

LTF ONLINE · TUESDAY 13 OCTOBER 2026

Is AI worth owning?

Strategic autonomy for law firms and legal departments

Raymond Blyd

CEO of legalcomplex · Founder of Sabaio

STRATEGIC AUTONOMY

Autonomy is control of the
data, the model and the
software

Implications

SPEND

Financial

Every question you ask AI costs a little money

A flat monthly fee cannot cover a bill that keeps growing

Buying your own computers costs a lot up front, and chips are scarce

INCOME

Commercial

Clients ask what happens to their data

AI vendors are changing how they charge

Being in control can win you clients

CONTROL

Legal

Client secrets stay your job, whoever runs the computer

If a foreign court asks, it asks your provider, not you

Your data goes wherever the AI runs, whatever the contract says

Firms are buying their way to control

Kirkland & Ellis

\$500M into its own AI platform

Latham & Watkins

Buying its own GPU servers

Harvey

Offers to build private models for firms

Thomson Reuters

\$40M and 20 months to train on Qwen3.6

PART 1 OF 3

Financial

What AI costs when it is metered, rented or owned

Financial

Commercial

Legal

Vendors earn about \$2 per million tokens. The model costs about \$15

\$2.08

Harvey earns per
million tokens

\$1.39

Legora earns per
million tokens

\$15

A frontier model costs per
million tokens

Like selling a \$15 item for \$2. That is why flat prices are giving way to usage prices

Two shifts change the cost equation

Usage-based pricing

Vendors are moving from flat seats to consumption as AI use grows

Open-weight models

Models like Kimi K3 are becoming good enough to rival the frontier. Custom models can run on-prem and on EU soil

Harvey's margin went from 50% to -50% once it served its own model (my calculation)

