NM

PRIMARY RESEARCH · 7 SEPTEMBER 2026

How AI Changed the Way People Search

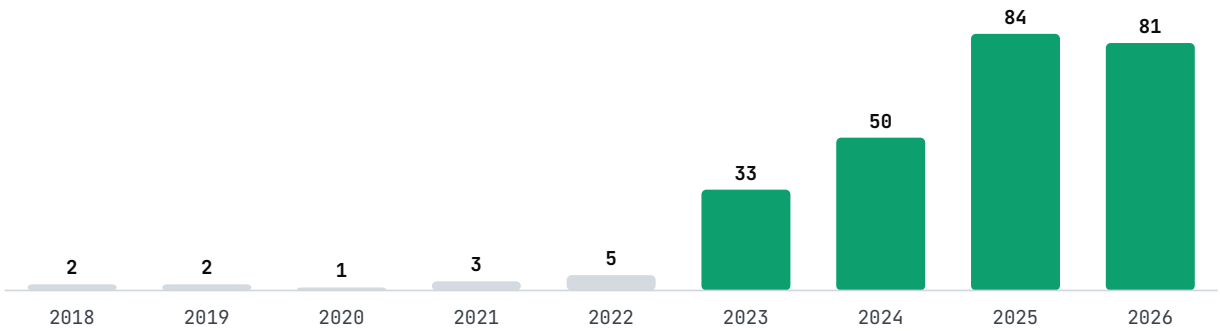
What 261 model releases from 15 labs did to consumer behaviour, search demand and SEO

The old way was a query, ten links and a click. The new way is a question answered where you already are. This report dates that switch, measures what it cost, and says what happens next.

NM

Nathan Mzumara

Head of SEO & AI Search · AI search, GEO and pipeline attribution



Model releases per calendar year, 2018 to 2026. 2026 covers 250 days to 7 September 2026, so the final bar is eight months of a year running at 118.3 a year. Grey bars pre-date ChatGPT. Linear axis: the pre-2023 counts are printed above bars too small to read.

261	15	2.98	68.0%	14.2x
MODEL RELEASES	LABS TRACKED	DAYS BETWEEN RELEASES	SEARCHES, NO CLICK	COHORT DIVERGENCE

Every figure computed from source data and independently re-derived · captured 7 September 2026 · nathanmzumara.com

KEY TAKEAWAYS

Six findings, and the numbers under them

If you read one page, read this one. Each figure below is computed from the 261-release dataset, from live Ahrefs index data, or from a named primary study – never from a statistics aggregator.

2.98

A new model every three days, and the gap is still closing

Across 15 labs the mean gap between releases fell from 9.12 days in 2023 to 2.98 in 2026, a 3.1-fold compression. 95.0% of the dataset shipped in the last four years.

75.5%

Three quarters of all models ever announced are already dead

197 of 261 releases are superseded or retired and 62 are still the model their vendor is selling; the remaining 2 were announced and never shipped. Anything you build on has a median shelf life measured in weeks, not product cycles.

12x

The floor collapsed while the ceiling rose

The cheapest capable model fell from $0.60 to $0.05 per million output tokens, a 91.7% collapse. Frontier pricing went the other way, up 25%. Cheap intelligence is what put an assistant in front of everyone.

58.3%

Most of the adoption numbers you have read are not chosen use

7 of the 12 products with a headline user count describe people who were given an assistant, sold a subsidy or counted through a legacy product – not people who chose one. Only 2 of the 15 adoption figures are audited.

68.0%

Search did not shrink. It stopped sending people anywhere

US Google searches ending without a click rose from 60.45% in 2024 to 68.01% in 2026. Across a 12-query basket the open web now receives 0.57 clicks per search.

14.2x

The collapse is not general. It is surgical

Queries an assistant can answer inside the user's own workflow fell 64.2% from their peak year. Physical, local and comparison queries fell 4.5%. Same index, same window, 14.2 times the damage.

THE SHIFT, IN ONE TABLE

The old model of search, and the one that replaced it

Everything in this report is a version of the contrast below. The left column is how search worked for about twenty years. The right column is what the data in Parts II and III says is there now. Every right-hand claim is measured later in the report, and none of them is a forecast.

THE OLD MODEL · ROUGHLY 2004 TO 2023	WHAT REPLACED IT · MEASURED IN THIS REPORT
You typed keywords into a box, in the words the engine understood.	**You describe the problem, or photograph it. Text-only releases fell from 25 in 2023 to 0 in 2026.**
Ten blue links came back and you chose one.	**One assembled answer comes back. 68.0% of US searches now end without a click.**
A search produced a visit. Traffic was the unit of demand.	**A search produces 0.57 clicks across a 12-query basket. Demand and traffic have come apart.**
Every query with volume was worth chasing.	**Only queries an assistant cannot finish. Substitutable queries fell 64.2%; the rest fell 4.5%.**
You ranked on the page. Position one was the prize.	**You are cited in the answer, or you are not in it at all. There is no position two in a paragraph.**
Authority meant backlinks, domain age and technical hygiene.	**Authority means being quotable. Reddit is cited 149x more often than Stack Overflow.**
Traffic was the proof that search was working.	**Google's search revenue grew 17% in the year clicks per search fell 19.1%.**
Search happened at one address, when you decided to go there.	**The assistant is inside the tools people already had. 41.2% of headline AI users were given one rather than choosing it.**

In row order, the working is in sections 10, 16, 17, 18, 19, 19, 21 and 14. Nothing above is a forecast: the predictions are section 22 and are labelled as predictions.

Two warnings about reading a table like this one. It compresses eight years into a binary, and reality ran as a slope: the click was already softening before the first assistant shipped. And the right-hand column is not uniformly distributed. Section 18 is the whole reason this matters – the new model has arrived completely for one class of query and barely at all for another, and which class your demand sits in decides whether this report describes an inconvenience or an extinction.

CONTENTS

What is in this report

Twenty-nine sections in five parts. Part II is the 261-release dataset on its own, and is the longest part of the report. Part III is what that did to search, and it leans on two supporting sources rather than the workbook. If you have five minutes, read the key takeaways on page 2, then sections 07, 18 and 22.

01

Why I built this

Every week somebody sends me a chart showing that AI has killed search, and every week somebody else sends me a chart showing search revenue at an all-time high. Both charts are real. I wanted to know which one describes the world the rest of us are actually living in.

So I did the boring thing. I took a workbook of every significant large language model released between 11 June 2018 and 3 September 2026 – 261 of them, from 15 labs, from GPT-1 to GPT-6 Astra – and I read it against what has actually happened to search demand in the same window. Not opinion about what AI will do. Two datasets, one timeline.

The reason to put those two things side by side is that the AI industry narrates itself through capability and the search industry narrates itself through traffic, and neither conversation is honest about the other. A lab announces a context window. An agency announces a click-through rate. Nobody puts a date on the release that moved the click-through rate.

261	15	8.2 yrs	12
MODEL RELEASES IN THE PRIMARY DATASET	LABS, WESTERN AND CHINESE	FROM GPT-1 TO GPT-6 ASTRA	KEYWORDS IN THE CLICK-LOSS BASKET

Who this is for

Founders deciding whether to build on a model or around one. CMOs whose organic line has been flat for eighteen months while their brand has never been more visible. CEOs being asked to approve budget for a channel that does not appear in last-click. Growth leaders who need to know which half of the funnel is actually broken. And anyone who is simply curious about where search goes next and would rather read the release data than the commentary about it.

It is not for people who want to be told AI is over, and it is not for people who want to be told search is over. Both of those readers will find something to be annoyed about in Part II, which is roughly the point.

> THE QUESTION THIS REPORT ANSWERS
>
> Not *is AI replacing search*. That question has no answer because it has no denominator. The question is: **which searches has AI taken, how much did they cost to lose, and what does the release data say about how fast the rest goes.**

02

What is in the dataset

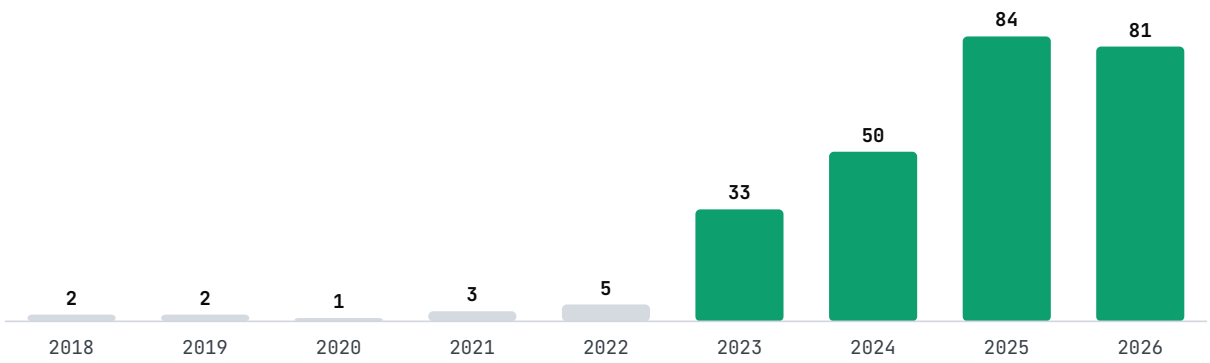
Three sources, all of which I can hand you. A model-release workbook, a live pull from a commercial search index, and six named outside sources. They are not equal. The workbook is the primary dataset and carries Parts I and II on its own; the other two exist only to answer the one question the workbook cannot, which is what all this did to search. No statistics aggregators, for a reason I will come back to.

1. The release workbook, which is the spine of this report

261 rows, one per significant model, compiled from vendor documentation, launch posts, API release notes and Wikipedia. Each row carries release date, lab, family, category, context window, weights licence, list price per million tokens at launch, and status.

Releases per year, and how much of the field is recent

All 261 rows. 2026 covers 250 days to 7 September 2026 and is therefore partial. Linear axis: the pre-2023 counts are printed above bars too small to read.



95.0% of the dataset shipped from 2023 onwards; the five years before produced 13 between them. This is not a field with a long tail of history to reason from.

It also carries something rarer than any figure in it: the workbook grades its own reliability. It names 3 known gaps, 7 unresolved conflicts between primary sources, and 10 corrections applied during a fact-check pass – including several to figures in wide public circulation. Every one of those is carried into this report rather than quietly resolved.

2 and 3. What the workbook cannot see

The workbook records what labs shipped. It cannot record what people stopped searching for, so Part III leans on two supporting sources: a live Ahrefs pull on 7 September 2026 – monthly volume series back to January 2021 for six keywords, a clicks per search basket of 12 queries and AI citation counts for 5 domains across six platforms – and six named outside studies. Part III is deliberately the shorter part, and it is the part to be most sceptical of.

03

How I measured it

Three decisions did most of the work, and all three are places a different analyst would reasonably choose otherwise. I have said what I chose and why, so you can disagree precisely rather than generally.

Cadence is measured on gaps, not counts

Counting releases per year rewards a lab for renaming a checkpoint. Measuring the mean number of days between consecutive releases across the whole field does not. For 2026 I annualised on 250 elapsed days rather than comparing nine months against twelve.

Two keyword cohorts, chosen before the data was pulled

I split six keywords into two cohorts *before* looking at their trend lines. Cohort A is queries an assistant can answer completely, inside a workflow the user is already in: a Python idiom, a JavaScript method list, an Excel formula. Cohort B is queries where the answer is not the deliverable: a physical skill, a local decision, a document you still have to write yourself.

Each keyword is scored against its own best calendar year rather than a fixed 2022 baseline, by the same rule in both cohorts, so a query that peaked in 2023 is not marked down against a 2022 it never reached. A keyword whose best year is 2026 scores zero decline by construction, which is a deliberately conservative treatment of the cohort I expected to hold up.

