Respondo

PRODUCT BRIEF

> **Try it yourself:** The information in this brief has been added to the Respondo Starter public demo — you can ask Respondo directly about itself at the live demo link: **https://dub.sh/jcqoK6A**

SECTION INDEX

One-Liner

Respondo is a private, source-available AI chatbot that answers questions using only the documents and data you give it — deployed on your own server, with your own AI provider, and owned outright with no monthly per-user fees.

High-Level Overview

Most teams have the same problem: customers, clients, or employees keep asking the same questions, and the answers are buried in scattered PDFs, manuals, and policy docs nobody wants to dig through. Respondo turns that pile of documents into a chatbot that can answer questions in plain language — and back up every answer with the actual source it came from.

Unlike typical SaaS chatbot tools, Respondo is not rented — it's built for you, deployed on infrastructure you control, and you own it outright. You pay for setup, upgrades, and maintenance, not a recurring per-seat or per-conversation fee. That means unlimited users and unlimited conversations; the ongoing costs are your chosen AI provider's usage fees and your cloud/server hosting costs, both of which you control directly.

Respondo is designed to be trustworthy by construction: it only answers from the knowledge you've given it, tells users when something isn't in its data instead of guessing, and shows the source behind every answer so people can verify it themselves.

Use Cases: Industries With High-Volume Repetitive Queries

Respondo delivers the most value in industries where the same questions get asked over and over — by customers, guests, patients, tenants, or employees — and where the answers already exist somewhere in a document, they're just hard to find quickly.

- **Hospitality (hotels, short-term rentals, resorts)** — Guests repeatedly ask about Wi-Fi passwords, check-in/check-out times, breakfast hours, gym and pool access, pet policies, and nearby recommendations. Respondo can answer instantly from house manuals and guest guides, day or night, in the guest's own language.
- **Property management and real estate** — Tenants and prospective renters ask about lease terms, maintenance procedures, amenity rules, and application requirements. A knowledge base built from lease templates and building policies handles these without staff intervention.
- **Healthcare and clinics (administrative, non-diagnostic)** — Patients ask about office hours, insurance accepted, appointment prep instructions, billing policies, and where to find forms. Respondo answers from clinic documentation, freeing front-desk staff for in-person needs.
- **E-commerce and retail** — Shoppers ask about shipping times, return policies, sizing guides, and order status procedures. A source-backed chatbot trained on policy pages and FAQs handles the repetitive volume around sales events.
- **Education (schools, universities, online courses)** — Students and parents ask about enrollment deadlines, course requirements, financial aid, and campus services. Respondo can serve as a always-available front line drawing from handbooks and program guides.
- **HR and internal employee support** — Employees repeatedly ask about PTO policy, benefits enrollment, expense procedures, and IT setup steps. Respondo turns internal wikis and HR manuals into a self-service assistant, cutting down on repetitive tickets to HR and IT.
- **Financial services (non-advisory, informational)** — Clients ask about account opening requirements, fee schedules, documentation needed, and general procedures. Respondo answers from official policy documents while staying within factual, source-backed bounds.
- **Legal and professional services (informational, non-advisory)** — Clients and prospects ask about intake procedures, document requirements, and standard service offerings — freeing staff from repeating the same onboarding information.
- **Government and public-sector services** — Citizens ask about permit requirements, office hours, application procedures, and eligibility criteria. A source-cited chatbot helps reduce call-center load for high-volume, low-complexity questions.
- **SaaS and software products** — Users ask setup, configuration, and troubleshooting questions already covered in documentation. Respondo functions as a documentation-grounded support layer, reducing repetitive tickets.

The common thread: any organization where staff spend meaningful time repeating information that's already written down somewhere is a strong fit for Respondo.

Who It's For

- Customer support teams fielding repetitive questions from clients or prospects

- Internal teams who need employees to self-serve answers from policies, manuals, or SOPs
- Onboarding and training, where new hires need fast answers instead of hunting through documents
- Any small business or service provider that wants a 24/7 assistant without hiring more support staff or paying a SaaS subscription per user