PART 2 OF 3

Commercial

What is worth owning, and where your edge is

Financial

Commercial

Legal

Open versus Closed models

2,739

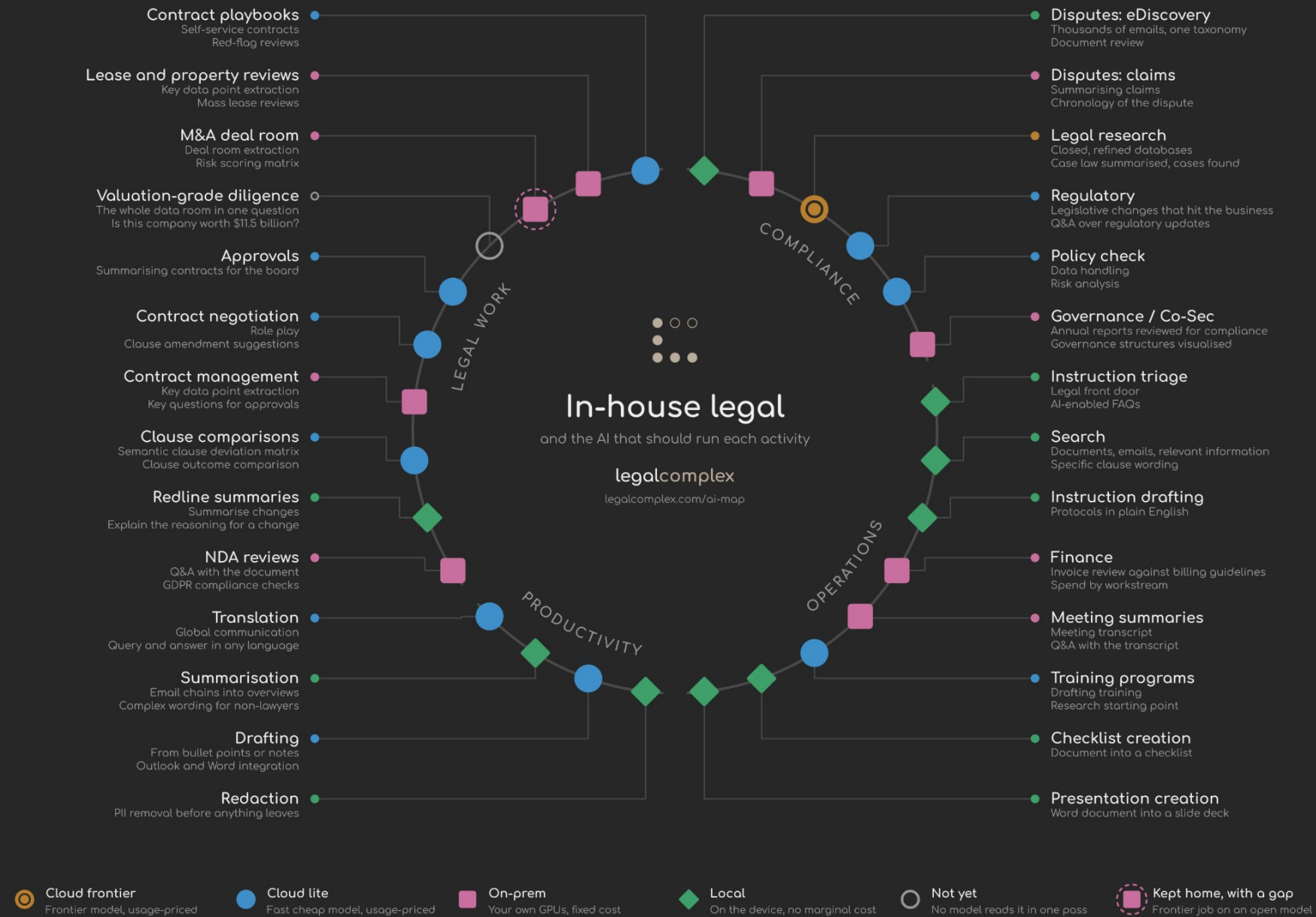
Legal AI models on Hugging Face

1M+

Downloads of Legal-BERT since 2022

Route each activity by a rule

If the activity is	It runs
A simple task a rule already solves	No AI at all
Past a million tokens of context	Nothing runs it yet
Classified	On your own hardware, never the cloud
Must stay in your domain, or a topic a hosted model may refuse	On the device if routine and small, on-prem otherwise
Everything else	By capability: frontier, cheap cloud model, on-prem at volume, or the device



1. Past a million tokens of context: nothing runs it yet.
2. Classified: your own hardware, never the cloud.
3. Must stay in your domain, or a topic a hosted model may refuse; the device if routine and small, on-prem otherwise.
4. Everything else by capability: frontier work to the cloud frontier, skilled work to a cheap cloud model, or on-prem at volume, routine work to the device.

In-house legal: 28 activities, one needs the frontier

1

Cloud frontier

Legal research

9

Cloud lite

Skilled work on a
cheap model

17

Under your
control

9 on the device, 8
on-prem

1

Not yet

Valuation-grade
diligence

What is worth owning

Own

Your data and who can see it

The routing rule

Hardware, only where the map says on-prem

Rent

The frontier model, for the few tasks that need it

A cheap cloud model for skilled work

Skip

Training your own frontier model

AI for tasks a rule already solves

PART 3 OF 3

Legal

Who can reach your data, and the duty that stays with you

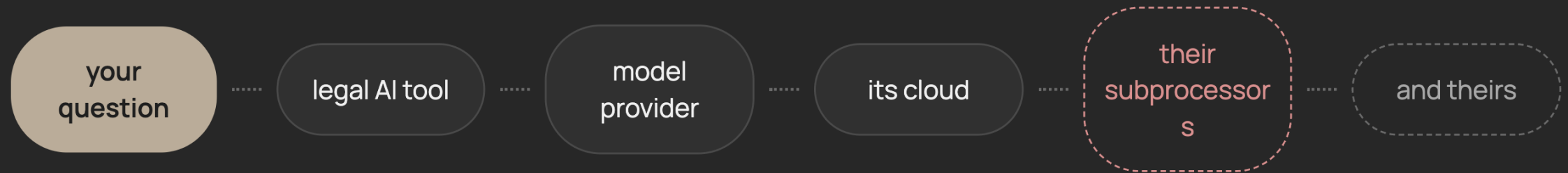
Financial

Commercial

Legal

Every subprocessor has subprocessors.