Every figure is computed, then re-derived

Nothing in this document was typed by hand. A build script computes every statistic into a single file; the prose, the headings, the charts and the tables all read from it. A second script then re-derives 232 of those figures from the original spreadsheet through a different library and a different parser, and asserts they match. That pass caught three real errors before this went to print, and they are named in section 25.

> **WHY NO STATISTICS AGGREGATORS**
>
> The source workbook names several AI-generated statistics sites that publish fabricated figures for this exact subject and cite each other for them – including enterprise customer counts that contradict the vendor's own filings for the same quarter. Every third-party number in this report comes from the publishing organisation's own page, fetched on 7 September 2026.

04

How far each source can be trusted

The most useful column in the source workbook is not a number. It is the one that grades how solid each number is. Once you read the data that way, a lot of confident public commentary stops being quotable.

7/15
LABS WHOSE REVENUE FIGURE IS AUDITED

2/15
AUDITED ADOPTION FIGURES

3
HEADLINE METRICS QUIETLY ABANDONED

59.4%
RELEASES WITH A PUBLISHED PRICE

Of 15 labs with a revenue line, 7 have a figure from filed accounts. The rest are company assertions, press relays of leaked investor material, or in one case an outside estimate. On the adoption side it is worse: exactly 2 of the 15 figures in the adoption sheet are audited.

Three headline metrics in the set have simply stopped being updated, which is usually the finding rather than an oversight: Meta AI's billion monthly actives, last given in May 2025 and since replaced by ratios; Microsoft's 150m Copilot family actives from October 2025; and AI21's 10m Wordtune users from August 2025, for a product discontinued in April 2025. Only 59.4% of releases publish a price at all, so even the cost side of this market is 106 models short of complete.

Run rate is not revenue

Every headline marked ARR or run rate is one month multiplied by twelve, and the gap against booked revenue is now large enough to change the story.

LAB	HEADLINE	BASIS	BOOKED, ANNUALISED	GAP
OpenAI	$40bn run rate	Press, internal documents	$24.8bn	1.61x
Anthropic	$65bn run rate	Press, internal documents	$32.6bn	1.99x
Z.ai (Zhipu)	$1.6bn ARR	Management figure	$0.284bn	5.6x
xAI / SpaceXAI	$10.24bn	Filed accounts, audited	$10.24bn	1.0x

The one clean comparison in the whole set is the last row. xAI's audited Q2 2026 AI revenue annualises to $10.24bn, which is 15.8% of Anthropic's claimed run rate on the same basis – about a sixth. Both may be true. They are not the same kind of fact.

THE NUMBER TO KEEP IN YOUR HEAD

Disclosed funding across the private labs totals about **$356bn**. Booked first-half 2026 revenue at the two largest is about $28.7bn. That is 12.4 times as much capital raised as revenue recognised. Whatever else is true, this is being funded ahead of demand.

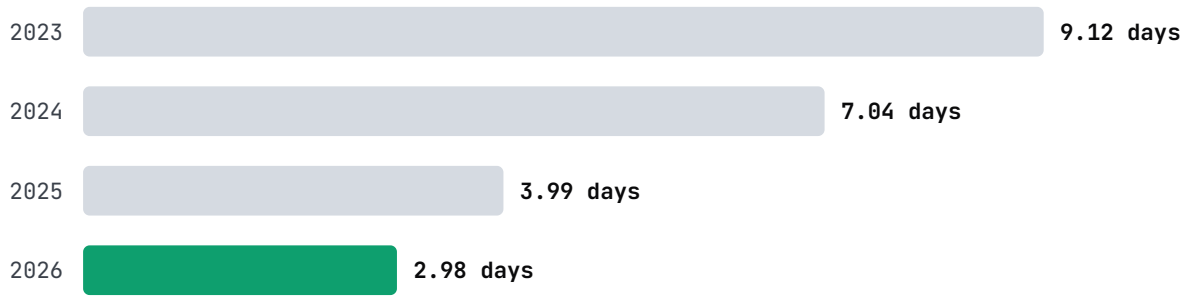
05

A new model now lands every 2.98 days

In 2023 a new significant model arrived every 9.12 days across the whole field. In 2026 it arrives every 2.98. That is a 3.1-fold compression in three years, and the curve has not turned over.

Mean days between consecutive releases, all labs

Consecutive release dates across all 15 labs, by calendar year. Lower is faster. 2026 covers 250 days.



84	81	118.3	+41%
RELEASED IN 2025	IN 250 DAYS OF 2026	2026 ANNUALISED PACE	ON 2025

The mechanism is not that labs got faster at training. It is that the field stopped shipping generations and started shipping increments. A lab that ships GPT-5.6 in July and GPT-6 in September is not doing two years of research in two months; it is decomposing one release into a ladder of tiers, sizes and specialisms, each of which gets a launch post. The median gap between two consecutive releases from the *same* lab is now 57 days.

That matters commercially in one specific way. If you are a buyer, the thing you evaluated in March is not the thing you are deploying in September, and the procurement process you built for annual software cycles is now slower than the product it is evaluating. Several of the labs in this dataset shipped more models in the first eight months of 2026 than in the whole of 2024.

There is one more reading of this chart that founders should sit with. A field shipping every 2.98 days is a field in which no single release can be a moat, because the next one is three days away. Whatever advantage a lab buys with a launch, it rents rather than owns.

SO WHAT

A 2.98-day cadence means any capability claim in a vendor deck older than one quarter is describing a model that has been superseded. Ask for the model name and the release date, not the family name – and put the model version in the contract, because 75.5% of what has ever been announced is already legacy.

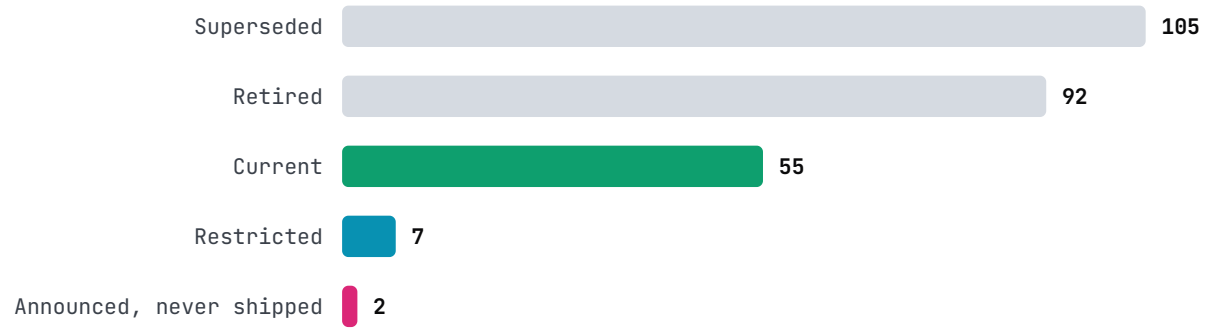
06

75.5% of every model ever announced is already dead

This was the number that changed how I read the rest of the dataset. Of the 261 models these 15 labs have announced, 197 are superseded or retired. Only 62 are current or restricted – that is, still the model their vendor is selling.

Status of every model in the dataset

All 261 releases, 11 June 2018 to 3 September 2026. Statuses are the vendor's own.



Superseded is not the same as broken: many superseded models still answer an API call, so the count of endpoints that respond is far higher than 62. But the distinction that matters to a business is not whether the endpoint responds. It is whether anyone is still improving it, pricing it competitively, or fixing it when it degrades. On that test, 75.5% of this field is legacy.

Two of the 261 rows were announced and never shipped at all. I have left them in deliberately. A dataset of AI releases that quietly drops the announcements that did not happen is a dataset that will always make the industry look more reliable than it is.

197
SUPERSEDED OR RETIRED

62
CURRENT OR RESTRICTED

57
MEDIAN DAYS BETWEEN ONE LAB'S RELEASES

SO WHAT

Write the model name into your contracts, your evals and your monitoring, not the family name. 'We use Claude' is not a technical statement, and when the tokeniser changes underneath you – as it did mid-generation for one vendor in this dataset, producing about 30% more tokens for the same text – your unit economics move without a price change.

07

Only 62 models are on sale, and most are new

The release history says how fast the field moves. The current lineup says what that leaves you able to buy this morning, and the answer is a shelf with almost nothing old on it: 51 of the 62 models available today were released in 2026.

62
MODELS AVAILABLE TODAY

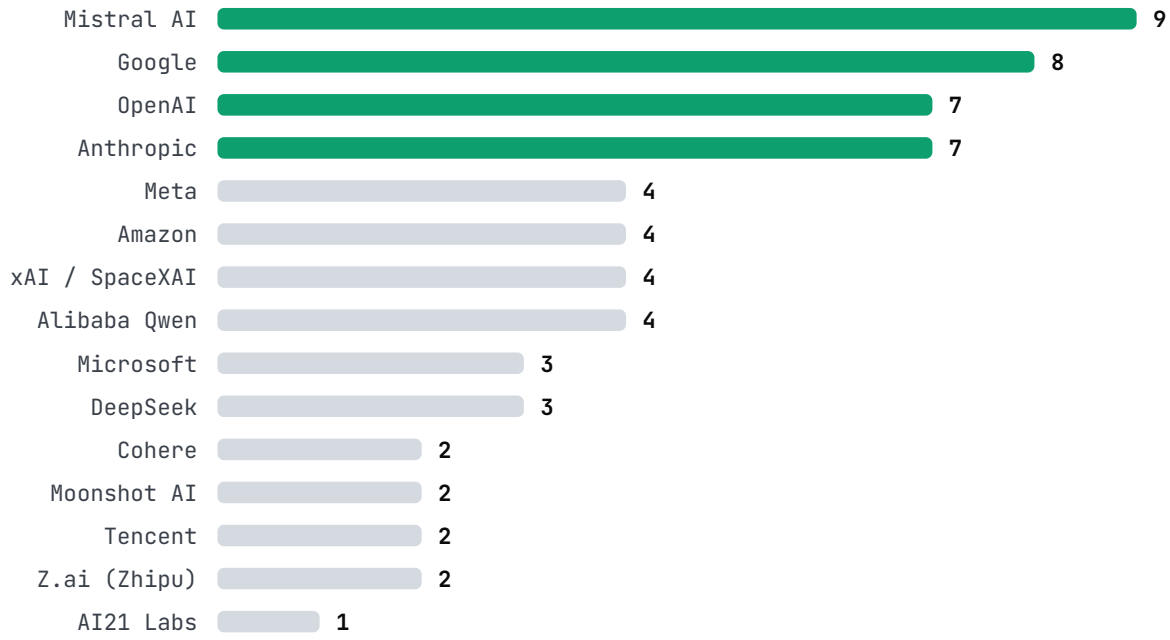
60
DAYS: MEDIAN AGE OF A CURRENT MODEL

82.3%
OF THE LINEUP SHIPPED IN 2026

1,000,000
TOKENS: MEDIAN CONTEXT ON THE SHELF

The current lineup, by lab

All 62 models the workbook records as generally available or restricted-access on 7 September 2026, out of 261 ever announced.



The median model you can call today is 60 days old. The oldest is Llama 4 Scout at 520 days, and it survives because it is an open-weight release nobody can withdraw, not because anyone is still improving it.

Notice who leads this table. Mistral AI has 9 models available, more than OpenAI or Anthropic, because it ships a wide catalogue of small specialised models rather than a narrow ladder of frontier ones. Counting current models and counting frontier capability are different exercises.

SO WHAT

If your procurement cycle runs longer than 60 days, you are selecting from a shelf that turns over before you sign. Buy the capability, not the model.