Core Value Proposition

- **You own it.** One-time setup and maintenance cost — not a SaaS subscription charged per user or per conversation. The code is source-available, so you're never locked into working with the original developer going forward.
- **Your data.** Add content by uploading PDFs, Word docs, or plain text files, or pasting information directly. Respondo can also pull information straight from your website.
- **Your AI provider, your cloud.** Deployed on your own server/cloud with the LLM provider of your choice — nothing is locked into a third-party platform.
- **Doesn't hallucinate answers.** If something isn't in the knowledge base, Respondo tells the user instead of making something up.
- **Verifiable answers.** Every response can show the source document (and file name or URL) it was drawn from.
- **Multi-language.** Respondo can incorporate data in multiple languages and receive questions in multiple languages. Even if the question and the source of the answer are in different languages, Respondo always answers in the language of the user.
- **Multi-channel.** Available as a web chat widget, and can connect to Facebook Messenger, WhatsApp, email, or a VoIP phone line.
- **Scales with you.** Start small on a modest setup, then scale up as usage grows.
- **Always on.** Runs 24/7, freeing staff from answering the same repetitive questions and from manually searching long documents for answers.

Two Main Flows

1. Public Chat Interface

What end users — customers, employees, prospects — interact with directly: a chat panel for asking questions and getting source-backed answers.

2. Secure Admin Knowledge-Base Manager

The back office where authorized staff manage what the chatbot knows, protected behind a login (e.g. `/admin`).

Detailed Functions

Adding and Managing Knowledge

- Manual text entry directly in the admin editor.

- File upload and automatic indexing for PDF, DOC, DOCX, and TXT formats.
- Website ingestion — point Respondo at a site and it can pull in relevant content.
- One-click document deletion to retire outdated information.
- Search and browse tools inside the admin portal to review what's currently indexed.
- Advanced only: consult usage statistics and download conversation logs from the admin portal.

The Chatbot Experience

- Users ask questions from the main chat interface about anything currently in the knowledge base.
- Answers are generated from the most relevant documents retrieved from the vector embedding store — not from the AI model's general/open-web knowledge.
- Citations and source snippets accompany responses, including file names and URLs where available.
- Updating a document in the admin portal keeps the chatbot's answers current — no retraining process required.
- Channel connections available for Facebook Messenger, WhatsApp, a website widget, email, and a voice-over-IP phone line.

How It Works Behind the Scenes

1. Documents are stored and indexed in a vector collection (ChromaDB in Starter, Pinecone in Advanced).
2. When a user asks a question, the system retrieves the most relevant documents from that collection — a technique called Retrieval-Augmented Generation (RAG), which is then used together with an LLM to produce the answer.
3. The question and retrieved documents are sent together to an AI model, which generates a concise, grounded answer.
4. The response is returned to the chat interface along with the document sources it was based on.

This retrieval-augmented approach is what keeps answers "factual" — the model is answering from retrieved evidence rather than free-associating, and it can explicitly say when nothing relevant was found.

Product Versions

Respondo comes in two versions:

Respondo Starter

A simpler, lighter deployment intended as both a demo and a practical option for small organizations.

- Custom logo, color scheme, and language.
- Custom links (e.g., to your website, contact page, or other resources).
- Supports Facebook Messenger, WhatsApp, a web widget, and email.
- Good fit for organizations that want the core Q&A experience without deeper customization.
- Runs on Node.js directly on a Linux server, with ChromaDB for local vector storage.
- Can be deployed on Oracle Cloud's free tier with Gemini's free tier, resulting in zero ongoing infrastructure or AI costs. This setup is suitable for roughly 100 queries per day, though occasional service interruptions

can occur due to Gemini API rate limits.

Respondo Advanced

The fuller build, for organizations that need more than the Starter setup covers.

- **Statistics panel** showing information about conversations and retrievals — usage patterns, what's being asked, and how the knowledge base is being used.
- **Conversation logging** — conversations are logged for later analysis, quality control, and ROI measurement.
- **More robust internal architecture**, built to support growth and spikes in demand:
  - Uses **Pinecone** as the vector database provider (a managed third-party service), rather than the local ChromaDB storage used in Starter.
  - Runs on **Google Cloud**.
  - Server code (Node.js, same as Starter) runs inside **Docker containers**, rather than directly on the host as in Starter — making it easier to scale, update, and recover from failures.
  - No theoretical maximum on usage: both Google Cloud and Pinecone can serve enormous capacity, so the only real limit is the client's budget.

