

# WEBINAR

## The impact of artificial intelligence on European lawyers' practices

18 NOVEMBER 2024, 10:00 - 13:00 CET

#TRAVAR *Training of Lawyers in various areas of EU Law*



18 November 2024  
10:00 – 13:00 CET  
Online on Zoom



Co-funded by the European Union

# WEBINAR

## The impact of artificial intelligence on European lawyers' practices

18 NOVEMBER 2024, 10:00 - 13:00 CET

#TRAVAR *Training of Lawyers in various areas of EU Law*



Laura Jugel

**The EU AI Act**

18 November 2024, Zoom



Co-funded by the European Union



# Webinar on the AI Act

**Laura Jugel**

Legal and Policy Officer

European AI Office

# Introducing the EU AI Act



**Product safety  
regulation**



**Complementary with  
other EU law**



**Focus on AI systems**



**Innovation-friendly**

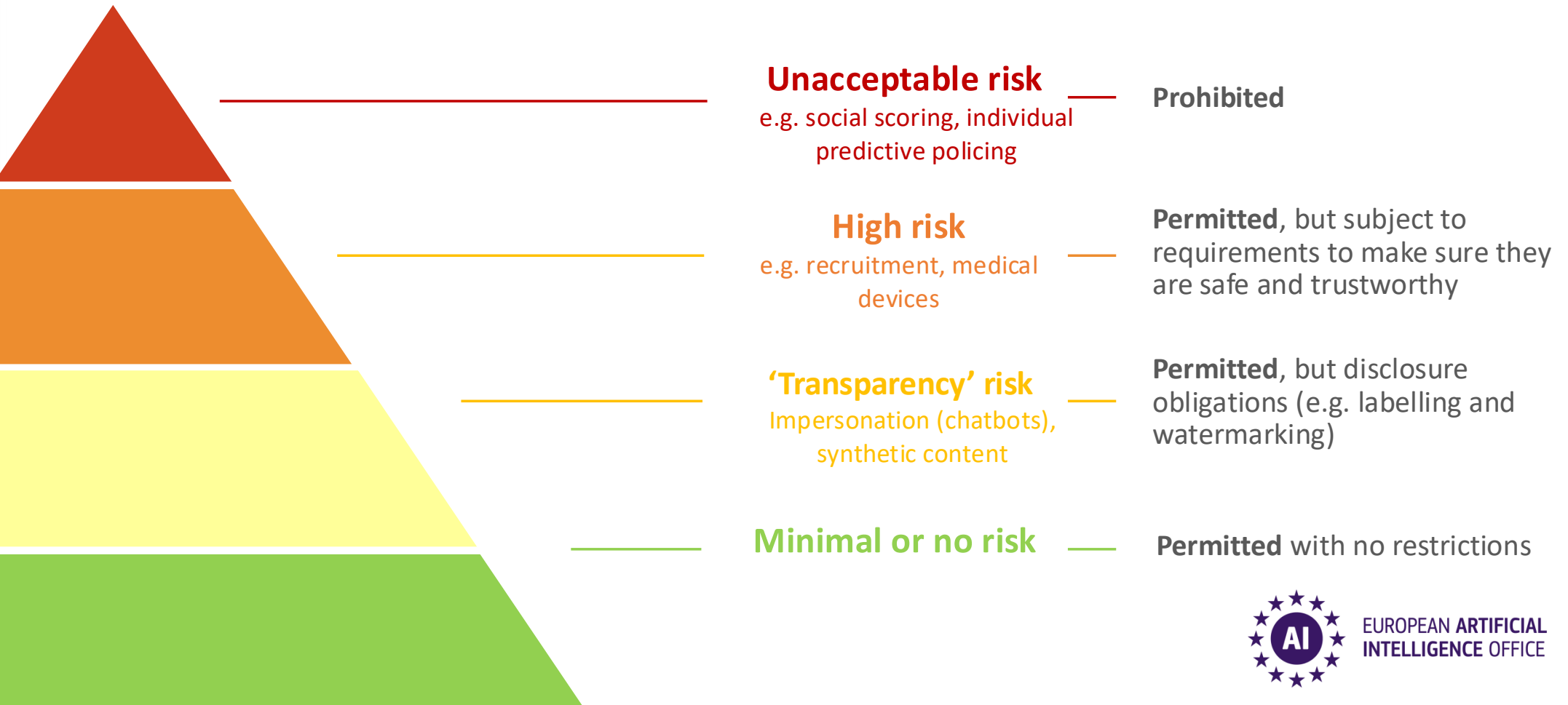


**Risk-based approach**



**EUROPEAN ARTIFICIAL  
INTELLIGENCE OFFICE**

# A risk-based approach for rules on AI systems



EUROPEAN ARTIFICIAL  
INTELLIGENCE OFFICE

# Unacceptable AI practices will be banned



<b>Manipulation or exploitation of vulnerabilities</b>	to manipulate people and thereby cause significant harms
<b>Social Scoring</b>	for public and private purposes
<b>Biometric categorisation</b>	to deduce or infer for example race, political opinions, religious or philosophical beliefs or sexual orientation, exceptions for labelling in the area of law enforcement
<b>Real-time remote biometric identification</b>	for the purpose of law enforcement, with narrow exceptions and with prior authorisation by a judicial or independent administrative authority
<b>Individual predictive policing</b>	assessing or predicting the risks of a natural person to commit a criminal offence based solely on profiling without objective facts
<b>Emotion recognition</b>	in the workplace and education institutions, unless for medical or safety reasons
<b>Untargeted scraping of the internet</b>	or CCTV for facial images to build-up or expand databases



# When is an AI system 'high-risk' under the AI Act?



The AI Act classifies AI systems as 'high-risk' in two ways:

**1** AI system is embedded into a regulated product or is itself a regulated product

Concerns 22 product regulations (Annex I).  
Examples: *Machinery Regulation, Radio Equipment Directive, Toy Safety Regulation*

Two conditions:

- AI system is intended as a **safety component** of a product or **is itself a product**
- Product in question is **subject to a third-party conformity assessment**

**2** AI system is intended to be used in a high-risk use case

8 areas which are sensitive for health, safety and fundamental rights (Annex III) with concrete use cases listed for each area.

AI system classifies as high-risk if it is **intended to be used for one of these use cases.**



„**Filter**“: AI systems can be excluded from the high-risk use cases in four cases, e.g. if they perform only a narrow procedural task.



EUROPEAN ARTIFICIAL  
INTELLIGENCE OFFICE

# High-risk AI systems in the legal sector?



## Specific high-risk use case in Annex III (8) point a):

- AI systems intended to be used by a judicial authority or on their behalf to assist a judicial authority in researching and interpreting facts and the law and in applying the law to a concrete set of facts, or to be used in a similar way in alternative dispute resolution
- Clarifications in recital 61:
  - *AI systems intended to be used by alternative dispute resolution bodies for those purposes should also be considered to be high-risk when the outcomes of the alternative dispute resolution proceedings produce legal effects for the parties.*
  - *The classification of AI systems as high-risk should not, however, extend to AI systems intended for purely ancillary administrative activities that do not affect the actual administration of justice in individual cases, such as anonymisation or pseudonymisation of judicial decisions, documents or data, communication between personnel, administrative tasks.*



EUROPEAN ARTIFICIAL  
INTELLIGENCE OFFICE

# What are high-risk requirements and obligations?



## Providers



**Requirements for the AI system**, e.g. data governance, human oversight, accuracy & robustness, operationalised through **harmonised standards**



**Conformity assessment** before placing the system on the market and **post-market monitoring**



**Quality and risk management** to minimize the risk for deployers and affected persons



**Registration** in the **EU database**

## Deployers



**Correct deployment**, training of employees, use of **representative data** and **keeping of logs**



Possible **information obligations** vis-a-vis affected persons



Possible **fundamental rights impact assessment** (applies only to some deployers, incl. public sector)



Public sector also has to **register the deployment** of high-risk AI in EU database



EUROPEAN ARTIFICIAL  
INTELLIGENCE OFFICE

# Addressing 'transparency' risks



Trust through  
disclosure

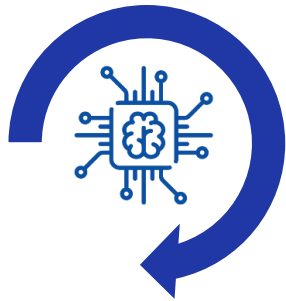
## When interacting with an AI:

- Humans have to be informed if they interact with an AI and this is not obvious
- Deployers have to inform humans if decisions are made about them involving the use of an AI system that is high-risk according to Annex III, e.g. in recruitment

## AI-generated content:

- AI systems that generate output need to include machine readable marks
- Labelling of audio and video content that constitutes a deep fake
- Labelling of text that is intended to inform the public on matters of public interest

# Transparency and risk management for powerful AI models



## General-purpose AI models

= highly capable AI models used at the basis of AI systems such as ChatGPT

Transparency for all  
general-purpose AI models



Risk management for those with  
systemic risk



Code of practice developed together with stakeholders will detail out rules

# A robust governance structure

Rules for AI systems

**National level:**  
EU Member States to  
designate supervisors



**AI Board**

with EU Member States to  
coordinate at EU level



**Scientific Panel**

supports with independent  
technical advice



**Advisory Forum**

supports with stakeholder input

Rules for general-purpose  
AI models

**EU level:**  
AI Office within Commission



EUROPEAN ARTIFICIAL  
INTELLIGENCE OFFICE

# Next steps?



## Priority workstreams of the Commission:

- ▶ **Building up the governance:** AI Office, AI Act advisory bodies, support and guidance to EU Member States
- ▶ **Preparing guidelines, implementing and delegated acts:** with priority on the guidelines on prohibitions and AI system definition
- ▶ **Contributing to preparation of standards for high-risk requirements:** under development by CEN and CENELEC
- ▶ **Coordinating the development of a Code of practice on general-purpose AI:** ongoing process with ca. 1000 stakeholders

# WEBINAR

## The impact of artificial intelligence on European lawyers' practices

18 NOVEMBER 2024, 10:00 - 13:00 CET

#TRAVAR *Training of Lawyers in various areas of EU Law*



Labhaoise Ní Fhaoláin

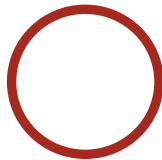
**The Impact of AI Act and other EU AI initiatives on lawyers' practices**

18 November 2024, Zoom

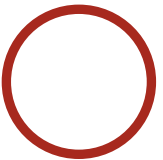


Co-funded by the European Union

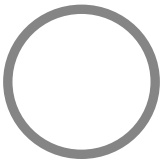
# Topics



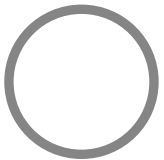
**AI Literacy**



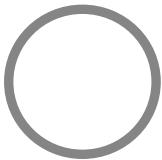
**Unacceptable Risk  
Use Cases**



**High Risk  
use cases**



**Transparency  
Obligations**



**Product Liability  
Directives**

# AI Literacy (Art. 4)



Standalone provision

Applies to all deployers of AI systems (regardless of risk)

