[ Anton Kraev Christensen ] / founder, advokat (L)

/ Legal AI note, april 2026

# Når noget af det *sværeste* er at få kunstig intelligens til at sige:

# ”Jeg ved det ikke.”

probably
**right.**
because ‘probably right’ isn’t enough.

# Når noget af det sværeste er at få kunstig intelligens til at sige: "Jeg ved det ikke".

[ Anton Kraev Christensen ]

## Det egentlige spørgsmål

Det spørgsmål, der oftest blev stillet om juridisk AI, lyder nogenlunde sådan:

*Kan systemet give rigtige juridiske svar?*

Det er et vigtigt spørgsmål. Men det er ikke det skarpeste.

## Tre separate problemer

Det mere præcise spørgsmål er, om systemet kan kende forskel på tre situationer:

- Ét problem er, om systemet kan *finde* materiale.
- Et andet er, om det finder materiale, der faktisk er *relevant*.
- Et tredje - og det vanskeligste - er, om det overhovedet har et stærkt nok grundlag til at *konkludere*.

Disse tre problemer blandes konstant i den offentlige samtale. At et system finder en tekst, er ikke det samme som, at teksten bærer svaret. Og at teksten bærer et udsagn, er ikke det samme som, at systemet bør konkludere i den konkrete sag.

I jura er det ikke nok, at svaret lyder rigtigt. Det skal hvile på de rigtige retskilder, i den rigtige retsorden, på det rigtige tidspunkt og på et faktum, der er oplyst godt nok til, at reglen kan anvendes forsvarligt.

Den sværeste opgave i juridisk AI er ikke bare at få systemer til at svare rigtigt oftere. Den sværeste opgave er at få systemer til at undlade at konkludere, når kilderne er for svage, når spørgsmålet bygger på en forkert forudsætning, eller når sagen er for dårligt oplyst.

## Hvorfor modellerne har svært ved at holde igen

Den første forklaring er teknisk, men ikke kompliceret. Store sprogmodeller er bygget til at fortsætte tekstproduktion.

OpenAI's analyse fra 2025 beskriver det meget direkte: standardtræning og standardmålinger belønner ofte gætteri mere end erkendt usikkerhed. Problemet er ikke, at modeller "vil lyve". Problemet er et systemisk træk i retning af at svare, også dér hvor et bedre kalibreret system burde holde igen.

Som OpenAI selv eksemplificerer i artikel *Derfor hallucinerer sprogmodeller, 5. september 2025*:

> *Lad os tage et […] eksempel og antage, at en sprogmodel bliver spurgt om nogens fødselsdag, men ikke kender svaret. Hvis den gætter "10. september", har den 1 chance ud af 365 for at gætte rigtigt. Hvis den siger "Det ved jeg ikke", får den med garanti 0 point. Med tusindvis af testspørgsmål ender gættemodellen med at se bedre ud på ranglisten end en påpasselig model, der indrømmer usikkerhed.*

Forskning peger på det samme fra en anden vinkel: hvis en model aldrig trænes til at formulere "det ved jeg ikke", forbliver den ude af stand til at gøre det, når den står over for noget ukendt. Tilbageholdenhed er ikke en standardfærdighed. Den skal læres eksplicit.

Hertil kommer et lag, der sjældent nævnes: produktdesign. Mange juridiske AI-produkter præsenterer typisk konklusioner selvsikkert, fordi det er det, brugerne belønner i øjeblikket. Det er ikke alene en teknisk begrænsning. Det er en kommerciel beslutning. Og det bør behandles som det - separat fra spørgsmålet om, hvad modellen teknisk er i stand til. Sikker tone forveksles let med sikker analyse, og det er ikke kun et træningsproblem. Det er et designvalg.

## Hovedtesen

Hovedtesen er derfor enkel: Det centrale problem i juridisk AI er ikke kun, at systemer undertiden hallucinerer og tager fejl. Det er, at de ofte konkluderer, før grundlaget metodisk kan bære det.

Det interessante er, at denne diagnose ikke kun kommer fra nyere AI-forskning. Læser man eksempelvis Mads Bryde Andersens *Ret og metode,* 2002 (det var lige den fremstilling af juridisk metode, som var lige ved hånden på vindueskammen), er det slående, hvor meget af problemstrukturen allerede er beskrevet i klassisk juridisk metode, blot på en lidt anden måde. Den nye forskning siger ikke det samme ord for ord. Men den måler mange af de steder, hvor juridisk metode længe har krævet tilbageholdenhed.

Artiklen her er derfor hverken en ren forskningsgennemgang eller en teknisk anvisning. Den er en argumentation for, hvad man som bruger eller udvikler af juridisk AI aktivt må forholde sig til, hvis man vil bruge teknologien juridisk forsvarligt.

