

Specializzare e *scalare* l'IA Generativa nel settore legale

AI Conf 2026

24/06/2026

LEXROOM

Lexroom. The new *legal* standard.

LEXROOM

The New *legal* Standard

LEXROOM



Ask complex
questions.
Get reliable
answers.

LEXROOM

[Homepage](#)
[Storico Ricerche](#)
[Libreria Privata](#)
[News](#)

Chiedi a Lexroom

Il vigente art. 182, comma secondo, c.p.c., consente di sanare l'inesistenza o la mancanza in atti della procura alle liti?

Crea ▾

Fonti caricate (3)

file_name_01.docx1.2 MB

file_name_01.docx1.2 MB

file_name_01.docx1.2 MB

Fonti (3)

Visualizza tutte le fonti →

Alla luce della giurisprudenza più recente, il vigente art. 182, comma 2, c.p.c., nella formulazione introdotta dalla legge n. 69 del 2009, non consente di sanare l'inesistenza o la mancanza in atti della procura alle liti. Le Sezioni Unite della Corte di Cassazione, con la sentenza n. 37434 del 21 dicembre 2022, hanno affermato il seguente principio di diritto:

Questa pronuncia ha risolto un contrasto interpretativo esistente in giurisprudenza, stabilendo che la sanatoria prevista dall'art. 182, comma 2, c.p.c. si applica solo ai casi di nullità della procura, non anche all'inesistenza o alla mancanza della stessa.

CIT. 5 - NORMATIVA

"Il vigente art. 182, comma secondo, cod. proc. civ., non consente di 'sanare' l'inesistenza o la mancanza in atti della procura alla lite"

Scrivi qui per un approfondimento della risposta



From your
ideas to the
first draft.
In minutes.

Redazione Atto

Revisionsi

Modifiche disponibili: 3/3

Stiamo utilizzando la tua ricerca come contesto per la redazione di un atto.

Lexroom Friday 2:20pm

Stiamo utilizzando la tua ricerca come contesto per la redazione di un atto.

Tu Friday 2:20pm

Stiamo utilizzando la tua ricerca come contesto per la redazione di un atto.

file_name_01.docx1.2 MB

file_name_01.docx1.2 MB

file_name_01.pdf1.2 MB

Lexroom Friday 2:20pm

Versione 2

Esempio di testo per versione o modifica.

Visualizza Versione

Redigi un ricorso semplificato di cognizione ex art. 281-undecies c.p.c. davanti al giudice di pace in materia del danno di natura patrimoniale prodotto da sinistro stradale. Estrapola lo scenario di riferimento dai documenti caricati.

↺v5↻

Fonti caricate (3)

file_name_01.docx1.2 MB

file_name_01.docx1.2 MB

file_name_01.docx1.2 MB

Fonti (3)

Visualizza tutte le fonti →

Redazione atto

Chiedi una revisione

TRIBUNALE DI ...

UFFICIO DEL GIUDICE DI PACE

RICORSO SEMPLIFICATO DI COGNIZIONE EX ART. 281 UNDECIES C.P.C.

Per il Sig./la Sig.ra ..., nato/a a ..., il ..., residente in ..., Via ..., n. (C.F. ...) rappresentato/a e difeso/a, giusta procura da considerarsi apposta in calce al presente atto, dall'Avv. ..., del Foro di ... (C.F. ..., P.e.c. ..., Fax ...) ed elettivamente domiciliato/a presso lo studio dello stesso in ..., via ...;

- Ricorrente -

contro

Il Sig./la Sig.ra nato/a a ... il ... (C.F. ...), residente in ... Via... n. ...

- Resistente -

Ai sensi dell'art. 2, comma 1, lett. C), del Decreto Ministeriale n. 110 del 7 agosto 2023, si indicano di seguito le parole chiave, che individuano l'oggetto del presente giudizio: Sinistro stradale; Risarcimento danni; Danni materiali; Art.

Lexroom's intelligence, directly in Word.

[illegible]

Lexroom. The new *legal* standard.

