# Diritto, LLM e ricerca

## Fondamenti teorici, protocolli operativi e criteri di validità

**MARCO PAGANI**

Lidia
Adaptive Legal Assistant

*In questo saggio si delineano le condizioni per integrare i Large Language Model (LLM) nella ricerca giuridica senza cadere nelle illusioni della plausibilità linguistica alla conoscenza giustificata. Si assume che gli LLM siano dispositivi di modellazione statistica del linguaggio; ne consegue che l'impiego è legittimo solo entro un impianto metodologico che subordini ogni affermazione rilevante a fonti identificabili, verificabili e correttamente citate (regola cite-or-silent). Il contributo si articola in tre direttrici: (i) un quadro teorico che distingue competenza apparente, inferenza probabilistica e giustificazione normativa; (ii) protocolli operativi per issue spotting, governance delle fonti (primarie, secondarie, giurisprudenza), impiego di sistemi RAG e tracciabilità mediante audit trail; (iii) criteri di validità e metriche di qualità, sia ex ante (idoneità e copertura delle fonti) sia ex post (replicabilità, controllo degli errori, accountability). Ne risulta un'architettura della fiducia fondata sulla separazione dei ruoli: definizione dell'oggetto, delle ipotesi e dell'interpretazione alla competenza del giurista; supporto esplorativo-organizzativo all'LLM (classificazione, deduplicazione, normalizzazione, sintesi con citazioni). Obiettivo finale è un metodo trasparente, replicabile e professionalmente esigente, capace di coniugare efficienza tecnica e rigore epistemico nella comunità legale.*

# Gli LLM non sono ricercatori legali

I modelli linguistici di grandi dimensioni non svolgono ricerca giuridica nel senso proprio del termine. La loro straordinaria fluidità testuale deriva dall'addestramento statistico sul linguaggio, non da un processo di selezione, qualificazione e controllo delle fonti secondo gerarchie normative e canoni interpretativi. In breve: competenza linguistica non equivale a competenza epistemica; la plausibilità formale non garantisce validità giuridica.[1]

Gli studi empirici più recenti mostrano che, anche quando gli strumenti "verticali" per la ricerca giuridica integrano meccanismi di recupero documentale, permangono errori sottili e insidiosi: citazioni fuorvianti, applicazione decontestualizzata di principi, difficoltà nel riconoscere l'overruling o nel rispettare l'ordine delle fonti. L'effetto è particolarmente grave in domini ad alto rischio, dove la forma citazionale non basta senza verifica sostanziale.[2]

Un ulteriore fattore di fragilità è la dipendenza dall'incorniciamento del compito: istruzioni incomplete, ridondanti o sbilanciate possono indurre risposte eccessivamente sicure o, all'opposto, rinunce immotivate, generando schemi di errore sistematici.[3] Queste distorsioni di ragionamento, ben note nella pratica con i sistemi conversazionali, non sono anomalie contingenti ma conseguenze attese del modo in cui i modelli sono addestrati e valutati: l'architettura premia la "risposta plausibile" anche quando l'incertezza dovrebbe imporre cautela.[4]

Neppure l'integrazione di basi documentali tramite recupero (RAG) neutralizza il problema alla radice: la qualità dell'output resta condizionata dalla pertinenza del recupero, dalla corretta qualificazione della fonte e dalla sua attualità. In assenza di una regia umana che filtri, qualifichi e contestualizzi, gli LLM producono spesso "pseudo-conoscenza": enunciati ben scritti ma epistemicamente deboli.[5]

**Chiosa operativa.** Gli LLM possono assistere (classificare, riassumere, proporre contro-tesi), ma non sostituire le funzioni critiche di qualificazione e validazione. La responsabilità epistemica — tracciabilità integrale delle fonti e controllo della loro forza normativa — resta in capo al professionista.[1]

---

1. *L. Paseri – M. Durante, Examining epistemological challenges of large language models in law, Cambridge Forum on AI: Law and Governance, 1 (2025), e7, doi:10.1017/cfl.2024.7. L'articolo argomenta, tra l'altro, la distinzione fra rappresentazione linguistica prodotta dai modelli e processi di validazione propri del diritto (qualificazione, affidabilità, rapporto con verità e accuratezza).*
2. *F. Surani – V. Magesh – M. Dahl, Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools, Journal of Empirical Legal Studies, 2025, doi:10.1111/jels.12413.*
3. *Z. Ling et al., Instruction Boundary: Quantifying Biases in LLM Reasoning under Various Coverage, arXiv:2509.20278 (v1, 2025).*
4. *A. T. Kalai – O. Nachum – S. S. Vempala – E. Zhang, Why Language Models Hallucinate, preprint (4 settembre 2025).*
5. *F. Surani – M. Dahl – V. Magesh, Legal RAG Hallucinations (materials e rubriche di valutazione).*

