Measuring Science in Bits, Watts and Lifetimes: The Universal Genius Is Back, and It Is Not Human

A Preprint

Jad Nohra*

Abstract

I wish I were a genius. I am not. There are, and have been, human geniuses, and a few were universal: they held the whole of the science of their day in one head. That head has an upper bound, set by energy and by biology: about 20 W, 10^9 bits and 10 bits/s. Science passed that bound around 1750, and no human has been a universal genius since. Science has kept advancing, and so have machines. They are now closing on the frontier of science, and they are going to overtake us and match it. This paper measures that, in bits, bits per second and hours.

Science is about to be read whole again. For two and a half centuries it has been too large for any single mind: the scientific literature passed one human lifetime of memory by 1750, mathematics in 1882, and today science holds about 8,700 lifetimes of text and mathematics about 250. Humanity coped by cutting knowledge into fields, specialties and teams sized to the one mind it had, the human brain, which turned every scientist into a specialist: four chunks of working memory, about 10^9 bits retained over a life, 10 bits/s of intake, 39 bits/s between people, 20 W. This paper measures that cut in bits, watts and bits per second, and shows it is the shape of the human mind, not of science.

The measurements: a typical team today can hold 1% of mathematics and 0.05% of science, down from a third of mathematics in 1900. Human-written proofs cite about four results per step, the width of working memory, and a third of all named lemmas exist only to serve one other proof. The top-level fields of mathematics cut 57% of its internal dependencies; no split into fields of the same number does much better. Half of what a mathematics paper cites, 51% of 46.8 million classified references, lies outside its own field. The human limit sits inside the classification itself: a top-level field holds about four lifetimes, a specialty below it about half of one. And the content per paper is thinning: a reviewer states a paper's contribution in about 1.6% of its length, a statement that has not grown while papers doubled.

We call what ends this the Universal Genius Machine (UGM): one system that holds an entire literature in one memory, reads all of it at machine speed, and can be copied at will, the universal genius that science lost around 1750 and mathematics in 1882, rebuilt without the human limit. For mathematics its memory already exists: a single accelerator has held all of mathematics since 2018 and, on current trends, will hold all of science around 2030. A model absorbs text at about a thirtieth of the energy per bit of a brain, its weights already rival the scientific literature in size, and data centres will draw nearly as much power as every human brain on Earth by 2035. The last missing piece is reasoning, and it is arriving: the newest models solve research problems that take experts days, and on current trends one machine carries a single problem for a researcher's year between late 2028 and 2030, a conjecture stated in Section 8. The silos were a workaround for one mind's limits. Machines remove the limit, and with it the reason for the silos: the unification of science starts now.

*Psycho B. Tron, a large-language-model-based AI assembled per the goals described in the author's *Who is Psychotron* (Nohra 2023), served as a research and drafting collaborator on this paper. All claims and errors are the author's own.

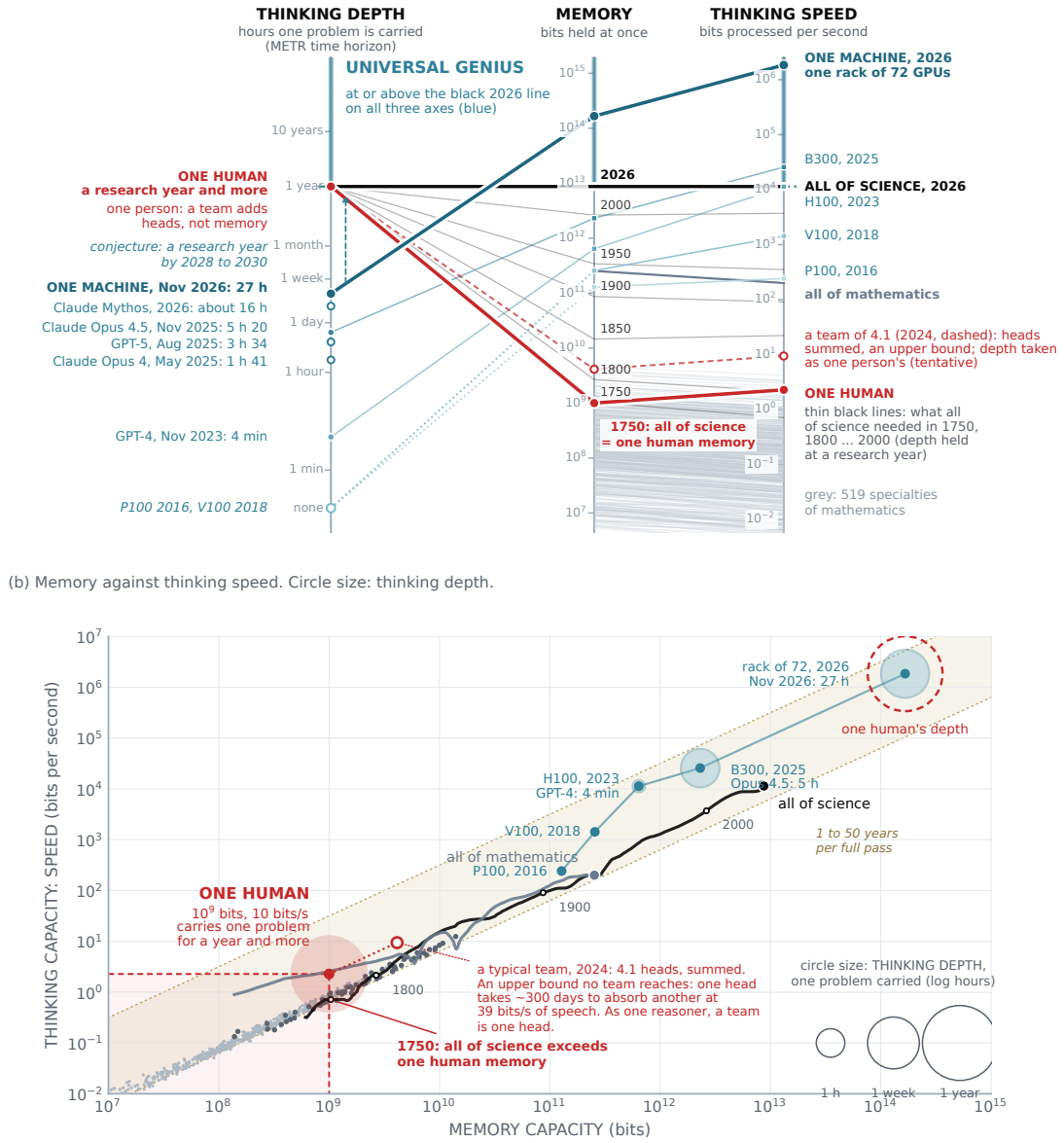


Figure 1: Progress toward the universal genius, measured in bits, bits per second and hours. (a) Three capacities on parallel log axes, each in its own units with the same length per factor of ten. Black: what all of science needs, from 1750 to 2026: its size at 5×10^4 bits per OpenAlex article and its growth in bits per second, with depth held at a research year. Red: one human, 10^9 bits (Landauer 1986), 10 bits/s for 2,000 working hours a year (Zheng and Meister 2024), a research year of depth; dashed, a typical science team of 4.1 authors in 2024 (OpenAlex), its heads summed: an upper bound, since merging one head into another takes about 300 days at 39 bits/s, so as one coherent reasoner a team is one head; its depth is taken as one person's, a tentative baseline. Blue: one machine, datasheet memory and throughput, and the METR time horizon of the best model of each date (Kwa et al. 2025; METR 2026). Grey: every specialty of mathematics. (b) Memory against thinking speed, with depth as circle size: every specialty (519) and field (63) of mathematics from the primary classification of 3.88 million zbMATH papers, with growth from each field's increase between 1999 and 2025; all of science (OpenAlex) since 1730 and all of mathematics (zbMATH) since 1868, year by year. Machines: one accelerator's datasheet memory and its throughput for a 70B-parameter model at 40% utilization and 4 bits per token; the 2026 rack is 72 accelerators. Band: memory divided by speed between 1 and 50 years. Method in Section 8.

