# Case Study: Launching a Custom Hybrid LLM Chatbot Platform (2025)

Background:

In today's rapidly shifting business landscape, organizations in Arkansas and across the country face a critical crossroads: **those who innovate and adopt artificial intelligence will lead, while those who fall behind risk being overtaken by larger competitors and more agile, tech-forward peers.** Despite this, Arkansas ranks startlingly low in actual AI adoption, a challenge that Enterprise Pursuits was founded to solve. Our mission is clear: make world-class no-code, low-code, & AI tools accessible for local businesses, SMBs, and nonprofits, equipping them to compete at a national level and deliver heightened value for their clients and constituents. For nonprofits, AI adoption can streamline operations, enhance donor engagement, and improve service delivery, ultimately leading to greater impact and sustainability.

We've journeyed through every major roadblock common to "DIY" AI deployments, misleading open-source promises, paid API traps, SSL certificate quirks, platform bugs, and frustrating hardware minimums. Rather than settling for limited "retrieval bots" masquerading as AI, we built and selected intelligent solutions that leverage cutting-edge models like Ollama3, fully open-source frameworks, and secure AWS cloud infrastructure, prioritizing both capability and cost efficiency. Our systems enable real cognitive chat experiences, handling complex questions, integrating custom business knowledge, and learning over time from new uploads. And all of this is managed securely in the backend by our experienced consulting team, who are always available to provide support and guidance.
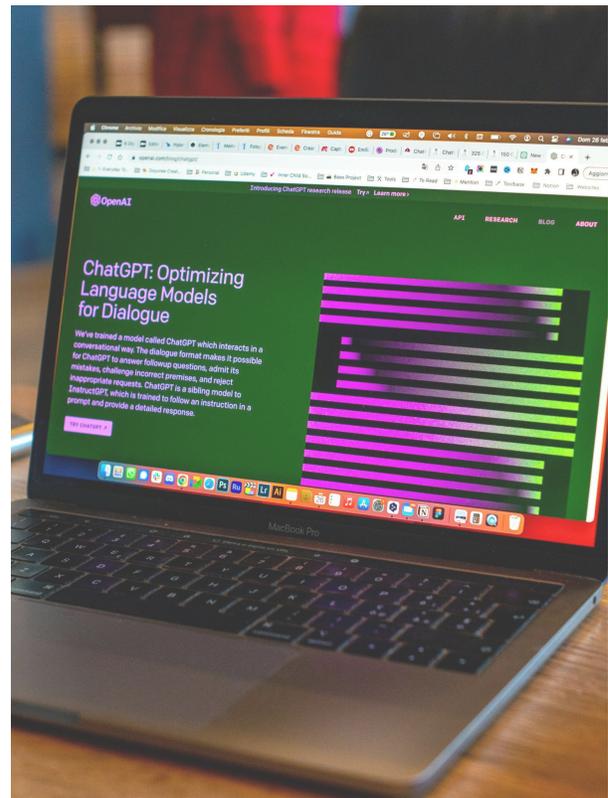
Enterprise Pursuits

Unlike many off-the-shelf vendors, Enterprise Pursuits goes beyond a quick installation. We design, optimize, and maintain AI chatbots tailored to your unique needs, whether you're engaging with clients, training staff, or supporting internal operations. Our approach is rooted in a long-term partnership, offering project-based setup fees that fit your budget, paired with monthly consulting support to keep your AI secure, updated, and ahead of the curve. We ensure you own your data, control your environment, and maximize every ounce of value AI brings to your organization, avoiding vendor lock-in and hidden recurring costs.

As business leaders and nonprofit executives look to the future, the message is simple: **AI is no longer just a buzzword, it's a competitive necessity.** With Enterprise Pursuits as your local and remote open source AI partner, your organization gains the power of genuine intelligence, not empty marketing claims. We understand the unique challenges and opportunities of the Arkansas business community, and we're committed to helping you future-proof your business, deliver transformative client experiences, and propel Arkansas's business community to the front of the national AI revolution.

Ask Enterprise Pursuits AI

You: Tell me in one sentence that this website is for?

This website, Enterprise Pursuits, is an Organizational Change Management boutique consulting firm that specializes in digital transformations for small to medium-sized businesses (SMB) and nonprofit organizations.