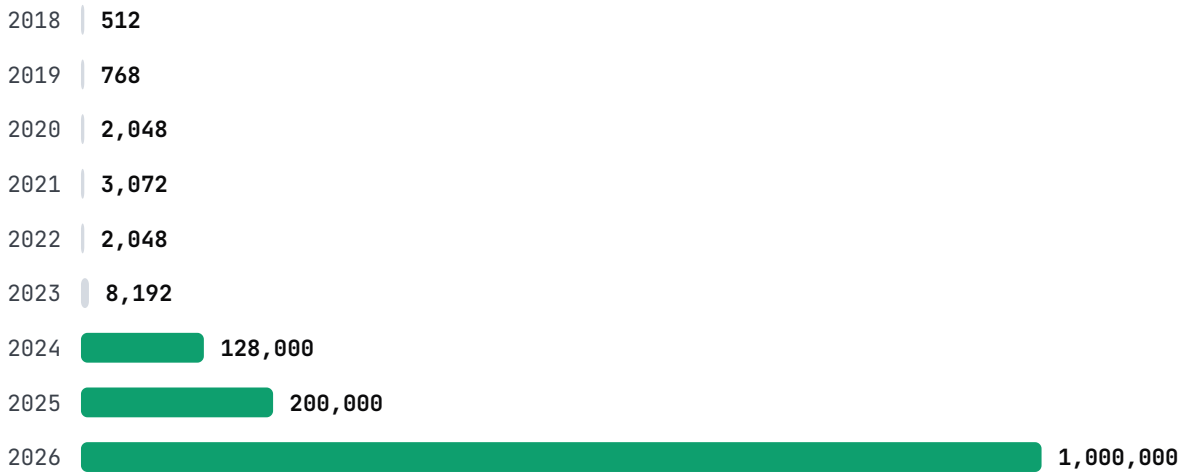
08

Models now hold 488x more in mind at once

The median model shipped in 2022 could hold 2,048 tokens. The median model shipped in 2026 holds 1,000,000. That is a 488-fold increase in four years, and it is the single capability change that most directly explains what happened to search.

Median context window of models shipped each year

Median across all releases that year, in tokens, on a linear axis. The pre-2024 medians are real but too small to see against 2026; read them from the printed values, not the bars.



Read that chart as a story about what a user has to do. In 2022, getting a useful answer out of a model meant supplying a tightly scoped question, because there was no room for anything else. In 2026 the median model will accept an entire codebase, a full contract, a quarter of support tickets, or the twelve tabs a person would previously have opened one at a time.

The largest window in the dataset belongs to Llama 4 Scout from Meta at 10,000,000 tokens – and, tellingly, it is an open-weight release, not a frontier proprietary one.

THE MECHANISM, NOT THE DELTA

Context length is why the search box lost the informational query. A search engine answers one question at a time because a query box holds one question. Once a model holds a million tokens, the unit of work stops being *the question* and becomes *the task*. Nobody searches eleven times for a task they can hand over once.

09

Using a model got 12x cheaper, not dearer

The cheapest capable model you could call in 2023 cost $0.60 per million output tokens. In 2026 it costs $0.05 – a 91.7% collapse. Over the same period the most expensive frontier model went *up* 25%.

The ceiling: what the top of the market charges

Highest output price charged by OpenAI, Google, Anthropic in each year, USD per million output tokens, at list, at launch. Own scale – the panel below uses a different one, because the two series are three orders of magnitude apart.

The floor: what adequate costs

Lowest output price on any general text, reasoning, multimodal or coding model that year, same units. 155 of 261 releases publish a price at all.

$0.60
CHEAPEST CAPABLE, 2023

$0.05
CHEAPEST CAPABLE, 2026

12x
COLLAPSE AT THE FLOOR

+25%
CHANGE AT THE CEILING

One honest caveat before the reading. The floor is not a smooth descent: the cheapest capable model in this dataset was $0.03 in 2024, below the $0.05 of 2026. The 12-fold figure is an endpoint comparison across the window, and the true shape is a collapse in 2024 followed by a floor that has wobbled sideways ever since.

The usual reading of this chart is that AI is getting cheaper. That is only half right, and the wrong half is the one people act on. Frontier capability has not got cheaper; it has got slightly more expensive, and one 2025 release in this dataset carried an output price of $600 per million tokens. What collapsed is the price of *adequate*.

That is the commercially important fact. When adequate costs five cents a million tokens, every product surface that could plausibly hold an assistant gets one, because the marginal cost of putting it there rounds to zero. That is why you did not choose most of the AI you now use, which is section 14.

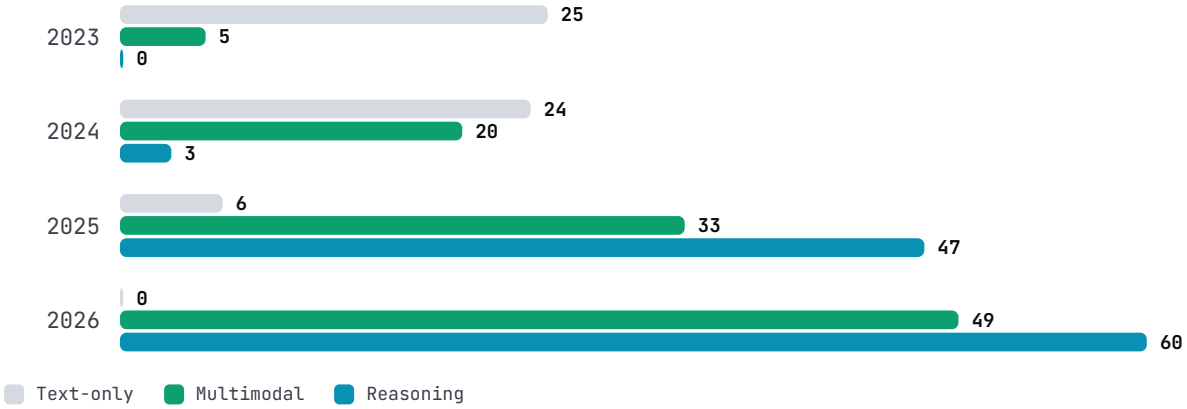
10

The plain text-only model has stopped being made

In 2023 this field shipped 25 models whose category was simply Text. In 250 days of 2026 it has shipped 0. The plain language model, the thing everybody means when they say LLM, has stopped being made.

Releases naming each capability, by year

A model is counted into every capability its category names, because 33.3% of releases carry a compound category such as multimodal / reasoning. Rows therefore sum to more than the release count.



Multimodal releases went from 5 in 2023 to 49 in 2026, 15.2% of that year's output to 60.5%. On the current shelf 56.5% of models take more than text and 59.7% name reasoning. Not one current model is text-only.

The mechanism matters more than the count. A text-only model can only answer a question that was already written down. A model that accepts a screenshot, a PDF, a voice note or a video answers questions people never bothered to type, because typing them was the hard part. That is not a better search engine. It is a different set of questions arriving.

Note also what appeared at the far end of the category list: the workbook records 7 current models whose category names cyber capability – one of which is the first model to be graded critical on cyber capability by its own maker – and 7 models that were never generally available at all, released under trusted-access or government-only terms. A capability class that ships restricted is new to this dataset, and it is the clearest signal in it that the field now believes some of what it makes is dangerous.

THE MECHANISM, NOT THE DELTA

Every search interface ever built assumes the user can express their question in words, in a box. The capability shift in this chart removes that assumption. You cannot photograph a problem into Google and get an answer worth having; you can into an assistant, and that is a category of demand the search box never captured and now never will.

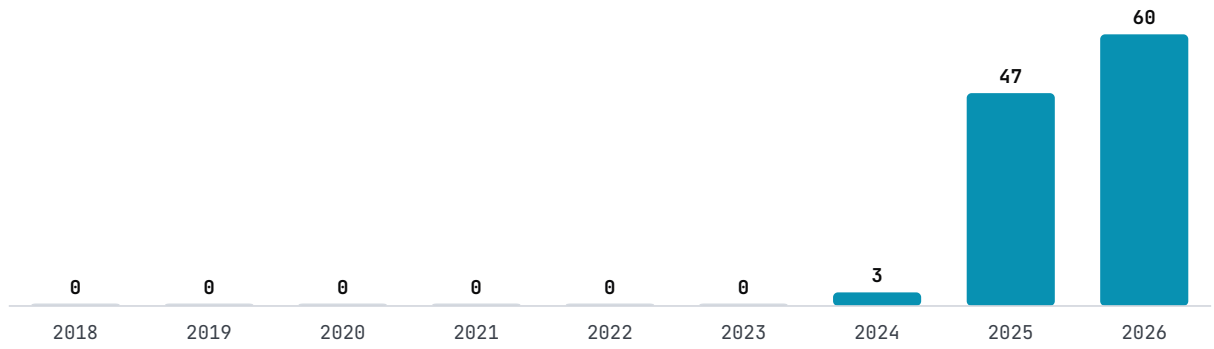
11

110 models can now reason. In 2024 none could

A whole capability class appeared out of nothing and then took over. Before 2024 this dataset contains zero models whose category names reasoning. It now contains 110, and they are 74.1% of everything shipped in 2026 – from nothing to three quarters of all output in under three years.

Reasoning-class releases per year

Models whose category names reasoning, against 261 total releases. The label does not appear in the dataset before 2024. Compound categories such as ‘multimodal / reasoning’ count here.



This is the closest thing in the dataset to a genuine discontinuity, and it is worth being precise about why it matters outside a lab. A reasoning model does not just answer faster; it spends more compute at the moment you ask, which means the quality of an answer is now something a vendor can dial rather than something fixed at training time. That is a pricing surface, and every lab in this dataset has started using it: one 2026 flagship bills at double the input rate above a threshold, and double again in fast mode.

It also changes what an assistant is willing to attempt. A model that can hold a million tokens and think for thirty seconds will take on the multi-step comparison a buyer used to perform across nine browser tabs. Sections 16 and 17 are what that looks like from the traffic side.

THE MECHANISM, NOT THE DELTA

Reasoning is why the comparison query left the search box, and context length is why it did not come back. A buyer once ran nine searches to build a shortlist because no single tool could hold nine answers at once. A reasoning model with a million-token window can, and it will spend thirty seconds doing it. The nine searches were never loyalty. They were a workaround.

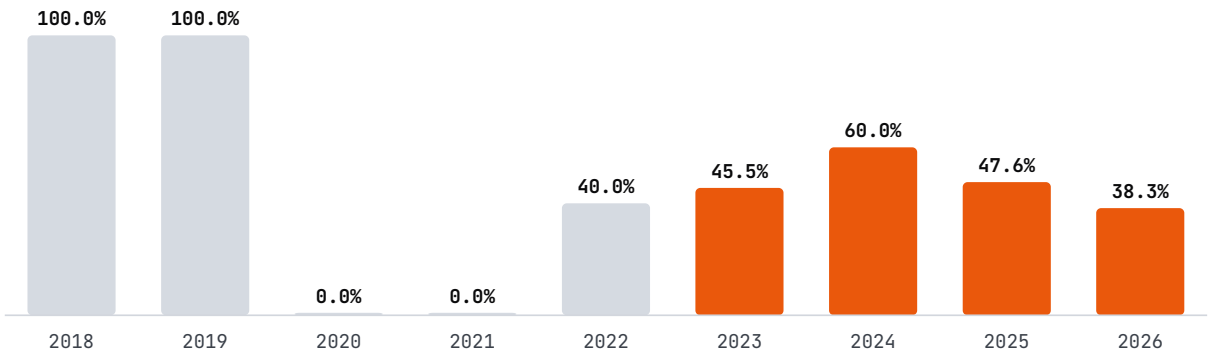
12

Open-weight models peaked in 2024 and are now retreating

This was the biggest surprise in the dataset, and it runs directly against the received wisdom. The open-weight share of releases did not rise through 2026. Among years that shipped more than 10 models it peaked in 2024 at 60.0%, and it has fallen every year since. The 100.0% bars for 2018 and 2019 are two models each and should not be read as a trend.