Your contract covers one vendor. Your data meets every vendor behind it.



ONE REAL CHAIN, FROM PUBLISHED LISTS

1 · YOUR VENDOR

A legal AI tool built
on OpenAI

2 · ITS SUBPROCESSOR

OpenAI

3 · THEIRS

Microsoft, CoreWeave,
Oracle

4 · AND THEIRS

Microsoft keeps its own
subprocessor list

Zero data retention is not zero data processing

Nothing is stored. The data is still read to answer you.

RETAINED

No

OpenAI Astra for Law offers no data retention, as the vendor describes it.

PROCESSED

Yes

The model has to read your data before it can answer.

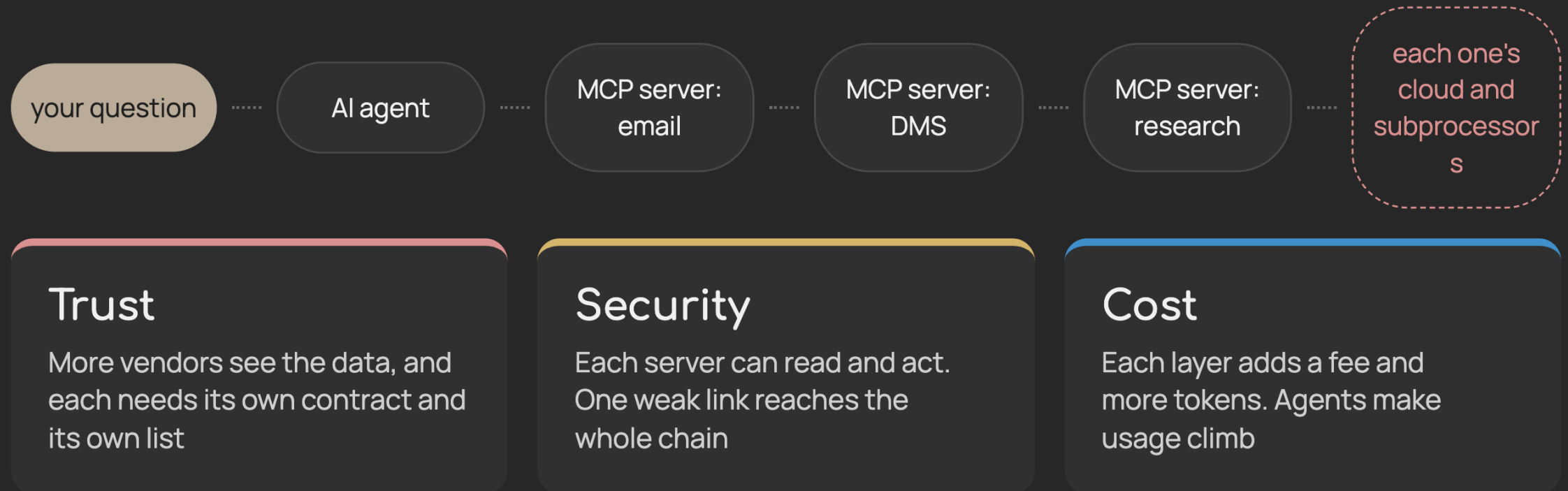
OPEN QUESTION

Can it be produced?

If a judge asks to see the data, what is there to show?

Every MCP layer is another subprocessor

Agents take actions, not just answers. Each connector is one more vendor your data passes through.



Five questions to take back to your vendors

FINANCIAL

What happens to my bill when you move to usage pricing?

COMMERCIAL

Can you show me the routing for each activity?

COMMERCIAL

If I want to leave, what do I take with me?

LEGAL

Which model ran this task, and why that one?

LEGAL

Where does my data go, and who can read it?

