



Defensible review. Inside your firm.

Review you can defend.

Defensible e-discovery on AMD Instinct™ MI350P. Relevance, privilege and PII coding across millions of documents, on one Liquid UltraStack server, on a model version you control.

20 GPUs

MI350P on one host, room for 30

2.88 TB

HBM3E on one server

9.5x

coding scaling, 1 to 10 GPUs

0 hops

network hops between GPUs



LIQID GPU chassis
10 to 30 MI350P over PCIe Gen 5

One host
Dual-socket AMD EPYC™ 9005

Inside your firm
No data egress

The Defensibility Dilemma

Send first-pass review to a vendor, or bring it in-house on a model you can explain.



Volume Pressure

Review grows with every matter
First-pass budgets go to outside vendors, priced per document.

AI in use, protocols missing
61% use AI-assisted review; under a third document protocols.



Court Pressure

Courts ask how AI was used
Model, configuration, validation and error rates.

Hosted APIs shift underneath you
A model that changes without notice is hard to defend.



Privilege Pressure

Privileged documents can't leave
Client-confidential data has to stay in your environment.

Attorneys own the call
Every AI decision has to link back to its source document.

Pin It. Run It. Defend It.

One 20-GPU Liquid UltraStack server runs the whole first-pass pipeline. Small models code every document, a 70B model drafts the privilege log and checks the calls, and a reviewer model reports error rates.

- » Pinned, logged model: the review can be reproduced
- » No per-token or per-document fees
- » Grow from 10 to 20 to 30 GPUs on the same host

Pinned, logged models · one review pipeline

12

Coding fleet

6

Privilege log + QC

2

Validation sampling

LIQID GPU chassis · 20 AMD Instinct™ MI350P · 2.88 TB HBM3E

PCIe Gen 5 connection · chassis to host

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

Coding Fleet · 12 GPUs

Mistral Small 3.2 and Llama 3.1 8B code every document for relevance, privilege and PII.

9.5x

scaling, 1 to 10 GPUs

44,383

tok/s on 10 GPUs, 7B-class

Privilege Log + QC · 6 GPUs

Llama 3.3 70B, one full model per GPU, drafts log entries and checks coding calls.

1,256

tok/s, 72B-class, interactive, 8 GPUs

1 GPU

per 70B model at FP8

Validation Sampling · 2 GPUs

gpt-oss-120b reviews statistical samples and reports error rates.

144 GB

HBM3E per MI350P

1 card

holds gpt-oss-120b

Measured results: Liquid Phase-3 testing on 1 to 10 MI350P GPUs on one host. Test conditions and notes on page 2.

Built for

Litigation support · e-Discovery · Legal innovation · Knowledge management

Proven on

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

Size Your Next Matter

sales@liquid.com · +1 303.500.1551

Proof: Measured on MI350P

No vendor, no cluster: one server, a pinned model and a validation record the court can follow.

Result	What was measured	Why it matters here
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	Coding throughput grows with every GPU in the fleet
1,256 tok/s on 8 GPUs	72B-class model, FP8, under interactive limits: P99 TTFT ≤ 2 s, ≥ 30 tok/s per user	Privilege log drafting at interactive speed for reviewers
1 GPU per 70B model	Llama 3.3 70B is ~70 GB at FP8 and fits on one 144 GB MI350P with room for cache	Each stage runs on its own GPUs, with no model split across cards

What it means	Every document gets a first pass	Attorneys review flagged calls, not every page
---------------	----------------------------------	--

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. Mistral Small 3.2, Llama and gpt-oss-120b performance is projected from these results; 20-GPU performance is not yet published.

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

Firm-wide contract review and deal-room surge

Scale contract review across practice groups and absorb M&A diligence spikes.

MODEL LAYOUT

6 × Llama 3.3 70B interactive review
4 × bulk diligence, moved to interactive when a deal closes

WHO BUYS, WHY NOW

CIO / Director of Legal Innovation.

Adoption beyond one practice group; deal volume.

20 GPUs

2.88 TB HBM3E

START HERE

Defensible first-pass e-discovery

Relevance, privilege and PII coding across millions of documents, on a version-pinned model.

MODEL LAYOUT

12 × Mistral Small 3.2 / Llama 3.1 8B coding fleet
6 × Llama 3.3 70B privilege log and QC
2 × gpt-oss-120b validation sampling

WHO BUYS, WHY NOW

Litigation Support / e-Discovery Director.
Large matter, review vendor cost, court scrutiny of AI.

30 GPUs

4.32 TB HBM3E

Whole-matter reasoning

Long-context analysis across deposition transcripts, pleadings and exhibits.

MODEL LAYOUT

16 × Mistral Large 3 (675B, BF16, 256K context)

14 × extraction and summarization workers

WHO BUYS, WHY NOW

Managing Partner / Chief Knowledge

Officer. Bet-the-company litigation; no data egress.

At 30 GPUs the frontier model runs at full BF16 precision for defensibility, with workers alongside. Starting with one job? See the 4-GPU Contract Review 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 documents are in your next large matter, and what will first-pass review cost?**

Bring Us Three Numbers

- 1 **Volume:** documents per matter, matters per year.
- 2 **Deadline:** the production turnaround required.
- 3 **Concurrency:** reviewers working at peak.

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. Outputs support, and do not replace, attorney review; review protocols and validation remain the firm's responsibility. Adoption data: EDRM 2025.