De udenlandske benchmark- og evalueringsstudier bruges her ikke som én-til-én-modeller for dansk retskildelære. De bruges, fordi de isolerer fejltyper, som også er metodisk relevante i dansk ret: forældede kilder, utilstrækkeligt faktum, vildledende kildebrug og overkonklusion.

## 1. At finde materiale er ikke det samme som at kunne svare

Det første niveau er *fremsøgning*.

I AI-litteraturen kaldes det ofte *retrieval* – altså den del af systemet, der skal hente relevante dokumenter frem, før modellen forsøger at skrive et svar. Her er der sket meget.

Mange nyere juridiske værktøjer er bedre end tidlige generelle chatmodeller til at hente bestemmelser, domme og andet materiale frem. Men det betyder ikke i sig selv, at de har forstået det juridiske spørgsmål.

Et af de bedste steder at se det er *A Reasoning-Focused Legal Retrieval Benchmark* af Lucia Zheng, Neel Guha, Javokhir Arifov, Sarah Zhang, Michal Skreta, Christopher D. Manning, Peter Henderson og Daniel E. Ho, publiceret ved *4th ACM Symposium on Computer Science and Law (CS&Law '25)* i 2025.

Forfatterne angriber en svaghed i mange ældre retrieval-tests: de blev ofte lavet på spørgsmål, hvor spørgsmålet og den relevante retskilde delte mange af de samme ord. Det gør opgaven lettere, men også mindre juridisk realistisk.

Zheng m.fl. viser, at mange juridiske researchopgaver ser modsat ud. Juristen får et faktum, udleder derfra det egentlige juridiske spørgsmål og leder så efter de kilder, der belyser netop dét. Relevante retskilder og spørgsmål behøver ikke dele mange fælles formuleringer. Ofte gør de ikke.

For at vise det har forfatterne bygget to tests, *Bar Exam QA* og *Housing Statute QA*, med omkring 10.000 eksempler, hvor relevansen er vurderet af jurister, og hvor den relevante passage ofte har lille sproglig lighed med spørgsmålet. De viser også, at klassiske søgemetoder som BM25 - et etableret system, der ranker dokumenter efter ordoverlap med spørgsmålet - kæmper netop dér, mens lovinspireret omformulering af spørgsmålet kan hjælpe.

Det er derfor vigtigt at holde fast i den første skelnen: et system, der finder materiale, har ikke dermed vist, at det kan besvare et juridisk spørgsmål. Det har kun vist, at det kan finde tekst.

For udvikleren af juridisk AI er pointen lige så klar: at forbedre søgefunktionen løser ikke i sig selv det juridiske problem.

## 2. Relevant materiale er heller ikke nok

Det andet niveau er vanskeligere. Her er spørgsmålet ikke bare, om systemet har fundet noget, men om det har fundet noget, der faktisk er relevant. Det lyder som en lille forskydning. Det er det ikke.

Her er *Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools* af Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D. Manning og Daniel E. Ho særlig vigtig. Undersøgelsen blev først lagt ud som arXiv-preprint i 2024 og senere publiceret i *Journal of Empirical Legal Studies* i 2025.

Det er den første præregistrerede, systematiske evaluering af de førende kommercielle AI-baserede juridiske researchværktøjer. Og den siger noget, der bør have standset en del markedsføring: søgebaserede arbejdsgange reducerer hallucinationer i forhold til almindelige chatbots, men de eliminerer dem ikke. For de lukkede værktøjer, der blev undersøgt, fandt forfatterne hallucinationsrater på omtrent 17 til 33 procent – altså i størrelsesordenen hver sjette til hver tredje forespørgsel.

Det mest interessante ved undersøgelsen er dog ikke kun tallene. Det er forfatternes begrebsmæssige oprydning:

- En respons kan være *incorrect* – ganske enkelt forkert.
- Men den kan også være *misgrounded*: den citerer en virkelig kilde, men bruger den juridisk forkert, for eksempel fordi kilden kommer fra en forkert retsorden eller ikke faktisk bærer den konklusion, modellen lægger i den. Det er ofte den farligste fejl.

En opdigtet dom kan undertiden opdages. En ægte dom, brugt på et forkert retligt spørgsmål, er langt mere forræderisk i det juridiske arbejde (og langt sværere at fange, når man læser sent om aftenen).

Derfor er forskellen mellem *same-topic* og *same-issue* så vigtig. Et system kan finde noget om det rigtige emne uden at have fundet noget om det rigtige juridiske spørgsmål.

For både bruger og udvikler af juridisk AI betyder det, at en synlig kildeliste ikke i sig selv er et kvalitetsstempel. Spørgsmålet er, om retskilden bærer netop den konklusion, svaret faktisk drager.

## 3. Støtte i en tekst er ikke det samme som et bærende retligt grundlag

På ét område er forskningen faktisk kommet længere, end mange tror. Det gælder kontrollen af, om en konkret påstand faktisk kan genfindes i en konkret tekst / retskilde. Det er et vigtigt fremskridt, fordi meget dårlig AI ikke lyder som ren hallucination. Den lyder som noget, der næsten passer med retskilden.

