

Don't let your agent answer from

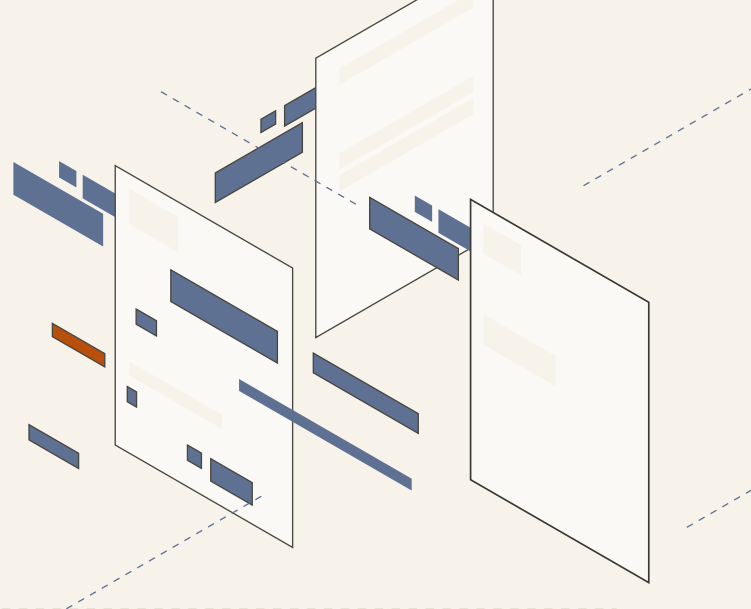
a draft

an old version

a contradiction



What's inside



01 Where bad answers come from

02 Old versions: A 46% accuracy gap

03 Contradictions: The right document isn't enough

04 Metadata coverage: Questions no single document can answer

05 How to fix unstructured data quality that jeopardizes accuracy

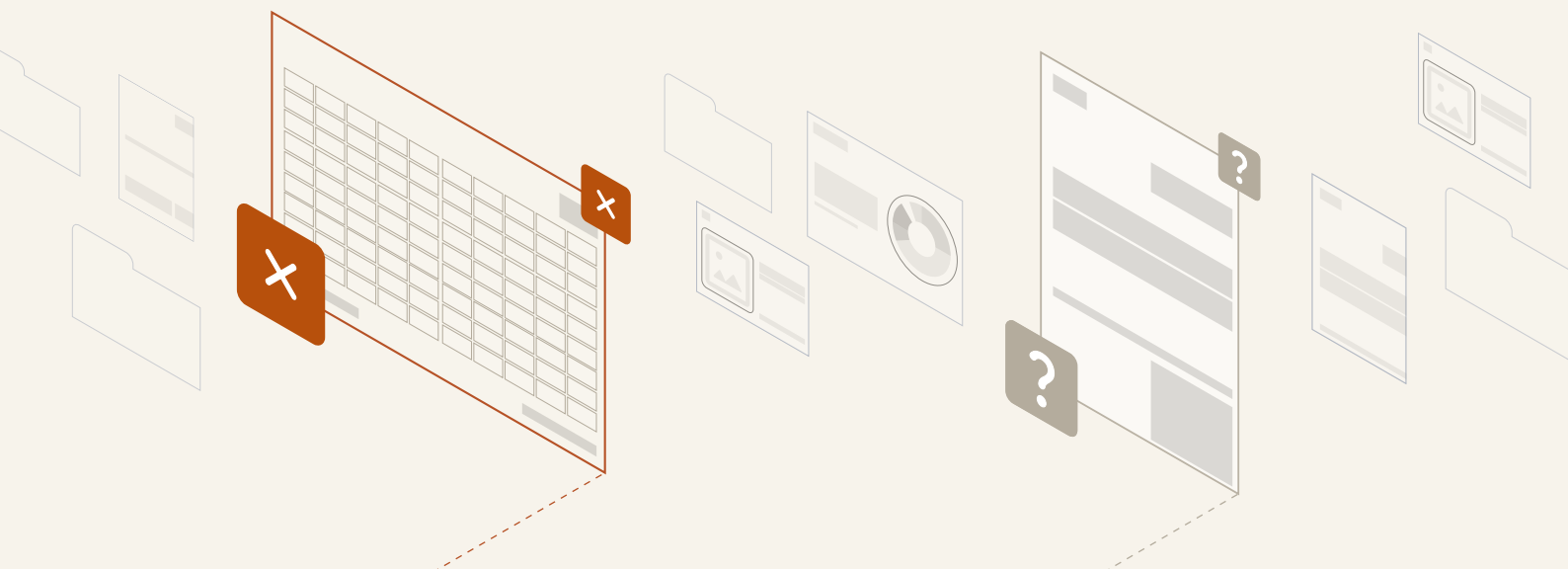


Your agent is only as right as what it reads from

The default fix for poor RAG accuracy is tuning, changing the embedding model, a reranker, a bigger context window.

But embedding models and rerankers score how well a document matches the question. They don't know which version of a policy is current, so both versions of the same policy look like equally strong matches.

