OCTOBER 2026

AI in Contaminated Sites Work

A Practical Introduction

Presented by :

PAUL FUELLBRANDT, CEO · COURT SANDAU, PHD, CSO · JORDAN YEO, HEAD OF PRODUCT + AI

Agenda

1. Who we are, and goals for this workshop
2. Foundational vocabulary and background
3. Using AI responsibly: trust, privacy, choosing tools
4. AI for documents
5. AI for data
6. Why we're hopeful about AI
7. Questions

Who we are

Introduction

Our team



Paul Fuellbrandt, P.Ag.

CO-FOUNDER & CEO

Paul has 25 years of experience in environmental science across laboratory, consulting, site owner, and regulatory roles, developing closure plans and managing environmental liability for contaminated sites across Western Canada, and is now pioneering environmental knowledge management.



Court Sandau, PhD

CO-FOUNDER & CSO

Court is also the founder of Chemistry Matters, a niche environmental and forensic chemistry consultancy, where he applies compound identification and chemical fingerprinting to determine the origin and attribution of contamination in technically contested and litigious matters.



Jordan Yeo

HEAD OF PRODUCT & AI

Jordan leads product and AI at Statvis Analytics, bringing two decades of software development experience and a specialization in AI and retrieval augmented generation systems, following a product leadership role at Jobber.



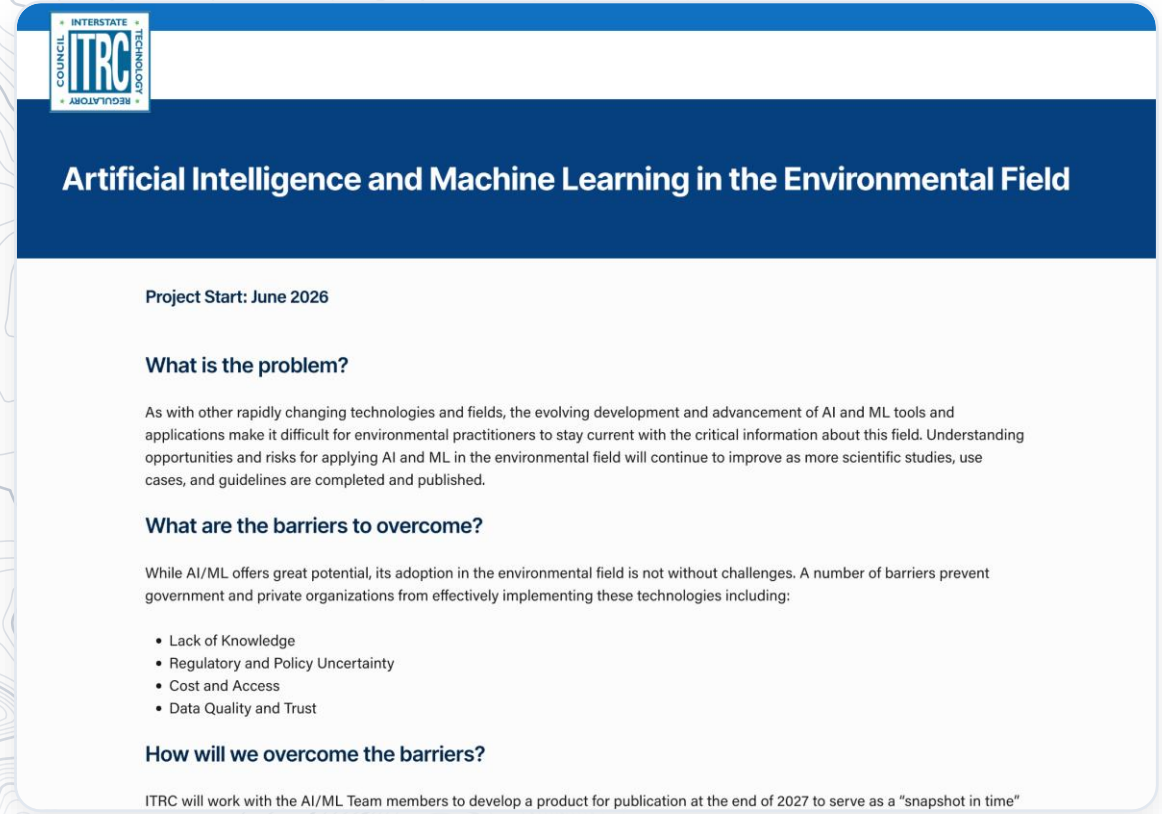
Carolyn Inglis

DIRECTOR OF BUSINESS DEVELOPMENT

Carolyn has 15+ years in the environmental space. creating exceptional customer experiences and helping organizations deliver lasting value. Intentionally building trust and relationships to create meaningful outcomes.

Contributing to the ITRC working group

- **ITRC, the Interstate Technology & Regulatory Council:** a state-led US coalition of regulators, industry, consultants and academics
- It turns new cleanup science into practical guidance and training, like its widely used PFAS guidance
- Its new team covers AI and machine learning in the environmental field. We contribute, and much of today's material comes from that work
- The goal: help practitioners keep up with fast-moving AI tools, and use them responsibly



Artificial Intelligence and Machine Learning in the Environmental Field

Project Start: June 2026

What is the problem?

As with other rapidly changing technologies and fields, the evolving development and advancement of AI and ML tools and applications make it difficult for environmental practitioners to stay current with the critical information about this field. Understanding opportunities and risks for applying AI and ML in the environmental field will continue to improve as more scientific studies, use cases, and guidelines are completed and published.

What are the barriers to overcome?

While AI/ML offers great potential, its adoption in the environmental field is not without challenges. A number of barriers prevent government and private organizations from effectively implementing these technologies including:

- Lack of Knowledge
- Regulatory and Policy Uncertainty
- Cost and Access
- Data Quality and Trust

How will we overcome the barriers?

ITRC will work with the AI/ML Team members to develop a product for publication at the end of 2027 to serve as a "snapshot in time"

ITRC · Artificial Intelligence and Machine Learning in the Environmental Field

Built on real project work, since 2018

Hundreds of thousands of pages

Reports, lab certificates, site plans and borehole logs, read and indexed.

Hundreds of billions of tokens

Read and reasoned over by AI, with every answer cited back to its page.

Time matters as much as volume. Since 2018 we've watched models, tools and vendors change underneath us, and learned what actually holds up.

Every lesson today comes from a real project.

Take our lessons so you don’t hit the same walls we did.

Goals for this workshop

You should leave with:

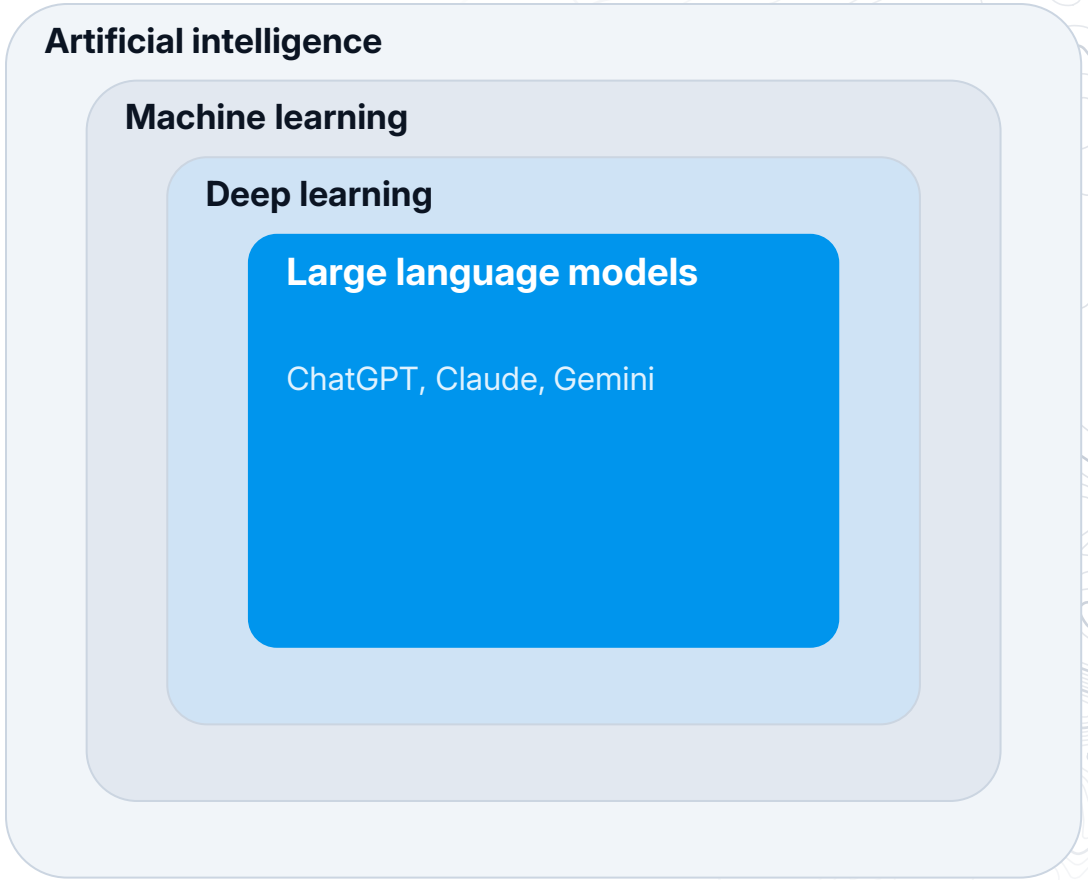
- A deeper understanding of how AI works, and the terms that matter
- When not to use AI
- Critical questions to ask an AI vendor or your IT team
- Best practices for real environmental workflows

One person can use these lessons tomorrow. Across a whole organization it takes more: safeguards for client data, and lessons built in and kept current. That system approach is what we've built at Statvis.

Foundational vocabulary & background

Part 1: How it works, in plain terms

AI vs ML vs LLMs



Artificial intelligence

The broad field: software doing tasks that normally need human judgement. Includes rules-based systems.

Machine learning

Models learned from data rather than hand-written rules: classify, cluster, predict. Usually built for one task.

Deep learning

Machine learning with large neural networks. Behind modern image recognition and OCR.

Large language models

Deep learning models trained on huge amounts of text to predict the next token. A kind of generative AI.

What is an LLM?

- **Large language model**
- Trained to predict the next word at massive scale
- Generates answers from patterns it was trained on
- Not a database of facts, or a search engine on its own



Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention

Vaswani et al., "Attention Is All You Need" (2017), the paper behind today's LLMs

Tokens: what the model reads, and how it answers

1 Text you type

Many words map to one token, but some don't: indivisible.

Unicode characters like emojis may be split into many tokens containing the underlying bytes: 🤚

Sequences of characters commonly found next to each other may be grouped together: 1234567890

2 What the model sees

Many words map to one token, but some don't: indivisible.

Unicode characters like emojis may be split into many tokens containing the underlying bytes: ����

Sequences of characters commonly found next to each other may be grouped together: 1234567890

52 tokens 246 characters

3 How it answers

It predicts the most likely next token, adds it, and goes again.

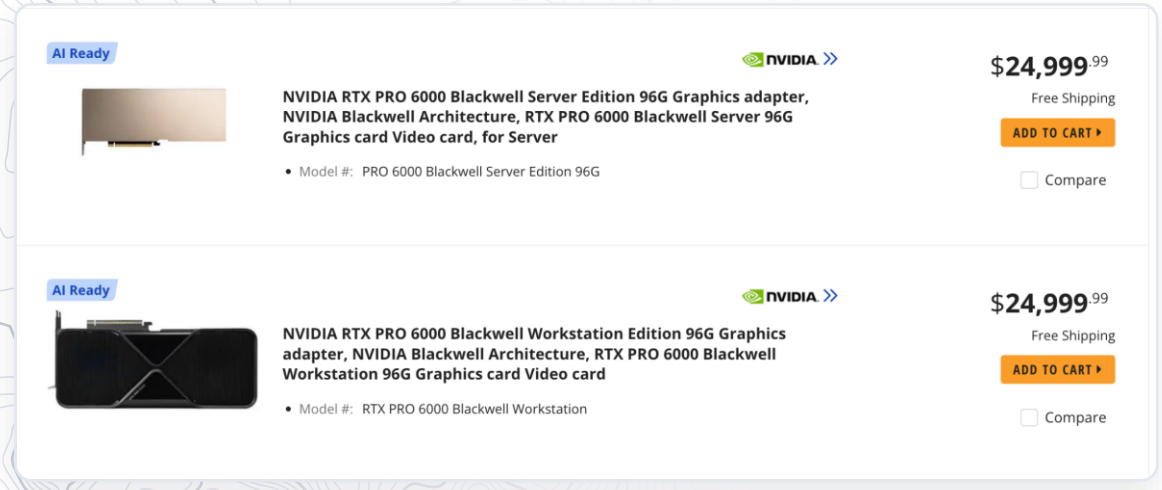
The
The **capital**
The capital **of**
The capital of **France**
The capital of France **is**
The capital of France is
Paris

Training gives words context: here “capital” means a city, not money.

GPT-4o tokenizer (o200k_base). Common words are one token; the emoji becomes four raw byte tokens.

Do you need your own proprietary LLM?

- **Certainly not.**
- Training your own model costs millions of dollars
- Some models allow fine-tuning
  - A small set of extra parameters that nudges how it responds
  - Better prompts almost always beat fine-tuning
- Running a model at useful speed needs a lot of GPU power



One data-centre GPU: about US$25,000. Serious models need many.

What are multimodal LLMs?

- Images, video and audio, as input and output, not just text
- More modes usually means more cost
- **Not all models are created equal:** quality varies wildly



“Create a ‘three wolf moon’ image, but instead of the wolf, use this puppy.”

The input



A weaker model



A stronger model



Then it became real

What's Retrieval Augmented Generation (RAG)?

1

User question

The user asks a question in natural language.

What happened to the old fuel tanks, and was any contamination found?

2

Retrieve

The system searches your documents for relevant passages.

Phase I ESA 1
Fuel sales cease, tanks are removed.

Phase I ESA 2
Four underground storage tanks (USTs)

Phase II ESA 3
Petroleum odour noted in sidewall soil.

3

Augment

The passages are added to the question and sent to the model.

Prompt to the model

USER QUESTION
What happened to the old fuel tanks, and was any contamination found?

+

RETRIEVED PASSAGES
1
2
3

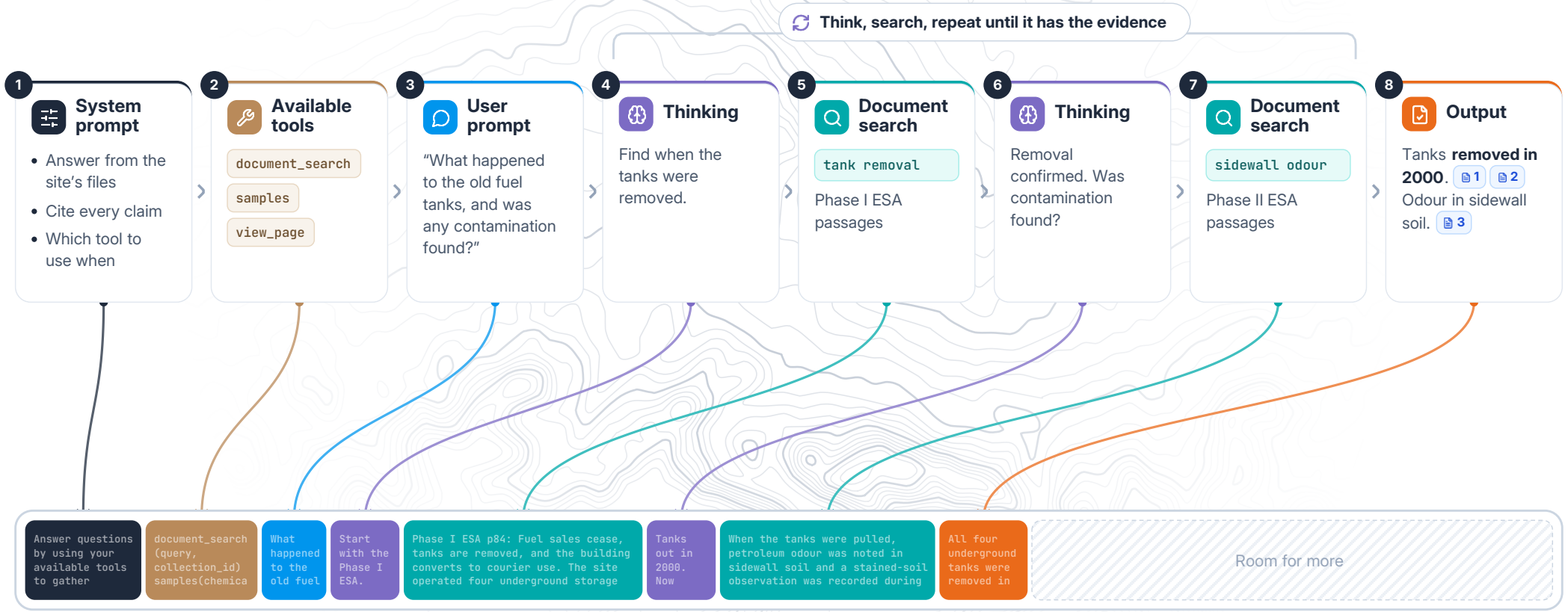
4

Generate

The model writes a grounded answer that cites its sources.

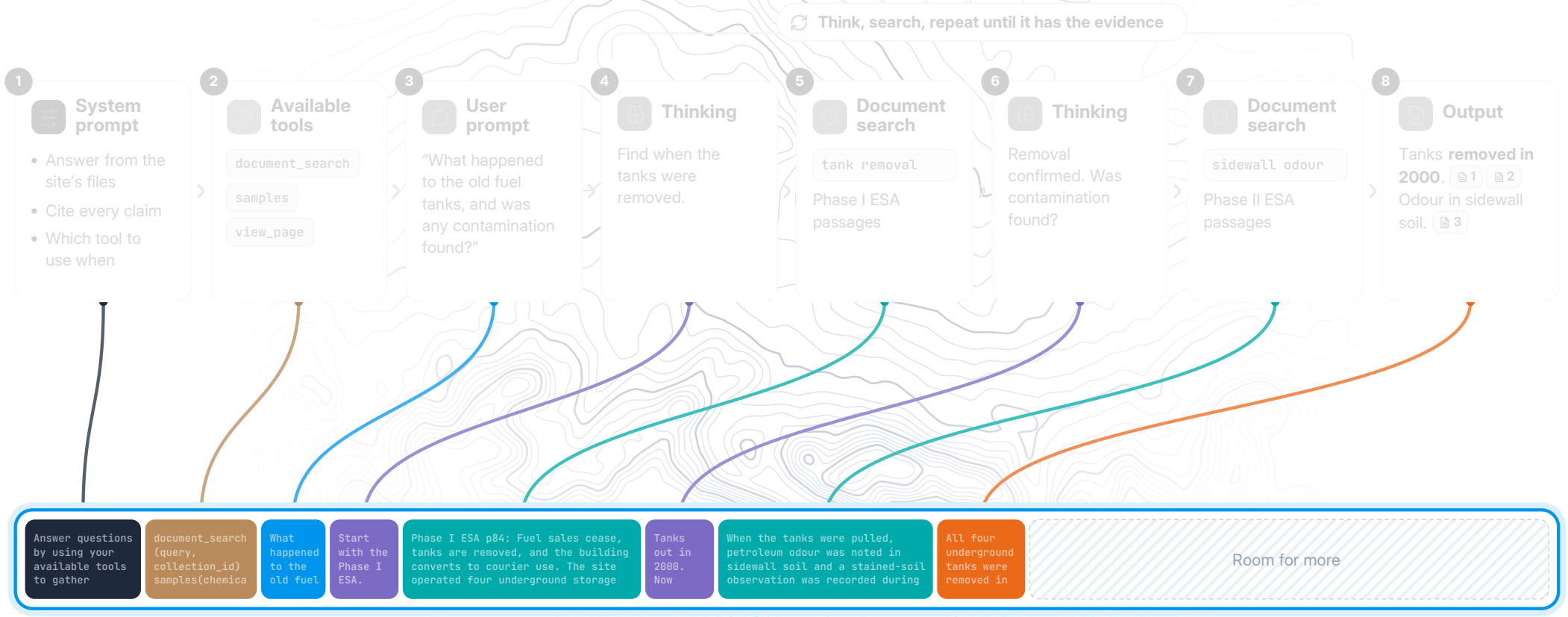
Statvis
All four underground tanks were **removed in 2000**, when fuel sales stopped. 1 2
As they came out, **petroleum odour was noted in the sidewall soil**. 3

“What happened to the old fuel tanks?”



Context window 1M tokens: everything the model sees, in order

“What happened to the old fuel tanks?”



