by construction.

Table 5: Equivalent inference budgets in millions of dollars, using the allocation-year column of Table 2. Arithmetic precision does not imply measurement precision.

Class	$s = 1$	$s = 10$	$s = 100$
ECDLP	148.45	14.85	1.48
All lattices	259.10	25.91	2.59
MLWE / structured	145.90	14.59	1.46
Plain LWE	113.20	11.32	1.13
Syndrome decoding	82.85	8.29	0.83
Factoring / FF-DLP	277.78	27.78	2.78
Isogenies	29.15	2.92	0.29
Multivariate	52.55	5.26	0.53

The comparison is an inference allowance, not an all-in AI project cost. Human supervision and verification, tool compute, infrastructure, and training are outside that allowance. Historical human research compute likewise remains outside cH . An all-in comparison would require explicit accounting on both sides and a revised productivity assumption.

The largest conceptual assumption concerns the transfer from success to scrutiny: we apply a productivity multiplier to the mixture of useful results, failed approaches, and incremental work. We have no observation establishing that AI scales all of these activities similarly. The scenario formulation makes that assumption visible instead of concealing it in a fitted ratio from a selected success.

6.2 Inverse conversion: an AI budget to human-year equivalents

For an AI expenditure or stated reference-price valuation D , the inverse of Equation (5) is

$$H^{\text{equiv}}(D; s) = \frac{sD}{c}, \quad L^{\text{equiv}}(D; s) = h_0 \frac{sD}{c}. \quad (6)$$

This is the comparison most directly relevant to claims about “human years” represented by recent AI research. Table 6 gives concrete examples. The multiplier increases the assigned years for a fixed budget because it assumes more research per dollar.

Table 6: Assumed human-year equivalents of selected disclosed budgets. These are not observed human labor, measured task difficulty, or estimates of labor actually displaced.

Example	Budget basis	$s = 1$	$s = 10$	$s = 100$
Anthropic HAWK	\$100,000	0.4	4	40
Anthropic reduced-round AES	\$100,000	0.4	4	40
OpenAI ten advances (combined)	\$2,000 valuation	0.008	0.08	0.8

For HAWK, the corresponding nominal hours are 800, 8,000, and 80,000. Under the 100x assumption the assigned 40 years happen to be close to the 35–38 paper-based historical years. This proximity reflects the choice of scenario and accounting boundary; it is not an independent validation. The ten-advance valuation yields 16, 160, and 1,600 hours across the selected results, with the source’s reference-price and selection qualifications retained.

6.3 Tokens and a token-only example

Dollar and token comparisons need a fixed valuation rule. For uncached input, cached input, and output fractions f_i, f_c, f_o summing to one, with prices p_i, p_c, p_o per million tokens, define

$$p_{\text{mix}} = f_i p_i + f_c p_c + f_o p_o, \quad T_{\text{reference}} = 10^6 D / p_{\text{mix}}. \quad (7)$$

A hypothetical mixed price of \$10 per million tokens maps one research-year to 25 billion, 2.5 billion, or 250 million reference tokens under the three scenarios. This example quotes no actual model price. Real comparisons must specify model version, date, mix, caching, and treatment of billed reasoning tokens.

For the Navier–Stokes disclosure, we can instead value only the reported 130 billion *output* tokens.[17] If p_o is an explicitly hypothetical output price,

$$D_{\text{output}} = 130,000 p_o, \quad H_{\text{output}}^{\text{equiv}} = 0.52 s p_o. \quad (8)$$

This exercise illustrates the sensitivity to price and productivity assumptions. It does not estimate historical human time spent on Navier–Stokes, for which this study has no bounded, audited literature-effort estimate. The reported 300 billion tokens across the wider campaign are a different scope and must not be added to the 130-billion subset. Elapsed hours and approximate concurrent agent counts do not reconstruct an exact agent-hour integral or a human work history.

Table 7: Illustrative output-only valuation of 130 billion tokens. Prices are assumptions, not the internal model’s prices or reported costs. Input tokens and other project costs are omitted.

\$/million output tokens	Output valuation	Years at 1x	Years at 10x	Years at 100x
10	\$1.3M	5.2	52	520
50	\$6.5M	26	260	2,600
100	\$13M	52	520	5,200

7 Selected changes in attack methods

Research effort and attack progress are separate quantities. To make the distinction concrete, a supplementary track records selected published attack analyses, their assumptions, and corrections. It compares leading cost expressions where a defensible common scale exists, and otherwise records qualitative events. This is not a complete history of the best attack at every date, an estimate of research productivity, or a forecast of the next breakthrough.