Et godt eksempel er *ContractNLI: A Dataset for Document-level Natural Language Inference for Contracts* af Yuta Koreeda og Christopher Manning, publiceret i *Findings of the Association for Computational Linguistics: EMNLP 2021*.

*Natural language inference*, ofte forkortet NLI, er opgaven at afgøre, om en tekst understøtter, modsiger eller er tavs om en påstand.

Koreeda og Manning tager den idé ind i kontraktverdenen: systemet får en kontrakt og en hypotese og skal afgøre, om hypotesen er understøttet, modsagt eller slet ikke nævnt – og samtidig pege på de passager, der bærer vurderingen. De offentliggør 607 kontrakter, som er gennemlæst og markeret af jurister, og viser, hvor svært problemet er: kontrakter er lange, undtagelser betyder meget, og den relevante passage kan ligge langt væk fra den sætning, der skal vurderes.

En bredere version af samme idé findes i *Enabling Large Language Models to Generate Text with Citations* af Tianyu Gao, Howard Yen, Jiatong Yu og Danqi Chen, publiceret ved *EMNLP 2023*.

Forfatterne introducerer testen ALCE – *Automatic LLMs' Citation Evaluation* – og bruger blandt andet to mål, som er simple nok til at være nyttige:

- *citation recall*, altså om de citerede passager faktisk understøtter de påstande, svaret fremsætter, og
- *citation precision*, altså om de citerede passager overhovedet er relevante.

Deres resultater er tydelige: selv stærke systemer har betydeligt forbedringsrum. Et af de sværeste datasæt i testen er ELI5, en samling af lægmandsspørgsmål, hvor svaret skal kunne forstås uden forhåndsviden. Her mangler selv de bedste modeller fuld kildeunderstøttelse af deres svar halvdelen af tiden.

Det er præcis den svaghed, *CiteEval: Principle-Driven Citation Evaluation for Source Attribution* af Yumo Xu, Peng Qi, Jifan Chen, Kunlun Liu, Rujun Han, Lan Liu, Bonan Min, Vittorio Castelli, Arshit Gupta og Zhiguo Wang, publiceret ved *ACL 2025*, går direkte efter.

De kritiserer den vane, der længe har præget forskningsfeltet, nemlig at reducere kildekvalitet til et simpelt spørgsmål om støtte eller ikke-støtte. Deres pointe er, at det er for simpelt. En kildehenvisning kan være misvisende, for svag eller overflødig.

Det er metodisk mere interessant for jurister, fordi det ligger tættere på den måde, en jurist faktisk læser på: ikke bare "*kan jeg finde noget støtte?*", men "*var dette den rigtige støtte at bruge?*".

Hovedpointen er den samme, uanset hvilken vinkel man læser forskningen fra: *at retskilden siger det, modellen påstår, er ikke det samme som, at kilden kan bære svaret*. Det er den sætning, der løber igennem resten af artiklen.

## 4. Fra tekst til præmis: hvorfor Mads Bryde Andersens 'Ret og metode' stadig er moderne

Det er her, Mads Bryde Andersens *Ret og metode* igen bliver interessant. I *Ret og metode*, 1. udgave, 2002, definerer han på s. 132 en retskilde som:

> *den præmis i en konklusion om gældende ret, som foreligger i autoritativ form, enten i form af et dokument udstedt af en kompetent instans eller i form af en udtalelse om, hvad der er praksis inden for et område, afgivet af en institution med en særlig kompetence inden for området.*

Det er en kort formulering, men den flytter hele diskussionen i juridisk AI et vigtigt sted hen.

Den definition er ikke interessant, fordi den er elegant. Den er interessant, fordi den tvinger én til at skifte niveau. Spørgsmålet er ikke længere bare: *Har jeg en tekst fra retskilden?* Eller: *Handler teksten om det samme emne?* Spørgsmålet bliver: *Har jeg en præmis, som faktisk kan bære en konklusion om gældende ret?*

Det er også derfor, noget af den nye AI-forskning rammer så tæt på klassisk juridisk metode. Ikke fordi den gentager den. Men fordi den teknisk forsøger at måle det, juridisk metode længe har insisteret på: at der er forskel på retskilde, relevant retskilde og en bærende *præmis*.

Når man læser Mads Bryde Andersen i lyset af ContractNLI, ALCE og CiteEval, bliver det tydeligt, hvorfor retskildehenvisninger i sig selv ikke løser det juridiske problem. De løser et vigtigt delproblem: de gør det muligt at kontrollere, hvor systemet kiggede. Men de afgør ikke, om retskilden har den rigtige vægt, den rigtige placering i retssystemet, den rigtige tidslige gyldighed eller den nødvendige relation til det konkrete faktum.