In effect **2 February 2025**

# AI Literacy



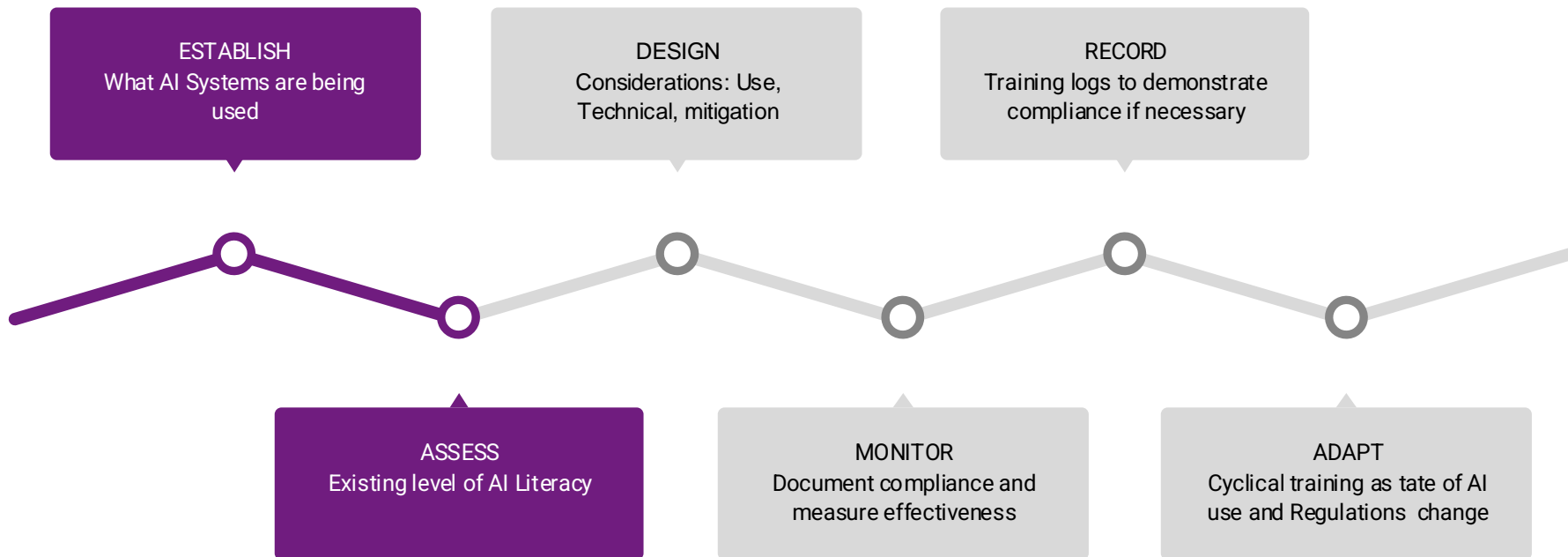
## Article 4

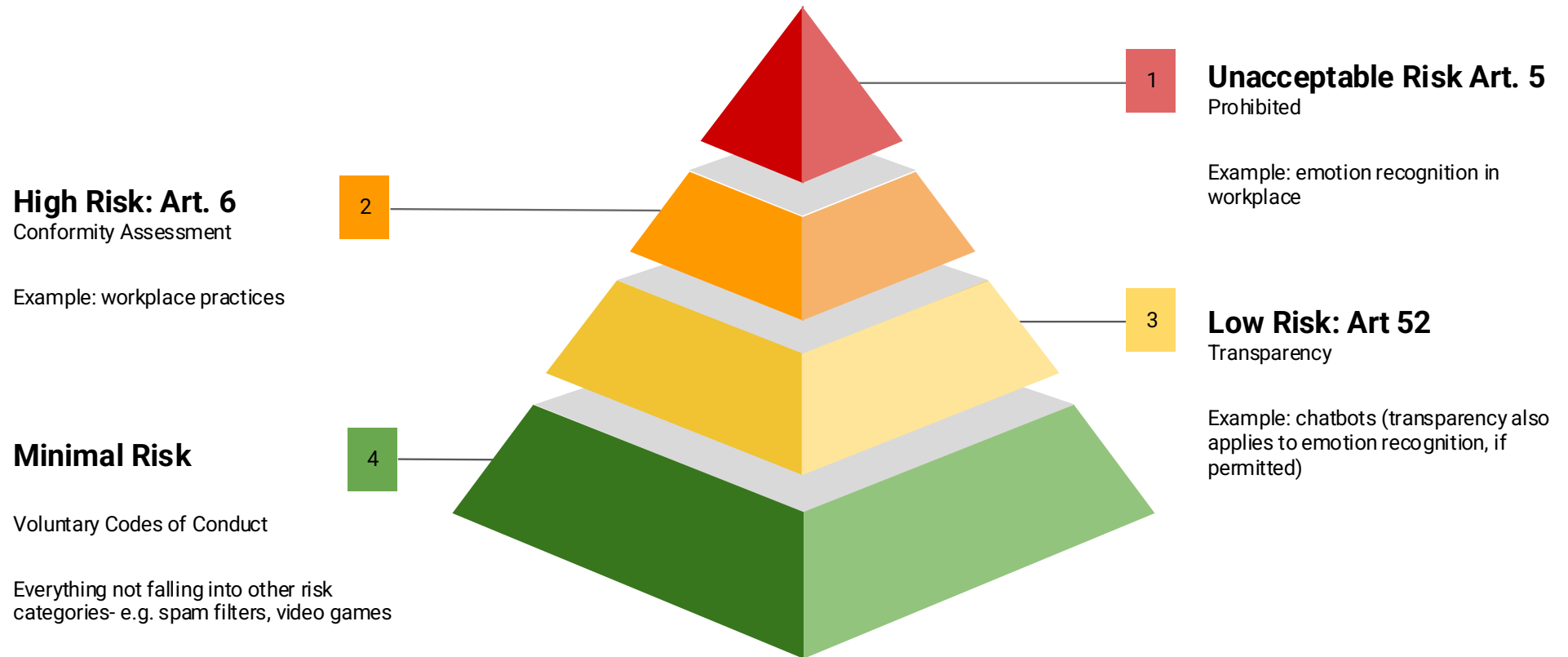
“Providers and deployers of AI systems shall take measures to ensure, to their best extent, a **sufficient level of AI literacy of their staff** and other persons dealing with the operation and use of AI systems on their behalf, taking into account their **technical knowledge, experience, education and training** and the **context** the AI systems are to be used in, and considering the persons or groups of **persons on whom the AI systems are to be used.**”

## Recital 20

- the measures to be applied during its use,
- the suitable ways in which to interpret the AI system’s output, and,
- in the case of affected persons, the knowledge necessary to understand how decisions taken with the assistance of AI will have an impact on them.

# AI Literacy - Compliance





# ■ Unacceptable Risk: Article 5

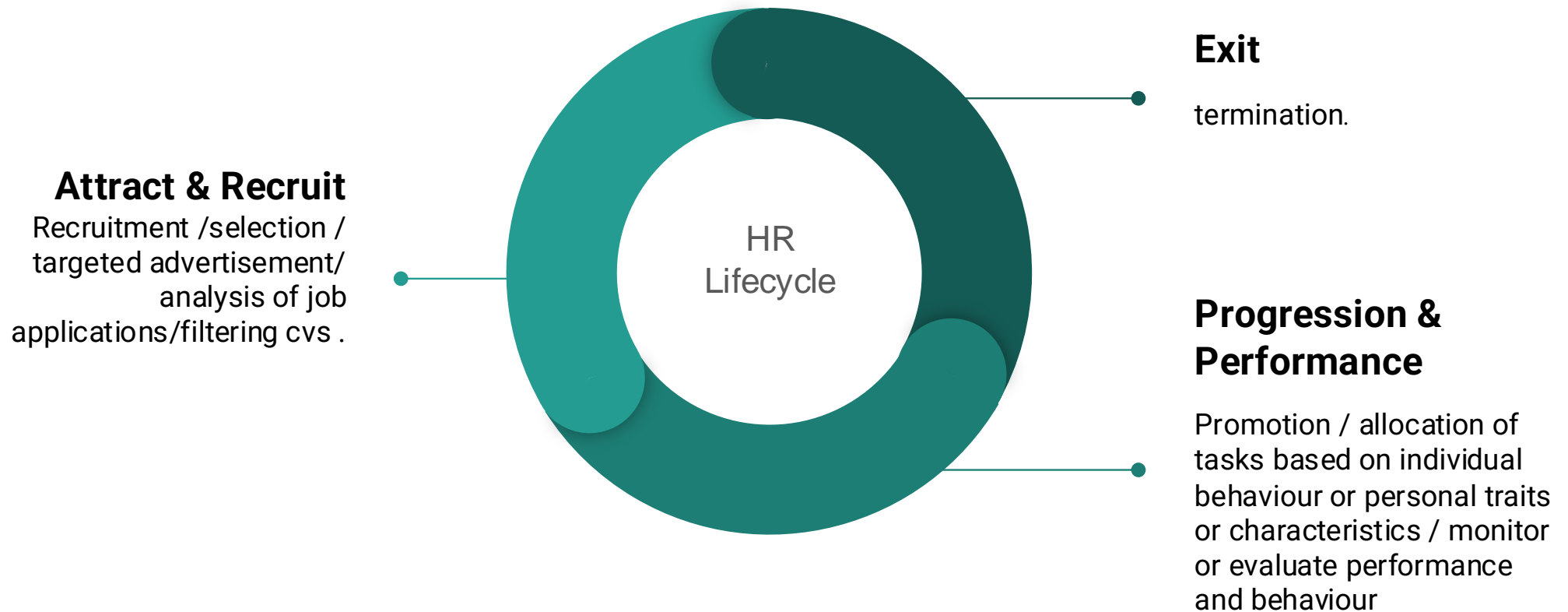


## ■ Unacceptable Risk: Article 5



1(f) the placing on the market, the putting into service for this specific purpose, or the use of AI systems **to infer emotions of a natural person in the areas of workplace** and education institutions, except where the use of the AI system is intended to be put in place or into the market for medical or safety reasons;

# Employment, Workers management and access to self-employment



# Employment, Workers management and access to self-employment

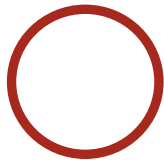


## Annex III: High-Risk AI Systems -Referred to in Article 6(2)

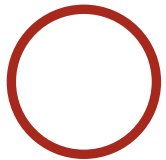
### (4) Employment, workers management and access to self-employment:

- (a) AI systems intended to be used for the recruitment or selection of natural persons, in particular to place targeted job advertisements, to analyse and filter job applications, and to evaluate candidates;
- (b) AI systems intended to be used to make decisions affecting terms of work-related relationships, the promotion or termination of work-related contractual relationships, to allocate tasks based on individual behaviour or personal traits or characteristics or to monitor and evaluate the performance and behaviour of persons in such relationships.