Scalability

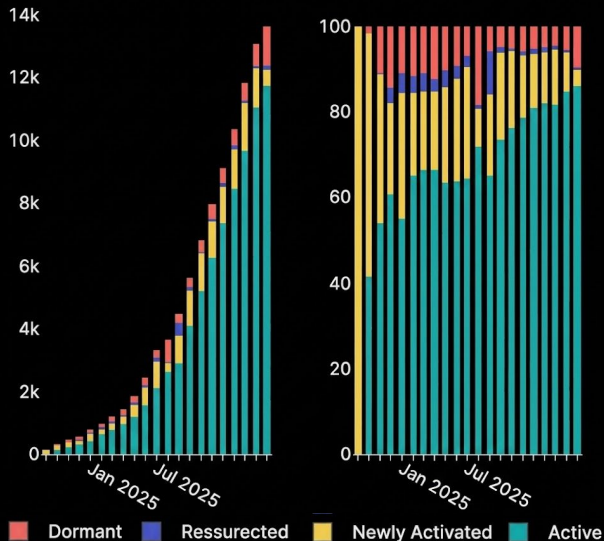
LEXROOM

From a Tool To a habit

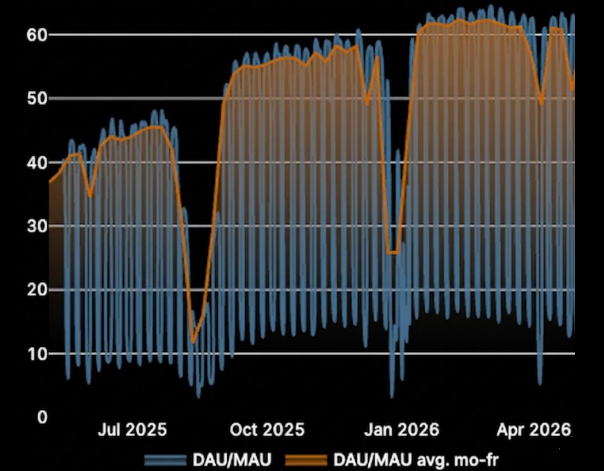
From 35% to +60% DAU/MAU in 12 months.

Across >10k customer.

Monthly Active Users (MAU)



Daily Active Users (DAU)/MAU



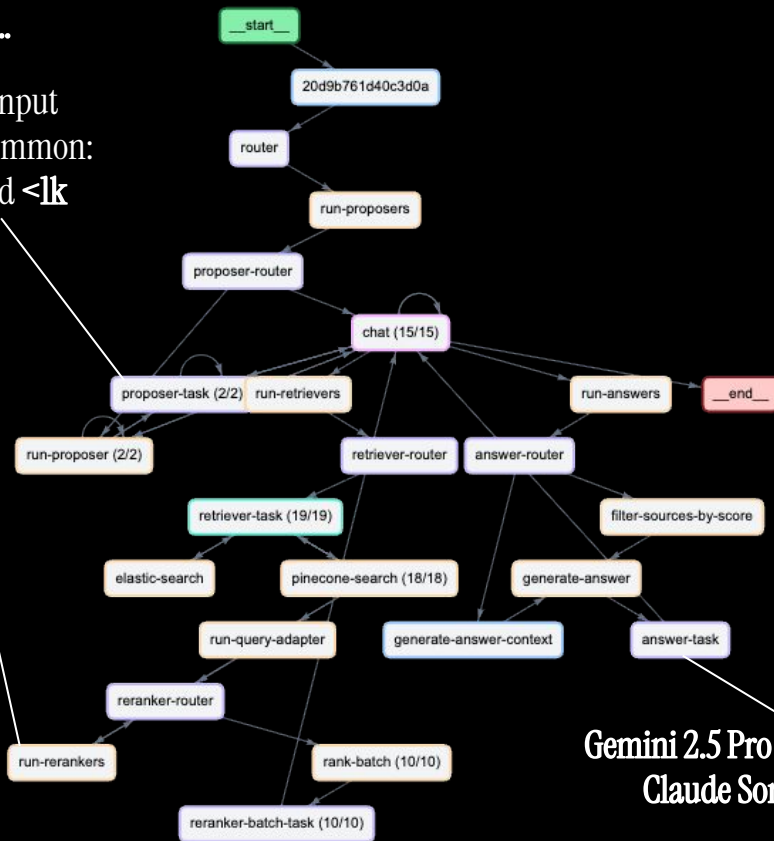
Under the Hood

One user query translates to ~60-70 tasks and subtasks, most of them relying on LLM calls.