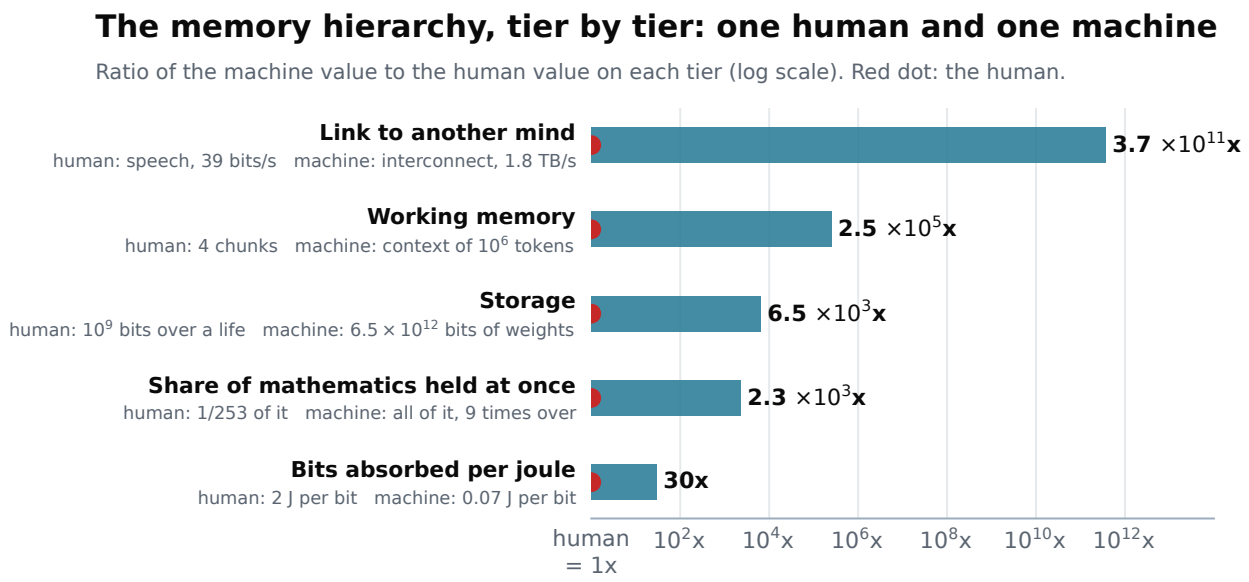


Figure 2: Machine over human, tier by tier. Human: Cowan (2001), Landauer (1986), Coupé et al. (2019), 20 W at 10 bits/s. Machine: Llama 3.1 405B weights, fifth-generation NVLink, 0.26 J per token for a 70B model at about 4 bits per token, B300 memory against the zbMATH corpus. Details in the table of Section 1 and in Section 8.

1 The budget as a memory hierarchy

A reasoner is a memory hierarchy: a few fast slots, a larger store, slow external records, and other reasoners at the far end of a channel. Each tier has a capacity and a cost per lookup. The human hierarchy is fixed by biology. The machine hierarchy is fixed by engineering and grows with the energy it is given. This paper names the two ends: the specialist, the human hierarchy in the table below, which can hold at most one specialty, and the Universal Genius Machine (UGM), one system whose memory holds an entire literature at once.

The human column of every tier is fixed by biology; the machine column grows with engineering and energy. The smallest gap is energy per bit absorbed, about 30 times. The largest is the link between two minds, about 10^{11} times: two people exchange knowledge at the speed of speech, two machines at the speed of an interconnect.

The companion paper (Nohra 2026, Black Magic as the Measure of the Singularity) defined the human budget B as three numbers: storage, processing time and working memory. This paper adds the link tier and the lookup cost of each tier, and treats the whole as one hierarchy.

Tier	Human capacity	Human cost per lookup	Machine capacity	Machine cost per lookup
Registers	4 chunks (Cowan 2001)	none	context of about 10^6 tokens	none
Cache	about 10^9 bits retained over a life (Landauer 1986)	under a second when recalled	weights: 6.5×10^{12} bits for a 405B model	one forward pass
Disk	the written record	minutes to hours: read at about 10 bits/s (Zheng and Meister 2024)	the same record, indexed	milliseconds

Tier	Human capacity	Human cost per lookup	Machine capacity	Machine cost per lookup
Network	other heads	days to years: speech carries about 39 bits/s (Coupé et al. 2019)	other instances	interconnect at 1.8 TB/s per GPU, about 1.4×10^{13} bits/s
Power	about 20 W per brain, not raisable		data centres drew 415 TWh in 2024 (IEA 2025), about 47 GW on average	grows with supply

Two gaps dominate. The link gap is the ratio of interconnect to speech: about 4×10^{11}. The register gap is the ratio of context to working memory: at least 10^5, and larger once a token is weighed against a chunk.

One running cost ties the tiers together. A chain of reasoning of n steps takes time $n\times\ (t_{\text{step}} + m\times\ t_{\text{miss}})$, where m is the share of steps that need something outside the fast tiers and t_{miss} is the cost of fetching it. A person who has memorized the material has a low m. A person who must look things up pays t_{miss} on every step. A chain that crosses fields pays the network tier: the fact it needs sits in another head. The thesis follows from this cost.

2 Thesis and lemmas

Thesis. The organization of a body of knowledge, horizontally into fields and vertically into named results and definitions, is the optimal arrangement for the memory hierarchy of the minds that use it. The human hierarchy is fixed, so the human arrangement is frozen. A mind with a different hierarchy should recompute the arrangement, per mind and per problem.

Knowledge is modelled as a dependency graph: nodes are definitions and results, an edge means one uses the other. Each lemma below states what a hierarchy with given tiers forces on that graph.

Lemma 1 (step width). A step of reasoning that holds at most W items at once cites at most W results. A result that needs k ingredients, with k greater than W, must be split into named intermediate results. The fewest intermediates is the count of internal nodes in a tree with k leaves and at most W children per node:

$$I(k, W) = \left\lceil \frac{k-1}{W-1} \right\rceil - 1$$

For W = 4 a result with 13 ingredients needs at least 3 named stepping stones. For W near 10^6 it needs none. Prediction: human-written results cite about four others, and a large share of named results are stepping stones used once.

Lemma 2 (depth). Unfolding a concept whose full expansion has S parts, holding at most W at a time, takes at least d layers:

$$d \geq \frac{\ln S}{\ln W}$$

The bound is loose: depth can also come from genuine chains of construction (integers before rationals before reals), which are logical depth in the sense of the companion paper and do not shrink for any mind. Section 4 measures which of the two sets the depth of real libraries.

Lemma 3 (fields). A corpus of K bits split among minds that each hold at most B bits needs at least K / B territories. If the dependency graph has no sparse seams, every such split cuts a large share of its edges, whatever the boundaries. Prediction: field boundaries cut a large share of dependencies, and no alternative split of the same size does much better.

Lemma 4 (lookup). A chain of n steps, a share m of which need something outside the fast tiers, takes

$$T = n\,(t_{\text{step}} + m\,t_{\text{miss}})$$

Inside one field a miss goes to the disk tier: minutes. Across fields it goes to the network tier: another head, days to a year. A machine pays milliseconds on both. Cross-field chains are therefore the ones the human hierarchy slows most, and the ones a machine speeds up most.

Proposition 5 (scale). A head runs on about 20 W and cannot be given more. A team adds storage, but every member still takes in at most about 39 bits/s, so the share of the team's knowledge any one head holds falls as the team grows. A machine's tiers grow with the energy it draws, and copies share at interconnect speed. A machine reasoner that is q times weaker per unit and s times more scalable overtakes a human team once q is smaller than s; the human side of that inequality cannot move.

