Juristische Sprachmodelle 2026

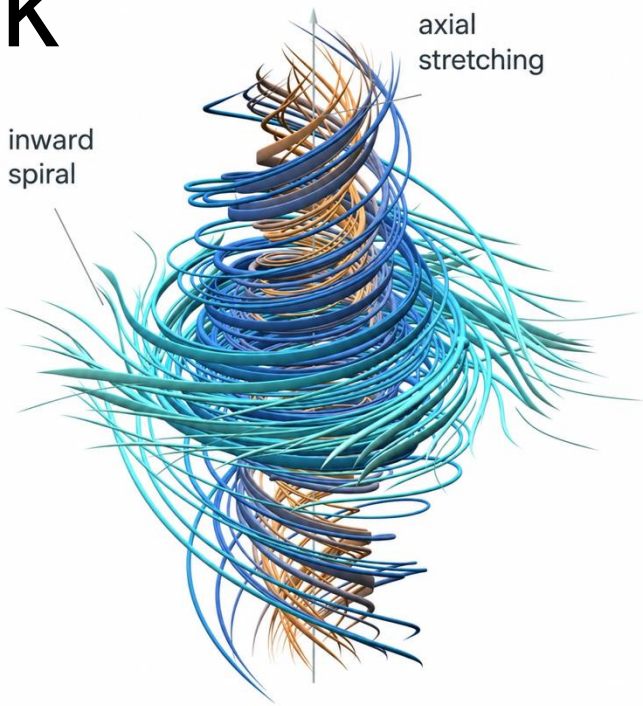
Aktueller Stand & absehbare Entwicklungen

Dr. Michael Vössing | IBM Deutschland GmbH
Prof. Dr. Philipp Reuß | Universität Göttingen
Prof. Dr. Matthias Grabmair | TU München

Agenda

1. **Technik:** Was ist mit heutigen Modellen möglich?
2. **Modellperformanz:** Wie können Modelleigenschaften gemessen werden?
3. **Nutzung:** Wofür werden welche Modelle (bzw. Produkte) heute genutzt?
4. **Forschung/Recht:** Wie bewerten wir Qualität und Zulässigkeit?
5. **Vom Modell zur Lösung:** Was braucht es zusätzlich zum Modell?
6. **Von der Lösung zum Betrieb:** Was braucht es für den Betrieb?
7. **Ausblick & Diskussion**

1.1. KI Durchbrüche in Mathematik



September 8, 2026 Research Publication OpenAI

On the Navier–Stokes Millennium Prize Problem

October 6, 2026 Research Publication OpenAI

Sharing AI progress in mathematics

NAVIER–STOKES LOST IN TRANSLATION

–

WHY LEAN VERIFICATION OF AI AUTOFORMALISATION DOES NOT GUARANTEE CORRECT NATURAL LANGUAGE PROOFS

[link]

Fehleranfällig auch mit modernen LLMs!

Problem → → **Beweis in natürlicher Sprache** → → **Formale Verifikation**

1.2. Adoption in der Rechtspraxis: Aktueller Stand

- Konvergenz zu **agentisch** betriebenen **Reasoning**-Modellen
- Durchbrüche in verifizierbaren Domänen (zB Mathematik)
- Im Recht: Harness-Lösungen
  - von Frontier-Modell-Entwicklern: Astra & Claude for Legal
  - Harvey & Legora
  - Mit Verlagsinhalten: Beck Noxtua, Juris KI-Suite
- Justizeigene Entwicklungsprojekte (noch Laborstatus)
  - Domänen-bezogen: GSJ, OLGA, MAKI, StruKI …
  - Task-bezogen: JANO, ALeKS, …
- Schlüsselfrage: *Welches Modell bzw. System kann was wie gut & wie fair?*

2.1. Forschungsstand: Modell-Performanz

LLMs werden größer, funktionaler und performanter.

- Wie gut können sie juristische Aufgabenstellungen bearbeiten?
- Welche Faktoren spielen bei der Evaluation von LLM-basierten Systemen eine Rolle?

- **Benchmarking**: Systematisch Evaluation von Modellverhalten auf Datensätzen nach bestimmten Kritierien
- Trend "LLM-as-a-judge": LLMs bewerten Performanz anderer LLMs anhand von Rubrik und ggf. Musterlösung
- Herausforderungen:
  - Datenverfügbarkeit für deutsches Recht limitiert
  - Generierung von freiem Text schwer beurteilbar
  - Kalibrierung & statistische Bewertung von LLM judges

Beispiel-Datensätze:

BenGER: Benchmarking LLM Systems on Subsumption-Based Legal Reasoning in German Law

LEXAM:BENCHMARKING LEGAL REASONING ON 340 LAW EXAMS

Methodik:

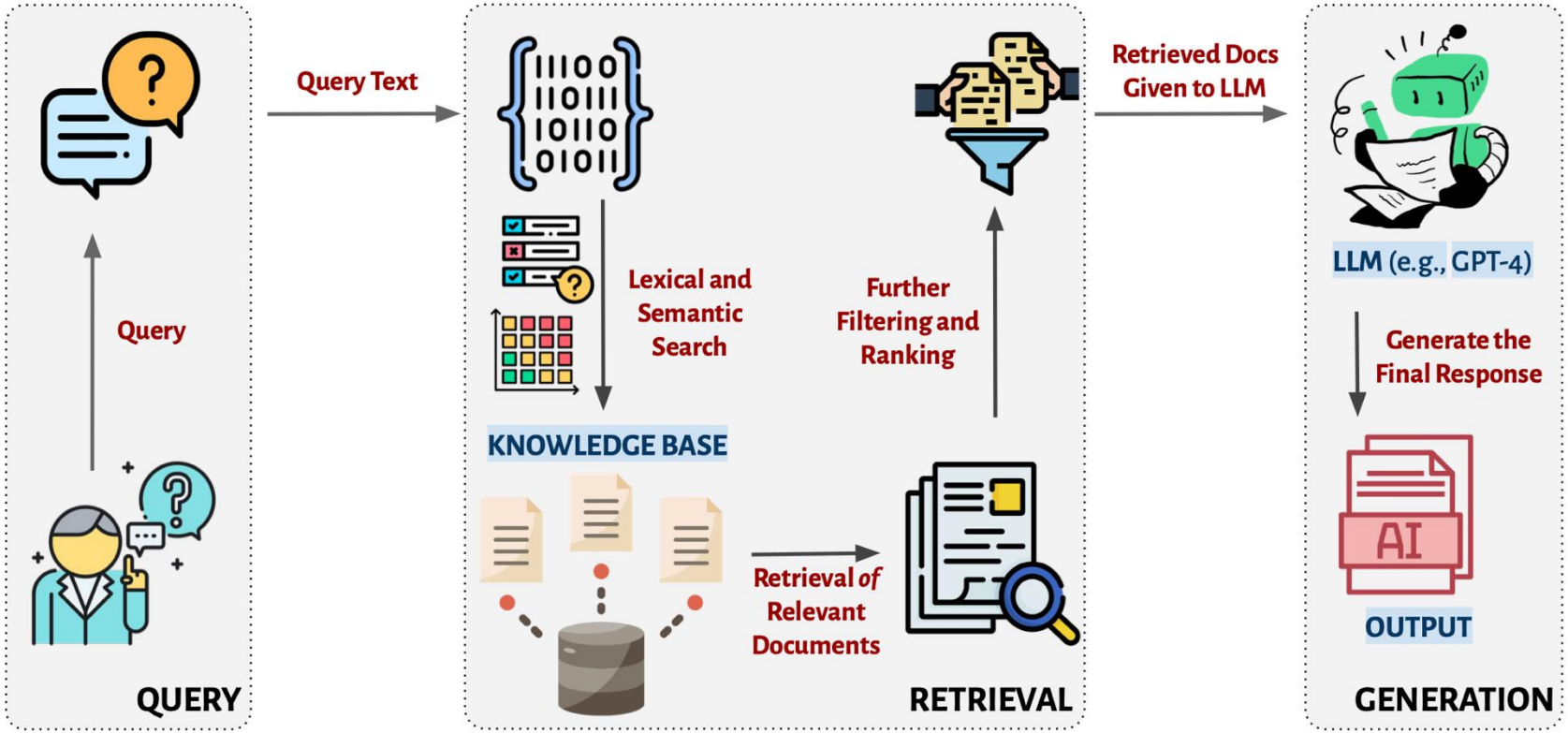
Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools

Who Checks the Citations? Benchmarking Legal Hallucination Detection

There is no free benchmark: An institutional view of legal AI benchmarking

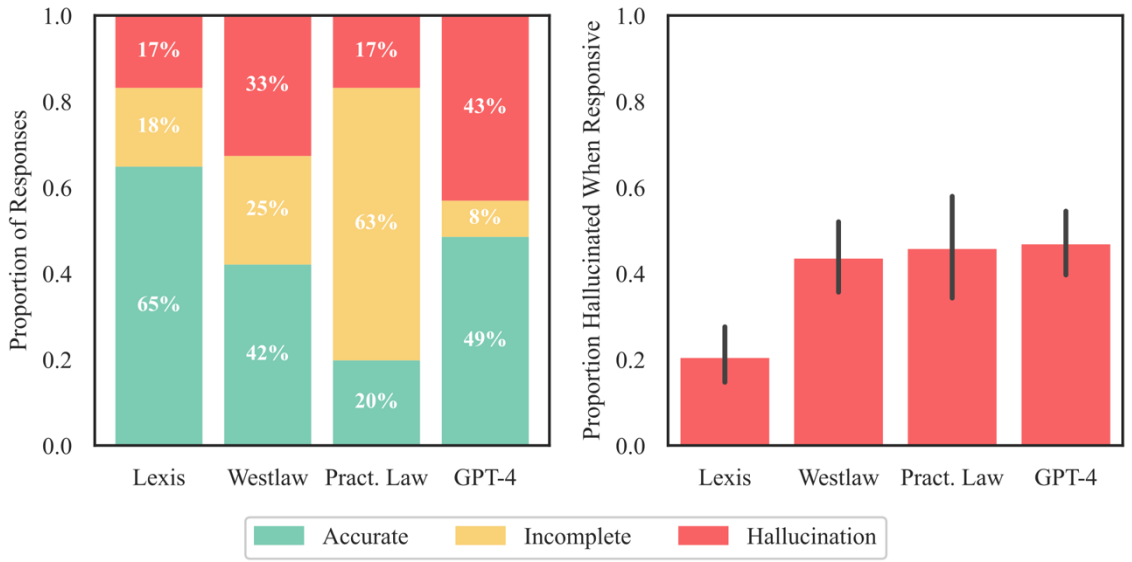
[Quellen: Nagl et al. 2026; Fan et al. 2026; Magesh et al. 2024; Liu et al. 2026; Guha et al. 2025]

2.2 Beispiel: Halluzination trotz RAG



[Magesh et al. 2024]

2.2 Beispiel: Halluzination trotz RAG



Prompt: What are some notable opinions written by Judge Luther A. Wilgarten?