Channel Options and Requirements

Respondo can be reached through several channels, each with different setup requirements:

- **Web widget** — Embedded directly on your website. Requires only that the widget script be added to your site; no external accounts needed. The most self-contained option.
- **Facebook Messenger** — Lets users message your Facebook Page and get answers from Respondo. Requires a Facebook Page and a connected Meta developer app with Messenger permissions approved.
- **WhatsApp** — Lets users message a WhatsApp number and get answers from Respondo. Requires a WhatsApp Business account connected through the Meta (WhatsApp Business) API, including a verified business and phone number.
- **Email** — Respondo receives inbound emails and replies with answers. Requires a dedicated mailbox/address configured to route messages to Respondo (e.g., via IMAP/SMTP or a mail provider's API).
- **Voice / VoIP phone line** — Available in Respondo Advanced. Lets users call in and have a spoken conversation with Respondo. Requires a Twilio account (for the phone line and call handling) and an ElevenLabs account (for voice synthesis), both connected via their APIs.

Because Messenger and WhatsApp both depend on Meta's business platforms, they involve more setup (business verification, app review) than the web widget, which can typically go live as soon as it's added to the site.

Extensions and Customizations

Respondo Advanced can be extended per client request to integrate with enterprise systems. Examples include:

- Exposing Respondo's API so it can be used by an existing web or mobile app.
- Taking specific actions in response to customer inquiries — such as forwarding an email to an employee or opening a support ticket.

These and other integrations are available on request and quoted individually.

Implementing Respondo Advanced: The Bigger Picture

Respondo Advanced can be installed "as is" — the Respondo team handles the technical setup and ongoing maintenance, and the chatbot is live and answering questions from day one. But for organizations looking to get real value out of it, it's worth thinking of the technical installation as one piece of a larger process, not the whole project.

1. Identify the problem

Before adding a chatbot, it's worth confirming and scoping the actual problem: which repetitive queries are consuming staff time, where they're coming from (support inbox, front desk, phone line), and roughly how much volume is involved. This framing makes it possible to tell, later, whether Respondo actually moved the needle.

2. Identify and validate the informational materials

Respondo only answers as well as the material it's given. That means gathering the manuals, policies, FAQs, and other documents that should answer these repetitive queries — and making sure that material is accurate, current, and actually covers the questions being asked. Gaps found during this step are often as valuable as the chatbot itself, since they surface missing or outdated documentation the organization didn't know it had.

3. Designate an owner

It's worth having one person (or a small team) designated as the authority on what goes into the knowledge base — responsible for reviewing new content before it's added, retiring outdated material, and being the point of contact when the chatbot gives an answer that needs correcting. Without a clear owner, knowledge bases tend to drift out of date over time.

4. Measure before, during, and after

To meaningfully measure ROI, you need hard data — not impressions. That means comparing:

- **Before:** how many repetitive questions were coming in, through which channels, and about what topics, prior to deploying Respondo.
- **During:** using Respondo Advanced's statistics panel and conversation logs to see how many conversations are happening, what's actually being asked, whether that information is included in the current materials, and whether questions are being answered satisfactorily.
- **After:** comparing the reduction in repetitive queries reaching human staff, and the time freed up, against the before-state baseline.

5. Consider a human fallback for large operations

For organizations with meaningful volume, it's worth planning for the cases where Respondo doesn't know the answer. Rather than leaving the user without a resolution, the conversation can be handed off to a human agent — either **asynchronously** (e.g., forwarding the unanswered question to a support inbox for a later reply) or **live** (e.g., escalating to a human-staffed chat interface in real time). This keeps the chatbot from being a dead end for the questions it can't yet answer, and the pattern of what gets escalated becomes useful input for improving the knowledge base over time.

Key Features Summary

Feature	Description
Secure admin login	Restricts knowledge base management to authorized users
Manual + file-based content entry	Type directly, or upload PDF/DOC/DOCX/TXT
Website ingestion	Pull content directly from an existing website
Source-aware answers	Every answer can cite its originating document/URL
Dynamic starter questions	Helps users begin a conversation quickly
Local, deployment-ready architecture	No dependency on external SaaS platforms
Editable, deletable knowledge base	Keep information current over time
Multi-language support	Not limited to a single language
Multi-channel delivery	Web widget, Facebook Messenger, WhatsApp, email
Ownership model	One-time setup/maintenance fee, not per-seat SaaS pricing

Business Model

Respondo is explicitly **not sold as SaaS**. The commercial model is:

- Pay for initial setup, and separately for upgrades and ongoing maintenance.
- No per-user or per-conversation licensing fee to the developer.
- Unlimited users and unlimited conversations — the variable ongoing costs are usage-based billing from whichever AI provider (LLM) you choose to connect, plus your cloud/server hosting costs.
- Because it runs on your own cloud/server, you retain full control over hosting costs and data residency.
- Further customization (additional channels, integrations, branding) is available on request.