7.1 Indices, dates, and scope

For an analysis j , let $\tilde{W}_j(s)$ be its cost expression after the stated simplifications, evaluated at a reference scale s . We display

$$I_j(s) = \log_2 \tilde{W}_j(s), \quad I_{S,t}(s) = \min_{j \in S: t_j \leq t} I_j(s), \quad (9)$$

where S is the explicitly selected set of analyses. The minimum is an envelope of those models, not a lower bound on attack cost. Dropping lower-order terms can distort finite-size comparisons, and memory requirements and success guarantees differ. Even within a track, a change in the index is not automatically an equal change in practical security. Values across tracks use different operations, assumptions, and reference scales.

Dates refer to the quoted analysis, using a public preprint or a specified checked revision when appropriate. An improved analysis of an old algorithm is not backdated to the algorithm’s invention. Incorrect historical claims are retained as contextual records rather than plotted as supported bounds. Thus this view does not reconstruct what researchers believed at each historical date. Gaps in the selected record and unchanged leading terms are not evidence that practical attacks stopped improving.

Table 8: Selected numerical tracks. Windows delimit the plotted endpoints, not the age of an assumption. Endpoints are leading-term indices at each track’s own reference scale; they are not comparable security ratings. Qualitative events are excluded.

Track	Window	First index	Latest index
ECDLP	1985–1998	128.3	127.8
SVP sieving	2008–2015	168.5	118.7
Syndrome decoding	2012–2022	137.2	128.0
Factoring	1982–1993	184.3	137.2
Boolean MQ	2004–2020	161.0	127.8

The elliptic-curve track concerns generic discrete logarithms in prime-order groups; weak-curve attacks are contextual. Factoring uses selected leading asymptotic expressions for balanced moduli. The factoring curve

starts with the quadratic-sieve analysis, omitting the incomplete earlier history. Neither the table’s cost indices nor their inversions replace concrete parameter recommendations. A polynomial-time classification alone also does not imply that no finite input could require a large amount of work.

7.2 Three boundaries that change the interpretation

Lattices. The numeric track compares selected heuristic classical sieve bounds at SVP dimension 406. That dimension comes from Kyber’s coarse core-SVP convention, not a complete ML-KEM attack. The specification’s separate refined estimate uses blocksize 413 and sieve dimension 375; repeated oracle calls and concrete work costs also matter.[2] Worst-case enumeration bounds are kept outside the sieve curve, avoiding the earlier spurious comparison that suggested hundreds of bits of lost ML-KEM security. Locality-sensitive filtering improves the selected leading sieve coefficient, but its memory tradeoffs must accompany its time bound.[5] Dimensions-for-free subsequently produces a dimension-dependent subexponential saving: preserving the leading coefficient does not make the saving a constant factor or make fixed-size work unchanged.[9]

Decoding. The scalar series concerns random binary codes at full Gilbert–Varshamov distance, maximizing each algorithm’s optimized exponent over the code rate. The maximizing rate can change; this is not one rate-1/2 code or a fixed deployed instance. The length scale $n = 1346$ is merely a normalization. The selected corrected endpoint uses $c \approx 0.0951$ in $2^{cn+o(n)}$, from the 2022 reassessment of Both–May, rather than the original erroneous claim.[10, 12] Its date denotes a correction to the analysis, not a new attack invented that year. Hidden factors, exponential memory, and later rate-specific results limit what this scalar history establishes.

Boolean quadratic systems. The selected record includes early XL and hybrid methods, not only exhaustive search and later parity-counting. Conditional algebraic estimates for random square systems must be distinguished from randomized bounds valid for every input. BooleanSolve’s assumptions and Dinur’s general-input guarantee are different, even when both are evaluated on a square Boolean reference system.[4, 8] A corrected hybrid coefficient enters at the checked analysis version; it is not retrospectively attributed to the original proposal.[6] Taking the smallest selected coefficient does not remove these differences or establish the security of structured signature schemes.

7.3 Breaks that should remain events

Small-characteristic finite-field DLP illustrates why incompatible units should not be joined into a curve. The BGJT result has heuristic quasi-polynomial complexity $n^{O(\log n)}$ in field bit-size, which remains super-polynomial.[3] Separately, the concrete reassessment of a supersingular pairing construction estimates about 59 bits of work for $\mathbb{F}_{2^{4 \cdot 1223}}$, in the authors’ modular-multiplication units. This is neither a measured computation nor an estimate for the base field $\mathbb{F}_{2^{1223}}$. [13] SIDH provides a practical key-recovery break exploiting auxiliary torsion images; it does not break the general isogeny-path problem.[7] We retain that event without manufacturing an operation count from wall-clock time.

