

LLMs for patent and IP work: the dos and don'ts

Why an LLM gets IP facts wrong, the architecture that reduces that risk, and what to demand from your data and workflow.

Kacper Gorski · Lighthouse IP



INTRODUCTION

Why we are all here

More IP teams are building their own AI workflows, and the flows are getting more complex.

In-house

Teams are building their own workflows, not waiting for a vendor to ship one.

Proprietary

What you search, what you compare, which fields you trust. That is your know-how.

Deliberate

The edge goes to the team that decides where the model sits and where the data comes from.

WHY NOW

Much of IP work lives in proprietary know-how: how a team searches, compares and judges. Teams are not looking to replace that with AI. They are looking to extend it.

DO

Decide where AI sits in your process, and what data it reaches, before you pick a tool.

DON'T

Hand a proprietary process to a general-purpose chatbot and hope.

LLMs are not patent databases

Training on patent text does not make a model a patent database.

It predicts the next token. Low temperature reduces sampling variation. It does not guarantee the same answer twice.

Without retrieval, its knowledge is a partial training snapshot. The Pile's patent slice: US grant background sections, 1976 to 2020. No claims, no other offices.

Without retrieval, a citation in the answer was generated, not looked up.

DO

Use it for language: drafting, summarising, restructuring what you give it.

INSIDE THE MODEL

Language patterns. A snapshot to a date. No lookup, no index, no record.

INSIDE AUTHORITATIVE PATENT DATA

Publication numbers. Bibliographic fields. Claims. Descriptions. Dates. Office-published events. Every fact tied to a source record.

DON'T

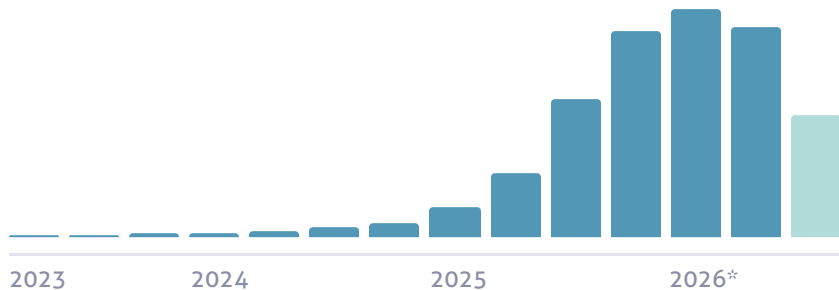
Ask it what a patent says from memory.

The hallucination problem has changed

Law runs on precedent and chains of cited information. A hallucination breaks the chain.

2,041

court decisions involving AI hallucinations, per quarter



Source: AI Hallucination Cases Database, Damien Charlotin, CC BY 4.0, 14 Sept 2026. *Q3 2026 partial.

OLD PROBLEM

The model invents the source. *Mata v. Avianca*, June 2023: six invented decisions, the first widely reported sanctions. A 2023 study found fabricated citations in 55% of GPT-3.5 answers and 18% of GPT-4.

NEW, ADDITIONAL PROBLEM

Retrieval reduces it. It does not remove it. Stanford RegLab tested the commercial legal research tools sold as 'hallucination-free' RAG: Lexis+ AI and Westlaw AI-Assisted Research hallucinated on 17% to 33% of queries. Lexis+ AI, the best performer, answered 65% accurately.

'It searched the web' is not 'it checked the register'. Neither is 'it used RAG'.

DO

Ask where each fact came from before you ask whether it is right.

DON'T

Let 'it retrieved something' replace 'it checked the register'.

Six ways an IP AI workflow can fail

Probabilistic generation is only one failure mode. The others sit in the data, retrieval and workflow around the model.

Probabilistic generation

The model predicts a distribution over next tokens. Lower temperature reduces sampling variation; it does not make unsupported facts true.

Data cleanliness

OCR errors, stray characters, duplicate records in the source. The model reproduces what it was given.

Data linkage

Missing or incorrect family, priority or legal-event links send retrieval to the wrong sibling, parent or version.

Source and provenance errors

A mis-keyed date at the office, a mislabelled field at an aggregator, a wrong paste in the prompt.

Missing or overloaded context

The needed record is not retrieved, is truncated, or is buried among distractors. The model may answer anyway.

No retrieval

Nothing authoritative was looked up. Patent facts come from parametric knowledge rather than a current record.

DO Diagnose which layer failed before changing the model.

DON'T Assume every wrong answer is a model problem.