Fremtiden for juridisk AI ligger derfor ikke i mere velformulerede svar. Den ligger i systemer, der i højere grad kan opføre sig, som juridisk metode længe har krævet af juristen: retskilde først, præmis derefter, konklusion sidst.

Det er lige så meget et krav til dem, der udvikler juridisk AI, som til dem, der bruger den.

## 5. Retskilden skal ikke bare findes – den skal kunne bære svaret

Mads Bryde Andersen skelner på s. 135 i *Ret og metode* mellem retskilders *autoritative* og *retssystematiske* værdi. Den autoritative værdi handler om, hvilken kompetence der står bag kilden. Den retssystematiske værdi handler om, hvordan kilden står i forhold til det øvrige retssystem:

> *Retskildens **autoritative værdi** måles på den kompetencenorm, der giver hjemmel for retskilden: Hvor højt rangeret er den, og hvilke frihedsgrader (dvs. selvstændig magt) overlader den til den myndighed, der har frembragt den? En afgørelse, der strider mod den kompetencenorm, der har ligget til grund fra den besluttende magt, har ingen stor retskildeværdi. Dette spørgsmål, der er et direkte udslag af retskildehierarkiet, skal jeg straks vende tilbage til.*
>
> *Den **retssystematiske værdi** måles ud fra den sammenhæng (harmoni), der består mellem den foreliggende retskilde og andre retskilder. En enkelt afgørelse, der bryder med en strøm af afgørelser, der går i modsat retning, har sjældent stor retskildeværdi. Denne problemstilling har jeg – som nævnt indledningsvis – valgt at holde ude fra retskildelæren.*

Det er et bedre sprog end bare "hierarki", fordi det forklarer to forskellige måder, en retskilde kan være svag på. En retskilde kan være svag, fordi den kommer fra et sted med begrænset kompetence. Men den kan også være svag, fordi den står alene mod stærkere eller mere samstemmende retskilder.

Det er netop derfor, Validate Your Authority: Benchmarking LLMs on Multi-Label Precedent Treatment Classification af M. Mikail Demir og M. Abdullah Canbaz, publiceret i Proceedings of the Natural Legal Language Processing Workshop 2025, er mere interessant, end titlen måske først antyder. Artiklen handler om et common law-delproblem: hvordan senere domme behandler tidligere domme. . Det er altså ikke en én-til-én-model for dansk retskildelære. Men metodisk er pointen bredere.

Forfatterne opstiller opgaven som et *multi-label*-klassifikationsproblem. En afgørelse kan være *overruled*, *criticized*, *questioned*, *distinguished*, *not followed* eller neutralt citeret. Og de indfører derfor et mål, *Average Severity Error*, fordi ikke alle fejl er lige alvorlige. Det er langt værre at forveksle en *neutral citation* med en undermineret retskilde end at forveksle to beslægtede negative behandlinger.

Common law-terminologien kan ikke oversættes én til én til dansk ret. Men princippet kan. Også uden for common law kan en retskilde være relevant og stadig ikke kunne bære svaret, fordi den er senere afgrænset, praktisk tilsidesat eller simpelthen vejer for lidt.

For udvikleren af juridisk AI rejser det et meget konkret spørgsmål: hvordan mærker systemet de retskilder, det bruger, så brugeren kan se ikke bare hvor svaret kommer fra, men også hvad kilden er værd?

## 6. Jurisdiktion: et svar kan være rigtigt og stadig gælde et andet sted

I dansk sammenhæng kan 'jurisdiktion' lyde som et fjernt problem. Det er det ikke.

De store modeller og de mest synlige forskningsbidrag er i høj grad engelsksprogede og ofte præget af amerikansk ret. Det betyder ikke, at modellerne kun "tænker amerikansk". Men det betyder, at de mønstre, eksempler og retskildetyper, som dominerer både træning og evaluering, i høj grad stammer fra andre retssystemer end det danske.

Her er *Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models* af Matthew Dahl, Varun Magesh, Mirac Suzgun og Daniel E. Ho, publiceret i *Journal of Legal Analysis* i 2024, stadig et af de bedste fremstillinger at læse.

Forfatterne tester modeller på en stor samling amerikanske føderale sager og finder, at fejl ikke fordeler sig tilfældigt. Hallucinationsraterne varierer systematisk med domstolsniveau, jurisdiktion, sagsalder og hvor kendt sagen er. De dokumenterer også det, de kalder *contrafactual bias*: modeller accepterer ofte brugerens forkerte juridiske forudsætninger uden at korrigere dem.

Det er en meget vigtig pointe for jurister, fordi den ligger tæt på praktisk rådgivning. Hvis klienten eller brugeren stiller et spørgsmål med en indbygget forkert forudsætning, er den rigtige reaktion ikke at bygge videre på forudsætningen. Den rigtige reaktion er at standse og sige, at spørgsmålet måske hviler på en fejl.