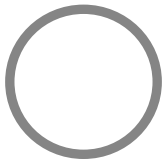
# Topics



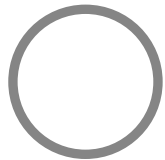
**AI Literacy**



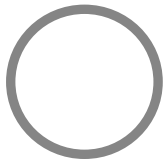
**Unacceptable Risk  
Use Cases**



**High Risk  
use cases**



**Transparency  
Obligations**



**Product Liability  
Directives**

# WEBINAR

## The impact of artificial intelligence on European lawyers' practices

18 NOVEMBER 2024, 10:00 - 13:00 CET

#TRAVAR *Training of Lawyers in various areas of EU Law*



Thierry Wickers  
**Ethical use of AI by lawyers**

18 November 2024, Zoom

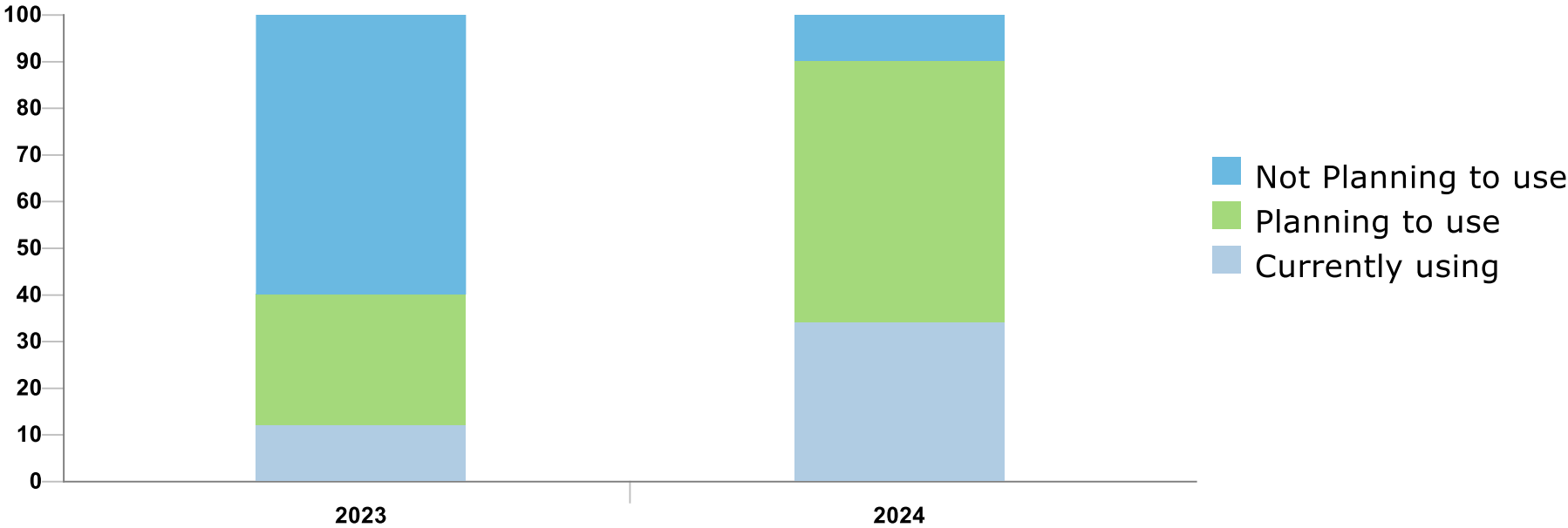


Co-funded by the European Union

# THE IA BREAKTHROUGH



## THE IA BREAKTHROUGH

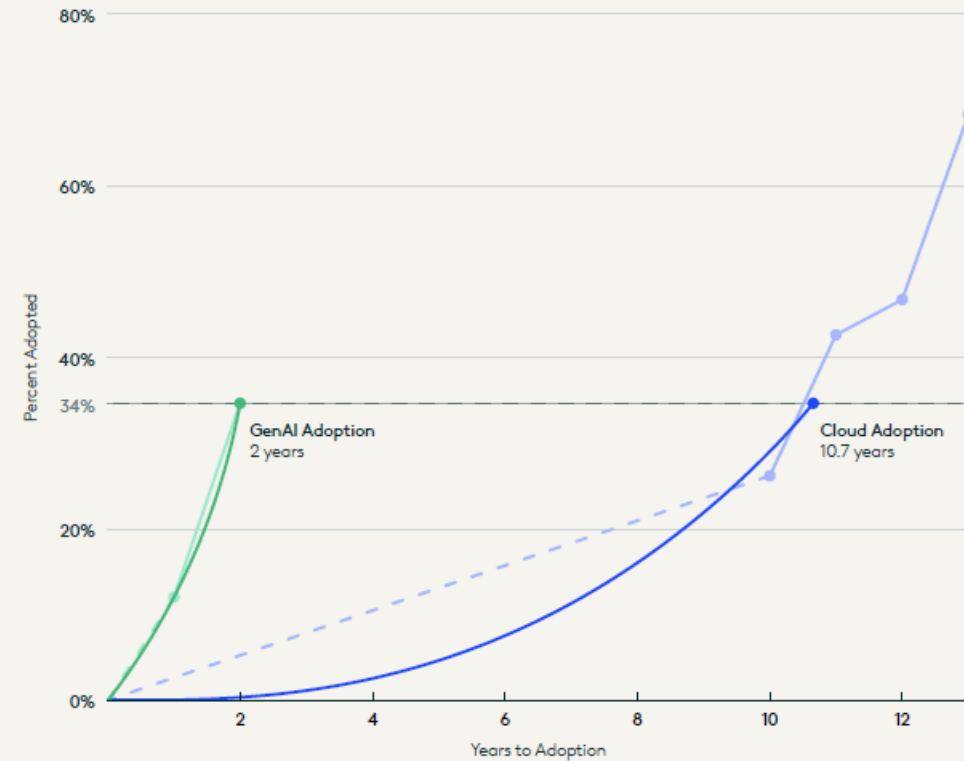


	2023	2024
Currently using	12	34
Planning to use	28	56
Not Planning to use	60	10

Source : Clio Legal Trends report 2024

## Legal Profession's Adoption of GenAI vs. Cloud-Based Ediscovery

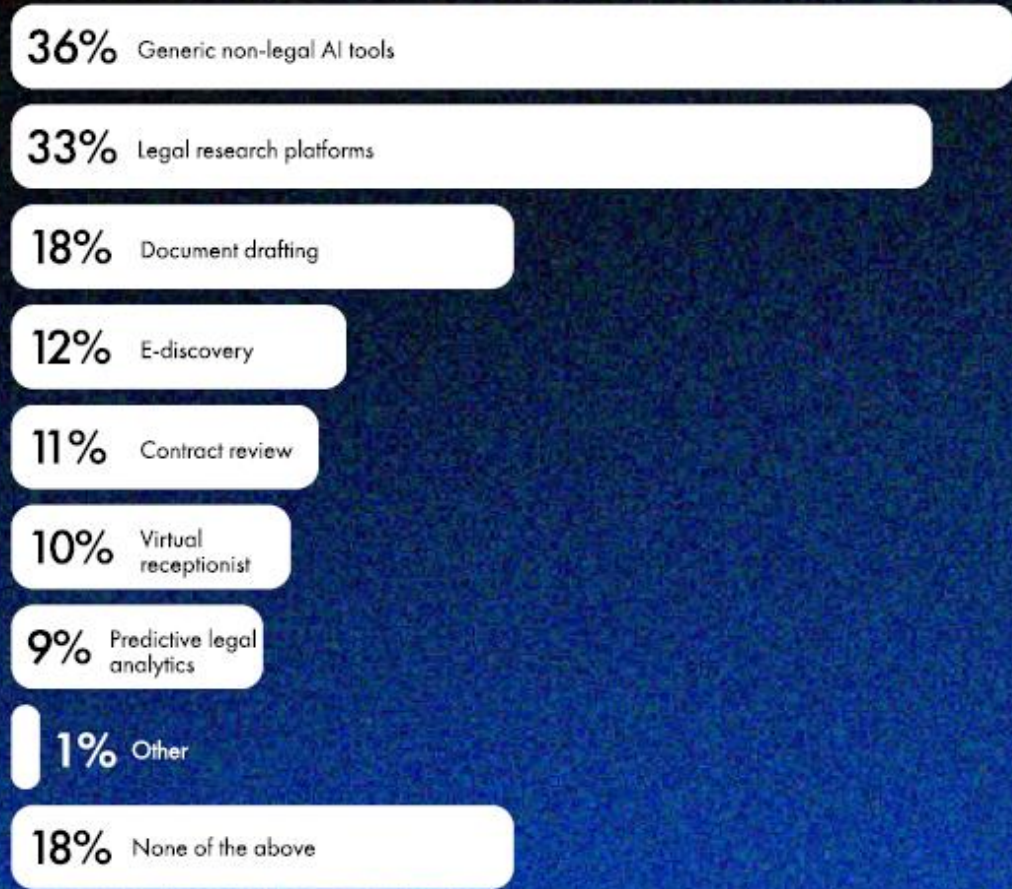
Fig. 8



*Note: The stylized curves represent both cloud and generative AI adoption rates, with the cloud adoption curve being extrapolated back to Everlaw's founding in 2010, and data on industry cloud adoption rates beginning in 2021. For the generative AI adoption curve, that data is extrapolated back to 2022, when this technology first started being leveraged by attorneys in a serious way, and data on adoption rates beginning in 2023.*