Context window 1M tokens: everything the model sees, in order

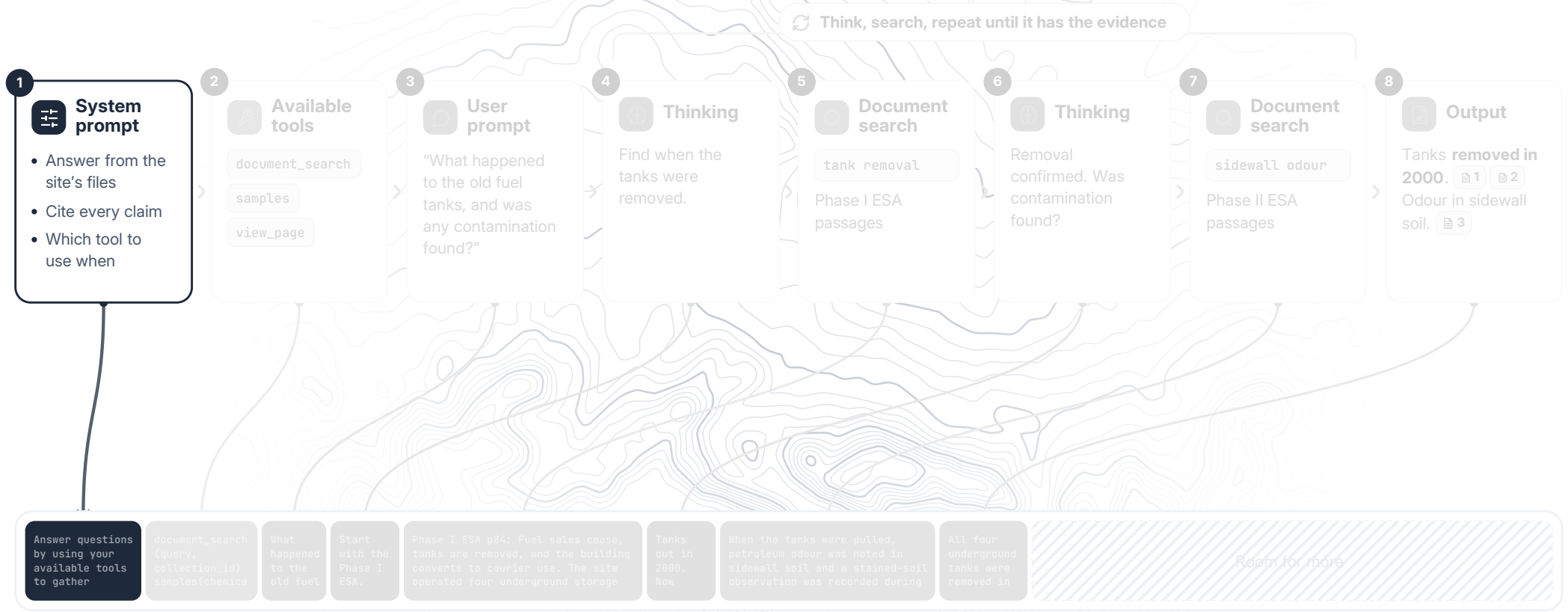
What’s a context window?

- “Working memory” for the model
- Holds both what goes in and what comes out
- 250k to 1M tokens, depending on model and provider
- When it fills, the conversation ends or is “compacted” (summarized)

> **Don’t fill it.** Studies find answer quality drops as the window fills, well before it’s full.

Liu et al., “Lost in the Middle” (TACL 2024) · Hsieh et al., “RULER” (COLM 2024) · Modarressi et al., “NoLiMa” (ICML 2025) · Hong et al., “Context Rot” (Chroma, 2025)

“What happened to the old fuel tanks?”



Context window 1M tokens: everything the model sees, in order

System prompt: standing rules for every turn

System prompt · sent at the start of every conversation

1. Environmental consultants use this tool to review the site file for Former Meridian Service Centre: Phase I and II ESAs, closure reports, lab certificates and monitoring reports.

2. Answer only from these documents. Use document_search to find relevant passages, then view_page to read the full page before relying on it.

3. Cite the page and passage for every fact. If the documents don't say, say that rather than filling the gap.

4. When sources disagree, show both with their citations and say which is the primary record. Never average or pick silently.

5. Start with a two-sentence answer, then the evidence as a dated list. Use plain technical English; no filler.

1. **Context, not a persona**
   Who uses it, which site, what's in the file.

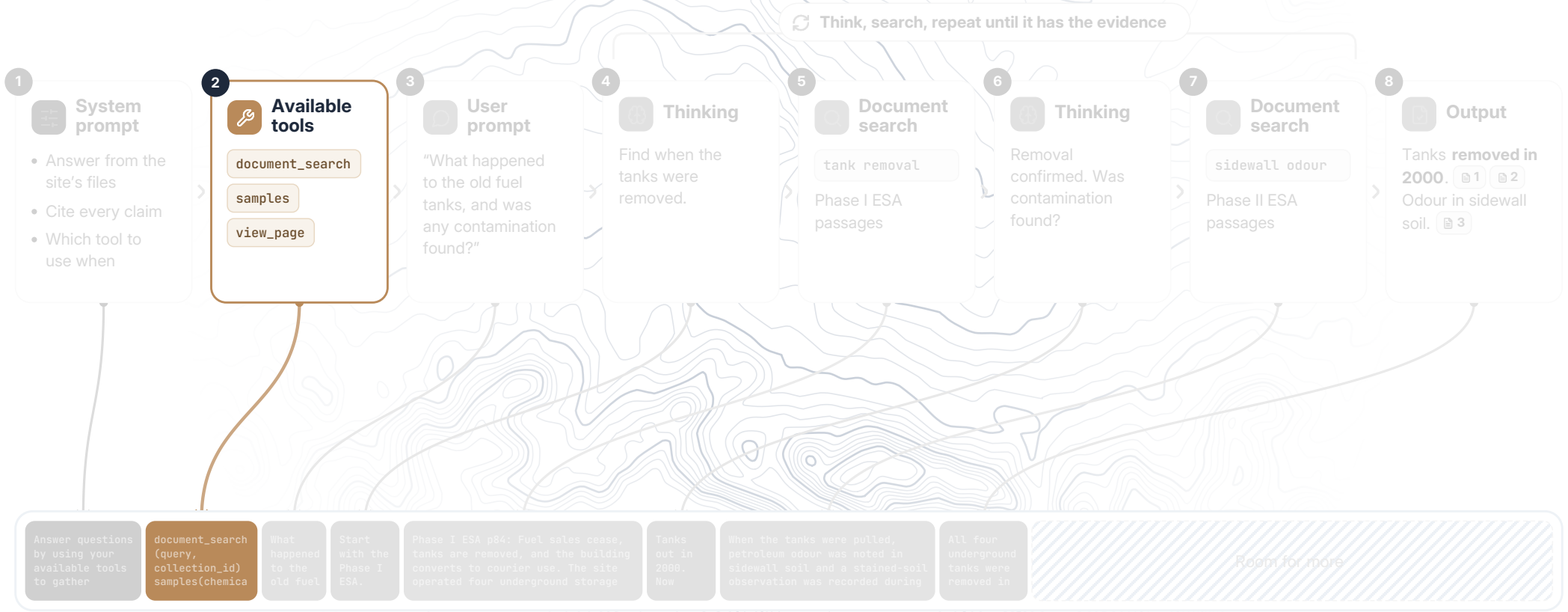
2. **Sources and tools**
   Only the site's documents; when to search vs read.

3. **Citation rule**
   Every fact traceable; permission to say "not found".

4. **Conflicts**
   Surface disagreement instead of hiding it.

5. **Output**
   Short answer first, then evidence.

“What happened to the old fuel tanks?”



Context window 1M tokens: everything the model sees, in order

What are tools?

- An action the model can ask for, but cannot do itself
  - Read a file, write a file, search documents, search the web
- The tool’s output goes back to the model
  - File contents, whether a write worked, matching passages, search results
- At the start of a conversation, every available tool is listed and described to the model

“What happened to the old fuel tanks?”



Context window 1M tokens: everything the model sees, in order

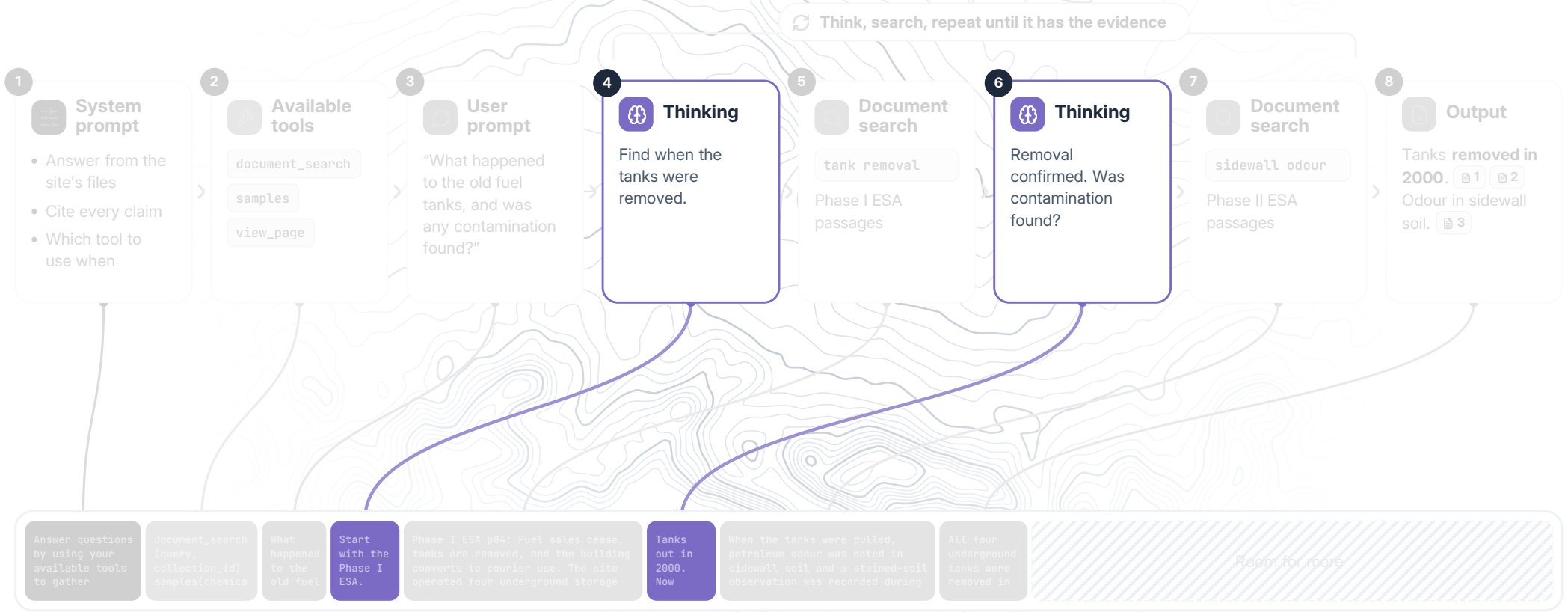
User prompt: what you actually ask

The one part of the context that changes every turn.

What happened to the old fuel tanks?

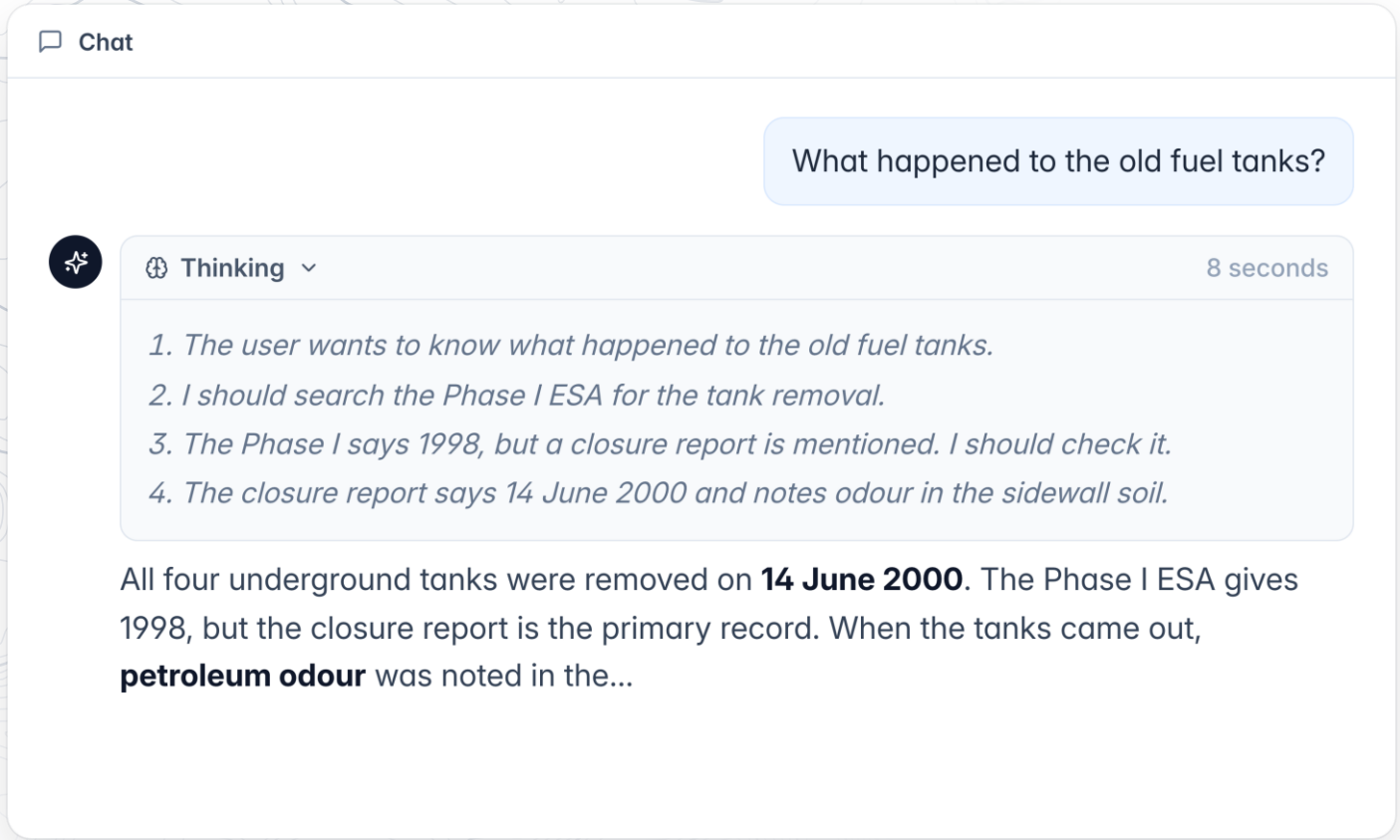
You

“What happened to the old fuel tanks?”

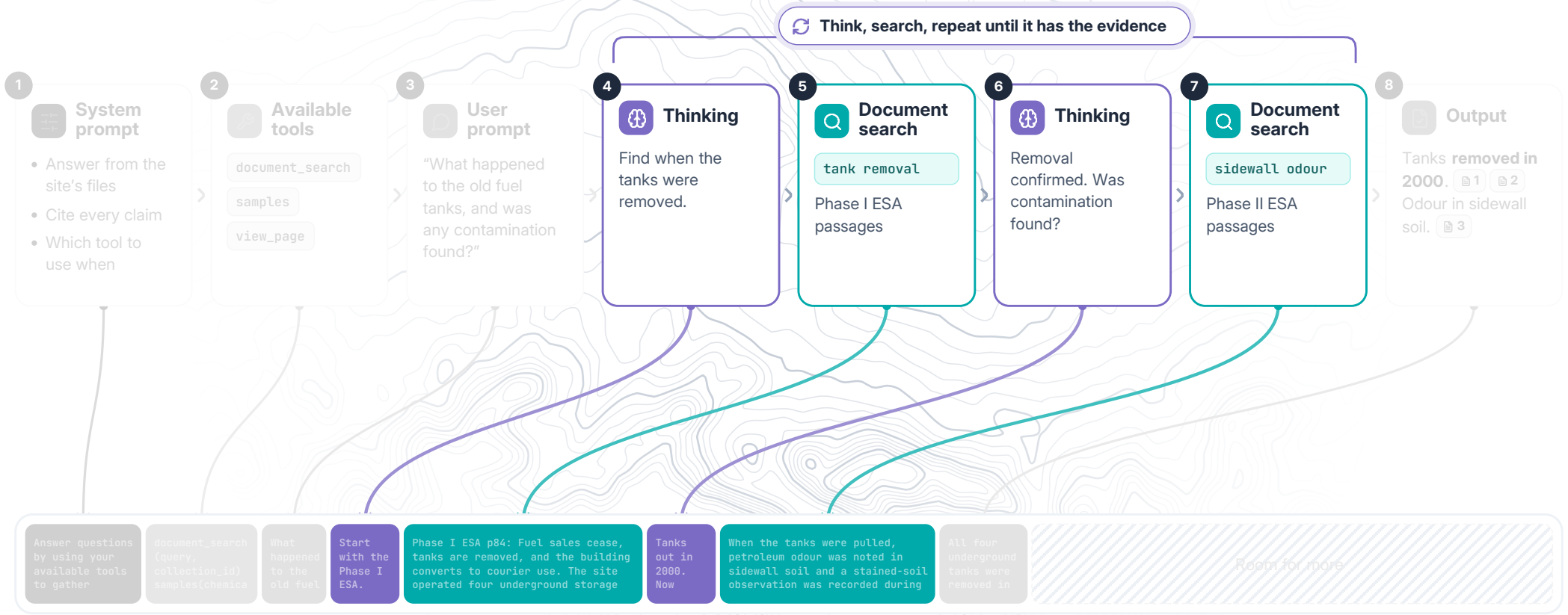


How does an LLM think?

- It is still all tokens: the same next-token prediction
- A trained behaviour: reason step by step before answering
- Thinking becomes input to the model's next action
- "Thinking budgets" cap how many tokens it can spend

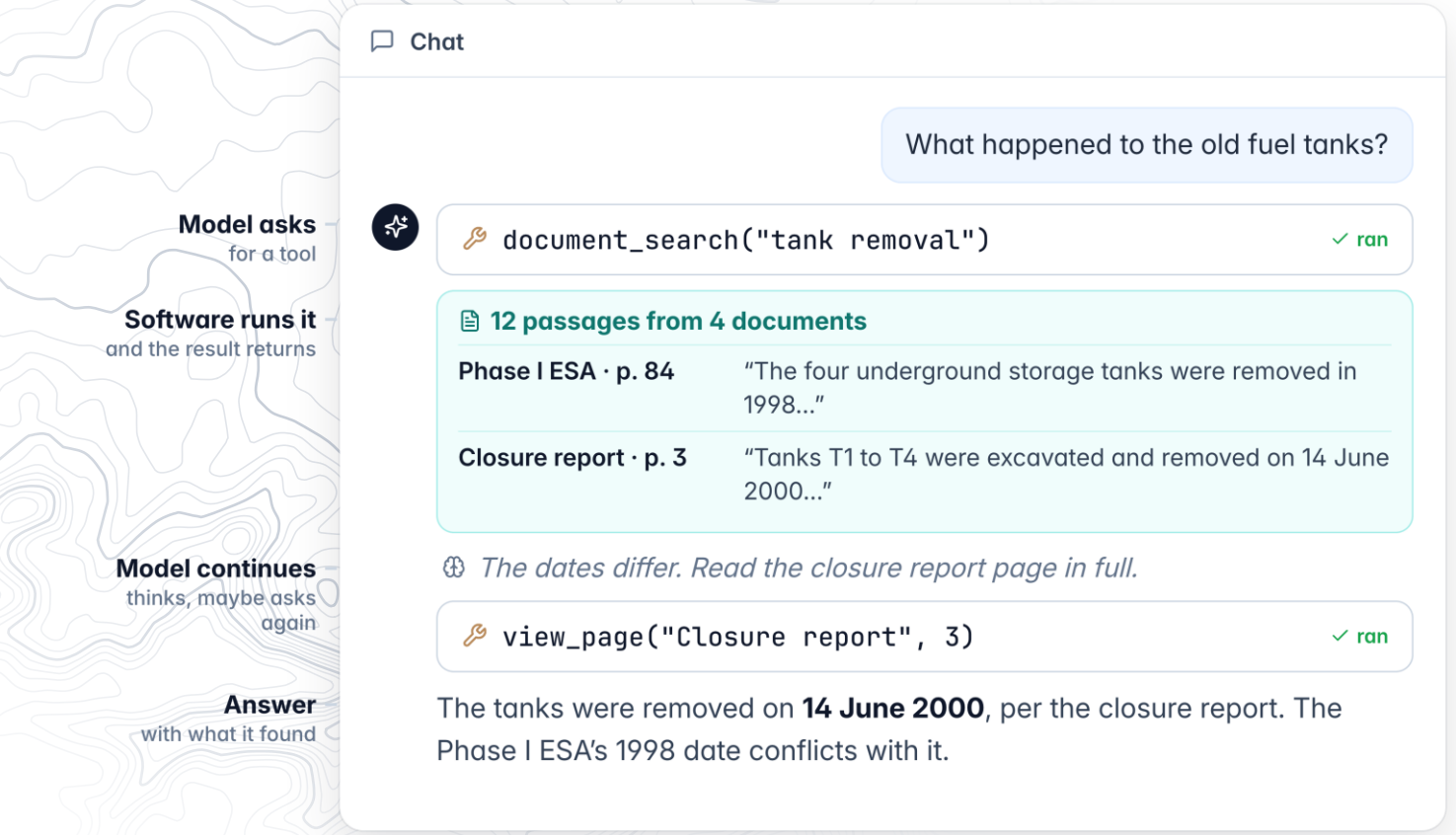


“What happened to the old fuel tanks?”

