Progress Agentic RAG Tokens: You're not buying a token. You're buying a pipeline.

DATA SHEET

PROGRESS AGENTIC RAG TOKEN DETAILS:

- ✓ One billing unit
- ✓ No embedding fee
- ✓ No storage fee
- ✓ No LLM pass-through markup

This lands fast: "Your token is $0.008. OpenAI's is $0.000025. You're 320x more expensive."

Wrong comparison. An OpenAI input token is one forward pass-through one model. But a Progress® Agentic RAG normalized token is the unit of work for a complete managed retrieval pipeline, including vector index, hybrid search, context assembly, multi-large language model (LLM) orchestration, governance and quality scoring. Comparing them is like pricing a managed Snowflake query against one CPU cycle on bare metal—different abstractions with different value.

What's Included in a Progress Agentic RAG Token

The Progress Agentic RAG solution charges normalized retrieval tokens measured when you ask a question, not when you load data. The consumption formula includes: input tokens + output tokens, image tokens and reasoning tokens. Question processing, context assembly, prompt creation, LLM processing and answer generation. That's the scope of what's metered.

Inside that token is hybrid retrieval across your indexed corpus, multi-LLM routing (the right model per task, per pipeline action), governance and access controls, citation tracking back to the source paragraph and REMi (RAG Evaluation Metrics Intelligence) quality scoring on every run. The exception is specialist processing at ingestion (e.g., vision models on images, speech-to-text on audio) where the underlying model has its own cost profile. Everything else is retrieval time only.

The reason this matters at scale: a typical enterprise corpus is uploaded once and queried tens of thousands of times. Charging at upload bills you for static work. Charging at retrieval bills you for the work that actually produced an answer.

The Real Comparison

Line up the managed platforms and the picture changes fast:

- **Vectara:** Ingestion plus retrieval plus storage, Mockingbird by default
- **Pinecone:** Vector storage plus read units, LLM and governance on you
- **Amazon Bedrock KB:** Per-page ingestion, per-query retrieval, LLM pass-through at AWS list rates and one cloud, full stop
- **Microsoft Azure AI Search:** Multiple SKUs, multiple bills, one working RAG stack if you assemble it right
- **Google Gemini Enterprise Agent Platform (formerly Google Agent Search and Vertex AI Search):** Per-document ingestion, storage, per-query search and LLM pass-through at Vertex rates; GCP-only, Gemini-first

Every competitor's pricing page shows the floor, so by the time you add the line items, you're well above it. The token within the Progress Agentic RAG solution is the whole number including at retrieval orchestration, multi-LLM routing, governance, citations and quality metrics, one unit, at retrieval time only.

Platform	What You Pay For	Approximate Unit Cost
Vectara	**Ingestion, storage, retrieval** Mockingbird LLM by default, model flexibility costs extra	~$2.50/1k queries + storage
Pinecone	**Storage, read units,** LLM pass-through with DIY architecture: you wire the LLM, governance and evaluation layer	$50-500/month + LLM tokens
Amazon Bedrock KB	**Ingestion per page, retrieval, LLM at AWS list rates**; AWS-only Token rates above API pricing with no multi-cloud offering	$.10/query + LLM tokens
Microsoft Azure AI Search	**Tiered subscription, index storage, query units, LLM tokens**, multiple SKUs to assemble before you have a working pipeline	$250-2K/month + LLM tokens
Google Gemini Enterprise Agent Platform	**Ingestion per doc, storage, search queries, LLM at Google rates**; GCP-only, Gemini-first	$2.50/1k queries + storage + LLM tokens
Progress Agentic RAG	**Retrieval only;** hybrid retrieval, multi-LLM orchestration, citation tracking, REMi quality metrics, connector licenses – one normalized unit, billed at retrieval time	$700/month + $0.008/token

The $0.008 token isn't a line item—it's the alternative to building and operating all of this yourself.

Vector index, hybrid retrieval, multi-LLM orchestration, governance, citation tracking, quality metrics assembled, managed and billed as one number, at the moment it delivers an answer. Not when you *upload* or when you *index*, but when you get a *result.*

The question was never: *Why is your token expensive?* The question is: *What does a grounded, measured, cited answer cost you today and what does it cost you when it's wrong?*

Try it yourself with a 14-day free trial