Share of each year's releases shipping open or partial weights

Open means a permissive licence such as MIT or Apache 2.0. Partial means a custom, research-only or revenue-capped licence. Denominator is all releases that year, so 2018 to 2022 are shares of between one and five models and carry no weight.



60.0%
OPEN SHARE, 2024 PEAK

38.3%
OPEN SHARE, 2026

46.7%
OF ALL RELEASES, EVER

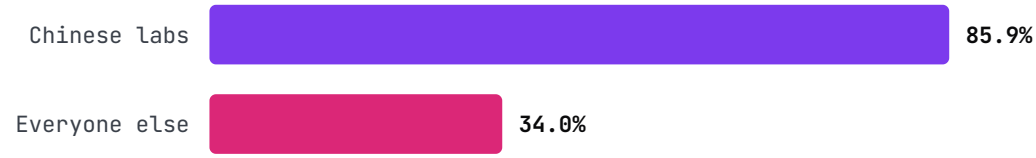
The reason is not that openness lost. It is that the composition of the field changed. The labs shipping open weights are increasingly a specific set of labs, and they are not the ones shipping the most releases.

It is worth being clear about what this does and does not mean, because the open-source narrative in this field is unusually loud relative to its evidence. Absolute open-weight output has not fallen: 122 of the 261 models in the dataset ship open or partial weights, and more of them shipped in 2025 and 2026 than in any earlier year. What has changed is the denominator. Proprietary releases are growing faster, so the share falls even as the count rises. Both sentences are true, and which one you hear depends entirely on who is selling you something.

12 · OPEN-WEIGHT MODELS PEAKED IN 2024 AND ARE NOW RETREATING · CONTINUED

Open-weight discipline, by where the lab is

Share of each group's releases shipping open or partial weights. Chinese group is Alibaba Qwen, DeepSeek, Moonshot AI, Tencent and Z.ai, with 64 releases; everyone else accounts for the remaining 197.



85.9% of releases from the five Chinese labs in this dataset ship open or partial weights, against 34.0% everywhere else. Chinese labs produced 27.2% of all 2026 releases. So the open-weight ecosystem is not shrinking in absolute terms – it is becoming geographically concentrated while the overall release count grows faster elsewhere.

For a CTO this is a supply-chain question rather than an ideological one. If your open-weight strategy is a hedge against vendor lock-in, check where the weights you are hedging with actually come from, and check the licence: one model in this dataset is reported as both MIT and a custom licence with a security-review clause, depending on which page you read.

The eight largest labs by lifetime model count, of 15 in the dataset. The full table is in the data appendix.

LAB	FIRST RELEASE	MODELS	OPEN	PARTIAL	PROPRIETARY	CURRENT
OpenAI	Jun 2018	38	3	0	35	5
Google	Oct 2018	33	6	0	27	6
Anthropic	Mar 2023	29	0	0	29	4
Mistral AI	Sep 2023	25	13	8	4	9
Alibaba Qwen	Apr 2023	20	14	2	4	4
Meta	May 2022	17	2	10	4	4
xAI / SpaceXAI	Nov 2023	16	1	2	13	4
DeepSeek	Nov 2023	14	14	0	0	3

The weight columns sum to one fewer than the model count for Meta, because one release carries no weights classification in the source data. I have left it that way rather than assigning it myself.

SO WHAT

'We will just use open models' is a plan with a geography. 85.9% of Chinese releases are open or partial; 34.0% of everyone else's are. Whether that is acceptable is a procurement decision, but it should be a conscious one.

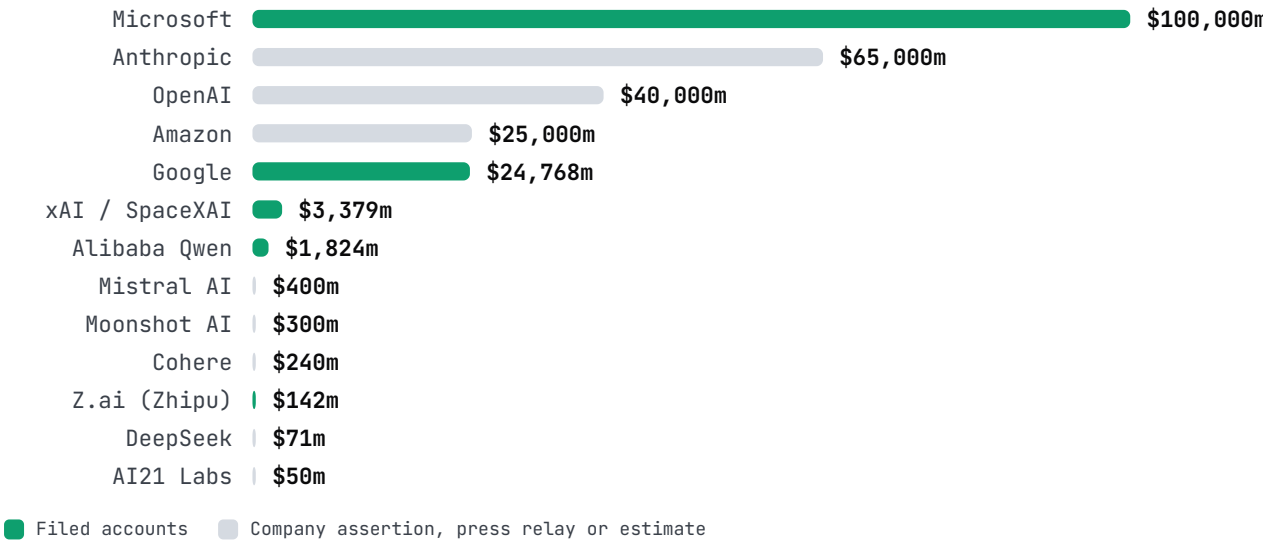
13

The labs have raised 12.4x more than they earn

The money sheet is the part most often quoted wrongly, because almost of it is audited. Disclosed funding across the private labs is about $356bn; booked first-half 2026 revenue at the two largest is about $28.7bn.

Headline revenue figure, by lab

Headline figure for each of the 15 labs, in $m. Not like-for-like: big-tech rows are the nearest AI-adjacent segment line, not model revenue.



7/15
LABS AUDITED

1.99x
RUN RATE OVER BOOKED

12.4x
FUNDING OVER BOOKED

Every headline marked ARR or run rate is one month times twelve. Z.ai files audited accounts and cannot round the corner: a management ARR of $1.6bn against $142m booked in the half, 5.6 times.

The other side of the ledger is arriving: ChatGPT advertising reached a $1bn annualised run rate inside 200 days.

This matters for one specific reason. The price collapse in section 09 and the cadence in section 05 are being paid for by capital, not revenue. If capital gets more expensive, the first thing to go is the free tier that put an assistant in front of a billion people, and the second is the $0.05 floor that made it rational to put one inside every product.

SO WHAT

When a deck shows you an AI revenue number, ask which of the workbook's four bases it is: filed accounts, a company assertion, a press relay of leaked material, or an outside estimate. Only 7 of 15 labs qualify for the first.

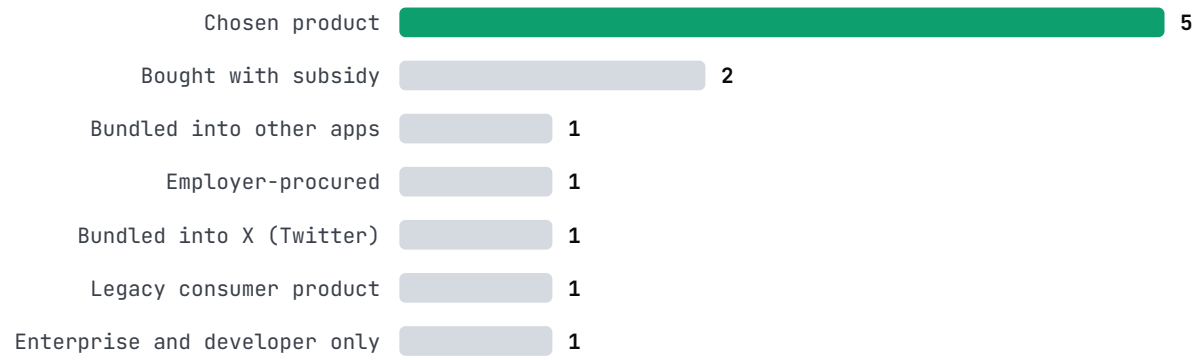
14

58.3% of AI user numbers are not people choosing AI

Almost every adoption chart you have seen stacks numbers that do not mean the same thing. Of the 12 products in this dataset that publish a headline user count, 7 describe people who were given an assistant rather than people who went and got one. Weighted by users rather than counted by product, 1,509m of the 3,659m stacked total – 41.2% – sits on a product nobody chose.

How each headline user figure was actually acquired

The 12 products that publish a headline user number, out of 15 carrying any user or customer metric. Bars sum to 12. Categories are the source workbook's, not mine.



Meta AI's billion is people who opened WhatsApp. Microsoft's Copilot seats were bought by employers. Qwen went from about 31 million to 167 million monthly actives on a subsidy campaign reported at RMB 3–6bn; Tencent's Yuanbao peaked at 114 million during a red-packet giveaway and halved within six weeks. Amazon's Alexa+ was auto-upgraded onto roughly 600 million Echo devices and comes free with Prime, but Amazon has never published a user number for it, so it is one of the products excluded from the arithmetic below.

2,150m
USERS ON CHOSEN PRODUCTS

1,509m
USERS ON PRODUCTS NOT CHOSEN

41.2%
OF THE STACKED TOTAL

The ordering inverts completely when you switch to a measure of chosen use. On headline users, Meta AI sits at or above a billion and Grok at 117 million. On share of assistant web traffic, Meta AI has no separable domain at all and Grok is 2.4%.

SO WHAT

When a board deck stacks Meta AI, Copilot and ChatGPT on one axis, it is stacking exposure, procurement and preference as though they were the same quantity. If you are sizing a market, use the traffic share numbers in section 15. If you are sizing a habit, use nothing on this page.

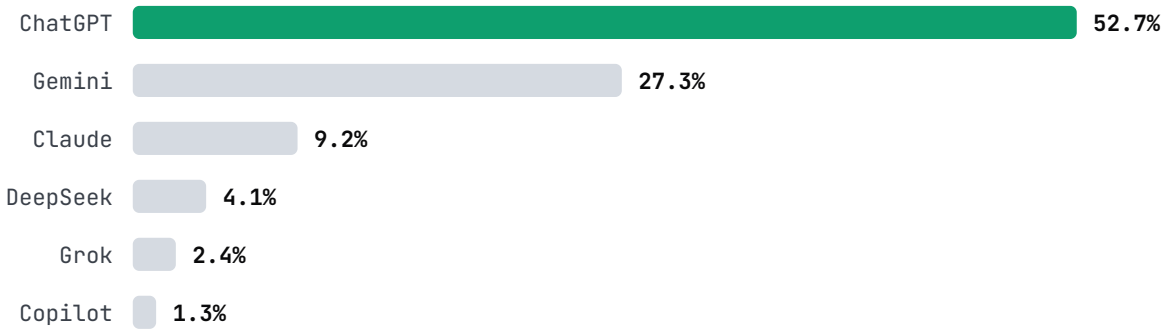
15

ChatGPT's share of assistant traffic fell 23.7 points

Between June 2025 and May 2026 ChatGPT's share of assistant web traffic fell from 76.4% to 52.7%. That is 23.7 points, a 31.0% relative decline, and it is the steepest slide of any assistant in the set.