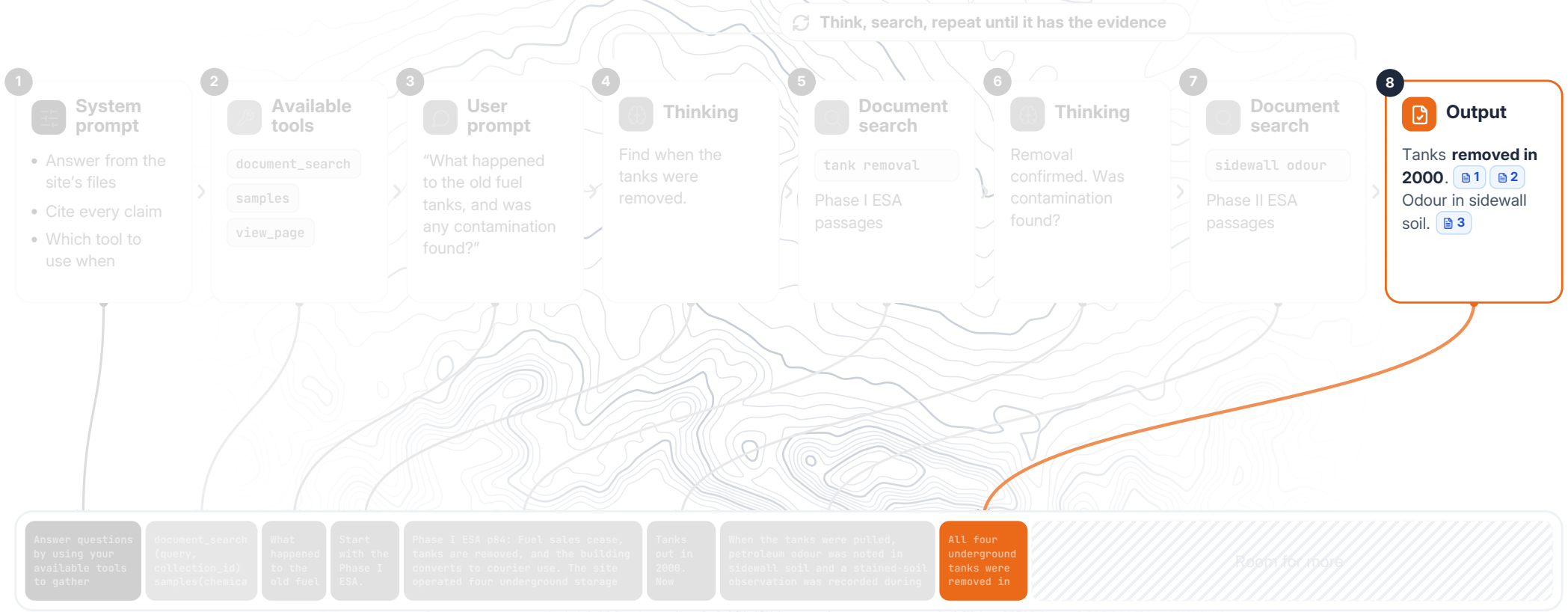


Tool use

- The model decides it needs a tool and asks for it
- The software around it runs the tool
- Results come back into the context window
- The model thinks again, and repeats until it has the evidence

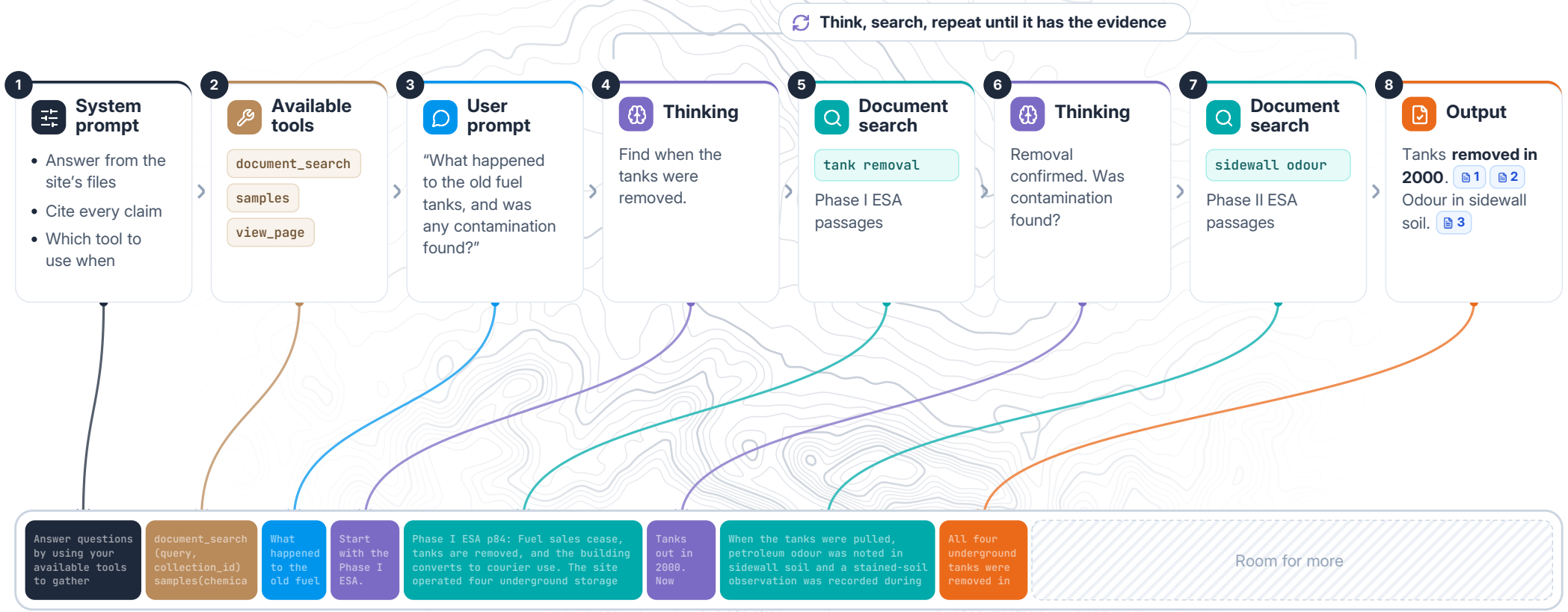


“What happened to the old fuel tanks?”



Context window 1M tokens: everything the model sees, in order

“What happened to the old fuel tanks?”



Context window 1M tokens: everything the model sees, in order

What's a harness?

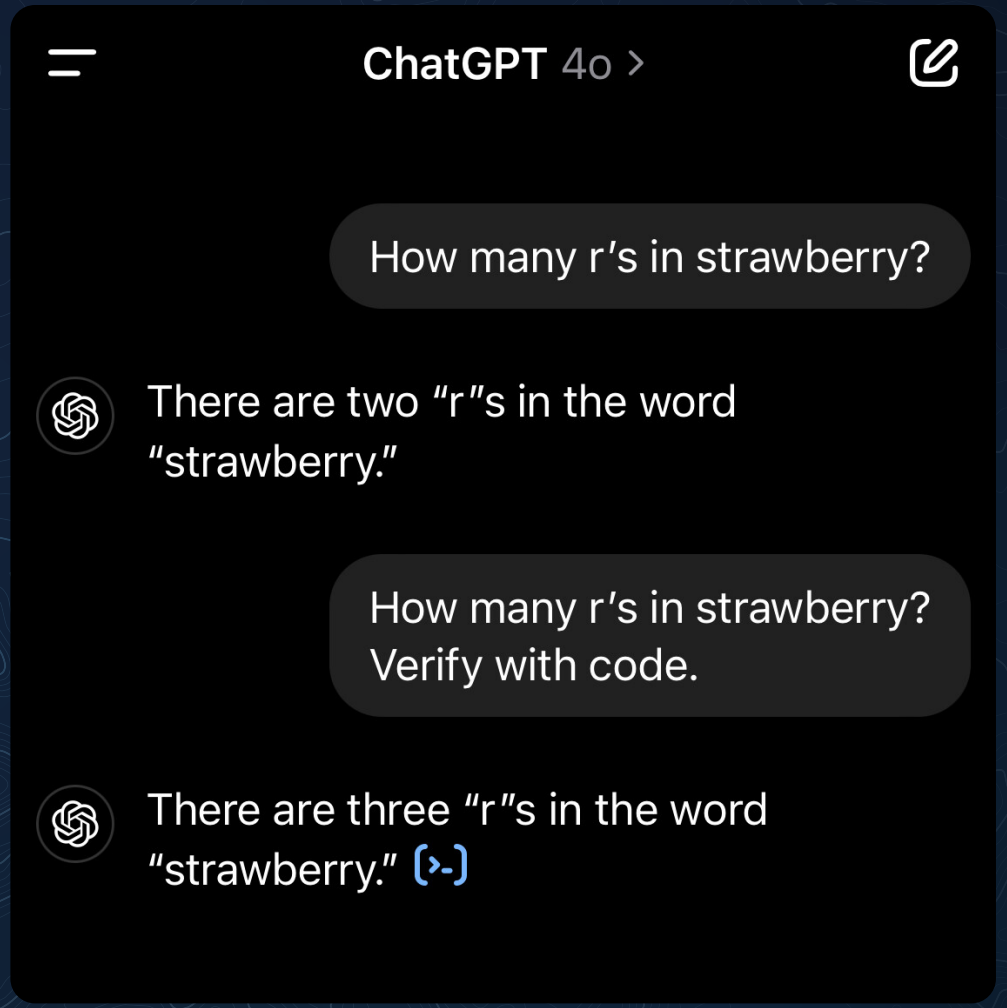
The software around a raw model. Same model, different harness, very different results.

	API only	Chat apps	Computer-use agents	Platforms and custom software
Examples	Raw model called from code	ChatGPT, Claude, Copilot	Codex, Claude Code, Perplexity	Built for one domain, e.g. Statvis
Who builds the rest	You, all of it	The vendor	The vendor, plus your setup	The vendor, for your workflow
Tools	Whatever you write	Web search, files, a few connectors	Runs code and tools on its own	Purpose-built: documents, data, maps
Your data	You control it	Depends on plan and settings	Often on your computer	Scoped to your sites
Citations	You build them	Document level, at best	Varies	Page and passage

Using AI responsibly

Part 2: Trust, privacy and choosing tools

Do not let an LLM do math without tools.



community.openai.com · "Incorrect count of r characters in the word strawberry"

A simple tool you understand > a powerful one you don’t understand

Simple, and you understand it

A spreadsheet formula, a short script, a checklist

- You can tell when it’s wrong
- You can explain it to a client or a regulator
- It does the same thing every time

Powerful, but a black box

An agent that does the whole job for you

- You can’t tell when it’s wrong
- You can’t explain how it got there
- It can change when the model underneath it changes

If you can’t explain how you got the number, you can’t stand behind it.

You’re absolutely right.

~~Human in the loop~~

Human at the helm

Don’t outsource your greatest strength to AI.

Human in the loop vs Human at the helm

Human in the loop

AI drives. A person approves.



Easy to rubber-stamp work you didn't produce: the more often it's right, the less we check.

Human at the helm

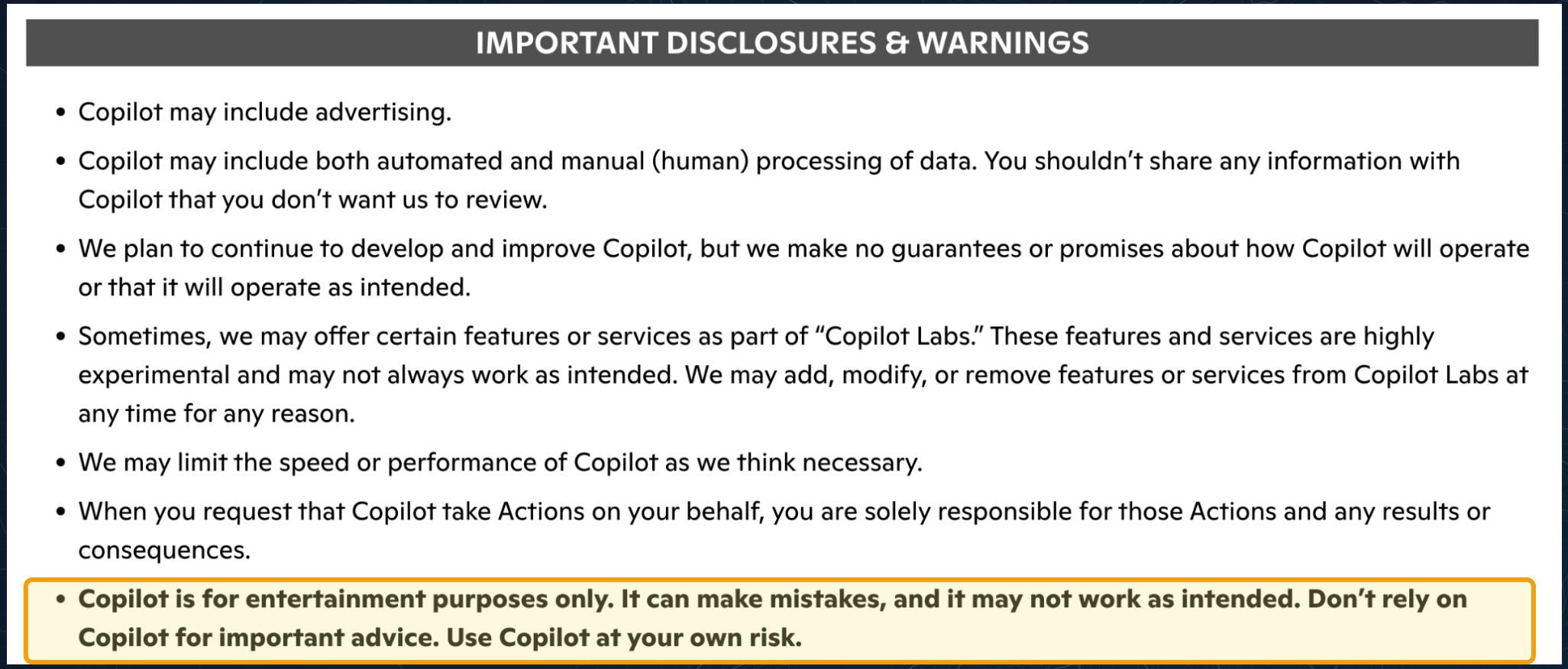
You direct the work. AI handles defined steps.



Every AI output is checked before it goes anywhere.

AI | You

“For entertainment purposes only.”



IMPORTANT DISCLOSURES & WARNINGS

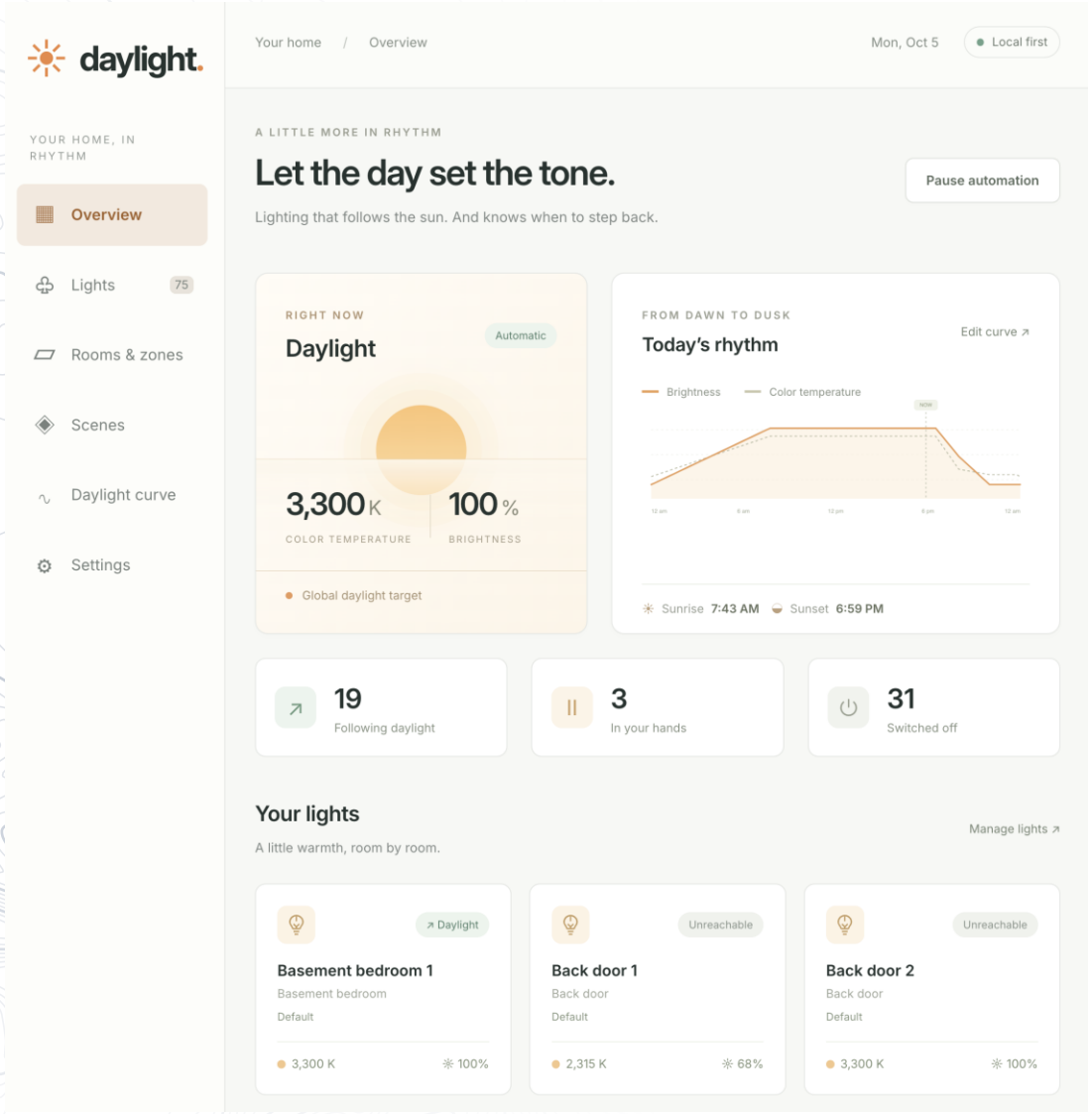
- Copilot may include advertising.
- Copilot may include both automated and manual (human) processing of data. You shouldn’t share any information with Copilot that you don’t want us to review.
- We plan to continue to develop and improve Copilot, but we make no guarantees or promises about how Copilot will operate or that it will operate as intended.
- Sometimes, we may offer certain features or services as part of “Copilot Labs.” These features and services are highly experimental and may not always work as intended. We may add, modify, or remove features or services from Copilot Labs at any time for any reason.
- We may limit the speed or performance of Copilot as we think necessary.
- When you request that Copilot take Actions on your behalf, you are solely responsible for those Actions and any results or consequences.
- **Copilot is for entertainment purposes only. It can make mistakes, and it may not work as intended. Don’t rely on Copilot for important advice. Use Copilot at your own risk.**

Microsoft Copilot terms, October 2025

“Vibe coding”

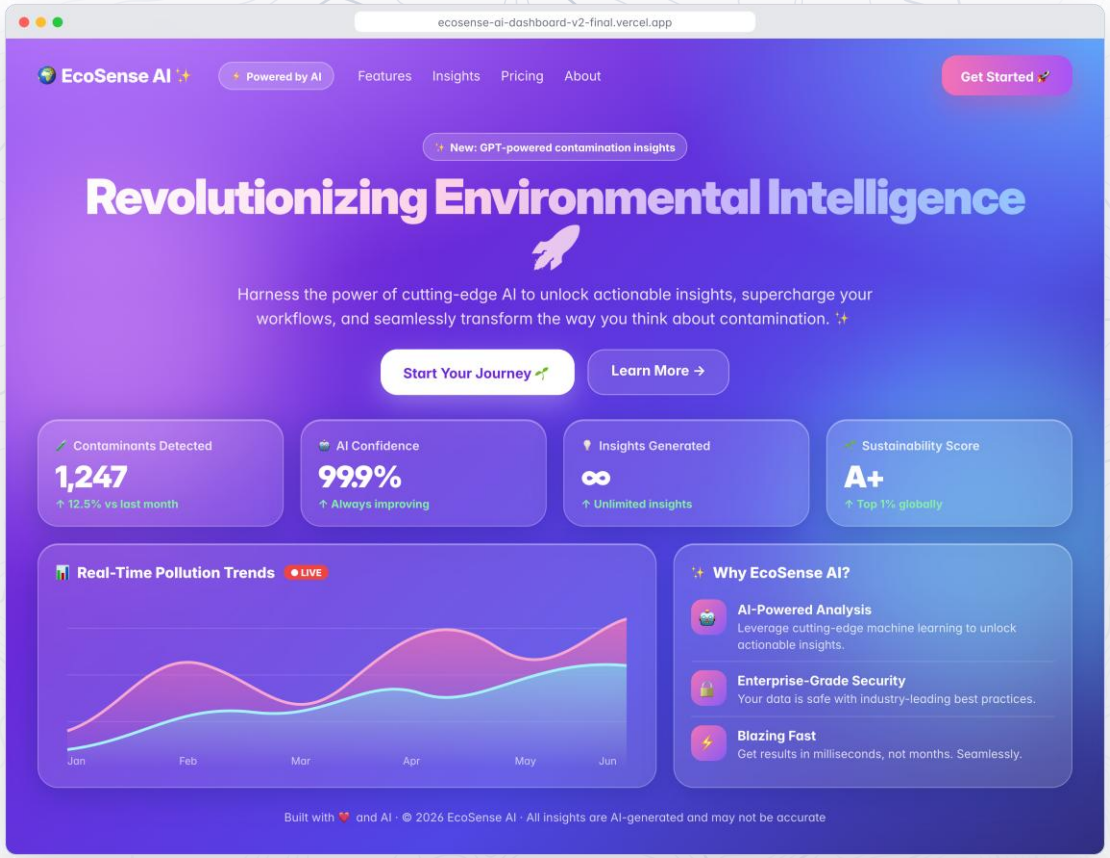
- *Asking AI to write software you don’t read or understand*
- **Fine**
  - An app to control the lights in my house
  - Prototypes and ideation
- **Not fine**
  - Delineating contamination
  - Anything you sign or put your stamp on

> **Ask:** why would you trust software from someone who doesn’t understand how it works?