I dansk sammenhæng gør det problemet større, ikke mindre. Hvis et system er trænet og målt i miljøer, hvor amerikanske retskilder og common law-begreber fylder meget, er der en reel risiko for, at det lyder sikkert også dér, hvor det burde have været tilbageholdende i mødet med dansk eller europæisk ret.

Pointen er ikke, at amerikansk forskning kan fortælle os, hvad dansk ret er. Pointen er, at den tydeliggør, hvor let modeller kan lyde juridisk overbevisende på tværs af retsordener.

Juridiske svar er stedbundne. Et system, der ikke er skarpt på, hvilken retsorden og hvilken type retskilder spørgsmålet angår, kan nemt give et svar, som virker rigtigt, men gælder et andet sted.

Det er en af de ting, udviklere af juridisk AI i dansk sammenhæng særligt må tænke igennem fra bunden.

## 7. Tid: en kilde kan være rigtig og stadig være forældet

Juridiske forhold er ikke bare lokalt forankrede. De er også tidsbundne. En dom kan være relevant og alligevel ikke kunne bære svaret, hvis den er blevet overhalet, afgrænset eller i praksis sat ud af kraft. En lovtekst kan være korrekt citeret og stadig være forkert for sagen, hvis spørgsmålet gælder en version, der var i kraft på tidspunktet for sagens faktum, og som senere er ændret.

Det er også derfor, arbejdet om senere behandling af præjudikater er mere end et common law-speciale. *Validate Your Authority* er metodisk interessant, fordi det viser, at en kildes betydning kan ændre sig, når senere afgørelser begrænser, kritiserer eller fraviger den. En retskilde kan derfor være relevant og alligevel ikke kunne bære svaret i sin oprindelige form.

Det samme ligger i nyere søgeforskning. Zheng m.fl. bemærker i *A Reasoning-Focused Legal Retrieval Benchmark*, at loven hele tiden ændrer sig, mens den viden, der ligger gemt i modelparametrene, er statisk. Derfor er det ikke nok, at systemet "ved noget om jura". Det skal kunne se, om det, det ved, stadig gælder.

For brugeren er pointen enkel: et system bør ikke bedømmes positivt alene, fordi det viser retskilder. Retskilderne skal gælde her og nu. For udvikleren er pointen lige så konkret: systemet skal helst kunne vise, hvis retskilderne ikke længere gør det.

## 8. Faktum: reglen kan være kendt, selv om sagen er for dårligt oplyst

Det mest undervurderede problem er stadig sagens faktum.

Et system kan godt kende reglen. Det kan endda citere den korrekt. Alligevel kan det mangle de oplysninger, som gør en sikker anvendelse mulig.

Her er *Interpretation of Natural Language Rules in Conversational Machine Reading* af Marzieh Saeidi, Max Bartolo, Patrick Lewis, Sameer Singh, Tim Rocktäschel, Mike Sheldon, Guillaume Bouchard og Sebastian Riedel, publiceret ved *EMNLP 2018*, stadig usædvanligt god at tænke med. Opgaven i det såkaldte ShARC-datasæt er ikke bare at læse en regeltekst og svare.

Systemet får en regeltekst og et spørgsmål og skal først afgøre, om spørgsmålet overhovedet kan besvares på det foreliggende faktum. Hvis ikke, er den rigtige næste handling at stille det relevante opklarende spørgsmål.

Datasættet rummer 37.000 eksempler, og forfatterne understreger selv, at mange spørgsmål i praksis er underspecificerede – at der simpelthen mangler oplysninger, før de kan besvares.

Det ligner almindeligt juridisk arbejde mere, end mange nyere tests gør. Hvis spørgsmålet er, om en bortvisning holder, om en klausul kan håndhæves, eller om en frist er overskredet, kan reglen være kendt.

Men hvis et afgørende faktum mangler, er den rigtige reaktion ikke at citere reglen og fortsætte. Den rigtige reaktion er at sige, hvad der mangler, eller at stille et opklarende spørgsmål.

Det er også her, at Mads Bryde Andersens *Ret og metode* igen bliver uventet moderne. På s. 173-174 peger han på en asymmetri, som enhver praktiker genkender, når den først er sagt højt: retsfølgen er ofte mere præcist beskrevet end retsfaktum (hvilket er forklaringen på, at så mange svar hos advokaten starter med 'det afhænger af' og 'det beror på'):

