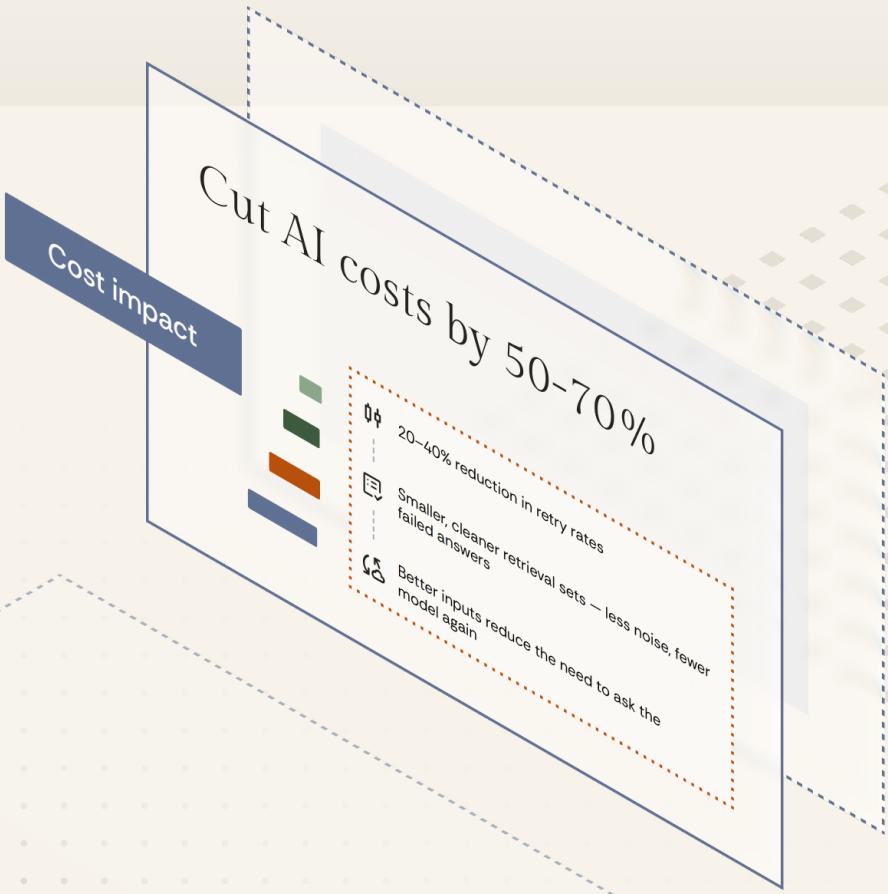
Deasy Labs

How to save 50-70% on your AI costs with data curation



A Deasy Labs field guide

Based on observed enterprise AI deployments across industries.

What's inside

01 Where enterprise AI costs hide

02 Lever 1: Curate what enters the pipeline

03 Lever 2: Retrieve at the chunk level

04 Lever 3: Skip the LLM for structured tasks

05 Retries: The multiplier most teams miss

06 What changes with metadata-driven retrieval

See methodology and assumptions below each chart.

Where enterprise AI costs hide

Every token you send a model costs money. When your retrieval pipeline pulls irrelevant documents, outdated files, or an entire contract when the answer is sitting in a one-sentence clause, you end up paying for noise at enterprise scale.

