Webb Local AI

PRIVATE AI FOR PUBLIC DATA

GLAMDRING

A Modular Expert AI Architecture for Secure, Local Deployment

WHITE PAPER

Version 1.0 • December 27, 2025

Webb Local AI, LLC

Private AI for Public Data

Executive Summary

Glamdring is an innovative AI architecture designed for secure, private, and efficient deployment on consumer-grade hardware. It combines a lightweight, rule-based routing layer with a collection of compact, specialized expert models to deliver expansive AI capabilities without the resource demands of large monolithic language models.

By activating only the most relevant expert model for each query—and limiting concurrent inference to a single quantized model (typically 7–12 billion parameters)—Glamdring achieves high-performance reasoning, tool integration, and multimodal features while operating entirely within constrained environments. This makes it ideal for privacy-sensitive applications, such as local government services, where data must remain on-premises and operational costs must remain low.

Glamdring represents a practical implementation of mixture-of-experts principles optimized for local execution, offering scalability, transparency, and ethical safeguards that align with public-sector requirements.

Introduction

The rapid advancement of large language models has demonstrated remarkable general intelligence, but their deployment in sensitive or resource-constrained environments remains challenging. High computational requirements, cloud dependency, data privacy risks, and lack of domain specialization limit their adoption in sectors like local government.

Glamdring addresses these limitations through a modular, locally deployable architecture that prioritizes:

- **Privacy and Sovereignty:** All processing occurs on client-controlled hardware.
- **Efficiency:** Minimal resource footprint suitable for consumer or small-server hardware.
- **Specialization:** Targeted performance via domain-specific expert models.
- **Transparency:** Deterministic routing for explainable behavior.

Named after the legendary sword from Tolkien's works—known for its ability to detect and respond to specific threats—Glamdring intelligently "detects" the nature of a query and directs it to the appropriate expert, ensuring precise and efficient responses.

Problem Statement

Traditional AI deployment models present several barriers for local, privacy-critical use cases:

1. **Resource Intensity:** Modern frontier models require high-end GPUs and significant memory, making on-premises deployment prohibitively expensive for small organizations.
2. **Privacy Exposure:** Cloud-based solutions transmit sensitive data to third-party servers, creating compliance and security risks.
3. **Generalization Trade-offs:** Large general-purpose models excel at breadth but often underperform in specialized domains and can produce inconsistent or hallucinated outputs in narrow contexts.
4. **Opacity:** Black-box inference complicates auditing and trust, particularly in public-sector applications.
5. **Scalability Limits:** Adding new capabilities to monolithic models requires retraining or fine-tuning at massive scale.

These constraints have left many local governments unable to benefit from AI advancements, perpetuating reliance on manual processes and outdated tools.

The Glamdring Solution

Core Architecture

Glamdring employs a two-tier design:

1. **Routing Layer**

   A lightweight, rule-based orchestration system analyzes incoming queries and selects the most appropriate downstream expert model. This layer uses deterministic logic—combining keyword matching, intent classification, and contextual signals—to ensure transparent, auditable routing decisions. No large neural inference is required at this stage, keeping overhead minimal.

2. **Expert Model Layer**

   A modular library of compact, quantized language models (7–12 billion parameters), each optimized for specific domains or task types. Examples include: administrative and regulatory compliance, financial analysis and budgeting, educational planning, code generation and technical support, and general reasoning (fallback). Only one expert model is loaded and active at a time, dramatically reducing memory and power requirements.

This approach enables an expansive effective capability surface while maintaining the resource profile of a single small-to-medium model.

Key Technical Advantages

- **Quantized Inference:** Models are aggressively quantized (typically to 4-bit or 5-bit precision) with minimal accuracy degradation, enabling efficient execution on CPUs or modest GPUs.
- **Single-Model Concurrency:** By design, only one expert processes a query, eliminating the memory bloat associated with parallel expert activation.
- **Horizontal Scalability:** New expert models can be added independently without retraining the router or existing experts.
- **Tool Integration:** Experts interface with safe, sandboxed tools (code execution, document parsing, controlled web access) to extend functionality beyond pure generation.
- **Local Execution:** The entire stack runs offline after initial setup, ensuring zero data egress.

User Interface

The reference implementation features a richly themed, web-based chat interface that supports: real-time streaming responses, multimodal input (text, documents, images), voice interaction, code block rendering with syntax highlighting, optional thinking trace disclosure, and image generation via integrated diffusion models. The interface is designed for intuitive use by non-technical staff while providing advanced features for power users.

Technical Overview

Routing Mechanism

The routing layer evaluates queries against a configurable rule set. Rules may include: keyword presence/absence, topic categorization, user-specified overrides, and query complexity heuristics. This deterministic approach ensures consistent behavior, facilitates debugging, and supports regulatory auditing—advantages over learned routers that introduce non-determinism.

Model Management

- Models are stored in compressed, quantized formats.
- Loading is lazy and exclusive: when a new expert is selected, the previous one is unloaded from memory.
- Context is preserved across switches via a shared conversation history buffer.

Performance Profile

On typical consumer hardware (e.g., modern laptop with 32 GB RAM and optional mid-range GPU):

- Peak memory usage: ~12–18 GB during inference
- Response latency: Comparable to local 7–13B class models
- Power consumption: Suitable for sustained desktop or small-server operation

These metrics enable deployment on existing municipal IT infrastructure without dedicated AI hardware budgets.

Security and Ethical Design

Glamdring embeds privacy and ethics by design:

- **Zero Data Exfiltration:** No telemetry or logging is sent externally.
- **On-Premises Only:** All components run locally; no cloud dependency.
- **Prohibited Practices:** Architecture explicitly avoids techniques associated with psychological manipulation or dark patterns.
- **Auditability:** Routing decisions and model selections are logged for review.
- **Controlled Tool Access:** External integrations (when enabled) operate within strict sandboxes.

These features align with public-sector requirements for transparency, accountability, and citizen trust.

Use Cases in Local Government

Glamdring's modular design excels in diverse municipal applications:

1. **Citizen Services:** Route inquiries about permits, taxes, or utilities to specialized administrative experts.
2. **Budget and Finance:** Direct financial queries to models trained on governmental accounting standards.
3. **Compliance and Legal:** Use regulatory-focused experts for policy interpretation and document review.
4. **Education Administration:** Support school districts with curriculum planning and compliance tools.
5. **Custom Extensions:** Deploy client-specific experts for unique local needs (e.g., water management, public works scheduling).

Conclusion

Glamdring redefines what is possible with locally deployed AI. By combining intelligent routing with specialized, efficient expert models, it delivers broad capability within strict resource and privacy constraints. This architecture enables small and medium-sized organizations—particularly local governments—to adopt powerful, secure AI tools without compromising sovereignty or budget.

As AI continues to evolve, Glamdring provides a future-proof foundation: extensible, transparent, and firmly under the control of its operators.

Webb Local AI, LLC is committed to advancing Glamdring as the premier platform for autonomous, ethical AI in the public sector.

For technical discussions or collaboration opportunities, please contact Webb Local AI.

Disclaimer: *Certain implementation details of the Glamdring architecture are proprietary and not fully disclosed in this document. All rights reserved.*