What to take back

SPEND

Financial

Vendors earn about \$2 on a model that costs about \$15, so prices are moving to usage

Know which tasks need the frontier before your bill does

INCOME

Commercial

Own your data and the routing rule

Rent the frontier only where it is needed: 1 of 28 in-house activities

CONTROL

Legal

The duty to keep client secrets stays with you, whoever runs the server

Hold the infrastructure, and every order has to come through your door

BACKUP

Detail and sources

The evidence behind each part, in the same order, then the timeline and every link

Financial

Commercial

Legal

Timeline and links

You cannot buy your way into autonomy today

Memory

There is a shortage right now.
Anyone buying in will wake up
from a nightmare

Compute

Not enough in the world to
build the best model alone.
Nvidia tried

Talent

Not enough to get you the
best solution. Meta tried

The sources behind the financial lens, 1 of 2

DISCLOSED

Harvey tokens: 1T to 14.6T a month

Harvey, via Artificial Lawyer, 8 July 2026
January to June 2026, a 14.6 times rise.

MY CALCULATION

Revenue per million tokens: \$17.75 to \$2

Legalcomplex, from Harvey ARR and the token curve

ARR interpolated between Dec 2025 (\$195M) and May (\$300M).

MY CALCULATION

Harvey gross margin: 50% to -50%

Legalcomplex, 24 September 2026

Once Harvey served its own model.

The sources behind the financial lens, 2 of 2

MY ESTIMATE

Harvey \$2.08, Legora
\$1.39 per million tokens

Legalcomplex, July 2026

ARR over estimated token volume.
The model costs about \$15.

MY ESTIMATE

\$3.00 paid, \$0.09 to \$0.25
on rented GPUs

Legalcomplex estimates, August 2026

Not vendor-disclosed. \$3.00 is a
frontier list price.

DISCLOSED

Thomson Reuters: \$40M
and 20 months

Steve Hasker with Richard Tromans,
Artificial Lawyer; Legalcomplex, 28
August 2026

To train a model on Qwen3.6.

