Friedrich-Alexander-Universität
Erlangen-Nürnberg

Unser Prototyp und das Projekt des Bayerischen Justizministeriums ALeKs (Anonymisierungs- und Leitsatzerstellungs-Kit zur smarten Veröffentlichung von Gerichtsentscheidungen)

Konferenz der Informationsfreiheitsbeauftragten in Deutschland
Potsdam, 10.06.2026

Prof. Dr. Axel Adrian | Rechtstheorie und -gestaltung
Philipp Heinrich, M.Sc. | Korpus- und Computerlinguistik

FAU

Es gibt nicht die eine KI

Es gibt nicht die eine KI

Analogie: Gehirn und KI-Systeme

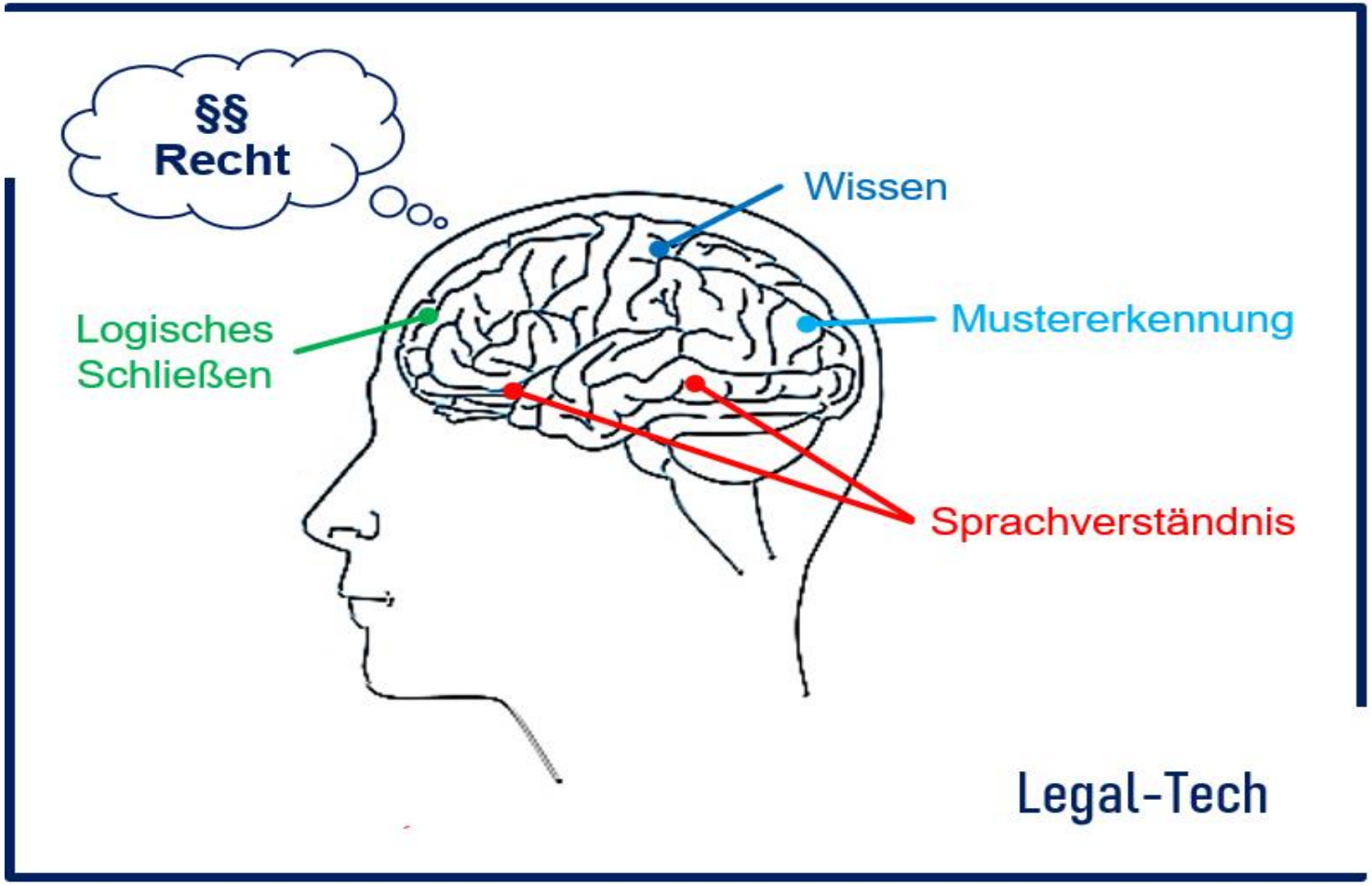
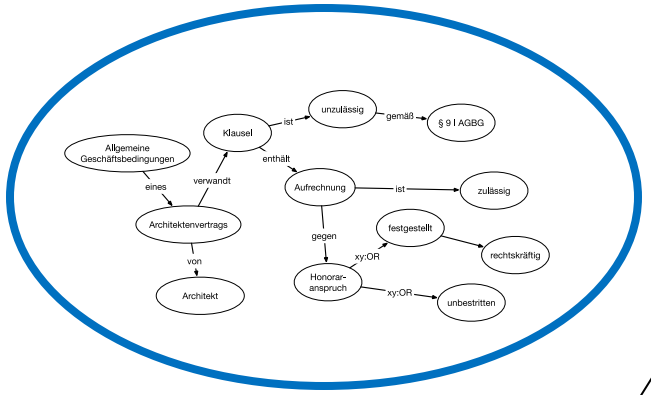
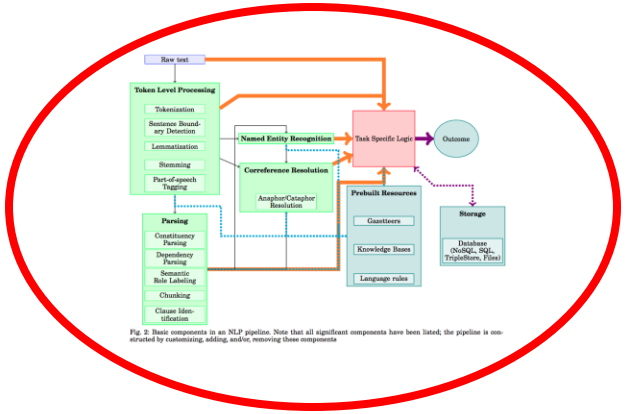