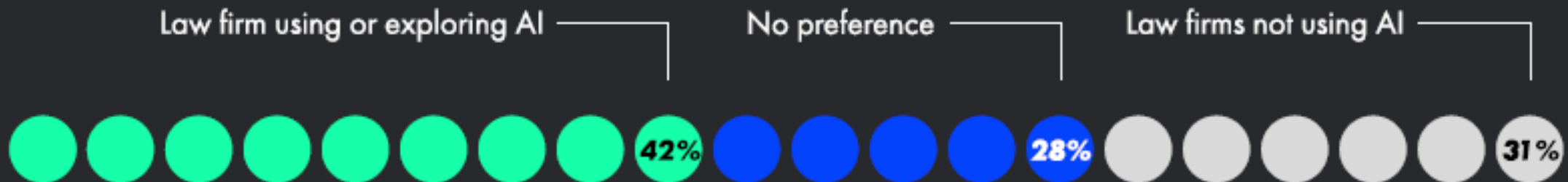
# TOOLS USED BY LAWYERS

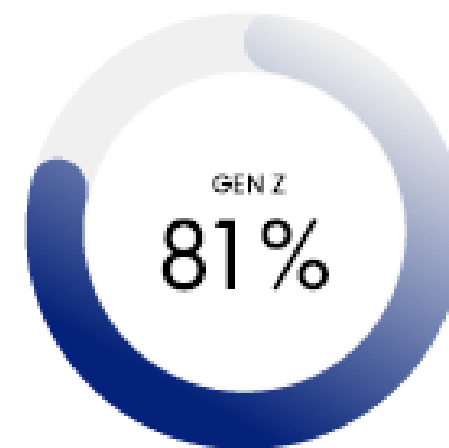
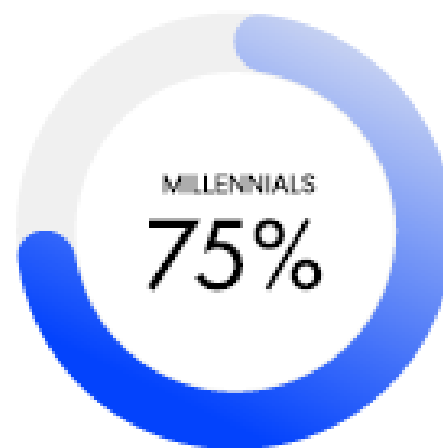
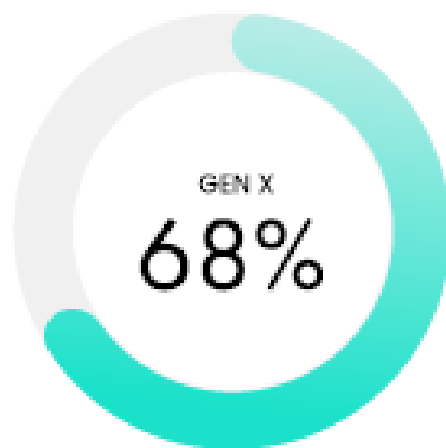
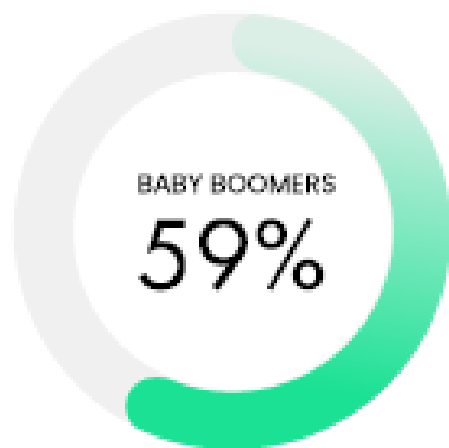
## Top AI-powered solutions in the legal industry



# If you were looking to hire a lawyer, which of the following would you prefer?

## Client preferences on AI usage





Openness to law firms using AI

**Steven A. Schwartz and Peter LoDuca (New York, 2023): Submitted a legal brief with fictitious case law citations generated by ChatGPT.**

**Zachariah Crabill (Colorado, 2023): Used ChatGPT to draft a motion that included references to non-existent court cases.**

**An unnamed Melbourne lawyer (Australia, 2024): Provided a judge with a list of AI-generated cases that did not exist.**

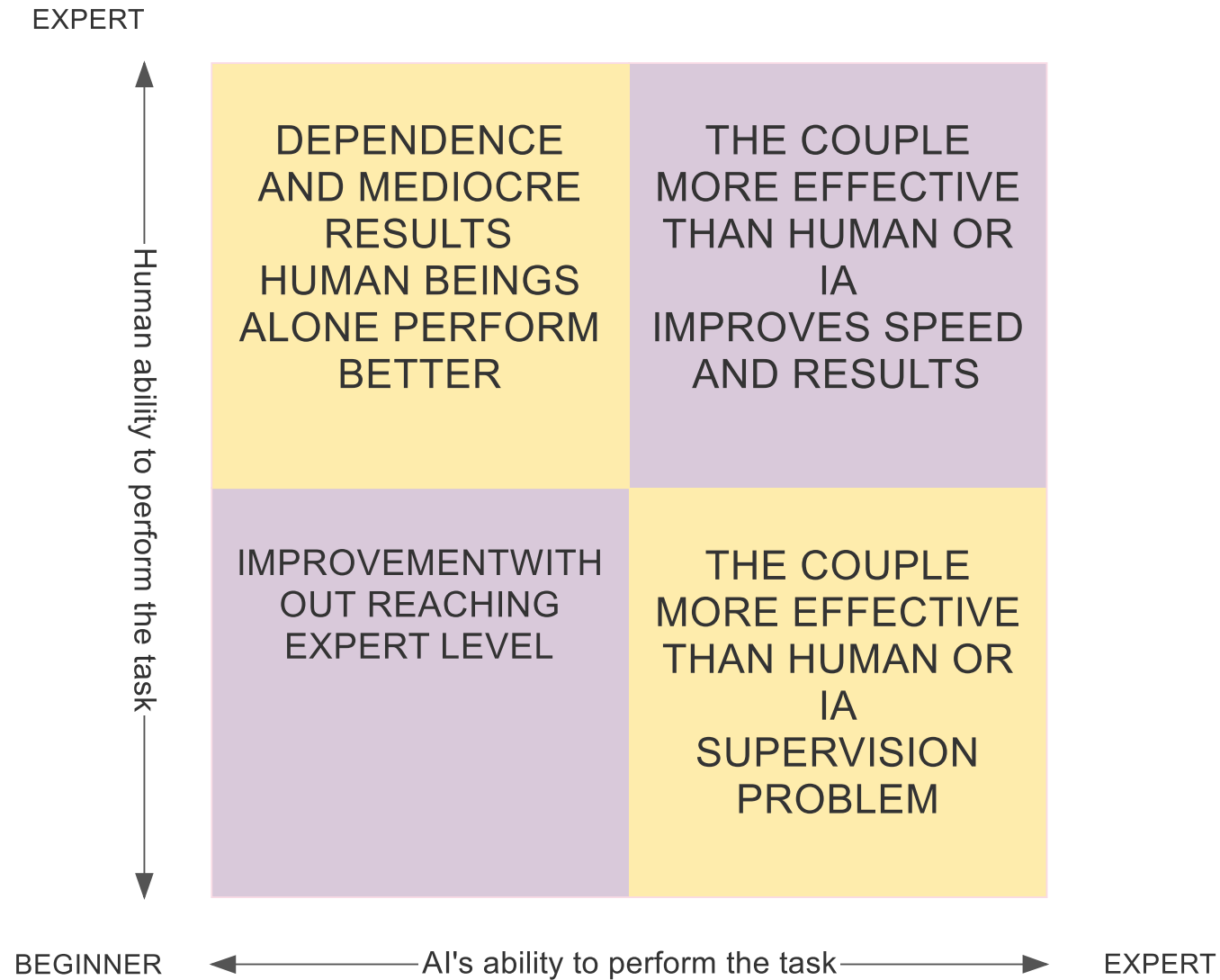
**An unnamed lawyer in Brazil (2023): Used ChatGPT to draft a legal petition containing non-existent case law citations.**

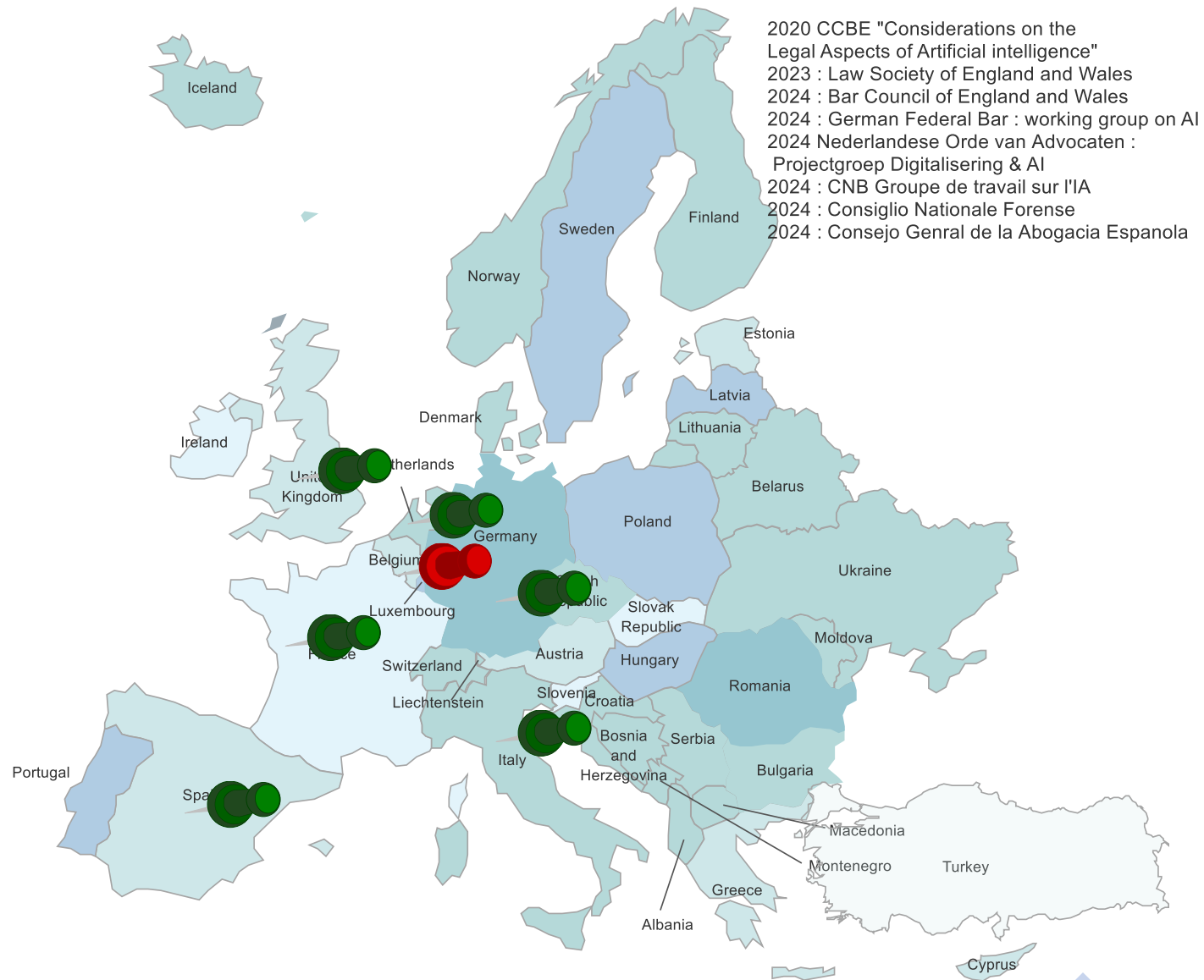
**An unnamed lawyer in Argentina (2023): Submitted a legal document generated by AI with references to non-existent laws.**