# La percezione dell'AI nella pratica legale

Nonostante la crescente diffusione degli LLM nella professione legale, la loro percezione risente ancora di rappresentazioni semplificate. In particolare, l'idea che questi strumenti siano capaci di svolgere attività "intellettuali" in modo analogo a un praticante esperto si è rapidamente diffusa, anche in contesti altamente qualificati. Questa visione, pur comprensibile, non coglie la natura tecnica degli LLM: modelli statistici in grado di generare testo plausibile, ma privi di consapevolezza normativa o capacità di verifica autonoma (vedi Appendice 1.2, A1).[1]

Un elemento centrale è l'interfaccia conversazionale: quando un LLM restituisce una risposta ben formata e coerente, l'utente tende a sovrainterpretarne l'autorevolezza. Questo fenomeno – talvolta descritto come user illusion – è tanto più forte quanto minore è la familiarità con le logiche di funzionamento del modello. Il rischio non è l'errore isolato, ma la fiducia automatica che può innescarsi anche tra professionisti esperti, specie sotto pressione o in assenza di un controllo strutturato del processo (vedi Appendice 1.2, A2).[2]

Un secondo aspetto riguarda la cosiddetta sycophancy: il modello tende ad assecondare le premesse della domanda, anche se errate. In questo senso, l'output può rinforzare un'ipotesi non fondata anziché metterla in discussione. È un comportamento tecnico atteso, ma poco intuitivo per chi si aspetta un ragionamento critico simile a quello umano (vedi Appendice 1.2, A3).[3]

Infine, il quadro è reso più complesso dalla limitata alfabetizzazione tecnica nel settore legale. Non si tratta solo di "saper usare" un LLM, ma di inserirlo in un processo strutturato che ne valorizzi i punti di forza e ne mitighi i limiti.

In mancanza di un playbook operativo, l'uso resta episodico, talvolta orientato da aspettative non realistiche, con rischio di inefficienza e confusione tra strumenti e metodo (vedi Appendice 1.2, A4).[4]

In sintesi, la percezione dell'AI nella pratica legale è ancora in via di assestamento. La consapevolezza tecnica cresce, ma è necessario rafforzare una cultura d'uso centrata sul processo, non sull'effetto. Questa percezione deformata è spesso amplificata dai claim dei fornitori: esaminiamo ora il marketing dell'AI legale.

---

1. *L. Paseri – M. Durante, Examining Epistemological Challenges of Large Language Models in Law (2025). Si veda anche A. T. Kalai – O. Nachum – S. S. Vempala – E. Zhang, Why Language Models Hallucinate (preprint, 4 settembre 2025)*
2. *Z. Ling et al., Instruction Boundary: Quantifying Biases in LLM Reasoning under Various Coverage (arXiv:2509.20278, 2025).*
3. *Z. Ling et al., Instruction Boundary (cit.).*
4. *F. Surani – M. Dahl – V. Magesh, Legal RAG Hallucinations (materials e rubriche di valutazione, 2025). Lo studio documenta che, senza un playbook che governi recupero, qualificazione e controllo umano, anche citazioni formalmente corrette possono sostenere argomenti decontestualizzati o superati*

# Le illusioni del marketing legale AI

Il marketing dei principali vendor legal tech ha contribuito in modo decisivo alla diffusione degli LLM nel contesto forense, ma spesso ha veicolato aspettative non allineate con le effettive capacità dei sistemi. Espressioni come “hallucination-free” o “trusted AI” sono state usate per suggerire un’affidabilità totale, quasi a indicare che il problema degli errori generativi fosse risolto una volta per tutte. Tali claim, pur efficaci sul piano comunicativo, non reggono al confronto con gli esiti di studi indipendenti.[1]

Un caso emblematico riguarda le promesse di accuratezza citazionale. Lexis+ AI è stato presentato come in grado di fornire “citazioni legali collegate prive al 100% di allucinazioni”; eppure, in uno studio indipendente del 2025, pubblicato sul Journal of Empirical Legal Studies, il sistema ha prodotto informazioni errate o riferimenti fuorvianti in circa un sesto delle risposte esaminate (≈17%), mentre gli strumenti di Thomson Reuters (Ask Practical Law/Westlaw AI) hanno superato un terzo di errori di questo tipo (≈33–34%), nonostante l’uso di fonti verificate e tecniche avanzate di recupero.[2]