Abbildung: Eigene Darstellung.

Verschiedene technische Disziplinen

Für die Rechtswissenschaften relevante Teildisziplinen der KI-Forschung





Symbolische
und
subsymbolische
KI

Wissensrepräsentation

Natural Language Processing

Maschinelles Schließen

Mustererkennung (Machine-/Deeplearning)

(E ∧ B ∧ ¬R) → A
∀x [∃y. ∃z. Eyx ∧ Bzx ∧ ¬Rzxy → Ayxz]
O(E ∧ B ∧ ¬R) → OA
□E ∧ □(B ∧ ¬R) → A

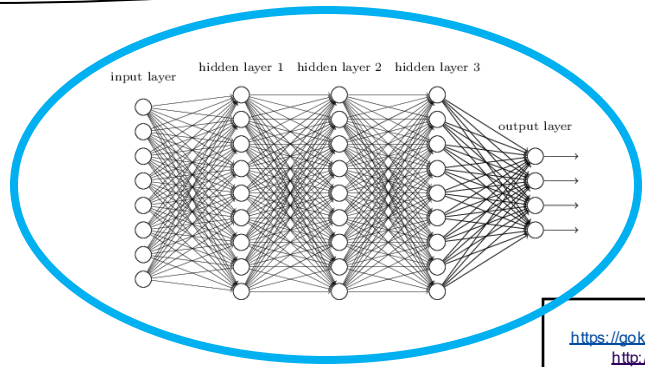


Abbildung:
https://gokulchittaranjan.files.wordpress.com/2015/09/pipeline1.png
http://neuralnetworksanddeeplearning.com/images/tikz36.png
Alexey Cheptsov (HLRZ Stuttgart)
Eigene Darstellung

Friedrich-Alexander-Universität
Erlangen-Nürnberg

FAU

Unser Forschungsteam

Unser Forschungsteam an der FAU

Drei Fakultäten, Principal Investigators, Wissenschaftliche Mitarbeiter

FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG
FACHBEREICH RECHTSWISSENSCHAFT

Prof. Dr. Axel Adrian
(Rechtswissenschaft und Wissenschaftstheorie)

Osman Anil Basaran, Ass. jur., LL.M. Legal Tech *(Regensburg)*
Michael Gritz, Notarassessor, Dr. Verena Stürmer, Notarassessorin, Dr. Helmut Birner, Notarassessor
Julia Mosebach, Referendarin, Julian Werner, Referendar,
Melanie Rosa, Referendarin
Daniel Odorfer, Referendar, Dr. Michael Keuchen

Legal

Lehrstuhl für Korpus- und Computerlinguistik

Prof. Dr. Stephanie Evert
(Natural Language Processing)

Philipp Heinrich, M.Sc.
Steffen Bothe, M.A.
Bao Minh Doan Dang, M.A.
Naveed Unjum, M.Sc.