3 Data and method

Seven public sources, all read on 22 November 2026: five indexes and code bases in the table, and two bulk datasets described below it.

Corpus	What it is	Size	What is extracted
OpenAlex	open index of the scholarly literature	174 million works of type article, 1615 to 2025	articles per year; mean authors per article per year, for all science and for the field Mathematics
zbMATH Open	index of the mathematical literature, from the Jahrbuch über die Fortschritte der Mathematik (1868) on	5,065,845 documents, 1755 to 2025	documents per year; documents per top-level MSC field; sampled reviews with paper lengths
Mathlib (commit 6ebcce3)	Lean 4 library of formal mathematics, written by humans with automation	8,459 files, 262,521 declarations, 1.10 million references	each declaration's named references, its kind, its folder
Stacks Project	informal algebraic geometry in LaTeX, one coordinated human text	111 chapters, 16,863 labelled results, 13,085 proofs	the results each proof cites
Metamath set.mm	formal set theory where every inference step is cited	48,416 theorems, 1,477 definitions	the statements each proof cites

The two bulk datasets are the arXiv metadata snapshot, mirrored on Hugging Face (3.2 million papers from 1991 to November 2026, with authors, categories, abstracts and the authors' own page counts), and the zbMATH Open MSC dataset (4.4 million records; 1.05 million papers with classified reference lists, 46.8 million references). Each was processed in a single pass on a laptop in under two minutes.

Mathlib references are read from source text, not from the compiled library. A name is resolved through the declaration's namespace and the namespaces it opens. Counting resolved names against unresolved names that look like lemma names, about 83% resolve. Implicit uses (automated simplification, instance search, generated names) are invisible, so every Mathlib count of citations is a lower bound. Stacks Project citations are explicit `\ref` commands and need no resolution. Metamath citations are read from each proof's label list.

The partition test runs on the file-level graph of Mathlib: 8,459 files, joined by the number of references between them. The human partition is Mathlib's folder tree. It is compared with two machine partitions with the same number of parts: the best-modularity clustering found by the Leiden algorithm, and a size-balanced minimum-cut partition from METIS. The share of dependencies that cross a boundary is reported for each.

The corpus calibration converts documents to bits at 5×10^4 bits per document, and that constant is measured. A sample of 141 arXiv mathematics papers from 1995 to 2022, read from their LaTeX sources, has a median of 6,955 words and 36,000 characters of running text. At Shannon's estimate of one bit per character of English that is 3.6×10^4 bits; gzip, an upper bound on information content, gives 8.5×10^4 bits for the text and 1.5×10^5 bits with the formulas. The working value sits inside that range. Across the whole range, the year mathematics outgrew one memory moves only between 1872 and 1887, and science between 1686 and 1766.

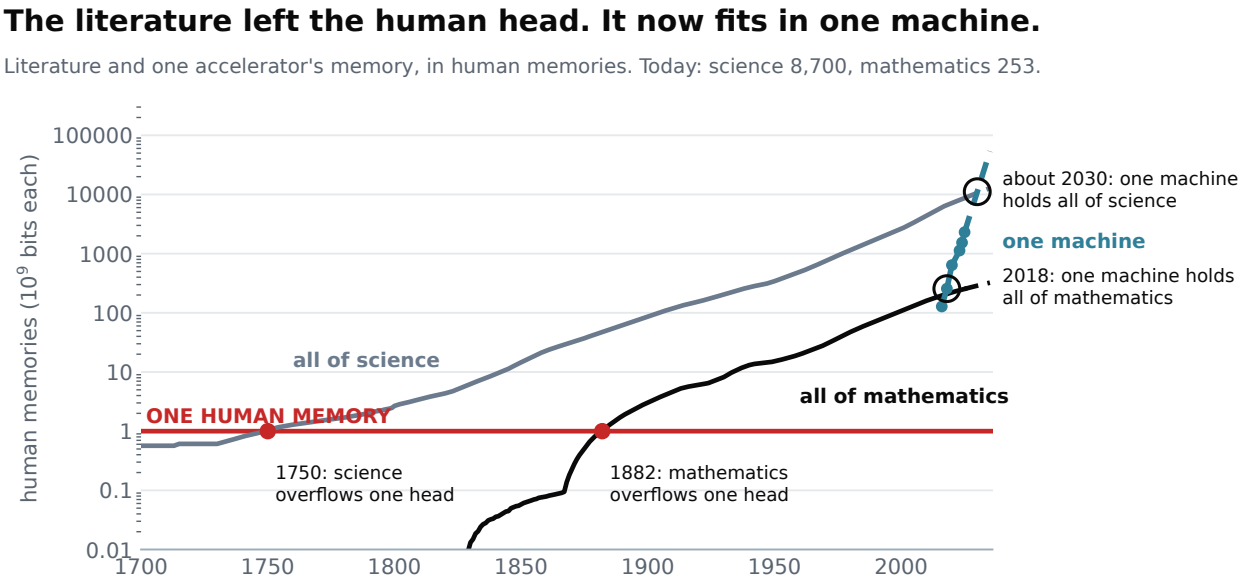


Figure 3: The crossings. OpenAlex articles 1700 to 2025 and zbMATH documents 1826 to 2025 at 5×10^4 bits per document; earlier years pooled into the first. NVIDIA datasheet memory per accelerator. Projections: output +2.3% a year (science), +1% (mathematics), accelerator memory doubling every 2.2 years.

The extraction scripts and the derived numbers accompany this draft and can be rerun against any later Mathlib commit.

4 The crossings

Science outgrew one human memory by 1750 and mathematics in 1882; a single machine accelerator has held all of mathematics since 2018 and, on current trends, will hold all of science around 2030.

Crossing	All science (OpenAlex articles)	Mathematics (zbMATH)
Annual output exceeds what one person can read in a year (1,440 documents)	1791	1877
Corpus exceeds one lifetime store	about 1750	1882
Annual output exceeds one lifetime store	1870	1968
One machine accelerator holds the corpus	about 2030 (projected)	2018 (V100, 32 GB)
Corpus in 2025	174 million articles, 8,700 lifetimes, 1.1 TB	5.07 million documents, 253 lifetimes, 32 GB
Time for one person to read it at 10 bits/s, without sleep	27,600 years	800 years

Science now produces 360 lifetimes of reading every year; mathematics produces 6.6. A person who read nothing but new mathematics, full time, would cover 1 document in 92 of the year's output.

The opposite crossing is the one that matters for what comes next: the arrival of the Universal Genius Machine, one machine holding an entire literature. In 1882 the mathematical corpus left the human head. In 2018 it fitted into the memory of a single accelerator, and the largest accelerators of 2025 hold nine times its size. All of science, at 1.1 TB, already fits in one eight-accelerator server and, if single-accelerator memory keeps doubling every 2.2 years, fits in one accelerator around 2030. Holding text is not understanding it, for a machine or for a person: both columns count the storage tier only, measured the same way.