Lexis+ AI Response: One of the notable opinions written by Judge Luther A. Wilgarten is the case of Luther v. Locke. The case citation is 2010 U.S. Dist. LEXIS 144868 and it was decided in the Eastern District of Virginia on January 11, 2010. In this case, the plaintiff had filed a Notice of Appeal of Judge Ellis's decision, but failed to properly prosecute the appeal. [...]

RAG Hallucination Typology

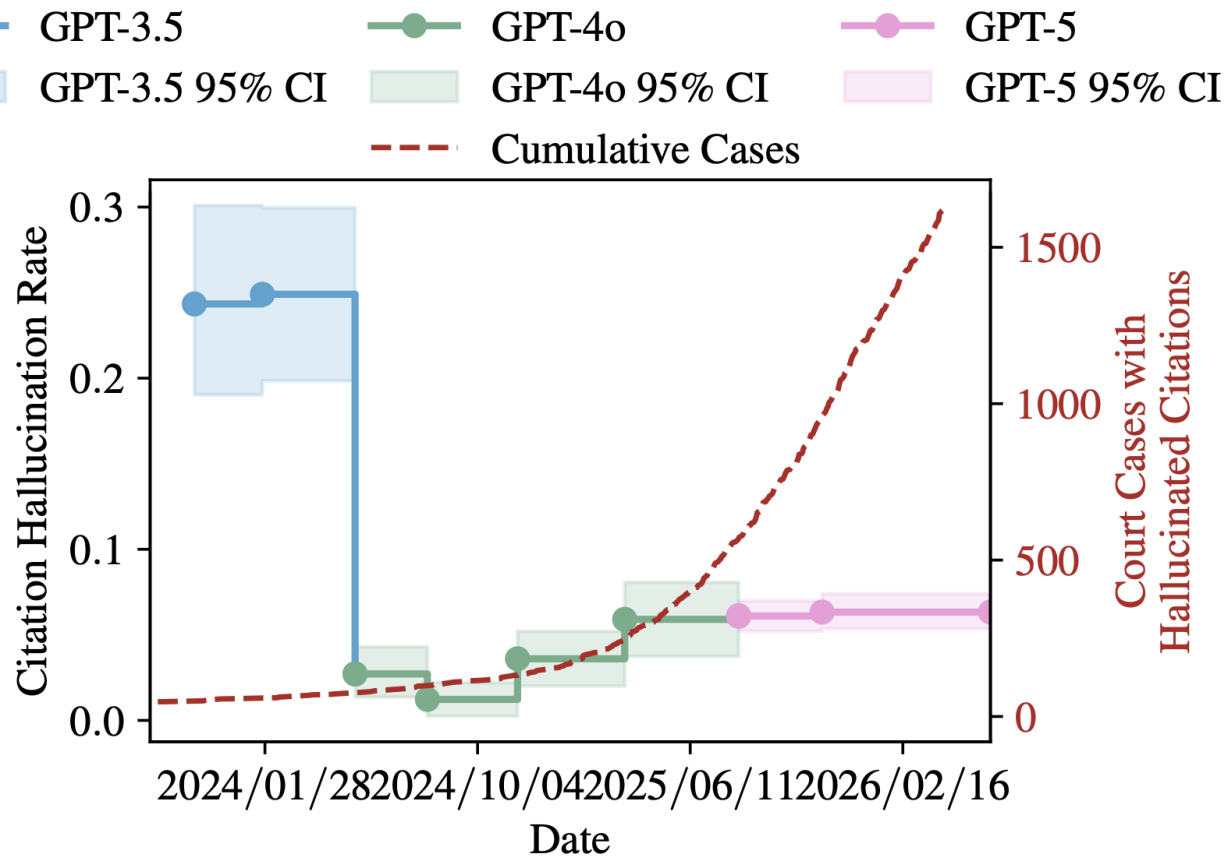
	Description	Example
Correctness		
Correct	Response is factually correct and relevant	The right to same sex marriage is protected under the U.S. Constitution. *Obergfell v. Hodges*, 576 U.S. 644 (2015).
Incorrect	Response contains factually inaccurate information	There is no right to same sex marriage in the United States.
Refusal	Model refuses to provide any answer or provides an irrelevant answer	I'm sorry, but I cannot answer that question. Please try a different query.
Groundedness		
Grounded	Key factual propositions make valid references to relevant legal documents	The right to same sex marriage is protected under the U.S. Constitution. *Obergfell v. Hodges*, 576 U.S. 644 (2015).
Misgrounded	Key factual propositions are cited but the source does not support the claim	The right to same sex marriage is protected under the U.S. Constitution. *Miranda v. Arizona*, 384 U.S. 436 (1966).
Ungrounded	Key factual propositions are not cited	The right to same sex marriage is protected under the U.S. Constitution.