CCL

kwarc

Prof. Dr. Michael Kohlhase
(Wissensrepräsentation)

Max Rapp, M.Sc. (phil.), Ms.Sc. (logic)
– Masteranden außerhalb des Projekts –
Stephan Prettner, M.Sc.
Sven Grünke, IT & Lab
Sebastian Wind, M.Sc.

KWARC

Pattern Recognition Lab

Prof. Dr. Andreas Maier
(Mustererkennung)

Florian Frank, M.Sc., Merlin Humml, M.Sc.,
Johannes Lindner (stud. B.Sc.), studentische Hilfskraft
Moritz Blöcher, B.Sc. (stud. M.Sc.), Masterand

ML

T.CS INF8 - THEORETIC COMPUTER SCIEN

Prof. Dr. Lutz Schröder
(Theoretische Informatik)

Prof. Dr. Christian Bergler (Mustererkennung)
Lucca Baumgärtner, B.Sc., stud (M.Sc.)
Aurelius Adrian, stud. (B.Sc.)

T.SC

Sonstige Beteiligte

Das AnGer-Projektteam

im Herbst 2024

Friedrich-Alexander-Universität
Erlangen-Nürnberg

FAU

Verordnung (EU) 2024/1689 (KI-VO)

Die Verordnung (EU) 2024/1689 (KI-VO)

KI-Systeme in der Justiz sind Hochrisiko-Systeme i.S.d. KI-VO

Bei Einsatz eines KI-Systems in der Justiz sind nach der KI-VO insbesondere die Vorgaben für KI-Hochrisikosysteme gemäß Anhang III Nr. 8 lit. a KI-VO zu beachten: „KI-Systeme, die bestimmungsgemäß von einer oder im Namen einer Justizbehörde verwendet werden sollen, um eine Justizbehörde bei der Ermittlung und Auslegung von Sachverhalten und Rechtsvorschriften und bei der Anwendung des Rechts auf konkrete Sachverhalte zu unterstützen, oder die auf ähnliche Weise für die alternative Streitbeilegung genutzt werden sollen“

Hochrisiko-KI-Systeme müssen entsprechend ihrer Zweckbestimmung eingesetzt werden und es gelten z.B. gem. Art. 3, 10, 15 KI-VO hohe Anforderungen für

- Transparenz,
- Datenqualität,
- Genauigkeit,
- Robustheit
- etc.

Anonymisierungs-/Pseudonymisierungs-KI-Systeme für gerichtliche Urteile, Dokumente oder Daten sind nach Erwg. 61 KI-VO: KEIN KI-Hochrisikosystem

Artikel 111 (Übergangsbestimmungen) und 113 KI-VO (schrittweises in Kraft treten):
Am 21. August 2024: Erlass der EU KI-VO.
Ab 2. Februar 2025:
* Kapitel I: Allgemeine Bestimmungen.
* Kapitel II: Verbotene Praktiken.
Ab 2. August 2025:
* Kapitel III: Hochrisiko-KI: Abschnitt 4: Notifizierende Behörden und notifizierte Stellen.
* Kapitel V: GPAI (generative KI). Ausnahme: Art. 111 KI-VO für unterschiedliche Integrations- und Komplexitätstiefen
* Kapitel VII: Governance.
* Kapitel XII: Sanktionen (ohne Art. 101: Geldbußen für Anbieter von GPAI).
* Kapitel IX: Beobachtung nach dem Inverkehrbringen, Informationsaustausch und Marktüberwachung: Abschnitt 3: Durchsetzung: Art. 78: Vertraulichkeit.
Ab 2. August 2027:
* Kapitel III: Hochrisiko-KI: Art. 6 Abs. 1 und entsprechende Pflichten.

Ignoriert man diese Normen, drohen Rechtsfolgen wie Bußgelder und Nutzungsverbote.

Die KI-VO

Datenqualität i.S.d. KI-VO

Die wichtigsten Anforderungen an die Datenqualität gem. Art 10 KI-VO sind:

- Relevanz, Repräsentativität, Fehlerfreiheit und Vollständigkeit der Daten (Was ist der Goldstandard)
- Minimierung von Verzerrungen (BIAS), um diskriminierende Ergebnisse zu vermeiden
- Dokumentation und Nachvollziehbarkeit der Datenquellen
- Einhaltung von Datenschutzgesetzen und Datensicherheit (Anonymisierung, realistische Maskierung)
- Kontinuierliche Überwachung und Anpassung im Lebenszyklus des KI-Systems