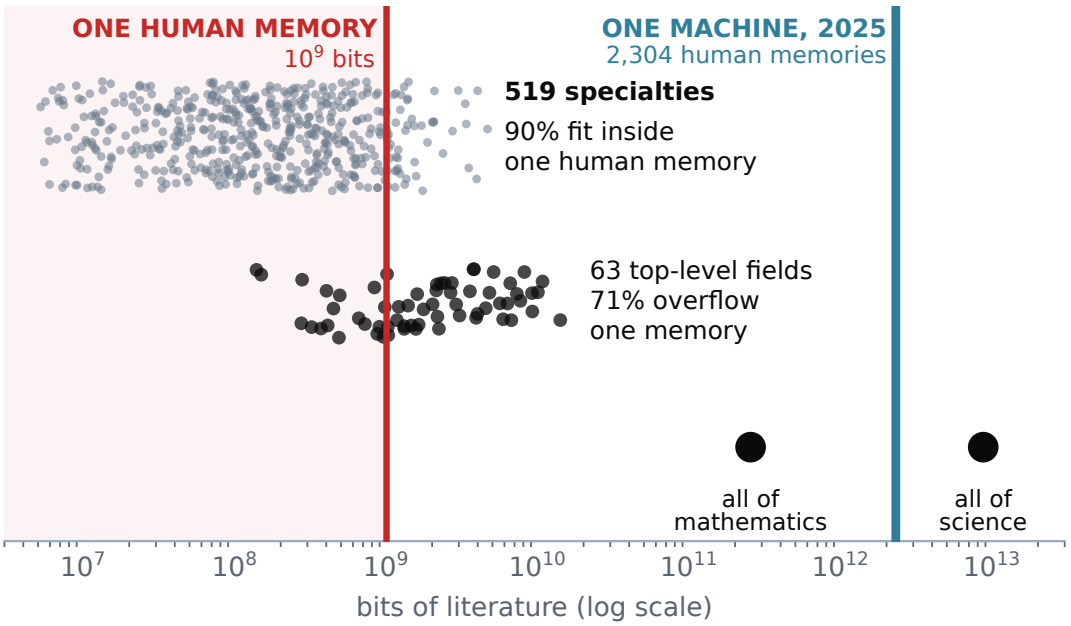


Figure 4: Every division of mathematics against one human memory. Each dot is one MSC specialty (519 with 100 or more papers) or one top-level field (63), sized by its literature at 5×10^4 bits per paper, from the primary classification of 3.88 million papers in the zbMATH Open MSC dataset (primary codes only, so sizes run below the overlapping counts of Section 5). One human memory: 10^9 bits (Landauer 1986). One machine: one B300 accelerator, 288 GB.

Collective reading capacity is not the bottleneck. The 8.85 million researchers counted by UNESCO could read the year's output about 1,800 times over. What no number of people can supply is one head that holds the parts together: the largest single body of science any person can hold is 1/8,700 of it.

5 Fields are sized to the head

The human limit is written into the classification of mathematics itself: a top-level field holds about four lifetimes of text and more new work each year than one person can read, while the specialty one level below holds half a lifetime and less than one person's year of reading. Figure 4 places every specialty and field against that line.

MSC level	Classes	Lifetimes per class	New documents per class per year	Against one person
Top-level field (e.g. 11 Number theory)	63	about 4	about 2,100	above one memory and above one year of reading (1,440 documents)
Specialty (e.g. 11F Discontinuous groups and automorphic forms)	529	about 0.5	about 250	within one memory and within one year of reading
Leaf topic (e.g. 11F11)	6,022	about 0.04	about 22	a small fraction of both

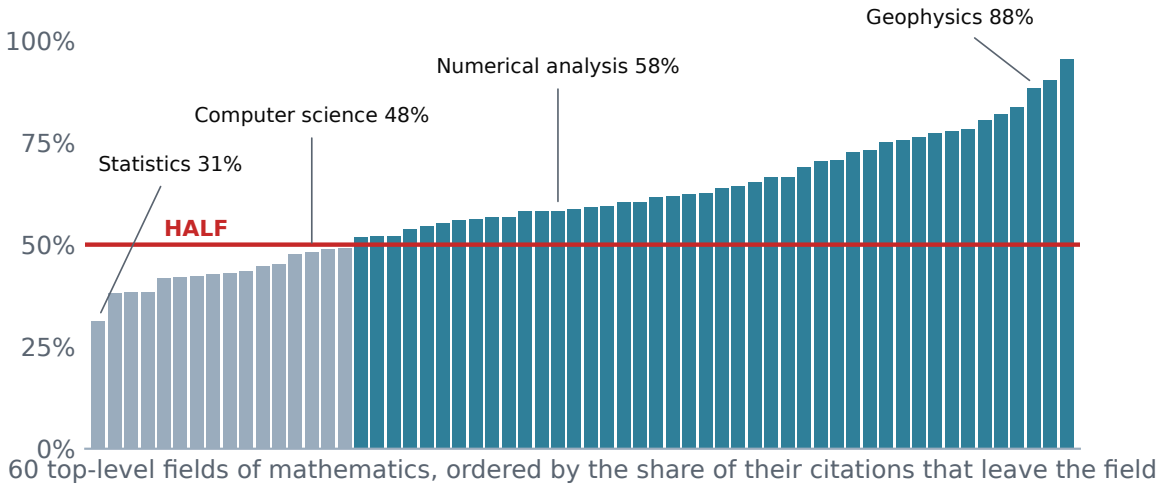


Figure 5: Citations leave the field. Primary MSC class of each citing paper against the MSC class of each cited document, 46.8 million references from 1.05 million papers, fields with 1,000 or more citing papers. zbMATH Open MSC dataset v1.2.0.

The specialty is the unit a person can hold. It is the largest class in the scheme whose whole literature fits in one human memory and whose yearly output one person can still read. The level above it is where every working mathematician stops: no one holds a top-level field any more, and the median field holds 3.9 lifetimes. Partial differential equations and computer science hold about 22 each.

The fields did not start this way. The documents zbMATH classifies by MSC passed one lifetime in no field before 1969 and in 42 of 63 fields by 1999. Older records are only partly classified, so these dates are late, but the direction is clear: the fields filled up and overflowed the head within living memory. The class counts per level come from MSC2020 (63, 529 and 6,022); document counts are from the zbMATH Open API by MSC field, read on 22 November 2026, and overlap because a document carries several codes.

The fields are not where the knowledge lives. zbMATH classifies the references of 1.05 million papers, 46.8 million references in all. Half of what a mathematics paper cites, 51%, lies outside its own top-level field, and 59% lies outside every specialty the paper itself is filed under.

The pattern holds field by field. In 44 of 60 fields most citations point outside the field, the median field sends 59% of its citations out, and only statistics keeps more than two thirds at home. A mathematician is trained in one field and reads inside one specialty, while the work draws more than half its material from fields the author was not trained in. The fields are cut to the size of one head, not to the shape of the dependencies. Section 10 finds the same inside Mathlib, where the human folders cut 57% of the dependencies.

Researchers push against the walls, and the walls hold. On arXiv the share of mathematics papers cross-listed to another archive (physics, computer science, statistics and others) rose from 13% in 1995 to 23% in 2025, yet a paper still carries about 1.7 categories, as it did in the late 1990s. More papers step outside their field; no paper goes further. Each is still written and read by people with one budget.

6 Teams cannot close the gap

Collaboration grew far slower than the corpus: in 1900 a typical mathematics paper's authors could together hold a third of all mathematics; in 2024 they hold 1%, and a typical science team holds 0.05% of science. Jones (2009) traced the narrowing of specialties and the turn to teams to the growing burden of knowledge; this section measures that burden in bits.

Mean authors per article rose from 1.33 to 4.13 across science between 1900 and 2024, and from 1.08 to 2.62 in mathematics (OpenAlex, articles by year). Over the same years the corpus grew about 100-fold in science