Six places an activity can run



Cloud frontier

Frontier model, usage-priced



Cloud lite

Fast cheap model,
usage-priced



On-prem

Your own GPUs, fixed cost



Local

On the device, no
marginal cost



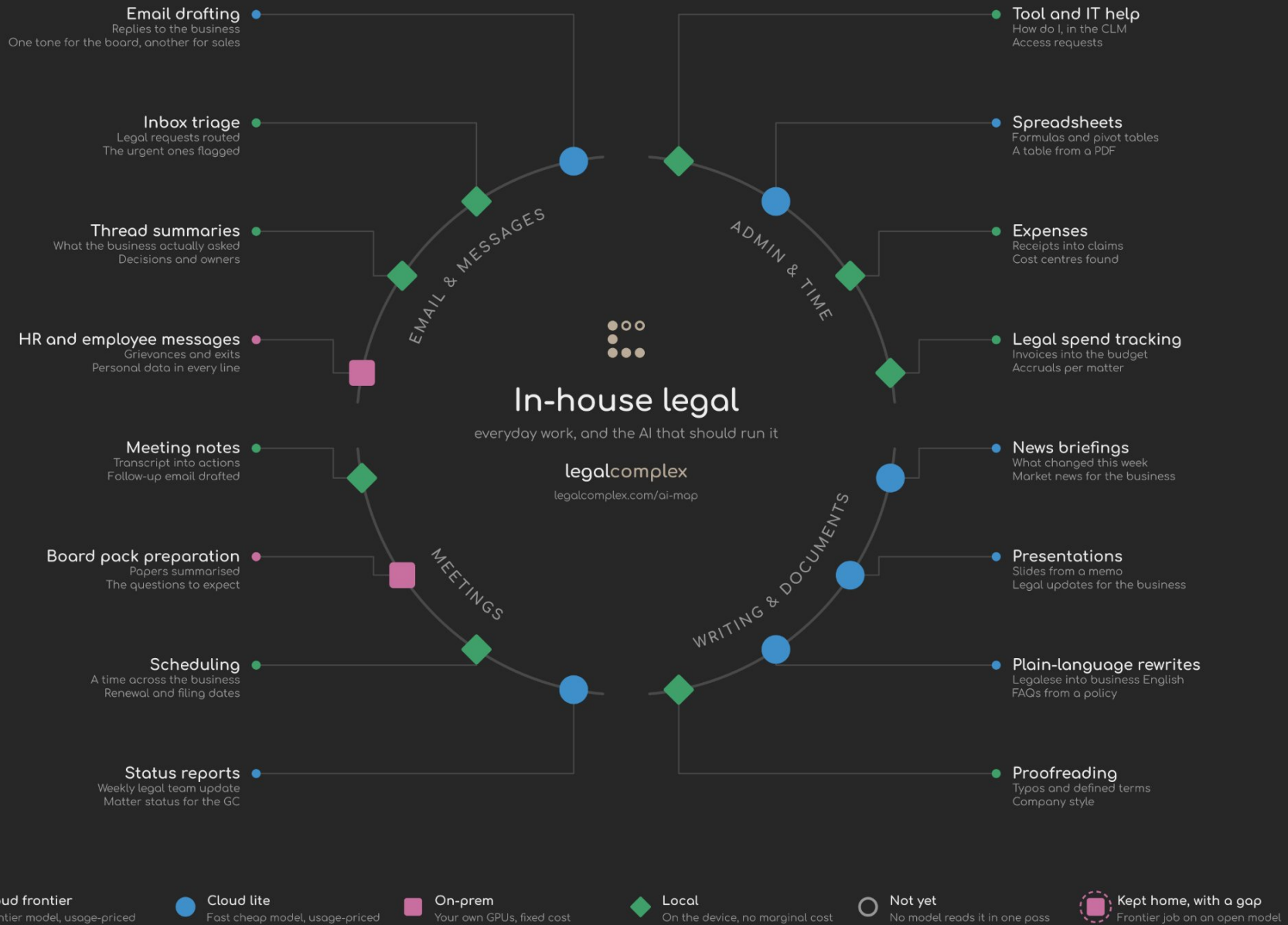
Not yet

No model reads it in one pass



Kept home, with a gap

Frontier job on an open model



1. Past a million tokens of context: nothing runs it yet.
2. Classified: your own hardware, never the cloud.
3. Must stay in your domain, or a topic a hosted model may refuse; the device if routine and small, on-prem otherwise.
4. Everything else by capability: frontier work to the cloud frontier, skilled work to a cheap cloud model, or on-prem at volume, routine work to the device.