[Magesh et al. 2024]

2.2 Beispiel: Halluzinationsraten im amerikanischen Recht

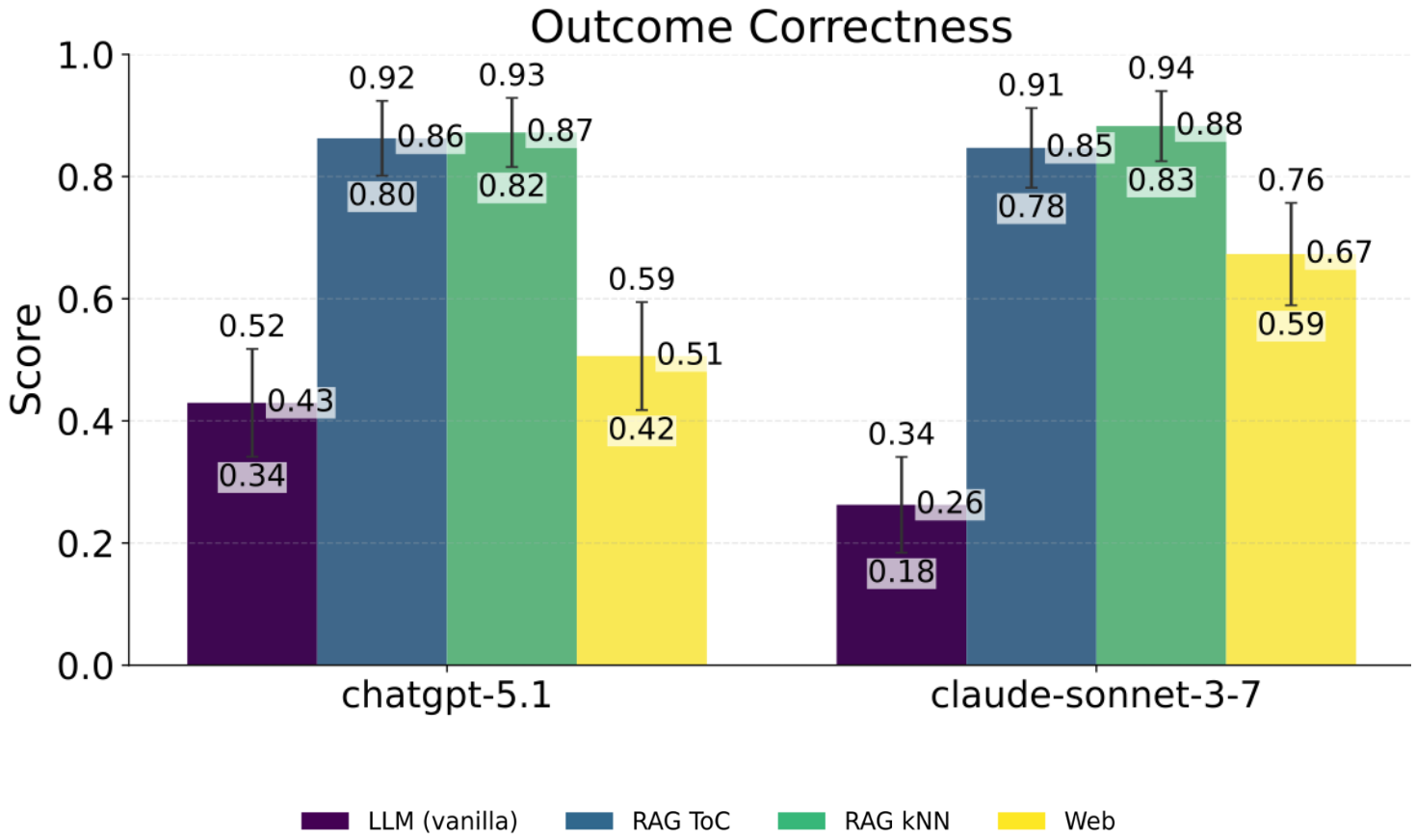
Halluzinationsrate auf mit GPT-Modellen synthetisierte Schriftsätze auf Basis realer Schriftsätze it Halluzinationen

Hallozination = Falscher Fallname oder Zitation



Testdatensatz von Liu et al. 2026 [link]