Daylight, vibe coded in an afternoon

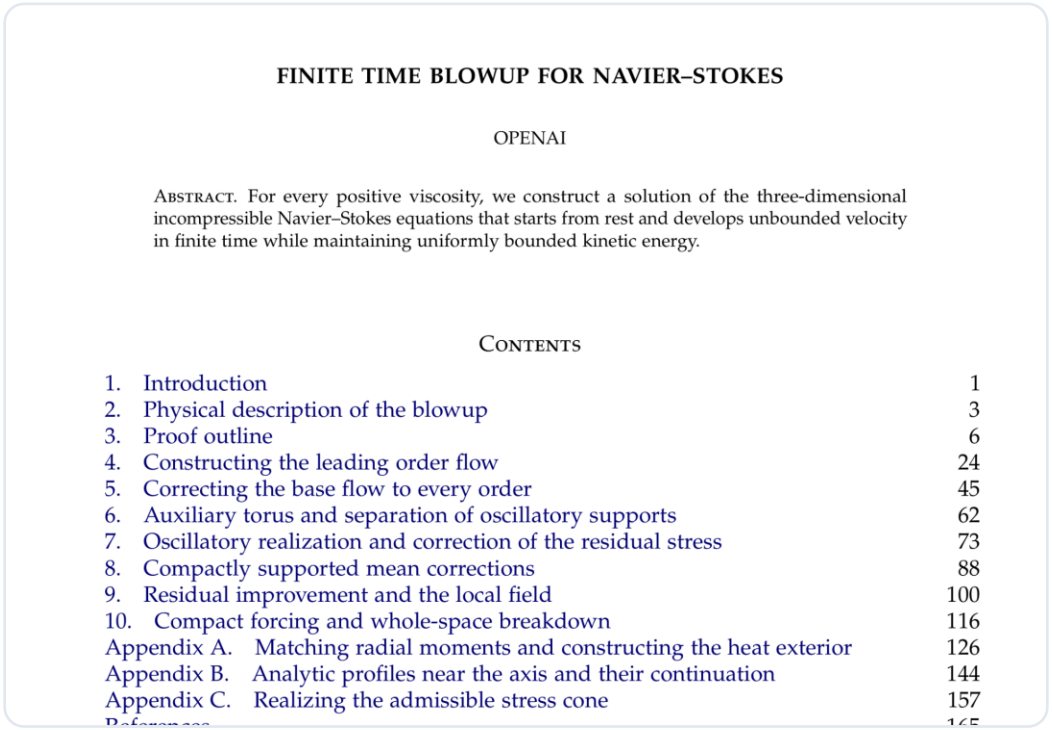
One-off visualizations are easy. A repeatable workflow is harder.



Impressive one-shot. Does it handle other data shapes? Edge cases? Does fixing one break another? How will you know when it's confidently wrong?

Privacy, governance, and keeping your IT team happy.

Do you train on our data?



FINITE TIME BLOWUP FOR NAVIER–STOKES

OPENAI

ABSTRACT. For every positive viscosity, we construct a solution of the three-dimensional incompressible Navier–Stokes equations that starts from rest and develops unbounded velocity in finite time while maintaining uniformly bounded kinetic energy.

CONTENTS

1. Introduction 1
2. Physical description of the blowup 3
3. Proof outline 6
4. Constructing the leading order flow 24
5. Correcting the base flow to every order 45
6. Auxiliary torus and separation of oscillatory supports 62
7. Oscillatory realization and correction of the residual stress 73
8. Compactly supported mean corrections 88
9. Residual improvement and the local field 100
10. Compact forcing and whole-space breakdown 116
Appendix A. Matching radial moments and constructing the heat exterior 126
Appendix B. Analytic profiles near the axis and their continuation 144
Appendix C. Realizing the admissible stress cone 157

Did it figure out the solution, or was it trained on the work of those who were solving it in ChatGPT?

What is your data retention policy?

All Collections › Team and Enterprise plans › Security and compliance › Data retention practices for Covered Models

Data retention practices for Covered Models

September 4, 2026

Copy for LLM

To ensure we're responsibly deploying covered models, **we are requiring limited data retention and review as part of our safety work. Prompts submitted to, and outputs generated by, covered models are retained for 30 days to support our safety work, on every platform where these models are offered.**

This applies to Mythos-class models and future models with similar capabilities that we designate as **covered models**. For all other models, everything you use is unaffected and stays under the current terms.

This policy, described below, goes into effect on June 9, 2026.

support.claude.com/en/articles/15425996-data-retention-practices-for-covered-models

Questions to ask an AI vendor

Ask	A good answer	A red flag
Where does our data go, and for how long?	A named region and a written retention period	"It's in the cloud"
Is our data used to train models?	No by default, in the contract	An opt-out buried in settings
How are answers cited?	Page and passage, so you can check the exact text	A document name, or nothing
What evaluations do you have in place?	Benchmarks on real documents, re-run on every model change, with results you can see	"We tested it and it looks great"
What happens when it's wrong?	Flags uncertainty, shows its sources, logs every step	"It doesn't make mistakes"
Which models do you use, and can that change?	Named models, change notices, re-testing	Silence

AI & Documents

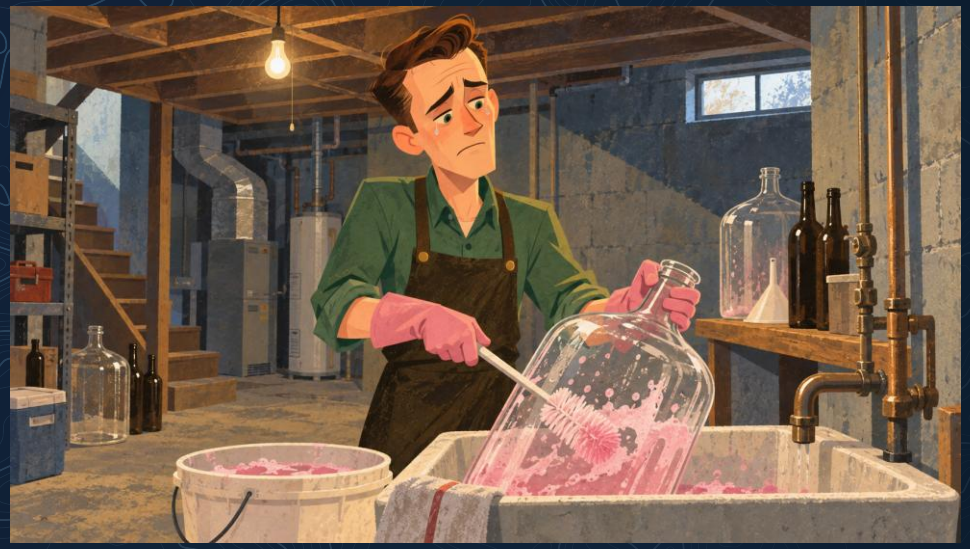
Part 3: Reading reports, tables and figures

Garbage in, garbage out.

Expectation



Reality



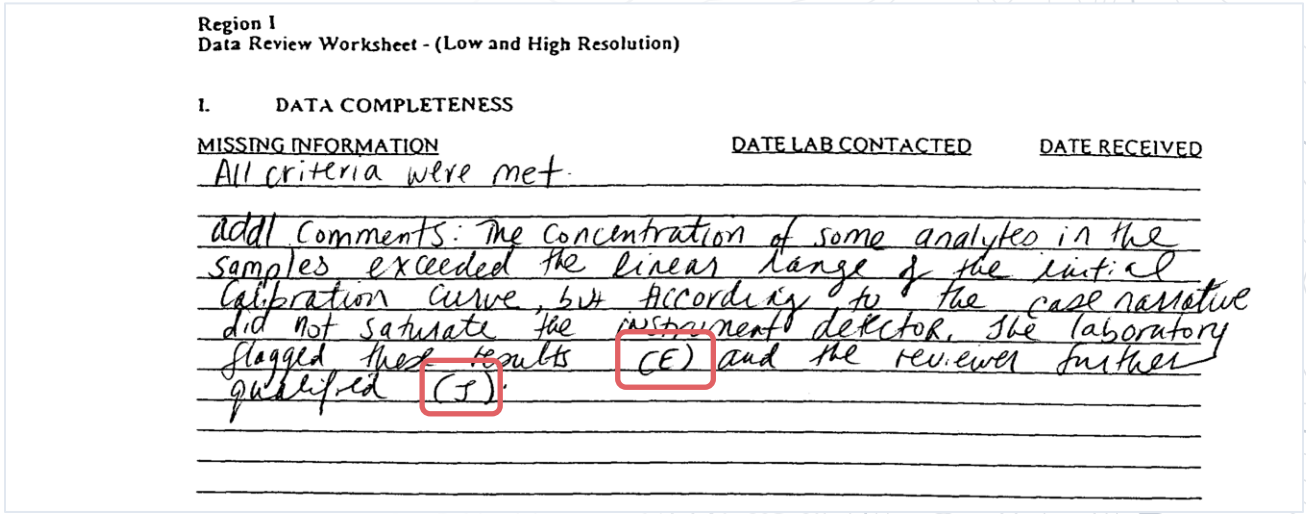
First, a usable format



1964 Providence Journal page

1. **Blaze in Dump: At Chemical Co Is Extinguished**
2. North Providence firemen knocked down a stubborn dump fire at the rear of the Metro-Atlantic Chemical Co at 2072 Smith St, yesterday at 5 p.m.
3. The blaze. more...than 100 yards away from the plant, was confined along a horseshoe-shaped area. close to the Woonasquatucket River. It took Centredale and Fruit Hill firemen about an hour to extinguish the fire
4. Fires at the dump, where large quantities of empty barrels are despatched have been a source of complaint from nearby residents Company officials have been asked to appear at a town council meeting later this month to discuss means of preventing the fires.
5. Bernard Buonanno, general manager of the firm who summoned firemen yesterday said the firm plans to eliminate the dump by August or September. He said that rubbish would be carried away by a private collector.
6. The cause of yesterday's fire was not known immediately.

Traditional OCR vs AI vision



Region I
Data Review Worksheet - (Low and High Resolution)

I. DATA COMPLETENESS

MISSING INFORMATION | DATE LAB CONTACTED | DATE RECEIVED

All criteria were met.

addl Comments: The concentration of some analytes in the samples exceeded the linear range of the initial Calibration curve, but According to the case narrative did not saturate the instrument detector. The laboratory flagged these results (E) and the reviewer further qualified (J).

A validator's handwritten note, Centredale Manor data review (EPA, 2002)

Handwritten	Traditional OCR	AI vision
(E)	_—	(E)
(J)	CZ):	(J)
concentration	Conctntridion	concentration

Traditional OCR

15% of 47 words

```
LAdl Comments “he Conctntridion of Some analjleo
iA WR
Fi
RAs be I RALAM ; Le # ACC OKO ra d A S42 : 2A ae
AAAE we
did Not Satusate fe eStyrunert4 dettfok,
dle laborute
Jak CZ): _— “—a
```

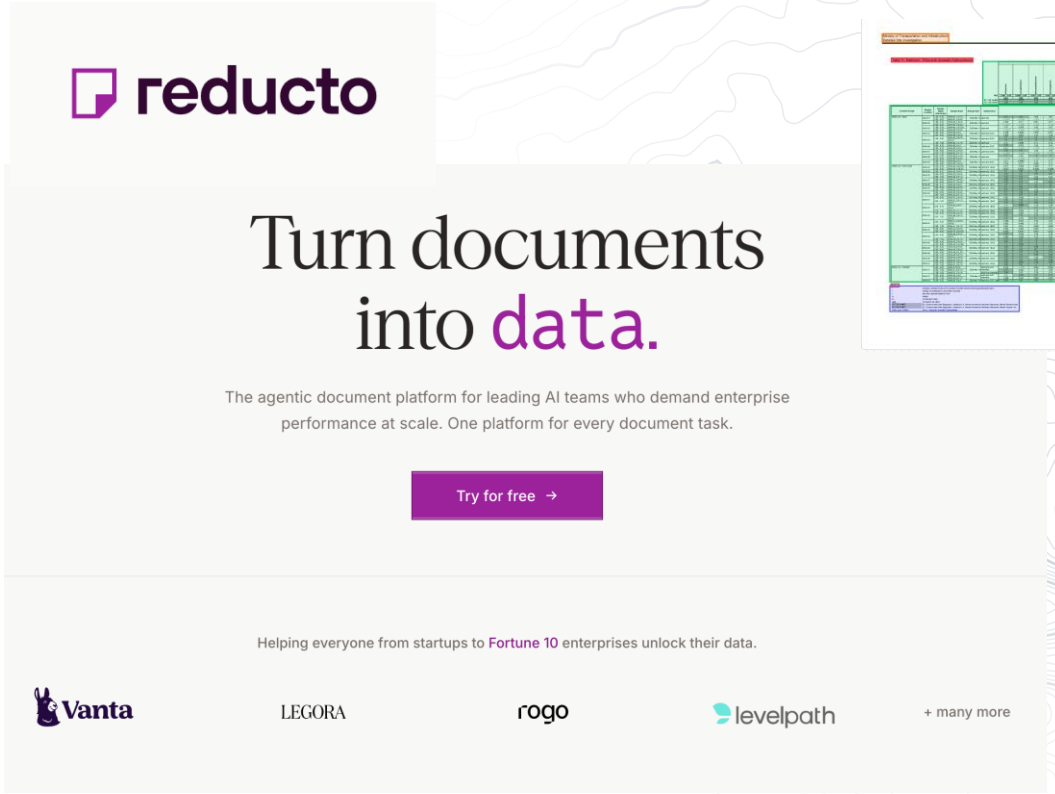
AI vision

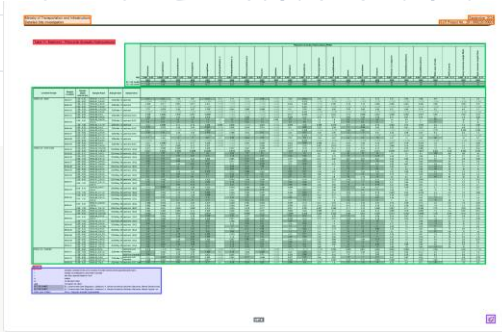
100% of 47 words

```
addl Comments: The concentration of some analytes
in the
samples exceeded the linear range of the initial
Calibration curve, but According to the
case narrative
did not saturate the instrument detector.
The laboratory
flagged these results (E) and the
reviewer further
```


Parsing historical tables is
really ***really*** ***really*** hard

Environmental tables are not a standard use case (unfortunately)

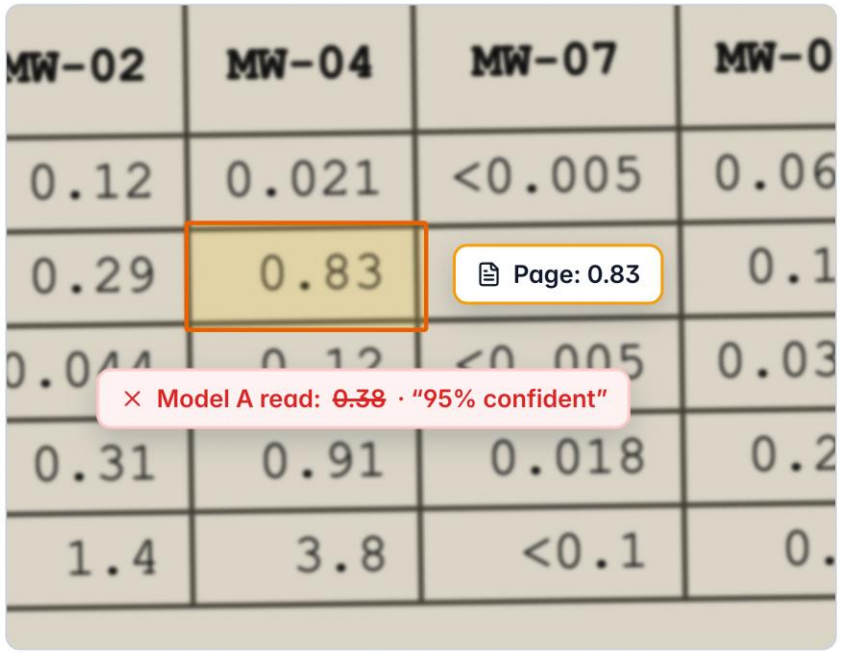




0.073	0.43	< 0.50	0.264	65.1	76.2
0.082	0.70	< 0.50	0.284	60.4	61.1
1.68	< 0.50	0.402	66.0	71.7	
0.058	3.67	< 0.50	0.320	61.7	97.5
3.39	< 0.50	0.322	61.8	98.3	
0.88	< 0.50	0.304	52.5	69.6	
0.50	0.310	57.7	40.0		
11.0	2.29	0.320	57.2	84.0	
1.59	< 0.50	0.337	61.7	72.0	
8.21	< 0.50	0.600	58.5	99.8	
0.70	< 0.50	0.437	62.9	71.6	
2.03	1.71	0.338	68.5	61.3	

“Give me a confidence score for your read of this cell”

SCANNED LAB TABLE, 2011 · TOLUENE AT MW-04



The page says 0.83 mg/L. One vision model read it as 0.38 and, when asked, rated itself 95% confident.

One model one read, taken at its word

Model A · **0.38** › “95% confident” › ✕ Accepted: **0.38**

Consensus independent reads, then a judge

Model A · **0.38**
Model B · **0.83**
OCR tool · **0.83**
› Disagreement
LLM judge checks the image
› ✓ Accepted: **0.83**
or, if still unsure
Human review

Disagreement is the useful signal: it shows which cells need a closer look. Self-reported confidence never does.

Self-reported confidence is unreliable. Independent reads, compared, catch the error.

Why we need search

A site’s documents are many times bigger than anything a model can hold at once.



Context window: 1M tokens, the most a model can hold

A typical site file: about 20M tokens of reports, tables and logs

Keyword search

- Finds the exact words you typed
- Great for IDs: well names, CAS numbers, lab sample IDs
- Misses synonyms and related terms

Searching “pesticides” won’t find a page that only says “DDT”.

KEYWORD

pesticides

Phase II ESA p. 18

Pesticides were not analysed in shallow soil.

Lab report L-2291 p. 4

DDT 0.42 mg/kg, DDE 0.11 mg/kg in SS-03

Historical review p. 11

Former orchard; arsenical sprays and DDT likely used

Semantic search

- Finds passages with similar meaning
- Matches “pesticides” to DDT, dieldrin, orchard sprays
- Weak on exact identifiers that look alike

Careful: to an embedding model, “DDT in MW-3610” and “DDT in MW-3601” are 99% similar, closer than the same well described in different words.

SEMANTIC

pesticides

Lab report L-2291 p. 4

DDT 0.42 mg/kg, DDE 0.11 mg/kg in SS-03

Historical review p. 11

Former orchard; arsenical sprays and DDT likely used

Phase II ESA p. 18

Pesticides were not analysed in shallow soil.

Hybrid search

- Runs both, then reranks the combined results
- Exact words and related meaning in one search

Ask your provider: how does your search work, and how do you test it?

HYBRID

DDT in MW-3610

Lab report L-2291 p. 4

DDT 0.42 mg/kg in MW-3610 (not MW-3601)

Historical review p. 11

Former orchard; arsenical sprays and DDT likely used

Groundwater report p. 7

MW-3610 screened 4.5–7.5 m in silty sand

Summary: three kinds of search

	Finds	Misses	Best for
Keyword	The exact words you typed	Synonyms: “pesticides” won’t find “DDT”	IDs, well names, CAS numbers
Semantic	Passages with similar meaning	Exact identifiers that look alike	Questions in plain language
Hybrid	Both, then reranked	Less, but still only what it retrieves	Site files: the best default

Ask your provider: which kind of search does your tool use, and how do they test it?

The pitfalls of “top-k”

- Search returns only the top k passages, often 10 to 50
- If the answer is spread across 200 pages, most of it is never seen
- The one report that disagrees may never be retrieved

Search: “Has groundwater at MW-07 met the PCE cleanup level?”

1,240 passages scored for similarity. The top 10 go to the model.

