IPURPOSE · RESEARCH GUIDE · 18 OCTOBER 2026

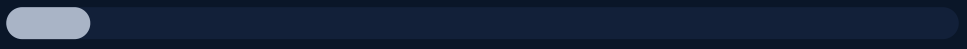
The two meters

What it really costs to make AI answer from your own documents, and where the value hides.

Indexing 720 million pages of research costs about $1 million. Answering questions about them could cost more than $100 million a year.

That gap is the whole story.

Estimates for a library the size of ResearchGate. Method and sources on page 14.



BUILD METER

Paid once per document. About a tenth of a cent a page.



QUESTION METER

Paid every time anyone asks anything. It never stops running.

$0.001
per page to build the index, once

$0.005 to $0.25
per question to use it, every time

$0
for the engine itself, released under the MIT licence

Jesse James
Founder, iPurpose

Intelligence with intent.

THE PROBLEM

The answer is on page 143

It is 4:40 on a Thursday. A councillor emails the finance office of a town of 9,000 people.

How much long-term debt do we carry, and how does that compare with the town next door?

The answer exists. It sits deep in the notes of a 312-page financial statement, and somewhere in the neighbour's. Someone has to find both pages before the meeting.

[Illustrative case, not a sourced event]

A new class of software promises to find them in seconds. One open-source engine, PageIndex, has drawn about 39,000 stars on GitHub by doing it without the machinery most AI search depends on. [Source: VectifyAI/PageIndex repository, Oct 2026]

This guide puts a price on it.

IN ONE MINUTE

01 **Building the index is cheap.** About a tenth of a cent per page, paid once per document.

02 **Asking questions is where the money goes.** Every question pays a model to reason, and that meter never stops.

03 **The advantage grows with length.** In small independent tests, it showed no clear accuracy win over simpler methods.

04 **The engine is free, so the engine is not the product.** The value sits in choosing the documents, keeping them current and standing behind the answers.

To see why the bill lands where it does, start with how a machine finds a page at all.

01 · HOW A MACHINE FINDS A PAGE

Three ways to look

Ask a computer where the long-term debt figure lives, and it has three ways to look.

Keyword search matches words. It finds every page containing “long-term debt.” If the statement says “non-current borrowings” instead, it finds nothing. It is fast, cheap and literal.

Vector search matches meaning. It turns every passage into a long list of numbers, called an embedding, and returns the passages whose numbers sit closest to your question’s. Synonyms land near each other. But close is not the same as relevant, and the search has no idea where in the document it is standing.

Reasoning-based retrieval reads the map first. It builds a table of contents for each document, then lets a language model decide which section to open, the way an analyst flips straight to the notes.

FIG. 01

Same question, three machines

“How much long-term debt does the town carry?”

KEYWORD	VECTOR	REASONING
Matches words	**Matches meaning**	**Reads the map**
FINDS: Pages with the exact phrase	FINDS: Passages that sound similar	FINDS: The section that should hold the answer
WEAK SPOT: Synonyms like “non-current borrowings”	WEAK SPOT: Where the passage sits in the document	WEAK SPOT: Cost and speed. Little edge on short documents
SPEED: Instant	SPEED: Under a second	SPEED: 3 to 8 seconds
COST PER LOOKUP: Close to zero	COST PER LOOKUP: Close to zero	COST PER LOOKUP: Several model calls

Speeds from a DataCamp test on a 60-page Federal Reserve report: 3 to 8 seconds per question for PageIndex, under one second for a cached vector index.

PageIndex is the third kind. What it keeps on disk is surprisingly small.

02 · WHAT THE ENGINE KEEPS

A table of contents that explains itself

There is no vector database. PageIndex saves one small tree per document, in plain text a person can read. [Source: PageIndex README]

Each branch holds a section title, a one-line summary, the pages it covers and the sections beneath it. That is the whole index.