EVERYDAY WORK

No everyday task needs the cloud frontier

16 activities: 8 on the device, 6 on a cheap cloud model, 2 on-prem

Drafting an email lands in three different places

Low firm

On-prem

Privileged client matter

In-house

Cheap cloud model

Replies to the business

Government

On the device

A reply to a citizen

Classified information:
your own hardware, never
the cloud

The map says which model. Orchestration says where it runs

1

Where it runs

The system you already work in, or a Legalcomplex service

2

What it reads

The inputs: the brief, the matter file, the case law

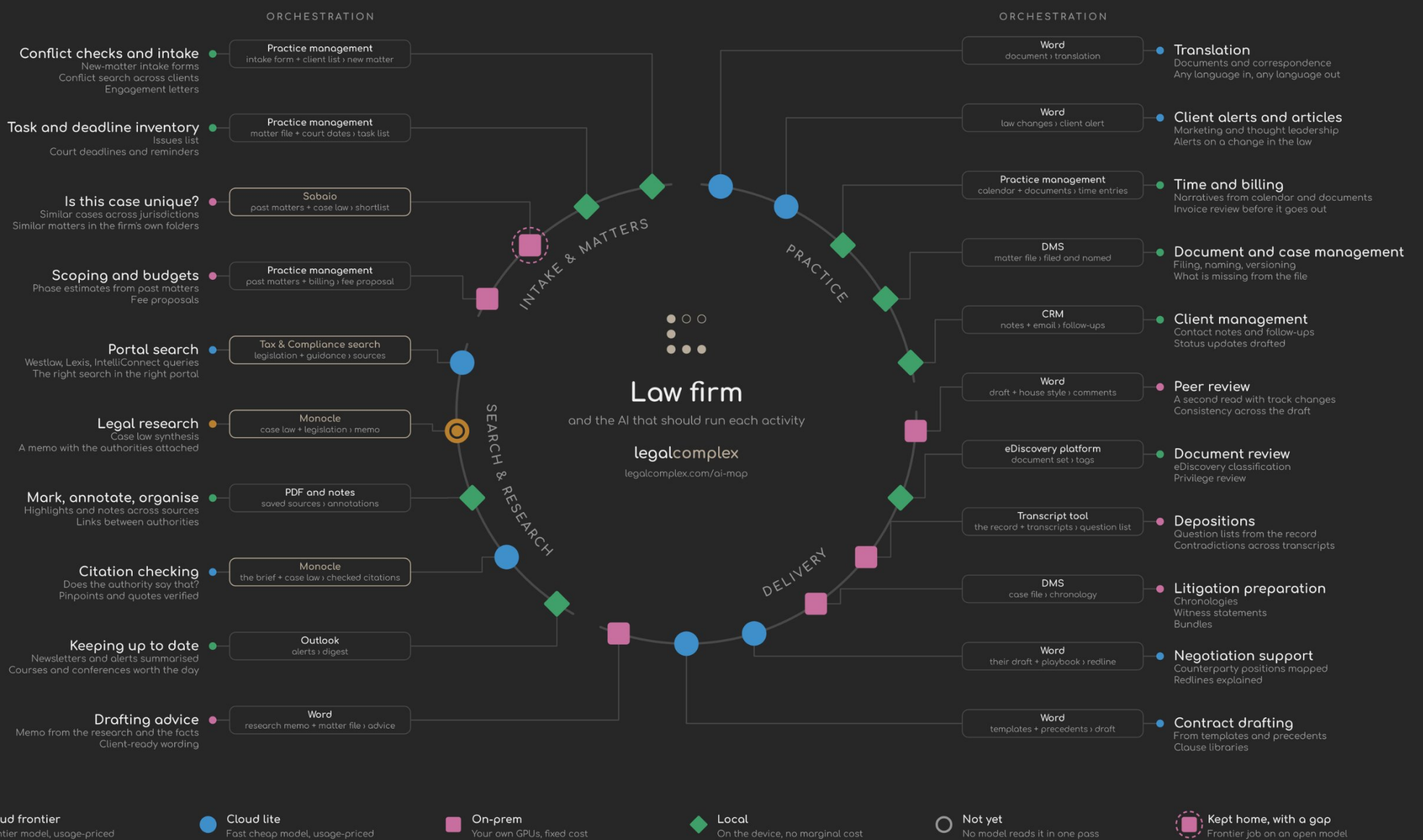
3

What it makes

The output that lands back in your day

Citation checking

Monocle · the brief + case law > checked citations



1. Past a million tokens of context: nothing runs it yet.
 2. Classified: your own hardware, never the cloud.
 3. Must stay in your domain, or a topic a hosted model may refuse: the device if routine and small, on-prem otherwise.
 4. Everything else by capability: frontier work to the cloud frontier, skilled work to a cheap cloud model, or on-prem at volume, routine work to the device.
 Orchestration: where each activity runs, then what it reads + what it makes. Champagne names a Legalcomplex service.

The sources behind the commercial lens, 1 of 2

OUR MAP

AI Activity Map: in-house legal

legalcomplex.com/ai-map. 28 activities, counts from the registry, 10 October 2026

One needs the frontier, 17 stay under your control.

PREVIEW

AI Activity Map: everyday work

Legalcomplex, 16 activities. Not yet on legalcomplex.com

No everyday task needs the cloud frontier.

SAMPLE

AI orchestration map: law firm

Legalcomplex, 21 activities. No in-house version yet

Where each activity runs, what it reads and what it makes.

The sources behind the commercial lens, 2 of 2

OUR RULE

Routing rule

Legalcomplex AI Activity Map

Verdicts are computed from the rule, not hand-picked.