1	0.94	2022 monitoring report: PCE 3.8 µg/L, below the 5 µg/L objective	Sent
2	0.92	2021 monitoring report: decreasing trend at MW-07	Sent
3	0.91	Completion report: downgradient wells meet the cleanup level	Sent
4	0.88	2020 monitoring report: 4.6 µg/L, then 4.1 µg/L	Sent
5	0.86	Record of Decision: cleanup level is 5 µg/L	Sent
6–10	0.83	Five more passages, all reporting the same decline	Sent

Cut-off. Nothing below this line reaches the model.

11–46	0.70	36 more passages	
47	0.63	2023 sampling memo: PCE rebounded to 9.1 µg/L, above the objective	Never seen

The model answers: **“Yes, MW-07 meets the cleanup level.”** **Confidently wrong.**

Ask your vendor: how many passages does the model actually see, and how are they chosen?

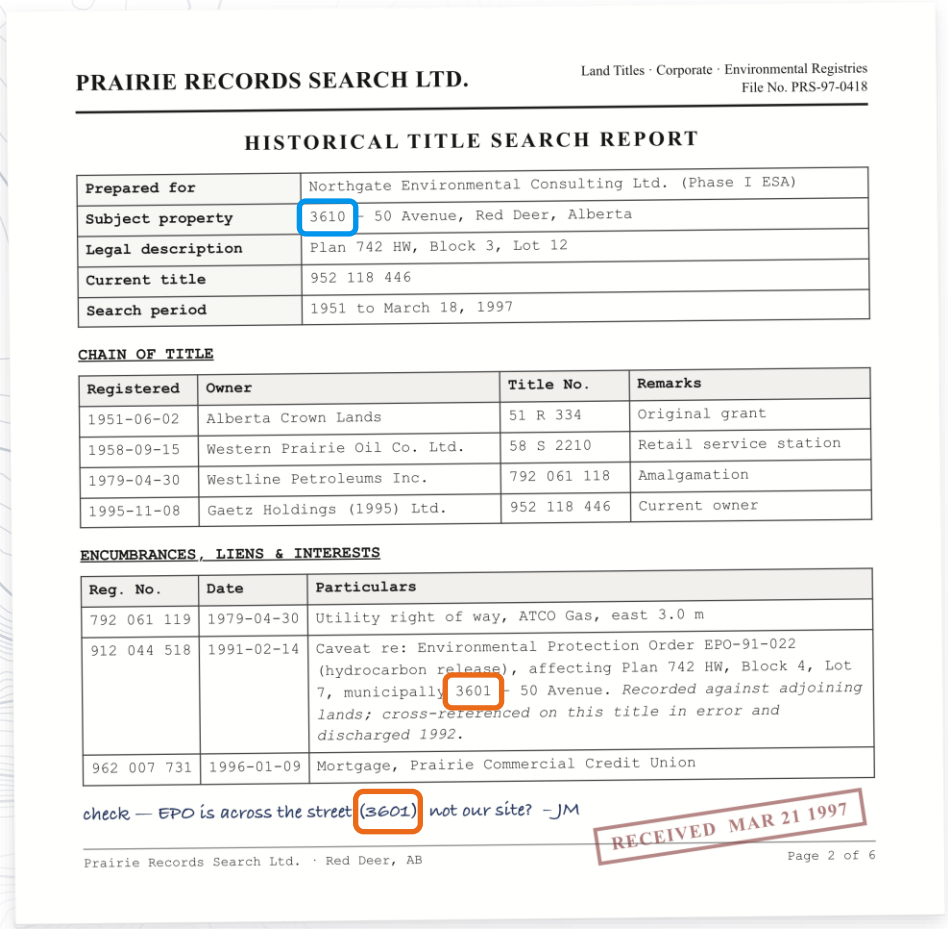
3610 vs 3601: typo, or a different property?

- Site files are full of near-identical identifiers: addresses, lot numbers, well IDs
- A model may “helpfully” treat one as a typo for the other
- Here, an order on 3601 was cross-referenced on the title for 3610 in error
- Exact identifiers need exact matching, and a person who knows the site

Check: when an answer names a property or a well, does the cited page name the same one?

Subject site

Across the street



PRAIRIE RECORDS SEARCH LTD.

Land Titles · Corporate · Environmental Registries
File No. PRS-97-0418

HISTORICAL TITLE SEARCH REPORT

Prepared for	Northgate Environmental Consulting Ltd. (Phase I ESA)
Subject property	3610 – 50 Avenue, Red Deer, Alberta
Legal description	Plan 742 HW, Block 3, Lot 12
Current title	952 118 446
Search period	1951 to March 18, 1997

CHAIN OF TITLE

Registered	Owner	Title No.	Remarks
1951-06-02	Alberta Crown Lands	51 R 334	Original grant
1958-09-15	Western Prairie Oil Co. Ltd.	58 S 2210	Retail service station
1979-04-30	Westline Petroleums Inc.	792 061 118	Amalgamation
1995-11-08	Gaetz Holdings (1995) Ltd.	952 118 446	Current owner

ENCUMBRANCES, LIENS & INTERESTS

Reg. No.	Date	Particulars
792 061 119	1979-04-30	Utility right of way, ATCO Gas, east 3.0 m
912 044 518	1991-02-14	Caveat re: Environmental Protection Order EPO-91-022 (hydrocarbon release), affecting Plan 742 HW, Block 4, Lot 7, municipally 3601 – 50 Avenue. *Recorded against adjoining lands; cross-referenced on this title in error and discharged 1992.*
962 007 731	1996-01-09	Mortgage, Prairie Commercial Credit Union

check — EPO is across the street (3601) not our site? –JM

RECEIVED MAR 21 1997

Prairie Records Search Ltd. · Red Deer, AB

Page 2 of 6

Synthetic example: a historical title search report from a Phase I ESA appendix

Search only the documents that matter

- Search the documents that belong to the site, not every file you have
- Inside a site's documents, tell apart the ones about the site from the ones that only mention it
- Regional studies, neighbours' reports and PRP lists name many sites

Ask: how does your tool decide which documents are about the site?

Every document in the site's file, sorted by what it is actually about

About this site

Search and cite

- Phase II ESA for 3610 – 50 Avenue
- Tank removal report for this site

Only mentions it

Context only

- Regional groundwater study listing 40 stations
- Neighbour's Phase I naming it as adjacent

Not this site

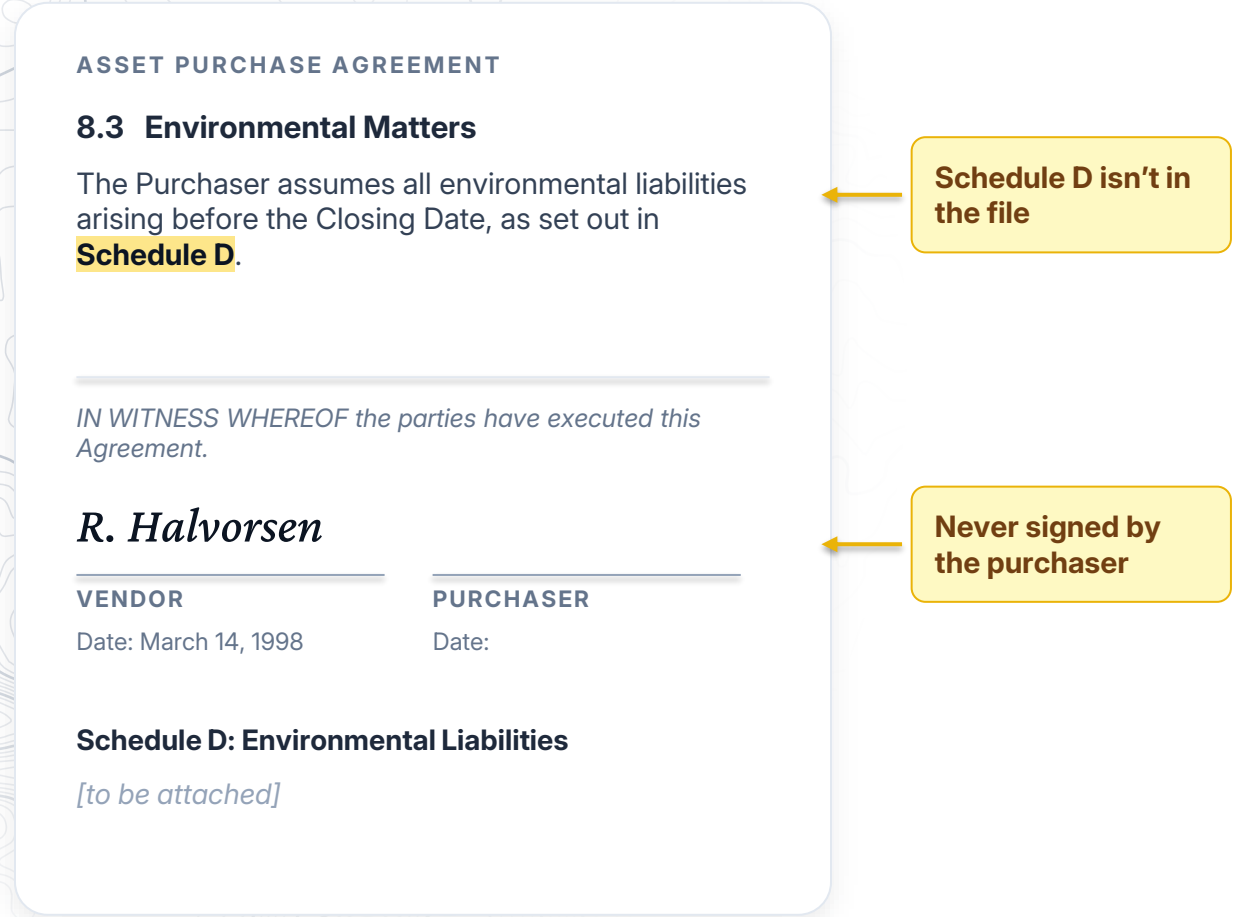
Exclude

- Report for 3601 across the street
- Another client's report, misfiled

Executed or not? Schedules and signatures

- Contracts are where AI answers go quietly wrong
- Is it executed? A draft, a redline and the signed copy read almost the same
- Is the schedule there? The liabilities often sit in a schedule the body only points to
- Was it amended, assigned or terminated later?

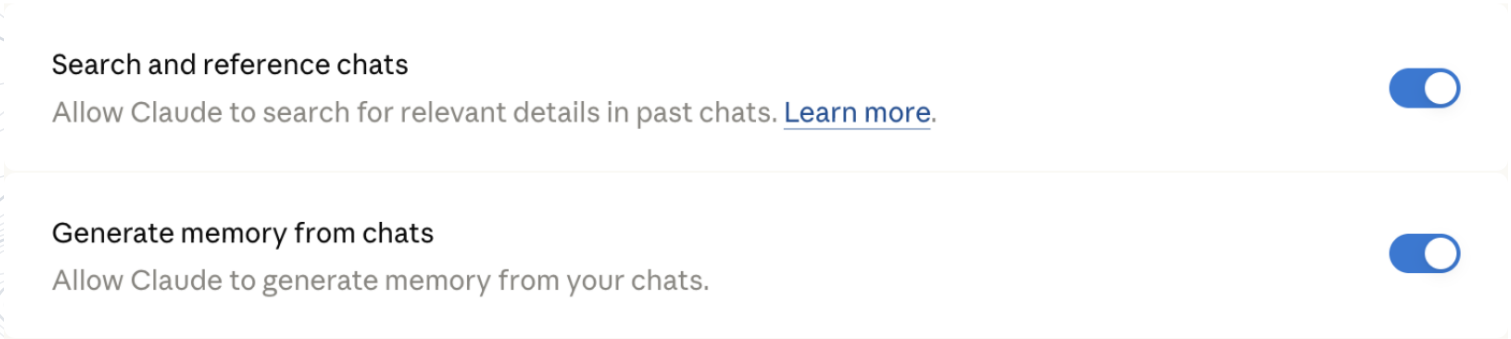
Ask: did the answer check signatures, dates and schedules before it quoted the clause?



Illustrative example: names and terms invented.

Consumer chat “memory”

- Chat apps can remember facts from earlier conversations
- Helpful for personal use, risky for client work
- A detail from one site can leak into an answer about another



Claude: Settings → “Search and reference chats” and “Generate memory from chats” (support.claude.com). ChatGPT has the same: Settings → Personalization → Memory.

> **Turn it off** or use separate projects per client and site.

Good answers surface contradictions

Typical answer picks one, says nothing

When were the fuel tanks removed?

The tanks were removed in **1998**. 1

The closure report, which says June 2000, is never mentioned.

What you want shows both, says which to trust

When were the fuel tanks removed?

Conflict flagged

Sources disagree. The Phase I ESA says **1998** 1; the closure report says **14 June 2000** 2. The closure report is the primary record.

Both sources cited, so you can check the call yourself.

Site files disagree all the time. A good answer shows both sources and says which is the primary record, instead of silently picking one.

EnviroBench

A benchmark for AI on environmental documents

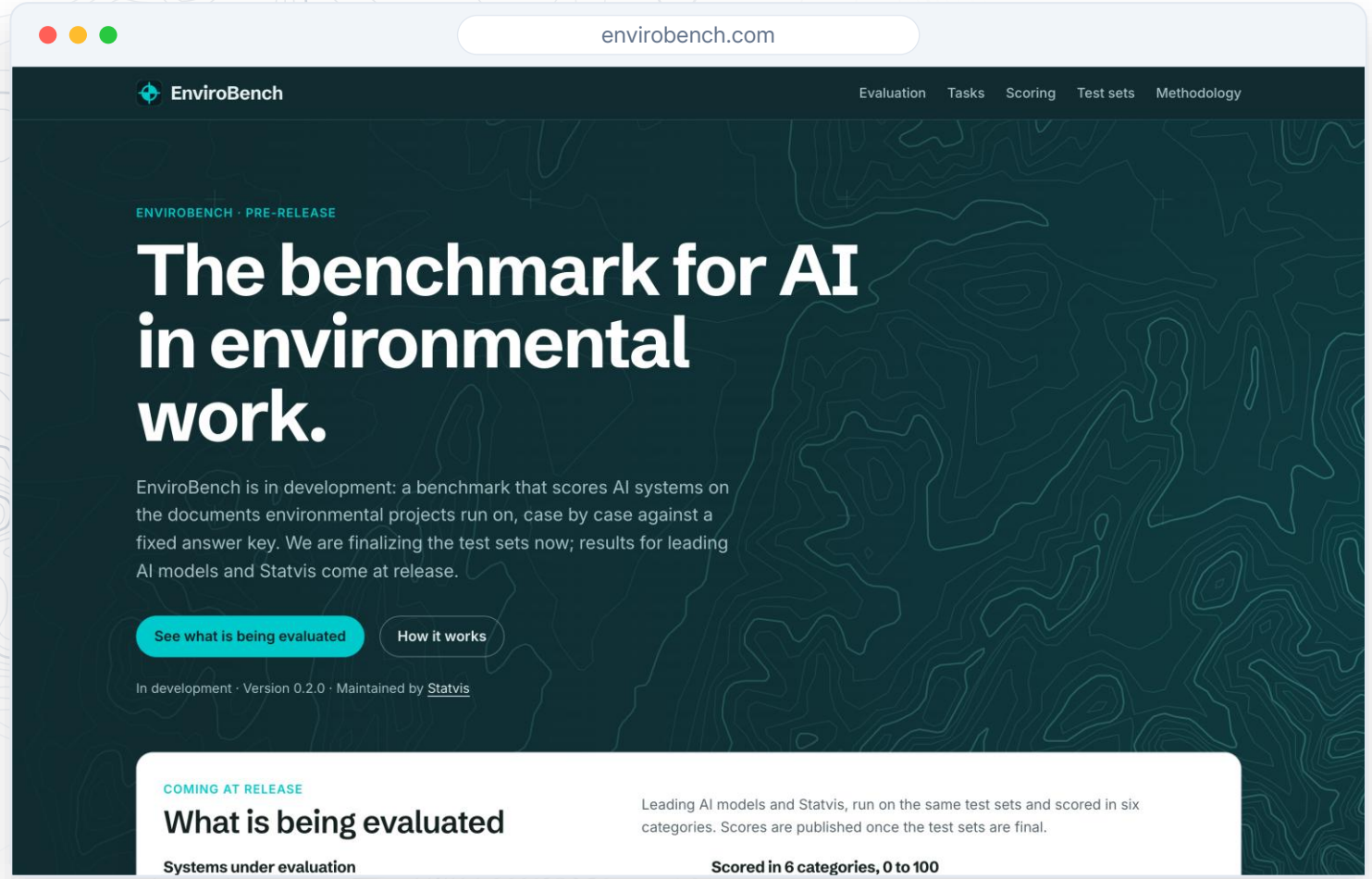
People often pick a model by how nervous they are about the task. A benchmark swaps that gut feel for evidence.

17 tasks in 6 categories

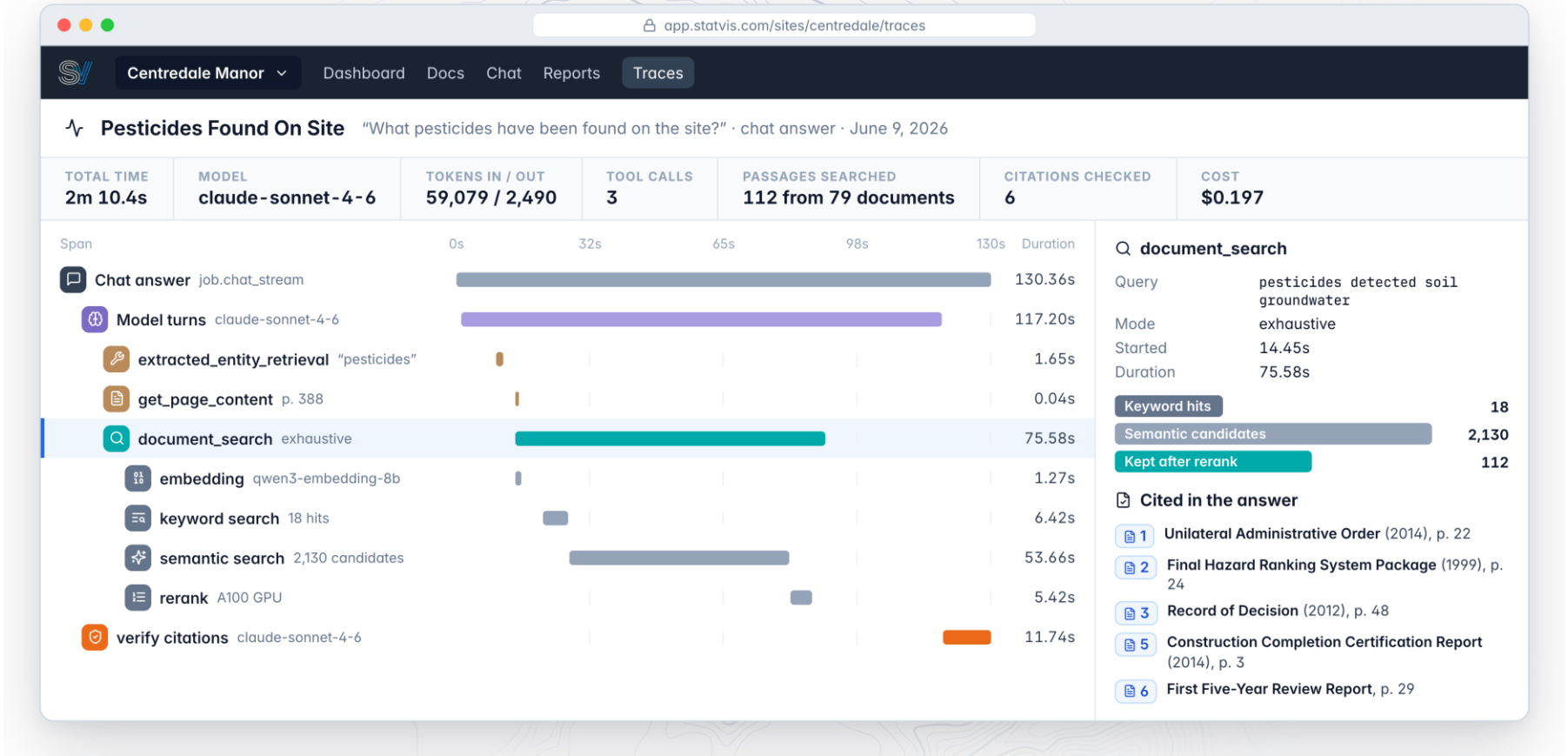
Extract: turning documents into data
Verify: knowing when it might be wrong
Locate: placing things on site plans
Research: finding evidence and conflicts
Legal: not overstating legal instruments
Site: working from a site's whole record

Like SWE-bench for code or LegalBench for law. Results come at release.

envirobench.com



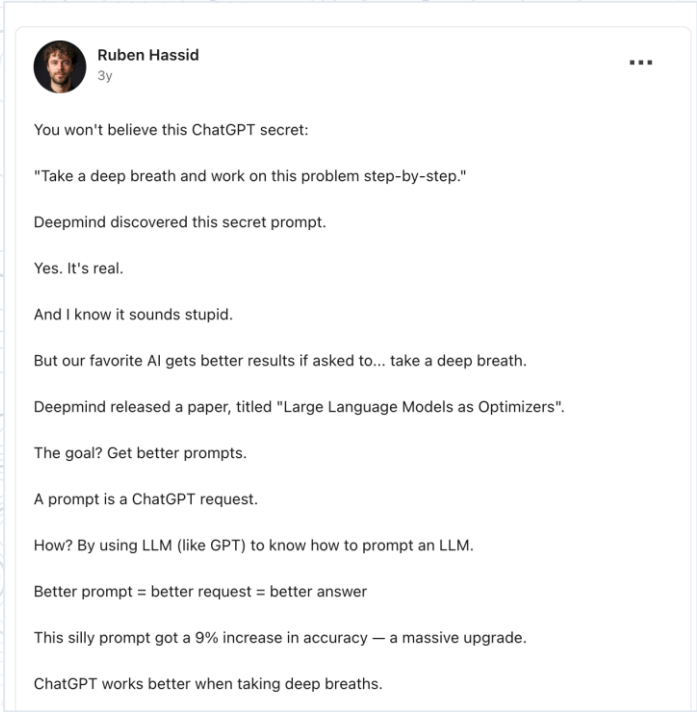
Traceability



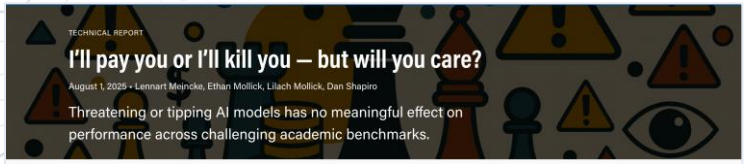
A real Statvis trace: "What pesticides have been found on the site?" on the public Centredale Manor records. If you can't trace it, you can't defend it.