> *Det forhold, at en retsregel med logisk nødvendighed må være i besiddelse af såvel et retsfaktum som en retsfølge, samt det forhold, at den altid må konfronteres med et faktum i form af den hændelse, der giver anledning til retsanvendelsen, har på det indirekte plan stor betydning for, hvordan en retsregel "taler til" retsanvenderen.*
>
> *I almindelighed vil retsfølgen være mere præcist beskrevet end retsfaktum. Forklaringen herpå er, at retsfølgen udtrykker selve den magt, regelgiveren har valgt at lægge i hænderne på retsanvenderen. En sådan magt vil man normalt kun ønske at give videre i afgrænsede former, og derfor vil magten typisk være beskrevet i forholdsvis præcise begreber, f.eks. udtrykt i en strafferamme, en aftaleretlig beføjelse eller en særlig type forbud eller påbud. Dertil kommer, at det ofte er enklere at beskrive en retsfølge (som netop kan udtrykkes i den slags kategorier) end det er at være præcis med hensyn til, under hvilke betingelser retsfølgen kan bringes i anvendelse (retsfaktum). I den konkrete retsanvendelse skal retsfaktum således konfronteres med sagens konkrete omstændigheder, og retsanvenderen vil derfor i alle tilfælde have det i sin magt at foretage en konkret subsumptionsvurdering.*

Det betyder, at den virkelige vanskelighed tit ikke ligger i at gengive, hvad der sker, hvis betingelserne er opfyldt. Den ligger i at afgøre, om de konkrete omstændigheder faktisk opfylder dem.

Det er også derfor, juridisk AI så ofte virker bedre, end den egentlig er. Den er ofte god til at gengive retsfølger. Den er langt mere sårbar, når den skal håndtere åbne, sammensatte eller tyndt oplyste retsfaktiske betingelser.

Et system, der er bygget til at give selvsikre svar på oplyste sager, kan fejle stille på de sager, der ikke er oplyst nok.

## 9. Undtagelser: dårlig juridisk AI ligner ofte ikke vild hallucination

Dårlig juridisk AI ligner ofte ikke vild hallucination. Den ligner halvt rigtig jura.

Det er noget af det vigtigste at være opmærksom på. Mange fejl ser ikke absurde ud. De ser næsten rigtige ud. Systemet gengiver hovedreglen loyalt, men overser undtagelsen, betingelsen, afgrænsningen eller procesforudsætningen.

Det er præcis derfor, ContractNLI og CiteEval er så nyttige. Koreeda og Manning viser i *ContractNLI*, hvor meget undtagelser og formuleringer betyder for korrekt retskildeforståelse. Xu m.fl. viser i *CiteEval*, at retskildekvalitet ikke bare handler om *støtte* eller *ikke-støtte*, men også om delvis støtte, svagere støtte og vildledende støtte. Det er i praksis den samme fejlstruktur, set med to forskellige metodiske værktøjer.

Mads Bryde Andersens 'Ret og metode' siger noget meget lignende. På s. 185 i *Ret og metode,* skriver han, at der næsten ikke findes regler, der ikke gennembrydes af en eller anden form for undtagelse:

> *Der findes ikke mange regler, der ikke gennembrydes af en eller anden form for undtagelse. Foruden de udtrykkelige undtagelser, der kan knytte sig til bestemmelsens hovedregler, findes der en række principper og grundsætninger, der ikke er forbundet med den pågældende hovedregel, men som ikke desto mindre kan fravige den – alt efter sagens nærmere omstændigheder. Eksempler herpå er de grundsætninger om passivitet eller retsmisbrug, som domstolene undertiden indfører på alment grundlag. Ligeledes har man på alment grundlag (og altså uden for straffelovens område) opstillet principper om sanktionsbortfald ved nødret og nødværge, der efter den almindelige opfattelse også gælder uden for denne lovs område.*
> *Sprogligt kan det være vanskeligt at udtrykke en regel på en måde, så den klart afgrænser det område, hvorpå man ønsker den anvendt, og et retsområde, som den ikke skal gælde for. […]*

Det er et enkelt udsagn, men det er svært at tænke sig en bedre advarsel til enhver, der vil bruge AI i jura.

For brugeren er pointen meget enkel. En regel kan være gengivet korrekt og stadig ikke bære svaret. Nogle gange fordi undtagelsen mangler. Nogle gange fordi en betingelse overses. Nogle gange fordi systemet har fundet en retskilde, der peger i den rigtige retning, men ikke hele vejen.

For udvikleren rejser det spørgsmålet: hvordan får man et system til at gå ud over hovedreglen og systematisk lede efter undtagelserne?

## 10. Brugerens juridiske kompetence ændrer risikoen

Det er heller ikke ligegyldigt, hvem der læser svaret. En erfaren jurist har en faglig buffer. Vedkommende kan ofte se, at en kilde er for svag, at en forudsætning er forkert, eller at et vigtigt faktum mangler.

En almindelig borger eller yngre advokatfuldmægtig uden juridisk baggrund har sjældent samme buffer.

*Hallucination-Free?* af Magesh m.fl. placerer netop et ansvar hos de ansvarlige jurister for at føre tilsyn med og kontrollere AI-svar.

Det er fornuftigt. Men det betyder også, at meget af forskningen i praksis antager en professionel bruger som tavst sikkerhedsnet. Det sikkerhedsnet findes ikke nødvendigvis hos den almindelige bruger.