Questa distanza fra promesse e performance non implica mala fede, ma evidenzia un nodo metodologico: mancano metriche condivise e replicabili per validare l’output AI in ambito legale. Nella pratica, ogni vendor adotta parametri propri — spesso non trasparenti né riproducibili da terzi — con la conseguenza che la valutazione comparativa risulta impervia e l’utente professionale fatica a distinguere il miglioramento incrementale dall’affidabilità strutturale. Gli stessi studi indipendenti segnalano, inoltre, quanto sia difficile valutare prodotti proprietari in condizioni realistiche d’uso.[3]

Si aggiunge una comunicazione talora ambivalente. Thomson Reuters, ad esempio, precisa che i propri strumenti non intendono sostituire l'avvocato, ma offrirgli una "sintesi iniziale" per accelerare il lavoro; al contempo, sottolinea riduzioni dei tempi di ricerca da 17–28 ore a 3–5,5 ore. Un messaggio di questo tipo può facilmente essere interpretato come promessa di sostituzione operativa, alimentando l'aspettativa che l'AI "risolva" la ricerca più che supportarla.[4]

In sintesi, il linguaggio promozionale contribuisce — spesso in modo involontario — a consolidare l'equivoco tra AI come strumento di supporto e AI come decisore. Per il professionista legale, il rischio non risiede soltanto nell'uso dell'AI, ma nel tipo di aspettativa che si costruisce attorno al suo impiego. La via d'uscita è riportare il discorso su basi verificabili: ancorare ogni uso al contesto operativo, a metriche trasparenti e alle evidenze disponibili.

---

1. *Cfr. Legal RAG Hallucinations.*
2. *Lo studio su oltre 200 interrogazioni legali reali (Lexis+ AI; Thomson Reuters: Ask Practical Law AI e Westlaw AI; confronto con GPT-4) rileva che "gli strumenti RAG-based per la ricerca legale ancora allucinano": oltre un sesto delle risposte di Lexis+ AI e circa un terzo di quelle Westlaw/Thomson Reuters contengono informazioni fuorvianti o non supportate. Il lavoro offre breakdown per tipo di query (giurisdizione, tempo, bar-style, false premise), con differenze significative nelle hallucination rates.*

# Prove alla mano: cosa dicono i test?

Il divario tra le promesse commerciali dell'AI legale e le sue performance effettive è stato oggetto di numerosi test indipendenti negli ultimi diciotto mesi. I dati convergono su una tendenza ricorrente: buoni risultati su compiti semplici e strutturati; decadimento progressivo in presenza di ambiguità, contesto normativo denso o necessità di ragionamento multi-step. La sintesi è chiara: l'LLM può assistere, ma non sostituire, il lavoro interpretativo e critico richiesto dalla ricerca giuridica avanzata.

GPT-4 ha ottenuto un punteggio nel decile superiore al bar exam statunitense, dato spesso addotto come prova della sua "competenza giuridica". Tuttavia, quel test si fonda su domande a scelta multipla e format rigidamente vincolati: in contesti di ragionamento aperto o multi-step, le performance del modello calano sensibilmente.

Uno studio del 2024 basato su LEXAM ha valutato le capacità di GPT-4 in compiti a risposta chiusa e aperta. Mentre nei test a scelta multipla — spesso ancorati a nozioni teoriche — il modello ha raggiunto risultati intorno al 75% di accuratezza, la performance è scesa drasticamente nei task che richiedevano ragionamento articolato su casi giuridici, con un tasso di errore del 43%. In tali prove, l'output è apparso formalmente coerente ma concettualmente fragile: mancata identificazione della norma rilevante, omissione di distinzioni giurisprudenziali, confusione tra ratio decidendi e obiter dictum (vedi Appendice 1.4, A1).[1]

Non solo i modelli generalisti. Anche strumenti specializzati, come Westlaw AI e Lexis+ AI, mostrano limiti analoghi. In un test indipendente del 2025, condotto su cinquanta prompt giuridici reali, entrambi i sistemi hanno generato almeno una citazione errata, irrilevante o superata nel 30–34% dei casi, nonostante l'integrazione con pipeline di recupero documentale su basi proprietarie (vedi Appendice 1.4, A2).[2]

In particolare, i sistemi hanno faticato con quesiti che richiedevano valutazioni di good law o il riconoscimento di overruling impliciti: inferenze che presuppongono competenza contestuale e qualificazione normativa.