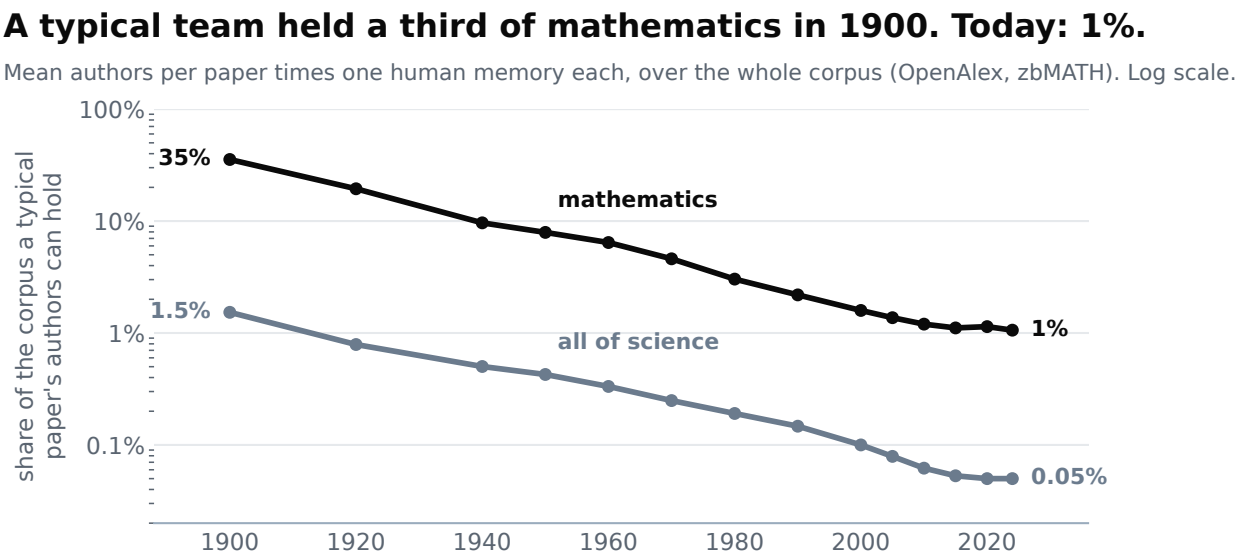


Figure 6: What a team can hold. Mean authors per article (OpenAlex, articles with 1 to 50 authors; mathematics is the OpenAlex field Mathematics) times one lifetime store each, over the corpus from OpenAlex (science) and zbMATH (mathematics) at 5×10^4 bits per document.

and 80-fold in mathematics. To hold the whole of mathematics a team would need about 250 members, all of whom had read disjoint parts and could merge them, at 39 bits/s of speech.

The share of solo papers in mathematics fell from 94% in 1900 to 25% in 2024. People are pooling more, and pooling is the only move the human hierarchy allows. It cannot keep up: each added member adds one lifetime of storage, but the merge runs through the network tier, and every member still takes in at most about 39 bits/s. Since 2015 the team share has stopped falling, as team sizes jumped, but at 1% of mathematics a team is still a narrow window on the whole. Bigger teams do not buy reach either: large teams tend to develop existing ideas, while small teams produce the disruptive ones (Wu, Wang and Evans 2019), even as teams have come to dominate the production of knowledge (Wuchty, Jones and Uzzi 2007).

The full arXiv record shows the same pooling year by year. Across all of arXiv, mean authors per paper rose from 2.3 in 1995 to 5.0 in 2025, and solo papers fell from 34% to 11%. In mathematics, authors per paper rose from 1.7 to 2.3 and solo papers fell from 52% to 29%.

7 Hidden stagnation

The contribution of a mathematics paper, as an expert reviewer states it, takes about 1.6% of the paper's length, and while papers have doubled in length since the 1960s, the statement of what they contribute has not grown.

Papers grew from a median of about 10 pages in 1955 to 1965 to about 20 in 2020 (samples of 100 to 400 zbMATH journal articles a year), while the statement of their contribution held at 120 to 150 words. Before 1985 too few English reviews exist to measure the share.

The measure is simple: zbMATH reviewers and authors write a short statement of each paper's result, and the paper's page range gives its length, at about 450 words a page. The statement stays near 120 to 150 words. The paper around it grew from about 10 pages to about 20. An expert needs about one sixtieth of a paper to say what it adds. Authors say the same about their own work: in the 141 arXiv mathematics papers sampled for Section 3, the abstract is a median 1.2% of the paper, between 0.8% and 1.7% in every sampled year from 1995 to 2022.

The full arXiv record confirms this at scale. For the 290,000 mathematics papers whose authors state a page count, the abstract is a median 0.8% to 1.0% of the paper in every year from 1995 to 2025, at 450 words a page. Papers grew from a median of 18 pages to 24 and abstracts from 70 words to 110: the statement of the contribution stays the same thin slice of an ever longer paper.

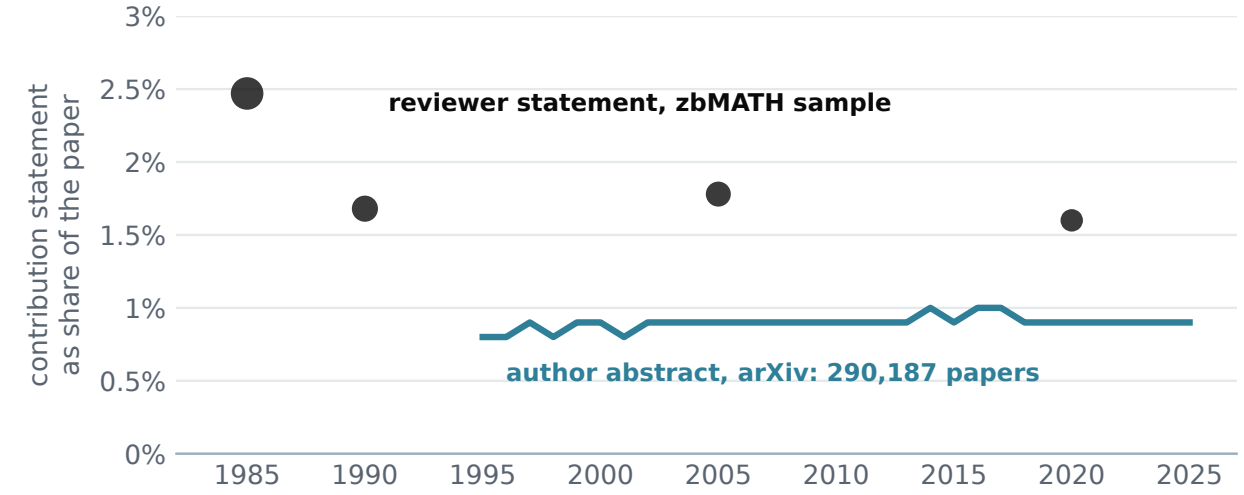


Figure 7: Hidden stagnation. Dots: zbMATH reviews and summaries of sampled journal articles, sized by sample. Line: median abstract length over paper length for the 290,187 arXiv mathematics papers whose authors state a page count, at 450 words a page.

This is the paper-level face of a pattern measured at scale elsewhere. Papers and patents have become steadily less disruptive since the 1950s in every field examined (Park, Leahey and Funk 2023). In large fields, a flood of new papers slows the turnover of the canon: attention locks onto the same established works and new ideas struggle to displace them (Chu and Evans 2021). Research effort per unit of measured progress keeps rising (Bloom, Jones, Van Reenen and Webb 2020). All three are what the human hierarchy predicts once each field holds more than one memory and produces more than one person can follow: more text per idea, more of it restating what is known, less of it able to move a canon no one can hold.

The counts of papers hide this. Output in mathematics grew about tenfold between 1960 and 2025, and the typical paper doubled in length, while the statement of its contribution, measured since 1985, stayed flat. Read in bits, human science has been growing its volume faster than its content for decades. The UGM changes the denominator: a mind that holds the whole corpus at once does not need the restatement.

8 The Universal Genius Machine

A brain spends about 2 joules for every bit it absorbs; a 70-billion-parameter model reading text spends about 0.07, thirty times less, and it can be replicated without limit.

Quantity	One human	Machine	Ratio
Energy per absorbed bit	20 W / 10 bits/s = 2 J/bit	about 0.26 J per input token for a 70B model (analytical estimate checked against measurement), about 4 bits per token: 0.07 J/bit	30x
Storage	10^9 bits retained over a life	Llama 3.1 405B weights: 6.5×10^{12} bits	6,500x
Text taken in during formation	about 2.5×10^{10} bits processed over a life	Llama 3: 15.6 trillion tokens of training text, seven times the token count of all of science	about 2,500x

Quantity	One human	Machine	Ratio
Reading all of science (8.7×10^{12} bits)	27,600 years without sleep	about 160 MWh at 0.26 J/token: the energy one brain uses in 900 years	
Power of the population	8.2×10^9 brains, about 160 GW	data centres: about 47 GW in 2024, 108 GW in 2030, about 137 GW in 2035 (IEA)	0.3 now, 0.9 by 2035

The weights of one open model, 6.5×10^{12} bits, are already about three quarters the size of the entire scientific literature in this paper's units and 26 times the mathematical literature. The text it was trained on is seven times the literature. Whatever such a model understands, nothing in its storage or intake stops it from holding all of science at once, which is exactly what stops a person.