Type your question...   Ask

Challenges:

- Standard AnythingLLM AMIs lacked turnkey API endpoints for direct connection to public website widgets.
- Early Docker/Open WebUI experiments on AWS EC2 encountered persistent Nextgen SSL certificate variable bugs, preventing models from loading reliably.
- Useful models could not run on minimum documented specifics and needed at least a t3.xl instance to function properly albeit a bit slow for production usage.

- The firm required a chat backend that could handle chat requests globally, parse uploaded files, and allow for no-code/low-code "agent" upgrades (custom skills, document ingestion).

One of the most significant hurdles encountered during this build arose when attempting to deploy AnythingLLM, a popular open-source LLM solution that promises out-of-the-box document retrieval and API integration. On the surface, AnythingLLM functioned flawlessly as a local application, allowing users to upload documents, interact with the model, and enjoy a clean interface. The real challenge surfaced when we attempted to integrate AnythingLLM with our website chatbot via its advertised REST API. Despite spending hours reviewing the configuration, inspecting backend logs, and troubleshooting connections with various low-code tools, it became clear that the public API was nonfunctional, a "spoof" that appeared legitimate but never established true connectivity. The underlying documentation confirmed our findings: the open-source version disables remote API calls by design, restricting all interactions to local or on-server usage only and foreclosing any practical option for external chatbot embedding or scalable cloud deployment.

During the Open WebUI implementation, we encountered a critical and time-consuming compatibility issue with the Nextgen SSL certificate. Initially, the deployment appeared structurally sound, but models stubbornly refused to load within the WebUI interface, stalling development for hours. After exhaustive troubleshooting, which included examining system logs, revisiting network settings, and testing container variables, it was discovered that the Nextgen SSL certificate introduced handshake and trust errors in the backend, effectively preventing Open WebUI from communicating securely with the Ollama model server. The WebUI failed to establish a secure connection, resulting in opaque errors and an unusable chat environment.
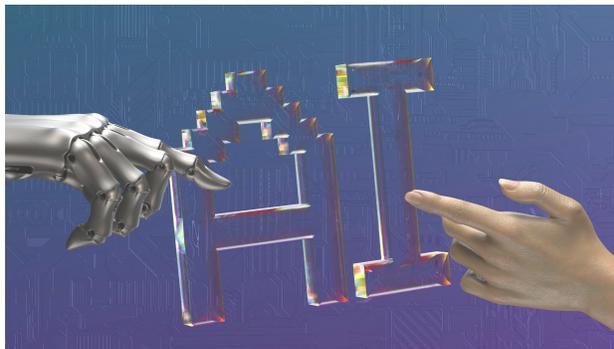
It became clear that Nextgen's SSL implementation for AWS environments was injecting certificate attributes and intermediate authorities that were not recognized or trusted by the LLM stack. This misconfiguration broke the model auto-discovery and failed every attempt to serve up chat completions. The solution came from switching away from NextGen to a more standardized Sucuri's free, managed SSL service certificate, which immediately restored handshake integrity and allowed secure traffic for both the Web UI and backend services. Once the certificate authority was corrected, model loading worked flawlessly, and chatbot operation returned to normal. For clients concerned about obscure cloud deployment errors that could derail their projects, Enterprise Pursuits isolates and resolves these advanced technical conflicts, delivering model connectivity, robust security, and seamless UI operations without wasted hours.

Despite what the official documentation indicated, we quickly learned that applicable language models would not run on the minimum recommended AWS instance types; such as t3.micro, t3.small, or t3.large, due to severe memory and compute constraints. After multiple attempts and extensive troubleshooting, it became clear that a minimum of a t3.xl instance was required to load and interact with models reliably, even if performance was marginal for production settings. On this larger instance type, we achieved basic functionality and demonstrated chatbot responses; however, the system remained slower compared to more robust hardware configurations. For organizations considering real-world chatbots and document parsing on self-hosted infrastructure, it's essential to budget for appropriately sized cloud resources. Enterprise Pursuits accelerates deployments by matching business needs with the correct instance specifications, saving clients time, money, and headaches when scaling up LLM-powered solutions.