Baseline example: Customer service AI assistant

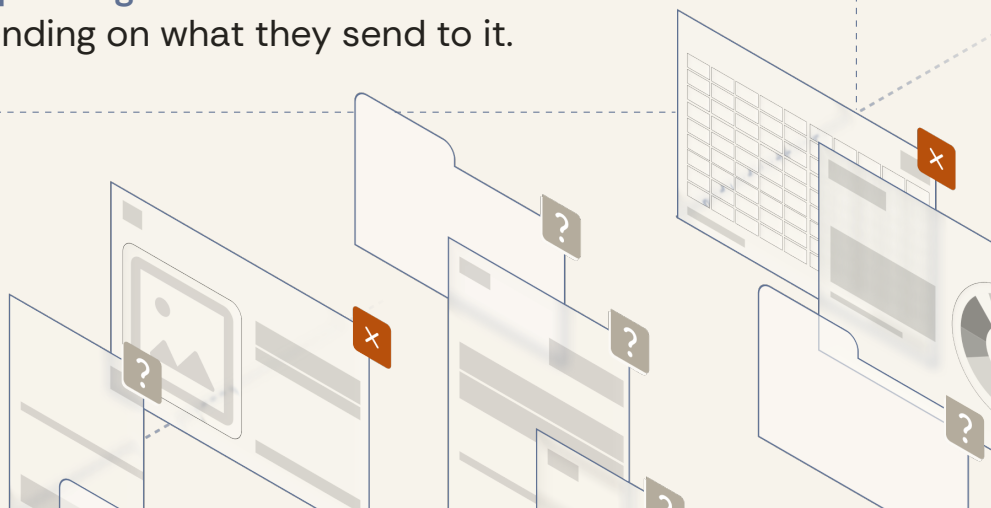
- Corpus size: 100,000 documents
- Daily query volume: ~2,000 queries/day
- Avg tokens per query: ~10,000 (input + output)
- Model cost (blended): $3–$10 per 1M tokens
- Estimated annual cost: $20K–$70K per AI project

Assumptions: GPT-4 class blended pricing; 1,000–5,000 queries/day; 8K–20K tokens/query. This baseline carries through every chart in this guide.

Storage is cheap; retrieval mistakes are expensive

In most enterprises, vectorDB storage runs $1K - 10K a year, but LLM usage runs 10 to 100x higher.

Most teams aren't overspending on AI.
Instead, they're overspending on what they send to it.



Lever 1

Curate what enters the pipeline

Cost reduction comes from the same three places in every deployment we've seen. Each works on its own, and together they compound.

The highest-leverage move is stopping irrelevant files from entering the pipeline at all. Enterprise data is never clean. There are duplicates, sensitive files, and often a huge volume of content that's completely unrelated to the AI project.

And in an even costlier setup, many teams use a multimodal LLM to ingest everything upfront.

30% of the data that seems relevant to your project at first glance isn't, so ingesting that data amounts to waste.

Deasy maps and filters your corpus before anything reaches a model, removing outdated and duplicate files, scoping for relevance, and excluding everything sensitive.

Cost impact

Filtering ~30% of irrelevant or duplicate data before multimodal ingestion

Roughly **$75K–$150K/year** in direct savings, and up to **$500K+** with reprocessing and infrastructure overhead.

Assumptions: ~5M documents/year at ~3K tokens each (enterprise-wide adoption). Model cost ~$0.006 per 1K tokens (multimodal, GPT-4o class). ~30% of documents irrelevant or duplicate, with 2–3× reprocessing over time from taxonomy changes, new AI projects, and model upgrades.

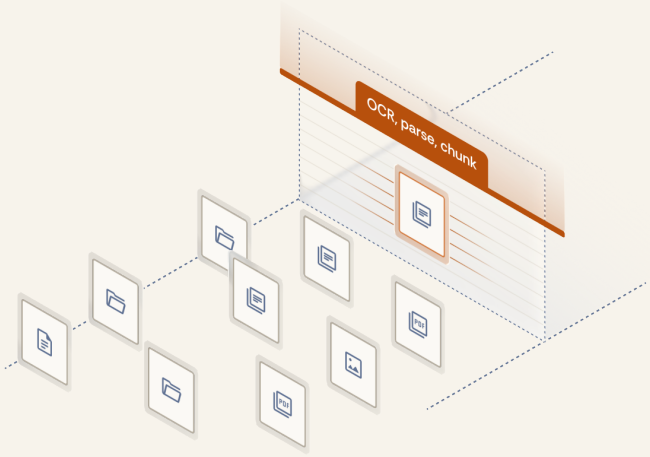
Lever 2

Retrieve at the chunk level, not the document level

A 40-page contract often has just one relevant clause to an answer your downstream AI system will need to pull. Or a 200-page manual has three relevant paragraphs. Sending whole documents bloats context and degrades performance. Deasy embeds at the page and chunk level and surfaces only what each query needs.

You don't need less data.

You need less irrelevant data per query.



Metric	Without optimization	With chunk-level retrieval
Chunks retrieved per query	~30	~8
Tokens per query	~10,000	~4,000
Reduction in token usage	n/a	40–70%

Assumptions: Same 100,000-document baseline; ~10,000 tokens/query before optimization. Chunk reduction of 60–80% with metadata.