Anche esami pratici non pubblicati confermano l'esigenza di intervento umano. In una serie di prompt sottoposti a GPT-4 e Claude 2, in modalità zero-shot e few-shot, si è reso necessario correggere manualmente 7 risposte su 10 per allineare riferimenti giurisprudenziali e logica argomentativa. È vero che la qualità migliorava quando il prompt era strutturato e corredato di riferimenti precisi — segnale che l'input incide profondamente sulla performance —, ma ciò non elimina la necessità di verifica ex post (vedi Appendice 1.4, A3).[3]

Rimane infine una criticità metodologica trasversale: le metriche usate per valutare l'output AI variano da test a test e mescolano indicatori eterogenei (esattezza testuale, pertinenza normativa, correttezza del ragionamento). Manca, ad oggi, uno standard condiviso per misurare la "precisione giuridica" intesa non come mera aderenza lessicale, ma come correttezza metodologica. L'assenza di criteri comparabili rende arduo confrontare i risultati tra modelli e complica la scelta informata da parte dei professionisti (vedi Appendice 1.4, A4).[4]

I dati indicano che nessun LLM attuale può essere impiegato come strumento di ricerca giuridica autonoma. Il valore aggiunto sta nell'assistere il professionista e accelerare fasi operative specifiche, ma solo all'interno di un workflow strutturato, con tracciabilità integrale delle fonti, regola "cita-o-taci" e controllo umano della qualificazione normativa.

---

1. *LEXAM — A Comprehensive Legal Reasoning Benchmark Derived from Real Law School Exams (preprint, 2024/2025).*
2. *Kostikova et al., A Data-Driven Survey of Evolving Research on Limitations of Large Language Models (2025)*

# Come funziona davvero un LLM: meccanismi, logiche e limiti cognitivi

Un modello linguistico di grandi dimensioni è, nella sua essenza, un predittore di parole: dato un contesto testuale, stima quale sequenza di simboli sia più probabile che segua. Questa architettura, basata su reti "a trasformatori" addestrate su grandi corpora, ottimizza una funzione statistica (minimizzare l'errore di previsione) e non una funzione veritativa (accertare ciò che è vero o giuridicamente valido). Da qui nasce l'ambiguità di fondo: una risposta plausibile non coincide necessariamente con una risposta corretta o valida dal punto di vista del diritto.[1]

Sul piano operativo, la produzione del testo è governata da decoding e campionamento (temperature, penalità di ripetizione, strategie come greedy, top-k, nucleus). Parametri più "creativi" aumentano varietà ed estensione argomentativa, ma anche il rischio di deviazioni non supportate; parametri più "conservativi" aumentano ripetizione e aderenza ai pattern più comuni, ma non garantiscono correttezza. In termini forensi, il modello massimizza la verosimiglianza linguistica, non la pertinenza normativa: per questo può citare correttamente una massima e al contempo applicarla in modo improprio, se il contesto d'uso non è qualificato.[2]

Negli ultimi anni sono stati introdotti strati di allineamento (ad es. instruction tuning e tecniche di preferenza umana) per rendere i modelli più utili e "docili" ai compiti dell'utente. L'effetto collaterale, ben documentato, è una tendenza alla conformità: il sistema privilegia risposte cooperative e sicure rispetto all'ammissione d'incertezza. Quando il compito è incorniciato con opzioni parziali o con presupposti impliciti, l'LLM asseconda il framing, riducendo l'"astensione virtuosa" (dire "non so" o "serve una fonte"). Questa dinamica spiega perché, senza un playbook preciso, si generino facilmente "certezze" apparenti.[3]

Il fenomeno comunemente chiamato “allucinazione” non è un incidente marginale, ma un esito atteso della combinazione fra addestramento su dati imperfetti, obiettivo di massima plausibilità e procedure di decoding. In assenza di un vincolo di verifica esterna (fonti, meta-controlli, calcolo), il modello colma i vuoti con enunciati coerenti con il contesto, anche quando mancano basi probatorie sufficienti. Le proposte recenti per mitigare il problema non agiscono “sulla natura del modello”, ma sull’ambiente di uso: restrizioni del compito, retrieval qualificato, meccanismi di astensione e audit dell’output.[4]

Le limitazioni strutturali sono particolarmente rilevanti per il diritto: (i) assenza di consapevolezza normativa (il modello non “sa” cosa sia una ratio decidendi, una norma abrogata, un overruling implicito; riconosce pattern su tracce testuali); (ii) fragilità temporale (l’addestramento non garantisce aggiornamento costante: la vigenza può mutare rispetto ai dati storici appresi); (iii) assenza di gerarchizzazione interna delle fonti (il modello non attribuisce “peso” alla fonte se non per come è rappresentata nel testo; l’ordinamento lo decide il giurista). Queste proprietà spiegano perché un’uscita ben formata possa essere epistemicamente debole.[5]