FIG. 02

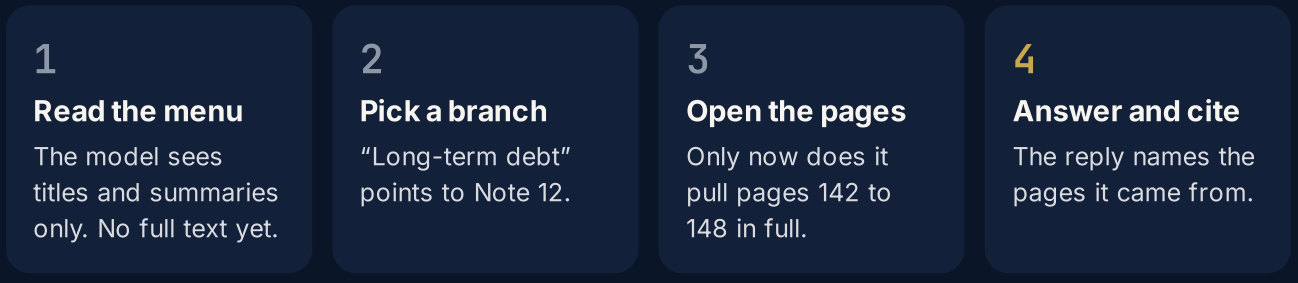
One document, one tree

Illustrative. Section names and page ranges are invented for the example.



FIG. 03

How a question walks the tree



Navigate first, read last.

That is how a competent person handles a 300-page report. It also explains the bill, because every step is a call to a model.

03 · THE TWO METERS

One purchase, one utility bill

Every system like this runs two meters.

The build meter runs once per document. A model reads the file, writes its tree and saves it. With a basic model, that costs about a tenth of a cent per page. A 1,000-page book runs a little over a dollar. [Source: PageIndex README, Oct 2026]

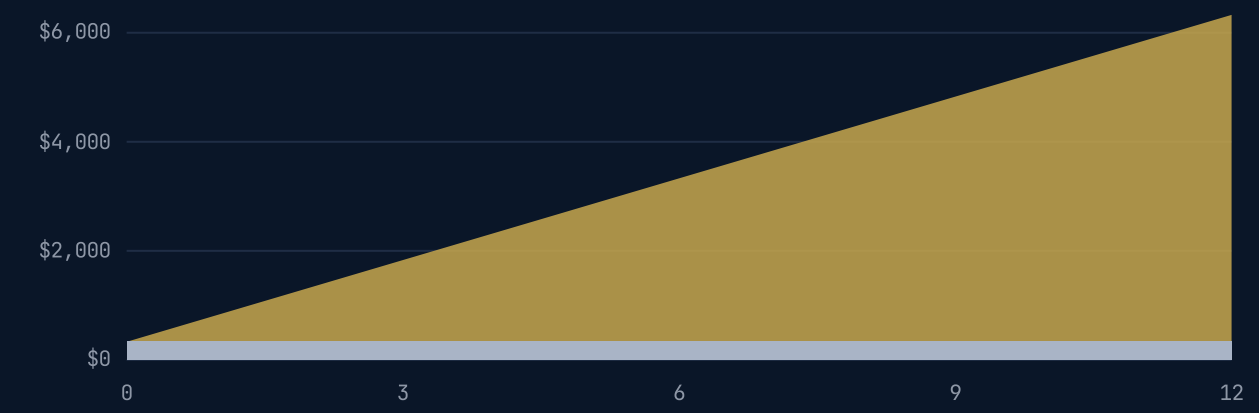
The question meter runs every time anyone asks anything. The model reads the tree, picks a branch, opens pages and writes an answer, and the model provider bills each step.

You pay the build meter again only for new documents, or when a source changes, such as an amended budget. Rent the hosted version instead and storage adds $0.001 per page per month. [Source: PageIndex pricing, Oct 2026]

FIG. 04

A year on both meters

3,000 municipal filings, about 300,000 pages, on the free engine. 10,000 questions a month on a mid-tier model. Cumulative cost by month.



Estimate. Build: 300,000 pages at $0.0011 a page. Questions: 10,000 a month at about $0.05 each. See page 7 for per-question ranges.

This is not training

Nothing about the model changes. The index is a librarian placed in front of an off-the-shelf model, handing it the right pages. Call it training in front of an engineer and they will expect something you did not build.