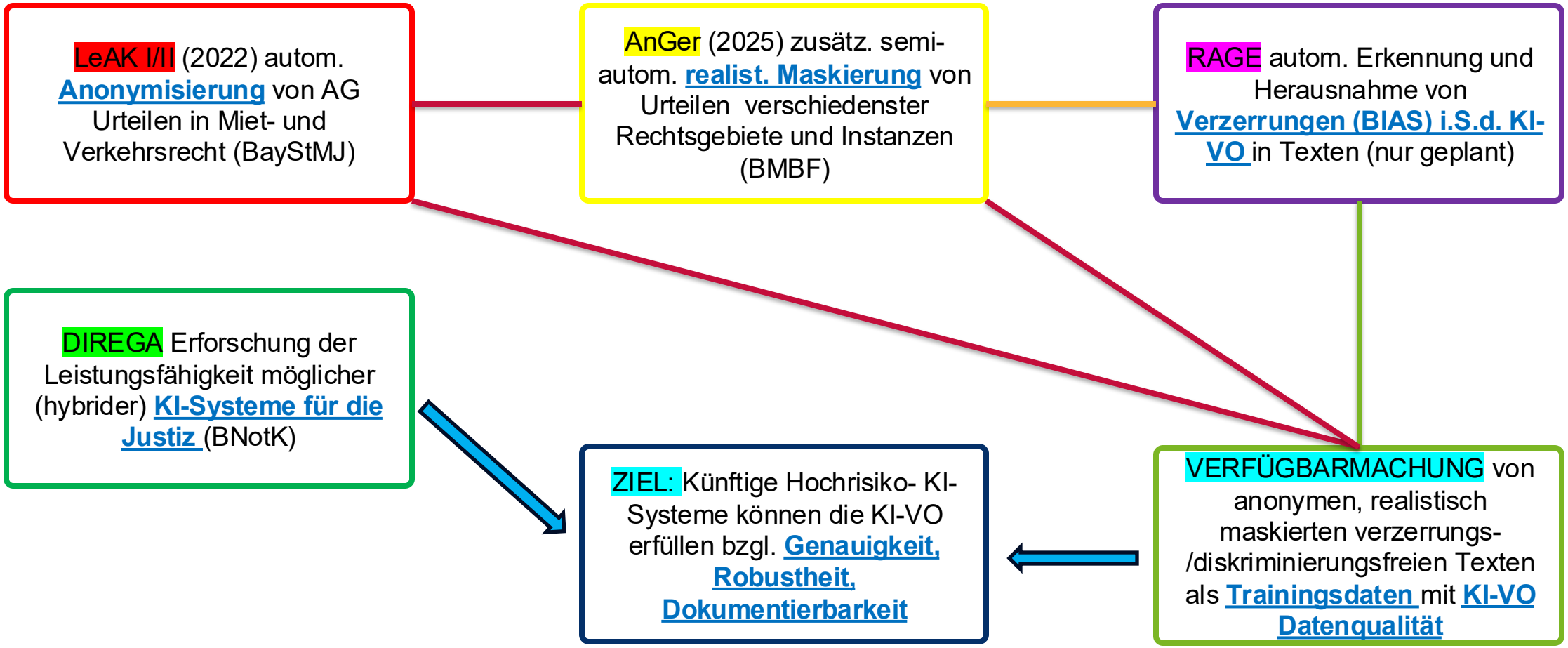
Friedrich-Alexander-Universität
Erlangen-Nürnberg

FAU

Unsere Forschungsprojekte

Unsere Forschungsprojekte an der FAU

bauen aufeinander auf



Friedrich-Alexander-Universität
Erlangen-Nürnberg

FAU

KI-Hochrisikosysteme in der Justiz

Die EU-KI-VO

Rechtsbegriffe müssen auch technisch ausgelegt werden

- „**Zweckbestimmung**“ (Art. 8 II, Art. 3 Nr. 12) verlangt, (nur) die Verwendung, für die ein KI-System laut Anbieter bestimmt ist

  Art 3 XII: „…„Zweckbestimmung“ die Verwendung, für die ein KI-System laut Anbieter bestimmt ist, einschließlich der besonderen Umstände und Bedingungen für die Verwendung, entsprechend den vom Anbieter bereitgestellten Informationen in den Betriebsanleitungen, im Werbe- oder Verkaufsmaterial und in diesbezüglichen Erklärungen sowie in der technischen Dokumentation;“

- „**Genauigkeit**“ verlangt Angabe und Transparenz von Genauigkeitskennzahlen

- „**Robustheit**“ verlangt Widerstandsfähigkeit gegenüber Risiken im Zusammenhang mit den Systemgrenzen

  Art 15 I, III: „ (1) Hochrisiko-KI-Systeme werden so konzipiert und entwickelt, dass sie ein angemessenes Maß an Genauigkeit, Robustheit und Cybersicherheit erreichen und in dieser Hinsicht während ihres gesamten Lebenszyklus beständig funktionieren… (3) Die Maße an Genauigkeit und die relevanten Genauigkeitsmetriken von Hochrisiko-KI-Systemen werden in den ihnen beigefügten Betriebsanleitungen angegeben.“

KI-Projekte in der Justiz

müssen an Hand der EU-KI-VO in spezifischer Weise beurteilt werden

z.B.