Accuracy problems start upstream, in the data. That shows up in two big ways.



Something's there that shouldn't be

An old version, conflicting information between two files, sensitive details

Something's missing

For some question types, agents need metadata to get a right answer



Removing old versions raised answer accuracy by 46%

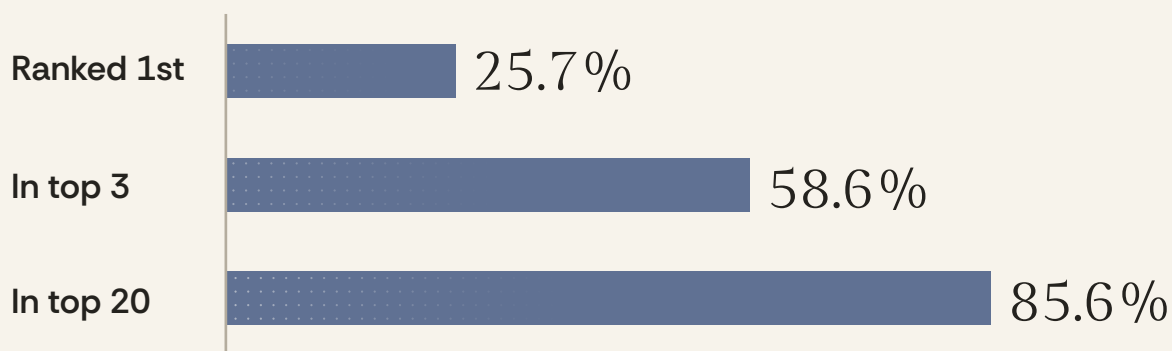
Enterprise repositories keep everything. When a policy doc gets updated in SharePoint, the old version almost always stays.

When two versions are nearly identical, retrieval can't tell which one is current, so it pulls both. We found that in a quarter of queries, the old version was actually the top result.

46%

relative accuracy gain after Deasy removed the outdated versions. Correct answers went from 27.2% to 39.7%.

Where the outdated version shows up in retrieved results



905 IBM support documents from RAGBench-TechQA plus 300 outdated copies, 319 questions, identical stack in both runs. [Full study](#)



Having the right document isn't enough when the old version is still there

Even when retrieval finds the correct policy doc, the old version usually comes along with it.

72%

fewer contradictory answers
once the old version
was removed

What the model reads
raw corpus

Current policy doc



Old policy doc, older facts



Conflicting facts

A hedged or wrong answer

What the model reads
after cleanup

Current policy doc



One current fact

The right answer

Same benchmark as section 02, measured only on questions where the correct document was already retrieved. Correct answers rose from 36.2% to 50.9% under deliberately strict scoring. [Full study](#)



Metadata coverage decides which questions your agent can answer

Some questions don't have an answer in any one document. "How many of our contracts auto-renew?" needs a count across every contract. To answer it by reading, an agent has to open each one and keep a tally.

We gave the same agent 56 questions like that about 483 S&P 500 companies. One version read the documents, the other queried metadata Deasy extracted from them.

50% more

questions answered correctly with metadata

With metadata 45/56

Reading the documents 30/56

How many of these companies are headquartered in Texas?

The right answer is 49

With metadata

1 query
17,601 tokens

49

Reading the documents

51 files
1.3M tokens

38

21.6x

fewer tokens, too

Across all 56 questions, the agent using metadata spent 21.6 times fewer tokens than the agent reading the documents.

Same agent, model, and budget throughout with deterministic scoring. [Full study](#)



How to fix unstructured data quality

Every accuracy gain in this guide came from fixing the data before retrieval. Here are the seven dimensions of unstructured data quality and what to do about each one.

Dimension	What to do	What it prevents
Uniqueness	Find duplicates and near-duplicates. Keep the canonical copy, flag drafts and superseded versions so they stay out of the index	Answers from an old version. +46% accuracy in our benchmark
Freshness	Flag expired and stale files using expiry tags and last-modified dates	Answers from content that's no longer true
Conflicting information	Find documents that state contradictory facts, determine which one the agent should trust	Hedged or two-answer responses
Relevance	Score every file against the use case and drop what doesn't serve it	Irrelevant files crowding out the right chunk. About 30% of content that looks relevant isn't
Sensitivity	Detect PII, PCI, PHI, and credentials across 40+ classifiers, and gate them per use case	Sensitive details reaching someone who shouldn't see them
Metadata coverage	Make sure the tags your agent filters and counts on are present on every file	Questions the agent can't answer without metadata. 50% more answered in our benchmark
Content completeness	Catch truncated extractions, missing pages or signatures, and empty sections	Answers built from a partial doc

Curating this by hand is a big lift for one project, and it doesn't scale to millions of files.

Deasy does all seven automatically and writes the metadata back to SharePoint, S3, and OneDrive, so you never rebuild a curation pipeline.



See what your agent is seeing

Book 30 minutes with us, and we'll dig into your accuracy and the unstructured data that's determining it.

[Book a demo →](#)