04 · THE STRESS TEST

Run both meters on 720 million pages

To see where the meters break, run them on the biggest research library most people know.

ResearchGate lists more than 160 million publication pages and more than 25 million members. [Source: ResearchGate, Oct 2026]

Not every page carries a paper. A 2015 study found full text on 56% of the pages it sampled, and about 19 million full texts existed that year. [Sources: Thelwall and Kousha, 2015; Jamali, 2017] The company publishes no current count, so the figures below are estimates built on those anchors.

TABLE 01 · BUILD METER AT RESEARCHGATE SCALE (ESTIMATE)

SCENARIO	FULL-TEXT PDFS	PAGES	FREE ENGINE, ONCE	HOSTED, ONCE	HOSTED STORAGE, PER YEAR
Low 10 pages per paper	30M	300M	$0.33M	$3.0M	$3.6M
Middle 12 pages per paper	60M	720M	$0.8M	$7.2M	$8.6M
High 18 pages per paper	100M	1.8B	$2.0M	$18M	$21.6M

Free engine at about $0.0011 a page. Hosted at list price: $0.01 a page to index, $0.001 a page per month to keep active. A buyer this size would negotiate. OCR for older scanned papers adds roughly $20,000 to $540,000.

Even the high case comes to about $2 million on the free engine, paid once.

[My view] For a platform of that size, that is small money.

> *The build meter is a rounding error. Then people start asking questions.*

04 · THE STRESS TEST, CONTINUED

The meter that never stops

A question costs whatever the model charges to read the tree, open the pages and answer.

PageIndex's own chart puts a question between $0.003 and $0.10. An independent test measured $0.0056 to $0.0091 on a budget model, and about $0.09 to $0.10 on a faster mid-range one. [Sources: PageIndex; Jivraj-18 benchmark, Sept 2026]

TABLE 02 · COST PER QUESTION (ESTIMATE)

MODEL TIER	ONE PAPER	ACROSS MANY PAPERS	VECTOR SEARCH, FOR COMPARISON
Budget	~$0.005	~$0.02	~$0.001
Mid-tier	~$0.05	~$0.15 to $0.25	~$0.013
Premium	~$0.25	~$0.75 to $1.25	~$0.065

Model prices change monthly, so this guide uses tiers instead of model names.

TABLE 03 · REASONING VERSUS VECTORS ON THE SAME LIBRARY (ESTIMATE)

METER	REASONING (PAGEINDEX)	VECTOR SEARCH
Build, 720M pages One-time indexing	~$0.8M	~$33K to $65K
Storage Managed hosting	~$100s/mo	~$1.4K to $3K/mo
Per question Same model tier	4 to 5× more	Baseline
Speed Per question	3 to 8 seconds	Under 1 second

Vector build: about 700 tokens a page with a large embedding model at batch pricing. Hosting at managed serverless rates before read charges. [Sources: OpenAI and Pinecone pricing via 2026 trackers; DataCamp]

Vectors win on both meters. Reasoning buys a citable path to the right page in a long document, which is worth the premium only when the answer is.