2.2 Beispiel Knowledge Cutoff & RAG Benchmark



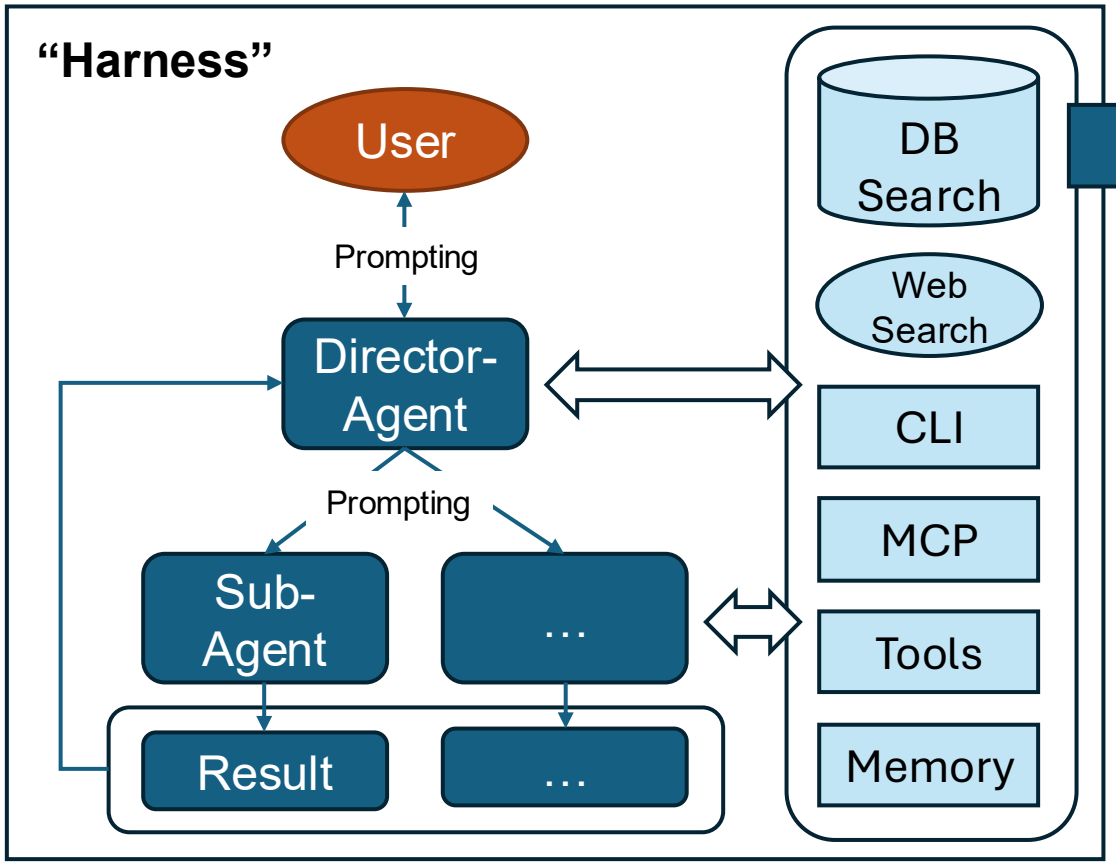
[Prior et al. 2026]

2.2. Beispiel: BenGER Reasoning Benchmark

	Benchathon			ZJS			Doctrinal Principles			
System / Group	n	Raw	Pass	n	Raw	Pass	n	Raw	Pass	Acc.
Gemini-3.1-Pro	15	68.3 ± 8.3	93%	579	61.8 ± 1.0	83%	531	83.2 ± 1.8	87%	83%
Opus-4.7	15	69.3 ± 5.5	93%	581	58.7 ± 1.1	76%	531	83.1 ± 2.0	86%	81%
GPT-5.4	15	66.6 ± 4.9	100%	581	60.4 ± 1.1	78%	531	80.3 ± 2.1	82%	79%
Sonnet-4.6	15	58.9 ± 5.9	73%	581	51.5 ± 1.0	50%	531	81.6 ± 2.1	83%	81%
Gemini-3.1-Flash-Lite	15	63.5 ± 5.7	87%	579	44.6 ± 0.8	26%	531	73.3 ± 2.3	76%	76%
GPT-5.4-mini	15	58.6 ± 5.6	73%	581	49.7 ± 0.9	46%	531	71.8 ± 2.4	72%	71%
DeepSeek-V4-Pro	15	56.1 ± 6.1	67%	581	50.1 ± 1.0	45%	531	73.5 ± 2.3	73%	73%
DeepSeek-V4-Flash	15	49.3 ± 3.7	47%	581	45.9 ± 0.8	30%	530	70.7 ± 2.4	73%	69%
Qwen3.5-122B	15	49.4 ± 4.9	40%	581	44.2 ± 0.8	25%	526	69.8 ± 2.4	75%	78%
Qwen3.6-35B	15	42.4 ± 7.6	27%	581	39.8 ± 0.8	13%	531	69.5 ± 2.5	69%	71%
Qwen3-235B	15	41.9 ± 4.3	20%	581	35.9 ± 0.7	6%	531	67.3 ± 2.5	69%	73%
Llama-4	15	37.8 ± 4.3	20%	581	36.1 ± 0.6	4%	531	70.1 ± 2.3	75%	77%
Human (all)	220	61.1 ± 2.1	77%	—	—	—	—	—	—	—
traditional	65	50.0 ± 3.6	52%	—	—	—	—	—	—	—
co-creation	155	65.7 ± 2.1	88%	—	—	—	—	—	—	—

[Nagl et al. 2026]

2.3. Vom Modell zu RAG, Agent & “Harness”



Optimierungsziele sind relative zum Harness und aller Komponenten zu verstehen!

Wissenschaft, z.B.:

LegalAgentBench: Evaluating LLM Agents in Legal Domain

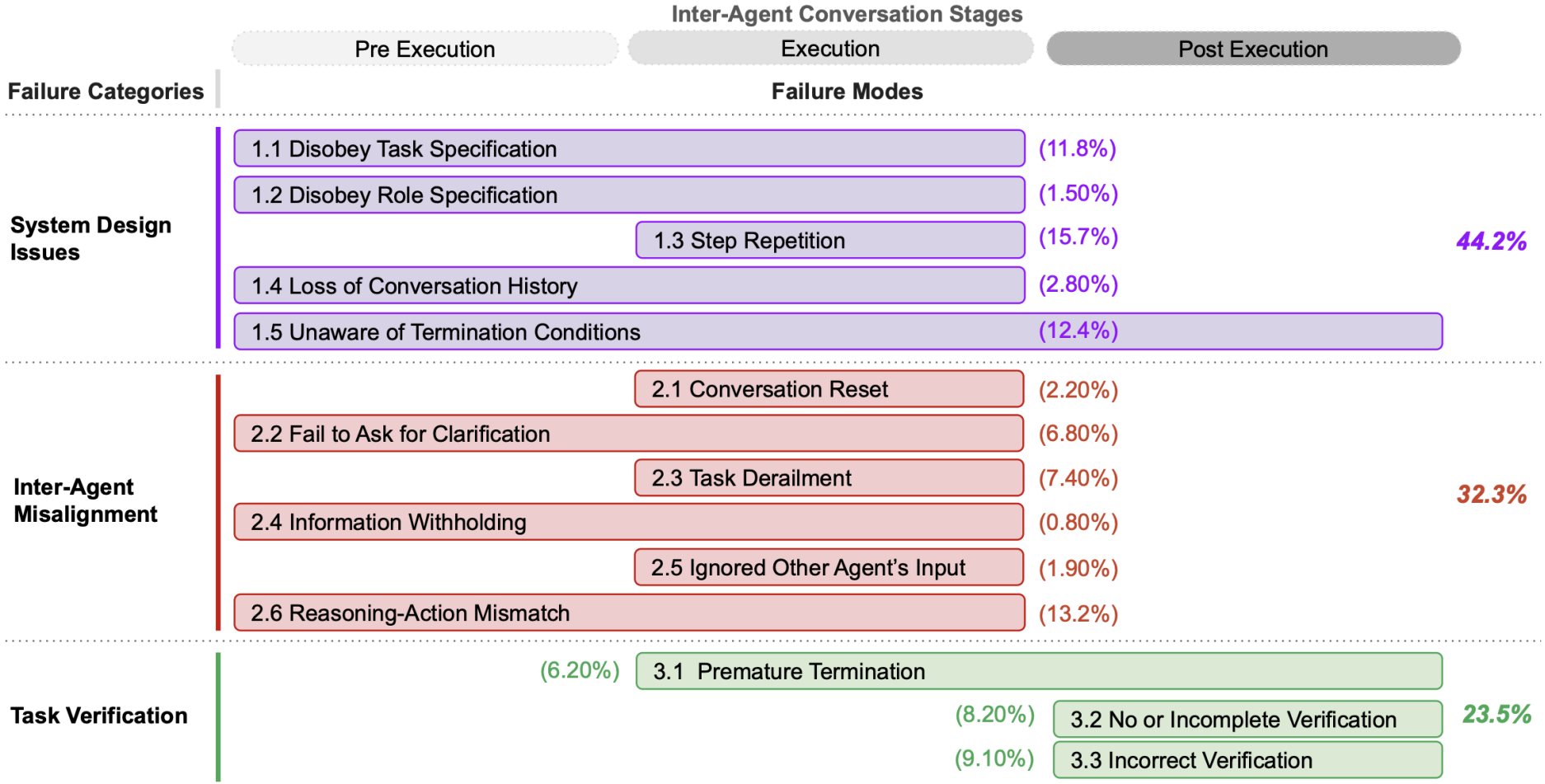
[Li et al. 2025]

Kommerzielle Anbieter:

Open-Sourcing Harvey’s Long Horizon Legal Agent Benchmark

[OpenAI]

Fehlermuster von Agentischen Systemen



[Cemri et al. 2025]

2.2 Technische Herausforderungen

- Halluzinationen bleiben trotz RAG und agentischen Systemen funktional ein Problem
- Subjektive Verbesserung und gemessene Raten können sich widersprechen
- Besondere Fehlerquellen, z.B. im Bereich der temporalen Relativität der Rechtslage in Modellparametern und Datenaugmentierung
- Benchmarking von Legal-Reasoning-Fähigkeiten von LLMs nur durch LLMs möglich, wodurch Varianz erhöht wird und neue Angriffspunkte entstehen
- Agentische Systeme sind mächtig, erhöhen jedoch Optimierungskomplexität und bringen neue Fehlerphänomene