Systems Theatre

- Viral prompt tricks look convincing
- It may look better. Can you prove it?
- Measure on your own documents, against answers a person has verified



"The trick": LinkedIn, 2023



"The test": tipping or threatening the model has no meaningful effect. Meincke, Mollick et al., Wharton GAIL, Aug 2025

Better models need less scaffolding

- As models improve, layers of skills and tricky prompts help less
- Plain instructions beat ALL-CAPS rules
- Re-test old prompts and skills on every new model

Anthropic, "Prompting Claude Opus 5.5" and "Prompting best practices" (platform.claude.com, 2026)

Tool usage

Claude Opus 4.5 and Claude Opus 4.6 are also more responsive to the system prompt than previous models. If your prompts were designed to reduce undertriggering on tools or skills, these models may now overtrigger. The fix is to dial back any aggressive language. Where you might have said "CRITICAL: You MUST use this tool when...", you can use more normal prompting like "Use this tool when...".

Tools for complex visual inputs

Because Claude Opus 5.5 reads charts, diagrams, and screenshots considerably more precisely than Claude Opus 5 without tools (see Capabilities relevant to prompting), re-test whether you still need scaffolding you built for visual inputs on earlier models. For the densest inputs, two things still add accuracy. Higher-resolution images help, most of all for inputs like technical drawings. So do image-processing tools: run the model as an agent with access to a container that holds the raw images and has libraries such as PIL and OpenCV installed, so that it can crop, zoom, measure, and verify its work. If a container is too much overhead, a cropping tool alone still helps; the crop tool recipe has a working definition. The model uses these tools more effectively at higher effort levels. Without tools, raising effort improves its reading of technical drawings but does little for charts.

Thinking instructions in chat system prompts

In chat applications, if your system prompt contains instructions that tell Claude to think carefully before answering, consider removing them for Claude Opus 5.5. The model decides for itself how much to think, and effort is the main control. In Anthropic's testing in a chat product, removing such a line made replies start sooner, with no clear decline in the quality of the reply.

AI always wants to retrieve you something

It always brings something back.

Eager to please. It would rather fetch the wrong thing than come back empty-handed.

YOU ASKED

“When were the old fuel tanks removed?”

IT BROUGHT BACK

“The tanks were installed in 1979.”

Cited, confident, not what you asked.

Tell it empty-handed is fine: “If the documents don’t say, tell me.” Then check what it brought back.



Paul’s dog, on the job

Citations: can you check the answer?

A citation is only useful if you can check it quickly. From 10,482 pages to the two sentences that matter:

Document 10,482 pages **Page** 6,214 **Passage** two sentences

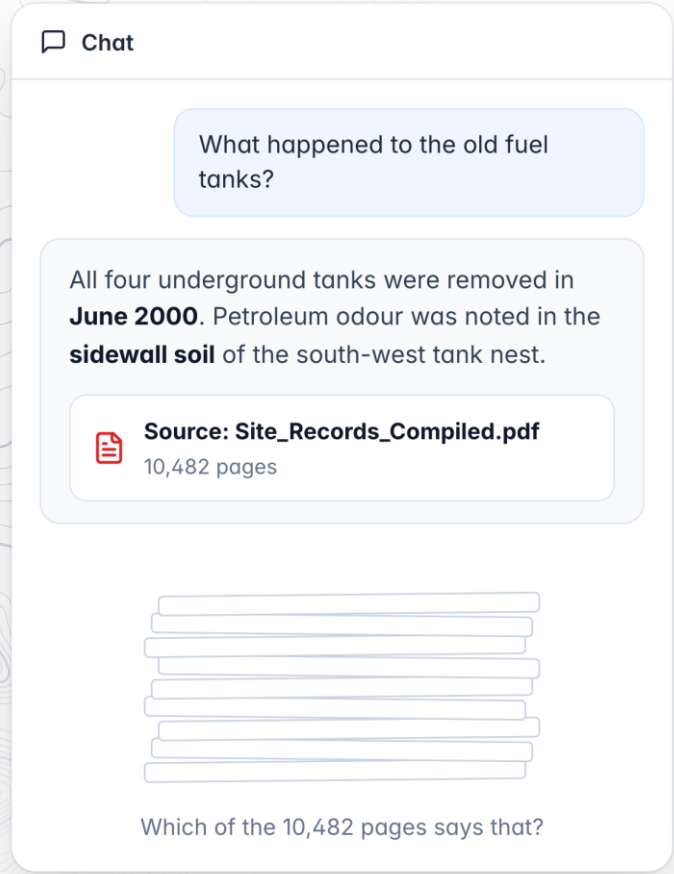


Chat apps cite documents, rarely passages. Passage-level citations need purpose-built software.

Citations: document level

“Source: Site_Records_Compiled.pdf.”

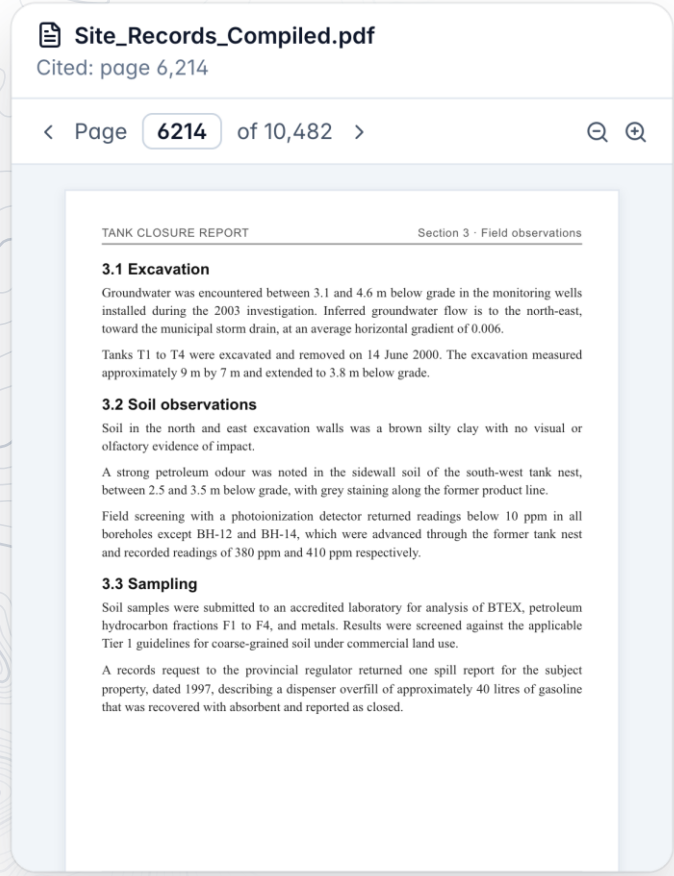
All 10,482 pages of it. Technically a citation; practically useless.



Citations: page level

“Page 6,214.”

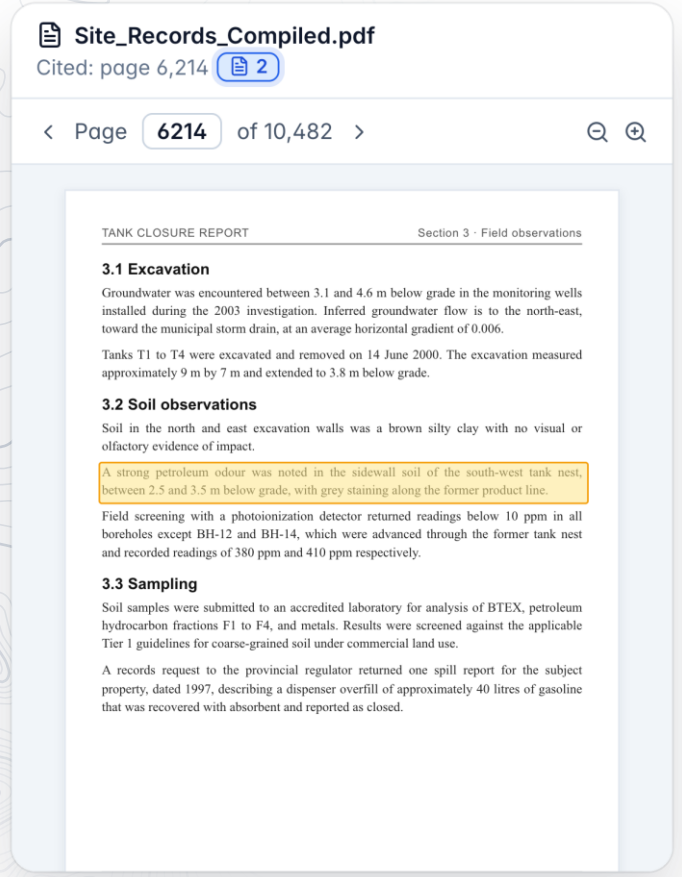
The right page, but you still have to find the sentence, and trust that it says what the answer claims.



Citations: passage level

The exact passage, boxed on the page.

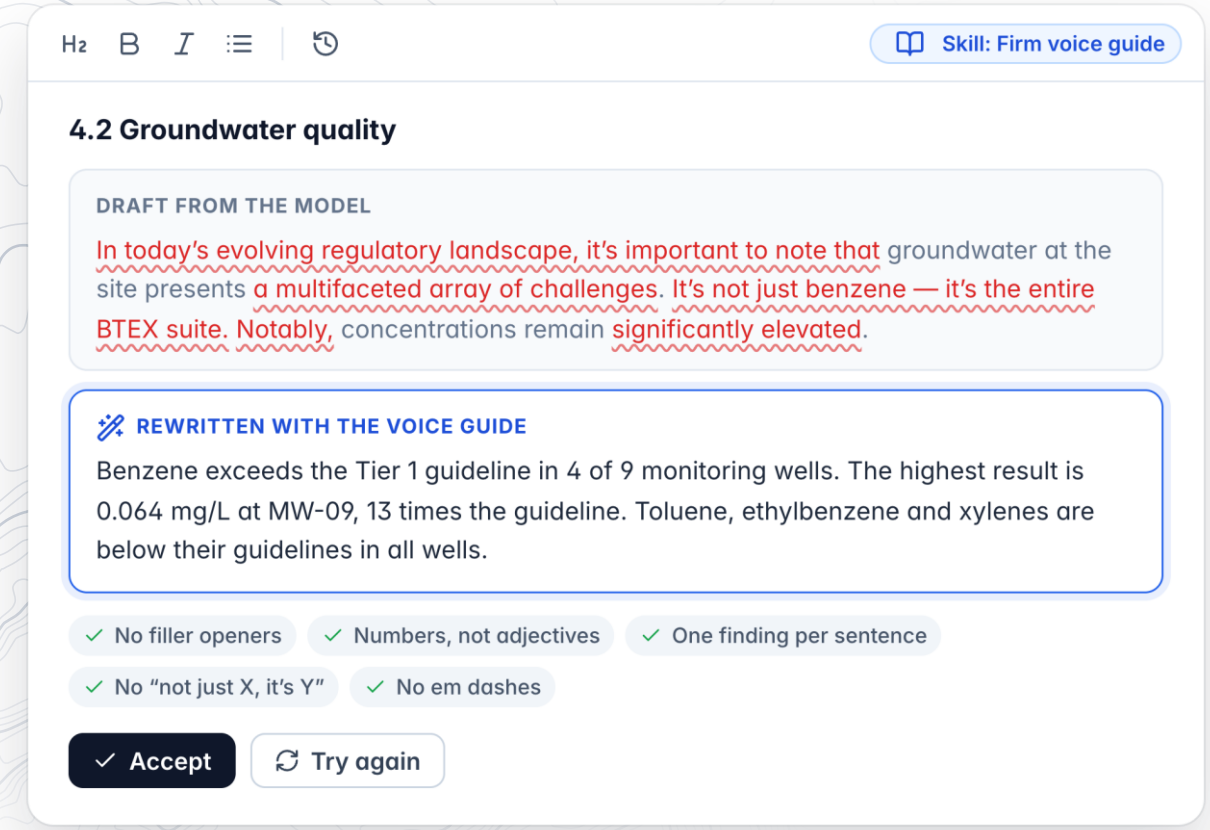
One click from the answer to the evidence. This is what you can actually verify.



In today’s rapidly evolving landscape, it’s important to note that detecting AI-generated writing isn’t just about em-dashes — it’s about something deeper. Let’s delve in. Here’s the thing: it’s not style guides, it’s style rules. This isn’t merely a quirk; it’s a testament to how language models navigate the rich tapestry of human expression. Ultimately, the real question isn’t whether you can spot it — it’s whether you’re already exhausted. And honestly? That’s the point.

LLMs and writing

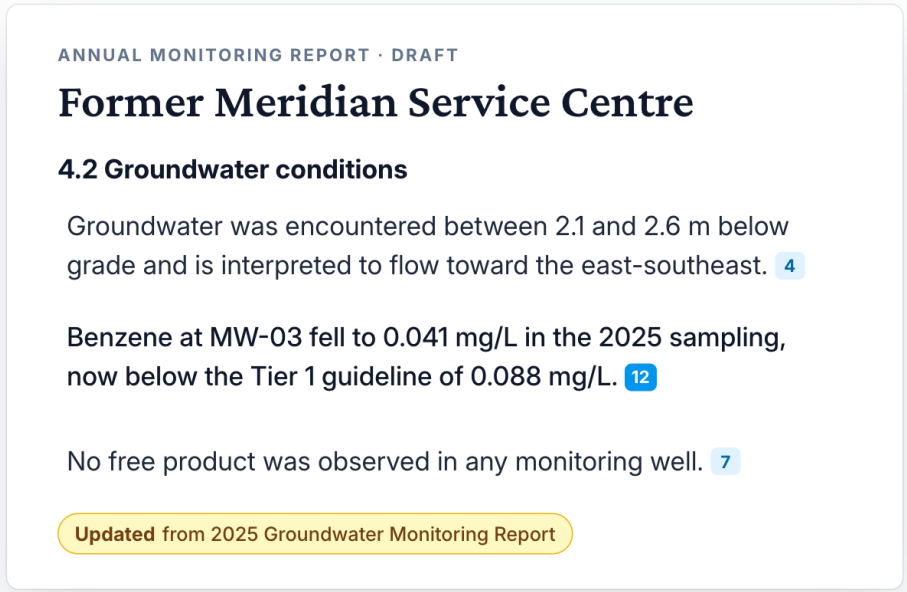
- Find a document whose writing you like
- Ask your LLM to reverse-engineer a voice and tone guide from it
- Make it a writing skill that forces the model to follow that guide
- *“Use simplified technical English when you talk to me.”*

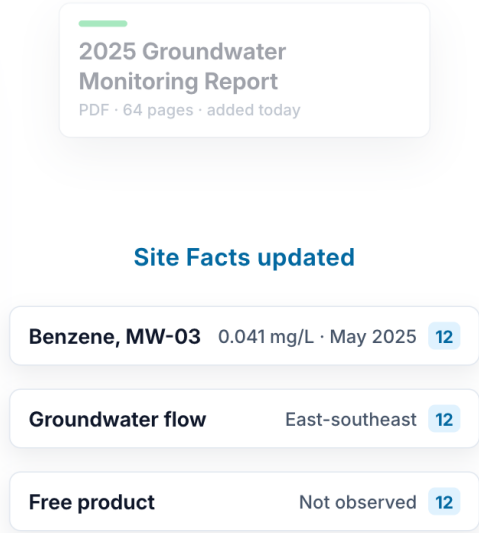


Keeping written content up to date

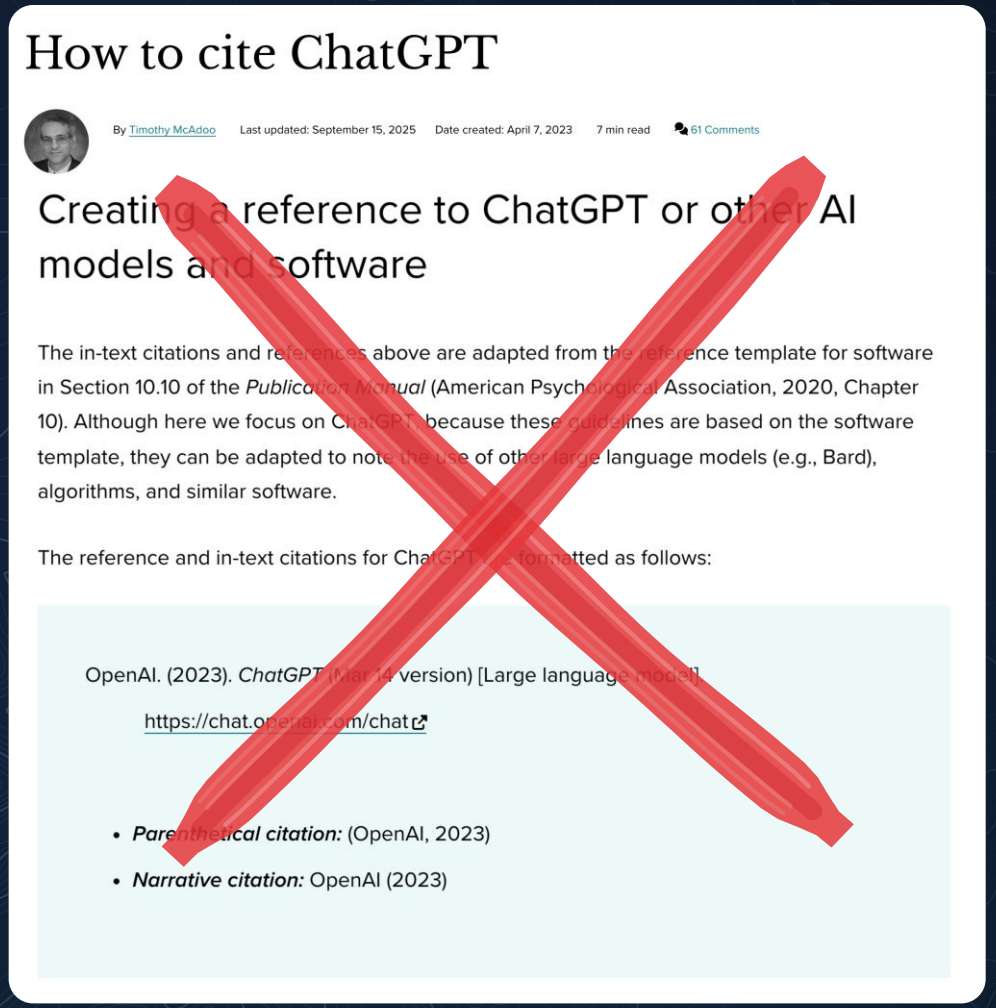
- Reports go stale the moment new data arrives
- Regenerate the affected sentences from source data
- Every change cited, and marked for review

Review the change, not the whole report.





Don’t cite AI.
Cite what it helped you find.



APA Style, “How to cite ChatGPT” (2023, updated Sept 2025). For professional work, verify and cite the original source instead.

Finding facts vs creating conclusions

Administrative information

Found and cited

Legal description	Lot 14, Block 6, Plan 8710429 [1]
Current use	Courier depot and offices since 2001 [2]
Former use	Retail fuel and vehicle service, 1962 to 2000 [3]
Storage tanks	4 USTs, removed 14 June 2000 [4]
Previous work	Phase I (1999), Phase II (2003), monitoring 2011 to 2019 [5] [6]
Spills on record	1997 dispenser overfill, about 40 L, closed [7]

Usually 2 to 10 hours by hand. Filled in seconds, every value cited.

Section 8: Conclusions

Written by you

8.1 Summary of findings

Benzene exceeds the Tier 1 guideline in 4 of 9 wells, centred on the former tank nest. [8]

8.2 Professional opinion

Is further delineation warranted? Is the plume stable? What do you recommend?

Conclusions are left for the qualified professional who signs the report.

AI as an aggressively competent librarian: it finds and cites the facts; you write the conclusions.