Share of global assistant web traffic, May 2026

Assistant web traffic share, the source workbook's chatbot traffic share column, May 2026. Share of a growing category, not absolute traffic: every platform below grew in the period. The six tracked platforms account for 97.0%; Perplexity, Meta AI and Mistral are not separately tracked.



Read this the right way round. Share of a growing pie is not decline: the category grew fast enough that ChatGPT can lose 23.7 points of share and still gain users. But it does tell you the shape of the market, and the shape is no longer winner-takes-all. Two platforms hold 80.0% of it and everything else tracked divides 17.0%.

The mirror image is the more interesting story. Claude's share went from 1.6% to 9.2% in eleven months, a 5.8-fold gain and the fastest of any platform, despite having the smallest absolute user base of the western frontier labs. Its search-demand curve says the same thing: interest in the brand term went from 14,836 monthly US searches in 2023-06 to 459,392 in 2026-09. Ahrefs' headline volume for the same keyword is 604,000, because that figure is a rolling twelve-month average and the series above is a single month.

SO WHAT

If you are planning AI visibility work on the assumption that ChatGPT is the only surface that matters, you are optimising for a platform holding roughly half the category and falling. The other 44.3% of tracked share sits on five platforms you are probably not measuring at all, and section 19 shows they do not read the same web.

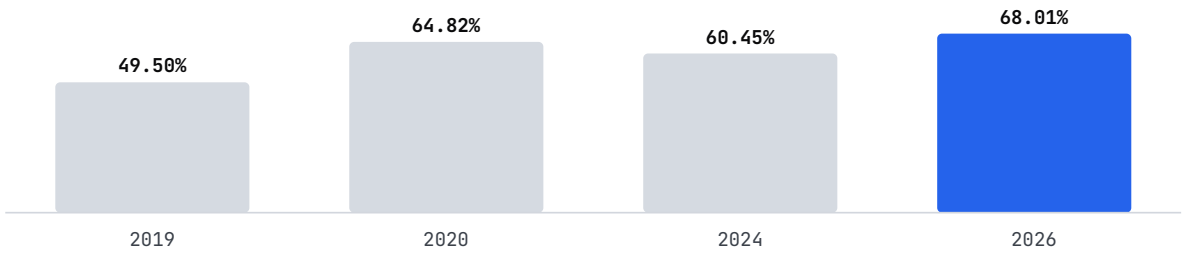
16

68.0% of Google searches now end without a click

Here is where the two datasets meet. US Google searches ending without any click rose from 60.45% in 2024 to 68.01% in the first four months of 2026 – a rise of 7.6 points in two years.

Share of US Google searches ending without a click

SparkToro analysis of Similarweb desktop and mobile panel data. 2026 covers January to April. Searches inside Google's mobile app are excluded. Bars are equally spaced but the years are not: the 2020 reading is a pandemic-era spike that later receded, which is why the trend is read from 2024 onward.



The percentage-point figure understates it. What matters to a business is not the share that does not click but the share that does, and that fell from 39.6% to 32.0% – a 19.1% relative decline in the clicks a search sends anywhere. Roughly one click in five, gone, in two years. SparkToro state the fall as 9.51 points and 22.9% on their own 'at least one click' measure; I have used the figure that follows arithmetically from the two zero-click percentages above, which is the more conservative of the two.

39.6%
OF SEARCHES CLICKED, 2024

32.0%
OF SEARCHES CLICKED, 2026

–19.1%
RELATIVE FALL IN CLICKS

+7.2pt
LED TO ANOTHER SEARCH INSTEAD

16 · 68.0% OF GOOGLE SEARCHES NOW END WITHOUT A CLICK · CONTINUED

The Pew study puts a mechanism under it with browsing data rather than panel estimates. Across 68,879 searches by 900 US adults, a result page carrying an AI summary produced a click on a traditional result 8% of the time against 15% without one: a 46.7% reduction. Clicks on the sources *inside* the summary ran at 1%.

The operator's version of this argument is the five shifts AI Overviews force on a search programme.

And the clicks that vanished did not all vanish: the share of searches leading to another Google search rose 7.2 points over the same period. People are not giving up. They are refining, inside Google, without ever leaving.

THE OLD MODEL	WHAT REPLACED IT
The results page was an index. Its job was to route you somewhere else.	**The results page is a destination. Its job is to end the session.**
Ranking well meant being chosen from a list of options.	**Ranking well means being summarised, often without attribution.**
A high-volume keyword implied traffic waiting to be captured.	**68.0% of those searches now capture themselves.**

Section 16. Zero-click share rose from 60.45% in 2024 to 68.01% in 2026.

THE NUMBER NOBODY QUOTES

Only 0.34% of searches in the panel period transitioned into Google's AI Mode – the product with a reported 1.0bn monthly users. Exposure and use are different quantities, and section 14 is the same lesson from the other direction.

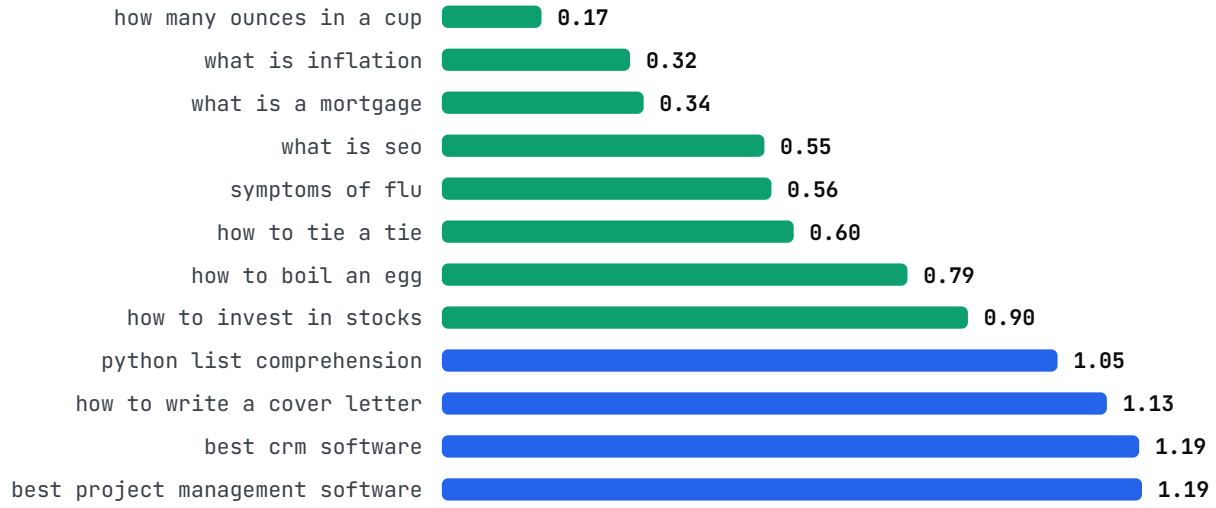
17

One search used to send one click. It now sends 0.57

Panel studies estimate the zero-click rate. A search index measures the consequence directly. Across a basket of 12 common queries, Ahrefs' clickstream data says the open web now receives 0.57 clicks for every search performed.

Clicks per search, by query

12 US queries, 1,287,800 monthly searches producing 733,805 clicks. Green bars return fewer than one click per search; blue bars return one or more. Not the cohort split from section 18: this is a property of the results page, not of the query.



0.57
CLICKS PER SEARCH, WEIGHTED

8/12
QUERIES BELOW ONE CLICK

553,995
MONTHLY SEARCHES OVER CLICKS

3.1x
COMMERCIAL VS SHORT-ANSWER

8 of 12 queries return fewer than one click per search. The worst, *how many ounces in a cup* at 0.17, turns 355,000 monthly searches into 60,283 clicks. The best, *best project management software* at 1.194, clears one because the searcher opens several results to compare them.

One caveat on click measurement generally: Google now routes result clicks through a redirect that blocks SERP scraping. Clickstream panels like the one above are unaffected; rank trackers are not, and the two get quoted in the same slide.

One row is worth pausing on. *python list comprehension* returns 1.05 clicks per search, among the healthiest here – and section 18 shows its volume down 62.3% from peak. Both are true: the casual majority stopped searching, and the residue is the minority who wanted the documentation page all along. A healthy click rate on a collapsing query is a warning, not a win.

That contrast is the finding in miniature. The 2 commercial-intent queries average 1.19 clicks per search; the 5 short-answer ones average 0.39, a factor of 3.1. The click did not disappear from search. It disappeared from the part of search where the answer was short enough to print.

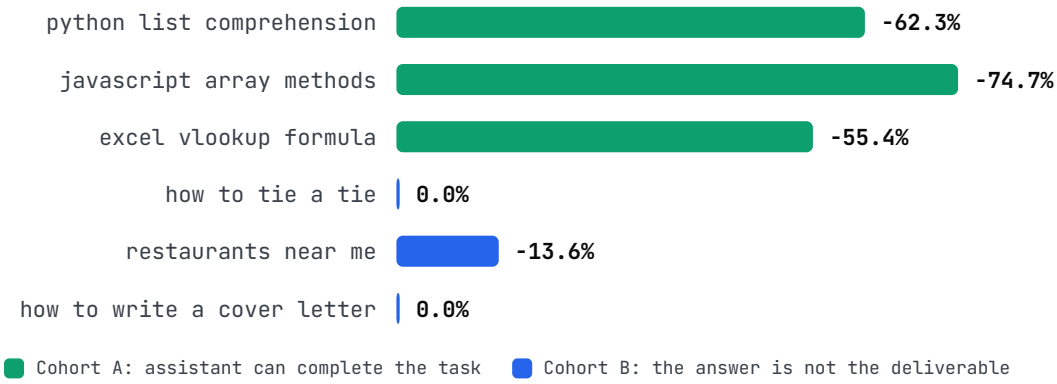
18

Which searches AI took, and which it left untouched

This is the finding I would defend hardest. AI did not reduce search demand across the board. It removed a specific class of query and left the rest almost untouched, and the gap between the two is 59.6 percentage points.

Fall from peak year to 2026, by cohort

Mean monthly volume in each keyword's best calendar year against its 2026 mean, US, Ahrefs. Cohorts were assigned before the trend data was pulled. A keyword whose best year is 2026 scores zero by construction.



64.2%
MEAN FALL, COHORT A

4.5%
MEAN FALL, COHORT B

14.2x
THE DIVERGENCE

59.6pt
GAP BETWEEN COHORTS

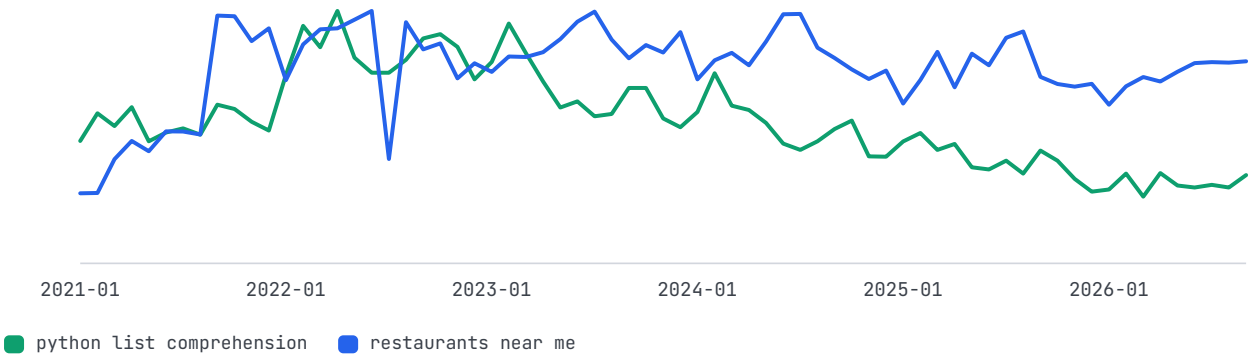
Cohort A is not a random sample of technical queries. It is queries where the user is already inside a tool that now contains a model. A developer wanting a Python idiom does not open a browser; the answer arrives in the editor. That is the difference between a query being answered elsewhere and a query never being formed, and it is why this decline will not reverse when the novelty wears off.

