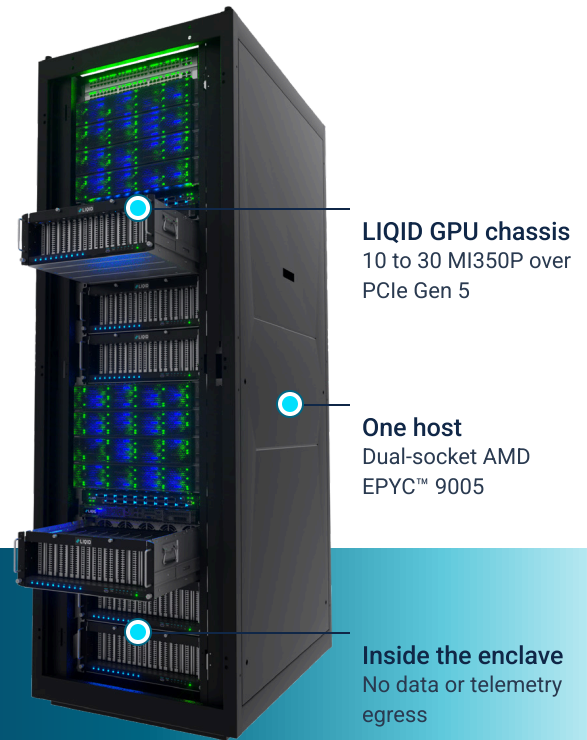


Every analyst. Cited answers.

Agency-wide document Q&A on AMD Instinct™ MI350P. Ask questions of your own records and get cited answers, on one Liquid UltraStack server, fully inside your enclave.



10 GPUs

MI350P on one host, room for 30

1.44 TB

HBM3E on one server

6 GPUs

of answer replicas, one model each

0 hops

network hops between GPUs

Inside the enclave
No data or telemetry egress

The Pilot-to-Agency Dilemma

Keep document Q&A with one analyst team, or scale it to thousands inside the boundary.



Demand Pressure

Use cases grew nine-fold

Federal generative AI use cases went from 32 to 282 in a year.

One pilot, thousands of users

A Q&A pilot that works for one team becomes agency-wide demand.



Infrastructure Pressure

Infrastructure is the barrier

Eight of twelve agencies cite insufficient technical infrastructure.

Every server adds scope

Each server, switch and cable in the boundary must be authorized.



Enclave Pressure

No hosted APIs

Classified and air-gapped networks can't reach them.

Analysts verify the source

Every answer has to link to the passage behind it.

Index It. Ask It. Cite It.

One 10-GPU Liquid UltraStack server runs agency-wide document Q&A. A RAG system indexes your records and gpt-oss-120b answers with citations, on an open stack that installs with no internet.

- » Air-gap ready: no data or telemetry leaves the enclave
- » A smaller boundary: one host, no network between GPUs
- » Fixed capacity, no per-token fees as users grow

Open-weight models · open-source RAG stack

6

Answers

2

Retrieval

2

Ingest + failover

LIQID GPU chassis · 10 AMD Instinct™ MI350P · 1.44 TB HBM3E

PCIe Gen 5 connection · chassis to host

One dual-socket AMD EPYC™ 9005 host · one OS image

Answers · 6 GPUs

gpt-oss-120b replicas, one full model per GPU, serve concurrent users.

~65 GB

gpt-oss-120b, fits one card

1,256

tok/s, 72B-class, 8 GPUs

Retrieval · 2 GPUs

Embedding and reranking models find the right passages.

9.5x

scaling, 1 to 10 GPUs

44,383

tok/s, 7B-class, 10 GPUs

Ingestion + Failover · 2 GPUs

Index new records and cover failover for the answer fleet.

144 GB

HBM3E per MI350P

1 host

in the accreditation boundary

Measured results: Liquid Phase-3 testing on 1 to 10 MI350P GPUs on one host. Model size is a published spec. Notes on page 2.

