The Model Came to the Data

A local open-weight deployment that made the crown-jewel workflows economically viable.

Dr. Tim Nedyalkov | Chief AI Officer & Cybersecurity Executive
tim@cyberkeynote.com | August 2026

The highest-value work was frozen.

The more sensitive the workflow, the larger the upside, the harder it was to touch.

~40,000
documents
KYC packs, contracts, and claims

7 weeks
rolling queue

~40 minutes
analyst time per batch

HIGH-VALUE WORK

Crown-jewel documents held the largest automation upside.

→

MANUAL CAPACITY

Analyst time was absorbed by a queue that kept rolling.

→

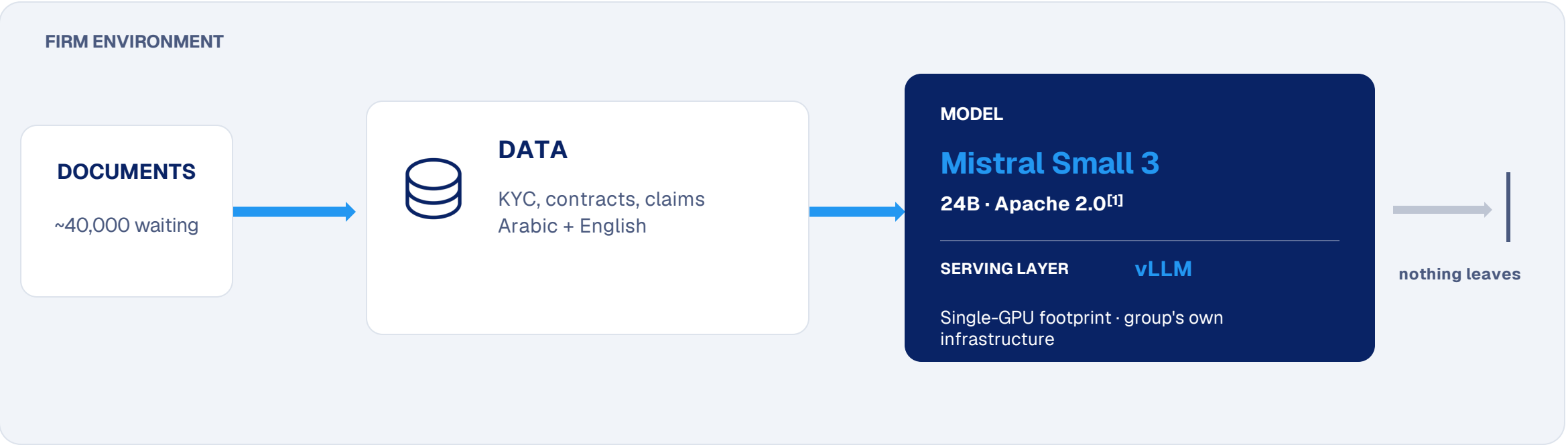
VALUE WAITING

The biggest prizes stayed frozen.

The opportunity was already inside the environment. It needed a production path.

We brought the model to the data.

Mistral Small 3 ran extraction and summarization directly against the existing document workflow.



The model, serving layer, and data all stayed inside one controlled production boundary.

[1] Mistral Small 3 - Mistral AI, 2025.

SELECTION CRITERIA

The firm selected against four requirements.

Governance rigor narrowed the field before any vendor preference entered the decision.

1

OPEN WEIGHTS

Self-hostable

2

SINGLE GPU

Deployable footprint

3

ARABIC

Capability to evaluate

4

ENGINEERING SUPPORT

Forward-deployed

FIELD EVALUATED

Llama · Phi-4 · gpt-oss · Mistral

Selected for workload fit, not blanket model superiority.

WHY MISTRAL

European provider + Apache 2.0 [1]

Single-GPU footprint · deployment engineering support

The result was a stack the firm could own, run, evaluate, and keep improving.

[1] Mistral Small 3 - Mistral AI, 2025.

Processing time per document fell by ~90%.

One primary metric, one backlog-burn illustration, and one downstream business effect.

PRIMARY OUTCOME

~90%

LESS PROCESSING TIME PER DOCUMENT

BACKLOG BURN

7 weeks → about a day

DOWNSTREAM
~3 to 7x onboarding throughput

WHAT FRONTIER DOCUMENT AI ACHIEVES TODAY

98%+
printed OCR

96%+
dense tables

Lower-cost models matched more expensive siblings on extraction. The larger gap appeared on reasoning. Human review handles exceptions. [2]

OPTIMIZATION SCENARIO

FIXED CAPACITY	INFERENCE MARGINAL	CROSSOVER
Low thousands	**Near zero**	**Few thousand**
Hardware-only	Excludes operating costs	Documents vs per-doc API

Unit cost: ~$3 to ~$0.20 per document. Reuse spreads fixed capacity across more workflows.

Freed analyst capacity moved to revenue work while the local pattern became reusable.

[2] Nanonets IDP Leaderboard - Nanonets, 2026.

Selected workloads are moving in-house.

The opportunity is shifting from renting the best model to placing each workload where it creates the most value.

GENERAL CAPABILITY GAP [3]

~4 MONTHS

~8 ECI POINTS

Epoch measures the overall open-to-closed frontier gap. It does not decide a structured extraction workload by itself.

SELECTIVE REPATRIATION [4]

PLACE THE WORKLOAD

AlixPartners frames repatriation as selective placement for predictable workloads, AI inference, sovereignty, cost, and control.

Task fit matters more than a blanket parity claim. Extraction is evaluated separately from general reasoning. [2]

[2] Nanonets IDP Leaderboard - Nanonets, 2026. [3] Open/Closed Model ECI Gap - Epoch AI, 2026.
[4] Cloud Repatriation - AlixPartners, 2026.

Build the pattern once, then scale it.

Build the local-deployment pattern once and a whole class of previously off-limits, high-value workflows opens across the business.

LOCAL MODEL
On your infrastructure

EVALUATION
Real cases and clear thresholds

HUMAN SIGN-OFF
Accountable production release

NEXT WORKFLOW
Reuse and keep improving

REUSE THE PATTERN

- CLIENT ONBOARDING
- KYC PACKS
- CONTRACTS
- CLAIMS

> **Put AI to work exactly where the value and the sensitive data already live.**

Going local was one decision. The mandate covers them all.

A fractional Chief AI Officer, embedded in your leadership team: governance, decision rights, spend, evidence, and the board report, owned end to end.

- **In your leadership team in one week or less**
- **First results by day 30, first board report delivered**
- **Four artifacts: decision rights, controls, evidence, board reporting**
- **Outcomes like this case: quality held, 90% less processing time**

CHIEF AI OFFICER AS A SERVICE

Dr. Tim Nedyalkov

Chief AI Officer & Cybersecurity Executive

Book a free 30-minute consultation

tim@cyberkeynote.com

cyberkeynote.com/caio

Dubai based · travels for selected engagements