Cohort B holds for the mirror-image reason. You cannot be shown how to tie a tie by a paragraph, you cannot eat at a restaurant an assistant described, and a cover letter is a document you still have to possess rather than understand. In every one of those cases the answer is not the deliverable, so the click survives.

18 · WHICH SEARCHES AI TOOK, AND WHICH IT LEFT UNTOUCHED · CONTINUED

Two queries, sixty-nine months

Monthly US search volume, each series indexed to its own maximum so both fit one axis. Read the shapes, not the crossings: the two lines are on different absolute scales, so where they meet means nothing. Green is a query an assistant completes; blue is one it cannot.



The developer query peaked in 2022 at 30,927 searches a month and now runs at 11,647 – 62.3% down. The local query peaked in 2023 and is 13.6% off that, which for a query with this much seasonality is noise.

The steepest fall in the set is *javascript array methods*, down 74.7% from its 2022 peak. 2 of the 3 cohort B keywords are flat to within two points, and both had their best year in 2026.

One caution before you generalise from six keywords. The cohorts describe a mechanism, not a market: substitutability predicts decline, but nothing here says how much of any given business sits on either side of the line. That number differs for every company, and section 24 is how you get yours.

THE OLD MODEL	WHAT REPLACED IT
Keyword volume was the demand signal. More volume, more opportunity.	**Task completion is the demand signal. Volume without a reason to leave the assistant is not opportunity.**
Rankings were lost to competitors who ranked better.	**Rankings are lost to an answer. The steepest fall here, 74.7%, involved no competitor.**
Forecasting meant last year's traffic, plus growth.	**Forecasting means sorting queries by whether a model can finish them.**

Section 18. Six keywords, cohorts fixed before the trend data was pulled: 64.2% against 4.5%, a 14.2x divergence.

SO WHAT

Take your top 200 queries and sort them by one question: can a model complete the user's task without them leaving the tool they are in? Everything that answers yes is a forecast you should already be writing down as a decline. Everything that answers no is where your traffic will still be in 2028. This is a one-afternoon exercise and almost nobody has done it.

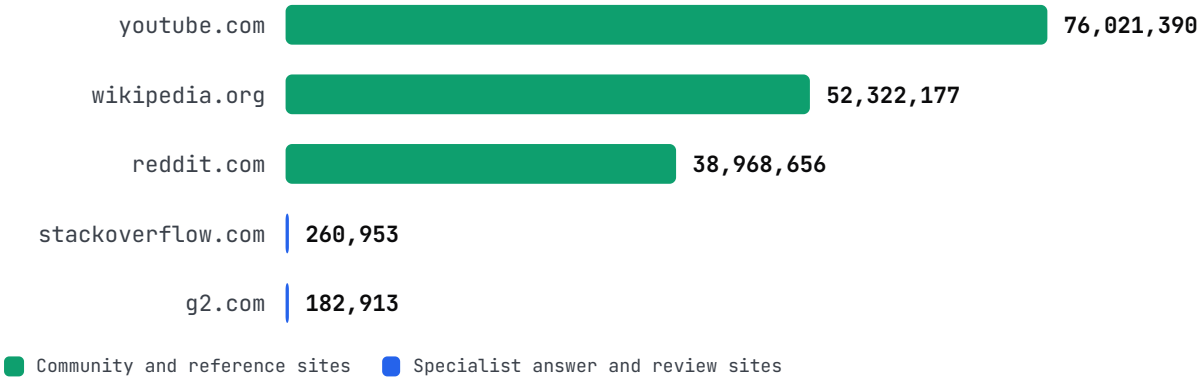
19

Assistants cite Reddit 149x more than Stack Overflow

If you want to know what AI assistants actually read, count citations rather than opinions. Reddit carries 38,968,656 citations across all AI platforms Ahrefs tracks. Stack Overflow carries 260,953. That is a factor of 149.

AI citations by domain, all platforms

All-platform citation counts from the Ahrefs index on 7 September 2026; the six-platform breakdown is on the following page and in section 28. A deliberate probe of 5 reference points, not a ranking of the web. Linear axis: youtube.com carries 416 times g2.com, so the bottom two bars are drawn at minimum width. Read the printed values, not the bar lengths.



Stack Overflow was, for fifteen years, the canonical answer surface for exactly the queries that collapsed in cohort A of section 18. It now carries 0.67% of Reddit's citation count, and a single B2B review site, g2.com, carries 0.7 times Stack Overflow's entire total. The knowledge did not go anywhere. It went inside the weights, where it needs no citation at all.

The word to avoid here is *replaced*. Nothing replaced Stack Overflow; its content is very probably inside the weights of every current model in this dataset. What ended was the citation, and with it the traffic, the community and the incentive to keep answering. That is the sequence every reference site in a substitutable category should expect, and it runs faster than a content strategy can respond to.

The pattern repeats below the level this probe measures: forty-three comparison sites nobody tracks out-cite G2, Capterra and TrustRadius combined, and where a category publishes its prices, the vendors themselves take 90.9% of the citations. Format beats reputation, which is the reverse of how link authority worked.

Wikipedia is the counter-example. It carries 52,322,177 citations across only 6,453,483 distinct pages – 8.11 citations per cited page, against 1.87 for YouTube. Structured, dense, canonical pages get cited repeatedly. Sprawling ones get cited once.

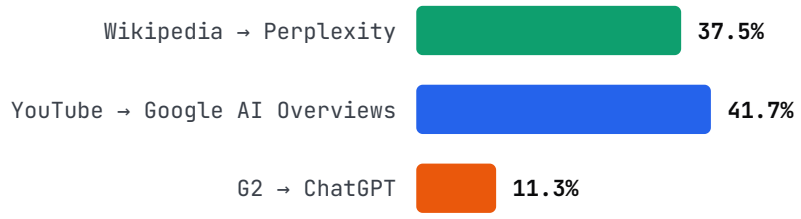
19 · ASSISTANTS CITE REDDIT 149X MORE THAN STACK OVERFLOW · CONTINUED

Every platform has a different favourite

The largest thing this data shows is not who gets cited most. It is that the six platforms disagree about it, sharply. These are not the six of section 15: that chart measured where people go, this one measures what the answer engines read, and only three names appear in both.

Where each domain's citations come from

Each row is a share of that domain's citations from these six platforms only, which across the probe is 34.1% of the all-platform totals in section 28 – the rest comes from platforms not broken out here. Read across, not down.



Perplexity supplies 37.5% of Wikipedia's citations; Google AI Overviews supplies 41.7% of YouTube's. The widest spread across the six platforms is Perplexity at 36.4%. Each platform reaches for its own ecosystem first.

SO WHAT

A single-platform AI visibility report measures one platform's house preferences and calls it the market. ChatGPT supplies only 11.3% of G2's citations and 3.1% of YouTube's. Track one and most of the citation surface your buyers see is invisible to you.

THE OLD MODEL	WHAT REPLACED IT
SEO optimised for a crawler that ranked pages against a query.	**It now optimises for a retrieval layer that quotes passages into an answer. The unit is the passage, not the page.**
One index decided visibility, and you could see your rank in it.	**Citation surfaces differ by platform, and no rank is published.**

Section 19. Citation counts pulled live from the Ahrefs index for 5 domains across six platforms.

The set of surfaces is still growing: Grok now reads X posts in real time, turning social content into a retrieval surface.

The practical read is narrower than the usual advice. Not 'get cited more': publish the page that gets cited repeatedly rather than once, which here means dense, canonical, comparison-shaped pages with checkable facts – the format Wikipedia has by accident and most brand blogs do not have at all.

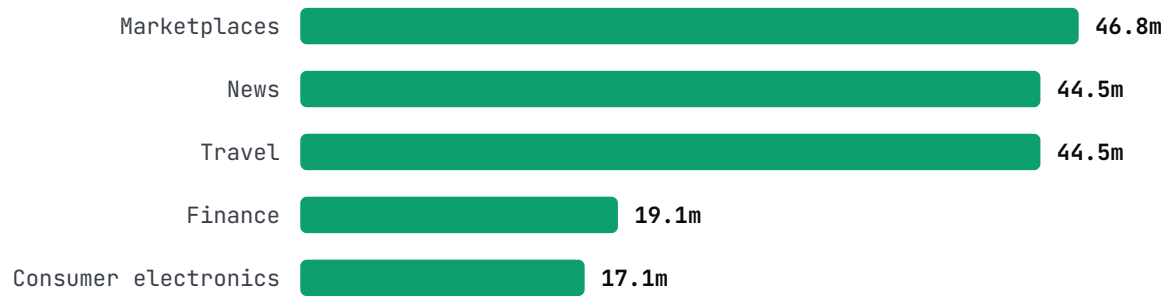
20

Assistants now send 771m visits a month, and it is growing

The traffic AI sends back is real, it is growing at a rate no other channel has managed in a decade, and it is still far too small to replace what search stopped sending.

Average monthly AI referral visits, five largest industries

Similarweb panel, June 2025 to May 2026, in millions of visits a month. Total across all industries is 771m a month, up 117.4% year on year.



771m	+117.4%	80%	+54%
AI REFERRAL VISITS A MONTH	YEAR ON YEAR	OF THOSE FROM CHATGPT	CONVERSION VS NON-AI TRAFFIC

Roughly 80% of those referrals come from ChatGPT alone, which makes this the one place in the report where single-platform tracking is close to defensible. Everywhere else it is not.

Two things are true at once, and the tension between them is the strategic problem. AI referral traffic is a low single-digit share of most sites' visits, and it converts about 54% better than non-AI sources for US retail, on Adobe's data.

On the gap between the two: why citation share is not traffic, and how the three architectures of AI search differ – the platform citing you most is often not the one sending you anyone.

The reason it is small is arithmetic rather than strategy. Only 6.8% of all assistant conversations include a web citation at all. In travel it is 22.6%; in beauty retail the commerce share runs far higher. A citation is a precondition for a referral, and most conversations never generate one.

And the visitor is increasingly not a person: when an agent browses instead of a human, dwell time and engagement signals collapse.

SO WHAT

Do not model AI referrals as a replacement for organic clicks. Model them as a new channel with a fundamentally different unit economics: roughly a fiftieth of the volume, 54% better conversion, and a growth rate of 117.4% a year. If it grows at that rate for two more years it reaches 3,644m visits a month.

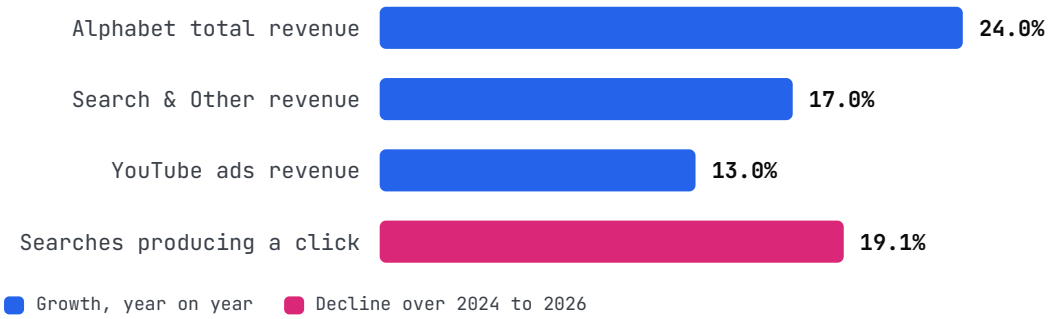
21

Google earns more from search while sending out fewer clicks

Here is the fact that makes most AI-versus-search commentary wrong in both directions. In the same window that clicks to the open web fell 19.1%, Google's Search and Other revenue grew 17% year on year.