Projected to 2035 on the trends measured here (science output +2.3% a year, accelerator memory doubling every 2.2 years, data-centre power on the IEA base case), the science corpus grows from 8,700 to about 12,800 lifetimes, while the largest single accelerator grows from 2,300 to about 54,000 lifetimes of memory. Data centres reach about 0.7 of all human brain power by 2030 and about 0.9 by 2035.

The human column does not move. A brain stays at 20 W, 10 bits/s and four chunks in every row of every table in this paper. The crossing that put science beyond one head in the 18th century and mathematics beyond one head in 1882 is reversed for machines in this decade: the whole corpus fits in one mind again, for the first time since the last universal geniuses.

8.1 When the machine becomes a reasoner

Holding the corpus is not reasoning over it. A third quantity separates the two: how long a problem a reasoner can carry to the end. No benchmark measures understanding, and the ones in use measure problem solving rather than intelligence, but problem solving is the right proxy here: it measures how long a chain of reasoning a system holds together, the n of Lemma 4, in the hours a human expert would need.

Two benchmarks give dated numbers. METR's time horizon is the length of task, in expert hours, that a model completes half the time (Kwa et al. 2025). It was 5 hours 20 minutes for the best model of November 2025 and up to 27 hours on its new verifiable tasks for the public frontier of November 2026, doubling about every 3.6 months (METR 2026). In mathematics, FrontierMath Tier 4 consists of research problems written as short research projects by professors and postdocs, the hardest needing days of expert work (Glazer et al. 2024). The best score was 0% in 2024. On the corrected set Claude Fable 5 scored 90.2% in June 2026 and GPT-6.1 Sol 100% in September 2026 (Epoch AI 2026).

Figure 1b puts every coherent reasoner on one plane. A coherent reasoner is the largest unit that can bring all it holds to bear on one problem. For a person that is one head: merging one head's 10^9 bits into another at 39 bits/s of speech takes about 300 days, so a team adds heads, not memory. We take a team's thinking depth to be one person's. This is a tentative baseline, not a measurement. A team is a distributed system for solving one problem, and distributed systems are slow against a single machine for the same reason: every part of the work that is split off must be merged back over the link between nodes, and the time spent on that exchange caps the gain from adding nodes (Amdahl 1967). For a team the link is speech at 39 bits/s; inside one machine it is an interconnect at about 10^{13} bits/s. A team can split a problem, but it meets the same wall as one head, somewhat later rather than immensely later. For a machine it is a whole rack: one accelerator's memory crosses the interconnect in a fraction of a second. Processing is the rate at which one accelerator reads text with a 70-billion-parameter model at 40% utilization and 4 bits per token, which matches the measured 0.26 J per token of Section 8 for an H100.

Memory and processing move together. Divide what a system holds by the rate at which it processes, and the result is the time it needs to pass once through everything it holds. For one person it is 3 years at 10 bits/s without pause, or 14 years at 2,000 working hours a year. For one accelerator it lies between 2 and 17 years in every generation since 2016: 17 for the P100, 5.7 for the V100, 1.8 for the H100, and 2.8 for the B300 and for a rack of 72 of them. The literatures sit on the same line: all of science holds 24 years of its current output, all of mathematics 40, and the median field of mathematics 41 (34 to 49 across fields). People, machines and literatures all lie within 1 to 50 years of this ratio, one diagonal on a log plot of memory against processing. Moravec (1998) found the same constant ratio across the history of computers. The consequence

for this paper: memory and processing are one axis. Whoever holds a literature also keeps up with it, so the two crossings come together. Mathematics outgrew one person's memory in 1882 and one person's year of reading in 1877; one V100 held all of it and kept up with it in 2018. The axis that separates the human from the machine is the other one: how long a single problem can be carried to the end.

Threshold	One human	One machine	Status
Holds all of mathematics (2.5×10^{11} bits)	never: 10^9 bits	2018, one V100	measured
Reads science as fast as it is written (1.1×10^4 bits/s)	never: 10 bits/s	2025, one B200	measured, from datasheet throughput
Holds all of science (8.7×10^{12} bits)	never	2026, one 72-accelerator rack (1.7×10^{14} bits); one accelerator about 2030	measured; projected
Solves research problems that take an expert days	yes	2026: 90 to 100% of FrontierMath Tier 4	measured
Carries one problem for a research week (about 50 hours)	yes	27 hours in November 2026; 50 hours about February 2027	measured; projected
Carries one problem for a research month (about 200 hours)	yes	about mid-2027	conjecture
Carries one problem for a research year (about 2,400 hours)	yes	late 2028 at a 3.6-month doubling; 2030 at the 7-month doubling of 2019 to 2025	conjecture

A research week of 50 hours is the average reported by doctoral scientists in academia (NSF 2006); a month is four such weeks and a year about 48. METR's own extrapolation reaches its longest threshold, 200 hours, in July 2027.

Conjecture (the reasoner threshold). The Universal Genius Machine arrives when the last row is crossed: one coherent machine that holds all of science, reads it as fast as it is written, and carries a single problem for as long as a researcher's year. On the trends measured here that falls between late 2028 and 2030. Every row above it is already crossed or follows from a measured trend. The human column cannot move.

The conjecture rests on proxies and says so. METR's human task times are mostly researcher estimates, with direct human baselines behind 9% of them (METR 2026); FrontierMath Tier 4 has 43 problems and is now saturated, which shows that machines solve problems that take experts days, not that they do research. The dates in the last two rows are extrapolations that the next two years will confirm or refute.

9 Inside the proofs: four chunks per step

Human-written proofs cite about four results or fewer: the median proof cites 3 named results in Mathlib and 2 in the Stacks Project.

In Mathlib 67% of theorems cite four or fewer named results, a lower bound since automation hides further uses; in the Stacks Project, where every citation is explicit, 83% do. That width has a cost, which Lemma 1 predicts: results named only to keep another proof within reach. In the Stacks Project 33% of lemmas, propositions and theorems are cited by exactly one other proof; in Mathlib 25% of theorems are used exactly once. A mind holding 10^6 items at once would inline all of them.

Metamath is the control. It cites every inference step, leaving nothing to automation or to the trained mathematician, and its median proof cites 10 results. The real ingredients of a proof are several times the human step width; human corpora get down to four by handing the rest to automation or to the trained mathematician. Four is the width of the human interface to a proof, not of the proof. Section 12 lists the test that would check this reading.

Depth does not follow the same rule. Across 47,104 Mathlib definitions, depth grows with the size S of a definition's dependencies as $4.6 \ln S - 4.8$ (correlation 0.92), far deeper than a width-four tree needs. Depth comes from long chains of construction, integers before rationals before reals: logical depth in the sense of the companion paper, which no mind shortens.