The milestone ledger in `data/attack_history/` records sources, assumptions, selection status, and corrections; `analysis/attack_stability.py` generates Table 8. Source checking here is an agent review, not independent mathematical verification or the repository’s separate human-verification flag. This track helps distinguish algorithmic discoveries, revised estimates, and implementation gains. It supplies no exchange rate between those outcomes and the literature-derived person-time estimates.

8 What can be concluded, and what remains uncertain

The most supportable findings concern the public literature. Multi-source coverage materially improves the representation of older elliptic-curve and factoring work. Lattice research is larger and more recent than ECDLP research in the principal views, while syndrome decoding is smaller cumulatively but has recently attracted substantial attention. These are statements about the selected public corpus, not comparisons of cryptographic security.

Under the study’s human-effort conventions, major assumption classes represent hundreds to roughly a thousand allocation years, or hundreds of thousands to a few million nominal person-hours. Under the adopted productivity scenarios, their budget equivalents range from hundreds of thousands to hundreds of millions of dollars. A disclosed AI research campaign may fall within that broad scale under some assumptions. That observation establishes neither equivalent coverage of attack approaches nor a probability of a practical break.

Several limitations dominate any finer interpretation:

- **Coverage and labels.** Candidate screening, incomplete abstracts, author identity errors, older sources, duplicate versions, and ambiguous scope affect all downstream quantities. Model agreement does not resolve correlated mistakes.
- **Publication and time.** Unpublished failures are invisible; output share is not necessarily time share. Shared effort priors and research-time fractions can shift whole classes together.
- **Selection and accounting.** Public AI results emphasize successes. Attempts, retries, tool use, and verification costs are not consistently disclosed. A reference-price token valuation is not an internal compute bill.
- **Different outputs and starting knowledge.** Historical research built the tools available to a later AI system. A weaker theorem, stronger theorem, implementation, and formally checked proof are distinct deliverables.
- **Transfer.** A single HAWK-style advance need not predict progress on mature ECDLP, generic decoding, or a different lattice structure. The common multiplier is assumed, not learned.
- **Unobserved activity.** Private and classified work, nonpublishing contributors, and undocumented AI campaigns are outside the measured corpus. Public visibility is not total effort.

The computational-record track provides a complementary measure of demonstrated attack computation, including normalized hardware-effort ranges and gaps to target parameters. It is kept separate here: executing an attack, discovering an attack, and estimating human research time are different accounting problems. Most individual records remain unverified by a human, so a stronger quantitative synthesis of those records should await review.

For this working draft, the productive next step is targeted validation of high-impact labels and effort definitions, not a costly rediscovery campaign. The earlier HAWK pilot produced no completed mathematical evaluations because its attempts were interrupted by model safeguards; it supplies no success-rate estimate. Even a successful controlled rediscovery would leave the dominant cross-problem transfer assumption unresolved. The present scenario comparison is complete without that experiment.

9 Conclusion

A literature search can make public cryptanalytic attention more concrete, but it cannot turn publications directly into observed hours. Explicit person-time conventions produce useful orders of magnitude while preserving that distinction. Recent AI mathematical disclosures can be placed beside those magnitudes through clearly labeled budget and productivity assumptions. The resulting human-year equivalents are a framework for discussing research scale; they are not measurements of years replaced by AI or forecasts of cryptographic failure.

A Reproducibility and source map

The LaTeX source is `paper/whitepaper.tex`; references are in `paper/references.bib`. From the repository root, run `make -C paper`. This first rebuilds the attack-history export and generated tables, then compiles with `latexmk` and copies the PDF to `output/pdf/cryptanalytic-effort-whitepaper.pdf`. Compilation requires a LaTeX installation with the packages declared in the source. No model call or network request is part of the build.

Tables 1, 2, 3, 5, and 8, Figure 1, and Appendix B are generated from repository outputs. The source ledger `paper/generated/manifest.json` records input SHA-256 hashes, constants, and derived values. The source snapshot is dated; updates should regenerate tables and trigger review of nearby prose.

Table 9: Repository provenance for this draft. Paths are relative to the repository root.

