

Documents in. Data out.

On-premises AI extraction on AMD Instinct™ MI350P. Loan, KYC and contract files become validated JSON, on one Liqid UltraStack server inside your data center.

4 GPUs

MI350P on one host, room for 30

576 GB

HBM3E on one server

2.2x

throughput, 2 to 4 GPUs, measured

0 hops

network hops between GPUs



LIQID GPU chassis
4 to 30 MI350P over PCIe Gen 5

One host
Grows to 30 GPUs, no new servers

Inside your data center
No data egress

The Re-Keying Dilemma

Keep analysts re-keying documents by hand, or send regulated files to a public AI API.



Manual Pressure

Analysts re-key every field
From PDFs and scans into core systems, by hand.

Slow and error-prone
Every manual field is a delay and a chance for a mistake.



Cost Pressure

Per-token fees at volume
Public API costs grow with every document you process.

Spikes become budget spikes
Month-end and deal volume drive token bills up.



Data Pressure

Regulated files can't leave
Loan, KYC and contract files can't go to a public AI API.

Output feeds core systems
Every record has to arrive as clean, validated fields.

Read It. Extract It. Validate It.

Four MI350P GPUs on one Liqid UltraStack server run a 70B model that reads each document and returns validated JSON (parties, amounts, dates, covenants, identity fields) into your workflow.

- » Data never leaves your environment
- » Fixed capacity replaces per-token API fees
- » Standard PCIe GPUs, open-source software stack

Llama 3.3 70B (FP8) · validated JSON output

2

Extraction

1

Triage + classify

1

Failover / peak

LIQID GPU chassis · 4 AMD Instinct™ MI350P · 576 GB HBM3E

PCIe Gen 5 connection · chassis to host

One host server · one OS image · grows to 30 GPUs

Extraction · 2 GPUs

Llama 3.3 70B replicas read each document and return validated JSON.

~7,800

docs/hr on 4 GPUs*

~2.2

docs/sec on 4 GPUs*

Triage + Classify · 1 GPU

Sorts incoming loan, KYC and contract files before extraction.

1 GPU

per 70B model at FP8

144 GB

HBM3E per MI350P

Failover + Peak · 1 GPU

Covers failover, or joins extraction for maximum throughput.

~18

interactive users*

552

tok/s on 4 GPUs

Measured: Liqid Phase-3 testing, 72B-class model on MI350P. *Illustrative; assumptions on page 2.

Built for

Loan files · KYC files · Contract files · Validated JSON output

Proven on

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

Size Your Volume

sales@liqid.com · +1 303.500.1551

Proof: Measured on MI350P

No re-keying, no public API: one server turns regulated documents into structured data.

Result	What was measured	Why it matters here
552 tok/s on 4 GPUs	72B-class model, FP8, under interactive limits: P99 TTFT ≤ 2 s, ≥ 30 tok/s per user	Interactive extraction on the 4-GPU configuration
Near-linear 2 $\rightarrow$ 4 $\rightarrow$ 8 GPUs	254 tok/s on 2 GPUs, 552 on 4, 1,256 on 8, one replica per GPU	Add GPUs and extraction throughput grows with them
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	No multi-GPU model splitting, so every GPU is its own replica
What it means	Every document becomes data	Analysts check exceptions instead of re-keying fields

*Illustrative. Assumes 1,024 input and 256 output tokens per document; ~ 18 = concurrent interactive users. 72B-class: FP8, one replica per GPU. Source: Liquid Phase-3 testing. Llama 3.3 70B performance is projected from same-class 72B dense-model results and has not been measured directly. Results vary with document length and output size.

Grow on the Same Server: 4 to 30 GPUs

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

4 GPUs 576 GB HBM3E Regulated document extraction Loan, KYC and contract files into validated JSON. MODEL LAYOUT 2 $\times$ Llama 3.3 70B extraction 1 $\times$ triage and classification 1 $\times$ failover or peak load BUILT FOR Lending, KYC and contract operations. Re-keying backlogs; documents that can't go to a public API. START HERE	10 GPUs 1.44 TB HBM3E Financial crimes investigation agent Triage, evidence, narrative and QA for every alert. MODEL LAYOUT 4 $\times$ Llama 3.1 8B triage 4 $\times$ Llama 3.3 70B evidence and narrative 2 $\times$ gpt-oss-120b typology and QA WHO BUYS, WHY NOW Head of Financial Crimes / BSA Officer. Alert backlog, exam findings, investigator attrition.	20 GPUs 2.88 TB HBM3E Bank-wide private model hub One server for every line of business: triage, extraction, KYC, complaints, credit memos and research. MODEL LAYOUT 1–2 GPUs per job from the 4-GPU briefs 6–8 GPUs of shared Llama 3.3 70B and gpt-oss-120b, interactive by day, batch by night WHO BUYS, WHY NOW CIO / Head of AI Platform. 5+ departmental pilots competing for GPUs; API bills growing.
--	--	---

At 30 GPUs, champion and challenger model validation runs on the same host. See the Financial Crimes Investigation brief for the full 10, 20 and 30-GPU path.

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:** 4 to 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$\times$ more tokens per dollar	Up to 1.8$\times$ better tokens per watt	Up to 50% lower power
--	-----------------------------------	---	---	---------------------------------

Two questions settle the size of the opportunity: **how many documents do your analysts re-key each day, and what does each one cost?**

Bring Us Three Numbers

- 1 **Volume:** documents per day and the peak backlog.
- 2 **Document size:** pages and fields per file.
- 3 **Concurrency:** users and systems calling extraction 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. Docs/sec, docs/hr and user figures are illustrative and vary with document length and output size.