- Frauke
- OLGA
- JUKION (Codefy+KI)
- JANO
- PfüB-KI
- ….

Perplexity-Abfrage am 4.9.2025: „Für das KI-System FRAUKE werden in der verfügbaren Fachliteratur und den Berichten keine spezifischen exakten Metriken zur Messung der Genauigkeit (wie z. B. Accuracy, Precision oder Recall) öffentlich angegeben. Allerdings ist bekannt, dass die Bewertung der Genauigkeit bei KI-Systemen im Justizbereich typischerweise mit standardisierten Maschinenlern-Performance-Metriken erfolgt, wie sie etwa im Bereich KI-Sicherheit und bei anderen KI-Anwendungen verwendet werden“

- Künstliche Intelligenz auf dem Prüfstand: Anforderungen ... https://pmc.ncbi.nlm.nih.gov/articles/PMC12287159/
- Maschinelles Lernen – Metriken als Gradmesser für KI ... https://safe-intelligence.fraunhofer.de/artikel/maschinelles-lernen-metriken-als-gradmesser-fuer-ki-sicherheit-reicht-das-aus

Friedrich-Alexander-Universität
Erlangen-Nürnberg

FAU

Zweckbestimmung

Zweckbestimmung i.S.d. EU-KI-VO

am Beispiel unserer Anonymisierungsprojekte: Anonymisierung einer bestimmten Textsorte

KI-VO erfordert, ein KI-System nur zu dem Zweck einzusetzen, zu dem es laut Anbieter bestimmt ist:

- Unser KI-System (LeAK 2022) wurde entwickelt, um Texte automatisch zu anonymisieren
  Es geht nur um Anonymisierung = Erkennung datenschutzrelevanter Textstellen
- Unser KI-System (LeAK 2022) wurde nur auf
  Amtsgerichtsurteilen trainiert
- Die Trainingsdaten bestanden nur aus Urteilen zum
  Wohnraummietrecht und
  Verkehrsrecht

Nur das LeAK 2022-Modell wurde durch die Justiz in **ALeKS** integriert. Im Projekt „**50K**“ (https://www.lto.de/recht/justiz/j/anonymisierung-gerichtsentscheidungen-per-ki-bayern-aleks-anonymisierung) werden auch Urteile anderer Instanzen und Rechtsgebiete bearbeitet. Die Kompensation des Recall-Verlustes soll durch „human in the loop“ erfolgen.

Friedrich-Alexander-Universität
Erlangen-Nürnberg

FAU

Genauigkeit

Genauigkeit i.S.d. EU-KI-VO

Recall und Precision am Beispiel unserer Anonymisierungsprojekte

Art. 15 verlangt Angabe von **Genauigkeitskennzahlen** für ein eingesetztes KI-System