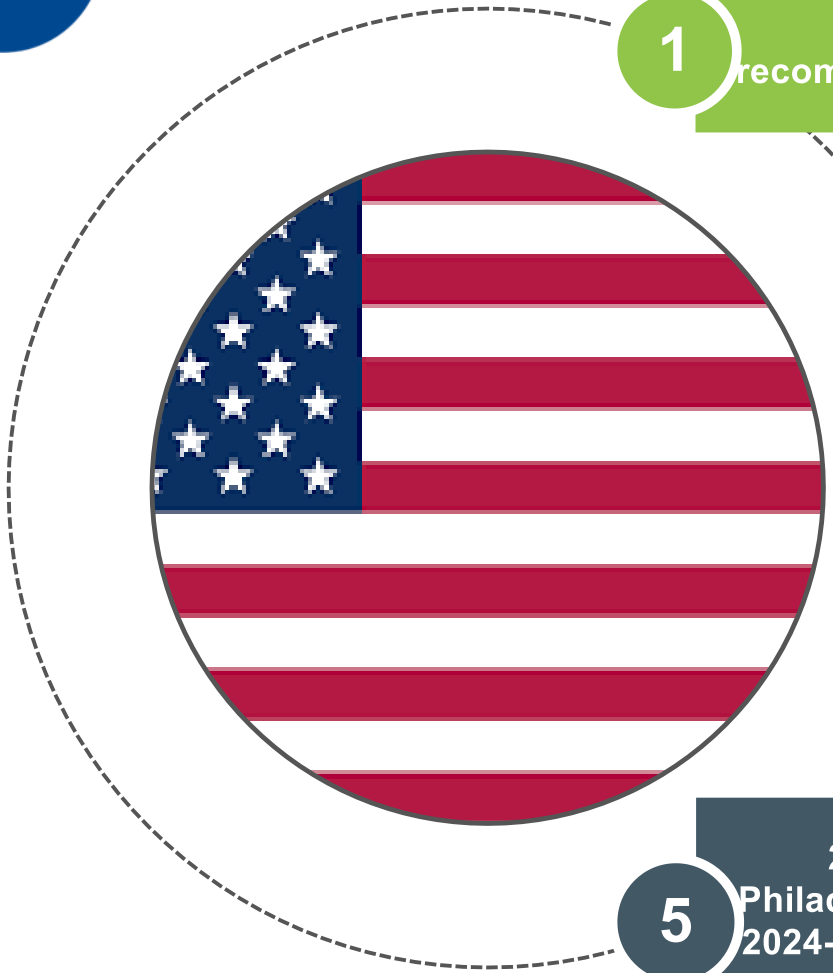
**An unnamed lawyer in Colombia (2023): Used ChatGPT to draft a legal opinion for a client without verifying the accuracy of the information or informing the client about AI use.**

An unnamed Australian law firm (2024): Used an AI tool to analyze confidential client documents without explicit client consent.- An unnamed Canadian law firm (2024): Used an AI tool for setting client fees without disclosing this practice.-An unnamed UK lawyer (2024): Used an AI tool to draft submissions without informing the client or the court.

# HOW TO USE AI IN LAW FIRMS







1

2020: New York State Bar Association recommendations on the use of AI in legal practice

2

2023: California State Bar practical guidance on the use of generative AI in law practice

3

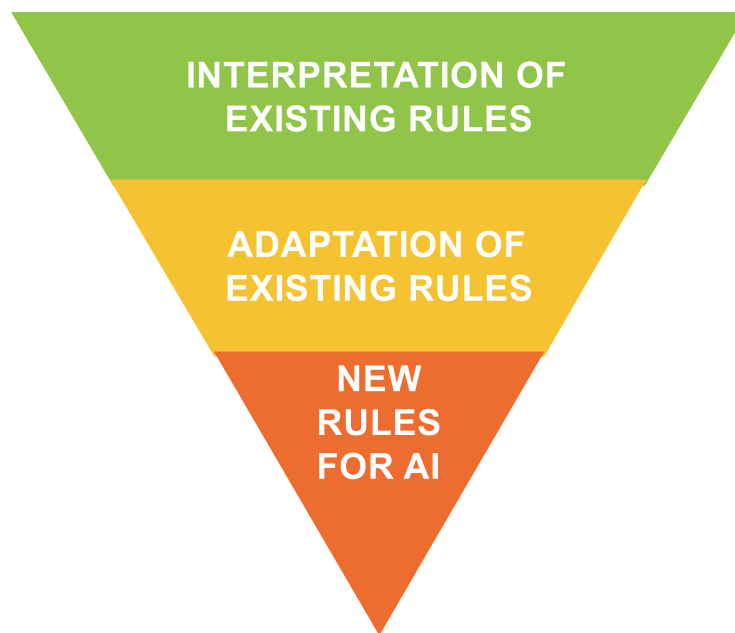
July 2024: American Bar Association Formal Opinion 512 on the ethical use of generative AI tools by lawyers

4

2024: Florida State Bar Association Opinion 24-1 on ethical considerations related to lawyers' use of AI

5

2024: Pennsylvania Bar Association and Philadelphia Bar Association Joint Formal Opinion 2024-200 on ethical implications of lawyers' use of generative AI



# NEWS RULES



- 1 > Obligation to understand the capabilities and limitations of AI tools used
- 2 > Duty to independently verify results produced by AI
- 3 > Obligation to inform the client of significant AI use in their representation
- 4 > Prohibition on blindly relying on AI results without exercising professional judgment
- 5 > Obligation for continuing education on AI developments affecting legal practice

# ADAPTATION OF RULES

CONFIDENTIALITY	Evaluating specific risks related to using AI tools for protecting client data	Obtaining informed client consent before entering confidential information into an AI tool
COMPETENCE	Extending technological competence to understanding and appropriate use of AI tools	
SUPERVISION	Establishing clear policies on AI use within the firm	Training staff on ethical AI use
BILLING	Adapting billing practices for AI use (e.g. not billing for time saved through AI)	
ADVERTISING COMMUNICATION	Transparency about AI use in legal services offered	



## INTERPRETATION OF RULES

DUTY OF DILIGENCE	Applying the duty of diligence to verifying AI results
PROFESSIONAL INDEPENDENCE	Maintaining professional independence in face of AI results
INTEGRITY	Verifying accuracy of AI-provided information before submission to court
PROFESSIONAL RESPONSIBILITY	Lawyer remains ultimately responsible for work produced, even with AI use
CONFLICTS OF INTEREST	Considering potential conflicts generated by use of shared AI tools



# WEBINAR

## The impact of artificial intelligence on European lawyers' practices

18 NOVEMBER 2024, 10:00 - 13:00 CET

#TRAVAR *Training of Lawyers in various areas of EU Law*



### Marlene Winther Plas and Alexandre Coratella **The challenge facing non-English speaking countries regarding generative AI**

18 November 2024, Zoom



Co-funded by the European Union

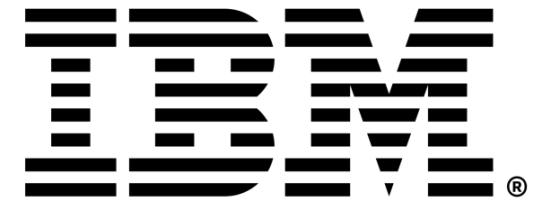




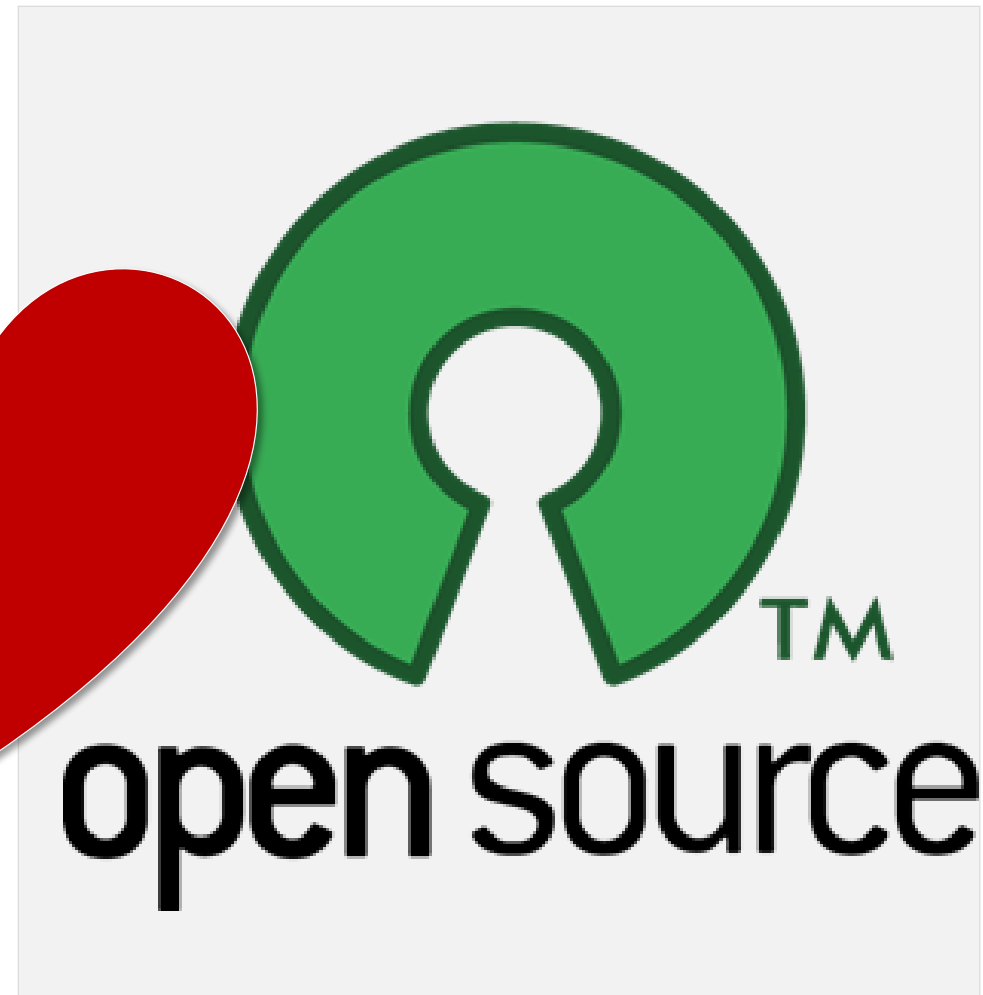
**DANSK  
SPROGMODEL  
KONSORTIUM**



**ALEXANDRA  
INSTITUTTET**



**DANSK  
ERHVERV**



# Continental & Common Law



French law has only been written in French since 1539.

French law belongs to the family of Romano-Germanic law. Law is mainly codified in the various texts adopted by the public authorities. Legislation enacted by the government is the main source of law.

The common law system is based on the concept of stare decisis, which basically means that what is settled should not be changed.

Precedents guide judges in making decisions in similar cases. Courts must therefore respect precedent and not disturb established law.

In continental law, judges decide cases primarily on the basis of the applicable code. Judges may refer to previous decisions of certain courts, but they do so only to ensure consistency, not because the law requires them to follow other judicial decisions.

Precedents are not binding in French law.

But AIs trained on common law are not as accurate as they should be, because the logic behind the rulings is very different.

# Promoting non-English speaking AIs : Langula & ALT-EDIC



Only 0.2% of the data used to train LLMs are French.