L’integrazione con retrieval (RAG) migliora la tracciabilità ma non sostituisce il giudizio. Due nodi restano decisivi: copertura (le fonti recuperate sono davvero le migliori disponibili per il quesito?) e qualificazione (quelle fonti, nel sistema delle fonti e della giurisprudenza, che peso hanno e a quale data?). Senza questi passaggi, si ottengono risposte con citazioni “vere” ma decontestualizzate, idonee a sostenere conclusioni fuorvianti. Per questo i protocolli evidence-first (cita-o-taci, source-of-record, controllo umano ex ante/ ex post) sono condizioni d’uso, non “optional” metodologici.[6]

Sul versante della valutazione, è utile distinguere tra: (a) correttezza fattuale/giuridica dell'enunciato; (b) grounding (tracciabilità alla fonte idonea); (c) pertinenza normativa (la fonte è davvero quella giusta nel tempo, nello spazio e nell'oggetto?); (d) qualità del ragionamento (scomposizione del caso, identificazione del decisum, confronto con tesi alternative). I test che misurano soprattutto la forma o la memoria non informano adeguatamente sulla affidabilità giuridica. La metrica deve riflettere il lavoro del giurista, non solo la scorrevolezza del testo.[7]

In un flusso professionale, l'LLM è uno strumento di calcolo linguistico al servizio di un processo governato dal giurista: va impiegato con compiti stretti, fonti tracciabili, standard di uscita e diritto di astensione esplicito. Quando il quesito esige qualificazione, interpretazione o bilanciamento di fonti, l'LLM assiste; non decide. Questo è il confine metodologico che consente di sfruttare la velocità del modello senza rinunciare alla responsabilità epistemica.

---

1. *L. Paseri – M. Durante, Examining Epistemological Challenges of Large Language Models in Law (2025).*
2. *Kostikova et al., A Data-Driven Survey of Evolving Research on Limitations of Large Language Models (2025)*
3. *Z. Ling et al., Instruction Boundary: Quantifying Biases in LLM Reasoning under Various Coverage (2025).*
4. *A. T. Kalai – O. Nachum – S. S. Vempala – E. Zhang, Why Language Models Hallucinate (preprint, 4 settembre 2025).*
5. *Paseri – Durante, op. cit.; si veda anche la riflessione sulla distinzione ratio/obiter e sul rischio di "pseudo-conoscenza" quando le strutture testuali sostituiscono le qualificazioni normative.*
6. *F. Surani – M. Dahl – V. Magesh, Legal RAG Hallucinations (2025).*
7. *Pipitone – Alami, LegalBench-RAG (2025): proposta di benchmark che separa la valutazione del recupero da quella della generazione in compiti legali; Kostikova et al., op. cit., per la necessità di metriche riproducibili su pertinenza normativa e qualità del ragionamento.*

# Finestre ampie ≠ comprensione: limiti del long-context

L'estensione della finestra di contesto non risolve, di per sé, i problemi di comprensione a lungo raggio. Valutazioni su compiti reali mostrano degradazioni marcate all'aumentare della lunghezza dell'input e una sensibilità alla posizione delle evidenze: gli stessi modelli che ricordano dettagli "in vetrina" faticano quando l'informazione rilevante è distribuita, gerarchizzata o richiede integrazione tra più parti del testo.[1] In altri termini, "più token" non equivale a "più ragionamento": l'effetto pratico è un decadimento progressivo della qualità quando la prova richiede di tenere insieme frammenti distanti e di disambiguare fra molte etichette plausibili.[2]

Questa dinamica ha implicazioni operative dirette. Primo, dare più testo introduce rumore e competizione attentiva tra segmenti; con semplici troncamenti (ad es. prime frasi) le metriche si appiattiscono dopo 7–8 frasi, segnalando un collo di bottiglia intrinseco nella gestione di sequenze lunghe.[3] Secondo, i guadagni più affidabili si ottengono fuori dal modello, con pipeline che pre-selezionano e comprimono il materiale: sifting mirato, riassunti estrattivi/diversi e ri-ordering delle porzioni salienti prima del prompting. Su compiti realistici, tali procedure aumentano l'accuratezza fino al 50% e riducono costi e latenza rispettivamente fino al 93% e 50%.[4]

In sintesi, la finestra lunga è un mezzo di trasporto del contesto, non una garanzia di ragionamento distribuito. Nel dominio giuridico — dove le evidenze sono gerarchiche (fonti), temporali (vigenza) e relazionali (rinvii, overruling) — l'uso professionale dell'LLM richiede un pre-processing metodico del corpus: segmentazione tematica, riduzioni controllate e indicizzazione che rendano lieve e saliente ciò che il modello deve davvero considerare.