PUBLIC DATA

Open models: 2,739 legal models

Hugging Face model hub; Legal-BERT: over 1M downloads since 2022, LinkedIn post

Open models for legal work are not a niche.

REPORTED

Firms moving to control

LTF Online announcement, 2 October 2026; Legalcomplex, 28 August 2026

Kirkland, Latham, Harvey and Thomson Reuters.

EU tenancy has a footnote.

Data at rest stays where you put it. Processing follows the endpoint.

Azure OpenAI

EU OPTION

Data Zone EU, or a regional deployment

GLOBAL OPTION

Global Standard, where new models land first

GLOBAL MEANS

Processed in any geography where the model is deployed

Google Vertex AI

EU OPTION

EU or regional endpoints

GLOBAL OPTION

The global endpoint, for capacity and availability

GLOBAL MEANS

Any Google Cloud location, no data residency guarantee

AWS Bedrock

EU OPTION

EU cross-Region inference profiles

GLOBAL OPTION

Global cross-Region inference

GLOBAL MEANS

Routed to any supported AWS commercial Region

EU soil is not EU control.

The law follows the operator, not the building.

WHO RUNS IT

**A US cloud, in
any region**

US CLOUD ACT, 2018

Covers data in the
provider's control,
wherever it is stored.

WHO RUNS IT

**A Chinese
provider**

INTELLIGENCE LAW, 2017

Art. 7: organisations must
support, assist and
cooperate.

WHO RUNS IT

An EU provider

EU E-EVIDENCE,
AUG 2026

Production orders from
any member state, due in
10 days.

WHO RUNS IT

You

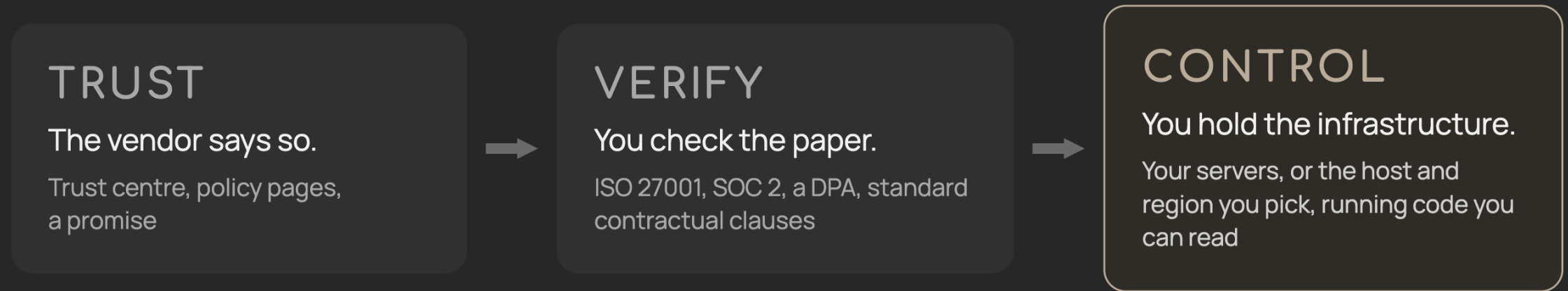
YOUR OWN COURTS

Any order has to come to
you. You see it, and you
can contest it.

EXAMPLE Amsterdam Westpoort: 78 MW, due 2028, on Dutch soil. Reported leased to Microsoft, so the CLOUD Act applies.

Trust, verify, or control.

Certificates and contracts verify a promise. Autonomy is physical control of the data.



Paper can be overruled. Microsoft France, under oath before the French Senate, 10 June 2025:

“Non, je ne peux pas le garantir.”

“No, I cannot guarantee that.”

The laws behind the legal lens, 1 of 2

UNITED STATES

CLOUD Act, 2018

18 U.S.C. § 2713

Covers data in the provider's control, wherever it is stored.

UNITED STATES

Non-disclosure orders

18 U.S.C. § 2705(b)

A US court can bar the provider from telling you.

CHINA

National Intelligence Law, 2017

Article 7

Organisations must support, assist and cooperate.