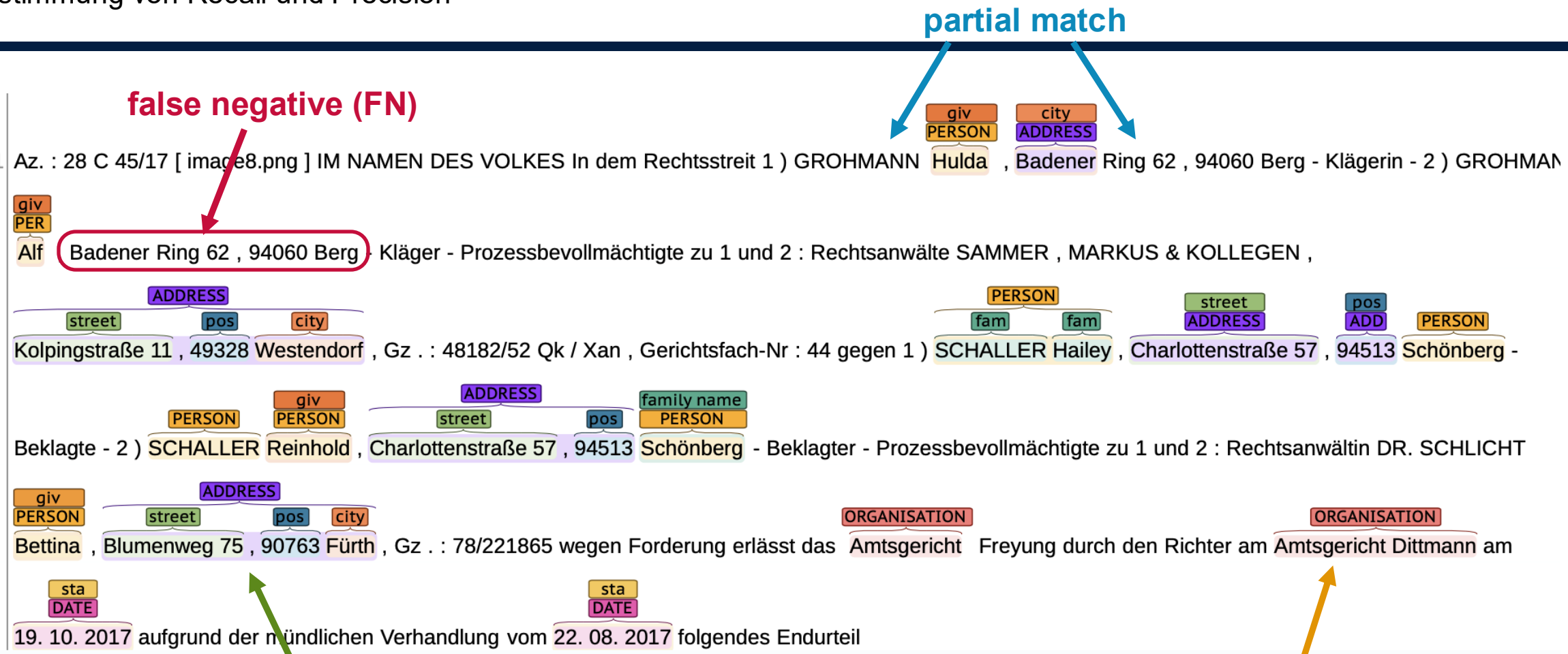
- Genauigkeit = Wie oft trifft das System richtige Entscheidungen?

Wie bestimmt man Genauigkeit? Was sind „richtige“ Entscheidungen?

- System zum Zweck der Anonymisierung:
  alle kritischen Textstellen müssen maskiert werden → **Recall**
  keine Textstellen sollen unnötig maskiert werden → **Precision**
- Berechnung durch Evaluation gegenüber manuell annotiertem **Goldstandard**
- Was soll gezählt werden: Token? ganze Textstellen? Texte mit Anonymisierungsfehlern?

Genauigkeit i.S.d. EU-KI-VO

Bestimmung von Recall und Precision



$$R = \frac{TP}{TP + FN} \qquad P = \frac{TP}{TP + FP}$$

https://mapa-demo.pangeamt.com/mapa/1.0/anonymization/model_showcase

Anforderungen an die Bestimmung von Genauigkeit

für ein System zur vollautomatischen Anonymisierung von hochsensiblen Texten

Anwendungsspezifische Anforderungen an die Evaluation:

- Evaluation auf Ebene **ganzer Textstellen** (ein nicht maskiertes Token genügt!)
- **Recall** wichtiger als Precision (→ F-Score kein nützliches Maß)
- anzustreben ist extrem hoher **Recall > 99%** für direkte Identifikatoren (PII)
  → immer noch mindestens 10.000 Anonymisierungsfehler in 1 Mio Urteile!
- Anforderungen an sonstige identifizierende Merkmale nicht offensichtlich, erforden eine Abschätzung des Deanonymisierungsrisikos (→ AnGer)
- höchste Anforderungen an **Qualität des Goldstandards**, um Recall > 99% zu messen (muss praktisch fehlerfrei sein ≠ computerlinguistische Goldstandards < 97%)
- Was ist mit teilweise erkannten Textstellen?
  Standard: strenge Evaluation akzeptiert nur exakt erkannte Textstellen
  CLUEval: entscheidend für Recall ist, ob alle kritischen Token einer Textstelle maskiert werden (auch wenn z.B. in zwei Textstellen zerlegt) – und analog für Precision

Evaluation: AG-Urteile im Miet- und Verkehrsrecht

LeAK 2022++ (im Rahmen von AnGer optimiertes LeAK-Modell)

Rechtsgebiet	Precision	Recall	Recall (PII)
Mietrecht	97.04%	96.05%	98.90%
Verkehrsrecht	97.41%	97.38%	99.11%
Gesamtergebnis	**97.25%**	**96.79%**	**99.03%**