Lever 3

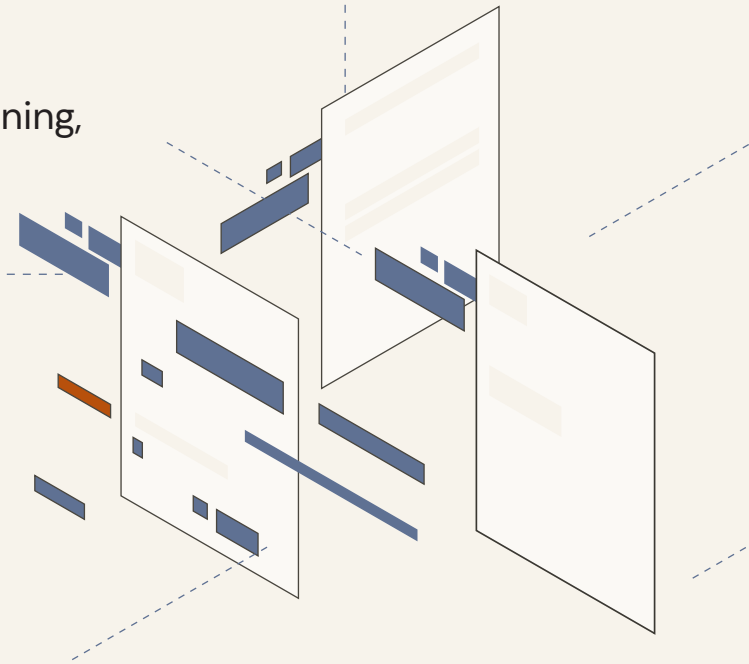
Skip the LLM entirely for structured tasks

A large share of classification tasks need no model call. Extracting dates, detecting PII, reading amounts: predictable and repeatable. Deasy handles these with regex pattern matching combined with contextual keyword matching. This method matches or beats LLM accuracy for far less.

The point is control. With Deasy, you decide where a model earns its cost. Nuanced work goes to an LLM — like reading sentiment in a call transcript — while deterministic work goes to pattern matching. Where you draw that line is always yours, and it is what sets your LLM spend.

Cost impact

- 30–60% of classification tasks bypass LLM calls entirely
- Cost per task reduced by 10–100×
- Model budget freed for reasoning, not pattern matching



The retry multiplier most teams miss

There is one cost this research's baseline misses entirely: retries. Enterprise AI rarely returns the right answer the first time, and 30 to 50% of queries trigger at least one retry, each one re-running retrieval and re-sending tokens. That turns the baseline into a multiple.

Scenario	Cost multiplier	Effective annual cost
No retries	1.0×	$20K–$70K
Moderate retries (30%)	1.5×	$30K–$105K
Heavy retries (50%)	2.0–3.0×	$40K–$150K+

Assumptions: Multipliers applied to the $20K–$70K baseline. Baseline retry rate of 30–50% of queries. Multiplied across dozens of AI applications, this becomes a multi-million dollar leak no one notices.

The three levers you just read through do more than cut tokens per query.

By improving retrieval, they also pull the retry rate down.

What changes with metadata-driven retrieval

The big unlock is combining all three levers with strong metadata: filter out what's irrelevant and duplicated, retrieve at the chunk level, and skip the LLM for structured tasks.

Metric	Without optimization	With Deasy
Chunks retrieved per query	~30	~8
Tokens per query	~10,000	~4,000
Retry rate	40%	15%
Effective tokens per query (incl. retries)	~14,000	~4,600
Cost per query	~$0.10	~$0.03
Annual cost (same use case)	~$70,000	$20,000–$30,000

Assumptions: Same baseline deployment. Chunk reduction 60–80%; retry rate cut from 40% to 15% with better retrieval. Cost reductions assume no drop in answer quality; in practice accuracy improves.

Net result:
50-70% reduction in LLM spend, with higher answer accuracy.

Deasy Labs

See it on your data

If you want to see this savings modeled on your corpus, book a 30-minute demo and we'll dig into your data and environment.

Book a demo →