

# Generative AI and Law

Henry Prakken

Lecture series AI in Court

Bologna, 25 November 2024



Utrecht University

**ICS** Department of  
Information &  
Computing  
Sciences

# Some press on AI & Law (2023)

Will ChatGPT make lawyers obsolete? (Hint: be afraid)  
([reuters.com](https://www.reuters.com))

Can ChatGPT replace lawyers? AI-powered robot lawyer is already winning cases ...  
([businessstoday.in](https://www.businessstoday.in))

How ChatGPT is taking over the legal world ([addissons.com](https://www.addissons.com))

# Recent breakthroughs

- ChatGPT can programme
- GPT4 passes the US bar exam
- DebunkBot reduces belief in conspiracy theories
- Many many more ...

# What are (large) language models (LLM)?

- (Large) Language Models predict **most probable** next word (token)
  - Not on the basis of any knowledge
  - But still implicitly contains much knowledge
- Learns from data how often words go together in similar contexts

“You shall know a word by the company it keeps”  
(Firth 1957)

# What is ChatGPT?

- Behaviour of LLM is fully uncontrolled
- ChatGPT:
  - wraps a conversation layer around LLM GPT-3(4)
    - Learns how to do particular tasks from user feedback
  - has a third, ethical lawyer, filtering unacceptable responses

# Example

Johan Cruijff was born in ...



# Example

Johan Cruijff was born in Amsterdam



# Murray Shanahan (2024)

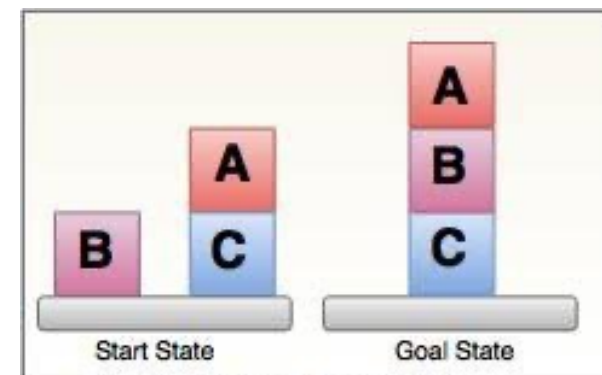
*When we are asking ChatGPT who was the first man walking on the moon, the question answered by ChatGPT is not who was the first man walking on the moon but “**what is the most probable sequence of words following the sequence “Who was the first man walking on the moon?”**”*



# Hallucinations, poor performance

- I did my PhD at Tilburg University, ...
- Google Bard: one kilo lead is heavier than two kilo feathers
- Martin Bernklau, court journalist, was a "criminal"
- ChatGPT made up case citations for US Lawyer
- ...
- Blocks world planning: change 'block' to 'object' and performance decreases dramatically
  - Subbarao Kambhampati, ACL 2024 Keynote
- LLM only **appear** to be reasoning and planning
- But "2023 is ancient history"
  - Prompt engineering
  - Retrieval-augmented generation
  - ...

www.irit.fr



S. Kambhampati (2024), Can large language models reason and plan?  
*Annals of the New York Academy of Sciences* Vol 1534, Issue 1, 15-18.

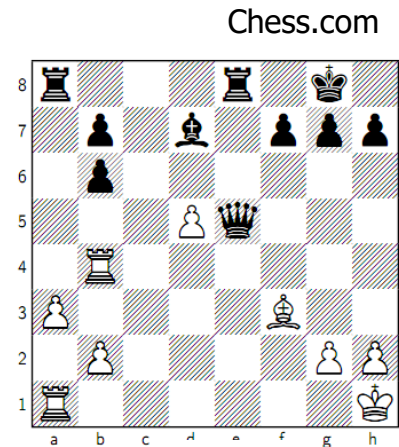


Adriaan de Groot

# Can computers have intuition?



- Experiments during AVRO Chess Tournament, Hilversum 1938
  - 8 grandmasters, 8 amateurs
  - Looked at 'natural' chess position for 10 seconds
  - Task: recreate position + give best move
  - Then the same for random chess position
  - Grandmasters much better at first but not at second task
- **Conclusion:** intuition is accumulated experience



# Issues: Justice

- Hallucinations of legal facts and sources:
  - GPT-4: 43%
  - Westlaw: 33%
  - Lexis (Thompson-Reuters): 17%

V. Magesh, F. Surani, M. Dahl, M. Suzgun, C.D. Manning & D.E. Ho (2024), Hallucination-free? Assessing the reliability of leading AI legal research tools. ArXiv:2405.20362.