Questo ponte tecnico anticipa la frattura epistemologica discussa nel §1.7 (plausibilità ≠ validità) e motiva le scelte di workflow evidence-first del §1.8.[24]

1. *T. Li et al., Long-context LLMs Struggle with Long In-context Learning, TMLR (03/2025):*
2. *P. Hosseini et al., Efficient Solutions for an Intriguing Failure of LLMs: Long Context Window Does Not Mean LLMs Can Analyze Long Sequences Flawlessly,*

# Epistemologia giuridica vs output LLM

Il diritto non è solo un insieme di testi, ma un sistema di conoscenza strutturato. L'epistemologia giuridica si fonda su processi di validazione che attribuiscono significato e valore a ogni fonte: distinguere tra norme primarie e secondarie, tra ratio decidendi e obiter dicta, tra orientamenti consolidati e minoritari. Questa pratica si regge su coerenza sistematica, verifica logica e interpretazione contestuale — requisiti che un LLM, per come funziona oggi, non può riprodurre autonomamente.[1]

Un modello linguistico predice parole, non fonda concetti. Genera testo plausibile a partire da regolarità statistiche, non da criteri normativi. Può apparire "competente" anche quando l'affermazione è falsa o priva di fondamento giuridico, perché la sua formulazione è formalmente coerente. Ne deriva un rischio specifico: la pseudo-conoscenza — contenuti che sembrano affidabili ma non soddisfano i requisiti minimi della validità giuridica.[2]

La pseudo-conoscenza non è un semplice errore di dettaglio: è una distorsione del processo conoscitivo. Nel diritto, dove la forza di un argomento dipende tanto dalla fonte quanto dalla forma, scambiare somiglianza linguistica per validità normativa indebolisce la decisione. Anche quando un output AI risulta utile, occorre chiedersi: secondo quale criterio è stato generato? Quale metodo ha guidato la selezione delle fonti? Che cosa è stato omesso lungo il percorso?

Questa frattura epistemologica non si colma con più prompt o più documenti nel retrieval. Il punto non è "dare più dati al modello", ma riconoscere che l'LLM opera con una logica diversa da quella del giurista: l'uno massimizza plausibilità, l'altro costruisce conoscenza tramite regole, interpretazioni e sistemi. Da qui l'esigenza di una regia umana forte, che impieghi l'AI come assistente e non come arbitro.[3]

Operativamente, ciò impone un cambio di paradigma: l'AI non sostituisce il ragionamento; supporta l'elaborazione delle ipotesi, l'esplorazione dei contro-argomenti e l'ampliamento dello spettro informativo. Il giurista resta il titolare del criterio di validità: decide quando un'informazione è utile, rilevante e corretta sul piano normativo. L'LLM è dunque uno strumento euristico, non un soggetto epistemico.

Solo chiarendo questa distinzione si evita la confusione tra linguaggio e diritto. Gli LLM possono generare contenuti giuridici, ma non conoscere il diritto; possono facilitare la ricerca, ma non garantire la correttezza normativa; possono accelerare la scrittura, ma non sostituire la verifica. In sintesi: possono molto, se il giurista conserva il controllo del metodo. Da qui la necessità di un workflow evidence-first: nella sezione successiva traduciamo questi principi in protocolli operativi.

---

1. *L. Paseri – M. Durante, Examining Epistemological Challenges of Large Language Models in Law (2025).*
2. *A. T. Kalai – O. Nachum – S. S. Vempala – E. Zhang, Why Language Models Hallucinate (preprint, 4 settembre 2025): spiega perché la massimizzazione della plausibilità linguistica, unita a dati imperfetti e strategie di decoding, generi enunciati coerenti ma non fondati. Si veda anche Paseri – Durante, op. cit., per l'applicazione di tale dinamica al dominio giuridico e per la distinzione tra "conoscenza" e "pseudo-conoscenza".*
3. *F. Surani – M. Dahl – V. Magesh, Legal RAG Hallucinations (2025); Z. Ling et al., Instruction Boundary: Quantifying Biases in LLM Reasoning under Various Coverage (2025).*

# Workflow evidence-first

L'integrazione degli LLM nella ricerca legale richiede un ripensamento radicale del workflow. Non è sufficiente usare modelli generativi come sostituti dei motori di ricerca o dei junior associate: ciò che cambia è il rapporto tra linguaggio e conoscenza. L'LLM non ragiona né verifica, ma predice. Di conseguenza, l'unico modo per mantenere controllo, tracciabilità e rigore è costruire un processo evidence-first: ogni informazione prodotta dall'AI deve avere un'àncora verificabile in una fonte giuridicamente valida. Questo approccio si traduce in una sequenza operativa articolata in nove fasi, ciascuna con un ruolo umano definito e un impiego dell'LLM disciplinato.