04 · THE STRESS TEST, CONTINUED

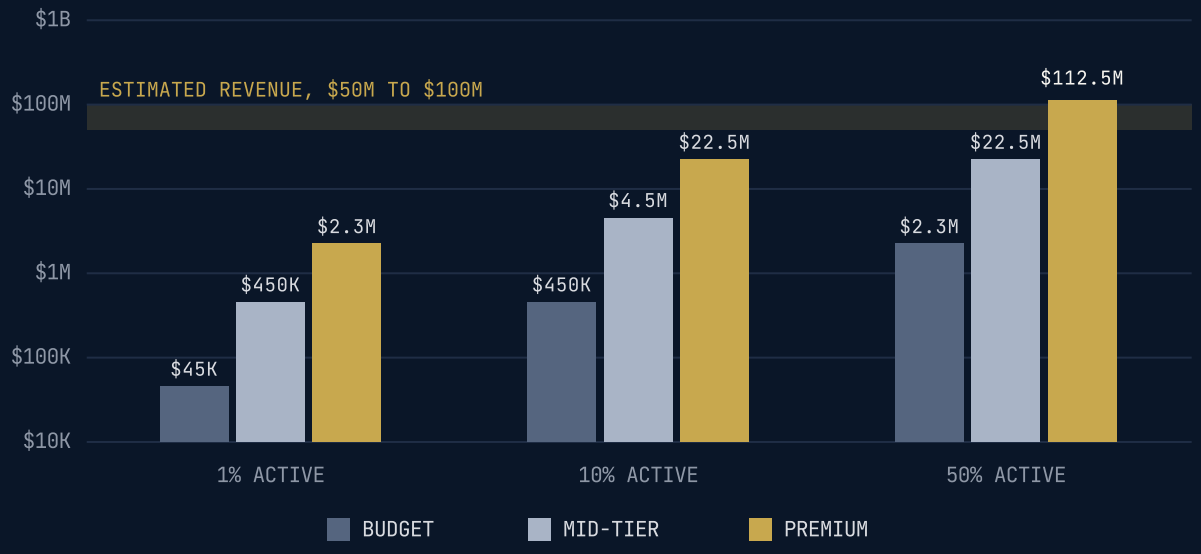
Where the bill meets the business

Now multiply. Assume each active member asks three questions a month, on single papers only.

FIG. 05

A year of questions at ResearchGate scale

25 million members, three questions per active member per month. Log scale: each gridline is ten times the one below.



Estimate. Per-question costs from Table 02. Revenue band is a third-party estimate; ResearchGate does not disclose revenue. [Source: LeadIQ]

At half the membership on a premium model, the question meter runs about $112 million a year. Offer premium answers free at that adoption, and the questions could cost more than the company is estimated to earn.

The fix is ordinary product economics: a cheap model for free users and a paid tier for heavy ones. At $10 a month, 250,000 subscribers bring in $30 million a year against about $4.5 million in mid-tier question costs, at 30 questions each a month. [Hypothetical]

Cost says whether it can ship. Accuracy says whether it should.

05 · THE EVIDENCE

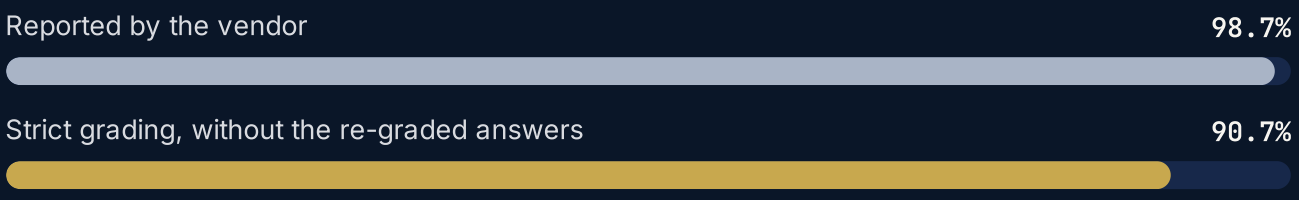
Where it wins, and where it doesn’t

VectifyAI, the company behind PageIndex, reports 98.7% accuracy on FinanceBench, a test built from questions about public company filings.

Twelve of those answers counted as correct only after the company’s own re-grading. Graded strictly, the figure is 90.7%. [Source: VectifyAI, Mafin 2.5 FinanceBench results]

FIG. 06

One benchmark, two scores



FinanceBench, vendor results. Independent replications use far fewer questions.

Independent tests are smaller and cooler.

INDEPENDENT PILOT · 28 QUESTIONS

No clear win

PageIndex never beat simpler methods. A plain search-and-read agent matched it at roughly a tenth of the cost, with gaps inside the noise.

[Source: Jivraj-18 benchmark, 30 Sept 2026]

DATACAMP · ONE FED REPORT

Two draws, one loss

Vector search won the third question. The tester concluded the advantage sharpens as documents get longer.

[Source: DataCamp PageIndex tutorial]

The vendor’s newer benchmark uses 62 text questions and leaves out tables and figures. In financial documents, that is often where the answer lives.