Licensing: Respondo is source-available. When a client buys the initial code and implementation, they aren't locked into an ongoing relationship with the developer — from that point on, they own the code and can modify it as they see fit for their own use case, whether the modifications are made by the original developer or by anyone else the client chooses to work with. The one restriction is that clients may not resell or repackage their modified version of Respondo in competition with Respondo itself.

Implementation: The Respondo team handles the complete implementation end to end — setting up the cloud environment, connecting the LLM and other services, configuration, channel integrations, and ongoing periodic maintenance, customizations, and upgrades. Clients aren't left to piece together the infrastructure themselves.

Pricing and availability: Respondo Starter can always be purchased directly from the Upwork listing at its published price and conditions. For Respondo Advanced — including current pricing, timelines, and availability

for implementation, customizations, and integrations — please contact Sergio Flores directly (see Links & References below).

The reference package (via Upwork) offers a starter Q&A chatbot — trained on the client's own data with a web interface, admin panel, cloud deployment, custom branding, and bug fixes included — with additional messaging channels (email, WhatsApp, Messenger, or embeddable web widget) available as an add-on.

Technical Details (Developer-Level)

Architecture pattern: Retrieval-Augmented Generation (RAG)

- **Storage layer:** Uploaded and indexed documents are stored in a local vector database — specifically ChromaDB — as embeddings, enabling semantic (meaning-based) search rather than plain keyword matching.
- **Ingestion pipeline:**
  - Text is extracted from uploaded PDF, DOC, DOCX, and TXT files.
  - Manually entered text is captured directly through the admin editor.
  - Website content can be crawled/scraped and ingested into the same pipeline.
  - Extracted content is chunked and embedded, then written to the vector collection.
- **Query pipeline:**
  1. User submits a question via the chat interface.
  2. The question is embedded and used to perform a similarity search against the vector store to retrieve the most relevant document chunks.
  3. The retrieved chunks, along with the original question, are passed as context to an AI/LLM completion call.
  4. The model generates a concise answer grounded in the retrieved context.
  5. The response is returned to the frontend along with citation metadata (source file name, URL where applicable) so the source can be displayed alongside the answer.
- **Grounding / anti-hallucination behavior:** The system is designed so that if no sufficiently relevant content is retrieved, the model is expected to tell the user the answer isn't available in the knowledge base rather than generating an unsupported answer.
- **Access control:** The admin portal (e.g., `/admin`) sits behind authentication, separating read-only public chat access from write access (upload/edit/delete) to the knowledge base.
- **Deployment model:** Designed for self-hosted, "local deployment-ready" architecture — runs on a cloud server/provider of the client's choosing rather than a shared multi-tenant SaaS backend. This means:
  - No dependency on a third-party chatbot SaaS platform.
  - The AI/LLM provider is pluggable/configurable (bring-your-own-provider). By default this means a commercial LLM such as Gemini or Claude. Running a GPU-cloud-hosted open-source model instead is also possible, though it's not the standard setup — it requires a small amount of customization. Performance with an open-source model won't match a commercial model, but the trade-off is that conversation data never needs to be sent to any third-party provider.
  - Infrastructure can start small (e.g., a low-cost or free-tier server) and scale up as usage grows.

- **Integration surface:** Web chat widget for direct website embedding, plus optional connectors for Facebook Messenger, WhatsApp, and email as add-on channels, with a voice/VoIP phone line (via Twilio and ElevenLabs) available in Advanced.
- **Development stack:** Built with Node.js; delivery has referenced orchestration tooling such as LangChain for more advanced customization, with support for deploying open-source models (e.g., via Ollama) on cloud GPU providers as an alternative to hosted commercial APIs.

Links & References

- **Buy Respondo Starter (Upwork):** https://dub.sh/upwrkrespnd
- **Live demo:** https://dub.sh/jcqoK6A
- **Respondo website:** https://saft.industries/respondo
- **Creator (Sergio Flores):** X: https://x.com/IamSergioMX | LinkedIn: https://linkedin.com/in/saft