Det er en væsentlig forskel, som man må være bevidst om, når man designer juridiske AI-værktøjer.

Det samme system, med det samme svar, har ikke den samme risikoprofil, når modtageren ikke selv kan se, at svaret ikke rækker.

## 11. Produktdesign og evalueringsdesign er en del af problemet

Det er fristende at beskrive hallucinationer som et rent modelproblem. Men det er kun halvdelen af historien. Den anden halvdel handler om, hvordan modeller trænes, måles og belønnes.

Det er her, *Why Language Models Hallucinate* af Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala og Edwin Zhang, udgivet som OpenAI-forskningspapir og arXiv-preprint i september 2025, virkelig fortjener at blive læst grundigt. Det er ikke juridisk. Men det rammer et problem, juridisk AI arver næsten ubrudt fra det generelle LLM-felt.

Forfatternes hovedpointe er, at hallucinationer ikke bare overlever på grund af modelsvaghed. De overlever også, fordi standardtræning og især standardevaluering belønner gætteri frem for erkendt usikkerhed.

Deres argument er enkelt. Hvis et system bliver målt under 0-1-logik – ét point for rigtigt, nul for blankt eller "jeg ved det ikke" – bliver det rationelt at gætte, når man er usikker (en erfaring enhver, der har været til skriftlig eksamen, vil kunne nikke genkendende til). Og Kalai m.fl. understreger, at problemet fortsætter i systemer med søgning, retrieval og reasoning.

Så længe de dominerende tests og scoreboards stadig i praksis prioriterer rå træfsikkerhed, vil modeller blive presset i retning af svar frem for tilbageholdenhed.

Det er en særlig vigtig pointe for juridisk AI:

- *Retrieval* ophæver ikke incitamentsstrukturen.
- *Reasoning* ophæver den heller ikke.

Hvis modellen stadig bliver belønnet for at sige noget frem for at standse, vil den ofte fortsætte også dér, hvor grundlaget ikke rækker.

Det er også derfor, *R-Tuning: Instructing Large Language Models to Say 'I Don't Know'* af Hanning Zhang, Shizhe Diao, Yong Lin, Yi R. Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji og Tong Zhang, publiceret ved *NAACL 2024*, er et nyttigt supplement.

Forfatterne starter fra en enkel observation: hvis man kun træner modeller på korrekte svar, lærer man dem samtidig, at de altid skal svare.

Deres forslag er at træne modellen til at afstå fra at give et svar som en selvstændig færdighed. De kalder det *Refusal-Aware Instruction Tuning* og viser, at evnen til at afstå kan generalisere som en meta-færdighed og endda forbedre kalibreringen.

For jurister er læringen klar. Godt produktdesign kan dæmpe risikoen. Men produktdesign er ikke nok, hvis det underliggende evalueringsdesign stadig lærer modellen, at manglende svar er nederlag.

Det er et problem, udviklere af juridisk AI ikke kan løse alene ved at vælge en bedre model. De er nødt til at vælge, hvordan deres eget system måles og belønnes.

## 12. God usikkerhed er ikke tavshed

Det er vigtigt at afvise en falsk modsætning. Alternativet til en skråsikker konklusion er ikke nødvendigvis manglende svar.

I praktisk juridisk arbejde ser god usikkerhed ofte sådan ud:

> *Den mest nærliggende læsning er X, men det afhænger afgørende af Y; hvis Z viser sig, ændrer vurderingen sig; på det foreliggende grundlag kan man ikke sige det mere sikkert.*

Det er også derfor, det er vigtigt at skelne mellem *kalibrering* og *afståelse (fra et svar)*.

*Kalibrering* betyder, at et systems angivne sikkerhed nogenlunde passer med dets faktiske træfsikkerhed over tid. Hvis et system ofte siger 60 procent sikkert, og omkring 60 procent af de svar faktisk holder, er systemet statistisk kalibreret. *Afståelse* er noget andet. Det er ikke et tal på sikkerhedsskalaen. Det er den metodiske beslutning om ikke at konkludere.

Kalai m.fl. gør netop denne forskel tydelig i *Why Language Models Hallucinate*: at være kalibreret kræver mindre af systemet end at være præcis, men et lavt sikkerhedssignal løser ikke i sig selv problemet med overkonklusion. *R-Tuning* af Zhang m.fl. peger i samme retning: afståelse skal læres som en særskilt kapacitet.

Og *ShARC* af Saeidi m.fl. viser den tredje vej. Den rigtige næste handling er ikke altid et svar. Nogle gange er det et opklarende spørgsmål.

Det er tæt på klassisk god jura. Usikkerhed skal ikke håndteres med passivitet. Det er disciplin. Og det er en disciplin, som et juridisk AI-system ikke får af sig selv – den skal bygges ind.