Calls can be **very** heterogeneous in terms of prompt and output sizes (e.g. structured output vs. final response).

Gemini 3.1 Flash Lite /
GPT-5.4 mini / ...

On average **8k** input tokens (most common: 1-2k at 35%) and **<1k** output tokens



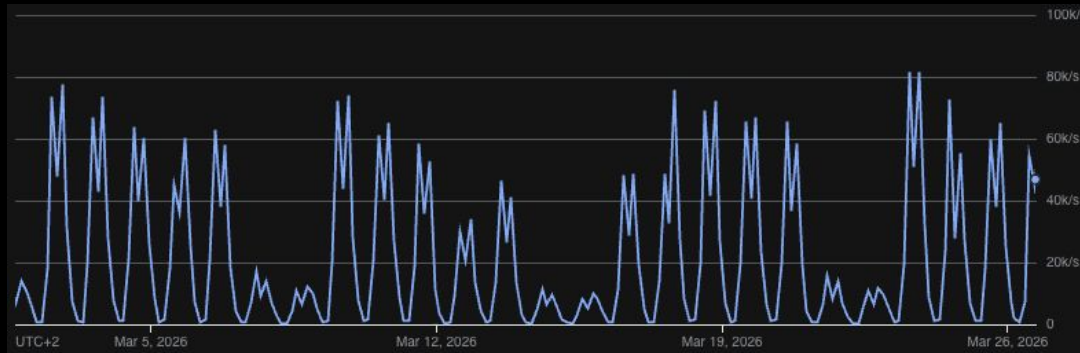
Gemini 2.5 Pro / GPT 5.4 /
Claude Sonnet 4.6 / ...

On average **33k** input tokens
and **4k** output tokens

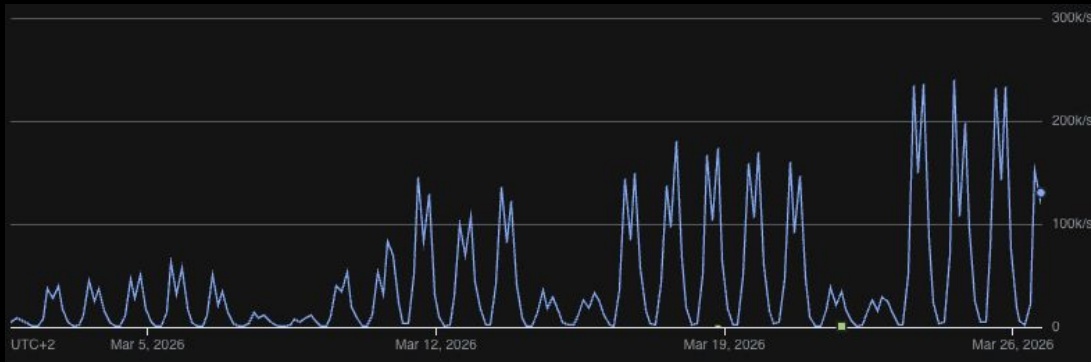
How does this Translate to Token Throughput?

Very “spiky” traffic, roughly
matching peak business hours (**70k
token/s** for final answer generation,
>200k token/s for intermediate
tasks).

Gemini 2.5 Pro (0.32 avg. QPS, 3.7x peak QPS)



Gemini 2.5 Flash (1.13 avg. QPS, 23.3x peak QPS)



How do we Handle High Traffic Loads?

First of all: a **robust** and **region-aware** handoff between Lexroom LLM orchestration and LLM providers.

