

Reg

The In-Boundary AI Companion for the Defense Industrial Base

Self-hosted · CUI/ITAR-safe · U.S.-origin models · Rickert-Areno Engineering, LLC



Reg is a fully self-hosted AI companion for organizations that cannot send their work to the cloud. It delivers frontier-class AI — conversational assistance, document workflows, and a complete FPGA/VHDL engineering toolchain — entirely inside the customer's security boundary, running exclusively on U.S.-origin models. Controlled Unclassified Information (CUI) and ITAR-controlled data never leave the enclave, so the Defense Industrial Base gets the productivity of modern AI without the compliance exposure that rules commercial assistants out. Reg is RAE's own production system — it runs the company's daily engineering work today.

The Problem

Commercial AI assistants process every prompt in a vendor cloud, placing CUI and ITAR/EAR-controlled data outside the customer's control and outside the compliance boundary. For the Defense Industrial Base this is disqualifying: NIST SP 800-171 and CMMC obligations, together with ITAR export rules, prohibit exposing controlled technical data to uncleared systems or foreign persons. At the same time, many of the strongest open models originate from foreign — notably PRC — developers, raising provenance concerns of their own. The organizations with the most demanding engineering problems are the ones effectively locked out of modern AI.

How Reg Works

Reg runs entirely on hardware the customer owns, air-gapped from any cloud service, and inverts the trust model of commercial AI: the data stays put and the models come to it.

- **Air-gapped inference.** All models run in-house on the customer's GPUs. No prompt, document, or design ever reaches an external service.
- **U.S.-origin models only.** Every resident model is from a U.S. developer (IBM, Microsoft, Snowflake, and U.S. open-weight releases). No PRC-origin weights are deployed, even air-gapped — a clean provenance posture for defense work.
- **Tiered escalation ladder.** A fast local router handles everyday chat and tool use, escalating the hardest reasoning to progressively larger in-boundary models. Only explicitly non-CUI work is ever eligible to reach a cloud model, and a routing guard enforces it.
- **Identity-aware and per-user.** Reg authenticates against the customer's Active Directory; each user gets a private, ACL-enforced file area, and every outbound action — email, calendar, deletions — passes through a draft-then-approve gate.

Why It Is Different

Existing options each leave a gap for regulated engineering work. Reg is the only known system that combines air-gapped, U.S.-origin, identity-aware AI with a verified multi-vendor FPGA/VHDL toolchain in a single deployable platform.

Approach	Representative offerings	Gap for defense / CUI work
Cloud AI assistants	ChatGPT, Copilot, Gemini	Every prompt leaves the boundary — CUI/ITAR-prohibited and exposes controlled data to uncleared systems and foreign persons
On-prem generic LLM chat	Open WebUI + Ollama, LM Studio, Jan	Chat only — no email/document/FPGA workflow, no per-user identity or approval gates, model provenance unmanaged
Cloud developer copilots	GitHub Copilot, Cursor, Claude Code (cloud)	Software code only, cloud-hosted; no VHDL or FPGA implementation sign-off; CUI-prohibited
DIY on-prem integration	In-house scripts around a local model	Months of glue code; typically no identity/ACLs, no verified engineering toolchain, no managed model provenance
Reg	Rickert-Areno Engineering	Air-gapped, U.S.-origin, identity-aware AI plus a verified FPGA/VHDL engineering toolchain across four vendors — one deployable system

Key Capabilities

- **CUI-safe by construction:** controlled data never leaves the boundary, and no foreign-origin model touches it.
- **Full engineering copilot:** chat, document authoring (Word/PDF/PowerPoint), and mail + calendar across Gmail, Microsoft 365, and iCloud — with a human-approval gate on everything outbound.
- **Verified VHDL, not just generated VHDL:** a generate-compile-simulate-repair loop against a real simulator, so only code that passes ships.
- **FPGA implementation sign-off across four vendors:** Microchip, Lattice, AMD/Xilinx, and Altera — synthesis through timing-closed bitstream, with hundreds of modules verified per vendor.
- **Deep reasoning on demand:** the escalation ladder brings 120B- and 400B-class in-boundary models to hard problems without leaving the enclave.

- **Deterministic guardrails:** timekeeping, file access, and pass/fail detection live in code, never in model output — the design principle that keeps Reg dependable in production.

Technical Specifications

Deployment	Self-hosted and air-gapped from the cloud; runs on customer-owned hardware inside the existing security boundary
Models (U.S.-origin only)	IBM Granite-4.1-30B (chat/orchestration), Microsoft Phi-4 (VHDL coder), gpt-oss-120B and Llama-4-Maverick-400B (reasoning tiers), Snowflake Arctic-Embed-2 (retrieval) — no PRC-origin weights, even air-gapped
Reasoning tiers	Four-tier escalation ladder — fast local router → 120B in-boundary → 400B-class frontier in-boundary → cloud (non-CUI only), with a routing guard that blocks any cloud call after in-boundary escalation
Application	Python / FastAPI web app over TLS; Active Directory / PAM authentication; per-user private file areas with Kerberos-enforced ACLs
Engineering toolchain	Verified VHDL generation (generate → compile → simulate → repair against QuestaSim/GHDL) and FPGA implementation sign-off across four vendors; 220+ integrated tools
FPGA sign-off	Microchip Libero (PolarFire), Lattice Radiant (MachXO5-NX), AMD Vivado (Zynq-7000), Altera Quartus Pro (Agilex 7) — synthesis → place-and-route → timing → bitstream
Additional modalities	Whisper speech-to-text; runtime-selectable image generation (SDXL / Flux.1 / Flux.2); document generation (Word, PDF, PowerPoint)
Hardware footprint	Scales from a single workstation GPU to a multi-node estate with dedicated deep-reasoning and frontier-model appliances
Compliance	CMMC Level 2 (self); ITAR / EAR-aware handling; built and operated by a U.S.-owned, woman-owned small business, all work performed in the U.S.

Deployment & Use Cases

Reg deploys as a self-hosted application stack sized to the customer's boundary — from a single GPU workstation to a multi-node estate with dedicated deep-reasoning and frontier-model appliances. Rickert-Areno Engineering delivers it as licensable software with integration and provisioning available as a service: Active Directory and identity wiring, model placement, and the engineering toolchain configured to the customer's FPGA vendors and target devices. The primary target is the Defense Industrial Base — contractors who need modern AI on CUI/ITAR engineering work but cannot use the cloud. The same platform serves critical-infrastructure and other regulated operators with air-gap requirements, and its FPGA/VHDL toolchain makes it a force-multiplier for any team doing hardware design under compliance constraints.

About & Contact

Rickert-Areno Engineering, LLC is a U.S.-owned, woman-owned small business delivering secure product architectures, licensable hardware-root-of-trust IP, and self-hosted AI systems to the U.S. Government, the Defense Industrial Base, and critical-infrastructure operators. Reg was designed, built, and is operated entirely in-house.

Dr. Matthew Areno, Chief Technology Officer · matt@rickertarenoengineering.com · (737) 406-6197