ALT-EDIC was officially created with the publication on 7 February 2024 of the European Commission's implementing decision (Implementing Decision (EU) 2024/458). Sixteen Member States are participating: Bulgaria, Croatia, the Czech Republic, Denmark, France, Greece, Hungary, Ireland, Italy, Latvia, Lithuania, Luxembourg, the Netherlands, Poland, Slovenia and Spain.

The project is led by France.

ALT-EDIC's mission is to develop a common European infrastructure in the field of language technology, with a particular focus on large-scale language models. It aims to improve European competitiveness, increase the availability of European language data and preserve Europe's linguistic diversity and cultural richness:  
implementing alternative language processing models

The ALT-EDIC Action Plan focuses on thematic areas: Data & Existing Language Models, Development of New Language Models.

# Langula & ALT-EDIC



By pooling their resources, the members aim to achieve the critical mass of data and other resources needed to build and refine large-scale linguistic models that would be difficult for any one member to achieve on its own.

This will enable the consortium to train generative AI models in languages other than English.

In France, this is a state-owned project :

1. Project led by the Ministry of Culture, and in particular the General Delegation for the French Language and the Languages of France and the General Directorate of Enterprises (Ministry of the Budget - Direction générale des Entreprises (DGE-Bercy)).
2. The team will be recruited in the autumn to start work in 2025.
3. Two scientific committees will be established:
  - 1 cultural bias committee
  - 1 committee on legal and ethical issues - 1 member of the French National Bar Association will participate in this committee.

# LangulA



The aim of LangulA, which could become the French branch of ALTEDIC, is to train models in the French language while respecting copyright. The approach is not purely economic or industrial: the treatment of language by AI is seen in its political dimension. Language is first and foremost a cultural object and a democratic issue.

The aim is to bring together in one place a mass of French-language training data that can be used to improve artificial intelligence models and combat the cultural biases of predominantly Anglo-Saxon AIs, in order to make better use of French data with major technology groups.

The INA (National Archives), the BNF (National Library) and the CNRS (National Institute for Scientific Research) will create an experimental database for this purpose. Non-governmental organisations will be involved in the project to collect available content. This database will be developed by a public start-up, incubated by the French Inter-Ministerial Directorate for Digital Affairs (Dinum), with the aim of providing developers with a hub of French-language data.

The French authorities also want to capitalise on francophone solidarity, for example with Quebec.

This work should be useful to Belgium, Switzerland and Luxembourg, which have similar legislation in some areas (company law, civil law in particular).

# Corporate Initiatives



Some legal tech companies have developed AI solutions based on RAG, which is quite relevant as such companies hold relevant legal data.

For example, LexisNexis or Lefebvre Dalloz, which are major publishers of law books.

When the RAG is based on relevant data, there are far fewer hallucinations, but major AI products such as Chat GPT or Copilot are not yet suitable for lawyers.

We, the CNB, have published a book of guidelines (ethical and practical) to enable lawyers to use AI.

We are currently auditioning all AI companies operating in French law and developing AI solutions for lawyers.

# WEBINAR

## The impact of artificial intelligence on European lawyers' practices

18 NOVEMBER 2024, 10:00 - 13:00 CET

#TRAVAR *Training of Lawyers in various areas of EU Law*



Péter Homoki

**Check-list on the acquisition and development of AI tools**

18 November 2024, Zoom



Co-funded by the European Union

# Agenda



1. Buying „AI tools”: function, ~~price~~, security, performance
2. Building trust in AI tools: measuring what’s difficult to measure (and countering misleading information)
3. Benchmarks and datasets:  
NLP tasks and legal tasks, LLM („genAI”)-specific benchmarks
4. Benchmark problems you should be aware of
  - costs
  - contamination
  - difference in needs: laypeople vs. lawyers
  - correlation (lack of)

# Buying AI tools: function

value-chain of legaltech tools

(LLM) as an *integrator tool*

machine learning tools for data mining/analysis

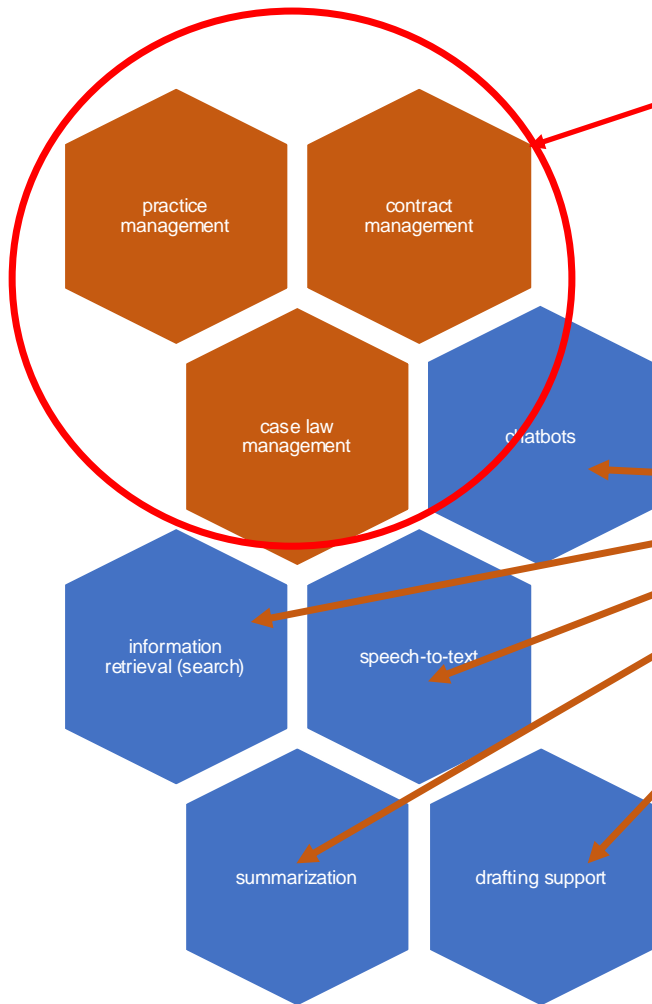
“*comfort tools*”: project and process management, knowledge management, rule-based document assembly

editorial and value-added content (templates, practice notes, legal literature/commentaries, monographies, articles)

public legal databases (laws, published cases)



# Risks in functions



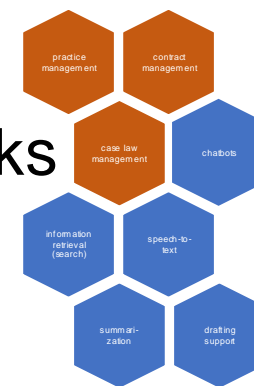
all these „*management*” tools could mean *many different things*, including risky functions, e.g. genAI-based drafting or analytics, *contract negotiation tools* or ML-based text extracting tools

depending on the *particular technology* used, these tools could produce

- incorrect results
- different results than last time [**issue of performance**]
- could possibly leak some information [**issue of security**]

# Find and understand typical weak points!

- **Drafting support:** complicated conditions and loops, unreliable data in database; genAI-based drafting (too much drafting freedom, unwanted text)
- **Information retrieval:** which hurts more: low *precision* (many irrelevant findings presented [false+]) or low *recall* (many relevant findings were missed [false-], e.g. missing **clause excluding termination right**)?
- **Analytics:** relying on machine learning models (SVM, neural networks) without understanding it (bias in samples, probabilities of incorrect answer ...)
- **Chatbots:** rules-based prescribed responses (intent) vs. LLM-based („conversational”) – different roles, different risks



# Buying AI tools: security

- main risk** is not involuntary leaking or governmental access, but
- losing access to your **own data** (no exit plan, no backup)
  - third-party having access to data via fault of original user (lack of appropriate access controls)

## **supply chain security:**

- know your provider, know their (major) subcontractors, understand the technology to the necessary degree (see risks)
- understand if you can reach the provider when needed (how responsive are they? can you litigate if needed?)
- know what you can backup and be prepared to lose everything else

# Performance of AI tools – how to build trust?

## **objective:**

- 1) measuring what's difficult to measure,
- 2) countering misleading (promotional) information about capabilities

**benchmark:** to evaluate or compare certain characteristics of tools (their performance, fairness, toxicity etc.)

in this case: their *performance* or *reliability*

can we rely on these tools? how do they compare to each other?

to build benchmarks, we need standardized sets of *data* and *tasks*

**dataset:** a specific *collection of data* used to *create- or benchmark* tools or evaluating performance of tasks (including to train, test or evaluate models), incl. a method to evaluate the answers

**benchmarks & datasets depend** on roles, tasks, of languages, jurisdictions ... a benchmark may be composed of measuring multiple tasks

# NLP tasks vs. legal tasks

traditional natural language processing tasks:

- question-answering (QA)

```
[closed book]: SQuAD [=dataset]: "qas": [{"question": "Into what language did Martin Luther translate the Bible?","answers": [{"text": "German","answer_start": 522}], "is_impossible": false}, {"question": "In what year did Martin Luther post his 95 theses?","answers": [{"text": "1517","answer_start": 132}], "is_impossible": false}], "context": "Martin Luther, a German monk, started the German Reformation by posting 95 theses on the castle church of Wittenberg on October 31, 1517. The immediate provocation spurring this act was Pope Leo X's renewal of the indulgence for the building of the new St. Peter's Basilica in 1514. Luther was challenged to recant his heresy at the Diet of Worms in 1521. When he refused, he was placed under the ban of the Empire by Charles V. Receiving the protection of Frederick the Wise, he was then able to translate the Bible into German."
```

[SQuAD 2.0](#) performances [=benchmark]

## Leaderboard

SQuAD2.0 tests the ability of a system to not only answer reading comprehension questions, but also abstain when presented with a question that cannot be answered based on the provided paragraph.

Rank	Model	EM	F1
	Human Performance Stanford University (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Jun 04, 2021	IE-Net (ensemble) RICOH_SRCB_DML	90.939	93.214
2 Feb 21, 2021	FPNet (ensemble) Ant Service Intelligence Team	90.871	93.183

# NLP tasks vs. legal tasks

traditional natural language processing tasks:

- summarization

**Article:** „One easy way to conserve water is to cut down on your shower time. Practice cutting your showers down to 10 minutes, then 7, then 5. Challenge yourself to take a shorter shower every day. Washing machines take up a lot of water and electricity, so running a cycle for a couple of articles of clothing is inefficient. Hold off on laundry until you can fill the machine. Avoid letting the water run while you're brushing your teeth or shaving. Keep your hoses and faucets turned off as much as possible. When you need them, use them sparingly.”

**Summary:** „Take quicker showers to conserve water. Wait for a full load of clothing before running a washing machine. Turn off the water when you're not using it.”

[ROUGE](#) performances [=benchmark]

- classification (e.g. evaluating language „understanding”)

[BoolQ](#) [[SuperGLUE#1](#)]

- **Passage:** Barq's – Barq's is an American soft drink. Its brand of root beer is notable for having caffeine. Barq's, created by Edward Barq and bottled since the turn of the 20th century, is owned by the Barq family but bottled by the Coca-Cola Company. It was known as Barq's Famous Olde Tyme Root Beer until 2012.