The reliability stack: RAG and beyond

Each layer addresses a different failure mode. A more capable model cannot compensate for a missing data layer.



If the answer does not carry the publication number and the field it came from, it is not evidence yet.

DO Build the stack first, then choose the model.

DON'T Buy a bigger model to fix a missing data layer.

How models actually get IP data

Four doors. Match the door to the user.

DISCOVERY

Web search

Broad and convenient. Often lands on an aggregator, not the issuing office.

PRODUCT BUILDERS

API

Query on demand. Freshness depends on the provider's update cycle. Metered.

CHAT USERS

MCP

An open standard, so any compatible client calls that API as a tool. One server, many clients.

DATA TEAMS

Bulk

Download, index and refresh it yourself. Most control, most engineering.

The public web is not the global patent system.

DO

Give chat users a governed connection instead of copy-paste.

DON'T

Build a connector per model when a standard exists.

Why the underlying corpus matters

A global patent corpus takes infrastructure that sits outside ordinary web search.

OFFICE-PUBLISHED BULK FULL TEXT

A handful of authorities*

The rest: one document at a time, licence-gated, aggregator copies, or paper.

THE GLOBAL PATENT SYSTEM

170+ issuing authorities

Each publishing in its own format, language and cadence.
Closing that gap is the work.

DO

Ask any data source: what, from when, which fields, refreshed how.

DON'T Accept 'AI-ready' without asking what is inside.

What it takes. Sourcing in whatever form each office publishes, then OCR, translation, one normalised format, and continuous work on the linkages.

Fields, not just documents. Legal events are published facts. A 'legal status' flag is a derived interpretation.

Freshness is not trust. A refresh timestamp says when a source was queried, not whether the record is safe to use.

* Internal assessment, June 2026.

What unit should AI retrieve? The Keeper example

Keeper is Lighthouse IP's retrieval layer: the corpus embedded as vectors for AI search. The chunk you embed is the chunk you get back, so we index at two levels.

DOCUMENT LEVEL

Title, abstract, claim 1

One vector per record across the corpus. Compact and fast. For landscaping, competitive intelligence and first-pass screens.

SENTENCE LEVEL

Every sentence of the full text

Points at the sentence that matched. For claim mapping, FTO and patentability. Heavier index.

Point lookup by publication number is easy. Open-ended FTO is a retrieval problem, and there the chunking strategy decides what the model sees.

DO Choose granularity by task, not by default.

DON'T Chunk everything at a fixed size and hope.

Privacy and governance

Consumer products do not automatically give confidential IP work the governance it needs.

CONSUMER ENVIRONMENT

Training, retention and deletion settings vary by vendor, tier and account configuration. Limited organisational governance and auditability.

APPROVED ENTERPRISE OR API ENVIRONMENT

Customer data typically excluded from training by default. Contractual data-handling controls. Defined retention, access control, zero-data-retention where supported. Logging of retrieval and verification.

USPTO, April 2024: AI use does not remove the practitioner's duty to review and verify. **ABA Opinion 512** says the same for lawyers. **NIST AI 600-1** names confabulation as a generative-AI risk to be measured and managed.

DO

Map each workflow to an environment before anyone pastes a document.

DON'T

Put confidential material into an environment whose terms you have not checked.

A possible future: own model, own data, own costs

Open weights are catching up. The whole stack can live inside your walls.

Own model

The best open-weight models trail the closed frontier by about four months on average (Epoch AI, May 2026), and they are yours to download and run.

Own data

A private model on your own hardware or tenancy, connected to your own patent database. Air-gapped if you need it. Prompts and documents stay inside your controlled environment.

Own costs

Trade variable API consumption for infrastructure and operating costs you control more directly. It pays at sustained utilisation, not for ten questions a week.

The model is the easy part to own. The corpus, the index and the refresh feed still have to come from somewhere.

DO

Decide the data first, then the hosting.

DON'T

Buy GPUs before you have a corpus to put behind them.

Five dos, five don'ts

DO

- 1 Treat the model as a writer and reasoner, not a patent database.
- 2 Ground answers in authoritative records and preserve provenance.
- 3 Retrieve only the records and matter content the question needs.
- 4 Verify before reliance or filing.
- 5 Use an approved environment appropriate to the sensitivity of the data.

DON'T

- 1 Ask a model what a patent says from memory.
- 2 Equate web search with checking the patent register.
- 3 Put confidential material into an environment whose terms you have not checked.
- 4 Accept 'AI-ready' without understanding corpus, fields, freshness and chunking.
- 5 Assume owning or self-hosting the model solves the data problem.