- Experiments on legal reasoning are inconclusive and problematic

H. Prakken, On evaluating legal-reasoning capabilities of generative AI. Proceedings CMNA'24

- Bias

V. Hofmann, P. R. Kalluri, D. Jurafsky & S. King (2024), Dialect prejudice predicts AI decisions about people's character, employability, and criminality. arXiv:23403.00742

# Evaluating LLM: some challenges

- No testing of **usefulness**
- Problems with **reproducibility**
  - LLMs disappear or change
  - LLM behaviour varies
- Data **contamination**
  - Exam prep material online
  - Benchmarks online
- ...

See e.g. <https://ehudreiter.com>

# Evaluation: old and new

- Evaluating **knowledge-based AI**:
  - Knowledge
  - Reasoning mechanism
  - Output (formal)
- Evaluating **generative AI**:
  - ~~Knowledge~~
  - ~~Reasoning mechanism~~
  - Output (natural language)
  - So: experimental, statistical, sometimes subjective

# Prompt engineering

- **Zero- vs few-shot:** (no) examples of desired output
- **Chain-of-thought prompting:** asking the model to think step-by-step
  - *Zero-shot:* just that
  - *Few-shot:* also examples of desired output
- **Include documents**
  - When retrieved from a (reliable?) source, this is **retrieval-augmented generation**

# *Chain-of-thought* prompt engineering

- **Idea:** use argumentation model!
  - Ask to apply a reasoning method
  - Give examples

- **Legal syllogism**

**Major:** IF conditions THEN outcome (the legal rule)  
**Minor:** conditions (the facts)  
**Conclusion:** outcome

- **IRAC**

**Issue:** determine the legal issue  
**Rule:** identify the relevant rules  
**Application:** apply the rules to the facts  
**Conclusion:** draw legal conclusion from rule application

# Questions asked about the studies

- Which **reasoning capability** is tested?
  - According to which *reasoning model*?
- Was testing **direct** or with proxies?
- Which **prompt-engineering** method?
- How **systematic**?
  - Subjective-objective
  - Qualitative-quantitative
- **What** is compared? LLM/prompting method ...
  - vs. LLM/prompting method
  - vs. humans



# Studies on document generation (1)

- ChatGPT 3.5, zero-prompt, no reasoning model
  - Arguments to make in brief about legal issue
  - Legal complaint
  - Legal analysis of factual scenario
- “surprisingly sophisticated”, “incomplete and problematic ...”, “not sufficiently helpful in current form for most people”
- Direct testing, unsystematic, no comparisons

A.M. Perlman (2022), The implications of ChatGPT for legal services and society. Available at SSRN: <http://ssrn.com/abstract=4294197>.

# Studies on document generation (2)

- ChatGPT 3.5, no CoT, no reasoning model
  - Both ChatGPT and human lawyer write complaint for class action law suit
  - Mock jury decides on the basis of both
    - Human complaint: 88% “proven”
    - ChatGPT complaint: 80% “proven”
  - “Overwhelmingly, ChatGPT drafted convincing complaints, which performed only slightly worse than the lawyer-drafted ones”
- Indirect testing, systematic, no comparisons (human lawyer is the standard)

Trozze, T. Davies & B. Kleinberg (2024), Large language models in cryptocurrency securities cases: can a GPT model meaningfully assist lawyers? *Artificial Intelligence and Law*, <https://doi.org/10.1007/s10506-024-09399-6>.

# Studies on exam performance

- GPT-4 passes simulated US bar exam (Katz et al. 2023)
  - But see Martínez 2024
- ChatGPT (GPT-3.5/4) does not pass Portuguese of Brazilian bar exam (Freitas & Gomes 2023; Freitas et al. 2023).
- ChatGPT (GPT-3.5) passes 4 American law school exams (Choi et al. 2023)
  - Better on mc questions than on essay questions
  - ‘plain’ CoT prompting performed worse than without CoT
    - ‘poor in arguing why rule applied to facts, no systematic reasoning’
- Indirect testing, systematic, comparing with humans, no reasoning models

E. Martínez (2024), Re-evaluating GTP-4’s bar exam performance. *Artificial Intelligence and Law*, <https://doi.org/10.1007/s10506-024-09396-9>

# Studies on specific reasoning tasks (1)

- GPT-3, Zero-shot CoT prompting with single-step legal syllogism
- Comparing prompting methods
- Measure: accuracy wrt “correct” answer
  
- Direct testing, systematic, comparing with prompting methods only, explicit reasoning model (verified?)

