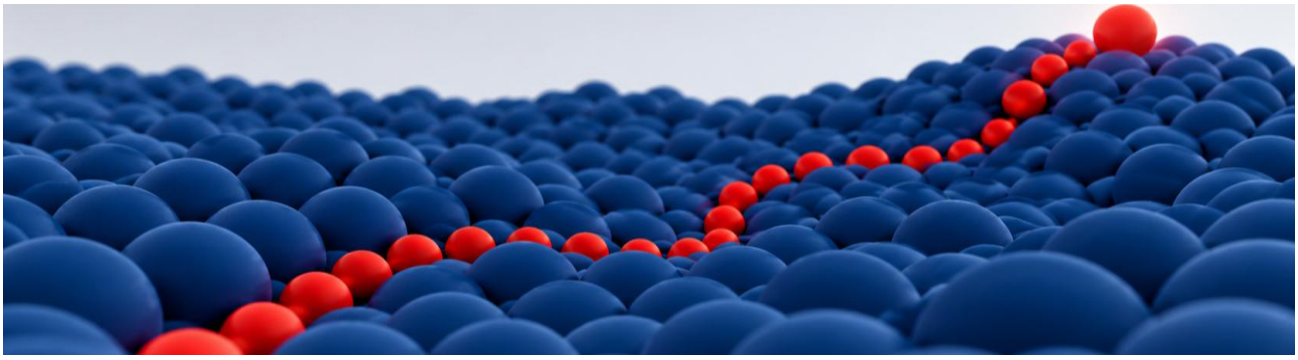
# How We Built a Smarter, Cheaper AI Assistant for Our Website

*PivotQ's website assistant*; answers that are accurate, fast, and affordable to run.

**PivotQ Insights | September 2026**



## The Problem with Most AI Chatbots

You have probably used a chatbot on a company website that gave you a generic, unhelpful answer. Or one that confidently told you something completely wrong. These failures are not accidents: they are the result of a common shortcut: asking an AI to answer questions purely from memory, without actually looking anything up.

AI models are trained on massive amounts of text, but they are not updated in real time. They do not know what your company actually does today, what your current pricing is, or what services you launched last month. If you want an AI assistant to represent your business accurately, it needs a way to look things up, just like a good employee would.

That is exactly what we built for PivotQ. This paper explains the three ideas behind it, in plain terms.

## Design choice 1: The AI Looks Things Up Instead of Guessing

Think of the difference between two employees answering a customer question. The first one improvises from memory. The second one says "let me check that for you" and looks it up in the company knowledge base before answering.

Our assistant works like the second employee.

Every time someone asks a question, the assistant searches through the actual content of the PivotQ website (all of it), finds the most relevant pieces, reads them, and then uses that information to form an answer. This is called Retrieval-Augmented Generation, or RAG. The "retrieval" part means looking things up. The "generation" part means writing a natural, conversational response based on what it found.

The result is an assistant that only says things that are actually true about PivotQ because it read them from the source before answering. When the website is updated, the assistant automatically knows the new information the next time it syncs.

## Design choice 2: We Store Knowledge the Smart Way: At a Fraction of the Usual Cost

To search through website content quickly, AI systems need to store it in a special format called a vector database. Think of it like a library where books are organized not by title, but by meaning. When you ask a question, the system finds content that means something similar to your question, even if the exact words are different.

The problem is that most vector databases are expensive. The most common options cost hundreds of dollars per month just to keep the lights on, even if nobody is using them.

We chose a newer approach: Amazon S3 Vectors. S3 is the same storage service that millions of companies already use to store files on Amazon's cloud. AWS recently added the ability to store and search vectors directly inside it.

Since it only charges for what you actually use, the cost for a site like PivotQ is less than a dollar a month, compared to hundreds of dollars per month that traditional options would cost. Same capability. No compromise on quality.

## Design choice 3: It Remembers Answers So It Gets Faster Over Time

The first time someone asks a question, the assistant does its full work: searches the knowledge base, thinks it through, writes a response. This takes about 10 to 14 seconds.

But here is the insight: visitors to a consulting website tend to ask similar questions. "What does PivotQ do?" and "Tell me about PivotQ's services" and "What kind of work do you take on?" are all asking the same thing in different words.

We built a system that remembers not just the exact questions people have asked, but the meaning of those questions. When a new question comes in, the assistant checks whether it has already answered something that means the same thing. If it has, it skips all the hard work and returns the saved answer immediately, in 1 to 2 seconds instead of 10 to 14.

This is called semantic caching. "Semantic" means relating to meaning. The cache gets smarter and faster as more people use it, because every answered question makes the next similar question faster.

## What This Means in Practice

These three ideas work together to create something that most AI assistants do not offer: answers that are accurate, fast, and affordable to run.

- Accurate because the assistant reads from the actual website before answering, not from general AI training data
- Fast because repeat and similar questions are answered from memory in seconds
- Affordable because every infrastructure choice was made with cost in mind, without cutting corners on quality

The entire assistant runs on Amazon's cloud with no servers to manage, no ongoing maintenance costs, and no monthly minimums. It scales automatically, whether one person asks a question or a thousand do.

## Why This Matters Beyond PivotQ

Every business with a website has the same problem: visitors arrive with questions, and there is rarely someone available to answer them at 2am on a Sunday. A generic chatbot does more harm than good if it gives wrong answers. Most custom AI assistants are either too expensive to run or too shallow to be useful.

The architecture described here solves all three problems simultaneously. It is accurate because it reads real content. It is fast because it learns from experience. And it is cost-effective because it uses the right tools for the job rather than the most expensive ones.

This is what modern, production-ready AI looks like: not a demo, but a system built to actually work.

## About PivotQ

PivotQ is a Silicon Valley AI consulting firm led by senior practitioners. We design and deploy AI applications, automation, and cloud solutions with security, governance, and measurable outcomes built in. Our work starts with a focused proof of value and a plan to operate and improve the solution.

Learn more at pivotq.ai.