**1. Formulazione del quesito e pianificazione della ricerca**
Il professionista inquadra il problema giuridico, individua le questioni di diritto rilevanti (issue spotting) e pianifica le aree da esplorare. In questa fase, l'LLM può offrire spunti generali, aiutando a espandere il perimetro di ricerca con suggerimenti su argomenti correlati. Tuttavia, la definizione della traccia argomentativa è prerogativa esclusiva dell'avvocato, che interpreta il contesto fattuale e gli obiettivi pratici della ricerca (vedi Appendice 1.7, A1).

**2. Ricerca multicanale e raccolta massiva delle fonti**
La raccolta delle fonti deve avvenire su più canali: banche dati giuridiche, repertori ufficiali, fonti aperte. Parallelamente, l'LLM può fungere da motore semantico ausiliario: interrogato con prompt mirati, può segnalare sentenze simili, dottrina rilevante o riferimenti normativi. La condizione vincolante è che ogni riferimento sia accompagnato da una fonte primaria accessibile. In assenza di fonte verificabile, l'output va scartato. L'obiettivo in questa fase è l'overcollection intelligente: raccogliere anche più di quanto servirà, per poi filtrare a valle

**3. Esplorazione delle fonti con LLM**

Le fonti raccolte vanno esplorate singolarmente, non aggregate. Ogni documento viene interrogato con prompt vincolati, che limitano l'ambito interpretativo: giurisdizione, data, materia. L'LLM viene usato come lettore aumentato, non come sintetizzatore globale. Questo riduce il rischio di allucinazioni e inferenze spurie dovute a contesto misto o incoerente (vedi Appendice 1.7, A2).

**4. Riassunto e indicizzazione**

Durante l'esplorazione, ogni fonte viene riassunta con l'assistenza dell'AI. Il giurista arricchisce questi riassunti con metadati: tipo di fonte, stance, rilevanza, correlazioni. Si costruisce così una knowledge base modulare, interrogabile e tracciabile. L'LLM svolge il ruolo di annotatore aumentato. Questa fase ha anche un valore euristico: aiuta a "vedere" il materiale in modo strutturato, con cluster tematici e contrasti emergenti.

**5. Estrazione di contenuti strutturati**

L'LLM è istruito a estrarre informazioni neutre: norme applicate, fatti accertati, argomenti centrali, citazioni chiave. Qui non si chiede all'AI di commentare, ma di elencare. Il prompting è orientato alla precisione, non alla fluidità. L'obiettivo è ottenere un dataset di contenuti grezzi ma organizzati, pronti per la fase di ragionamento.

**6. Ragionamento giuridico umano**

Sulla base del materiale estratto, l'avvocato effettua una prima selezione critica. Valuta la solidità delle fonti, la coerenza dei principi, la pertinenza al caso. È in questa fase che avviene la trasformazione da "documenti" a "blocchi argomentativi validati". Ogni elemento viene pesato in funzione dell'obiettivo strategico: convincere, difendere, prevenire.

**7. Confronto argomentativo assistito**
Il giurista sottopone all'LLM la propria teoria del caso o ipotesi argomentativa. L'AI viene utilizzata per testarne la coerenza interna e simulare obiezioni esterne: "Quali contro-argomenti possono emergere?", "Quali principi giurisprudenziali potrebbero confutare questa tesi?". Questo uso dell'LLM valorizza la sua capacità combinatoria senza delegargli la decisione. Le contro-tesi proposte vengono vagliate e, se rilevanti, diventano spunti per rafforzare l'argomentazione.

**8. Architettura del documento e verifica logica**
Prima di scrivere, si costruisce una mappa logica del documento: ogni sezione è collegata a fonti, contro-fonti, norme e ha uno scopo preciso. L'LLM può aiutare a verificare che non ci siano salti logici, ripetizioni o squilibri. Il documento è pianificato come una struttura argomentativa trasparente, che riflette la catena "fonte → tesi → implicazione".

**9. Stesura finale e controllo**
La scrittura avviene per sezioni, con prompt contestualizzati. L'LLM può riformulare, semplificare, uniformare lo stile, ma non generare interamente le sezioni critiche. Ogni citazione è verificata. Ogni passaggio viene confrontato con l'evidenza. Il giurista è autore pieno, l'AI è uno strumento avanzato di editing e supporto stilistico.