ABOUT US

Lighthouse IP, in one slide

You are probably already touching our data every day.



200

people, five continents

170

patent authorities

182M+

patent documents

Weekly

refresh, direct from the office

WHO WE ARE

A Dutch data company, founded 2006, part of the Fovea IP Group. We source patent, trademark and design data directly from the publishing authorities, in the countries where the offices are. Our data sits under an estimated 70% of the IP analytics market.

WHAT WE DO

Full text from 83 patent authorities, about 200 trademark registers, 100+ design authorities. One normalised format, enriched with legal events, ownership and corporate-tree links. Delivered as bulk files, then API, and now Keeper, our vectorised retrieval layer for AI search.

DO

Ask us for the coverage file and benchmark your current source against it.

DON'T

Take our word for it either. Ask us the same questions as any other supplier.

Do not ask only 'Which model are we using?'

Ask 'What data can it reach, where did that data come from, and can I verify every important fact?'

Questions

Kacper Gorski · Lighthouse IP

SOURCES

Further reading

DUTIES AND GOVERNANCE

USPTO: Guidance on use of AI-based tools in practice (11 Apr 2024)

[federalregister.gov](https://www.federalregister.gov) · 89 FR 25609

USPTO: Artificial intelligence initiatives

[uspto.gov/initiatives/artificial-intelligence](https://www.uspto.gov/initiatives/artificial-intelligence)

ABA Formal Opinion 512: Generative AI tools (29 Jul 2024)

[americanbar.org](https://www.americanbar.org) · [aba-formal-opinion-512.pdf](https://www.americanbar.org/files/other_publications/supplements/2024/07/aba-formal-opinion-512.pdf)

NIST AI Risk Management Framework

[nist.gov/itl/ai-risk-management-framework](https://www.nist.gov/itl/ai-risk-management-framework)

NIST AI 600-1: Generative AI Profile (Jul 2024)

doi.org/10.6028/NIST.AI.600-1

EVIDENCE

Magesh, Surani, Dahl, Suzgun, Manning and Ho: Hallucination-Free? Assessing the reliability of leading AI legal research tools (J. Empirical Legal Studies, 2025)

arxiv.org/abs/2405.20362

Charlotin: AI Hallucination Cases Database (live)

damiencharlotin.com/hallucinations

Walters and Wilder: Fabrication and errors in ChatGPT citations (Sci. Rep., 2023)

[ncbi.nlm.nih.gov/pmc/articles/PMC10484980](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10484980)

Whitlock and Wetzel: Generative AI patent searching performance (2026)

open.clemson.edu/jptrca/vol36/iss1/2

Chroma: Context rot, how input length degrades LLM performance (Jul 2025)

trychroma.com/research/context-rot

Thinking Machines: Defeating nondeterminism in LLM inference (Sept 2025)

thinkingmachines.ai/blog

STANDARDS AND ENGINEERING

W3C PROV: provenance data model, overview

[w3.org/TR/prov-overview](https://www.w3.org/TR/prov-overview)

Model Context Protocol (MCP), specification and docs

modelcontextprotocol.io

Anthropic: Effective context engineering for AI agents (Sept 2025)

anthropic.com/engineering

Epoch AI: open-weight models vs the closed frontier (May 2026)

epoch.ai/data-insights/open-closed-eci-gap

The Pile: an 800GB dataset for language modeling (2020)

arxiv.org/abs/2101.00027

The cost of getting it wrong

Consequences recorded in the AI Hallucination Cases Database, from a warning to a lost licence.

Warnings and struck filings

Where it usually starts: a show-cause order, a filing struck, a warning on the record.

Financial sanctions

Fines and fee awards against the party or the lawyer. *Mata v. Avianca*, June 2023: \$5,000, the first widely reported case.

Professional referrals

Referral to the bar or the regulator, in the US and elsewhere.

Suspension from practice

Suspension or loss of licence has followed in more than one jurisdiction, usually where the response after detection made it worse.

Reputational damage

The cases are public, searchable, and increasingly reported by name.

What turns a warning into a suspension. Usually not the hallucination. It is what happens after detection: refusing to own up, doubling down, blaming the intern. The database's own maintainer says so.

39%

of cases had a practising lawyer as the responsible party: 653 of 1,668, on 2 July 2026. Not only self-represented litigants.

DO

Verify before filing. If an error gets through, correct the record promptly.

DON'T

Double down, or blame the tool or the intern.