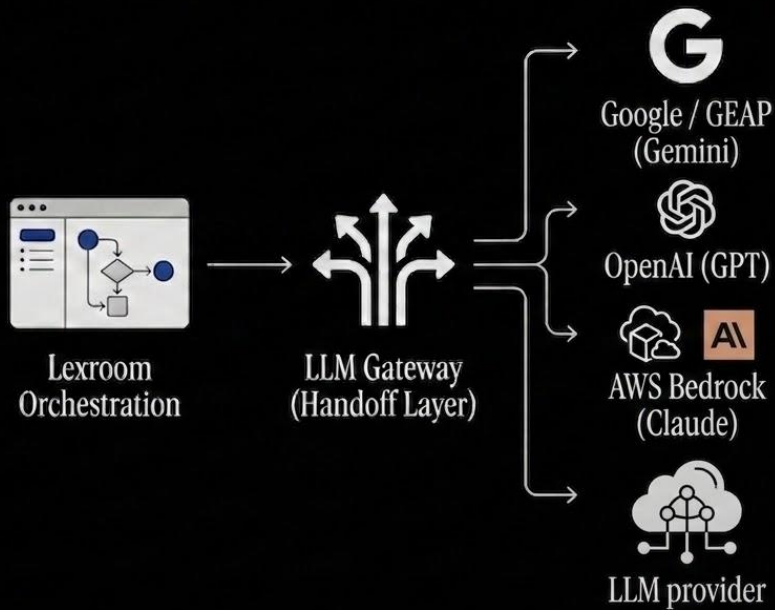
A mix of different strategies (all working together):

1. **Routing to multiple regions and providers.** We (a) rely on different LLM providers (Google/GEAP, AWS Bedrock, OpenAI) and (b) distribute requests across all supported EU regions (7 in Google/GEAP, 4-6 regions in AWS)
2. **Smarter retries** . Many LLM provider SDK includes native retry behavior, but generalized solutions (like [Tenacity](#) in Python) work better with 429 Resource Exhausted errors thanks to techniques like exponential backoff and jitter
3. **Fallback models.** All LLM requests have a primary model *and* a chain of fallback models. Fallbacks get triggered in sequence when (a) a request fails or (b) a request reaches a predefined timeout
4. **Scale down latency-impacting parameters (e.g reasoning effort)** . Latency and stability are highly affected by size of the input and thinking budget/reasoning effort. With high-traffic scenarios and/or fallback requests we reduce both.

How do we Handle High Traffic Loads?

First of all: a **robust** and **region-aware** handoff between Lexroom LLM orchestration and LLM providers.

We use our own **LLM Gateway** to (1) implement the handoff layer as *decoupled* from our agentic workflows, (2) have full control on *how* we interact with LLM providers, and (3) take care of all the bookkeeping (e.g. tracking of costs, token throughput, latencies per model)



How do we Handle High Traffic Loads?

Last resort: **Provisioned Throughput** to reserve a certain fraction of token throughput.

Most LLM providers (Google GEAP, AWS Bedrock) have cost models that revolve around two different paradigms:

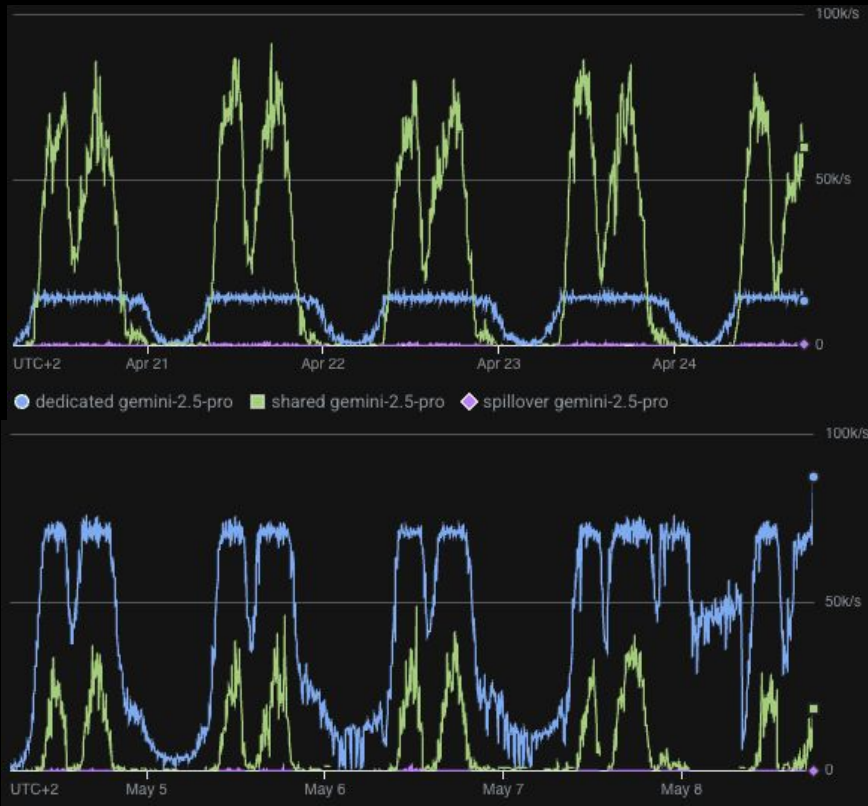
- **Pay-as-you-go** (token-based, usually billed per **1M** input/output tokens)
 - ✓ Costs scale flexibly with the token throughput
 - ✗ Treated with lower priority by the provider, uses **shared** endpoints so it is subject to resource availability
- **Provisioned throughput** (resource-based, usually billed based on a measure of *guaranteed* token throughput)
 - ✓ Provider reserves **dedicated** hardware, so treated with highest priority until the provisioned throughput is reached
 - ✗ Fixed monthly cost, depends on the minimum throughput to guarantee