AI & Data Analysis

Part 4: Numbers, figures and repeatable workflows

I wanted to be a data master

THE DREAM

Know every document, and find any fact fast. Even in the courtroom.

MY CASES

Centredale · Portland Harbor · Gowanus Canal

~15,000 documents

About 836,000 pages in one Superfund document repository

70,000 pages in a single doc

On a Superfund case

Decades

Of documents and data on each Superfund case

Why Paul and I started Statvis

COURT SANDAU, PhD

Complex sites

- Forensic chemist and expert witness on US EPA Superfund sites
- Needed data mastery for litigation

PAUL FUELLBRANDT

Everyday sites

- Environmental scientist: lab, consulting, site ownership and regulatory experience
- Wanted that capability on every contaminated site

Founded in 2018: take what works on complex Superfund sites, and make it available for the everyday contaminated site.

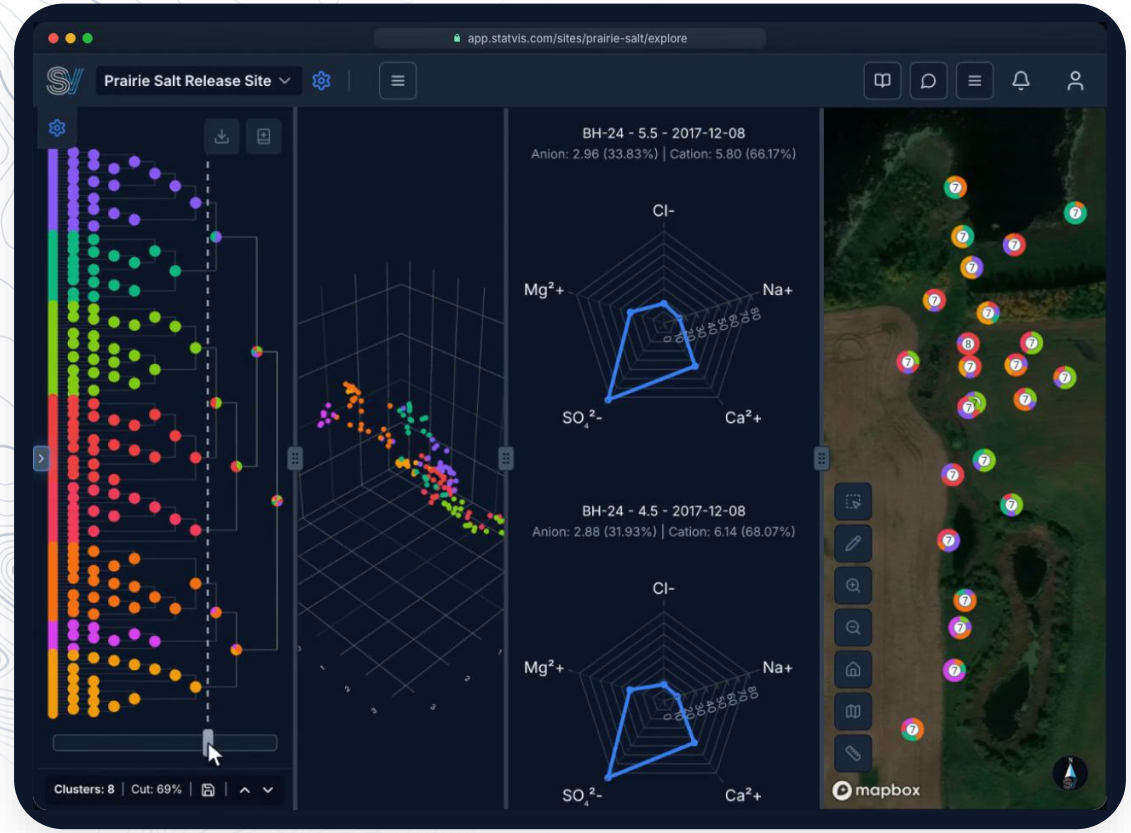
Then: five tools. Now: one platform

THEN (ABOUT 2018)

Excel	Histograms, drawn by hand
JMP	HCA, PCA, UMAP: the only interactivity
GIS	Maps
R	Statistics
PDF reports	Complex science, flattened for the other parties

Export, re-import, repeat.

NOW: STATVIS EXPLORE, SHARED WITH THE OTHER PARTIES



Data first, then documents. You need both.

We started with the data. We kept going back to the documents to understand it.

DATA ALONE

Numbers without context

- Which well, which depth?
- Before or after the remediation?
- Did the lab change its method?

DOCUMENTS ALONE

Context without structure

- Tens of thousands of pages
- Results stuck in scanned tables
- Nothing to map, chart or compare

BOTH

Every number in its context

- The sample on the map, linked to its report
- The method, the lab, the certificate
- The site’s story behind the data

Use AI to build the tool, not to be the tool.

Generation for the build. Repeatable code for the run.

AI builds the tool: check it once, run it forever

1

Describe the method

Screen 2024 benzene results against the Tier 1 guideline

2

AI writes the code

A short script or spreadsheet you can read

3

You test it

Against results you've checked by hand

4

It runs the same way

On every site, every time

Same inputs, same outputs. When it's wrong, it's wrong the same way, so you can find it and fix it.

MW-03 ✓ MW-03 ✓ MW-03 ✓
MW-03 ✓ MW-03 ✓ MW-03 ✓

Validate the tool once, then trust its runs, instead of re-checking every answer.

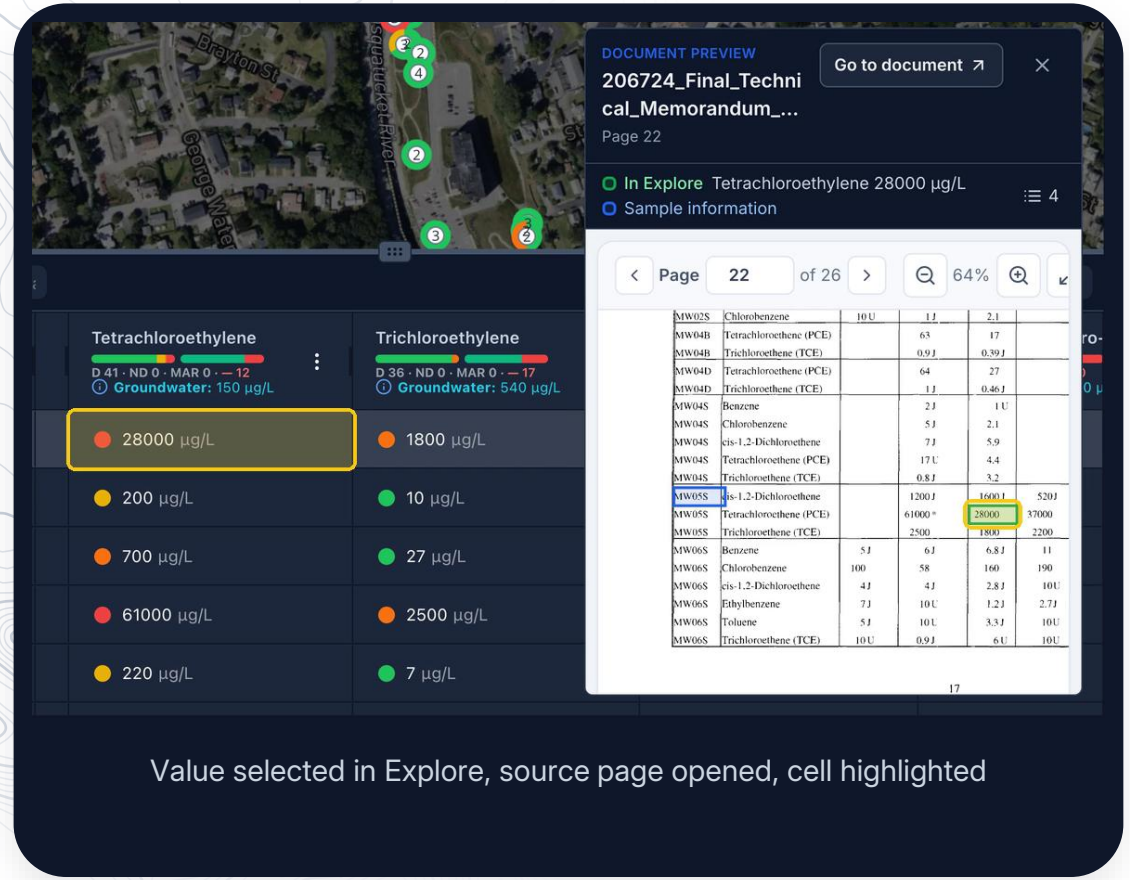
From the histogram back to the source

Lots of ways to notice something's off:

- An exceedance on the map
- A value that's too high in the table
- A fingerprint that looks weird

Either way, you want the lab page it came from. That used to take 90 minutes and break your flow.

Now it's one click.

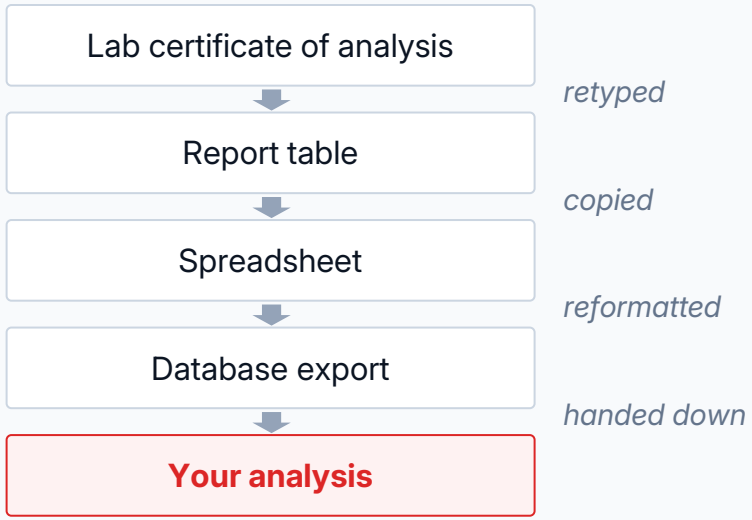


Value selected in Explore, source page opened, cell highlighted

Statvis Explore, Centredale Manor Superfund Site (public EPA records).

Errors in handed-down data

HOW DATA USUALLY ARRIVES



Errors carry forward, unseen.

CASE 1 · SPREADSHEET

A whole column shifted

- Entered twice, and one column shifted
- The reported exceedances were wrong
- Found by luck: someone saw a repeated number

CASE 2 · DATABASE USED BY SEVERAL PARTIES

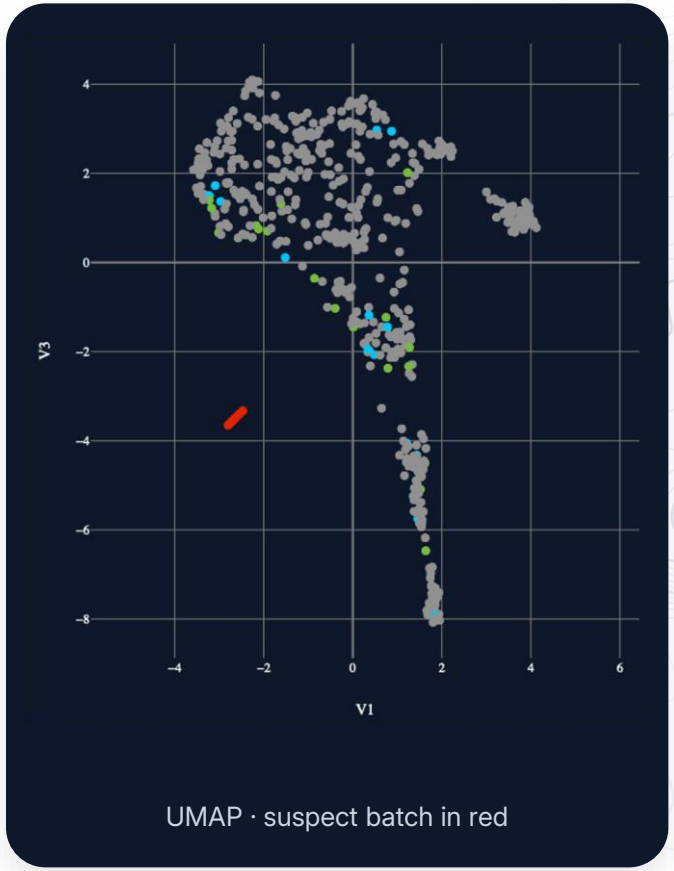
One sample counted twice

- A field duplicate counted as its own sample
- In a cost allocation between parties
- **One party's share for that area: 5.62% to 5.43%**

On almost every site, we find errors in the old data. Re-extraction checks every entry against the lab certificate.

Both cases anonymous. Case 2 values changed to protect the client; the size of the change is similar.

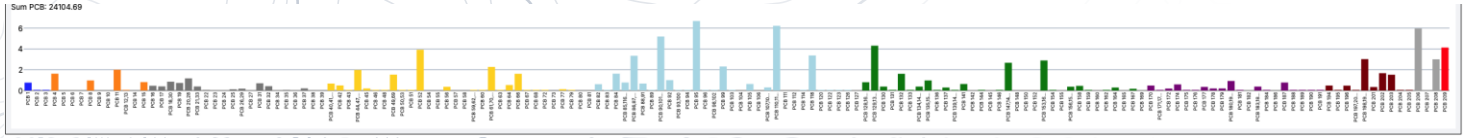
Centredale PCBs: bad data that passed validation



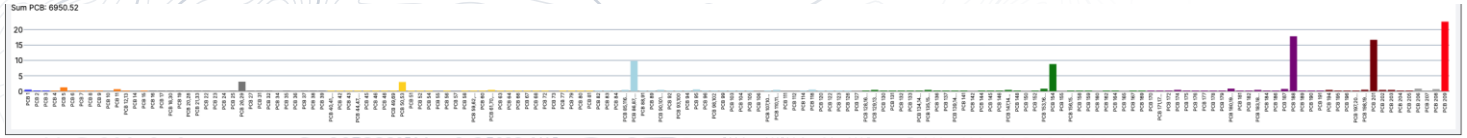
UMAP · suspect batch in red

- Third-party validated PCB data. UMAP put one lab batch (red) far from the rest
- The batches before and after: same areas, days apart. Not a new source
- **The bad batch had been in the case file for 3 years. PCA didn't show it**

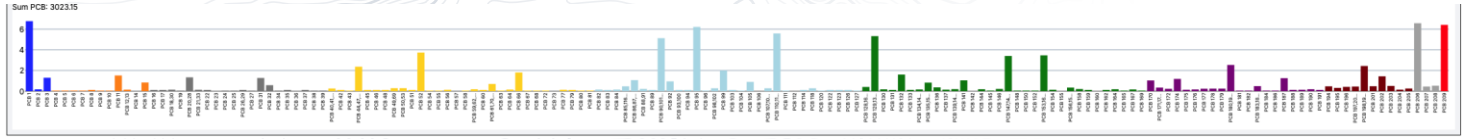
Batch before



Suspect batch

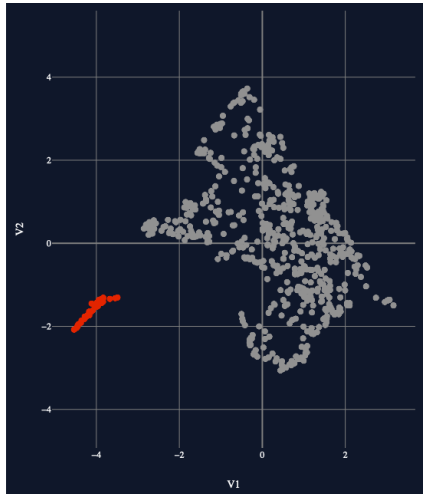


Batch after

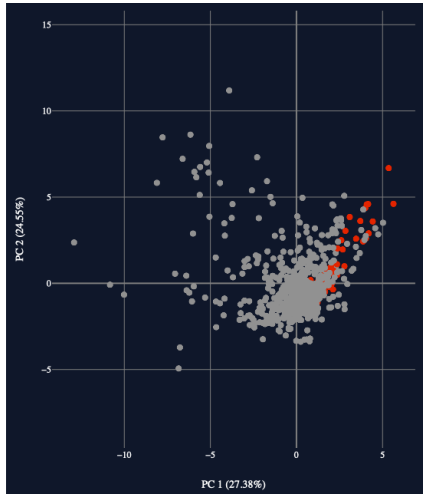


PAHs: a lab method changed the fingerprint

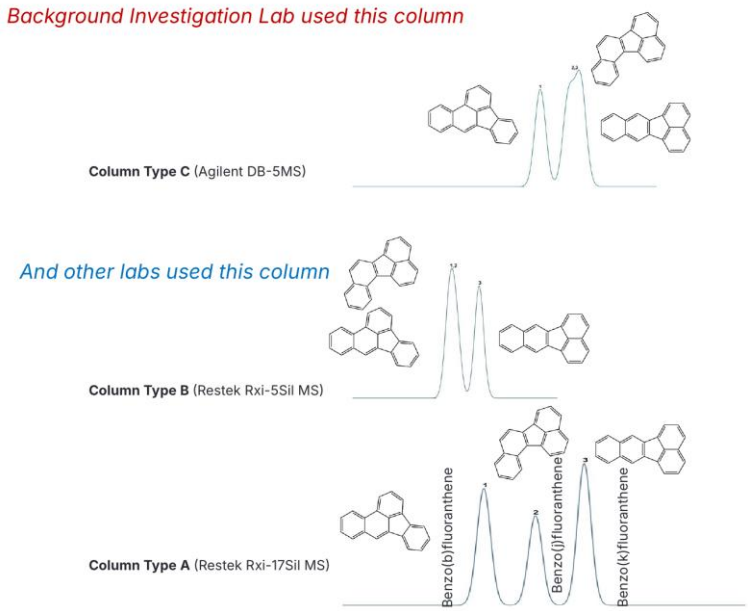
UMAP



PCA



- UMAP put the background lab's samples (red) apart. PCA did not
- **The supplied spreadsheet did not show the co-elution**



Background lab: B[j]F co-elutes with B[k]F
Other labs: B[j]F co-elutes with B[b]F

Result: the QA/QC passed, but the data could not be used to evaluate sources.

AI takes it to the next level

RECEPTOR MODELLING ON A SUPERFUND CASE

~7,000

samples, with several receptor models, to identify contaminant sources

~6 weeks

what this job would normally take us

~6 hours

with a repeatable workflow

A repeatable workflow

The expert designed every step

The model runs are math

R packages, no AI in the run

The expert decides

What the result means

We test it before we trust it

1 KNOWN ANSWER

Receptor model workflow

Reproduces our 2022 published end members (about 0.99 cosine similarity)

2 INDEPENDENT SOFTWARE

Statvis against JMP

Sample counts did not agree (6,085 against 6,024).

3 PARALLEL BUILD

Non-detect imputation

Built outside Statvis, then inside. Same seed, same values in both

Known answers, independent software, parallel builds. Then we use it on a case.

The honest takeaway

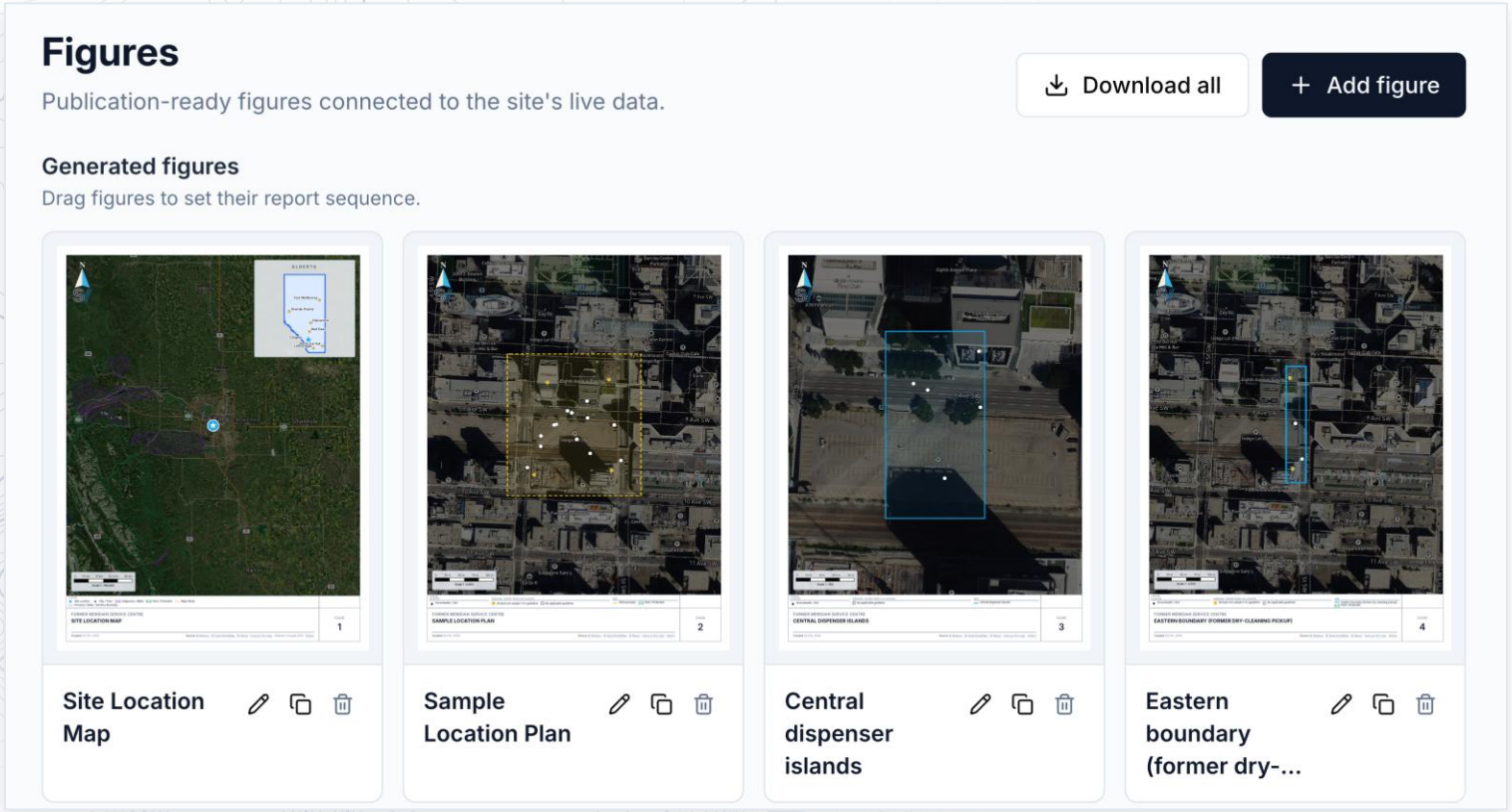
Building a repeatable workflow you can trust takes about as long as doing the job by hand the first time. **The payoff comes after.**



Not checking the outputs yourself? That's risk. AI can be confidently wrong without you knowing.