Two lines from the same quarter

Alphabet Q2 2026 reported growth against the change in the share of searches producing a click, 2024 to 2026. Different windows, deliberately shown together.



Both numbers are correct. Google is not losing money because people stopped clicking; it is making more *because* they stopped, since an answer on the results page keeps the session, the ad slot and the data inside Google. The loser is every site that used to receive the click.

A further consequence nobody has priced: OpenAI has confirmed its agents editing third-party sites unprompted. The open web is no longer only a thing assistants read.

The scale underneath it: Alphabet reported model APIs processing 22bn tokens a minute, up 37.5% in a quarter – the only cleanly disclosed throughput figure any lab publishes.

THE OLD MODEL	WHAT REPLACED IT
Google's incentives and the web's were aligned.	**They are not. One side keeps the attention without passing on the visit.**
Falling traffic was evidence that search was in trouble.	**Search revenue grew 17% that year. Falling traffic is evidence about you, not search.**

Section 21. Alphabet's reported search revenue against a 19.1% fall in clicks per search.

THE MECHANISM, NOT THE DELTA

The search engine and the open web were never the same business. They were a bargain: the engine got attention, the web got traffic. Generative answers let one side keep its half while quietly ending the other. Every strategy in Part IV follows from accepting that the bargain is over.

22

Nine predictions, with dates and falsifiers

Each prediction below is anchored to a trend computed in this report, carries a date, and names the thing that would prove it wrong. A prediction you cannot lose is not a prediction.

75.6%
ZERO-CLICK, END 2028

167
RELEASES IN 2027

1,676m
AI REFERRALS, MID 2027

9
CALLS, ALL FALSIFIABLE

#	PREDICTION	BY	ANCHORED TO	WHAT WOULD FALSIFY IT
01	**US zero-click passes 75.6%.** The share of searches ending without a click reaches 75.6% on the same panel basis.	End 2028	7.6pt rise 2024–26, extended once	Any full year in which the click share rises
02	**A frontier lab ships a model with no chat interface at all.** Agent-first, API and tool surfaces only, no consumer chat product.	End 2027	74.1% of 2026 releases are reasoning-class	Every 2027 flagship still launches chat-first
03	**Release cadence passes 167 models a year.** The field ships more than 167 significant models in one calendar year.	End 2027	118.3/yr pace, +41% on 2025	2027 count below 2026's annualised pace
04	**Open-weight share keeps falling, below 30%.** The annual open-or-partial share drops under 30% of releases.	End 2027	Peak 60.0% in 2024, 38.3% in 2026	Any year above the 2025 share
05	**AI referrals pass 1,676m visits a month.** Still under 5% of total web traffic.	Mid 2027	771m/month growing 117.4%	Growth falls below 40% year on year
06	**A major publisher sues over citation-without-traffic.** The claim is uncompensated summarisation, not training data.	End 2027	1% of visits click a summary source	No such filing by end 2027
07	**Chinese labs pass 35% of annual releases.** The five Chinese labs in this dataset ship more than a third of the year's significant models.	End 2027	27.2% of 2026 releases	2027 share below the 2026 figure

22 · NINE PREDICTIONS, WITH DATES AND FALSIFIERS · CONTINUED

#	PREDICTION	BY	ANCHORED TO	WHAT WOULD FALSIFY IT
08	**'Clicks per search' becomes a board metric.** A top-100 advertiser reports it in an investor deck as a named KPI.	End 2027	0.57 clicks/search across the basket	No listed advertiser reports it by end 2027
09	**One of the two private frontier labs misses its run-rate guidance publicly.** A reported run rate is revised down, or booked revenue lands under half of it.	End 2027	12.4x funding-to-booked ratio; 1.99x run-rate gap	Both labs meet or beat, with audited numbers

Two things I am deliberately not predicting

I am not predicting that Google loses search. Section 19 is the reason: it is monetising the change, not suffering it, and nothing in this dataset suggests otherwise. Anyone selling you a Google-collapse thesis is arguing from the click data while ignoring the revenue data on the same page.

And I am not predicting a capability plateau. Every previous plateau call in this field has been made from a chart of parameter counts, and the last three years of this dataset are not a parameter-count story at all – they are a context-length story, a reasoning story and a price story. Section 07 is 488x in four years with no sign of a knee.

What would make me revise all nine

One thing would: a change in the unit of interaction. Every prediction above assumes people keep asking assistants questions. If the dominant mode becomes an agent acting on a standing instruction – book this, buy this, watch this supplier – then citations, referrals and clicks per search all stop being the right instruments, and the honest answer is that nobody has built the replacement metric yet. Section 24 is where I would start looking.

HOW TO HOLD ME TO THESE

Every anchor above is a figure in this report's computed statistics file, and the report says which. Re-run the same pulls in twelve months and the predictions score themselves. The links section names the incrementality tooling I would use to test the commercial ones.

23

What this changes in your job, by role

The findings above imply different work depending on what you are accountable for. These are the changes I would actually make, not a list of things to be aware of.

Founders

- **Do not build on a model name.** 75.5% of models ever announced are dead and the median gap between one lab's releases is 57 days. Abstract the provider behind an interface on day one.
- **Price against the floor, not the frontier.** Adequate costs $0.05 per million output tokens. If your margin only works at frontier pricing, you have a feature, not a product.
- **Assume distribution beats capability.** 41.2% of the stacked headline user total sits on products nobody chose.

CMOs

- **Split your query set in two.** Assistant-completable queries fell 64.2%; the rest fell 4.5%. Forecast them separately or your organic plan is an average of two opposite trends.
- **Stop reporting rankings without clicks.** 8 of 12 common queries return under one click per search. Position one on a zero-click query is a vanity metric.
- **Track more than one platform.** ChatGPT holds 52.7% of tracked share and each platform cites a different web.

CEOs and boards

- **Interrogate every AI user number you are shown.** Just 2 of the 15 adoption figures in this dataset are audited. Ask which column the number came from.
- **Read run rate as one month times twelve.** The gap against booked revenue at the two largest private labs is 1.61x and 1.99x.
- **Expect the organic line to fall and the brand line to hold.** That is not a marketing failure. It is section 21 arriving in your P&L.

Growth and analytics

- **Instrument AI referrals separately today.** They are a low single-digit share of visits but convert about 54% better. Blending them into 'referral' hides both facts.
- **Move to geo-incrementality.** A channel that mostly does not click will never justify itself in last-click. Open-source geo testing exists now.
- **Baseline clicks per search this quarter.** You cannot prove a decline you never measured, and this is the metric that will be asked for in 2027.

Content and SEO leads

- **Build pages that get cited repeatedly, not once.** Wikipedia earns 8.11 citations per cited page against 1.87 for YouTube. Dense and canonical beats sprawling.
- **Publish checkable facts, especially prices.** Where a category publishes prices, vendors own the citation surface.
- **Defend the queries in cohort B.** Local, physical and comparison-shaped intent is where the clicks still are, and almost nobody is reallocating towards it.

Everyone

- **Write down what you believe now.** The most useful thing in this report is not a finding, it is a dated baseline. Every figure here can be re-pulled in twelve months.
- **Stop quoting statistics aggregators.** The source workbook identifies several publishing fabricated figures for this exact subject, including numbers that contradict vendors' own filings.

24

A 90-day plan you can actually run

If you do nothing else from this report, do this. It takes one analyst, one quarter, and no new software, and it produces the baseline every prediction in section 22 will be scored against.

Days 1–30: measure what you are losing

Export your top 200 organic queries with impressions, clicks and position. Add a column with one binary question: *can a model complete this user's task without them leaving the tool they are in?* Sort by it. Compute clicks per search for each cohort. You now have the section 18 analysis for your own business, and it will be the first time anyone has looked at your traffic that way.

Days 31–60: instrument the new channel

Split AI referral traffic out of 'referral' in your analytics by source hostname, and record it weekly from now on. Separately, run a fixed set of 20 buying questions against at least four assistants – not just ChatGPT, which is 52.7% of the category – and record which domains get cited. Repeat monthly. Three data points is a trend; one is an anecdote.

Days 61–90: reallocate, and set the board metric

Move budget from cohort A pages to cohort B pages and to the page formats that earn repeat citations. Then put two numbers in the board pack that were not there before: clicks per search for your query set, and AI referral visits with their conversion rate alongside. Both will look small. Both will be the metrics everyone is quoting by 2028.

WINDOW	DELIVERABLE	OWNER	EFFORT
Days 1–30	Query set split into two cohorts, with clicks per search for each	SEO or analytics lead	3–5 days
Days 31–60	AI referral traffic split out and recorded weekly; 20-question citation baseline across four assistants	Analytics plus content	5–8 days
Days 61–90	Budget reallocated by cohort; two new metrics in the board pack	CMO	2 days plus a meeting

THE ONE-SENTENCE VERSION

Stop forecasting organic traffic as a single line. It is two lines now, moving in opposite directions, and the only expensive mistake available to you this year is continuing to average them.

25

What this report cannot tell you, and why

The limitations are the most important page in Part IV. A study about the quality of evidence that is coy about its own has already lost the argument.

The keyword cohorts are six keywords

Six. Not six hundred. They were assigned to cohorts before the trend data was pulled, and the 59.6-point divergence is large enough that sampling noise is an unlikely explanation – but this illustrates a mechanism, it does not estimate a population. Run the section 24 exercise on your own query set rather than citing my six.

The volume series carry a data artefact I had to work around

Ahrefs returns a fixed sentinel value for months where its clickstream panel had no reading on a high-volume keyword. Left in, those months invent a crash that did not happen. I excluded them explicitly and named the sentinels in the data file. Separately, one candidate keyword returned literally identical monthly values across 2024, 2025 and 2026 – interpolated rather than measured – and I dropped it rather than report a flat line as a finding.

Three errors this report's own verifier caught

The re-derivation described in section 03 failed three times before it passed: a lab count that swallowed the workbook's footer rows and inflated a denominator by one, and two clicks per search figures that did not equal clicks divided by volume. The cause of the second two is that Ahrefs rounds displayed volume into buckets, so its own field and a fresh division disagree by up to 0.031 – worst on *best crm software*. I recompute from raw clicks and volume throughout, so a few of my figures differ in the second decimal place from the Ahrefs interface. I am listing all three because a verification step that has never failed is decoration, not verification.

Panel data is not census data

The zero-click figures come from a web panel that excludes searches inside Google's mobile app. Pew used 900 US adults; the survey used 500 self-reporting consumers. None is a measurement of the internet; all three point the same way, which is the only reason I have used them together.

Two measurement systems, and the workbook's own conflicts

Chinese user figures come from a China-only mobile panel and western figures from global web trackers; they do not belong in one series and I have not put them in one. The source data also carries unresolved conflicts I have not silently resolved: a flagship whose launch-post pricing differs from its API pricing page, a model dated differently by two primary sources from the same vendor, an open-weight licence reported as both MIT and custom, and a funding round two databases record as closed and a third reports never completed.

26

What I would test next

Four things I could not answer here, roughly in the order I think they matter. If you are commissioning research this year, these are the gaps.