The accuracy evidence is mixed. The cost evidence is not, and it turns on one variable: length.

05 · THE EVIDENCE, CONTINUED

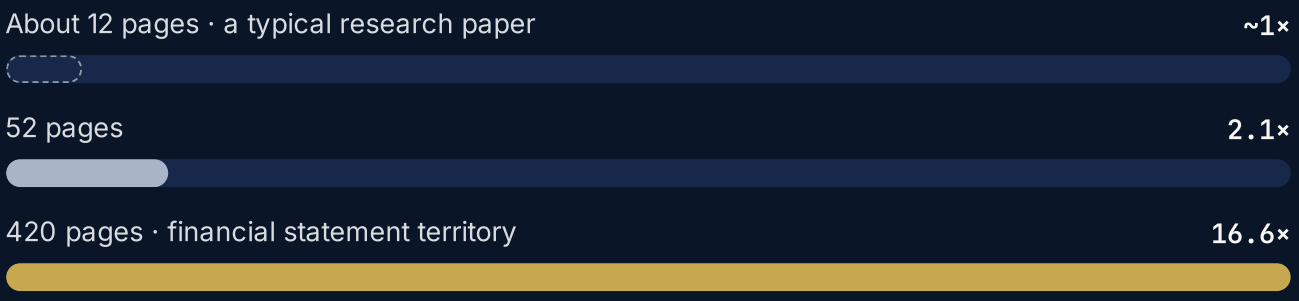
Length is the variable that matters

The simplest rival to any index is no index at all: paste the whole document into the prompt.

VectifyAI's own numbers show when that stops making sense. Sending a full 52-page PDF to the model costs 2.1 times as much as retrieving with PageIndex. At 420 pages, it costs 16.6 times as much. [Source: PageIndex README]

FIG. 07

Cost of pasting the whole document, relative to retrieval



52 and 420 pages: vendor figures. 12 pages: estimate, extrapolated. Dashed bar marks the estimate.

On a 12-page paper, pasting the whole thing is simpler, needs no index and costs about the same.

[My view] That makes a research library the wrong place to start.

On a 400-page financial statement, with notes, schedules and cross-references, the arithmetic flips. Financial statements, budgets and bond disclosures commonly run 100 to 500 pages, the terrain this engine was built for.

Seconds, not milliseconds

In DataCamp's test each question took 3 to 8 seconds, against under one second for cached vectors. [Source: DataCamp]

[My view] Fine for an analyst. Too slow for a search box.

> *Small documents don't need a map. Big ones do.*

06 · AT YOUR SCALE

A library one person can keep

Shrink the library to something one person could choose, check and maintain.

TABLE 04 · BOTH METERS, SMALL LIBRARIES (ESTIMATE)

LIBRARY	PAGES	BUILD, FREE ENGINE	BUILD, HOSTED	QUESTIONS, 10,000 A MONTH
A briefing's sources 50 government documents	5,000	~$5	$50 + ~$4/mo	$50 to $2,500/mo
1,000 PDFs 50 pages each	50,000	~$55	$500 + ~$49/mo	$50 to $2,500/mo
3,000 municipal filings 100 pages each	300,000	~$330	$3,000 + ~$299/mo	$50 to $2,500/mo

Hosted storage after the first 1,000 free pages. Question range runs from budget to premium tier. At 1.35 to 1.40 Canadian dollars per US dollar, the 3,000-filing library costs roughly C$450 to build on the free engine.

Price for the margin

A subscriber paying $50 a month who asks 100 questions on a mid-tier model costs about $5 in model calls. That leaves roughly 90% of the price before hosting and labour. Unlimited premium answers for free users is how the question meter eats a business. [Estimate]

Use the engine where it wins. Long financial documents first. Short council minutes and requests for proposals may be better served by plain search.

Older scanned records need the hosted version or your own OCR, because the free engine reads text-based PDFs only. [Source: PageIndex README]

Which leaves the uncomfortable question: if this is so cheap, what stops anyone from copying it?

