



JUSTIN BROWN

Local Small Language Model AI Assistant

Secure, Cost-Efficient Agentic Assistance

GOAL: POC of a local .NET-based SLM Assistant aiding with real-world business needs

This application is a .NET-based local AI agent designed to operate in a closed system without requiring internet access. The application acts as a recommendation and assistant tool for technical consulting firms, enabling natural language queries about projects and resources while providing intelligent recommendations; however, applications of a similar architecture could be used across the board and in many domains - even in privacy-sensitive environments, such as legal or healthcare organizations.

Using only native NuGet packages, the agent leverages Small Language Models (SLMs) to provide actionable recommendations and easy-to-use functionality using only the hardware on a standard developer laptop. Built for cost efficiency and security, this architecture ensures all data stays local and eliminates reliance on costly third-party AI services.

Highlights

- Purely .NET Stack
- Small Language Model
- Secure Local Deployment
- Cost-efficient and Privacy-first
- Clean, intuitive UI

★ Key Features

- Fully offline AI agent
- CPU-only execution with SLMs
- Router logic for intent detection
- RAG for context-aware answers
- Streamlined, modular .NET architecture for easy customization

🏢 Applicable Use Cases

- Delivering AI functionality in cost-sensitive contexts
- Navigating internal knowledge bases
- Assisting with project management
- Matching resources to projects
- Querying secure data in healthcare, legal, or other regulated environments



LET'S CONNECT! Follow up with us at Allab@nvisia.com or visit the website:



Explore what happens when engineers get curious

The **AI Lab** is where nvidia's engineers and our clients experiment, explore, and build with the latest in AI and emerging tech. See real-world demos from our lab — not just ideas but functional innovations in:

- AI-powered software solutions
- Intelligent automation
- Data science & machine learning
- Custom AI integrations



Engineering the Future with AI

Strategic Focus Areas

↳ Frameworks & Playbooks

Codifying reusable architectures and best practices for consistent, scalable AI delivery.

↳ Security & Compliance

Embedding data security, model integrity, regulatory alignment, and safe prompt engineering into every stage of the AI lifecycle.

↳ Human-in-the-Loop AI

Designing workflows where AI accelerates analysis and automation, while human oversight and expert review ensure accuracy, reliability, and real-world relevance.

↳ Responsible AI & Ethics

Creating principles, reviews, and tooling that ensure AI is transparent, fair, and aligned with organizational values.

↳ Driving Value

Prioritizing high-impact opportunities, measuring ROI, and ensuring AI initiatives deliver meaningful business outcomes.

↳ Improving Predictability

Using techniques such as Spec-Driven Development to enhance model accuracy, reproducibility, and trustworthiness.

Come Talk to Us

Ask questions, see the tech in action, or just chat with our team. We're here to share what we've learned — and to learn from you too!



Real demos



Ready to scale



Built by engineers

LET'S CONNECT! Follow up with us at AILab@nvidia.com or visit the website:

