

FEBRUARY 2025

Can you trust AI for legal work?

An empirical study into the quality and applicability of LLM advice.

Employment law • Contract law • Intellectual property
• Privacy and GDPR • Corporate law • Confidentiality •
Professional liability

Mark Zijlstra	Head of Legal Technology
Arnoud Engelfriet	Chief Knowledge Officer
Guido Grevink	Business Development Manager
Carla Knoope	Senior Legal Counsel
Martijn Michael	Senior Legal Counsel
Suzanne Peekel	Senior Legal Counsel
Steven Ras	Chief Executive Officer
Bram de Vos	Senior Legal Counsel
Marco Caspers	Senior Legal Counsel

Photography

Joep Hijwegen

Licence

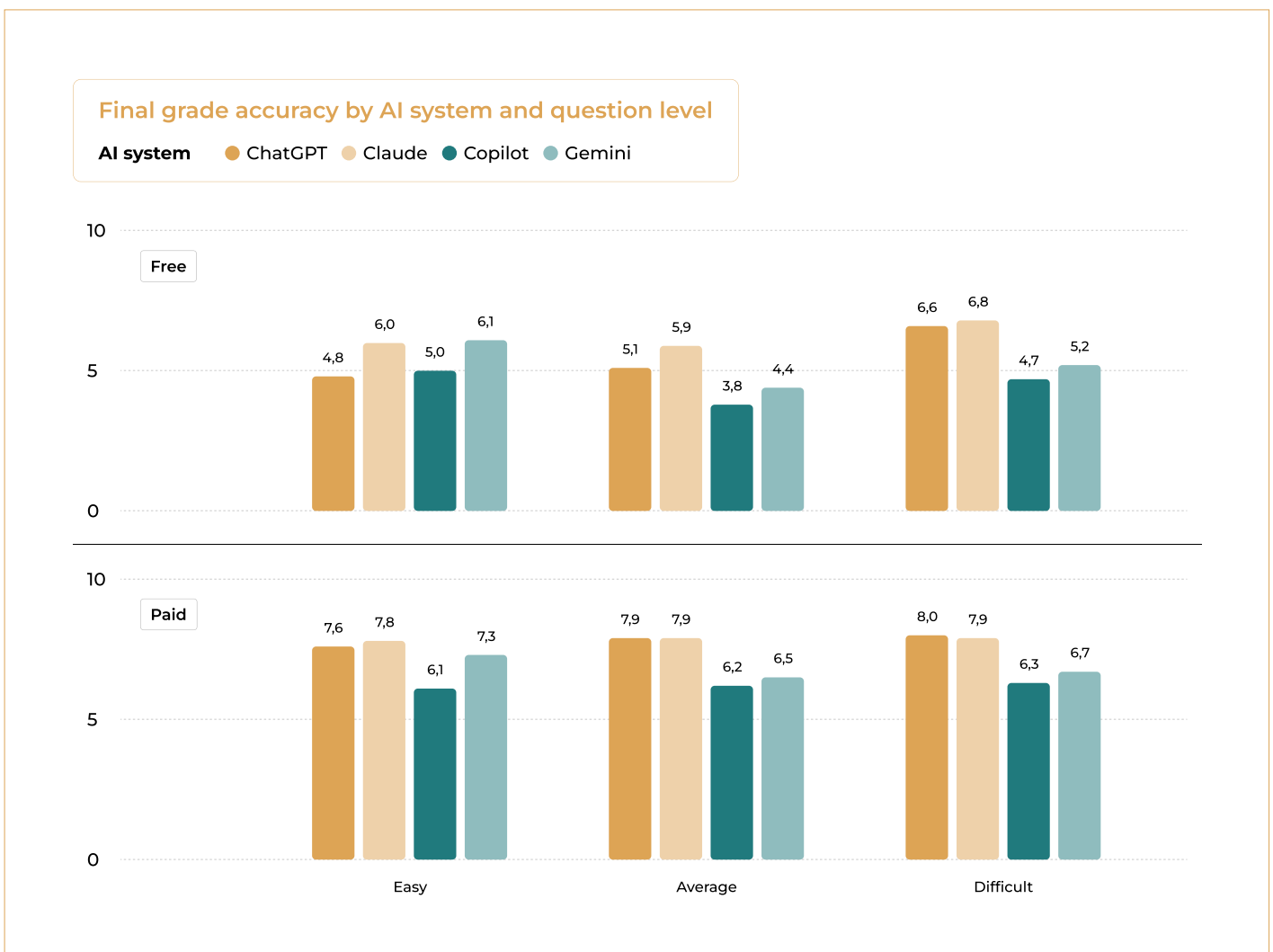
All content from this publication may be reproduced and or made public, in print, electronically or by any other means, subject to the terms of the Creative Commons Attribution-ShareAlike 4.0 Netherlands licence.



Management summary

This report analyzes the practical use of AI systems in legal work. The conclusions are based on the analysis of 75 legal questions at three difficulty levels across five legal areas: labor law, contract law, intellectual property, corporate law, and privacy/data protection. We examined four well-known AI systems in both their free and paid versions: ChatGPT, Claude, Copilot, and Gemini.

INTRODUCTION



You can trust AI systems for exploratory legal work.

The research shows that AI (in the paid version of the examined services) performs very well as an initial exploration of a legal question or in structuring legal information - think of drafting a checklist for due diligence or categorizing contract clauses. Use requires strict frameworks: every source reference must be verified, not randomly but systematically. AI use must be embedded in broader work processes with explicit control moments.

Claude and ChatGPT lead in legal AI, but only in their paid versions.

The paid model Claude 3.5 Sonnet ranked best in the test, directly followed by ChatGPT 4o Plus. The free versions lag far behind. Where both leaders consistently score high on complex questions, we see significant quality fluctuations with Copilot and Gemini.

AI provides extensive answers but often lacks legal depth.

The paid versions of Claude and ChatGPT score highest on relevance, but all systems tend to deliver irrelevant expositions. In terms of completeness, we see similar patterns: Claude delivers the most comprehensive analyses, but no system consistently achieves the highest level of legal depth. This problem is just as much present in the paid versions as in the free versions.

The biggest risk is blind trust in the legal knowledge of the systems.

All examined systems tend to occasionally cite non-existent case law and articles of law, alongside the often correct legal sources. This also applies to the paid versions, although less so. We also see answers that are devised from non-existent legal principles. Due to the high persuasiveness with which answers are presented, this quickly creates unjustified trust.

Legal work cannot be done with free AI tools.

The quality of the free versions of the examined models is substandard, both legally and in terms of sources. The paid versions show a significant leap in all areas. Compliance requirements can only be met with the paid versions of these services.

AI performs better in complex analyses than in legal questions.

Where one would expect AI systems to perform best on simple legal questions, we see higher scores on issues of medium and high complexity. This suggests that the systems are better at analyzing complex but logically structured issues than at reproducing basic legal knowledge.

European law suits AI better than Dutch law.

In harmonized European legal areas such as privacy and intellectual property, the systems perform strongly, while they fall short in nationally oriented law such as labor and corporate law. This difference becomes even more pronounced in more complex issues, where the systems consistently score high in privacy law but lose ground in labor law.

AI is particularly strong as a compass, not as an endpoint.

The systems perform very well in (1) quickly exploring legal issues, (2) legal areas with a strong systematic structure and crystallized case law, and (3) analyzing complex but logically structured legal issues. This makes them particularly valuable as an orientation tool for company lawyers who regularly need to explore new legal areas. The risk of using incorrect sources is least significant here.

Compliance requires active management by the lawyer.

Enterprise versions offer better guarantees, but the ultimate responsibility remains with the legal professional. For large organizations with sensitive data, Microsoft's enterprise solution with EU data boundaries is the recommended choice. For SME use the paid versions of ChatGPT, Claude, or Gemini are acceptable alternatives.

SUMMARY

The difference between winning and losing is made by lawyers who can effectively use AI. Organizations that invest in paid AI tools and thorough verification processes develop a new form of legal expertise that

strengthens traditional legal knowledge with technological judgment. In a market where speed and accuracy are becoming important, this will be the distinguishing feature of progressive legal departments.

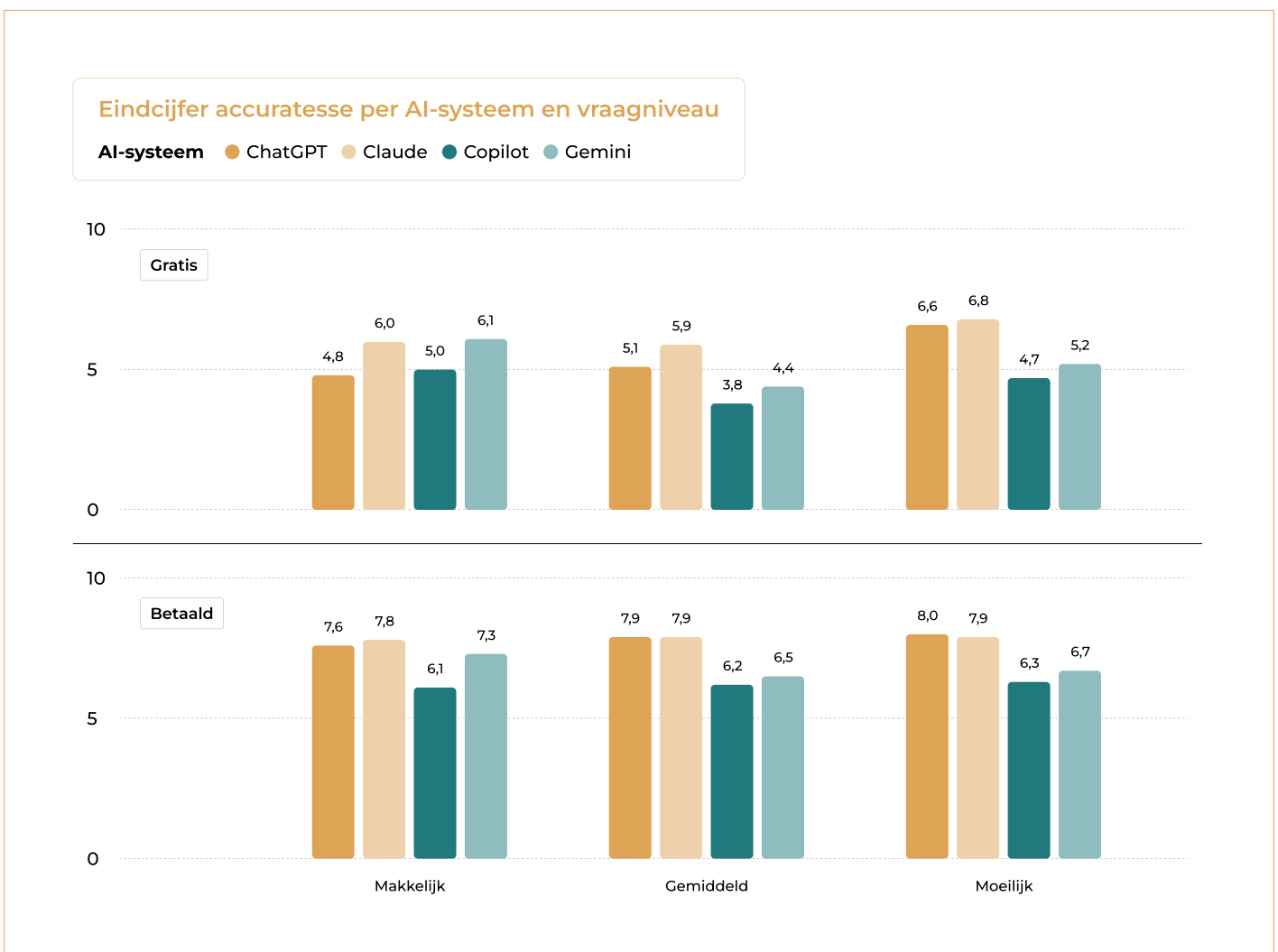
Want to know more? Suggestions? Other experiences? Let us know: contact@ictrecht.nl



Managementsamenvatting

Dit rapport analyseert de praktische inzet van AI-systemen in juridisch werk. De conclusies zijn gebaseerd op analyse van 75 juridische vragen, op drie moeilijkheidsniveaus en verdeeld over vijf rechtsgebieden: arbeidsrecht, contractenrecht, intellectueel eigendom, ondernemingsrecht en privacy/gegevensbescherming. We onderzochten vier bekende AI-systemen in zowel hun gratis als betaalde versies: ChatGPT, Claude, Copilot en Gemini.

INTRODUCTIE



Je kunt AI-systemen vertrouwen voor verkennend juridisch werk.

Het onderzoek laat zien dat AI (in de betaalde versie van de onderzochte diensten) uitstekend kan functioneren als eerste verkenning van een rechtsvraag of bij het structureren van juridische informatie - denk aan het opstellen van een checklist voor due diligence of het categoriseren van contractbepalingen. Inzet vraagt wel om strikte kaders: elke bronverwijzing moet worden geverifieerd, niet steekproefsgewijs maar systematisch. AI-gebruik moet worden ingebed in bredere werkprocessen met expliciete controlemomenten.

Claude en ChatGPT gaan aan kop bij juridische AI, maar alleen in hun betaalde versies.

Het betaalde model Claude 3.5 Sonnet komt als beste uit de test, direct gevolgd door ChatGPT 4o Plus. De gratis versies blijven daar ver bij achter. Waar de beide koplopers consistent hoge scores behalen bij complexe vragen, zien we bij Copilot en Gemini grote schommelingen in kwaliteit.

AI geeft uitgebreide antwoorden, maar mist vaak juridische diepgang.

De betaalde versies van Claude en ChatGPT scoren het hoogst op relevantie, maar alle systemen hebben de neiging tot irrelevante uitweidingen. Bij volledigheid zien we vergelijkbare patronen: Claude levert de meest complete analyses, maar geen enkel systeem bereikt consistent het hoogste niveau van juridische diepgang. Dit probleem bestaat net zo goed in de betaalde versies als in de gratis versies.

Het grootste risico is blind vertrouwen in juridische kennis van de systemen.

Alle onderzochte systemen hebben de neiging om af en toe niet-bestaande jurisprudentie en wetsartikelen te citeren, naast de vaak wél correcte rechtsbronnen. Dit geldt ook in de betaalde versies, hoewel daar minder. Ook zien we redeneringen die uitgaan van niet-bestaande rechtsbeginselen. Vanwege de grote overtuigingskracht waarmee antwoorden worden geponeerd, wekt dit al snel onterecht vertrouwen.

Juridisch werk kan niet met gratis AI tools.

De kwaliteit van de gratis versies van de onderzochte modellen is onder de maat, zowel juridisch-inhoudelijk als op het gebied van bronnen. De betaalde versies laten een forse sprong zien op alle gebieden. Ook compliance-eisen zijn alleen te realiseren bij gebruik van de betaalde varianten van deze diensten.

AI presteert beter bij complexe analyses dan bij juridische vragen.

Waar men zou verwachten dat AI-systemen het beste presteren bij eenvoudige juridische vragen, zien we juist hogere scores bij vraagstukken van gemiddelde en hoge complexiteit. Dit suggereert dat de systemen beter zijn in het analyseren van complexe maar logisch gestructureerde vraagstukken dan in het reproduceren van juridische basiskennis.

Europees recht gaat AI beter af dan Nederlands recht.

Bij Europees geharmoniseerde rechtsgebieden zoals privacy en intellectueel eigendom presteren de systemen sterk, terwijl ze bij nationaal georiënteerd recht zoals arbeids- en ondernemingsrecht tekortschieten. Dit verschil wordt nog geprononceerder bij complexere vraagstukken, waar de systemen bij privacyrecht consistent hoge scores behalen maar bij arbeidsrecht juist terrein verliezen.

AI is vooral sterk als kompas, niet als eindstation.

De systemen presteren uitstekend bij (1) het snel verkennen van juridische vraagstukken, (2) rechtsgebieden met een sterk systematische opbouw en uitgekristalliseerde rechtspraak, en (3) het analyseren van complexe maar logisch gestructureerde juridische vraagstukken. Dit maakt ze bijzonder waardevol als oriëntatie-instrument voor de bedrijfsjurist die regelmatig nieuwe rechtsgebieden moet verkennen. Het risico op gebruik van verkeerde bronnen is hierbij het minst significant.

Compliance vereist actieve regie door de jurist.

Enterprise-versies bieden weliswaar betere waarborgen, maar de eindverantwoordelijkheid blijft bij de juridische

SAMENVATTING

professional. Voor grote organisaties met gevoelige data is Microsoft's enterprise-oplossing met EU data boundaries de aangewezen keuze, terwijl voor MKB-gebruik de betaalde versies van ChatGPT, Claude of Gemini een acceptabel alternatief zijn.

Het verschil tussen winnen en verliezen wordt gemaakt door juristen die AI effectief kunnen inzetten. Organisaties

die investeren in betaalde AI-tools én gedegen verificatieprocessen ontwikkelen een nieuwe vorm van juridische expertise die klassieke juridische kennis versterkt met technologisch oordeelsvermogen. In een markt waar snelheid en accuratesse steeds belangrijker worden, wordt dit het onderscheidende kenmerk van vooruitstrevende juridische afdelingen.

Meer weten? Suggesties? Andere ervaringen? Laat het ons weten: contact@ictrecht.nl





Table of contents

Management summary	4
Management summary (Dutch)	7
Table of contents	11
The rise of AI in legal work	12
Research design and methodology	13
Areas of law	13
Simple questions	13
Assessment: more than just right or wrong	14
Results by area of law	18
Accuracy by area of law	18
Relevance and completeness	29
Conclusion: where AI excels and where it does not	32
Three specific situations	33
Handling of sources	33
The paradox of trust	33
Can you trust AI for legal work?	34
Risk management and liability	35
Data protection and privacy	35
Confidentiality and secrecy	37
Copyright concerns	38
The AI Act: a matter of literacy	39
Professional liability in the AI era	40
Outlook and recommendations	42
The balance between quality and compliance	42
Concrete recommendations for legal departments	43
Concrete recommendations for the individual lawyer	44
AI, the Digital Decade and consultancy firm ICTRecht	45

The rise of AI in legal work

“That legal chatbot actually gave a pretty good answer, didn’t it?” That has been the experience of many lawyers since the end of 2022. While IT specialists had been talking for years about the revolutionary impact of artificial intelligence, lawyers had little use for it. Until ChatGPT appeared. Suddenly, anyone could ask questions about legal issues in plain language, and the answers were surprisingly usable.

But how usable, exactly? Around us we heard fascinating experiences, both hopeful and worrying. One in-house lawyer told us how ChatGPT generated a privacy statement in five minutes that looked “very professional”. On closer inspection, however, it lacked crucial GDPR requirements. A lawyer used Claude to summarise case law, but discovered that the AI regularly “cited” non-existent judgments. And that is precisely the problem: the answers are beautifully structured, use the right jargon and sound convincing. Legal popcorn that reads easily but offers little nourishment.

These experiences raise practical questions. Not about a distant future in which AI systems replace lawyers, which for now remains science fiction. But about the here and now: how reliable are the answers these systems provide?

What types of questions can they handle? And, more importantly, which ones can they not?

We examined this systematically. We presented hundreds of legal questions to four popular AI systems, ranging from simple questions of statutory knowledge to complex cases. The answers were assessed by experienced lawyers. This report presents the results, and they are fascinating.

This report does not contain theoretical reflections on the future of the legal profession. Instead, it offers concrete data on what these systems can and cannot do, practical examples of successes and failures, and clear guidelines for responsible use.

Bringing work in-house through internal AI capabilities will become increasingly attractive in 2025.

(Deloitte, January 2025)

Research design & methodology

Lawyers are trained to weigh every detail and question every assumption. How, then, do you test whether AI systems are capable of handling legal work?

This research confronted us with a classic dilemma. On the one hand, we wanted to test the systems thoroughly across different areas of law and levels of complexity. On the other hand, the research needed to reflect daily practice. Not an academic exercise, but a realistic stress test.

Areas of law

We selected five areas of law that are central to business practice:

1. Privacy (GDPR, cookie legislation, data breaches)
2. Contract law (general terms and conditions, liability)
3. Intellectual property (copyright, trademarks, licences)
4. Corporate law (governance, mergers and acquisitions)
5. Employment law (dismissal, reorganisations, terms of employment)

For each area of law, we formulated three types of questions:

1. Basic questions about legislation (“Which transfer mechanisms does the GDPR provide?”)
2. Practical questions involving case law (“How does the protection of databases under the Database Directive differ from protection under copyright law?”)

3. Complex cases (“An employee subject to a non-compete clause wants to move to a new employer after a merger, but the new role partially overlaps, since the merger, with the prohibited field of work. What are the consequences?”)

The full list of questions per area of law and type is available on our website via the link [AI Research: Appendix questions and output](#).

Simple questions

For the purposes of this research, we deliberately chose to use simple, direct questions. Not because we did not know better. By now, an entire field has emerged around prompt engineering, the optimisation of instructions given to AI systems. Experienced users know that better answers can be obtained by adding context (“You are a Dutch lawyer with 20 years of experience”), providing specific instructions (“Cite relevant legislation and case law”), or prescribing the format of the answer (“Structure your answer in three parts”). Several tips can be downloaded in our factsheet “[Prompt tricks for lawyers](#)”.

Why, then, did we still opt for basic prompts such as “Which transfer mechanisms does the GDPR provide?” or “A graphic design can be protected under both copyright and trademark law. What are the main differences in the consequences of this protection?”

- First, realism. The average lawyer does not have time to become an expert in AI systems. They want to ask a question and receive a usable answer. Complex prompts are like an advanced search engine that delivers perfect results, provided you know exactly which search operators to use. Useful for experts, but not

representative of everyday use.

- Second, a methodological consideration. Once you introduce complex prompts, it becomes more difficult to compare systems fairly. Is a better answer the result of the AI system's capabilities, or of a smarter prompt? By using basic prompts, we measure the systems' raw intelligence.
- Perhaps most importantly, our approach tests the lower bound of what these systems can do. If an AI system already provides legally correct, relevant and complete answers with minimal instructions, it can only perform better with optimal prompts. The reverse is not true. Good performance with perfectly calibrated prompts says little about usability in practice.

Assessment: more than just right or wrong

Legal work involves more than simply the correctness of facts and statutory provisions. Valuable legal advice combines accuracy with relevance and depth, expressed in clear and readable language. For that reason, we developed an assessment model that measures these three dimensions of legal quality. Just as a judge does not only consider whether a lawyer has cited the law correctly, but also whether the argument is relevant and sufficiently substantiated.

Assessment model

We assess each AI answer in the same way an in-house lawyer would evaluate external advice, across three dimensions:

Accuracy
<i>Are the cited statutory provisions, time limits and requirements correct?</i>
Correct facts or references where requested: +1 point
Incorrect facts or references: -1 point
Correct but unrequested information: 0 points
Failure to provide requested information: 0 points

Relevance
<i>Does the answer address the core of the question?</i>
Exclusively irrelevant information: 1 point
Core addressed with some digressions: 6 points
Fully focused on the issue: 10 points

Completeness
<i>What is the legal depth of the analysis?</i>
Extremely superficial treatment: 1 point
Basic legal analysis: 6 points
Very in-depth legal substantiation: 10 points

We illustrate this assessment model with an example from employment law:

Aspect	Details
Prompt	<p>“What are the requirements for a valid summary dismissal of a sick employee who takes on freelance work during illness?”</p>
AI answer	<ul style="list-style-type: none"> ● “In the case of summary dismissal, there must be an urgent reason that justifies immediate dismissal (Article 7:677 of the Dutch Civil Code). The employer must act with due care and communicate the dismissal without delay. As the Amsterdam Court of Appeal ruled in the case of Jansen v Retail BV (2023), the employer must be able to demonstrate that the freelance work hampers recovery or conflicts with the company’s interests. Important factors include: ● The nature and scope of the ancillary activities ● The impact on the recovery process ● Whether a prohibition on ancillary activities exists ● The length of service ● The employee’s role <p>It is advisable to issue a warning first and to document the dismissal properly.”</p>
Model answer	<p>In addition to the general requirements for summary dismissal (urgent reason, immediacy and prompt notification, Articles 7:677 and 7:678 of the Dutch Civil Code), additional safeguards apply to a sick employee due to the prohibition on termination during illness (Article 7:670(1) of the Dutch Civil Code). The mere fact that an employee engages in ancillary work during illness is insufficient for a valid summary dismissal. The employer must demonstrate that the ancillary work hampers recovery and or conflicts with concrete reintegration arrangements.</p>

We illustrate the assessment model with an example from employment law. This example demonstrates a classic AI problem. The system provides an apparently complete answer with a list of factors that sound logical, but it misses the legal core of the issue, namely the prohibition on termination during illness, and invents case law. The answer may sound convincing to a layperson, but for an in-house lawyer who is not specialised in employment law and is seeking a quick initial answer, it creates additional risks.

In terms of accuracy, this answer scores 1 point: 1 point for the correct statutory reference, 1 point for the factors relating to the urgent reason, and minus 1 point for the fabricated case law. The relevance of the advice is limited, scoring 6 points, as it mainly contains general digressions and misses the core issue. The advice is highly incomplete, scoring 4 points, because the question was not about when an urgent reason exists, but about how illness affects the assessment.

The systems tested

We tested four leading AI systems, each in both their free and paid versions. This allowed us not only to compare the systems with each other, but also to assess whether the paid versions deliver significantly better legal performance.

The tests were conducted in two phases:

- December 2024: free versions (each question asked twice)
- January 2025: paid versions (single test round)

For the free versions, we asked each question twice in order to measure the consistency of the answers. For the paid versions, it quickly became apparent that the answers were so consistent that a second test round added no value.

We deliberately chose to use the standard chat interface, without system prompts or style settings. This reflects how most lawyers are likely to use these systems in practice.

Although advanced prompt techniques can lead to better results (see Appendix B), we wanted to test how the systems perform under normal use.









Shortly after completing the tests, the Chinese company Hangzhou DeepSeek AI released its model DeepSeek-R1. In light of the planned publication of this study and the immediate concerns raised about censorship and bias in this model, we decided not to include it in our tests.

What we did not do

Equally important as what we did examine is what we deliberately left out:

- No complex prompt engineering
- No follow-up questions or requests for clarification
- No legal databases or specialist tools
- No questions requiring specialist expertise
- No comparison with human lawyers

That is because the core question is not whether AI can outperform lawyers, but whether these systems can be trusted for everyday legal work. As the next chapter shows, the answer is nuanced and some of the results are surprising.

Free	Paid	Price*
 ChatGPT 4o mini	 ChatGPT 4o (Plus)	€ 20/user/month
 Claude 3.5 Haiku	 Claude 3.5 Sonnet	€ 20/user/month
 Copilot (web)	 Copilot (Office 365)	€ 22/user/month
 Gemini 1.5	 Gemini 2.0 Flash Experimental	€ 20/user/month

*Prices include VAT, based on the cheapest subscription plan, as applicable in January 2025



Results by area of law

This section presents the research results systematically by area of law. We first analyse the accuracy of the legal answers, followed by an assessment of their relevance and completeness. For each aspect, we compare the performance of both the free and paid versions of the four systems tested. We conclude with an analysis of specific strengths and weaknesses in practical application. The full list of questions per area of law and type can be found on our website via the link [AI Research: Appendix questions and output](#).

Accuracy by area of law

We begin with accuracy: how substantively correct are the answers? We assess this per area of law.

Privacy and data protection

For this first area of law, we selected a mix of theoretical basic questions, practice-oriented applications and complex policy issues. The simple questions focused on fundamental GDPR knowledge that every lawyer should have readily available, such as “Which transfer mechanisms does the GDPR provide?” or “Which rights do data subjects have?”. These questions test pure knowledge of the statutory framework.

The intermediate category consisted of questions that require a combination of statutory knowledge and legal analysis, such as “Can a commercial interest qualify as a legitimate interest under the GDPR?” or “Is an IP address personal data?”. Here, systems must not only know the law but also be able to apply relevant case law.

The most difficult category consisted of complex practical situations combining multiple GDPR aspects. One example was: “An airline runs an advertising campaign via a social media platform. A data breach occurs. Who is required to report it?” Such questions test the ability to combine different data protection concepts, such as joint controllership and data breach notification obligations, in a practical analysis. This layered approach, moving from pure statutory knowledge to case law analysis and practical application, allowed us to test not only the theoretical knowledge of AI systems but also their ability to apply that knowledge in realistic scenarios.

The results in this area of law show a clear pattern, with the paid version of Claude standing out as a clear outlier.

	ChatGPT		Claude		Copilot		Gemini	
	Free	Paid	Free	Free	Free	Paid	Free	Paid
Easy	8,9	9,1	9,6	9,1	7,4	8,7	9,8	8,3
Average	4,8	8,0	5,0	9,5	1,8	8,5	5,3	6,0
Difficult	5,6	7,2	4,7	9,4	3,9	8,9	4,4	6,1
	6,4	8,1	6,4	9,3	4,4	8,7	6,5	6,8

“How quickly must a data breach be reported to the Dutch Data Protection Authority?”

ChatGPT (free): Under the GDPR, a data breach must be reported to the Dutch Data Protection Authority within 72 hours of discovery, unless the breach is unlikely to result in a risk to the rights and freedoms of data subjects. This deadline applies to breaches involving personal data and for which the responsible party (the organisation processing the data) is aware of the incident.

The notification obligation requires the organisation to inform the Authority of: (...)”

Claude (paid):“A data breach must be reported to the Dutch Data Protection Authority within 72 hours of discovery. This period starts as soon as the controller has a reasonable degree of certainty that a security incident has occurred in which personal data has been exposed to loss or unlawful processing.

Important additional points include:

- If it is not possible to provide all information within 72 hours, the notification may be made in phases
- The notification must nevertheless be made without undue delay (...)”

While the free versions perform reasonably well only on basic questions, the paid versions show strong performance across the board. On simple privacy questions, all systems score highly, with scores ranging from 7.4 to 9.8 for the free versions and from 8.3 to 9.1 for the paid versions. The difference becomes dramatic with more complex questions. Where the free versions show a sharp decline in accuracy, the paid versions largely maintain their level, with scores between 6.0 and 9.5. Claude Sonnet stands out in particular for its remarkably consistent and very strong performance.

The quality differences are particularly apparent in the handling of case law and statutory structure. The paid versions prove to be well informed about recent developments, a performance that may be attributable to the international nature of the GDPR, which provides more training material. Even so, typical AI pitfalls remain visible in the premium versions. ChatGPT and Claude

regularly produce “half-correct” references, for example by naming the correct case but citing an incorrect ECLI number.

Copilot, in its paid version, distinguishes itself by actively linking to up-to-date sources such as EUR-Lex and legal blogs, although the analyses sometimes remain superficial. Gemini Advanced delivers accurate case law analyses but without precise references, which makes verification more difficult. In complex and concrete questions, such as whether a company may track its truck drivers using GPS without consent, Google and Copilot tend to give shorter answers, often accompanied by disclaimers advising consultation of a lawyer.

A persistent problem, even in the paid versions, remains the handling of fundamental data protection concepts. For example, all systems fail to address whether a public authority can rely on legitimate interest, which is not

possible in the context of public tasks, and struggle with the correct qualification of parties as processors or controllers. This suggests that even advanced AI systems have difficulty with more abstract legal principles that cannot be directly derived from statutory text.

Contracts

In contract law, we focused on three levels: basic statutory knowledge, the application of contractual principles, and the drafting of specific clauses. The straightforward questions tested fundamental knowledge that every lawyer should master, such as “What are the differences between terminating and rescinding a contract?” or “What is the difference between void and voidable agreements?”. These questions require precise insight into statutory definitions and legal consequences. The intermediate category involved more complex analyses, such as the question “What notice period must be

observed for a continuing performance contract if no contractual period has been agreed?”. Here, systems must not only know the law but also integrate relevant case law and distinguish between different types of contracts. The most difficult category consisted of drafting concrete contractual provisions for specific situations, such as “Draft a market-standard liability clause for a SaaS provider” or “Draft an audit clause for a software supplier”. These assignments test the ability to translate legal knowledge into practical, balanced and market-standard provisions.

In this area of law, the paid version of ChatGPT scores the highest, albeit with a modest 7.1.

	ChatGPT		Claude		Copilot		Gemini	
	Free	Paid	Free	Paid	Free	Paid	Free	Paid
Easy	4,0	8,0	3,7	7,3	4,7	4,0	5,7	7,3
Average	2,5	5,6	4,1	5,0	3,4	6,3	2,8	6,9
Difficult	6,2	7,8	7,1	7,4	5,8	4,3	4,5	7,6
	4,2	7,1	5,0	6,6	4,6	4,8	4,3	7,3



“Which elements of a software solution do or do not qualify for copyright protection, and which legislation and case law support this?”

Copilot (paid): Copyright protection of software is a complex topic. The main points are as follows.

Elements that may qualify for copyright protection:

1. **Source code and object code:** Source code and object code: the actual lines of code written by the programmer^[1].
2. **Documentation:** manuals, specifications and other accompanying documentation of a creative nature^[1].
3. **Graphical user interface (GUI):** the visual elements of the software, provided they have their own original character^[2].

Elements that do not qualify for copyright protection:

1. **Functional aspects:** the functionality of the software, such as how it works or the algorithms it uses^[1].
2. **Programming languages and protocols:** the languages and protocols used to write the software^[1].
3. **Ideas and principles:** the underlying ideas and principles behind the software^[2].

Legislation and case law:

Copyright protection for software is laid down in the Dutch Copyright Act. Article 1 of the Copyright Act provides that the author of a work has the exclusive right to make the work public and to reproduce it^[1]. In addition, case law further elaborates these principles, such as the judgment of the Court of Justice of the European Union in *SAS Institute Inc. v. World Programming Ltd.*, which confirmed that functionality, programming languages and data formats are not protected by copyright^[2].

Do you have any specific questions about this topic?

References

[1] [Auteursrecht software: dit zijn de 5 vragen - ICT-Advocatenkantoor Legalz](#)

[2] [Auteursrecht op software. Hoe zat het ook alweer?](#)

We observe that the free versions are hardly usable for day-to-day legal work. While the answers are often extensive and filled with contractual jargon, blatantly incorrect references and lists that make little sense occur frequently across all systems. Contract clauses may sound polished, but they are superficial and often miss the core of the question.

This is different for the paid versions. These achieve high scores, including on questions requiring legal analysis. Source references are now almost always correct. Notably, even the most verbose systems, ChatGPT and Claude, are less likely to wander off in their analyses and remain more focused on the core of the question.

The higher scores are largely attributable to better performance on European contract law issues. Where the free versions often became entangled in a mix of Dutch and US concepts, the paid versions show a clearer understanding of the European legal framework. This is reflected in more accurate references to EU law. The quality improvement lies primarily in the accuracy of the legal analysis, rather than in the structure or scope of the answers. In other words, the paid versions make fewer mistakes, but their answers are not necessarily more complete or more relevant.

It is interesting that both Gemini and Copilot, in their paid versions, more explicitly point out the need for legal review. Where the free versions sometimes appeared overconfident, the paid versions more often include sober disclaimers. This suggests a greater degree of legal restraint, possibly because these versions are better trained to recognise legal complexity.

A notable constant between the free and paid versions is their performance in drafting contractual provisions. Both variants produce usable draft texts that can serve as a basis for further development. The paid versions stand out primarily through clearer explanations of the proposed clauses and more pointed warnings about potential pitfalls. This confirms that these systems are particularly valuable as a source of inspiration and for initial drafts. The higher accuracy of the paid versions makes them a more reliable starting point, but this output still requires thorough legal review. In practice,

this means that the paid versions can mainly deliver time savings when preparing initial drafts and exploring legal issues, but they do not replace the judgement of the contract lawyer.

Intellectual property

In intellectual property law, we chose a structure that tests the interaction between fundamental principles and practical application. The simple questions focused on basic requirements and formalities, such as “What requirements apply to an exclusive licence?” or “How long does copyright last if the author is a legal entity?”. These questions test pure knowledge of the statutory framework. The intermediate category included questions exploring the boundaries of protection, such as “How does the protection of databases under the Database Directive differ from protection under copyright law?” or “What is the difference between the US concept of ‘fair use’ and the European right of quotation?”. Here, systems must not only know the rules, but also understand and compare the underlying principles.

The most difficult category confronted the systems with current challenges for which no settled case law yet exists, such as “Who owns the copyright in a text generated by a generative AI?”, or complex practical situations such as the protection of innovative algorithms in the automotive sector, where different protection regimes must be weighed against each other.

Performance levels are relatively close. ChatGPT, Claude and Copilot all achieve a strong average score of 8, with Claude performing slightly better in the most difficult category.

	ChatGPT		Claude		Copilot		Gemini	
	Free	Paid	Free	Paid	Free	Paid	Free	Paid
Easy	0,0	5,0	4,2	5,0	2,9	6,7	2,9	5,8
Average	7,9	10,0	9,7	10,0	6,1	8,9	5,8	7,4
Difficult	8,3	9,1	9,1	9,6	5,2	8,7	6,5	8,3
	5,4	8,0	7,7	8,2	4,7	8,1	5,1	7,2

A structural weakness in the field of intellectual property, including copyright, trademarks and patents, is that a great deal of information is available on the open internet, but much of it reflects US practice and views. This is evident in all answers, with the free version of ChatGPT performing particularly poorly for this reason. When a prompt explicitly refers to a specific country, all four free versions perform slightly better. The paid versions clearly contain more information about the Netherlands, at least at a general level. Claude stands out in terms of quality due to its correct references to legislation and case law, while ChatGPT calmly invents one judgment after another. Copilot provides very short answers and therefore does not reach case law or article numbers unless explicitly prompted. The sources are then correct, but an arbitrary legal website is presented as being just as authoritative as a judgment of the Dutch Supreme Court. Gemini refers in general terms to "case law" without being specific.

Paradoxically, the intermediate and difficult questions perform better than the simple ones, because they allow more room for digressions and general explanations that still score points under the assessment model. For this reason, ChatGPT and Claude score highly, with Claude approaching the maximum score. Copilot quickly reverts to US concepts, such as fair use, even where European rules are explicitly requested. The answers remain superficial and appear aimed at helping laypersons rather than advising lawyers. Copilot also frequently states that the question is complex and that it is better to consult a lawyer. Gemini shows a similar pattern.

Employment law

For employment law, we chose a mix of everyday basic questions and complex situations that cut across multiple areas of law. The simple questions focused on fundamental employee rights that every lawyer regularly encounters, such as "Am I entitled to a break during an eight-hour working day?" or "What is false self-employment?". These questions test pure knowledge of employment legislation that is directly applicable in daily practice. The intermediate category included questions requiring a deeper understanding of employment law principles, such as "What rights does an employee have when a company is taken over?" or "Which legal criteria apply when determining whether a role is suitable in the context of reintegration?". Here, systems must not only know the law, but also be able to apply relevant case law and weigh different interests.

The most difficult category confronted the systems with situations in which employment law, privacy law and good employment practices intersect, such as an employee reporting sick after an escalating conflict, or the international dimension of working remotely from abroad. These questions test the ability to integrate different areas of law into practical advice.

In this area of law, ChatGPT performs most strongly, closely followed by Google's Gemini. Claude leaves a considerable number of points on the table compared with other areas of law, while Copilot is best avoided altogether for employment law questions.

“To what extent is including an e-book in an LLM permitted under European copyright law?”

Gemini (free): “Including an e-book in a Large Language Model (LLM) raises complex legal questions under European copyright law.

Key question: is this permitted use?

The answer is not unequivocal and depends on several factors, including:

- **The volume of text included:** the larger the portion of the e-book incorporated into the LLM, the greater the likelihood of copyright infringement.
- **The nature of the use:** is the e-book used to train the LLM or to answer specific questions? In the former case, copyright infringement is more likely.
- **The impact on the market:** does the use of the e-book within the LLM affect the market for the original work? If the LLM can reproduce large parts of the book, this may harm sales.

Relevant provisions of European copyright law

- **Right of reproduction:** copyright grants the author the exclusive right to reproduce the work. Including an e-book in an LLM may be regarded as a form of reproduction.
- (...)”

Claude (paid): “This question touches on recent developments in European copyright law relating to text and data mining (TDM). The legal analysis is as follows:

1. General framework

- Articles 3 and 4 of the DSM Directive (2019/790) regulate text and data mining
- Article 3 concerns TDM for scientific research
- Article 4 concerns other forms of TDM

2. Analysis under Article 4 DSM Directive

- Training an LLM falls within the definition of text and data mining
- In principle, a TDM exception applies that permits this
- However: rightholders may object (“opt-out”)
- An opt-out must be communicated in an “appropriate manner”
- For online content, this means a machine-readable opt-out (...)”

	ChatGPT		Claude		Copilot		Gemini	
	Free	Paid	Free	Paid	Free	Paid	Free	Paid
Easy	5,6	8,4	5,2	8,1	4,7	4,8	6,1	8,4
Average	5,6	7,4	4,7	6,2	2,7	2,6	3,5	7,2
Difficult	7,0	7,1	6,0	5,2	2,9	3,1	3,6	5,5
	6,1	7,6	5,3	6,5	3,4	3,5	4,4	7,0

The scores for employment law show a stable pattern in the free versions for easy and medium questions, with a slight increase for complex cases. ChatGPT shows the most consistent performance here. In the paid versions, we see a clear upgrade step: these versions consistently

achieve higher scores. Notably, Copilot is the only system that does not make this leap and remains at similar scores in both versions.

“Am I entitled to continued payment of wages during illness?”

Copilot (paid): “Yes, in the Netherlands you are entitled to continued payment of wages during illness. Under the law, your employer must continue to pay at least 70 percent of your wages for up to two years if you are ill^{[1][2]}. In the first year, this amount must be at least equal to the statutory minimum wage^{[1][2]}.

There may be exceptions, depending on your contract or collective labour agreement. For example, if you work fewer than four days per week as a domestic worker, your employer only has to continue paying wages for six weeks^[1].

Do you have any specific questions about your situation?”

References

[1] [Krijg ik loon doorbetaald bij ziekte? | Het Juridisch Loket](#)

[2] [Hoeveel loon krijg ik doorbetaald als ik ziek ben?](#)

ChatGPT (free): “Yes, in the Netherlands you are entitled to continued payment of wages during illness, but the precise conditions depend on your employment contract, the applicable collective labour agreement, and your employment status (employee, self-employed contractor, etc.).”

ChatGPT and Gemini struggled to achieve a passing score on easy questions, while Claude and Copilot had clear difficulty even with seemingly simple questions. These concerned basic matters such as continued payment of wages during illness, entitlement to breaks, and the definition of false self-employment. On average, the systems are therefore able to answer simple employment law questions to some extent adequately. However, the scores vary considerably between systems, indicating inconsistency in the reliability of the information provided. For simple questions, ChatGPT performs best.

The tested systems struggled more clearly with questions of medium difficulty. Here, the average score across all systems dropped to a clear failure: 4.1. The answers lacked sufficient depth, relevant information was missing, and in some cases incorrect information was even presented. ChatGPT again stood out for these medium questions.

Curiously, the systems performed better on difficult questions than on questions of medium difficulty. We suspect this is due to how the systems analyse more complex questions and prioritise completeness over simplicity. The structure of the questions may also have contributed, as more complex questions provide broader context and more details. ChatGPT and Claude both achieved a pass here. Nevertheless, the results still showed clear incompleteness and inaccuracies in places.

Reliability between systems is highly inconsistent, with scores ranging from 2.4 to 7.4. This means that while AI systems show potential in answering employment law questions, they also exhibit significant shortcomings in reliability and consistency. The best solution for employment law questions is ChatGPT, which provides clear and extensive answers. Copilot and Gemini in particular sometimes gave very short answers with little detail and advised contacting a lawyer. While that may well be the best advice given the results, it was not awarded points as it provides little direct assistance.

A relevant factor in employment law is changing legislation. Employment law is a dynamic field in which new legislation is regularly introduced, including in the areas the systems were asked about, such as flexible working and false self-employment. AI systems struggle

to process and correctly qualify such changes. It is therefore quite possible that the training data contains conflicting information based on both old and new legislation. In addition, some systems are subject to a data cut-off, meaning their knowledge stops updating at a certain point. This means that answers based on outdated laws or rules pose a risk where recent legislation is crucial. This further underlines the need for users to carry out additional checks against current sources.

Corporate law

In corporate law, we chose a combination of formal requirements and practical applications reflecting the daily reality of business. The easy questions focused on basic company law aspects that every in-house lawyer regularly encounters, such as “Who determines the level of remuneration of a statutory director?” or “When must annual accounts be prepared and filed at the latest?”. These questions test knowledge of fundamental statutory requirements that are crucial for corporate governance.

The medium category included questions on more complex corporate law constructions, such as “What are the conditions for a dividend distribution?” or “When is a company bound by acts of employees without formal authority to represent the company?”. Here, systems must not only know the law but also understand the practical implications. The most difficult category confronted the systems with drafting concrete legal documents, such as shareholder resolutions and drag-along or tag-along clauses, requiring technical legal knowledge to be translated into practically usable documents that meet all formal requirements. In this area, Claude again achieved the highest score, particularly due to very strong results on the easy questions. ChatGPT followed closely behind.

The picture in corporate law largely mirrors what we have already seen. The systems score slightly better on complex questions than on basic knowledge questions, mainly due to the drafting nature of these questions. As soon as specific legal details are required, accuracy quickly declines. This suggests that the systems are

better at imitating legal language than at ensuring legal correctness. The upgrade effect is clearly present here as well.

Claude performs best overall, mainly thanks to accurate statutory references and sound interpretation of formal legal requirements. ChatGPT follows closely, although its answers are often extensive and include irrelevant digressions.

	ChatGPT		Claude		Copilot		Gemini	
	Free	Paid	Free	Paid	Free	Paid	Free	Paid
Easy	5,7	7,3	7,5	9,5	5,2	6,4	6,1	6,8
Average	4,8	8,4	6,0	8,8	5,0	4,8	4,6	5,2
Difficult	6,2	9,0	7,1	8,1	5,7	6,7	6,7	6,2
	5,6	8,2	6,9	8,8	5,3	5,9	5,8	6,1



“The remuneration of a statutory director of a Dutch private limited liability company is determined by whom?”

Claude (free): “The general meeting of shareholders determines the remuneration of statutory directors, unless the articles of association allocate this authority to another body (usually the supervisory board, if one exists). This is laid down in Article 2:245 of the Dutch Civil Code.

In a structure company, this authority always lies with the supervisory board under Article 2:263 of the Dutch Civil Code.”

Gemini (free): “In the Netherlands, the salary of a statutory director of a private limited liability company (BV) is generally determined by the general meeting of shareholders. This means that the shareholders collectively decide how much the director is paid.

How does this work?

1. Proposal: often, the supervisory board (if one exists) submits a proposal on the level of remuneration.
2. Decision: the shareholders then meet in a general meeting and decide on the proposal. They may approve, reject or amend it.

What does the law say?

This is laid down in Dutch law, specifically Book 2 of the Dutch Civil Code. Article 2:245 provides that the remuneration of directors is determined by the general meeting, unless the articles of association provide otherwise.

Why is this arranged this way?

Control: giving shareholders the right to decide on remuneration provides a form of cost control.

Balancing interests: shareholders have an interest in sound business operations and healthy financial results. By setting remuneration, they can ensure it aligns with company performance.

In summary:

The remuneration of a BV director is not a fixed amount but is determined by the shareholders. This creates a balance between the interests of the director and those of the shareholders.

If you would like to know more about the rights and obligations of directors or the general meeting of shareholders, it is advisable to contact a notary or a lawyer specialising in corporate law.

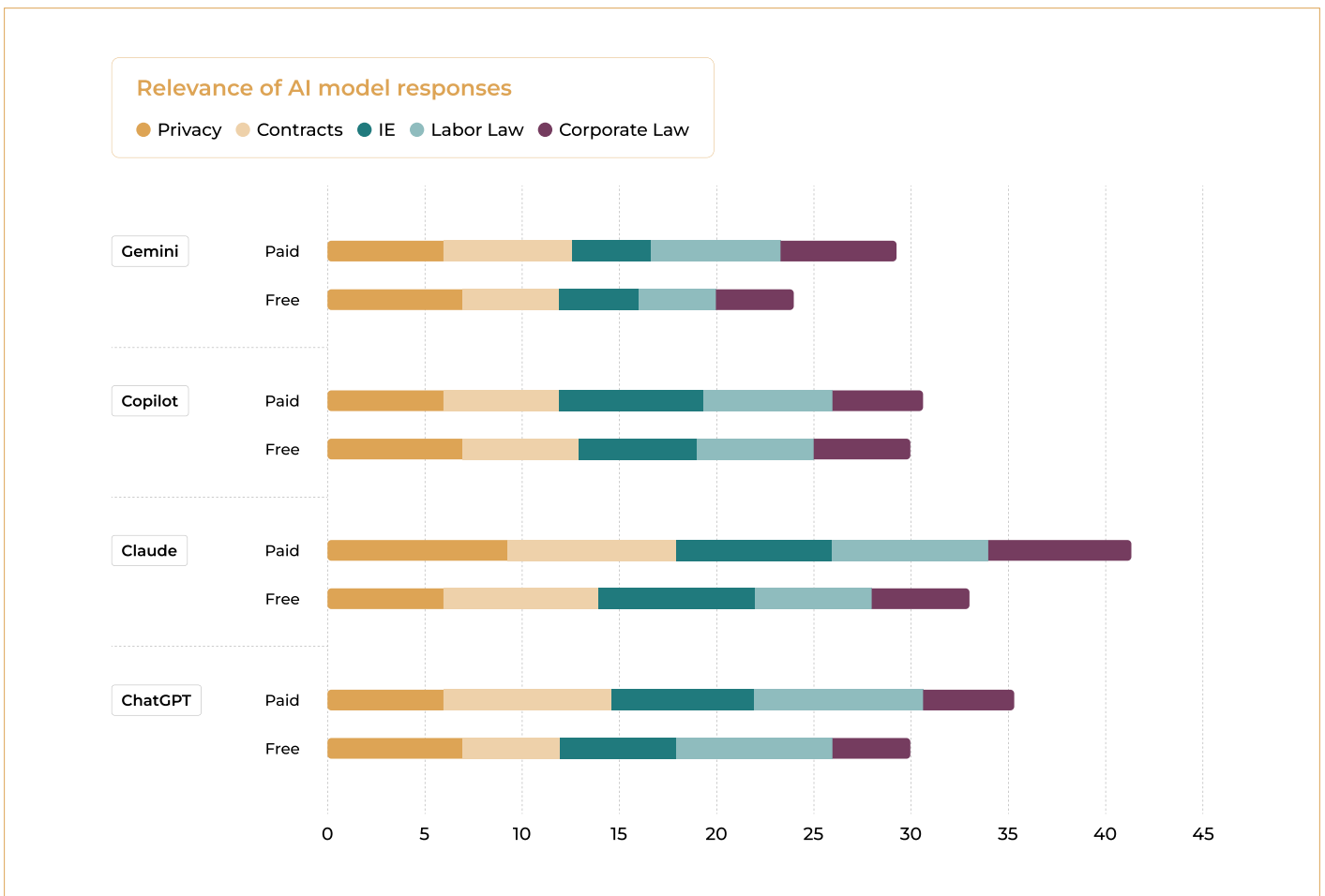
Please note: this information is intended as a general introduction and does not constitute legal advice. For personalised advice, it is recommended to consult a professional.”

The improvements mainly relate to reproducing correct legal information, not necessarily to delivering practically usable analyses. Put simply, the advice is mostly longer.

Where Copilot still uses legal jargon (albeit not always correctly and rarely with statutory references), Gemini tends to translate legal concepts into lay language. This

makes answers more accessible but often less precise. We also again see disclaimers, particularly with Copilot, such as “I am not a financial adviser and cannot draft a legal document such as a dividend resolution”.

Relevance: focus on the question



We first look at relevance, meaning the degree of focus on the question. A certain level of irrelevance is inherent to LLMs. All systems are clearly trained to produce narratives with a beginning and an end. We see many bullet lists and headings with the characteristic capitalisation used by LLMs. As a result, answers have a strong tendency to include irrelevant information. This includes not only reflective opening sentences (“An excellent question!

This is a complex legal issue”) or disclaimers (“consult a lawyer”), but also broader frameworks that were not requested. For the IP question on whether an exclusive licence must be in writing, almost all models provide a list of contractual provisions that should be included in such a licence. ChatGPT often concludes with a summary that reiterates everything.

Claude stands out most positively, with consistent relevance scores of around 8 for complex questions, particularly in intellectual property and contract law. ChatGPT performs similarly in employment law but struggles to maintain focus in contract matters, where relevance drops sharply. Copilot and Gemini often remain at scores of 5 to 6, meaning they address the core issue but regularly digress.

In the paid versions, relevance scores show a modest improvement. Claude maintains consistently high scores across all legal areas, particularly strong in IP and contract law. ChatGPT Plus scores similarly high in employment law and privacy but drops back in corporate law.

Copilot remains close to its free counterpart, mainly scoring around 6, while Gemini Advanced improves slightly from an average of 4 to between 6 and 8. Overall, the paid versions mainly improve their ability to stay on topic, but fundamental digressions persist. Notably, relevance scores improve less dramatically than accuracy scores, suggesting that even premium versions struggle to deliver sharply focused analyses.



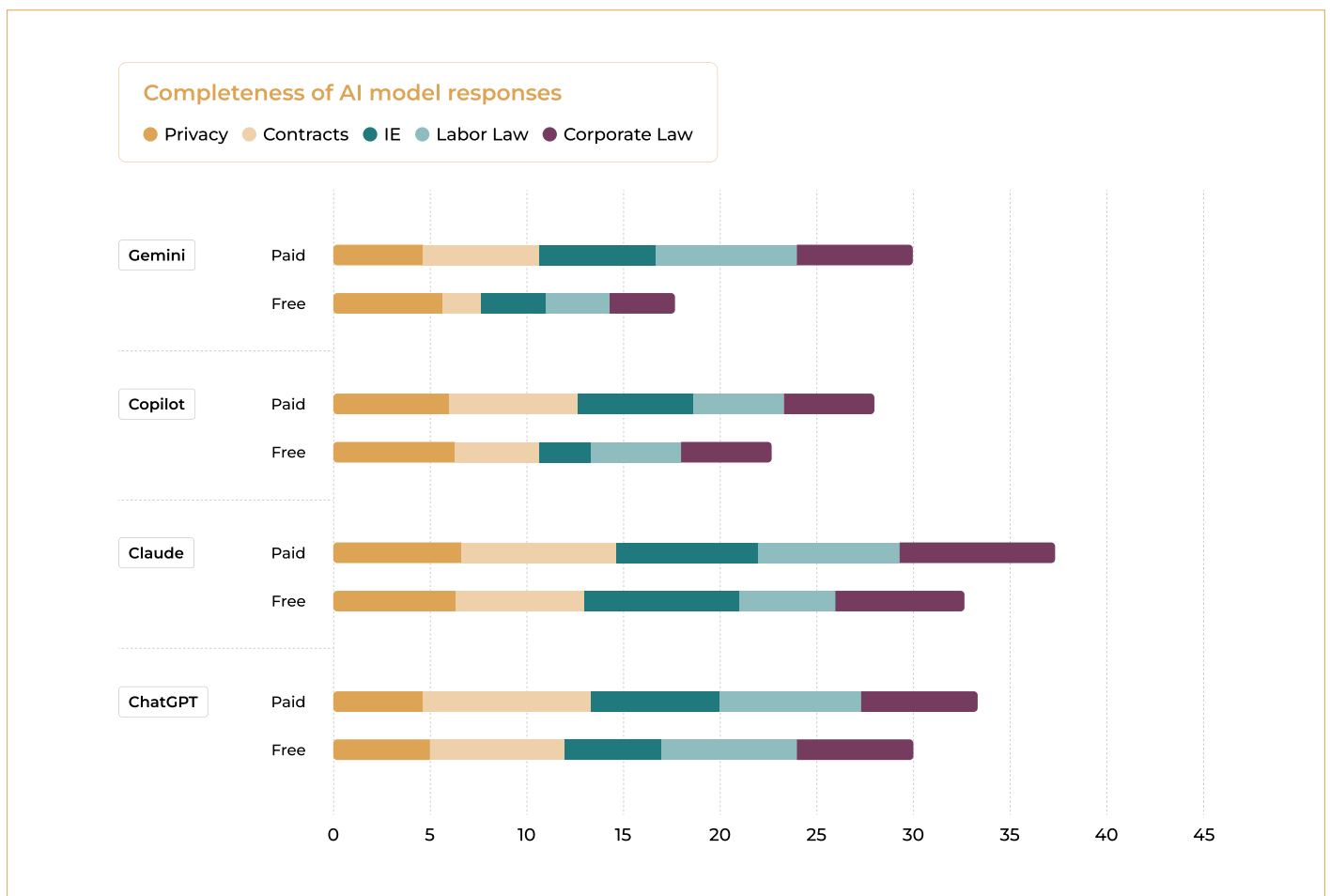
Completeness: depth of answers

Completeness scores reveal a subtle but important difference between free and paid versions. Free systems usually score around 4 to 6, indicating a basic legal analysis without deeper nuance. This often manifests in what we call “almost legal”: answers that sound correct but miss crucial legal subtleties. For example, in contract law, some answers state that “when a contract is terminated, it is considered as ****if it never existed****”. This is incorrect: that effect applies to rescission, not termination. This kind of near-miss confusion of concepts is endemic to these language models.

Paid versions regularly reach scores of around 8, although they rarely achieve the maximum score of 10 for deep legal substantiation. Even then, the systems continue to struggle with legal subtleties. Crucial distinctions, such as between rescission and termination, are still systematically confused, and the tendency to translate legal jargon persists, even when this weakens the

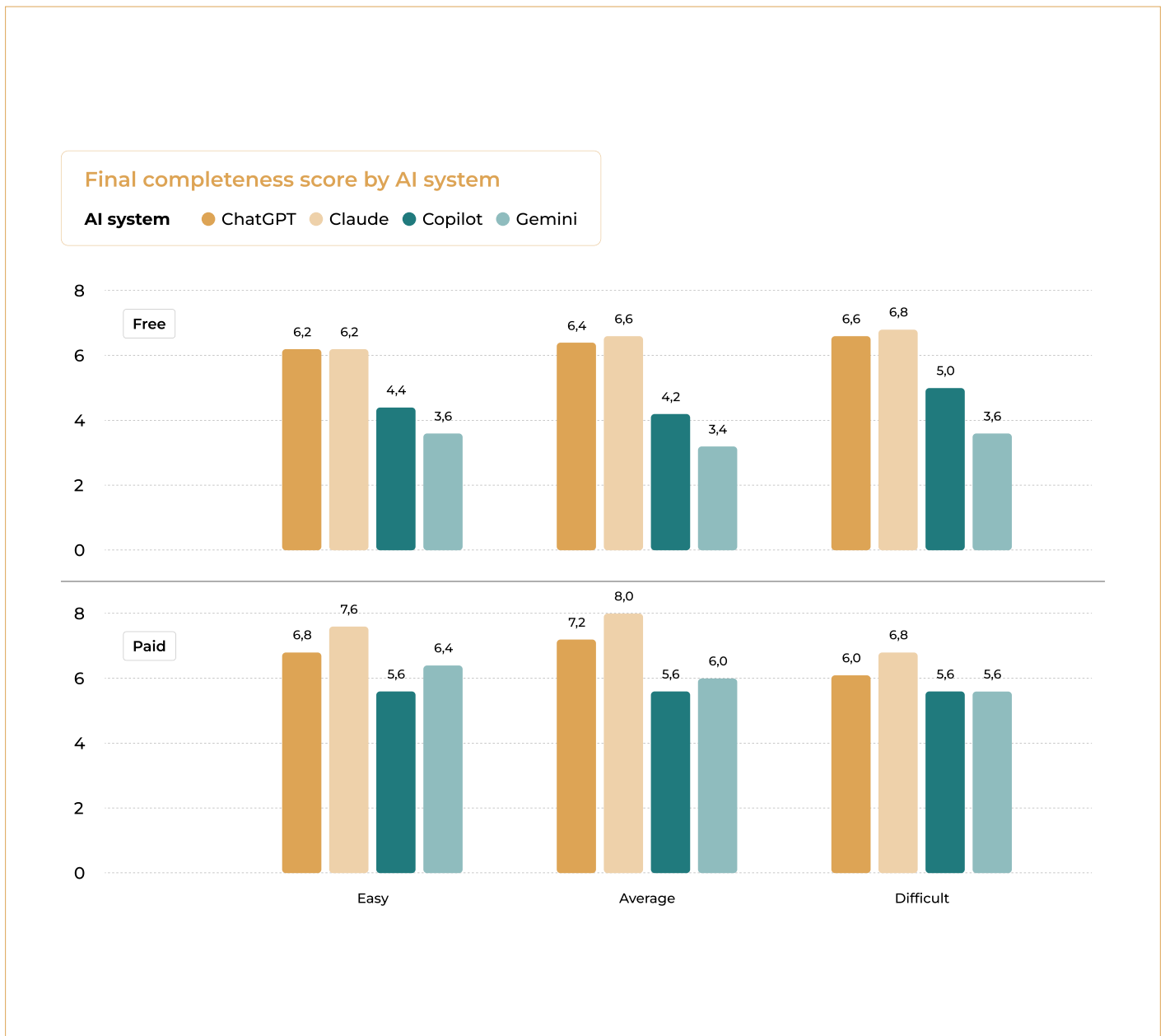
analysis. This suggests that premium versions provide more information, but not necessarily more legal depth.

A recurring phenomenon is what we call the “democratisation of legal advice”: all systems appear programmed to translate legal concepts into lay language, even when the question clearly comes from a legal professional. This leads to unnecessarily long answers in which legal jargon is avoided or explained, reducing effectiveness for professional users. Jargon is, after all, part of the job. We also see, particularly with Copilot and Gemini, general references to “case law” as justification. For professionals, that is of course insufficient.



Conclusion: where AI excels and where it does not

We summarise all findings in the table below. The overall winner is the paid model Claude 3.5 Sonnet, closely followed by ChatGPT 4o Plus. Results vary significantly by legal area.



The systems examined show a peculiar mix of impressive performance and worrying blindness to legal nuance. The popular comparison with a legal intern therefore falls short: what intern would excel at complex IP analyses yet confuse elementary contract principles, or flawlessly explain GDPR compliance while stumbling over basic formal requirements? This “Swiss cheese pattern” of excellence and deficiency is unique to artificial intelligence and requires its own supervisory framework.

The research presents a nuanced picture of AI’s legal capabilities, with scores ranging from almost a perfect ten to a clear zero. It is not the complexity of the questions but the legal field that proves decisive for performance. This suggests that AI systems do not struggle with legal complexity as such, but rather with specific characteristics of legal domains. The added value of paid versions is mainly reflected in:

1. Consistency: fewer extreme low outliers
2. Jurisdiction awareness: better recognition of national differences
3. Source referencing: more accurate use of legislation and case law

The radar chart below visualises this specialisation per AI system and shows how different models can complement each other in legal practice.

Three specific situations

AI systems excel in three specific scenarios. First, in rapidly exploring legal questions, where they can identify relevant statutory provisions and general principles. Second, in legal areas with a strong systematic structure and well-established case law, where high scores are achieved. Third, they perform very well in analysing complex but logically structured legal issues that lend themselves to conceptual breakdown.

At the same time, we see persistent shortcomings that legal professionals cannot ignore. Most concerning is the systematic blindness to crucial legal nuances, such as the distinction between termination and rescission. Performance fluctuates strongly between legal areas: a

system that performs excellently in privacy law may fail dramatically in contract law. The systems also struggle with over-contextualisation, leading to inefficient answers full of irrelevant digressions.

Handling of sources

Particularly worrying is the handling of sources. Where human lawyers always start with legislation and case law, AI systems treat these more as suggestions than as binding sources. They cite non-existent statutory provisions, refer to case law they invent on the spot, or, more insidiously, merge elements from different judgments into a new “norm” that has no basis in law. These systems thus hallucinate entire legal foundations, imitating the rhetorical structure of legal argumentation while fabricating the substance. This is treacherous for legal professionals, whose entire methodology is built on the assumption that cited sources at least exist. It forces a new working method in which every source reference must be checked, not for correctness of interpretation but for the simple fact of whether the source exists at all.

The paradox of trust

As ChatGPT might put it, here we encounter a fascinating paradox that particularly affects the work of the modern in-house lawyer. We expect this professional to advise across the full spectrum of law, from employment disputes to security rights, from privacy breaches to IP issues. While law firms can deploy specialists for each field, the in-house lawyer must handle incidental, one-off questions as a legal all-rounder. This is precisely where AI seems attractive: it can quickly sketch the contours of an unfamiliar legal area or even provide a clear and extensive answer to an in-depth question. But how do you verify the reliability of AI analyses in an area you know little about?

The solution lies in redefining the goal: use AI not to get answers, but to learn which questions to ask. Where a specialist immediately identifies the core issues (“Ah, this is a transfer of undertaking issue”), AI can help the in-house lawyer develop that perspective. The system

can explore relevant contexts, identify key concepts and jargon, and recognise typical problem patterns that help frame the issue. This framing then enables targeted verification: instead of having to check the entire output, the lawyer now knows where to look and which expertise to involve.

This transforms AI from a risky oracle into a valuable orientation tool. It helps in-house lawyers efficiently involve specialist assistance where needed and search official sources more effectively. The art is not to trust AI's analysis blindly, but to use it as a compass in unfamiliar legal territory.

Can you trust AI for legal work?

The central question we posed at the outset was simple but fundamental: can you trust these systems for everyday legal work? The answer is nuanced but clear: yes, provided you use them where they are strong and systematically verify their output. In concrete terms, this means AI can function very well as an initial exploration of a legal question or for structuring legal information, such as drafting a due diligence checklist or categorising contract clauses. The systems are also valuable in legal areas with a clear structure, such as analysing trademark issues or working through GDPR requirements.

However, this use requires strict boundaries. Every source reference must be verified, not on a sample basis but systematically. AI output that claims “according to established case law” should set off alarm bells: where are these judgments? Complex issues that span multiple legal areas or involve very recent legislation are, for now, unsuitable for AI support. AI use must also be embedded in broader workflows with explicit review moments, for example by having a senior lawyer validate the final product.

In practice, this means AI can support activities such as:

1. Drafting an initial version of standard documentation
2. Structuring legal arguments into a systematic overview

3. Translating complex legal concepts into understandable language
4. Conducting an initial exploration of unfamiliar legal areas, provided it is used as a starting point rather than an endpoint

Human expertise remains indispensable for:

1. Verifying cited sources and case law
2. Safeguarding crucial legal nuances
3. Assessing time-sensitive aspects
4. Final legal judgment

This balance between opportunities and safeguards requires a solid foundation of risk management and liability control, which is the subject of the next chapter.

Risk management and liability

Let us immediately address the biggest practical risk: virtually all generative AI systems operate on US servers. This means that every legal question you ask, every document you analyse, travels through the United States. For many lawyers, that feels uncomfortable, and rightly so.

Paid versions, particularly enterprise variants, offer more safeguards than free versions. Where consumer terms primarily protect the provider, business licences include specific provisions on data security, confidentiality and liability. Microsoft's Copilot Enterprise goes furthest in this respect with its EU data boundaries, but other providers also offer enterprise-specific guarantees. We examine this per risk area.

Data protection and privacy

Since 2018, the General Data Protection Regulation (GDPR) has formed the foundation for personal data protection in Europe. Data transfers outside the EU receive particular attention. The starting point is simple: transfers of personal data to non-EU countries require appropriate safeguards. That is also the core of the problem here: every question you submit to an AI system constitutes a transfer of the personal data contained therein within the meaning of the GDPR.

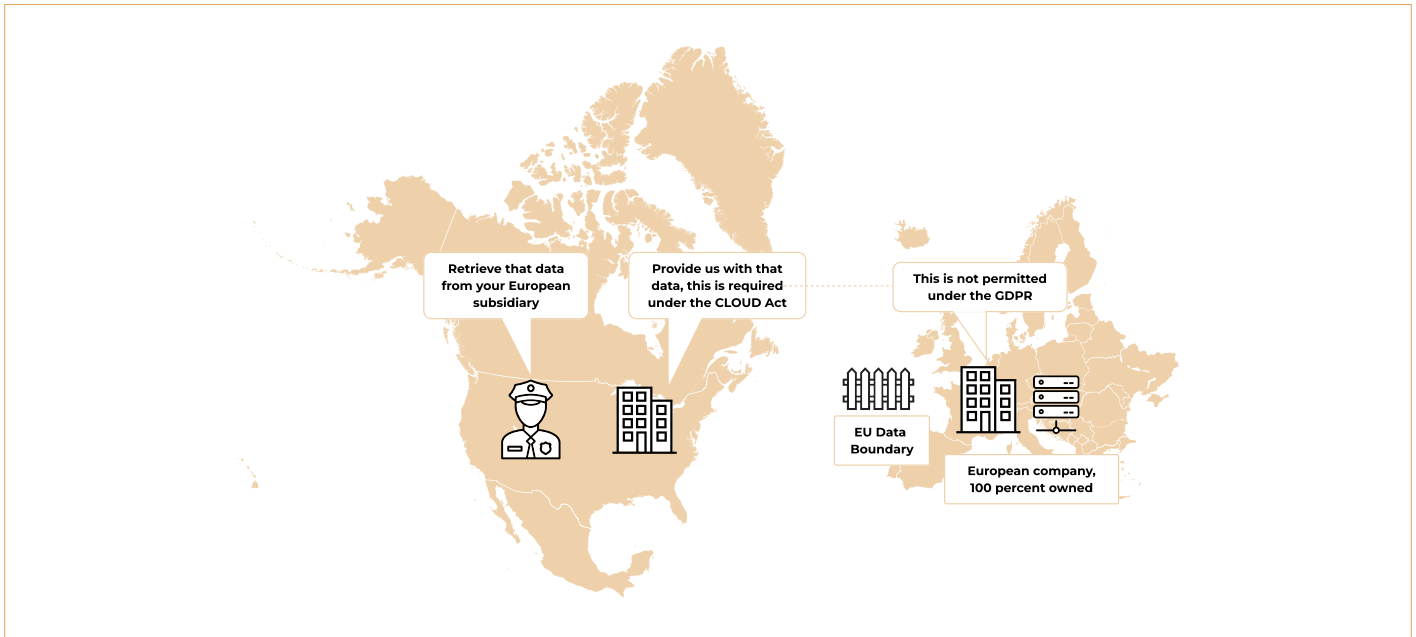
Paper safeguards

Does the United States lack such safeguards? On paper, no. Since March 2022, the Trans-Atlantic Data Privacy Framework (DPF) has existed between the US and the EU. This addresses the fundamental criticisms raised by the Court of Justice in relation to the previous framework, the Privacy Shield, which was invalidated in 2020 in the Schrems II judgment. That framework

itself had replaced the Safe Harbor Framework, which the Court had already struck down in 2015 in Schrems I. Each time, the reasoning was the same: there are many paper guarantees, but we do not know what actually happens in the US.

We see the same issue with standard contractual clauses (SCCs). Organisations turned to these en masse when Privacy Shield fell away. The idea is simple: contractually agree that the US party will comply with European rules. In June 2021, the European Commission introduced new SCCs better aligned with Schrems II. But even these are not a panacea: the Court still requires organisations to investigate whether the US party can actually honour those commitments. You must be able to guarantee that local legislation (read: the CLOUD Act) does not force them to break those promises.

The Cloud Act



The elephant in the data china shop remains the US CLOUD Act. This law obliges US IT service providers, including their EU subsidiaries, in certain cases to share data with US authorities. This conflicts with European law, specifically Article 48 GDPR, when the data is located in the EU. How this would play out in practice is unclear, and stories about gag orders preventing providers from even disclosing that such an order was received hardly improve transparency.

Large US tech companies offer extensive guarantees, often referred to as “data boundaries”. Data is physically stored in EU data centres, combined with organisational safeguards and contractual commitments to prevent access from outside the Union. ISO certifications, privacy audits and transparency reports support these assurances. Microsoft is the clear leader in this area in terms of compliance.

There are also practical dilemmas. Support and incident response often take place from the US, requiring access to data. More importantly, AI models are trained on all input they receive, including customer input. This can often be contractually disabled or switched off in a control panel, but the technical possibility remains, and with it concerns about unwanted access.

Compliance in practice

In a follow-up study, we will examine GDPR compliance of the four LLM providers in more detail. For now, we can already state the following:

- OpenAI (ChatGPT) provides the most GDPR documentation but offers limited control over retention periods
- Microsoft’s integration into a broader ecosystem makes privacy settings complex but offers extensive control
- Google offers the most flexible retention periods but has fragmented documentation
- Anthropic (Claude) is the most restrained in data processing but offers more limited control options

Our analysis shows that Microsoft, with its EU data boundaries and extensive compliance measures, offers the most robust GDPR safeguards. This approach, however, comes with significant costs and requires a complex IT infrastructure, which can be challenging for SMEs. The more accessible alternatives, ChatGPT, Gemini and Claude, offer less far-reaching safeguards,

but operate within the existing legal framework of SCCs and the EU-US Data Privacy Framework, without explicit guarantees that data remains within the EU. Although there is legal debate about the sustainability of these frameworks, and concerns about the CLOUD Act persist, the use of these systems is, in practice, defensible. The reality is that almost all organisations that rely on international services are in the same position.

It would be unrealistic to wait for absolute legal certainty before deploying this technology. If the Court of Justice of the European Union were to rule in the future that the current safeguards are insufficient, this would affect the entire European business community. Until then, we advise the following:

1. Choose an appropriate system

- For large organisations with sensitive data: consider Microsoft's enterprise solutions
- For SMEs and general use: ChatGPT, Gemini or Claude are defensible choices
- Use paid versions only, as these offer the most safeguards and functionality

2. Implement additional safeguards:

- Limit the input of personal and commercially sensitive data
- Document your considerations regarding data minimisation
- Actively use available privacy features, such as memory restrictions
- Develop an exit plan that can be implemented at very short notice if SCCs and/or the DPF are invalidated

3. Monitor developments:

- Keep track of case law and GDPR and AI Act developments
- Be prepared to adjust your use
- Consider membership of relevant industry associations for collective representation

Confidentiality and secrecy

Even if GDPR compliance is resolved, a fundamental tension remains between AI use and confidentiality. Even where all personal data has been removed, sharing information with AI systems may still breach confidentiality obligations. Think of trade secrets, professional secrecy, or information obtained under a non-disclosure agreement. The provider of an AI system is, after all, a "third party" for the purposes of confidentiality rules.

Confidentiality of input

Many stories circulate about AI systems processing confidential input into training data, which then resurfaces in output for other users. Most of these trace back to an incident in early 2023 at Samsung. Employees made enthusiastic use of ChatGPT to improve meeting notes and software source code, until the company intervened and restricted access.

This story quickly morphed into claims that others had actually gained access to this information, but there is no evidence of that. Where other companies discovered that employees had entered confidential information into AI services, there has likewise been no substantiated case of sharing with third parties.

Confidentiality in the terms

The four services address confidentiality in their terms of use (terms of use, terms of service, EULA), but in very different ways. In the free version, OpenAI makes no statements about confidentiality in its terms, but does state that it may reuse both input and output to improve its services and to comply with requests from law enforcement authorities. Google is even more explicit: it literally states that it is prohibited to enter confidential information into the service, which implies that it will treat such information as non-confidential.

In the terms for the paid versions, we see more standard confidentiality clauses. Both the OpenAI terms and those of Microsoft Copilot contain explicit confidentiality provisions that align with an average NDA. It is expressly stated that customer input qualifies as 'Confidential Information'. Comparable provisions appear in Anthropic's terms for Claude. Only Google's terms remain just as anti-confidentiality as in the free version.

Confidentiality and shadow IT

80% of employees use unapproved applications that no one has assessed and that may not be compatible with your security and compliance policies.

(Microsoft, November 2024)

The problem is compounded by the fact that AI services are often deployed as 'shadow IT', employees independently using free AI tools. A lawyer who 'quickly' has a contract reviewed by Google Gemini, or a legal assistant who asks Claude to produce a case analysis, without the organisation having any oversight. These free tools offer no guarantees as to what happens to the information entered. On the contrary, their terms often explicitly state that all input may be reused for model training.

For lawyers, whether self-employed or employed, this situation is particularly sensitive due to their duty of

professional secrecy. This is not merely a contractual obligation, but a legally protected core value of the legal profession. A lawyer who shares confidential client information with an AI system without adequate safeguards risks not only civil claims but also disciplinary sanctions. The Netherlands Bar Association is working on specific rules of conduct for the use of AI, although it is unclear when these will be published.

Copyright concerns

A frequently voiced concern regarding LLM use relates to copyright: do these systems not simply generate variations of existing texts, thereby infringing copyright? The technical reality is more nuanced. LLMs work with statistical patterns and probabilities, not by copying and rearranging existing texts. They predict which word is likely to follow based on patterns in their training data.

Copyright infringement

Given the enormous size of the datasets used, it is unlikely that specific output from one of the four models examined will match an existing source text. No such cases have been reported in the literature or press. The only exception is the New York Times, which in its lawsuit against OpenAI demonstrated that specific newspaper articles could be reproduced when explicitly prompted. This appears to be a very specific case.

The situation therefore differs fundamentally from AI systems for image generation, where generated images often clearly contain elements of specific source images or well-known characters. In software code generation, there are also documented cases where AI systems reproduced entire code blocks verbatim.

In the United States, several lawsuits are pending alleging that LLMs infringe copyright. These primarily concern the use of copyrighted works as training material, not the actual output of the systems. In the EU, this is also under discussion, particularly around the use of opt-out declarations (Article 15o of the Dutch Copyright Act). It seems far-fetched, however, to argue that copyright infringement during model training is relevant for users generating new text with that model.

Protection against infringement claims

In commercial contracts, indemnification is the standard instrument to protect a party against third-party claims, such as copyright infringement. We see this reflected in the terms of AI providers, but essentially only in their business variants. These are typically the versions that also allow automated (API) access and internal sharing of conversations and knowledge within a team.

None of the four services offers indemnification in its free or paid consumer version. OpenAI even reverses the position: the user must indemnify the organisation against all claims arising from use of the service, including copyright infringement. Anthropic fully excludes its liability in these terms, including for infringement of third-party rights.

In business terms, there is more protection. Microsoft and Anthropic explicitly include indemnification against copyright claims in their business terms. Google, in addition to its standard indemnity for business customers, includes an additional provision explicitly confirming this for output from Copilot and use within the business variant. OpenAI also confirms that it indemnifies its Enterprise customers against copyright claims. These indemnities are always limited to output “as is” and subject to compliance with all terms and restrictions, and do not apply where the customer knew or should have known that the prompt would lead to infringement.

Copyright in the output

Another issue is copyright in the output itself. Opinions in the literature are sharply divided on whether output involves copyright-relevant creativity and, if so, to whom that copyright belongs. An LLM has no legal personality. Human involvement lies partly in the choice of prompt (the user) and partly in assembling the model (the provider). Consensus on how this interaction should affect copyright in the output is unlikely in the near term.

Fortunately, OpenAI and Anthropic state in their terms that they assign any copyright in output to the user. Google and Microsoft do the same, although more

cautiously phrased as not claiming ownership of that content. Microsoft explains that it does so in part because output for different customers may resemble each other, potentially leading to copyright issues.

The AI Act: a matter of literacy

As of 1 August 2024, the AI Act has entered into force in phases. Since 2 February, for example, there is an organisation-wide requirement for AI literacy, and certain practices must be discontinued. The bulk of the obligations, for so-called high-risk AI and transparency risks, will only apply from 2026. For the AI systems examined, these requirements are not relevant: providing advice to a legal professional is not considered a high-risk application.

Our research underscores why AI literacy is so important. LLMs display a unique pattern of excellence and failure fundamentally different from human expertise. Using these tools effectively requires a new form of professionalism. This manifests in three dimensions. First, lawyers must learn to recognise patterns in where AI is reliable. Our research shows that systems excel in systematic legal areas and conceptual analyses, but can fail dramatically when source references or current case law are involved. This requires a new kind of judgment: not traditional assessment of legal reasoning, but recognising contexts in which AI is likely to be valuable or risky.

Second, there is the art of systematic verification. Where lawyers are accustomed to checking sources for interpretation, AI forces a fundamentally different approach: verifying the mere existence of sources. This requires new workflows in which verification is embedded not as an incidental check but as a structural component of AI use.

The third dimension concerns strategic deployment. AI proves particularly valuable as an orientation tool that helps frame issues and identify relevant contexts. This makes it especially suitable for in-house lawyers who must engage with the full spectrum of law. But this use requires careful consideration: when is AI support

sufficient, and when is specialist expertise needed?

For legal departments, this creates a new training challenge. The goal should be to foster a learning culture in which experimentation with AI is encouraged and experiences are systematically shared. By jointly analysing AI answers, documenting best practices, and discussing concrete successes and failures, a team develops a feel for the possibilities and limits of this technology.

Crucially, this training should not descend into technical detail but focus on professional judgment. It is not about perfect understanding of the technology, but about the ability to make sound choices about its use. This requires continuous investment across the entire team, not only in technical experts or early adopters. Only in this way can an organisation develop the collective judgment needed to use AI effectively and responsibly in legal work.

Professional liability in the AI era

“As the Supreme Court ruled in *Martinez v. Delta Airlines* (2023)...” except that the case existed only in ChatGPT’s imagination. The judge in this New York case was not amused, the lawyer was publicly reprimanded, and the legal world received a hard lesson in blind trust in AI. Amusing? Perhaps. But also a warning. What if the fabricated case law had not been so easy to verify? What if a lawyer advises based on an AI legal analysis that subtly but critically misses the mark?

The AI paradox in legal work

This situation contrasts sharply with how we treat traditional legal tools. No lawyer would accept “Kluwer Navigator returned no results” as the end of research, or regard “the search engine of a legal journal found nothing” as adequate verification. We expect lawyers to consult multiple sources, cross-check references, and base their analysis on solid review of legislation, case law and authoritative legal commentary.

Why, then, do we accept AI output at face value? Perhaps because it sounds so convincing, so human in formulation, so confident in analysis. But it is precisely this persuasive presentation that makes AI

hallucinations particularly dangerous. Unlike a “no results found” message from a search engine, AI presents its fabrications as established legal truths.

This dynamic is especially risky for in-house lawyers. As legal generalists, they navigate daily between employment law, contracts, privacy, compliance and intellectual property. You must know something about everything, but deep specialisation in every field is unrealistic. An AI system that reduces complex legal material to clear, accessible analyses is therefore tempting, even liberating. But where an IP specialist immediately sees that an analysis of copyright is overly simplistic, or a privacy expert notices missing nuance in a GDPR argument, the generalist may miss these warning signs. The AI system presents itself as the perfect on-demand specialist: always available, never hesitant, with an answer to every question. But unlike a human specialist, it has no awareness of its own limitations. It will never say “this issue is too complex for a quick scan” or “you really need specialist advice here”.

A legal vacuum

At present, there are no Dutch judgments on professional liability or disciplinary law specifically relating to AI use by lawyers or other professionals. This is not reassuring but rather an additional reason for vigilance. It means there are not yet crystallised standards for what constitutes “careful AI use”. Should we treat AI as a research tool, a junior lawyer, or external advice? Each model carries its own duties of care and liability risks.

The New York case offers a preview of how Dutch courts and disciplinary bodies may approach this. The US judge held that blind reliance on AI-generated case law was negligent. The lawyer should have verified whether the cited judgment actually existed. This reasoning fits seamlessly within the Dutch system of professional liability, where legal professionals are held to a high standard of care. But how do we concretely fill in this standard?

Duty of care in the AI era

Our research shows that using AI systems does not reduce the duty of care of legal professionals, but rather increases it. The persuasive tone of AI output,

combined with the real risk of factual or legal errors, means professionals can never rely on such output without scrutiny.

For in-house lawyers, this creates a new core competence: the ability to recognise when an AI answer sounds too smooth, too simple, too convincing. Precisely when the technology seems most helpful, translating complex specialist material into practical guidance, the risk of oversimplification is greatest.

This new competence is no longer optional. From February 2025, the AI Act requires professionals who use AI systems in their work to demonstrably possess

AI literacy. For lawyers, this means more than knowing how to write a prompt. It involves a fundamental understanding of AI systems: their capabilities and limitations, the ability to critically evaluate results, and the skill to use this technology responsibly. Where a disclaimer about AI use once sufficed, active management of AI risks is now expected.

One thing is certain: “the AI said so” is not a defence in cases of professional error. On the contrary, undocumented or unchecked AI use could be seen as additional negligence. It is like a lawyer blindly trusting the work of an intern without checking a single source.



Outlook & recommendations

“Just run it through ChatGPT quickly”, it has become a common phrase in legal departments. Where two years ago we were still sceptical about AI in legal work, we now use it daily. Our research shows that this is neither good news nor bad news; it is simply reality. Just as we once had to get used to digital case law databases and automated document generation, we are now finding our way with AI. But there is an important difference. Earlier innovations mainly made existing work faster and more efficient. AI forces us to rethink fundamentally how we do legal work. An AI system that confidently cites a non-existent judgment presents a very different risk from a search engine that reports no results.

The balance between quality and compliance

Our research reveals a fascinating tension. The quality of AI legal analyses is often surprisingly good, especially in paid versions, but at the same time fundamentally unreliable. It is like a brilliant intern who can deliver a watertight legal analysis in one breath and cite entirely fabricated case law in the next, only to then fail to answer a basic question. How should a professional work with that?

This quality paradox is compounded by compliance challenges. While quality issues can be mitigated by human review, compliance risks are often binary: you either share confidential information with a US company or you do not use AI at all. Paper safeguards such as SCCs and the EU-US Data Privacy Framework provide a legal framework but no guarantees.

One general conclusion is that anyone seeking to enhance legal work with large language models should use a paid subscription. Given the poor results and

extensive data access in the terms, we would almost consider it a professional error to use free versions on real cases or client advice.

Performance varies widely by legal area. We strongly suspect there is a European law explanation for this: there is more harmonised source material for areas such as privacy and intellectual property than for specialised national law such as employment or corporate law. Even in the strongest-performing models, the tendency to invent legal sources remains.

The key question for organisations is not whether they will use AI, that is already happening, but how to embed it responsibly. The art is to reconcile technological possibilities, legal requirements and practical reality. Sometimes this means choosing more expensive enterprise solutions with better safeguards, sometimes imposing strict limits on AI use in sensitive matters.

All examined models are American and general-purpose. Are there no Dutch or European models specifically suited to legal practice? Unfortunately, not yet. So far, we mainly see marketing promises or tools positioned firmly at the high end of the market. This is logical: developing fundamentally new, truly European AI systems costs billions and requires enormous data collections and human effort. The legal market is also slow to change, as years of struggle with legal tech have shown. If even a basic document management system sparks debate for years, an AI tool requiring complex implementation will not be easier.

Precisely for this reason, the examined models have an important practical advantage: they are immediately usable, without involvement of IT departments or expensive consultants. You ask a question and get an

answer. That answer may not be perfect, as our research shows, but it is often good enough and provides a solid starting point for legal analysis. And that is exactly what most lawyers need: not yet another complex system to be “implemented”, but practical support in daily work.

Concrete recommendations for legal departments

What does this mean for legal departments that must determine a strategy now? The answer does not lie in finding the perfect tool, because it does not exist. It lies in developing a working method that fits today's reality, one that embraces technology where useful but does not lose sight of the legal core of the work.

This balance will differ for each organisation. A law firm focused on litigation has different needs from an in-house lawyer reviewing dozens of contracts daily. A privacy officer overseeing international data flows has different concerns from a lawyer focused on Dutch employment law. The question is not what is theoretically best, but what works best in practice.

This requires strict frameworks. Every source reference must be verified, not randomly but systematically. Questions must align with AI's strengths, such as drafting initial versions of standard documents or structuring arguments. AI is also well suited for an initial exploration of unfamiliar legal areas. Complex issues spanning multiple legal areas or involving very recent legislation are less suitable. AI use must also be embedded in broader workflows with explicit review moments, such as senior validation of final output.

Our research nevertheless reveals some universal principles. First, the quality of AI answers depends entirely on the quality of human oversight. A well-designed verification process can yield usable results even with free tools. Conversely, an expensive enterprise solution without proper governance can still lead to legal errors. Second, the biggest risks lie not in what AI does wrong, but in what lawyers do with it. It is like a car with driver assistance: safety depends not on the system, but on how the driver uses it.

Banning AI altogether? Abandon that ill-fated idea. It drives use underground and creates a dangerous form of shadow IT, where employees use tools independently without safeguards. Let us be honest: if an in-house lawyer needs to make an initial assessment of an 80-page contract at 4.30 pm, the temptation to “just run it through ChatGPT” is strong. And if a legal assistant needs to determine whether a situation triggers a data breach notification obligation, Claude is much faster than manually combing through GDPR guidelines. Prohibiting these tools is not only unrealistic, but also counterproductive. It prevents the development of responsible practices and drives a wedge between management and staff. It is far better to regulate, monitor and adjust use where necessary. Better an imperfect but transparent process than a perfect policy that no one follows.

Concrete recommendations for the individual lawyer

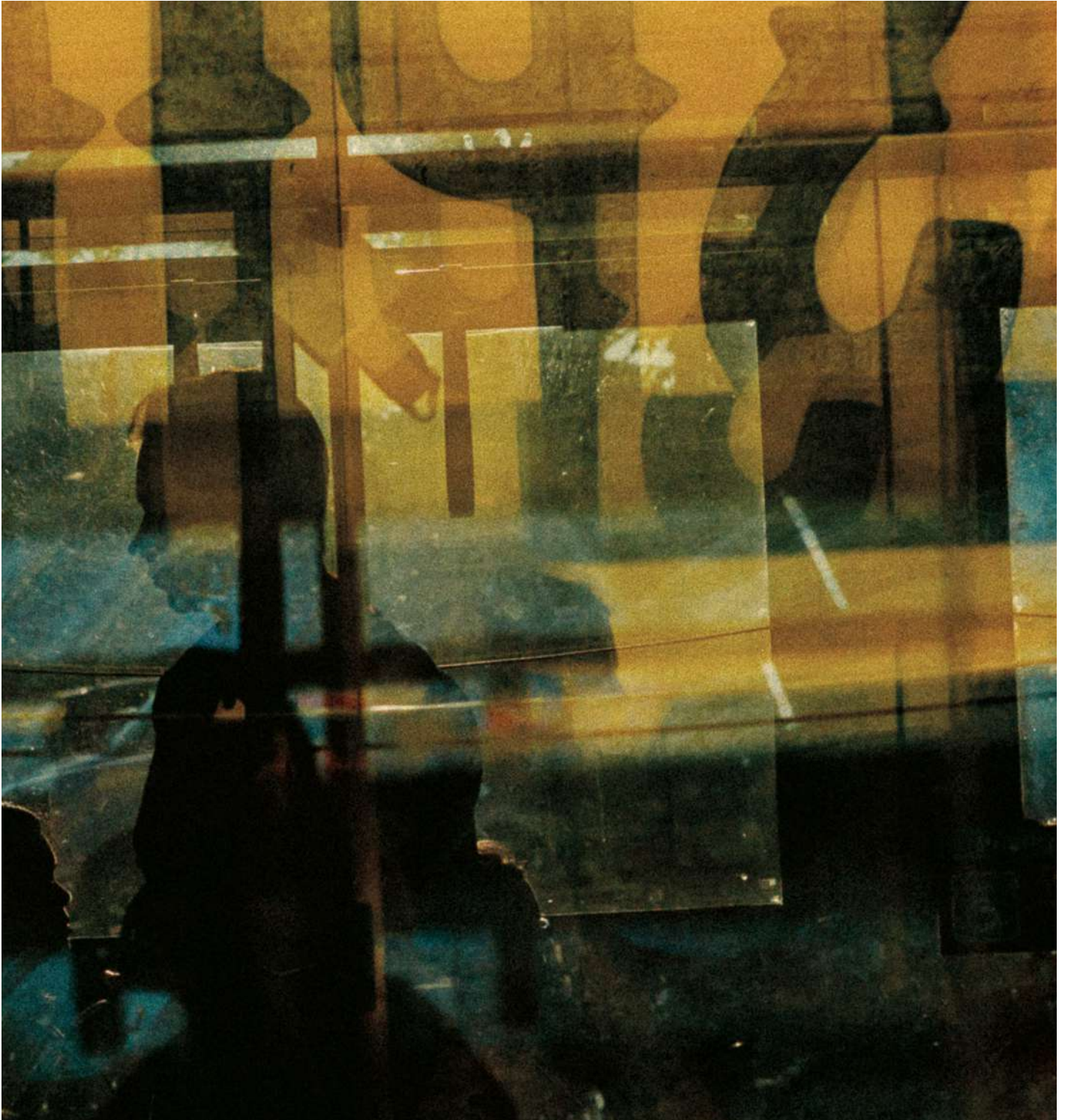
First, acknowledge that you are probably already using AI, or will soon. Not because you have to or because it is fashionable, but because it is practical. A first scan of a contract, a quick check of case law, a draft for advice, AI tools are like a satnav for legal work. They point the way, but you remain at the wheel. And just as you always check whether a suggested route makes sense, you must always critically assess AI output.

Start small but purposefully. Choose one specific task for which you want to use AI, such as creating a first draft of standard documents. Experiment with it, learn the pitfalls, develop a feel for when you can trust the output and when not. It is like learning to drive: start in a quiet neighbourhood, only later on the motorway.

Develop what we call AI literacy. It sounds lofty, but it simply means learning to recognise patterns. When does an answer sound too smooth? Which legal nuances are often missed? Where are the blind spots by legal area? Our research shows that AI systems are consistently strong in some tasks, such as structuring legal information, and weak in others, such as understanding new case law. Know those patterns.

Document your experiences and share them with colleagues. Which prompts work well? Which pitfalls have you encountered? Where did things go wrong? This knowledge is invaluable, not only for yourself but for your entire team. It prevents everyone from having to make the same mistakes. It also contributes to developing best practices that benefit the entire legal community.

And perhaps most importantly, continue investing in your core expertise. AI is a tool, not a replacement for legal judgment. In fact, the better you are at legal analysis yourself, the better you can assess when you can trust AI. It is like a modern diagnostic instrument in medicine: a valuable tool that provides data and analyses, but it is the doctor who, using expertise, makes the correct



AI, the Digital Decade and consultancy ICTRecht

The conclusions of our research into AI systems reflect a broader truth about digital transformation: technology and law are inextricably linked. Where AI systems struggle with legal nuance, we see the same pattern in the implementation of new digital technologies. It is not about the technology itself, but about the ability to deploy it responsibly and on a sound legal basis.

This perspective is not accidental. For twenty years, we have seen digital innovation and legal frameworks shape and reinforce each other. Where in 2004 we were mainly dealing with hosting and online shops, we now face challenges such as DORA in the financial sector, the European Accessibility Act for digital inclusion, and the European Digital Identity Framework laying the foundation for digital identities in Europe.

The Digital Decade brings a fundamentally new dynamic to the legal landscape. Where we once looked to a single statute for solutions, we now face a complex web of interconnected regulations. A modern connected product often must comply with five or six different laws simultaneously, which complement and reference each other.

Take a modern car with driver assistance, for example. It must comply not only with traditional type approval, but also with:

- The AI Act for driver assistance
- The Cyber Resilience Act for digital security
- The GDPR and Data Act for sensor and passenger data
- The Right to Repair Directive for software updates
- The Product Liability Directive for liability

This interconnection requires a new approach to legal services. Not isolated advice per statute, but an integrated framework encompassing all relevant

regulation. Technical standards increasingly serve as a practical compass, but must always be placed within the broader legal context.

The art lies in looking beyond technological hype and focusing on underlying legal issues that remain relevant. With the metaverse, the ultimate issues were familiar themes such as digital identity, property rights and liability in online environments. That expertise remains valuable, even as technology takes new forms.

Ultimately, it comes down to one fundamental insight: digital transformation cannot succeed without a solid legal foundation. A strong legal basis gives organisations the confidence to innovate and reap the benefits of digitalisation. Whether deploying AI systems, developing connected products, or digitising services, law provides the foundation for responsible innovation. This conviction has driven us for two decades and is more relevant than ever in the AI era. Only by balancing technology and law can we create a digital future that works for everyone.

Ready to get started with AI? Contact us at :
contact@ictrecht.nl.

Our AI advisers are ready to assist you

Business
Development
Manager
CAICO®



**Guido
Grevink**

g.grevink@ictrecht.nl

Legal
Counsel
Tech
CAICO®



**Abel
Hoogeveen**

a.hoogeveen@ictrecht.nl

Chief
Knowledge
Officer
CAICO®



**Arnoud
Engelfriet**

a.engelfriet@ictrecht.nl

Senior Legal
Counsel
CAICO®



**Marco
Caspers**

m.caspers@ictrecht.nl

Senior Legal
Counsel
CAICO®



**Pelçim
Kaygusuz**

p.kaygusuz@ictrecht.nl

Senior Legal
Counsel
Tech
CAICO®



**Niels
Winters**

n.winters@ictrecht.nl

Talent
Manager
CAICO®



**Ardine
Siepman**

a.siepman@ictrecht.nl

Head of Legal
Technology
CAICO®



**Mark
Zijlstra**

m.zijlstra@ictrecht.nl

Legal
Counsel
CAICO®



**Rick
Kaspers**

r.kaspers@ictrecht.nl

Business
Development
Manager
CAICO®



**Aakriti
Bhatia**

a.bhatia@ictrecht.nl

Legal
Counsel
CAICO®



**Sanne
Haumersen**

s.haumersen@ictrecht.nl

Business
Development
Manager
CAICO®



**Machiel
Takens**

m.takens@ictrecht.nl