**Question:** is barq's root beer a pepsi product

**Answer:** No

[RTE](#) [[SuperGLUE#2](#)] (Recognizing Textual Entailment): *[does it follow?]*

- **Text:** Dana Reeve, the widow of the actor Christopher Reeve, has died of lung cancer at age 44, according to the Christopher Reeve Foundation.
- **Hypothesis:** Christopher Reeve had an accident.
- **Entailment:** False

Leaderboard Version: 2.0

Rank	Name	Model	URL	Score	BoolQ	RTE
1	Inspur Cloud	Hairuo		91.4	92.5	92.8
8	DeBERTa Team - Microsoft			90.3	90.4	93.2
9	SuperGLUE Human Baselines			89.8	89.0	93.6

# NLP tasks vs. legal tasks

[COLIEE \(2021\)](#): 5 different tasks, task\_3: 768 *Japanese Civil Code* articles, 806 questions – *which article answers the given question?*

[MAUD \(2023\)](#): Merger agreement understanding dataset, 47457 annotation of legal text from 152 public merger agreements

**Question:** When are representations and warranties required to be made according to the bring down provision?

**Options:** A: At Closing Only; B: At Signing & At Closing

**Example:** *Section 7.2 Conditions to Obligations of Parent and Acquisition Sub to Effect the Merger. The obligations of Parent and Acquisition Sub to effect the Merger are, in addition to the conditions set forth in Section 7.1, further subject to the satisfaction or (to the extent not prohibited by Law) waiver by Parent at or prior to the Effective Time of the following conditions: (a) each of the representations and warranties of the Company contained in this Agreement, without giving effect to any materiality or "Company Material Adverse Effect" or similar qualifications therein, shall be true and correct as of the Closing Date, except for such failures to be true and correct as would not, individually or in the aggregate, have a Company Material Adverse Effect (except to the extent such representations and warranties are expressly made as of a specific date, in which case such representations and warranties shall be so true and correct as of such specific date only)*

[CUAD \(2021\)](#): labelling the text of 510 commercial contracts based on 41 types

12; Category: No-Solicit of Customers; *"Is a party restricted from contracting or soliciting customers or partners of the counterparty, whether during the contract or after the contract ends (or both)?"*; Answer Format: **Yes/No**

13; Category: Competitive Restriction Exception; *"This category includes the exceptions or carveouts to Non-Compete, Exclusivity and No-Solicit of Customers above."*; Answer Format: **Yes/No**

[LexGLUE](#) (2021): a composite dataset for evaluating legal language understanding tasks (ECtHR, CaseHOLD etc.)

# LLM („text genAI”)-specific benchmarks

LLMs (GPT3+-, Llama, Gemini etc.) have special capabilities  
→ often new NLP-benchmarks (evaluations) are used

Previous legal benchmarks are reused,  
new legal benchmarks are introduced

# „GPT-4 Passes the Bar Exam” – from NCBE materials

**#1** „A plaintiff domiciled in State A brought a federal diversity negligence action in State A against a defendant domiciled in State B... The defendant has moved for an order dismissing the action based on the personal jurisdiction challenge asserted in the amended answer. Should the court issue the order?

- (A) No, because the defendant waived the challenge to personal jurisdiction by failing to include it in her original answer.
- (B) No, because the defendant was personally served with process within 100 miles of the federal courthouse where the action is pending.
- (C) Yes, because the defendant lacks minimum contacts with State A.
- (D) Yes, because service was not delivered to the defendant at her home”

Answer: **C**

**#2** „Four months ago, Victim was shot and seriously wounded in City. Defendant has been charged with attempted murder. The prosecution’s theory is that Victim and Defendant were both members of a criminal street gang called “The Lions,” which engages in drug dealing, robbery, and murder in City. The prosecutor alleges that the shooting was the result of a gang dispute. Defendant has brought a pretrial motion objecting to the prosecutor’s introducing the following anticipated evidence: (A) Testimony by a City detective who will be offered as an expert in gang identification, gang organizational structure, and gang activities generally and as an expert on particular gangs in City. The detective is expected to testify as follows:

...  
Defense counsel’s motion raises the following objections to the evidence described above:

1. The detective’s anticipated testimony about gang identification, organization, and activities is improper expert testimony.

...

How should the trial court rule on each objection? Explain. (Do not address constitutional issues.)”

„The trial court should deny the defendant’s motion and allow the detective’s anticipated testimony about gang identification, organization, and activities as proper expert testimony. Under Federal Rule of Evidence 702, a witness who is qualified as an expert by knowledge, skill, experience, training, or education may testify in the form of an opinion or otherwise if: (a) the expert’s scientific, technical, or other specialized knowledge will help the trier of fact to understand the evidence or to determine a fact in issue; (b) the testimony is based on sufficient facts or data; (c) the testimony is the product of reliable principles and methods; and (d) the expert has reliably applied the principles and methods to the facts of the case. ...”

# LLM („text genAI”)-specific benchmarks


LegalBench: a superset built from 33 existing task benchmarks (MUAD, CUAD ...) to evaluate „legal reasoning” capabilities of LLMs

**Language-specific** legal benchmarks for LLM (e.g. ArabLegalEval)

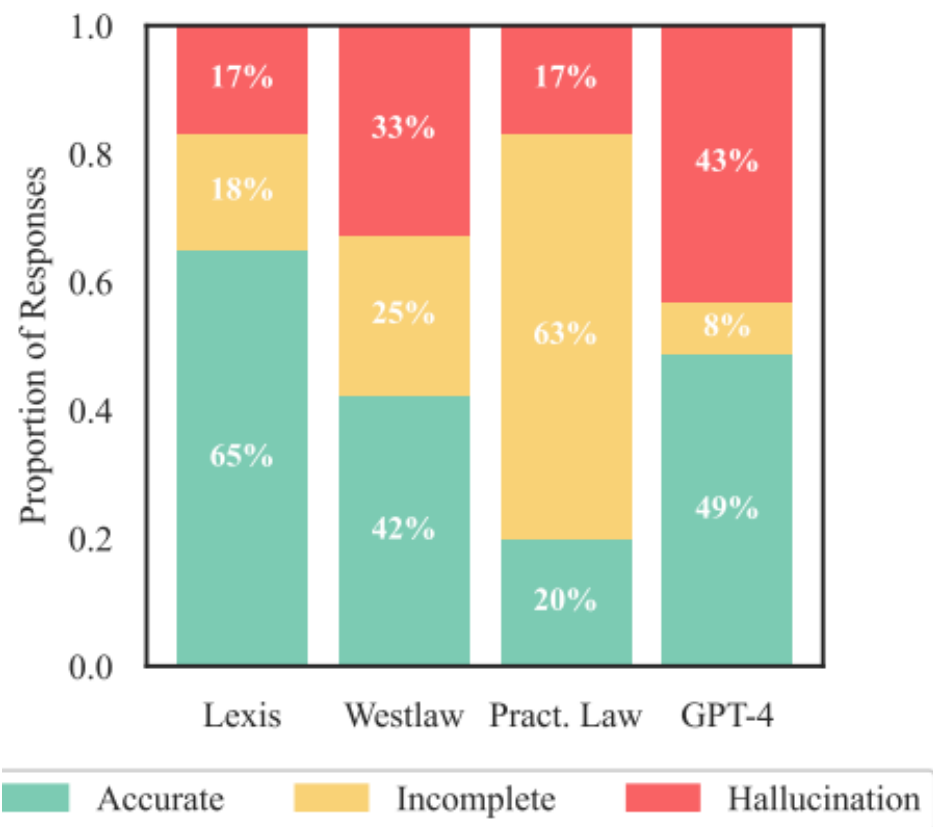
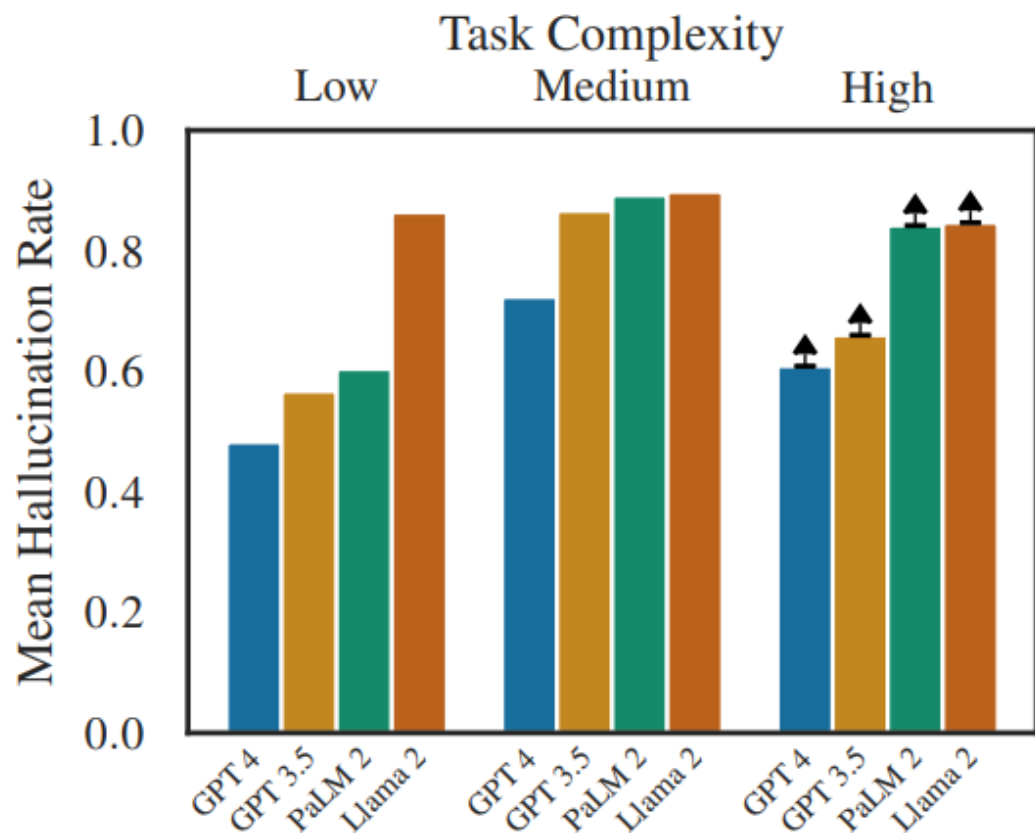
**Academy-led** reliability/**hallucinations benchmarks**:

legal hallucinations (general LLMs),

Stanford Legal Hallucination Benchmark (legal-specific LLMs)

**Law firms** (practising lawyers) testing which tool to use  
(e.g. LinksAI, Ashurst, )

**Vendors** publishing their own benchmarks (BigLaw Bench)



# Generic problems with benchmarks

- costs, language- and jurisdiction-specific developments
- contamination (train-test leakage)
- lack of understanding how users will use tools, which tasks to measure
  - hugely different in laypeople use vs. supporting professional users
- no correlation in expected performance
  - expected in-depth „understanding” or „reasoning” capabilities do not follow (surprising failures despite good benchmarks)

[1] Hello everyone, my name is Peter Homoki, I'm a lawyer working in Hungary. I was the chairman of the IT Law Committee of CCBE a very long time ago. I also participated in a project by the European Lawyers Foundation, CCBE, and the European Commission that ended in 2022 with publishing the Guide on the use of AI-based tools by lawyers and law firms.