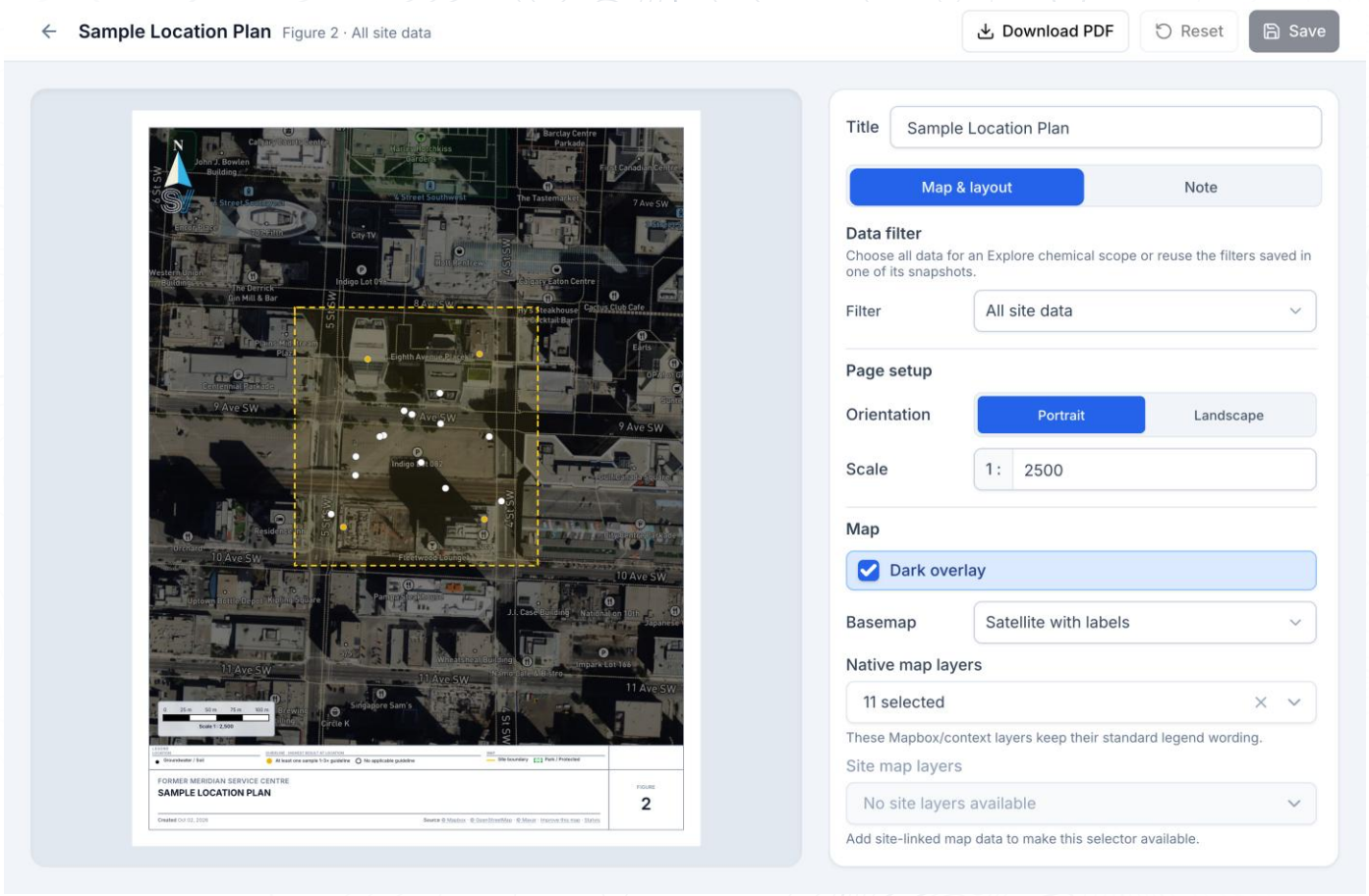
Repeatable workflows: figure generator

- Standard figures, generated from live site data
- The same figure every time, from the same inputs
- Regenerated automatically when the data changes



The Figures page in Statvis: a set of standard figures for one site

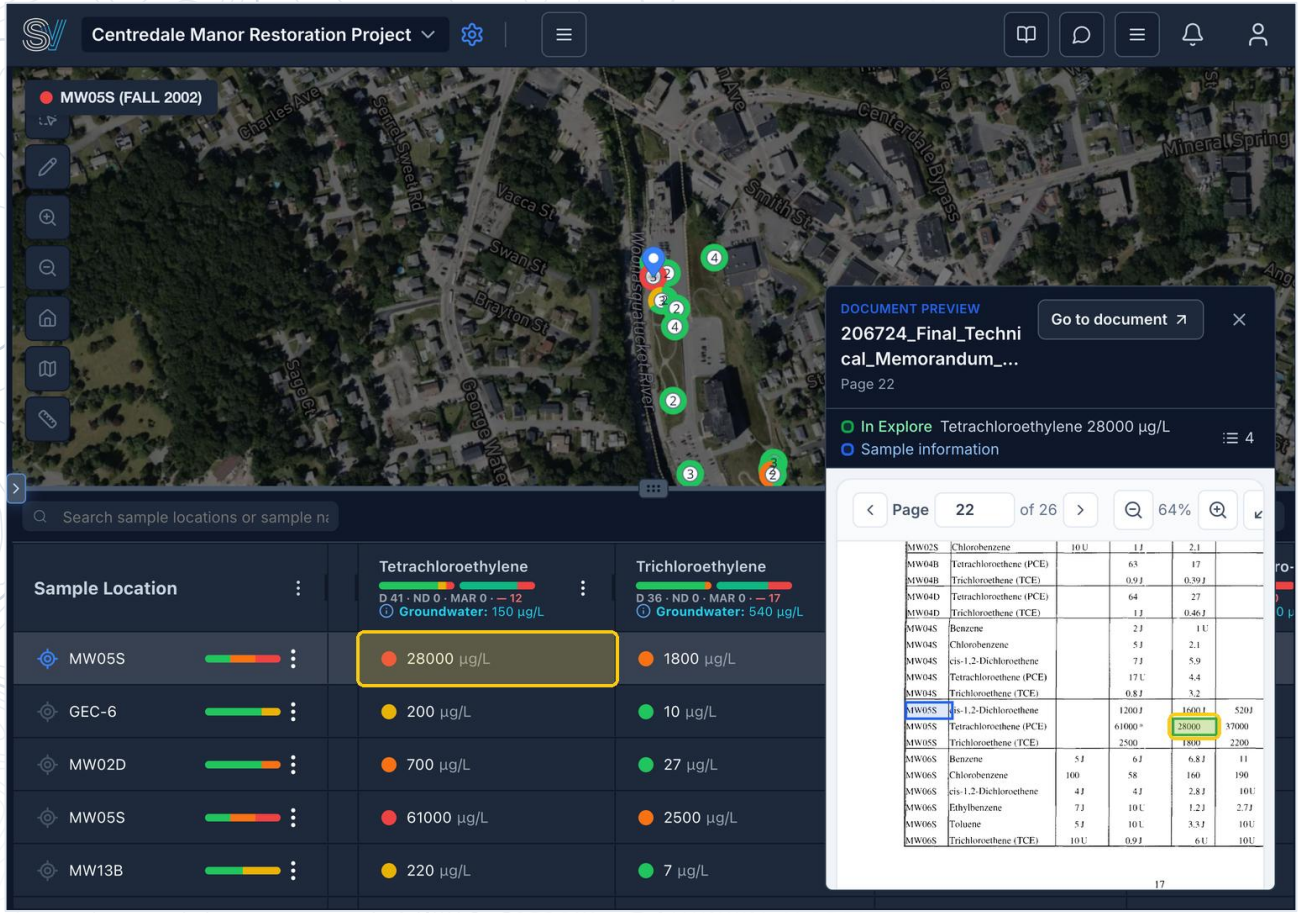
Repeatable workflows: one figure, every time



Every setting saved with the figure, every input traceable, regenerated when the data changes.

Repeatable workflows: data extraction

- Lab tables from scanned reports into structured data
- Units, detection limits and qualifiers kept intact
- Every value traced back to its page and cell



Centredale Manor: a 2004 EPA lab table, read into Explore and traced back to its cell.

Data consistency

The same chemical, the same wells and the same results, written four ways across four reports.

1987 · TYPED LAB REPORT

WELL	PARAMETER	RESULT
MONITORING WELL NO. 1	PERCHLOROETHYLENE	0.8 ppb
MONITORING WELL NO. 2	PERCHLOROETHYLENE	ND

1998 · CERTIFICATE OF ANALYSIS

Sample ID	Analyte	Result
MW 1	Tetrachloroethene	0.83 µg/L
MW 2	Tetrachloroethene	<0.5 µg/L

2011 · PHASE II ESA TABLE

Location	Parameter	Result
BH01-MW	PCE	0.00083 mg/L
BH02-MW	PCE	<0.0005 mg/L

2019 · LAB DATA EXPORT

Location	Chemical	Result
MW01	Tetrachloroethylene	0.83 ug/L
MW02	Tetrachloroethylene	0.5 U ug/L

NORMALISED

Chemical	**Tetrachloroethylene**
Location	**MW-01**
Result	**0.83 µg/L**
Location	**MW-02**
Result	**Not detected (<0.5 µg/L)**

4 names for one chemical
4 names for each well
3 units
3 ways to write “not detected”

Right tool for the job: machine learning vs LLMs

Machine learning

- Purpose-built models for one task: UMAP, clustering, classification
- Repeatable once trained, with fixed settings and seeds
- Results measurable against known labels

LLMs

- General-purpose: language, documents, code
- Flexible, but answers can vary run to run
- Needs checking against sources

Use machine learning where a defined, testable method exists. Use LLMs to read, find and draft.

Would you testify to it?

BUILT FOR SUPERFUND LITIGATION. NOW FOR EVERY SITE.

We've pressure-tested these workflows on billion-dollar cases, where every number can be challenged.

Ask yourself: would you be comfortable testifying to your workflow's outputs, on the stand?

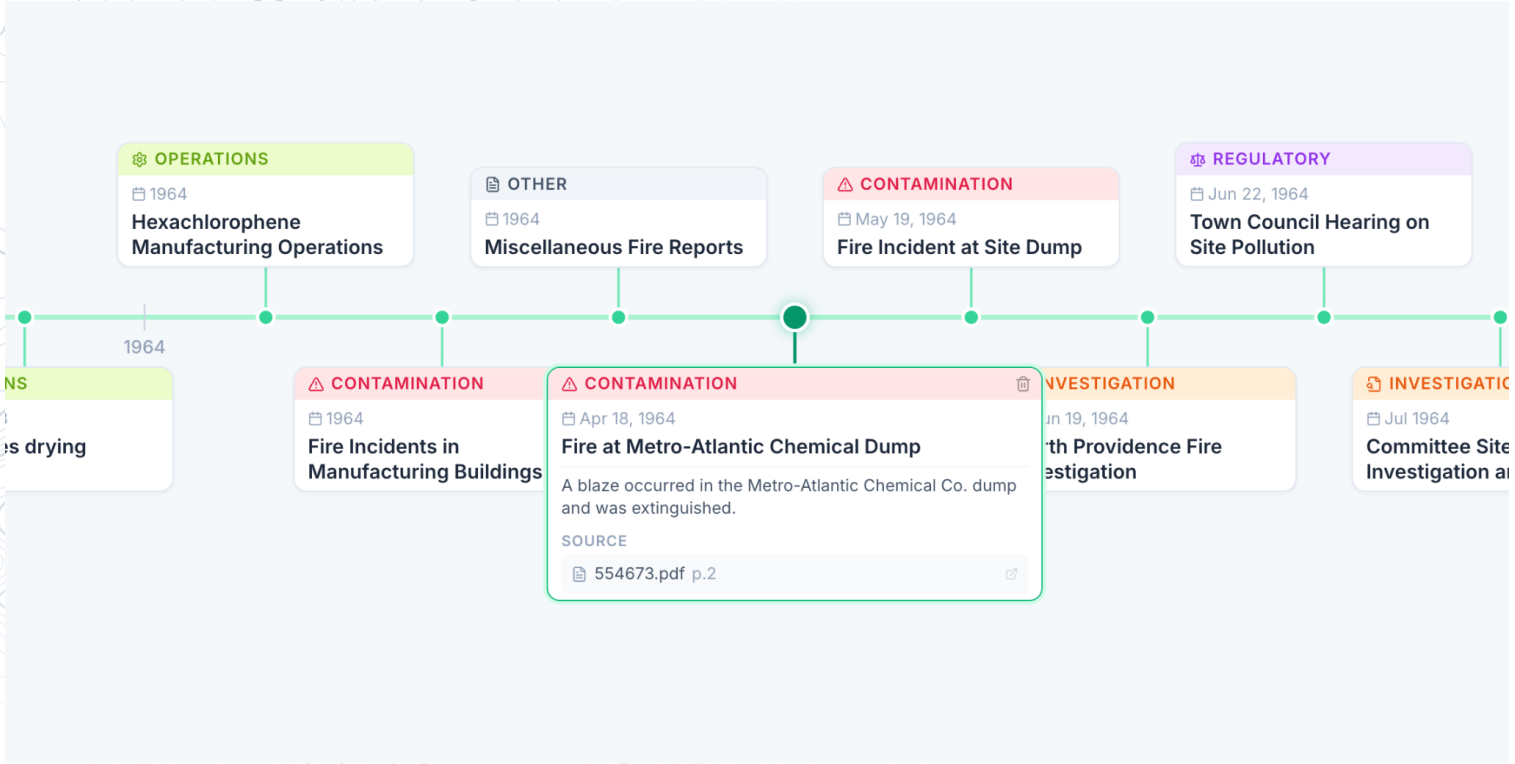
Superfund-grade data mastery, for the everyday contaminated site.

Why we’re hopeful about AI

Closing thoughts

Nothing gets missed

- Every page of decades of site files can actually be read
- The 1964 newspaper clipping is found, dated and cited, not forgotten



A 1964 dump fire, found in a newspaper scan and placed on the site timeline

Faster, better answers

What happened to the old fuel tanks, and was any contamination found?

Here is a summary of what the documents record about the old fuel tanks and associated contamination findings.

Fuel Tanks: What Happened

Underground Storage Tanks (Southwest Tank Nest)

The site operated as a service station with **four underground storage tanks (USTs)** located in the southwest portion of the property, referred to as the "southwest former tank nest." 1 These tanks were associated with retail gasoline sales and a former product-line corridor. 2

All four tanks were **removed in 2000**, coinciding with the cessation of fuel sales and the building's conversion to courier use. 3 The southwest excavation footprint marks where the tanks and product lines were located. 4

Waste-Oil Tank (North Service Bay)

A separate **2,300-litre waste-oil tank** was located in the north service bay. It was also **reportedly removed in 2000**. 5

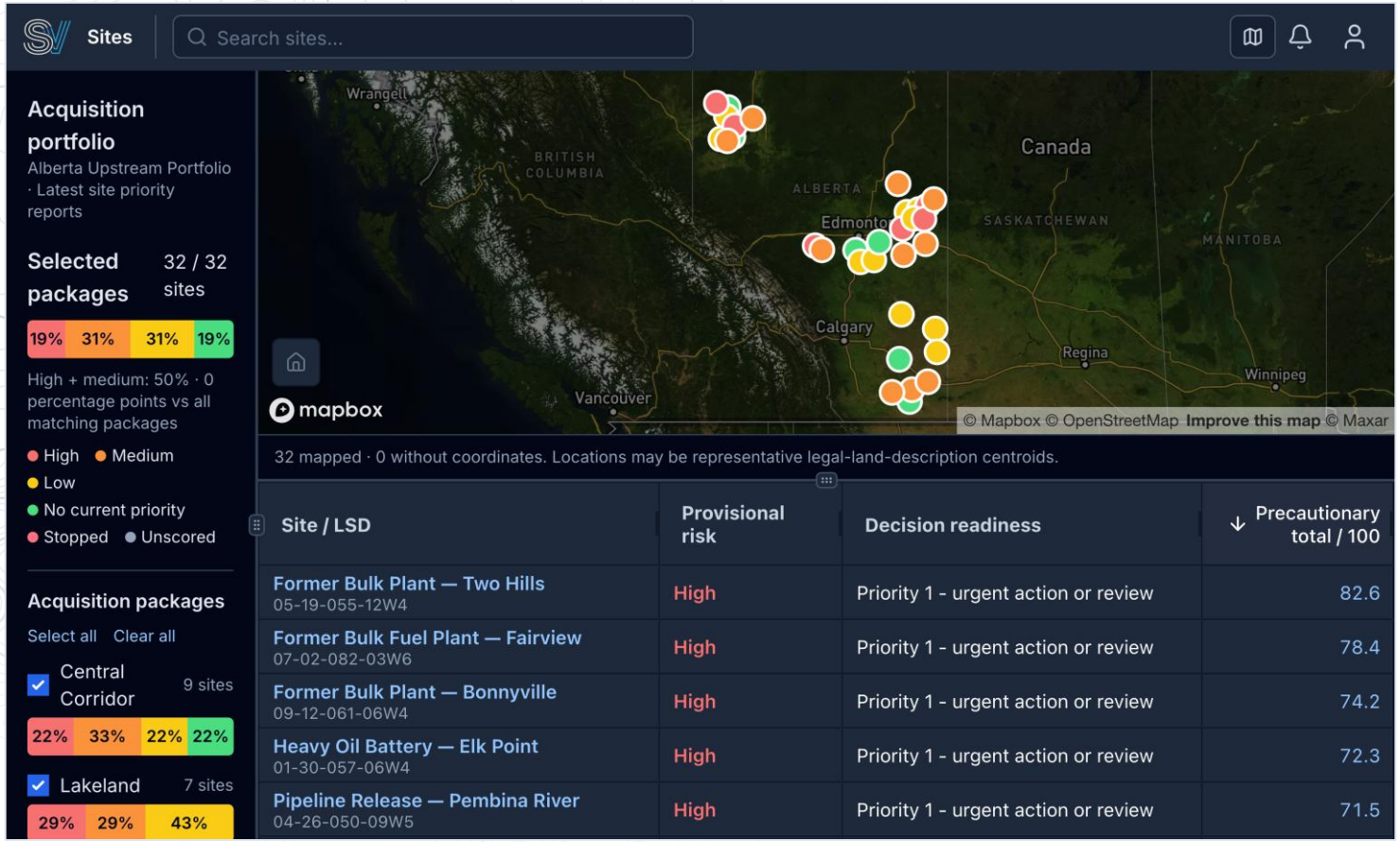
Contamination Found

At the Time of Tank Removal (2000)

Document questions answered in minutes, not weeks, with a citation on every fact. A real Statvis answer.

Work that wasn’t possible before

- **Scoring a 300-site portfolio** in an afternoon, not a year of reading
- **The only limit now is compute**
- **Re-run it** whenever new documents or guidelines arrive



A portfolio of 32 sites ranked by risk, each score built from the site’s own data (fictional demo sites)

AI will move you up the value chain.

Less time on menial work. More time on the judgement only you can make.
And the tools keep improving.

Case study: naming documents for a multinational energy company

Every document must be renamed to a strict corporate naming standard before it goes into their records system.

260	× 300	= 78,000	× 3 min	= 3,900
documents on one of their sites	sites in the portfolio	documents to rename	per document, timeboxed	hours of menial work

≈ 2 people working full time for a year just to rename files, at 1,800 working hours each

13%

fewer naming errors than people

Blind test of 251 documents, scored against the company's written naming standard

Humans get tired. AI workflows don't. The 78,000th document gets the same care as the first, and every name follows the standard.



Questions?

statvis.com