Questa sequenza non è rigida ma adattabile. Ciò che conta è il principio guida: costruire intelligenza, non chiederla. L'LLM è un acceleratore linguistico, non un motore epistemico. Solo inserito in un protocollo evidence-first può produrre valore senza minare la qualità della conoscenza giuridica.

Il metodo qui descritto non è una formula definitiva, ma un canvas operativo replicabile, capace di adattarsi a contesti diversi mantenendo costante il presidio umano sulla validità.
Ma questo approccio funziona solo se i ruoli umani e le competenze sono ben delineati: vediamo come progettarli per fase.

# Ruoli umani e strumenti LLM, per fasi

L'introduzione degli LLM nella ricerca giuridica non implica una sostituzione del ruolo umano, bensì una sua riconfigurazione funzionale. All'interno di un workflow evidence-first, ogni fase del processo richiede un presidio umano differenziato, in relazione alla natura epistemica dell'attività e al grado di affidabilità degli strumenti AI impiegati. La chiave non risiede nella sostituzione, ma nella distribuzione ottimale delle funzioni (task allocation) in base alle competenze e ai rischi associati.

Nelle fasi iniziali di framing e pianificazione della ricerca, il controllo è interamente umano. Il professionista – tipicamente un partner o un associate esperto – formula il quesito giuridico, definisce il perimetro dell'indagine e stabilisce i criteri di rilevanza. In questa fase, l'LLM può essere impiegato unicamente in funzione esplorativa, per suggerire possibili aree correlate, ma non interviene nella strutturazione logica della ricerca.

Durante la raccolta e organizzazione delle fonti, si configura una collaborazione. Il giurista seleziona le fonti autorevoli mediante canali tradizionali, mentre l'LLM, sotto la supervisione di un knowledge manager o di un associate dedicato, supporta l'annotazione, il clustering semantico e l'indicizzazione dei materiali. La tracciabilità delle fonti resta condizione necessaria per l'utilizzabilità dell'output.

Nella fase di analisi e sintesi, l'LLM può essere impiegato come assistente cognitivo per la pre-elaborazione del materiale: sintesi di documenti, estrazione di principi, comparazione automatica tra sentenze. Tuttavia, la valutazione interpretativa – ovvero l'attribuzione di rilevanza e valore giuridico – resta una prerogativa insostituibile del giurista.

Infine, nella fase di validazione argomentativa e redazione, il ruolo dell'LLM si limita a funzioni ancillari: verifica della coerenza interna, simulazione di obiezioni, assistenza stilistica. La responsabilità epistemica – tanto nella costruzione del ragionamento quanto nella sua esposizione – è integralmente umana.

Ogni fase richiede strumenti differenti, calibrati sul tipo di attività: modelli generalisti per il brainstorming, LLM verticali con capacità di RAG per il supporto documentale, ambienti protetti per la redazione e l'auditabilità. In tutti i casi, è la regia umana a determinare la misura e la modalità dell'intervento AI.

# Conclusione – Asimmetria epistemica, regia umana

L'integrazione degli LLM nella ricerca giuridica impone una distinzione non eludibile: tra generazione linguistica e validazione normativa. I modelli producono testo plausibile, ma non fondato; simulano coerenza, ma non costruiscono conoscenza. In questo scenario, il rischio non è l'errore isolato, ma la formazione di pseudo-conoscenza: affermazioni formalmente corrette ma epistemicamente deboli.

La risposta non sta nel rifiuto dell'AI, ma in una sua integrazione governata. Ciò richiede tre presìdi essenziali: (i) tracciabilità sistematica delle fonti, con applicazione rigorosa della regola cite-or-silent; (ii) attribuzione esplicita dei ruoli decisionali, distinguendo le fasi dove l'LLM può assistere da quelle che richiedono giudizio umano; (iii) adozione di metriche condivise per valutare l'output, fondate su pertinenza normativa, coerenza logica e validità contestuale.

Per lo studio legale, l'AI rappresenta un potenziale acceleratore cognitivo, non un sostituto della competenza. Il vantaggio competitivo non nasce dalla velocità del prompting, ma dalla qualità del metodo. L'LLM può suggerire, riassumere, sollecitare contro-tesi. Ma la responsabilità epistemica – fondamento della professione – resta esclusivamente umana.

MARCO PAGANI
Ottobre 2025

Lidia

+39 010 8991141
info@lidiatech.ai
www.lidiatech.ai
/company/lidiatech