- XLM-RoBERTa Large (Conneau et al. 2019) mit 500 M Parametern, vortrainiert auf 2,5 TB Webdaten
- Finetuning für die Erkennung von Textstellen (*span identification* per BIO-Tagging) über ~10 Epochen
- Lernrate optimiert mittels Grid Search (auf Development-Daten)
- Multitask-Modell zur Erkennung von Textstelle + Informationskategorie + Risikoniveau
- 50% Trainingsdaten, 25% Development-Daten, 25% Testdaten

Friedrich-Alexander-Universität
Erlangen-Nürnberg

FAU

Robustheit

Robustheit i.S.d. EU-KI-VO

Recall und Precision am Beispiel unserer Anonymisierungsprojekte

EU-KI-VO versteht technische **Robustheit** als Widerstandsfähigkeit gegenüber Risiken im Zusammenhang mit den Grenzen des Systems …

… was zum computerlinguistischen Verständnis des Begriffs passt:

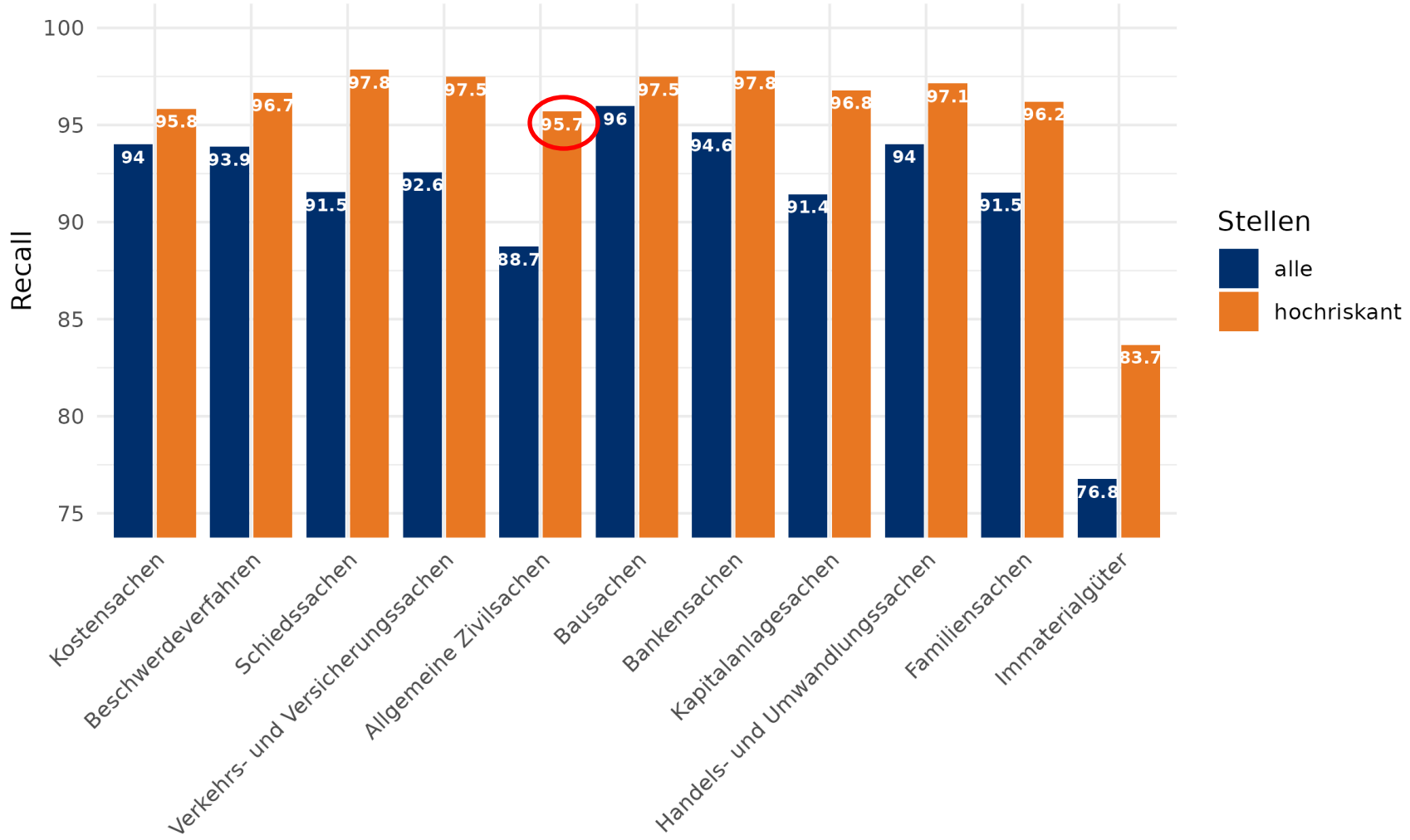
- Ein robustes System erzielt gleichbleibend gute Ergebnisse, auch wenn es auf Daten angewendet werden, die sich maßgeblich von seinen Trainingsdaten unterscheiden.
- messbar durch Evaluation über Textsorten und Domänen hinweg
- ggf. wird eine **Domänenanpassung** erforderlich