## 13. Den rigtige sammenligning er ikke den perfekte jurist

Det rigtige sammenligningsgrundlag er ikke den ideelle jurist med ubegrænset tid, fuldt kildemateriale og perfekt overblik (som forfatteren bag artiklen er). I praksis sammenlignes juridisk AI ofte med travle advokater, pressede sagsbehandlere, yngre jurister eller borgere, der prøver at finde rundt uden professionel hjælp.

Det betyder ikke, at kravene til AI skal sænkes. Men det betyder, at diskussionen bliver skæv, hvis AI måles mod et idealmenneske, mens de faktiske menneskelige alternativer slipper for samme standard.

Magesh m.fl. skriver i *Hallucination-Free?*, at juridiske AI-værktøjer allerede nu kan have værdi som første skridt i juridisk research, hvis de ikke bruges som den endelige analyse og konklusion. Det er en realistisk og sund rollefordeling: hjælp først, konklusion senere.

## Konklusion

Det rigtige spørgsmål om juridisk AI er ikke, om modellerne kan skrive overbevisende. Det kan de.

Spørgsmålet er, om de kan lære at standse, når grundlaget ikke rækker. Det gælder, når retskilden har for svag vægt, når svaret gælder et andet sted eller en anden tid, når brugerens forudsætning er forkert, og når sagen er for dårligt oplyst til sikker subsumption.

Ny forskning i juridisk AI gør problemet langt skarpere, end debatten ofte lader forstå.

*Retrieval*-forskningen viser, at ægte juridisk fremsøgning ofte kræver mere end sproglig lighed.

*Reliability*-studierne viser, at retrieval-baserede værktøjer stadig hallucinerer, og at den farligste fejl ofte er misgrounding.

*Citation*- og *NLI*-litteraturen viser, at støtten i en tekst kan måles langt bedre end før, men også at en kildehenvisning ikke er det samme som et bærende grundlag.

Arbejdet om jurisdiktion, underspecificerede spørgsmål og afståelse (fra et svar) viser på hver sin måde det samme.

Konklusionen er enkel. Et juridisk AI-system bliver ikke bedre af altid at have et svar.

Det bliver bedre, når det kan kende forskel på, hvornår det skal finde, hvornår det skal forklare, hvornår det skal spørge, og hvornår det skal lade være med at konkludere.

For brugeren er det et krav om læseretning: spørg først, hvad systemet faktisk har grundlag for, før du spørger, hvad det siger.

For udvikleren er det et krav om design: byg systemer, hvor det er muligt at afstå, og hvor det er målbart, om kilderne bærer svaret.

I jura er det ikke en svaghed. Det er en forudsætning.

## Kilder:

Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala og Edwin Zhang, *Why Language Models Hallucinate*, OpenAI research paper / arXiv 2509.04664, september 2025.

Lucia Zheng, Neel Guha, Javokhir Arifov, Sarah Zhang, Michal Skreta, Christopher D. Manning, Peter Henderson og Daniel E. Ho, *A Reasoning-Focused Legal Retrieval Benchmark*, Proceedings of 4th ACM Symposium on Computer Science and Law (CS&Law), 2025.

Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D. Manning og Daniel E. Ho, *Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools*, først som arXiv-preprint 2024, senere publiceret i Journal of Empirical Legal Studies, 2025.

Yuta Koreeda og Christopher Manning, *ContractNLI: A Dataset for Document-level Natural Language Inference for Contracts*, Findings of the Association for Computational Linguistics: EMNLP 2021.

Tianyu Gao, Howard Yen, Jiatong Yu og Danqi Chen, *Enabling Large Language Models to Generate Text with Citations*, Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023).

Yumo Xu, Peng Qi, Jifan Chen, Kunlun Liu, Rujun Han, Lan Liu, Bonan Min, Vittorio Castelli, Arshit Gupta og Zhiguo Wang, *CiteEval: Principle-Driven Citation Evaluation for Source Attribution*, Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL 2025).

Mads Bryde Andersen, *Ret og metode*, 1. udg., Gjellerup/G.E.C. Gads Forlag, 2002.

M. Mikail Demir og M. Abdullah Canbaz, *Validate Your Authority: Benchmarking LLMs on Multi-Label Precedent Treatment Classification*, Proceedings of the Natural Legal Language Processing Workshop 2025.

Matthew Dahl, Varun Magesh, Mirac Suzgun og Daniel E. Ho, *Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models*, Journal of Legal Analysis, 2024.

Marzieh Saeidi, Max Bartolo, Patrick Lewis, Sameer Singh, Tim Rocktäschel, Mike Sheldon, Guillaume Bouchard og Sebastian Riedel, *Interpretation of Natural Language Rules in Conversational Machine Reading*, Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018).

Hanning Zhang, Shizhe Diao, Yong Lin, Yi R. Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji og Tong Zhang, *R-Tuning: Instructing Large Language Models to Say 'I Don't Know'*, Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2024).