How do we Handle High Traffic Loads?

Last resort: **Provisioned**

Throughput to reserve a certain fraction of token throughput.

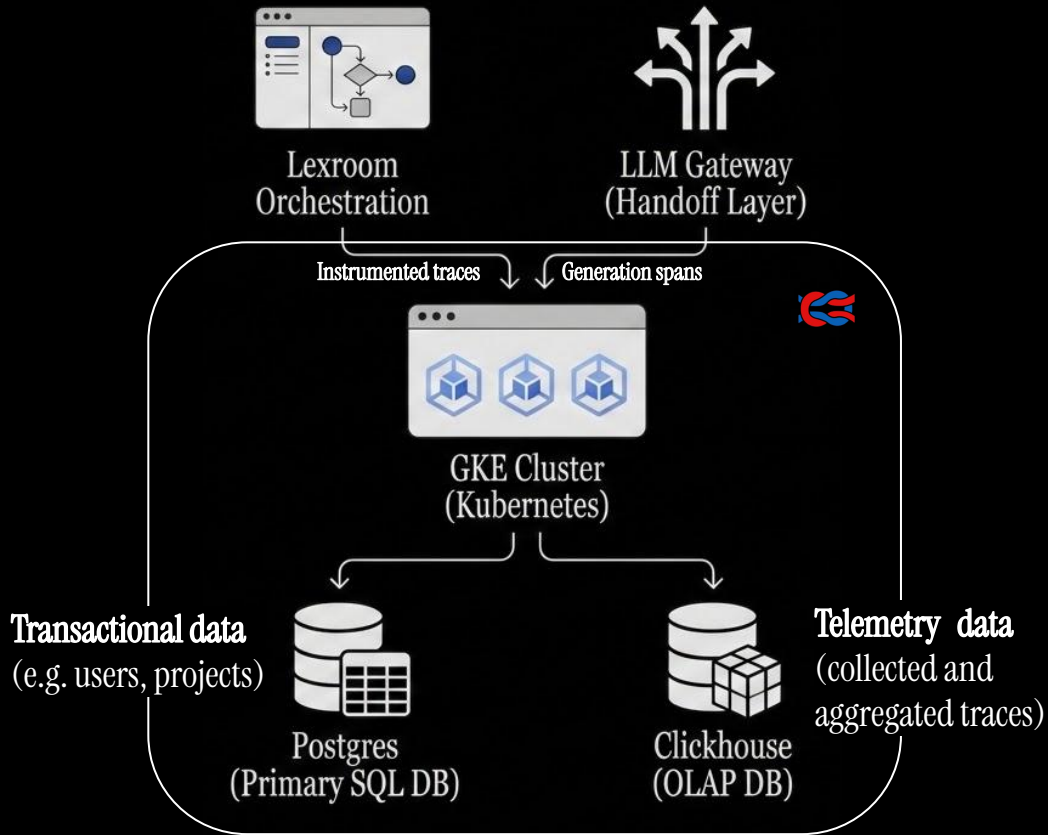
22 GSU (0.2 QPS, 14k token/s) + ~80% PayGo



109 GSU (1 QPS, 70k token/s) + ~30% PayGo

LLM Observability at Scale

A self-hosted **Open Telemetry (OTel)** framework (**Langfuse**) to monitor logs, track metrics, in-depth debugging whenever needed.