For business owners seeking to implement a reliable LLM-powered chatbot for their website, this distinction is critical. Many open-source LLM tools, while excellent for local experimentation, do not expose a real API suitable for cloud-based or multi-user use, essentially limiting their value to internal testing environments. As documented by AnythingLLM's own maintainers, the open-source edition is specifically designed to handle only local invocations from the desktop or server, rather than public web apps or commercial frontends. Such limitations can quickly bring ambitious AI chatbot projects

to a halt before they even launch. Enterprise Pursuits has the expertise to cut through these pitfalls, leveraging enterprise-ready cloud solutions with verifiable API access and proven extensibility. Entrusting your implementation to us ensures that your chatbot is robust, scalable, and most importantly, actually works from anywhere your business needs it.

Solution Steps:

Cloud Infrastructure:

Rather than purchasing paid "enterprise" AMI solutions like IronCloud LLM, which would have added $0.95 per hour on top of ongoing AWS server fees; Enterprise Pursuits prioritized truly free, open-source options for maximum client cost savings. We shifted from troubleshooting Docker containers to deploying the latest Open WebUI release directly on AWS EC2, leveraging community best practices and resourceful server setups.

WebUI and API Integration:

We deployed Open WebUI as our primary interface, taking advantage of its free, browser-based chat system and flexible configuration for private, on-premises use. Unlike paid options, Open WebUI required more technical configuration, managing network security and server permissions; but allowed us to maintain full control without recurring licensing costs. Our approach ensured that SMBs and nonprofit organizations could access LLM capabilities without budgetary constraints.

Frontend Chatbot Implementation:

Our team developed a lightweight HTML/JS chatbot widget, with pure JavaScript for fast, direct client integration, or optional support for low-code platforms like Bubble and Retool for rapid prototyping. The widget communicated securely with our self-hosted WebUI instance, sending/receiving user prompts so clients could access AI-powered responses directly from their own domains.

javascript

```javascript
async function sendChat(userMessage) {
  const response = await fetch('https://your-webui-url/chat', {
    method: 'POST',
    headers: { 'Authorization': 'Bearer XXXXXXXX',
'Content-Type': 'application/json' },
    body: JSON.stringify({ message: userMessage })
  });
  const data = await response.json();
  displayChatReply(data.reply);
}
```

Advanced Capabilities:

As LLM consultants specializing in SMBs and nonprofits, we recognize the importance of making document upload and internal knowledge retrieval seamless for custom chatbots. In the Open WebUI ecosystem (and similar free/open-source self-hosted platforms), document upload and analysis are only available in deployments where specific plugins, such as the "RAG" (retrieval-augmented generation) plugin or integrations like "File-Chat" and
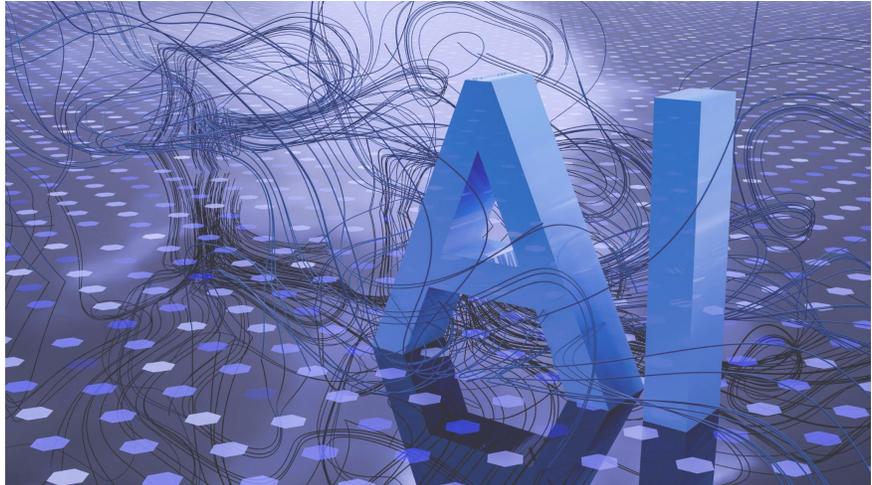
"Document QA", have been installed and properly configured. For basic deployments or those using the stock WebUI release, the capability may not be active by default due to security design, limited API support, or backend storage restrictions. For instance, uploading emails, Q&A documents, and training materials requires turning on the RAG plugin or a compatible "knowledge base" module, and then batch ingesting these files using the administrative or developer dashboard.
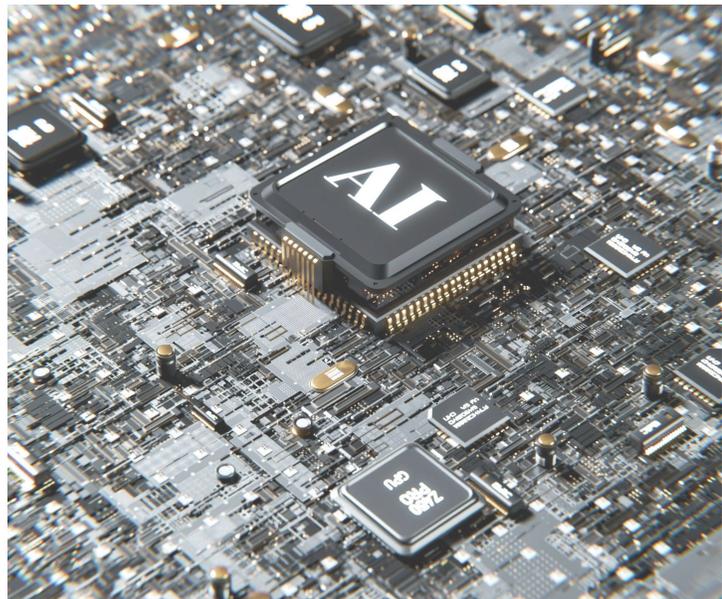
Enterprise Pursuits