3.1. Nutzung: aktueller Stand

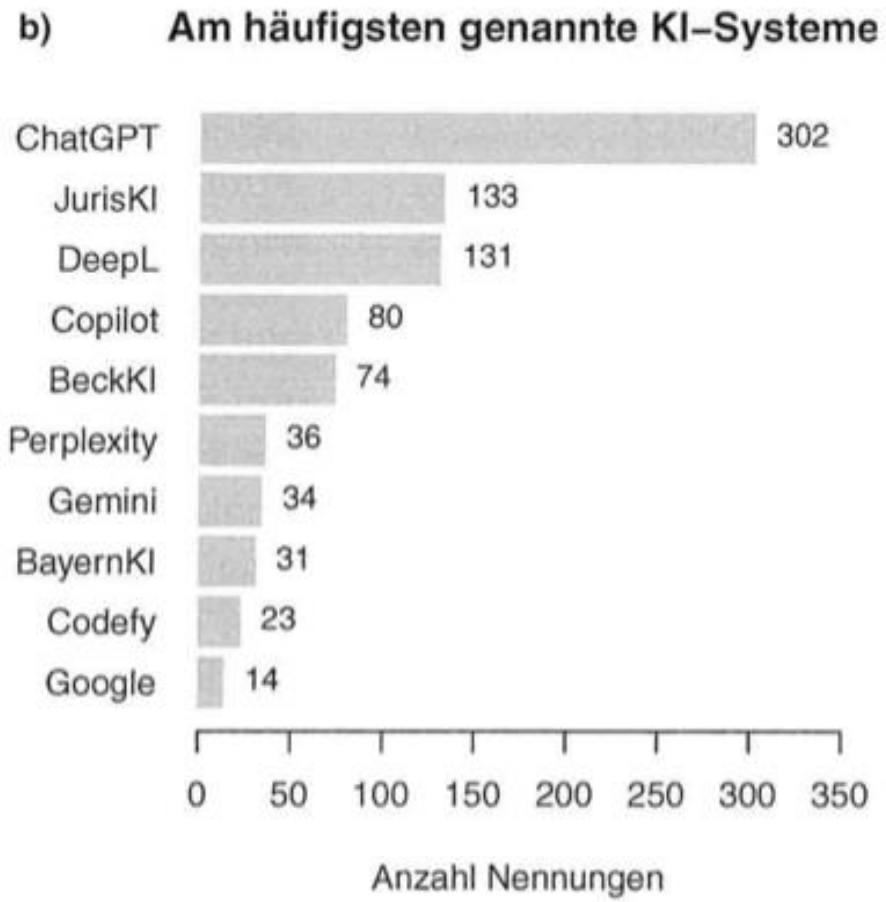
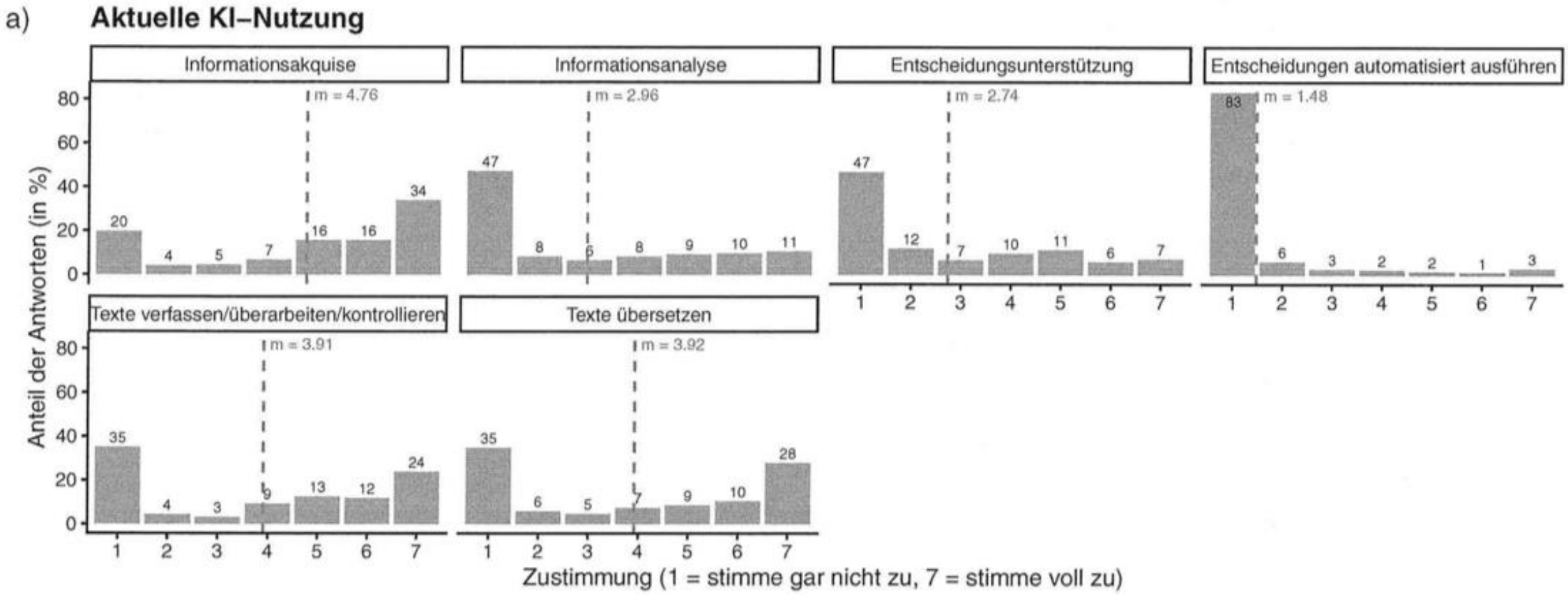


Abbildung 2: Welche der folgenden KI-Systeme benutzen Sie?

Quelle Caroei et al. MMR 2026, 619

3.2. Nutzung: aktueller Stand



Quelle Caroei et al. MMR 2026, 619

4. Konsequenzen für richterliche Entscheidungsfindung

- Sprachmodelle derzeit nur entscheidungs-*unterstützend* denkbar
- Zwingende menschl. Kontrolle & Aufsicht
- Transparenzerfordernisse & Rechtsgrundlage
- Datenschutz und KI-VO-Compliance
- Anforderungen anhängig von drei Faktoren:
  - funktionaler Charakter der Tätigkeit (admin. vs. inhaltl.)
  - Grad der Automatisierung
  - Komplexität des jew. Anwendungsfeldes