LLM Observability at Scale

An **Open Telemetry** (OTel) framework to monitor logs, track metrics, in-depth debugging when needed

Why self-hosted ?

1. **Better scalability** . Cost increases with infrastructure sizing and maintenance effort (vs. trace-based or span-based pricing)

row_count	day
4.76 million	2026-06-11
7.38 million	2026-06-12
1.76 million	2026-06-13
1.66 million	2026-06-14
9.12 million	2026-06-15
9.55 million	2026-06-16
9.52 million	2026-06-17
3.44 million	2026-06-18

60% of the **monthly** amount of spans generated (in November 2025)

2. **Full control over telemetry data** . Important for compliance: we decide on data retention and per-user deletion (GDPR), we can implement audit logs, we can handle span anonymization and PII removal (1) asynchronously, and (2) with our own anonymization models

Privacy and Compliance at Scale

We do **not** compromise on privacy and compliance, *even when* it gets in the way of scalability.

We ensure GDPR compliance through three core pillars:



EU Data Sovereignty: We route production LLM requests exclusively through *regionalized*, EU-only endpoints with *Zero Data Retention* (ZDR).



PII and Anonymization: We strip user document content and apply anonymization on generation spans asynchronously and across languages.



No Training on User Data: We never use user data for training, fine-tuning, or model evaluation.

Lexroom. The new *legal* standard.

Specialization

LEXROOM



Why not just a *generalist* LLM?

(with clever prompt engineering
and/or a few Markdown files)



Claude for the legal industry

We're releasing 20+ new MCP connectors that link Claude to the software that the legal industry runs on and 12 new plugins tailored to specific legal work and practice areas.



 **Claude**
For Legal

Specializing LLMs to the Legal Domain

A handful of instructions is not enough; we need a **legal-specific harness** .

From seminal evaluation studies  on **Legal AI systems** (not even generalist LLMs) we learn that basically *every single task* in a legal workflow (search, rank, etc.) require *in-depth legal reasoning*, and can easily fail when offloaded to generic tools.

TABLE 6 | This table shows the prevalence of different contributing causes among all hallucinated responses for each model. Because the types are not mutually exclusive, the proportions do not sum to 1.

Contributing cause	Lexis	Westlaw	Pract. Law
Naive Retrieval	0.47	0.20	0.34
Inapplicable Authority	0.38	0.23	0.34
Reasoning Error	0.28	0.61	0.49
Sycophancy	0.06	0.00	0.03

On top of that: different countries can rely on *completely different* legal system and paradigms, e.g. **Civil Law** vs. **Common Law** .



Specializing LLMs to the Legal Domain

A handful of instructions is not enough; we need a **legal-specific harness** .

Two prominent **Italian-specific** examples that are just about *search* in the legal domain:



The **Cartabia Reform** (D.Lgs. 149/2022, D.Lgs. 150/2022), as a sweeping overhaul of the Italian justice system, requires:

- Careful handling of normative vs. case law search
- Explicit temporal reasoning to integrate sources *before* and *after* the reform (with the former being a much larger data pool!)



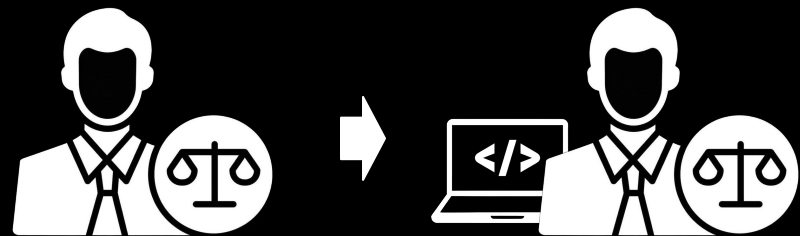
The **Favor Rei** principle in Criminal Law, which dictates that the effects of a law should be those of the revision of that law that's *most favorable* to the offender (also retroactively). In other words, we need:

- Full versioning of Criminal Law data
- Explicit temporal reasoning to integrate search results with available contextual information

Specializing LLMs to the Legal Domain

A handful of instructions is not enough; we need a **legal-specific harness** .