Where backend or environment constraints block direct file upload, such as locked-down server storage, API-disabled WebUI images, or frontend widgets lacking file input Enterprise Pursuits solves this by manually enabling advanced plugins, configuring file storage paths, and mapping input sources so the AI chatbot can ingest and reference client files dynamically. We either extend the WebUI with open community modules that support document QA or build bridging scripts that convert client data (emails, PDFs, docs) into formats ingested by the LLM's internal database. This means that even organizations with the tightest budgets can train and deploy chatbots that answer internal, context-rich business questions, turning their unique documents into a living knowledge resource, regardless of the limitations out of the box.

We configured integrations for document upload when needed, streamlined business workflows, and enabled basic compliance/privacy controls all hosted on the client's own AWS account and protected inside their VPC. While Open WebUI was less "point-and-click" for business plugins, our custom engineering ensured essential features like contextual chat and file-based retrieval, without paid "agent" or plugin requirements.

Model:

For this project, Enterprise Pursuits selected Ollama3 as the foundational AI model powering our custom chatbot solution. Unlike OpenAI's GPT and similar cloud products, which offer only limited free use before shifting users to paid plans and API restrictions, Ollama3 is completely free for unlimited self-hosted deployments. Functionally, Ollama3 delivers an experience on par with ChatGPT or Perplexity, capable of natural dialogue, deep document understanding, and precise Q&A, making it an ideal choice for SMBs and nonprofits looking to maximize value while avoiding recurring costs.

Enterprise Pursuits

We leveraged Ollama3 to build an advanced internal chatbot with our simple HTML/JavaScript web interface, enabling users to:

- Ask natural language questions and get actionable, context-aware responses.
- Upload business emails, training documents, Q&A guides, and internal workflows for instant retrieval and analysis in the backend.
- Train the bot by expanding its document library or updating knowledge links whenever processes or information evolve.
- Securely manage knowledge without transmitting proprietary data to outside providers.
- Integrate custom agent skills, plugins, and low-code features as needed for growth.

Under the hood, Ollama3 ingests client documents as vectorized knowledge sources, cross-references them intelligently, and uses its transformer engine to respond with highly relevant answers, just like mainstream models. The system is open for further "training" at any time, simply by uploading new files or refreshing indexes, no expensive ML pipelines or external subscriptions needed.

For teams needing specialized assistant capabilities, there are other free models, such as CodeLlama or StableCode. These may be less powerful in chat/Q&A but are excellent for building coding assistants, developer helpdesks, or technical support bots. Our flexible infrastructure enables us to match every client with the right free model, whether for business chat, document search, or programmer support. We always focus on low- or no-cost deployment that drives real results.

