Fine-Tuned, Low-Power and RAG(ged)

Sustainable, privacy-preserving AI language models for libraries using minimal hardware and Retrieval Augmented Generation (RAG)

David Meincke, Johnson & Wales University

QR Code to project website: github.com/drmein/acrl2025_llm_poster

The Challenge

- Students expect conversational interfaces, not traditional FAQs[1]
- Critical assistance needed when librarians aren't available
- Traditional FAQ searching can be awkward and ineffective

Library Impact: Users often abandon traditional FAQ searches without finding answers, leading to increased staff workload and user frustration.

Our Solution: Working RAG-Enhanced Local Model

Project Timeline:

Nov 2024	Feb 2025	Now	Summer 2025
Project Start	Backend Complete	Testing	Public Launch

Surprising Discovery!
RAG significantly outperformed fine-tuning for library FAQs, providing more accurate, specific responses while requiring less training resources.[2]

Built with real library knowledge:

- 500+ Q&A pairs from public FAQs and to answer common themes that occurred in chat transcripts
- Library-specific content (databases, services, resources)
- Easily maintainable by non-coding librarians via spreadsheets

Technologies:

Python · LlamaIndex · ChromaDB · Raspberry Pi

Running on minimal hardware: Raspberry Pi 5 (8GB)[4]

Key Benefits:

- Complete privacy control
- 80% less energy than cloud[3]
- Full control over knowledge
- Runs on affordable hardware

The Farm Metaphor

Implementing AI in libraries presents a choice:

Local AI: Sustainable Family Farm
- Lower but sufficient yield
- Environmentally friendly
- Self-sufficient & locally controlled

Cloud AI: Industrial Factory Farm
- Higher output
- Resource-intensive
- External dependencies & control

Core Values for Library AI

Keep patron data on your hardware · 80% less energy than cloud models · Library control over AI decisions

ARL Principle: "Libraries believe 'no human, no AI.' This principle underscores the importance of human involvement in critical decision-making junctures."
Association of Research Libraries, April 2024

[1]LIBRARYASSESSMENT.ORG. "University Library FAQ Usage Analysis." 2023.
[2]Experimental findings, Meincke. 2025.
[3]Alibaba Cloud. "Optimizing Energy Efficiency in AI Models." 2024.
[4]ITSFOSS.COM. "Running 9 Popular LLMs on Raspberry Pi 5." 2023.

AI Implementation Options for Libraries

Balance privacy, quality, and sustainability based on your needs

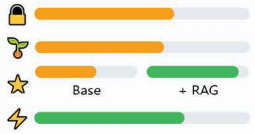
Small Local Models

- Complete privacy & control
- Minimal energy use[7]
- Very slow (15-60+ sec)
- SmolLM2, TinyLlama[6]
- RAG significantly improves quality[5]

Cost: $50-150 one-time, then just electricity

Medium Setup

Medium Cloud-Hosted Models



- Best overall balance
- Good privacy & response quality
- Moderate energy use
- Moderate speed (1-3 sec)
- Highest quality with RAG

Cost: $0.10-10/month based on usage

Medium Setup

Large Cloud API Models

- Highest base quality
- Still improves with RAG
- Fast responses (0.8-1.5 sec)
- Higher ecological impact
- GPT-4o, Claude, Mistral Large[8]

Cost: $0.10-10/month based on usage

Easy Setup

Privacy & Control · Sustainability · Response Quality · Response Speed · Low — High

[5]Soudani et al. "Fine Tuning vs. Retrieval Augmented Generation for Less Popular Knowledge." ACM SIGIR Conference 2024. [6]Ben Allal et al. "SmolLM – Blazingly Fast and Remarkably Powerful." HuggingFace Blog, 2024. [7]Alibaba Cloud. "Optimizing Energy Efficiency in AI Models." 2024. [8]OpenAI. "GPT-4 Technical Report." 2023.

See GitHub repository for full implementation guide

All About RAG

RAG (Retrieval-Augmented Generation) can connects AI models with your library's knowledge, enabling accurate, specific answers even with smaller models and is especially useful for conversational FAQs.[1]

Example: "How do I reserve a study room?"

Without RAG:
"...visit the library website... go to the 'Reserve a Study Room' tab... browse available rooms, check availability, and make reservations... contact... Ask a Librarian..." (General Guidance)

With RAG:
"...visit the library homepage... select the 'Study Rooms' tab... Reserve online for Downcity [link] or Harborside [link]... limited to 2 hours per day... confirm via JWU email..." (Specific Links & Details)

Without RAG (Base Model)
Relies only on pre-trained data and context from the prompt and chat. Answers may be generic, outdated, or incorrect.

With RAG (+ Library Info)
Accesses library's specific info. Answers are more accurate, detailed, and up-to-date.[2]

[1]Soudani et al. "Fine Tuning vs. Retrieval Augmented Generation for Less Popular Knowledge." ACM SIGIR Conference 2024. [2]Experimental findings, Meincke. 2025.

Key Concepts

Base Model
Pre-trained AI system that understands and generates human language
General knowledge before specialization. SmolLM-2 (small) vs. Mistral-7B (medium)

Local vs Cloud Models
Where AI processing happens: on-site or on remote servers
Local: like on-premise systems. Cloud: like subscription services

RAG
Retrieval-Augmented Generation: enhances AI responses with your specific documents
Like checking reference materials before answering questions

Parameters
Measure of AI model's knowledge capacity and complexity
SmolLM-2 (320M): compact collection vs. GPT-4 (1T+): massive archive

Next Steps: Sustainable AI for Your Library

1. Explore
Scan QR code for links to tools like Ollama and other frameworks for exploring local LLMs on your own computer

2. Prepare
Access our scripts and methods for organizing your library's knowledge base into retrieval-ready format

3. Implement
Use our shared code and implementation guide to build your own sustainable AI system

All resources freely available - take a handout!
github.com/drmein/acrl2025_llm_poster