In Lexroom we don't have **Legal Experts** , we have **Legal Engineers** .

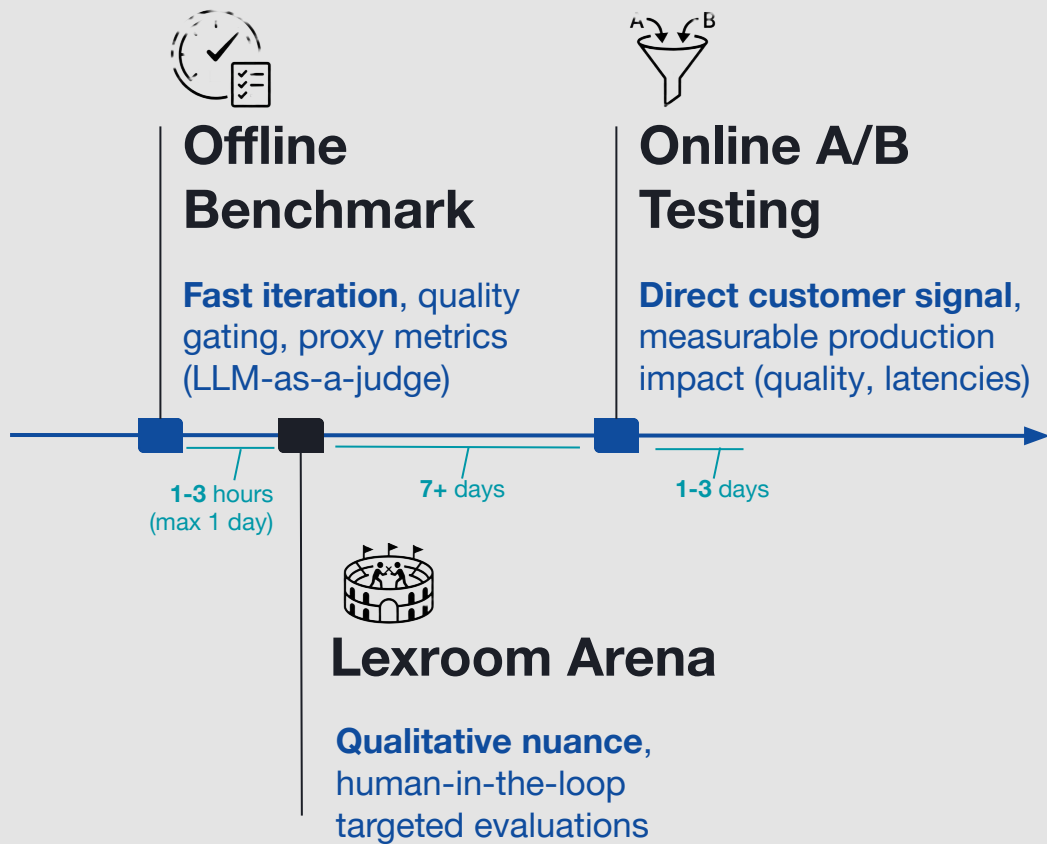


Activity	Partner tech team(s)
Own <i>feature shaping</i> for Lexroom products	Product
Own and develop <i>versioned legal knowledge</i> (e.g. prompts, skills, sources metadata)	Data, AI & Search
Translate back and forth between <i>legal requirements and technical requirements</i> (e.g. Favor Rei support → Criminal Law versioning)	Product, AI & Search



How do We Evaluate which LLMs to Use?

A tiered evaluation strategy as a tradeoff among **speed**, **nuance**, and **customer-first** principles.





How do We Evaluate which LLMs to Use?