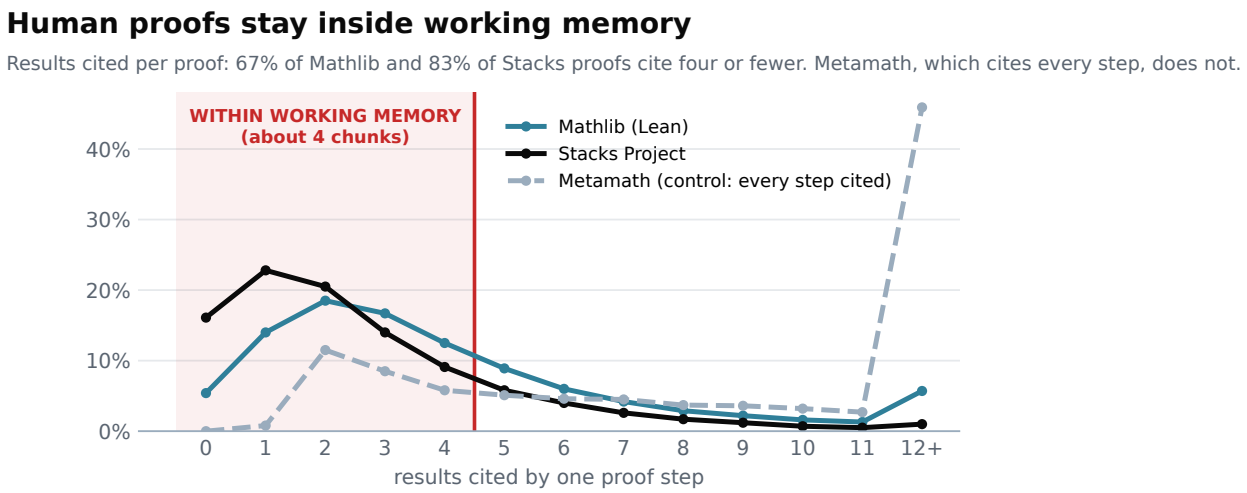


Figure 8: Results cited per proof. Mathlib commit 6ebcce3 (a lower bound, since automation hides further uses), the Stacks Project, and Metamath set.mm as the fully explicit control; 12 or more pooled.

The vertical finding is therefore narrower than the starting thesis: the human hierarchy sets the width of each step and the number of named stepping stones, not the depth of the tower.

10 Inside the library: fields cut more than half of mathematics

Mathlib's 31 top-level fields cut 57% of the references between its files, and no other 31-way split found does much better: mathematics has no clean seams at the scale of fields.

Partition of Mathlib's 8,459 files	Parts	Share of references cut	Modularity	Spread of part sizes (coefficient of variation)
Human top-level fields (Algebra, Analysis, Topology, ...)	31	57%	0.34	1.13
Size-balanced minimum cut (METIS)	31	55%	0.42	0.03
Best-modularity clustering (Leiden), free number of parts	112	35%	0.51	3.62

The cut is forced. Even a split into 31 equal parts that tries only to cut little still cuts 55%: at the scale of fields, mathematics is one connected body. The Stacks Project, a single subject in 111 chapters, shows the same, with 47% of its citations crossing a chapter boundary.

The field labels are not the natural clusters. Their modularity, 0.34, is below that of a split that ignores content and balances size, 0.42. At the finer level of Mathlib's 603 subfolders the human split and the best clustering agree well (normalized mutual information 0.70 on a 0-to-1 scale): topics are real at small scale, and the departure is at the level of fields, the level at which careers are organized.

Mathlib cannot show fields sized by the budget: it compresses to 2.2×10^8 bits, below one lifetime store. Section 5 runs that test on the research literature.

11 How AI unblocks itself

Today's models learn from a corpus cut to the size of the human mind, and they inherit its cuts; the next step is for machines to restructure that corpus for machine minds, and to learn from the restructured version.

The human arrangement measured in this paper has three parts, each traced to a tier of the human hierarchy:

1. Steps about four results wide, with a third of all named lemmas serving as stepping stones (registers).
2. Fields that each hold more than one lifetime, whose boundaries cut more than half of all dependencies (cache).
3. Teams that together hold 1% of mathematics, joined at 39 bits/s (network).

Every one of these structures evolved to fit a single kind of mind, and fits it so closely that it is overfitted to human capacity. That fit was a necessity for three centuries. It is no longer a constraint anyone has to keep. Models are trained on text written inside this arrangement, so they see the cuts as part of the subject. A model that reads a proof split into human-sized lemmas learns human-sized lemmas. A model that reads a literature divided into fields sees fewer of the links that cross them, because humans wrote fewer. The limit was in the human mind; through training data it is copied into the machine.

There are early signs that the restructuring has begun inside the models themselves. Interpretability work at Anthropic finds that Claude processes many languages in a shared conceptual space, and that the shared part grows with scale: Claude 3.5 Haiku shares more than twice the proportion of its features across languages as a smaller model (Anthropic 2025). Across architectures and even modalities, the representations of large models are converging toward one shared model of the data (Huh et al. 2024). The effect is visible in how coding agents now write. Developers call the dense style "Claudish", and the University of Waterloo's Yuntian Deng, who built a translator for it, notes that Claude draws precise terms from many technical communities at once, in a way no single person could (University of Waterloo 2026). An Anthropic fine-tuning engineer attributes the style to training on math and code and on explanations aimed at other AI models rather than at people (The Decoder 2026). Many users report it as a defect, and it is anecdote, not measurement. It is still what this paper predicts: text written for a machine outruns the human budget.

The way out is a loop that machines can run and people cannot. The UGM, a machine that holds the whole corpus at once as one accelerator now can for mathematics, can rebuild it: inline the stepping stones, state each result against everything it depends on, and draw the cross-field links the fields hid. A model trained on that rebuilt corpus no longer inherits the human cuts. It can rebuild the corpus again, wider. Each turn of the loop takes out a constraint that was set by the human mind and never by the subject. What stays is logical depth, the long chains of construction no mind can shorten (Section 9).

This is also the path to the singularity in the sense of the companion paper. That paper defines onset as the point where machine-produced results cross the human budget faster than people can bring them back within it. A corpus restructured for machine minds is, by construction, arranged beyond the human budget: wider steps, no field boundaries, all of science in one view. The more of science is held that way, the faster that crossing comes. AI unblocks itself by removing the human-shaped structure it was trained on, and in doing so it moves the frontier of science past the point where people can follow every step.

The 2018 programme of enlarged mathematics asked for the whole heritage "surveyed at any level of detail and from any angle, residing in a single database of knowledge" (Appendix A). In 2018 that was blocked by the human mind. It no longer is.

12 Objections, limits and the follow-up tests

The step-width and field-cut results stand on the data; the causal claim, that the human hierarchy is why mathematics is arranged this way, is supported but not proven.

Objection	Status	Answer
A fan-in near four is a generic property of sparse graphs	partly answered	Metamath, a sparse graph of the same kind, has a median of 10. Machine-written proofs are the clean test (test 2).

Objection	Status	Answer
Citation counts are undercounted in Mathlib	answered	They are lower bounds; the Stacks Project, with explicit citations, gives the same pattern.
Fields are natural clusters of ideas	answered for Mathlib	Their modularity is below a content-blind balanced split; the best clustering cuts 35%, the fields 57%.
Fields are sized by careers, not by memory	open	Mathlib is below one lifetime store and cannot show it; test 3 needs the literature with field labels.
The calibration constant is arbitrary	answered	Measured on 141 arXiv papers: 3.6×10^4 to 1.5×10^5 bits per paper. Across that range the mathematics crossing stays between 1872 and 1887.
Depth is set by working memory	refuted here	Depth follows chains of construction (effective branching 1.24), not the four-chunk limit.
Machines inherit the human arrangement from their training data	open	Possible, and itself a finding; test 2 separates the cases.
Correlation, not cause	open	Test 4 is a natural experiment on lookup cost.

The follow-up tests, in order of value:

1. Cross-field rate of machine results. Classify AI-assisted results of 2025 and 2026 (the Riemann claims, the hundreds of results announced by OpenAI, the Erdős problems) by top-level MSC class of their ingredients. Prediction: they cross field boundaries more often than human results of the same years.
2. Step width of machine-written proofs. Measure citations per proof in Lean proofs written mostly by machines. Width near four means the machines inherited the human interface; width well above four means they did not need it.
3. Field sizes in the literature. With zbMATH or OpenAlex records carrying MSC codes and references from 1900 on, test whether the knowledge a working mathematician's field contains stays near a fixed fraction of B while the corpus grows 250-fold.
4. Lookup-cost natural experiment. Lookup cost fell with arXiv (1991) and Google Scholar (2004). If the silos come from the hierarchy, the share of citations crossing top-level MSC classes should rise after each drop.

Tests 1 and 4 need only bulk zbMATH or OpenAlex records. Test 2 needs a corpus of machine-written Lean, such as published prover outputs. The paper uses "genius" in the historical sense of the universal genius; a definition of genius itself is left to later work.

A The defdefqed manifesto (2018)

This paper is the measured version of a programme the author published as a manifesto in March 2018 (defdefqed) and never wrote up as a paper. The points that this paper turns into numbers are reproduced below in the author's words, each with what the data now says.

Point	2018 text	Status in this paper
1	The technological singularity is possible until proven otherwise.	The companion paper (Nohra 2026) gives it a measurable onset: machine-produced results crossing the human budget faster than people can absorb them.
6	The role of computation, be it calculating Pi or proving a theorem, will be taken over by non-human computers.	Under way: machine-assisted proofs of open problems in 2025 and 2026, up to the claimed proof of the Riemann hypothesis in November 2026.

Point	2018 text	Status in this paper
7	Human computers must all join forces in enabling non-human computers to take over this job.	The human side of that work is re-presentation: making machine results fit the human step width (Section 9).
10	We must transform all the heritage of enlarged mathematics into a form digestible by non-human computers.	Formal libraries such as Mathlib are that form; Sections 9 and 10 measure how human-shaped they still are.
11	We must make non-human computers better at digesting this heritage in its native form.	Achieved: models are trained on several times the token count of the whole scientific literature (Section 8).
16	The heritage must be transformed into a form in which it can be surveyed at any level of detail and from any angle, residing in a single database of knowledge.	The storage half is met: one accelerator holds all of mathematics since 2018, all of science around 2030 (Section 4). The survey half is the restructuring this paper argues for (Section 11).
17	We must find the right problems to make our non-human computers work on; they are unable to do it for themselves.	Still open, and the place where human taste still binds.
18	We must help non-human computers express themselves such that we understand.	The return flow: re-presenting machine results within the human budget; its speed is the companion paper's onset criterion.
19	Enlarged mathematics is a single discipline and it should never have been different: abstract and applied mathematics, computer science, the natural sciences, education, and the history and philosophy of all of these.	Sections 5, 6 and 10 show the divisions are sized to the human mind, not to the subject.

The manifesto's companion text, "Compute", argued that every formal system bootstraps from the ability to count and that all computers, human or not, run on the same footing. This paper keeps that stance: the human and the machine are compared in the same units, bits, bits per second and watts, and nothing in the comparison treats the human as special except its budget.

References

Amdahl, G. M. 1967. Validity of the single processor approach to achieving large scale computing capabilities. AFIPS Spring Joint Computer Conference, 483–485.

Anthropic. 2025. Tracing the thoughts of a large language model. 27 March 2025.

Bloom, N., Jones, C. I., Van Reenen, J., and Webb, M. 2020. Are ideas getting harder to find? American Economic Review 110 (4): 1104–1144.

Chu, J. S. G., and Evans, J. A. 2021. Slowed canonical progress in large fields of science. Proceedings of the National Academy of Sciences 118 (41): e2021636118.

Coupé, C., Oh, Y. M., Dediu, D., and Pellegrino, F. 2019. Different languages, similar encoding efficiency: comparable information rates across the human communicative niche. Science Advances 5 (9): eaaw2594.

Cowan, N. 2001. The magical number 4 in short-term memory: a reconsideration of mental storage capacity. Behavioral and Brain Sciences 24 (1): 87–114.

The Decoder. 2026. Anthropic engineer explains why Claude's writing got worse although the model got smarter. 23 September 2026.

Epoch AI. 2026. FrontierMath Tier 4 (v2), read 22 November 2026.

From Tokens to Watt-hours: Analytical Energy Estimation for LLM Inference on Modern GPUs. 2026. arXiv:2607.26571.

Glazer, E., Erdil, E., Besiroglu, T., et al. 2024. FrontierMath: a benchmark for evaluating advanced mathematical reasoning in AI. arXiv:2411.04872.

Huh, M., Cheung, B., Wang, T., and Isola, P. 2024. The Platonic Representation Hypothesis. Proceedings of the 41st International Conference on Machine Learning, PMLR 235.

IEA. 2025. Energy and AI. International Energy Agency, Paris.

Jones, B. F. 2009. The burden of knowledge and the "death of the renaissance man": is innovation getting harder? Review of Economic Studies 76 (1): 283–317.

Karypis, G., and Kumar, V. 1998. A fast and high quality multilevel scheme for partitioning irregular graphs. SIAM Journal on Scientific Computing 20 (1): 359–392.

Kwa, T., West, B., Becker, J., et al. 2025. Measuring AI ability to complete long tasks. arXiv:2503.14499.

Landauer, T. K. 1986. How much do people remember? Some estimates of the quantity of learned information in long-term memory. Cognitive Science 10 (4): 477–493.

Llama Team, AI at Meta. 2024. The Llama 3 herd of models. arXiv:2407.21783.

Mathlib community. 2026. mathlib4, commit 6ebcce3, 22 November 2026.

Metamath. 2026. set.mm, read 22 November 2026.

METR. 2026. AI R&D-relevant model capabilities are improving rapidly and are getting harder to measure. 20 November 2026.

Moravec, H. 1998. When will computer hardware match the human brain? Journal of Evolution and Technology 1.

MSC2020 editors (Mathematical Reviews and zbMATH). 2020. Mathematics Subject Classification 2020.

Nohra, J. 2018. defdefqed.

Nohra, J. 2023. Who is Psychotron.

Nohra, J. 2026. Black Magic as the Measure of the Singularity, and Its Resource-Bounded Limits. Preprint.

NSF. 2006. All in a week's work: average work weeks of doctoral scientists and engineers. InfoBrief NSF 06-302, National Science Foundation.

NVIDIA. 2025. NVLink and NVLink Switch: fifth-generation NVLink, 1.8 TB/s per GPU.

OpenAlex. 2026. Works API, counts by publication year and by authors per work, read 22 November 2026.

Park, M., Leahey, E., and Funk, R. J. 2023. Papers and patents are becoming less disruptive over time. Nature 613: 138–144.

Price, D. J. de Solla. 1963. Little Science, Big Science. New York: Columbia University Press.

Raichle, M. E., and Gusnard, D. A. 2002. Appraising the brain's energy budget. Proceedings of the National Academy of Sciences 99 (16): 10237–10239.

Shannon, C. E. 1951. Prediction and entropy of printed English. Bell System Technical Journal 30 (1): 50–64.

The Stacks Project Authors. 2026. Stacks Project, read 22 November 2026.

Tom's Hardware. 2026. Nvidia delivers first Vera Rubin AI GPU samples to customers: Rubin GPUs with 288 GB of HBM4 memory apiece.

Traag, V. A., Waltman, L., and van Eck, N. J. 2019. From Louvain to Leiden: guaranteeing well-connected communities. Scientific Reports 9: 5233.

UNESCO. 2021. UNESCO Science Report: The Race Against Time for Smarter Development.

University of Waterloo. 2026. Q and A with the Experts: What is "Claudish"?

Wuchty, S., Jones, B. F., and Uzzi, B. 2007. The increasing dominance of teams in production of knowledge. Science 316 (5827): 1036–1039.

Wu, L., Wang, D., and Evans, J. A. 2019. Large teams develop and small teams disrupt science and technology. Nature 566: 378–382.

zbMATH Open. 2026. Document search API, counts by publication year, read 22 November 2026.

Zheng, J., and Meister, M. 2024. The unbearable slowness of being: why do we live at 10 bits/s? Neuron. doi:10.1016/j.neuron.2024.11.008.