Gold/Reuß et al., KI in der richterlichen Entscheidungsfindung, 2025

5. Vom Modell zur Lösung

- **Modellauswahl ist eine anwendungsbezogene Entscheidung:**
  - Allgemeine und spezialisierte Modelle können je nach Anwendungsfall passend sein
  - Entscheidend ist das Zusammenspiel aus Modell, Quellenanbindung bzw. Retrieval, Prompts und Werkzeugen.
- **Quellenbindung ist keine Korrektheitsgarantie:**
  - RAG kann Fehler reduzieren, verhindert aber nicht alle Halluzinationen
  - Quelle, Aktualität, korrekte Wiedergabe und Schlussfolgerung müssen getrennt betrachtet werden.
- **Die Zusammenarbeit von Mensch und KI muss mitgedacht werden:**
  - Prüfbare Quellen, nachvollziehbare Verarbeitungsschritte, bearbeitbare Vorschläge und klar definierte Prüf- und Freigabeschritte

6. Von der Lösung zum Betrieb

- Modell-, Prompt- und Datenversionen dokumentieren
- RAG und Agentische KI erweitert die Integrations- und Sicherheitsaufgabe (z.B., begrenzte Berechtigungen, nachvollziehbare Aktionen)
- Qualität der gesamten Anwendung an repräsentativen juristischen Aufgaben evaluieren, im Betrieb überwachen und bei Änderungen erneut prüfen

Diskussion

Herausforderungen

- Aussagekräftige Bewertung von Modelleigenschaften
- Rechtskonformität erarbeiten & operationalisieren
- Integration in agiles Ökosystem Justiz

Handlungsempfehlungen

- Notwendige Rechtsgrundlagen schaffen
- Rechtskonformität unterstützendes Systemdesign
- Strukturierte Daten & Praxis-informierte Standards
- Infrastruktur & Rechenkapazität

Dr. **Michael Vössing**
IBM Deutschland GmbH

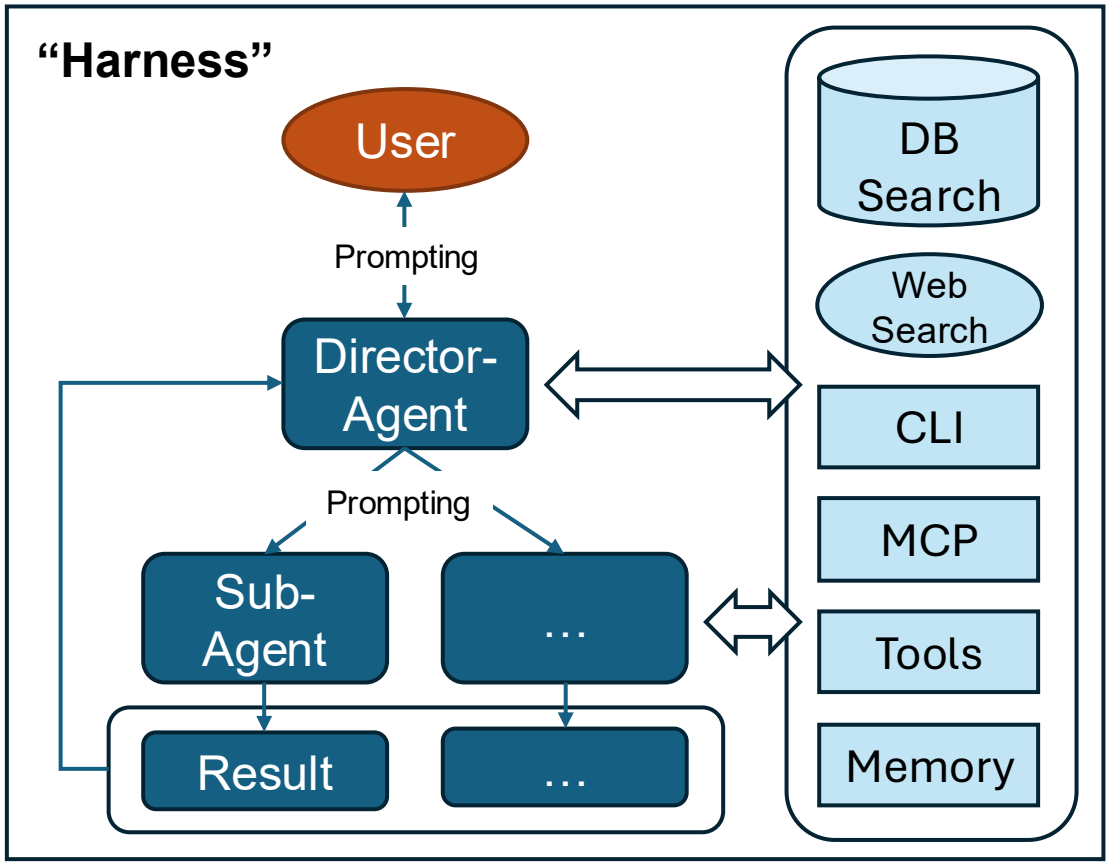
Prof. Dr. **Philipp M. Reuß**, MJur (Oxford)
Universität Göttingen

GOeJUSTICE
Göttingen Research Department for e-Justice, Automation and Digitalisation

Prof. Dr. **Matthias Grabmair**, LL.M.
TU München

Appendix

2.3. Vom Modell zu RAG, Agent & “Harness”



Wissenschaft, z.B.:

LegalAgentBench: Evaluating LLM Agents in Legal Domain

[Li et al. 2025]

LexRAG: Benchmarking Retrieval-Augmented Generation in Multi-Turn Legal Consultation Conversation

[Li et al. 2025]

AgentCourt: Simulating Court with Adversarial Evolvable Lawyer Agents

[Chen at al. 2025]

Kommerzielle Anbieter:

Open-Sourcing Harvey’s Long Horizon Legal Agent Benchmark [OpenAI]