Quantity	Source or method
Corpus and counts	<code>analysis/phase1_report.md</code> , <code>analysis/phase1_counts.txt</code> , <code>data/export/combined_included.csv</code>
Attribution and labels	<code>pipeline/attribution.json</code> , <code>pipeline/rubric.md</code> , <code>analysis/recall_benchmark.md</code>
Author effort	<code>data/authors/fte_by_year.csv</code> , 2026 cumulative columns; <code>analysis/author_map_method.md</code>
Problem literature effort	<code>data/ai_effort/calib/*/labeled_final.jsonl</code> ; priors in <code>pipeline/calib_estimate.py</code> and <code>analysis/calib_exchange_table.py</code>
Selected attack analyses	<code>data/attack_history/</code> , <code>analysis/attack_stability.py</code> ; numerical indices and qualitative events kept separate
AI disclosures	<code>data/ai_effort/ai_results.yaml</code> , <code>data/ai_effort/erdos_problems.csv</code> ; primary sources cited in the text
Adopted scenarios	<code>analysis/ai_effort_comparison_scoping.md</code> ; <code>analysis/ai_effort_hawk_case.md</code>

Before external circulation, review the strongest claims, especially source coverage, the person-hours interpretation, historical cutoffs in the mathematical samples, and validation status of the cited AI results. These tasks improve the evidentiary status of the draft; they do not require claiming a universal exchange rate.

B The selected Erdős literature sample

The existing study selected 12 problems with AI resolutions reported across the Epoch and AlphaProof Nexus collections and 10 sampled unresolved problems. This appendix retains all 22 literature estimates as a scope

disclosure, not a fresh evaluation of models. Different conjecture variants, proof dates, and AI-assisted publications require item-level review. The “items” column counts eligible category entries under the existing inclusion rule, including borderline records; it is not a count of successful attempts. No effort totals should be added across overlapping problem literatures without deduplication.

Table 10: Full selected Erdős sample: paper-based effort priors with borderline work at half weight. Ranges are prior-endpoint sensitivity sums.

Problem number	Midpoint years	Endpoint years	Included/borderline items
1	22.1	14.7–29.4	31
12	6.2	4.1–8.3	6
26	2.6	1.7–3.4	5
28	3.2	2.2–4.3	9
41	1.0	0.7–1.3	4
68	0.3	0.2–0.5	4
74	3.0	2.0–4.0	2
125	2.9	1.9–3.8	2
126	4.9	3.3–6.5	11
138	64.4	31.4–97.3	56
152	0.9	0.6–1.2	1
172	6.1	4.1–8.1	10
324	2.3	1.5–3.0	3
548	9.8	6.5–13.0	7
571	14.4	9.6–19.2	24
713	1.7	1.1–2.3	11
741	0.5	0.3–0.7	1
846	1.9	1.3–2.5	2
970	16.3	8.9–23.8	19
1003	1.9	1.3–2.5	5
1083	8.3	5.5–11.0	10
1206	0.8	0.5–1.0	2