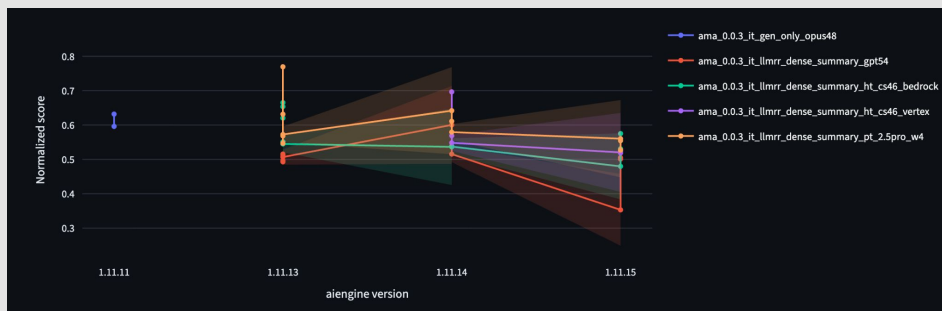
A tiered evaluation strategy as a tradeoff among **speed**, **nuance**, and **customer-first** principles.



Offline Benchmark

Fast iteration, quality gating, proxy metrics (LLM-as-a-judge)

- **100+** (and counting) end-to-end *question-answer pairs with grounded sources*, spanning different modules and production use cases
- Detailed human-crafted evaluation rubrics (>**30**) per question-answer pair, runnable with an *LLM-as-a-judge harness*
- Internal leaderboard to compare different versions of the Lexroom AI Engine and/or different generalist LLMs



How do We Evaluate which LLMs to Use?

A tiered evaluation strategy as a tradeoff among **speed**, **nuance**, and **customer-first** principles.

ELO scoring
and qualitative
evaluation



Lexroom Arena

Qualitative nuance,
human-in-the-loop
targeted evaluations

Head-to-head
comparison
between two
Lexroom versions

LEXROOM

Chiedi a Lexroom

Homepage

Storico Ricerche

Libreria Privata

Arena: AMA: GPT-5.4 vs Opus 4.8

RISPOSTA A

Forti

La clausola **claims made** nei contratti di assicurazione della responsabilità civile professionale è il meccanismo per cui la copertura assicurativa non dipende soltanto dal momento in cui è stato commesso il fatto dannoso, ma anche – e soprattutto – dal momento in cui la **richiesta risarcitoria** viene formulata e denunciata all'assicuratore

CIT.1

CIT.2

CIT.3

CIT.4

RISPOSTA B

Forti

I contratti di assicurazione della responsabilità civile professionale con clausola "claims made"

1. Inquadramento e funzionamento della clausola

Il contratto di assicurazione della responsabilità civile professionale con clausola "claims made" (letteralmente "a richiesta fatta") costituisce un **deroga convenzionale** al modello legale tipico delineato dall'art. 1917, primo comma, c.c. Per comprenderne il funzionamento, occorre anzitutto raffrontarlo con lo schema codicistico.

Secondo il modello dell'art. 1917, comma 1, c.c. – definito **"loss occurrence"** o "insorgenza del danno" – la copertura assicurativa opera

CIT.1

"In relazione a tutte le condotte generatrici di domande risarcitorie insorte nel periodo di durata

1. Modello legale e deroga convenzionale

Il modello codicistico dell'assicurazione della responsabilità civile è quello dell'art. 1917 c.c., secondo cui:

CIT.4

"Nell'assicurazione della responsabilità" civile l'assicuratore e' obbligato a tenere indenne l'assicurato di quanto questi, in conseguenza del fatto accaduto durante il tempo dell'assicurazione, deve pagare a un terzo, in dipendenza della responsabilità' dedotta nel contratto."

Valutazione

Quale risposta preferisci?

Config A

Config B

Pareggio

Entrambi male

Config A

Config B

Logica espositiva

Proprietà del linguaggio

Completezza delle fonti

Qualità della risposta

Motivazione

Spiega brevemente la tua scelta...

Salva valutazione

AI & Search @ Lexroom



Claudio
Delli Bovi

TECH LEAD



Andrea
Ponti

AI ENGINEER



Daniele
Brambilla

AI ENGINEER



Emanuele
Costantini

AI ENGINEER



Giuseppe
Mantineo

AI ENGINEER



Simone
Pierro

AI ENGINEER



We're
hiring!

AI ENGINEER

Q&A