The laws behind the legal lens, 2 of 2

EUROPEAN UNION

e-Evidence Regulation

Regulation (EU) 2023/1543, from
Aug 2026

Production orders from any member
state, due in 10 days.

EUROPEAN UNION

CJEU C-694/20

8 December 2022

The EU Charter protects what
passes between lawyer and client,
advice included.

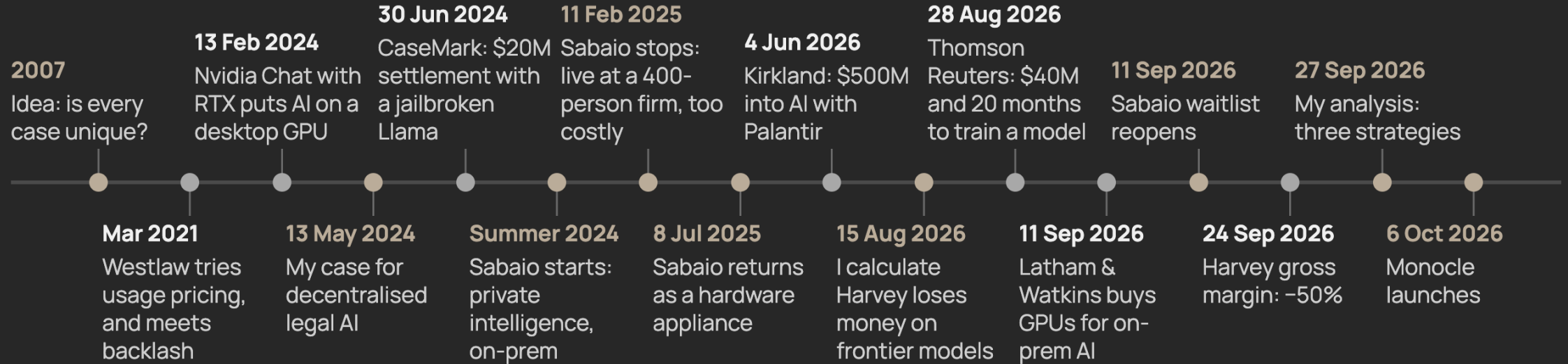
LEGAL PROFESSION

ABA Formal Opinion 512

2024

Outsourcing the server does not
outsource confidentiality.

How we got here: private AI, 2007 to today



Links: stories behind this talk (1 of 2)

6 Oct 2026

[How good is Jev at routing legal versus non-legal questions?](#)

2 Oct 2026

[Law firms want more control over their AI to protect their edge](#)

27 Sep 2026

[Private Intelligence: will on-prem legal AI strategy work?](#)

26 Sep 2026

[Monocle tracks token costs daily on all models in your harness](#)

Links: stories behind this talk (2 of 2)

24 Sep 2026

Harvey margins went from 50% to -50% once it served its own model

28 Aug 2026

Thomson Reuters spent \$40M and 20 months to train Qwen3.6

Jul 2026

Harvey earns \$2.08 per million tokens, Legora \$1.39, but the true cost is \$15

Live

The AI Activity Map

18 Sep 2026

OpenAI Astra for Law: no data retention, but prompt injection safeguards

Links: posts and legal sources

LinkedIn · timeline

[Private intelligence: the timeline from 2007 to Monocle](#)

LinkedIn · MCP

[Why will MCPs fail? They repeat the mistakes from OpenAI GPTs store](#)

LinkedIn · open models

[Legal-BERT downloaded over 1M times since 2022](#)

Hugging Face

[Open models on the Hugging Face hub](#)

Legal and technical sources

CJEU C-694/20, 8 December 2022 (lawyer-client confidentiality)

18 U.S.C. § 2705(b) and § 2713 (CLOUD Act)

ABA Formal Opinion 512, 2024

Regulation (EU) 2023/1543 (e-evidence)

PRC National Intelligence Law

French Senate inquiry hearing, Microsoft France, 10 June 2025

OpenAI sub-processor list

Microsoft Learn, Google Cloud and AWS documentation