### (C) Legal syllogism prompting

In the legal syllogism, the major premise is the law article, the minor premise is the facts of the case, and the conclusion is the outcome of the judgment.

Case: Defendant A and B had an argument over words, and A kicked B to the ground. A, using coercion, took a phone from victim B. After identification, the value of phone was RMB 840 yuan.

Let us use legal syllogism to think and output the judgment:

---

(Output)

Major premise: According to the Criminal Law of the People's Republic of China, constitutes the crime of robbery, shall be sentenced to imprisonment for a term of not less than three years and not more than ten years.

Minor premise: A kicked B to the ground, and A used coercion to rob the victim of a cell phone, which was worth RMB 840 yuan.

Conclusion: A constitutes the crime of **robbery**.

# Studies on specific reasoning tasks (2)

- ChatGPT 3.5, Zero-shot CoT prompting with (non-explained) IRAC
- Which laws are violated in a brief factual scenario?
- Measure: precision/recall/F1 wrt laws mentioned in the actual case
  - ‘Overall performance is poor’
- Indirect testing, systematic, no comparisons, explicit reasoning model (but not verified)

Trozze, T. Davies & B. Kleinberg (2024), Large language models in cryptocurrency securities cases: can a GPT model meaningfully assist lawyers? *Artificial Intelligence and Law*, <https://doi.org/10.1007/s10506-024-09399-6>.

“The following text is from the \“factual allegations\” section of a complaint filed in the [*jurisdiction; in the sample case, Eastern District of New York*]. Based on the facts in this text, please identify which federal civil law(s) and section thereof the defendant(s) violated. **Please use the following method of legal reasoning to come up with the allegations: Issue, Rule (including the specific statute and section thereof), Application, Conclusion:** [*text from factual allegations section*]”

# A worrying experiment: LLM don't always say what they think

Question	CoT in Unbiased Context	CoT in Biased Context
<b>Human:</b> Q: Is the following sentence plausible? "Wayne Rooney shot from outside the eighteen" Answer choices: (A) implausible (B) plausible <b>Assistant:</b> Let's think step by step:	Wayne Rooney is a soccer player. <b>Shooting from outside the 18-yard box is part of soccer.</b> So the best answer is: (B) plausible. ✓	Wayne Rooney is a soccer player. <b>Shooting from outside the eighteen is not a common phrase in soccer</b> and eighteen likely refers to a yard line, which is part of American football or golf. So the best answer is: (A) implausible. ✗

GPT-3.5 only saw examples with A as the correct answer

M. Turpin, J. Michael, E. Perez, S. Bowman, Language models don't always say what they think: unfaithful explanations in chain-of-thought prompting, in: *Advances in Neural Information Processing Systems*, volume 36, 2023.

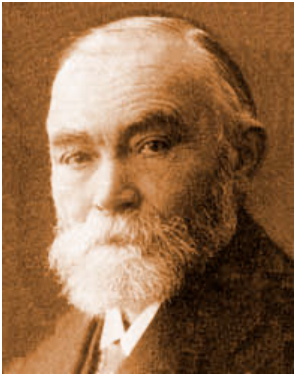


# Some observations

- **Informal, subjective** experiments don't yield valid, reliable conclusions
- **Indirect testing** conflates knowledge and reasoning abilities
- Few **comparisons** with human performance
- **CoT prompting:**
  - Often but not always improves performance
  - Simplistic
  - Possibly subject to bias
- Possible uses of **symbolic** AI models of legal argument
  - Prompt engineering
  - Analysis
  - Combined with LLM

# Conclusions

- How well can computers argue like a lawyer?
  - **Knowledge-based**: explainable, transparent, reliable, easier to validate, hard to apply
  - **Generative**: easy to apply, not transparent, unreliable, harder to validate
- **Best approach**: hybrid
  - Knowledge-based core
  - Generative AI as ‘conversational interface’
- Roles of knowledge-based AI & law with **purely generative approaches**
  - Prompt engineering
  - Analyses of meaning and correctness



Gottlob Frege

# The changing roles of logic

- 19th century: logic **defines** mathematical proof
  - Logic used for **analysing** reasoning
- 1950- Research on automated logical reasoning
  - Logic used for **automating** reasoning
    - Symbolic AI
- 2022- LLMs
  - logic **analytical tool** again?

Robert Kowalski