While our chatbot, powered by Ollama3, can deliver answers from internal business documents and emails, these files must first be ingested, indexed, and made searchable on the server by an authorized user. End users interacting with the chat interface cannot upload documents themselves for immediate inclusion in the bot's knowledge base. Instead, our process involves:

- Curating documents (emails, training guides, FAQs, PDFs) in advance
- Admins securely upload and register those files on the server, where the model's RAG (retrieval-augmented generation) or knowledge base plugin processes them.
- The chatbot then accesses these sources for context and provides answers without requiring direct file uploads from chat users.
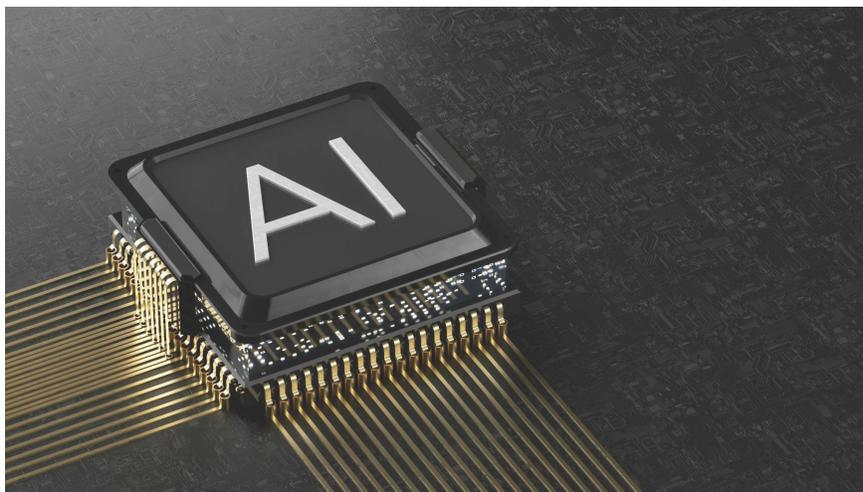
This design ensures data security, maintains model accuracy, and gives organizations control over what proprietary and sensitive information is added to the AI's memory. If your business requires new documents, please get in touch with your system administrator or Enterprise Pursuits. We handle backend uploads and retraining so your AI remains up-to-date, accurate, and secure.

Outcome:
Clients now benefit from 24/7 web-based AI help, cost-free for unlimited usage, with direct answers and business document support. Enterprise Pursuits delivered a scalable, upgradable "no-code" backend for AI chat, allowing small businesses and nonprofits to save on software costs and redirect their investments to core mission activities.

Lessons Learned & Our Value Proposition:
Through the course of building and deploying a custom AI chatbot using Ollama3 and Open WebUI, we have distilled several essential lessons that position Enterprise Pursuits as Arkansas's and the region's go-to provider for open-source, fully featured AI chat solutions for SMBs and nonprofits. First and foremost, we found that many products labeled "AI" are little more than glorified document retrievers or search bots, offering canned answers with no intelligence, context, or dynamic learning capabilities. Our standard surpasses these: we deliver real AI experiences capable of nuanced, conversational answers, contextual knowledge

retrieval, and continual improvement, all while keeping your proprietary data secure and your costs significantly lower than other commercial options.

Our choice of Ollama3, an unrestricted platform, and Open WebUI, a powerful and flexible platform, allows us to offer cost-effective solutions. Unlike other commercial options that may appear "free" but quickly lock out users or demand expensive subscriptions, our solutions bypass licensing fees, usage restrictions, and vendor lock-in. The backend handles secure ingestion of emails, internal Q&A, training manuals, workflows, and more, transforming your business documents into a living knowledge source. While document uploads and model training are managed securely by administrators, end users enjoy a seamless chat experience with enterprise-level capabilities, all through a simple interface tailored to your needs.

At Enterprise Pursuits, we offer transparent, project-based fees for setup and integration, as well as monthly consultation packages. This ensures that you receive continuous support, security updates, and upgrades as your needs evolve. Our solutions empower local Arkansas businesses and remote organizations alike with true AI, going beyond basic retrieval to deliver strategic insight, personalized answers, and lasting business value. If you're seeking an affordable, secure, and future-proof chatbot solution, Enterprise Pursuits is your dedicated open-source AI specialist.