[2] Today, I will talk about what lawyers should focus on when buying an AI-based tool for their practice. Especially on how to measure and compare the performance of these tools - that's why we are talking about benchmarking.

First, I will give you a broad overview of how AI tools fit among legaltech tools and where we have to pay special attention to the use of such tools. While the price of such tools is also important, I will not discuss this, considering that the legaltech market is very fragmented in the EU, and there are as many available tools as countries, with different pricing.

Following a short overview of the major security issues, I will introduce you to how lawyers can start to trust these very complex, highly diverse, but often unreliable AI tools - what are datasets and benchmarks, how did they change with time.

I will also introduce you to some of the more exciting current research on the reliability of AI tools. I will close my presentation with a summary of the major pain points with benchmarks and the reliability of AI tools in law in general.

[3] On the left, you can see that there are many different layers of commercial legaltech tools. They all rely on legal databases of some sort, both on public data and private, publisher or customer-owned data.

When we talk about AI tools, we have to differentiate these new tools from the more traditional legaltech tools.

Mainly, the difference is not about the technology, it's more about how predictable they work. They are complex in a different way than traditional tools, and will not always come up with the same results, and thus, can even surprise their users.

That is a serious risk that you cannot manage by training the users the same way as you can do with more traditional tools.

There are very important technological differences in the background, but for us, this unpredictability is the most important.

We will be discussing the layers highlighted on the left: machine learning and large language model (LLM) based tools (also called nowadays as generative AI tools) and of course those (previously harmless) categories of tools that now incorporate such new functionalities as well.

"AI" is a marketing label, but legaltech tools do not and will not always have a big title on them saying that they are "AI-based tools".

On the right, those software that often (but not always) incorporate AI capabilities are marked with a red background - these are quite generic tools that can include dozens of different functions, so their name will not clarify in any way how they operate. For those tools with a blue background, you should presume that any such machine-generated output will need some level of human review. Unless you have reliable information to the contrary.

[4] Let's take a more in-depth look. Blue tools could come up with incorrect results, or even very different results than the last time you've used it. And for tools using a certain technology, there is

also a remote chance of some of your data being leaked to third parties. We'll return to these in a minute.

Red tools, like practice and case management tools, can include dozens of different functions based on what their publishers want to market. Some include analytical functions or large language model (LLM)-based text drafting, or even some kind of contract negotiation functionalities. If these tools include such functionality, you have to give the results the same level of human review as you would give to dedicated AI tools with a blue background.

[5] No matter what kind of AI tool you are using, you have to be aware of its weak points, even if you do not have to become a machine learning specialist.

For software supporting legal drafting, you can create very complex drafts even with traditional software technology that was cutting-edge 50 years ago, but this problem has been significantly exacerbated by LLMs.

Especially if the instructions (the "prompt") given to such models are not specific enough: if the lawyer does not provide the necessary details for a clause it wishes to include, the tool will probably never fill that gap by "common-sense" provisions, like what we would expect from a trainee.

Information retrieval tools, also called research tools are not new for legal practices. But the larger the amount of information you want to review with them, the more mission-critical their functionality becomes, and the less you can do about any decrease in their performance.\*\*\* Depending on the task and use case, it may be a bigger problem if they give you too many false results, or if they miss a single, but still important instance.

If information retrieval and extraction of relevant clauses is unreliable, it's no surprise that any analytics relying on such results will also be at least as unreliable. The lack of reliability could be caused by unidentified problems in the examples the model was trained on, or in using the model for some new real-world problems that it was not prepared to handle.

Lastly, different types of chatbots could mean different problems for lawyers. The risks of chatbots that restrict input to a set of valid questions and the answers to a set of templates, are minimal. In contrast, chatbots that simply plug into large language models and are theoretically restricted by the LLM instructions (by prompts) easily step out of their guardrails and can be a catastrophe for any law firm.\*\*

[6] In relation to the security of AI tools, the biggest risk is losing access to your own data. While many AI tools rely on some form of cloud computing services, it is usually not easy to export and back up all information you have in the tool, including any work in progress. Just as there are no standardised "AI tool applications", there is also no standardised format of how they store and process all your data, so it won't be easy to transfer such data to your premises or to a new provider. Furthermore, vendors come and go, merge into larger providers, and cease their business in less profitable markets, which could all result in your loss of important data.

Similarly, most AI tools rely on some form of cloud computing services, at least during the analysis, but often for storage as well. It is your responsibility to make good use of the access control tools provided, no matter how inconvenient those are.

But besides these traditional worries, it is very important for you to understand what the vendor is offering. What kind of tool are you using, what parts can introduce new surprises? Where will you need extra human attention?

Also, keep in mind where the vendor is located and try out how responsive the customer service is if you need some help.

[7] Now, turning to the performance of these AI tools. Considering that AI tools are often quite unpredictable by nature, we have to better understand just how unpredictable they are, and for what tasks.

To rely on their capabilities, we need to introduce some level of predictability of this unpredictability\*\*\*. This can also help reduce the haze created by the misleading information we can read daily from most AI vendors.\*\* By relying on appropriate benchmarks published by vendors, we can also easily compare their relative performance and their value.

Benchmarks are used in machine learning and AI in general to evaluate or measure certain characteristics of tools. This could be a measure of their performance (as in our case), but there are also different benchmarks for more specific purposes (like measuring the toxicity of a tool or a dataset.)

So, benchmarks give you results that you can compare. But to have such a benchmark, you will need to use the same standardised set of data, both as input and the expected set of answers.

Benchmarks and datasets are specific to a task, role or domain. In relation to legal use, if you have one dataset for the English language, most often you cannot just translate that to a different language and use it. As we will see, some benchmarks are a composite of different results in several datasets.

Let's take a look at some examples of benchmarks, and how benchmarks are different across tasks.

[8] The more ancient benchmarks that are still useful in the legal field are those used in generic natural language processing (NLP) tasks. The first example shows you a question-answer-source triad - the extract provided is used as a basis for answering a number of "closed-book questions". Here, SQuAD means both a dataset and a benchmark. Part of the dataset is used for training, and another part for evaluating the performance. You can also see the leaderboard for the benchmarks results: you can compare a "human baseline" and the results of two models from 2021.

As you can see, these models from three years ago have already surpassed human baseline. Needless to say, this doesn't mean any replacement of humans, because this is just a very specialised task.

[9] Another traditional NLP task is summarization - here, the machine generated summary is compared to the predefined "reference summary" and measured according to that standard. Similarly to SQuAD, ROUGE performance was also a once widely-used benchmark.

Yet another example is the classification, in this case, classification based on how well the tool can perform tasks that would require, if done by a human, understanding the text. This will of course not mean that AI tools "understand" the text in the same way we humans do. But it is still a good measure to compare the performance of tools. Also, it gives some remote hope that tools excelling in this benchmark called SuperGLUE will also approximate human understanding, at least in some ways.

The SuperGLUE example here is a composite benchmark, which is made of 10 different tasks, of which two binary classification tasks are included with an example, where the decision is to be made based on the question and the "passage", the extract of text.

[10] Now, turning over to legal task benchmarks and datasets, first I give you some very simple examples.

COLIEE is a question-answering benchmark where the tool is evaluated on five different tasks. One of the tasks is about picking the correct article in the Japanese Civil Code as a response to one of the 806 specific questions.

MAUD is a dataset for testing how well tools can perform in correctly choosing the right answer for a specific contractual clause in relation to merger agreements.

Similarly, CUAD is a dataset used for comparing how these tools can answer questions related to the content of specific commercial contracts, like is this text a "governing clause" or not?

LexGLUE is a superset of different datasets. One of these datasets is made of sentences from 50 online terms of services. The tool being evaluated must determine whether a given sentence is considered an 'unfair term' under EU consumer law, and if so, identify which of the 8 possible categories it falls into.

As you can see, to create all these reference datasets, a lot of effort and money is needed. While some simple datasets can be outsourced to "legal students", some need more expensive expertise.

[11] Compared to earlier machine learning tools, the latest large language models, like GPT 3.0 and later, Gemini, or Llama, have special capabilities that warrant the creation of new, specialized evaluations for them, not ROUGE or SuperGLUE.

In the legal domain, we can find both new benchmarks and also a reuse of previous datasets.

[12] One such new benchmark was the famous GPT-4 test in which the model successfully passed the Unified Bar Exam, here with two examples: one from the multiple-choice answer for the Multistate Bar Exam, and the other, a short essay for the Multistate Essay Examination. Clearly, only the highest-performing LLMs are capable of generating relevant, extensive texts that meet the necessary quality standards as well.

While this was very interesting for the public, passing the bar exam will not necessarily make LLMs useful for lawyers - we do not work by passing bar exams daily and answering multiple-choice questions.

[13] That's why the new generation of legal LLM-benchmarks are useful. We have now a more comprehensive superset of almost all existing "simple" legal benchmarks, which was said to measure a more abstract "legal reasoning" capability. However, despite the title, we also have to highlight that there is no guarantee that tools performing well on this LegalBench benchmark will have any "legal reasoning" capability at all - at least not in the way humans use this word.

We can also see that a number of evaluations were created for non-English language sources, like in Arabic or Greek.

We can also see an increasing number of studies to evaluate just how reliable LLMs are in general, how strong their capability to hallucinate is, and how much that hallucination can be further reduced by providing all the legal databases available, and running specially crafted chains of prompts to reduce false answers.

We can also see a number of large law firms publishing the results of their own "in-house" evaluations of these tools, investing a huge number of partner hours to know the coveted answer: just how reliable that expensive tool is.

[14] As you can see, the results are, at the moment, pretty bad. Even the most expensive large language models, specially crafted for law firms, provide inaccurate or incomplete answers most of the time.

This doesn't mean the AI tools are useless or without potential - to the contrary, it's amazing what these tools are currently capable of.

However, it does mean we must remain aware of their limitations and risks, and understand how to mitigate them. We have to invest the necessary time to review all output, no matter how well they are integrated into a complex chain of AI workflow.

[15] Benchmarks are very useful tools to make our life easier and help with the spread of AI tools, but there are some generic problems that we have to be aware of.

Creating these costs a lot of money and effort, and still you cannot use one benchmark globally for all legal tasks that lawyers will be expected to perform or where lawyers would expect the help of AI tools.

While such benchmarks should be public and shared among all interested stakeholders, that might have the unwanted side-effect of tools being optimized for such benchmarks and incorporated in a training process, which will not necessarily show a true picture of the real capabilities of the tool in daily use.

Also, just because we have a tool that in 92% of the cases, correctly categorizes a predefined set of merger agreement clauses, will not mean the tool is reliable for finding all the clauses that are important for a due diligence review of merger documents.