References

- [1] Anthropic. Discovering cryptographic weaknesses with Claude. <https://www.anthropic.com/research/discovering-cryptographic-weaknesses>, 2026. 28 July 2026; accessed 9 October 2026. Primary disclosure of discovery cost and process.
- [2] Roberto Avanzi, Joppe Bos, Léo Ducas, Eike Kiltz, Tancrede Lepoint, Vadim Lyubashevsky, John M. Schanck, Peter Schwabe, Gregor Seiler, and Damien Stehlé. CRYSTALS-Kyber: Algorithm specifications and supporting documentation. Version 3.02, 4 August 2021. <https://pq-crystals.org/kyber/data/kyber-specification-round3-20210804.pdf>, 2021. Table 4 and Section 5.2 distinguish coarse core-SVP and refined estimates.
- [3] Razvan Barbulescu, Pierrick Gaudry, Antoine Joux, and Emmanuel Thomé. A quasi-polynomial algorithm for discrete logarithm in finite fields of small characteristic. <https://arxiv.org/abs/1306.4244>, 2013. First posted 18 June 2013; EUROCRYPT 2014. Heuristic quasi-polynomial complexity.
- [4] Magali Bardet, Jean-Charles Faugère, Bruno Salvy, and Pierre-Jean Spaenlehauer. On the complexity of solving quadratic boolean systems. <https://arxiv.org/abs/1112.6263>, 2011. First posted 29 December 2011; Journal of Complexity 29(1), 2013, pages 53–75. Conditional algebraic analysis.
- [5] Anja Becker, Léo Ducas, Nicolas Gama, and Thijs Laarhoven. New directions in nearest neighbor searching with applications to lattice sieving. Cryptology ePrint Archive, Paper 2015/1128. <https://eprint.iacr.org/2015/1128>, 2015. First posted 23 November 2015; SODA 2016. Section 7 distinguishes time–memory variants.
- [6] Daniel J. Bernstein and Bo-Yin Yang. Asymptotically faster quantum algorithms to solve multivariate quadratic equations. Checked manuscript dated 25 January 2018. <https://homepage.iis.sinica.edu.tw/papers/byyang/21806-F.pdf>, 2018. Classical analysis in Sections 1.2 and 4.5; ePrint 2017/1206 first posted 18 December 2017. The correction is not backdated to the original FXL proposal.
- [7] Wouter Castryck and Thomas Decru. An efficient key recovery attack on SIDH. Cryptology ePrint Archive, Paper 2022/975. <https://eprint.iacr.org/2022/975>, 2022. First disclosed 30 July 2022; EUROCRYPT 2023. Revised running times should not be attributed to the original implementation.
- [8] Itai Dinur. Improved algorithms for solving polynomial systems over $\text{GF}(2)$ by multiple parity-counting. <https://arxiv.org/abs/2005.04800>, 2020. First posted 10 May 2020; SODA 2021. Randomized general-input bound.
- [9] Léo Ducas. Shortest vector from lattice sieving: A few dimensions for free. Cryptology ePrint Archive, Paper 2017/999. <https://eprint.iacr.org/2017/999>, 2017. First posted 11 October 2017; EUROCRYPT 2018.
- [10] Léo Ducas, Andre Esser, Simona Etinski, and Elena Kirshanova. Asymptotics and improvements of sieving for codes. CWI publication 33753. <https://ir.cwi.nl/pub/33753/33753.pdf>, 2023. Table 1: full-distance decoding exponents and their differing maximizing rates.
- [11] Epoch AI. FrontierMath Erdős. <https://arxiv.org/abs/2609.25050>, 2026. Sections 4–5 and Tables 2–3; accessed 9 October 2026.
- [12] Andre Esser. Revisiting nearest-neighbor-based information set decoding. Cryptology ePrint Archive, Paper 2022/1328. <https://eprint.iacr.org/2022/1328>, 2022. First posted 6 October 2022; corrected analysis of the Both–May algorithm.

- [13] Robert Granger, Thorsten Kleinjung, and Jens Zumbrägel. Breaking ‘128-bit secure’ supersingular binary curves (or how to solve discrete logarithms in $\mathcal{U}_{2^{4 \cdot 1223}}$ and $\mathcal{U}_{2^{12 \cdot 367}}$). <https://arxiv.org/abs/1402.3668>, 2014. CRYPTO 2014. Concrete estimates in the authors’ modular-multiplication units.
- [14] Literature-search project. Cryptanalytic literature corpus, author map, and effort estimates, 2026. Repository snapshot, 9 October 2026. Sources and rebuild instructions in Appendix A of this paper. Provisional, machine-classified results.
- [15] Guilhem Mureau and Alice Pellet-Mary. Dimension reduction for SVP in Hawk: A trace-zero approach. Cryptology ePrint Archive, Paper 2026/1560. <https://eprint.iacr.org/2026/1560>, 2026.
- [16] OpenAI. Sharing AI progress in mathematics. <https://openai.com/index/sharing-ai-progress-in-mathematics/>, 2026. 6 October 2026; accessed 9 October 2026.
- [17] OpenAI. On the Navier–Stokes millennium prize problem. <https://openai.com/index/navier-stokes-solution/>, 2026. 8 September 2026, updated 10 September; accessed 9 October 2026. Research claim, not independently adjudicated in this study.
- [18] OpenAI. Ten advances in mathematics and theoretical computer science. <https://openai.com/index/ten-advances-in-mathematics/>, 2026. 1 August 2026; accessed 9 October 2026.
- [19] OpenAI. Mathematics research repository: README and manuscript catalogue. <https://github.com/openai/math>, 2026. Accessed 9 October 2026; catalogue and verification status may change.
- [20] Zygimantas Straznickas and Stephen A. Weis. HAWK- n key recovery reduces to SVP in dimension $n/2 + 1$. Cryptology ePrint Archive, Paper 2026/1593. <https://eprint.iacr.org/2026/1593>, 2026.