07 · THE TURN

Nothing stops them

The engine is MIT-licensed. Anyone can download it this afternoon. The index costs cents. The model is rented from the same handful of providers everyone uses.

So none of that is the product.

FIG. 08

What a competitor can copy, and how fast

Framework reflects the author's view.

The engine Open source, free to anyone	Copied in a day
The index Cents per page, rebuilt on demand	Copied in a week
The model Rented by the question	Same for everyone
The map Which documents, which versions, kept current, with a name behind every answer	Copied in years, if ever

The value sits in the bottom layer. In practice that means five disciplines:

- **Complete, versioned sources.** Every fiscal year, every amendment.
- **Fast intake.** New filings indexed soon after they are posted.
- **Clean labels.** Municipality, fund, fiscal year and document type on every file.
- **Accountability.** A public corrections log, and a person who signs off.
- **A test set.** Around 200 questions with known answers, rerun after every change.

[My view] Anyone can download the code in a day. Few will keep 3,000 municipal filings current for five years.

08 · BEFORE YOU BUILD

What can still go wrong

RIGHTS

Other people’s documents carry other people’s claims

Publishers sued ResearchGate in 2017. A Munich court found it responsible for copyright-infringing content on its platform in 2022, and a confidential 2023 settlement added automated copyright checks. Its terms of service do not mention AI. Public records avoid most of this. [Sources: C&EN, 2022; Chemistry World, 2023; ResearchGate terms, 2024]

DISTORTION

The right page can still produce the wrong summary

A 2025 study of about 4,900 AI summaries of science papers found overgeneralized conclusions in up to 73% of outputs for some models. Asking the models to be accurate made it worse. [Source: Peters and Chin-Yee, Royal Society Open Science, 2025]

FRAGILITY

Scans and failures

The free engine cannot read scanned pages. In one independent test it failed to index 2 of 24 filings. [Sources: PageIndex README; Jivraj-18 benchmark]

TRY IT THIS WEEKEND

1 **Pick five sources you have already cited**
Text PDFs, not scans, whose answers you know cold.

2 **Index them on the free engine**
A few hundred pages costs well under a dollar.

3 **Ask what you already know**
Check every answer and every page citation against your own work.

4 **Ask one thing you don’t**
Then decide whether you would have trusted it.

THE CLOSE

Two meters, one moat

The build meter makes the library. The question meter makes the bill. Neither one makes the moat.

[My view] The moat is the map: the documents you chose, kept right, with your name on every answer.

SOURCES

VectifyAI, *PageIndex* repository and README, github.com/VectifyAI/PageIndex. Accessed October 2026.

PageIndex documentation, pricing page; PageIndex blog, "PageIndex File System," May 2026.

VectifyAI, Mafin 2.5 FinanceBench results.

Jivraj-18, *benchmark-finance-pageindex*, GitHub, 30 September 2026.

DataCamp, "PageIndex Tutorial: Reasoning-Based RAG Without Vectors."

ResearchGate site statistics and Marketing Solutions page, October 2026; ResearchGate Terms of Service, updated 20 February 2024.

Thelwall, M. and Kousha, K. (2015), study of ResearchGate full-text coverage.

Jamali, H. R. (2017), "Copyright compliance and infringement in ResearchGate full-text journal articles," *Scientometrics*.

LeadIQ, ResearchGate company profile (revenue estimate).

C&EN (2022) and Chemistry World (2023) on ACS and Elsevier v. ResearchGate.

Peters, U. and Chin-Yee, B. (2025), on generalization bias in LLM summaries of science, *Royal Society Open Science*.

Embedding, vector database and OCR prices from provider pricing pages and 2026 pricing trackers.

METHOD

ResearchGate's full-text count, query volume and revenue are not public. Corpus sizes are estimates extrapolated from 2015 anchors, and revenue comparisons rely on a third-party range.

Per-question costs are modelled from vendor and independent figures, not measured on academic papers. Model prices change monthly, so this guide uses tiers rather than model names. Independent evaluations of PageIndex are small, between 28 and 62 questions, and none test retrieval across millions of documents. Nothing here is legal advice.

Jesse James
Founder, iPurpose

Intelligence with intent.