Built for

Defense · Intelligence community · Civilian agencies · Records and FOIA offices

Proven on

AMD Instinct™ MI350P · AMD EPYC™ 9005 · ROCm™ · 200+ Liquid deployments

Size Your Enclave

sales@liquid.com · +1 303.500.1551

Proof: Measured on MI350P

No cluster inside the boundary: **one host, ten GPUs, no GPU network, and room to grow to 30.**

Result	What was measured	Why it matters here
~65 GB gpt-oss-120b at MXFP4	Published model size: the full model fits on one 144 GB MI350P with room for cache	Six answer replicas, no model split across cards
1,256 tok/s on 8 GPUs	72B-class model, FP8, under interactive limits: P99 TTFT ≤ 2 s, ≥ 30 tok/s per user	Reference for interactive answer speed at scale
9.5× scaling 1 to 10 GPUs, 7B-class	4,653 tok/s on 1 GPU to 44,383 tok/s on 10, one replica per GPU	Retrieval and ingestion scale with every GPU
What it means	Thousands of users, one host	No inter-node network in the boundary

Model size is a published spec. 7B-class: FP4, one replica per GPU, concurrency 128. 72B-class: FP8, one replica per GPU, P99 time-to-first-token ≤ 2 s, ≥ 30 tok/s per user. Different test conditions; don't compare them. Source: Liquid Phase-3 testing. gpt-oss-120b performance is projected from these results and has not been measured directly.

Grow on the Same Server: 10, 20 and 30 GPUs

Add GPUs and a new job, not a new cluster.

10 GPUs
1.44 TB HBM3E

START HERE

Agency-wide air-gapped document Q&A
Scale document Q&A from one analyst team to thousands of users.
MODEL LAYOUT
6 × gpt-oss-120b answers
2 × embedding and reranking
2 × ingestion and failover
WHO BUYS, WHY NOW
CIO / CDAO of agency or command. Pilot success; user demand beyond one team.

20 GPUs
2.88 TB HBM3E

Records review and release (FOIA, redaction)
Classify, redact and draft release decisions with human sign-off.
MODEL LAYOUT
10 × Llama 3.1 8B / Mistral Small 3.2 classify and PII
6 × Llama 3.3 70B redaction rationale and drafting
4 × gpt-oss-120b reviewer and QA
WHO BUYS, WHY NOW
Records Officer / FOIA office. Statutory backlogs; litigation over delays.

30 GPUs
4.32 TB HBM3E

Frontier reasoning inside the enclave
A frontier-class open model plus analyst agents, with no cluster to accredit.
MODEL LAYOUT
8 × Mistral Large 3 (675B, FP8)
16 × analyst agents (70B, gpt-oss-120b)
6 × retrieval, translation and ingestion
WHO BUYS, WHY NOW
Program office / mission owner. Sovereign AI mandates; no hosted frontier model at your classification.

One host with 30 GPUs keeps the accreditation boundary small. Starting with one team? See the 4-GPU Air-Gapped Document Q&A brief.

What UltraStack Adds

- » **One model per GPU:** 144 GB HBM3E holds a full 70B model at FP8, or gpt-oss-120b.
- » **No network between GPUs:** peer-to-peer over PCIe Gen 5 to one EPYC host.
- » **Air-cooled:** about 22 kW for the full 30-GPU system.

Start Fixed, Add Flexibility

- » **One host:** one OS image, one license footprint, one thing to patch.
- » **Room to grow:** 10 to 20 to 30 GPUs on the same host.

What Is Proven Today

- » P2P bandwidth saturates the link; RCCL collectives pass, 1 to 10 GPUs.
- » ROCm integration complete, developed jointly with AMD.
- » 200+ production deployments, 85+ patents.

The Efficiency Case Targets, not measurements*	Up to 75% fewer servers	Up to 2.1× more tokens per dollar	Up to 1.8× better tokens per watt	Up to 50% lower power
--	-----------------------------------	---	---	---------------------------------

Two questions settle the size of the opportunity: **how many analysts want the pilot today, and how large is the record set behind it?**

Bring Us Three Numbers

- 1 Corpus:** documents indexed and yearly growth.
- 2 Users:** analysts at peak and questions per day.
- 3 Context:** the longest record an answer draws on.

Request a Sizing Session

We size it in one working session.
sales@liquid.com · liquid.com

*Efficiency targets compare against multi-server deployments of the same GPUs; they are Liquid internal projections using AMD published specifications, and results vary by configuration and workload. Performance figures for the AMD Instinct MI350P are based on AMD published specifications and Liquid Phase-3 testing from 1 to 10 GPUs, and are subject to change. Performance at 20 and 30 GPUs, and for 400B+ models, is not yet published; 20- and 30-GPU layouts and frontier-model fit are sizing arithmetic on published model sizes, not benchmark results. Models shown are US and EU open-weight models. Embedding and reranker models are selected per deployment from US/EU-origin open-weight options. This is a reference architecture, not an accredited system; authorization is the customer's responsibility. Federal AI data: GAO-25-107653.