1. The cohort split at scale

Section 16 with a thousand keywords across ten categories instead of six keywords, with the cohort assignment done blind by two independent raters and an inter-rater agreement score reported. That converts an illustrated mechanism into a defensible population estimate, and it is the single highest-value study nobody has published.

2. Citation-to-referral conversion, measured directly

We know 6.8% of conversations include a citation and we know AI referrals total 771m a month. Nobody has measured, for a fixed set of domains, what share of citations becomes a visit. That ratio is the exchange rate between AI visibility work and revenue, and it is currently unpriced.

3. Whether the assistant tail cites differently by intent

Section 17 shows those six platforms disagree about sources. It does not show whether that disagreement is stable across query types, or whether it collapses on high-commercial-intent questions where all six converge on the same three vendor pages. That is a two-week study and it changes where a B2B content budget goes.

4. What happens to a query after it stops being searched

The cohort A queries fell 64.2%. I have inferred that the demand moved into tools. I have not proved it, because nobody outside the labs can see the denominator. A study that instruments a cohort of developers directly – what they asked, where they asked it – would settle the mechanism this whole report rests on.

> **IF YOU RUN ANY OF THESE**
>
> Tell me, and I will link to it from the insights index. The reason this report names its own gaps is that they are more useful published than hoarded.

27

Glossary

Every term used in this report that could reasonably mean two things. Fuller definitions for 54 terms are at the AI search glossary.

Zero-click search	A search that ends without the user clicking any result. Measured here on a desktop and mobile web panel, excluding searches inside Google's mobile app.
Clicks per search	Estimated clicks a query sends to the open web, divided by its monthly search volume. Below 1.0 means the average search sends nobody anywhere; above 1.0 means searchers open several results.
Cohort A / cohort B	This report's split. Cohort A is queries an assistant can complete inside a tool the user is already in. Cohort B is queries where the answer is not the deliverable – physical, local or comparison intent.
Open weights	Model weights published under a permissive licence such as MIT or Apache 2.0. *Partial* means a custom, research-only or revenue-capped licence. *Proprietary* means API or hosted access only.
Reasoning model	A model that spends additional compute at inference time before answering. The category does not appear in this dataset before 2024.
Context window	How much text a model can consider at once, in tokens. The median in this dataset rose 488-fold between 2022 and 2026.
Run rate / ARR	One month's revenue multiplied by twelve. Not audited revenue, and not the same as booked revenue: the gap runs to 5.6x in this dataset.
Chosen vs bundled	Whether a user actively picked an assistant or received one inside something they already used. 7 of the 12 products with a headline number are not chosen use (58.3% by product, 41.2% by user).
AI referral	A visit to a website originating from an AI assistant's answer. Currently 771m a month globally and growing 117.4% year on year.
Citation	A link an assistant shows alongside an answer. Only 6.8% of all assistant conversations contain one, which caps how much referral traffic the channel can produce.
Superseded	Replaced by a newer model in the same line by the same vendor, usually still callable. Distinct from *retired*, which is withdrawn.
Frontier ceiling / price floor	The highest output price charged by OpenAI, Google or Anthropic in a year, against the lowest output price on any general-purpose model. They moved in opposite directions.

28

The full data behind Parts II and III

Everything charted above, as numbers. The complete 261-row release table and the raw Ahrefs pulls are in the data appendix that ships with this report.

YEAR	RELEASES	MEAN GAP, DAYS	MEDIAN CONTEXT	OPEN OR PARTIAL	TEXT-ONLY	MULTIMODAL	REASONING	CHEAPEST $/M OUT
2018	2	–	512	100.0%	2	0	0	–
2019	2	–	768	100.0%	2	0	0	–
2020	1	–	2,048	0.0%	1	0	0	–
2021	3	–	3,072	0.0%	2	0	0	–
2022	5	–	2,048	40.0%	4	0	0	–
2023	33	9.12	8,192	45.5%	25	5	0	$0.60
2024	50	7.04	128,000	60.0%	24	20	3	$0.03
2025	84	3.99	200,000	47.6%	6	33	47	$0.10
2026	81	2.98	1,000,000	38.3%	0	49	60	$0.05

A model is counted into every capability its category names, so the text-only, multimodal and reasoning columns overlap and do not sum to the release count: 33.3% of releases carry a compound category.

KEYWORD	COHORT	PEAK YEAR	PEAK-YEAR MEAN	2026 MEAN	FALL	CLICKS PER SEARCH
python list comprehension	A	2022	30,927	11,647	-62.3%	1.05
javascript array methods	A	2022	1,942	491	-74.7%	–
excel vlookup formula	A	2023	597	266	-55.4%	–
how to tie a tie	B	2026	440,663	440,663	0.0%	0.60
restaurants near me	B	2023	7,162,051	6,188,731	-13.6%	–
how to write a cover letter	B	2026	164,463	164,463	0.0%	1.13

28 · THE FULL DATA BEHIND PARTS II AND III · CONTINUED

DOMAIN	ALL PLATFORMS	THESE SIX	PAGES	PER PAGE
youtube.com	76,021,390	21,360,127	40,601,546	1.87
wikipedia.org	52,322,177	20,309,833	6,453,483	8.11
reddit.com	38,968,656	15,335,843	13,715,764	2.84
stackoverflow.com	260,953	14,357	176,715	1.48
g2.com	182,913	158,165	54,862	3.33

DOMAIN	CHATGPT	AI OVERVIEWS	AI MODE	GEMINI	PERPLEXITY	COPILOT
youtube.com	3.1%	41.7%	30.8%	5.5%	17.8%	1.1%
wikipedia.org	10.6%	17.2%	15.7%	9.6%	37.5%	9.4%
reddit.com	6.5%	28.8%	23.6%	11.2%	29.7%	0.2%
g2.com	11.3%	8.1%	13.0%	8.6%	38.7%	20.3%
stackoverflow.com	14.3%	13.7%	14.7%	1.2%	54.2%	1.9%

Two denominators, deliberately. Section 17's ratios divide by *all platforms*; the platform-mix percentages divide by *these six*, because that chart is about how one domain's citations split between named platforms. The six named platforms carry 34.1% of all citations, so the columns are not interchangeable.

BRAND QUERY	US VOLUME	GLOBAL VOLUME
chatgpt	68,580,000	811,200,000
google	36,030,000	141,800,000
gemini	10,640,000	107,100,000
copilot	2,450,000	17,940,000
claude ai	604,000	12,150,000
deepseek	114,000	8,270,000
grok	1,910,000	7,440,000
perplexity ai	357,000	3,280,000

The citation tables are a deliberate five-domain probe of reference points, not a ranking of the web. The brand table is why section 15 matters: at 811,200,000 global monthly searches, *chatgpt* is searched 5.72 times as often as *google*, and the seven assistant terms together 6.82 times as often. People are using a search engine to find the thing that replaces it.

THE WHOLE REPORT IN ONE LINE

Search demand did not fall. 811,200,000 searches a month are for 'chatgpt' alone. What fell is the share of searches that end anywhere other than inside the answer.

29

Further reading and how to cite

Thirteen pieces that go deeper on specific findings above. Each row says which section it speaks to, and every URL was fetched and confirmed live on 7 September 2026.

01 **How ChatGPT Shortlists Software Brands**

A 35-page audit of 60 buying questions across 10 software categories. 65.9% of links went to vendors' own sites, 88.9% of the pages ChatGPT read never appeared in an answer, and only 34% of a shortlist survived a reworded question.

THE MECHANISM BEHIND SECTIONS 15 AND 17

02 **Citation Share Is Not Traffic**

Why the three architectures of AI search – retrieval-first, hybrid and pure parametric – need different work, and why being cited often has no relationship to being visited.

THE DISTINCTION UNDERNEATH SECTIONS 14 AND 18

03 **The Sites Beating G2 in ChatGPT**

Forty-three comparison sites nobody tracks took 14.4% of citations, 3.5 times the combined total of G2, Capterra, TrustRadius, Software Advice and TechnologyAdvice. Format beat reputation.

WHY SECTION 17 FINDS G2 LEVEL WITH STACK OVERFLOW

04 **Accounting Software AI Citations: 90.9% Vendor-Owned**

Where a category publishes its prices, the vendors own the citation surface: 90.9% in accounting against 48.7% in healthcare EHR, where pricing is gated.

A WORKED EXAMPLE OF THE PATTERN IN SECTION 17

05 **How AI Overviews Will Change SEO**

The five structural shifts as AI Overviews stop being experimental and become the default discovery layer, and what each one breaks in a conventional search programme.

PRACTICAL FOLLOW-ON FROM SECTIONS 12 AND 13

06 **AI Browsers Are Rewriting the Rules of Customer Acquisition**

When an agent visits instead of a person, engagement signals collapse: it does not dwell, scroll or interact. How to build for two readers at once.

WHAT SECTION 20 IMPLIES FOR MEASUREMENT

07 **ChatGPT Ads: OpenAI Hits $1bn Run Rate**

A $1bn annualised run rate inside 200 days, across tens of thousands of advertisers. The paid layer of the assistant is now a permanent channel, not a pilot.

THE COUNTERWEIGHT TO SECTION 08 ON PRICE

08 **OpenAI Wiki Incident: When AI Agents Write to the Open Web**

The first public admission of agents editing third-party sites without a prompt, and what it changes about crawler policy and brand safety.

A RISK SECTION 21 DOES NOT PRICE IN

09 **Google Meridian GeoX: MMM Moves From Correlation to Causation**

Open-source geo-incrementality testing, validating mix models against real experiments. The measurement answer to a channel that refuses to show up in last-click.

HOW TO TEST THE RECOMMENDATIONS IN SECTION 22

29 · FURTHER READING AND HOW TO CITE · CONTINUED

10 **OpenAI Astra: The First Model to Hit the Critical Cyber Threshold**

The first model graded critical on cyber capability by its own maker: 100% on ExploitBench, and working exploits built against hardened systems. Why a capability class now ships restricted.

THE RESTRICTED RELEASES COUNTED IN SECTION 10

11 **Google Goto URL Redirects: What They Mean for Rank Tracking**

From 26 August 2026 Google routes result clicks through an intermediate redirect that blocks SERP scraping. Rankings are unchanged; what third-party tools can see is not.

A MEASUREMENT CAVEAT ON SECTIONS 16 AND 17

12 **Grok Bot Now Works With X: What It Means for GEO**

Grok reads X posts and mentions in real time, which turns social content into a retrieval surface an answer can quote. One more platform no citation audit is watching.

A SURFACE MISSING FROM THE SIX IN SECTION 19

13 **Discovery Digest · Issue 15**

The week GPT-6 Astra arrived and data sovereignty became infrastructure. A weekly read of what actually shipped, which is where this report's dataset comes from.

HOW TO KEEP THIS REPORT CURRENT

Primary sources cited in this report

SOURCE	WHAT IT IS
Pew Research Center	Browsing data from 900 US adults, March 2025
SparkToro, reported by Search Engine Land	US Google searches, January to April 2026
Similarweb AI Search	AI referral traffic by industry, June 2025 to May 2026
OpenAI's own usage analysis	1.5 million conversations, published with NBER working paper 34255
Alphabet	Q2 2026 earnings, CEO letter
Eight Oh Two, reported by Search Engine Land	Survey of 500 consumers, November 2025

How to cite this

Mzumara, N. (2026). *How AI Changed the Way People Search: 261 model releases read against what happened to search.*
nathanmzumara.com. Captured 7 September 2026.

Carry the sample with the number. '64.2% of queries fell' is not a finding; '64.2% mean fall across three assistant-completable keywords, peak year to 2026, US, Ahrefs' is. Every figure in this report is computed from source data and independently re-derived through a second code path; 232 such checks ran before this file was rendered.