Hier: Evaluation des AG-Modells LeAK 2022++ auf **AnGer-Goldstandard**

- ca. 1500 Urteile des OLG München mit insg. ca. 5 M Token
- aus 11 verschiedenen Rechtsgebieten
- zwei Robustheitsdimensionen: Instanz + Rechtsgebiet

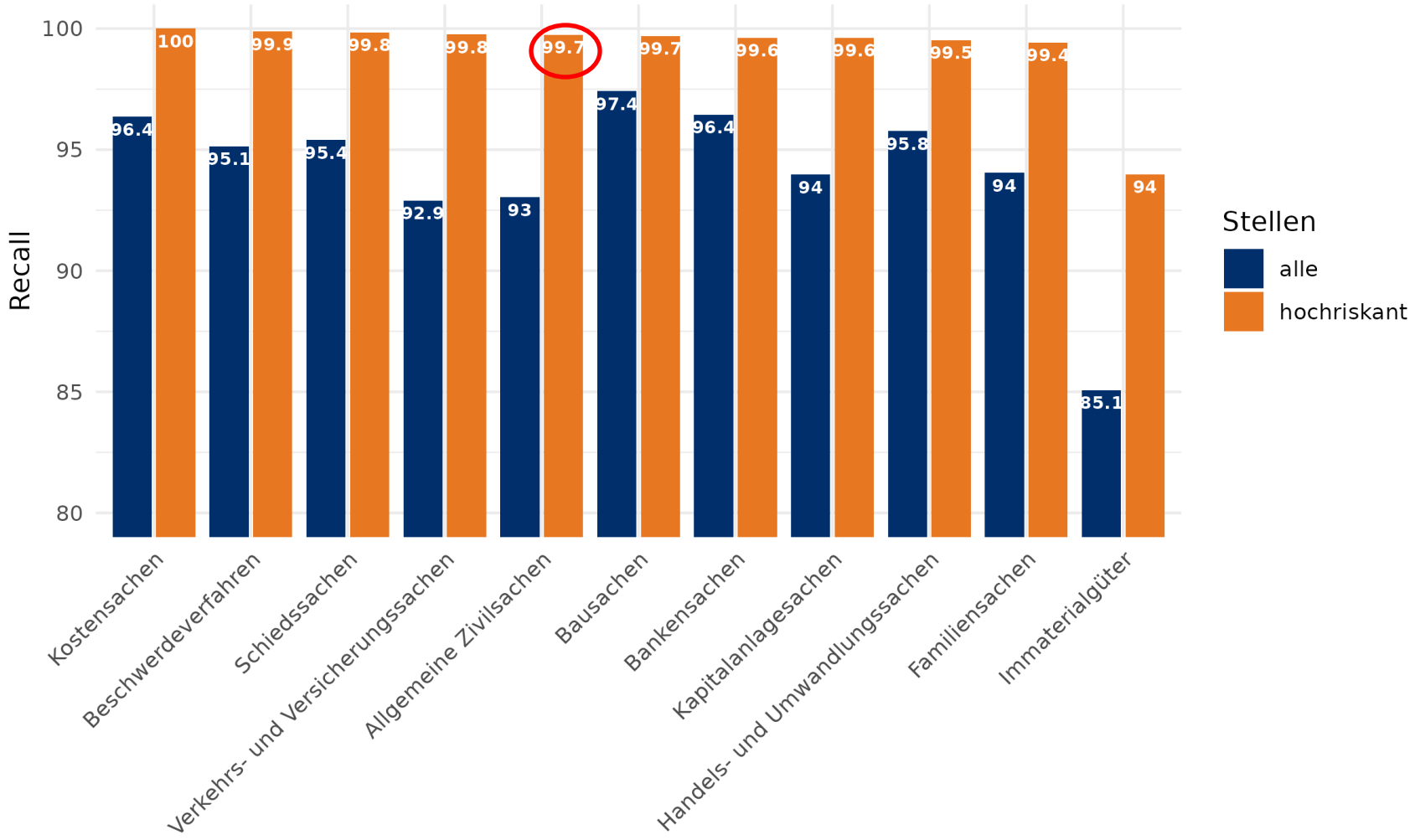
Robustheit: Evaluation von LeAK 2022++ auf OLG-Goldstandard

aufgeteilt nach Risikoniveau und Rechtsgebieten



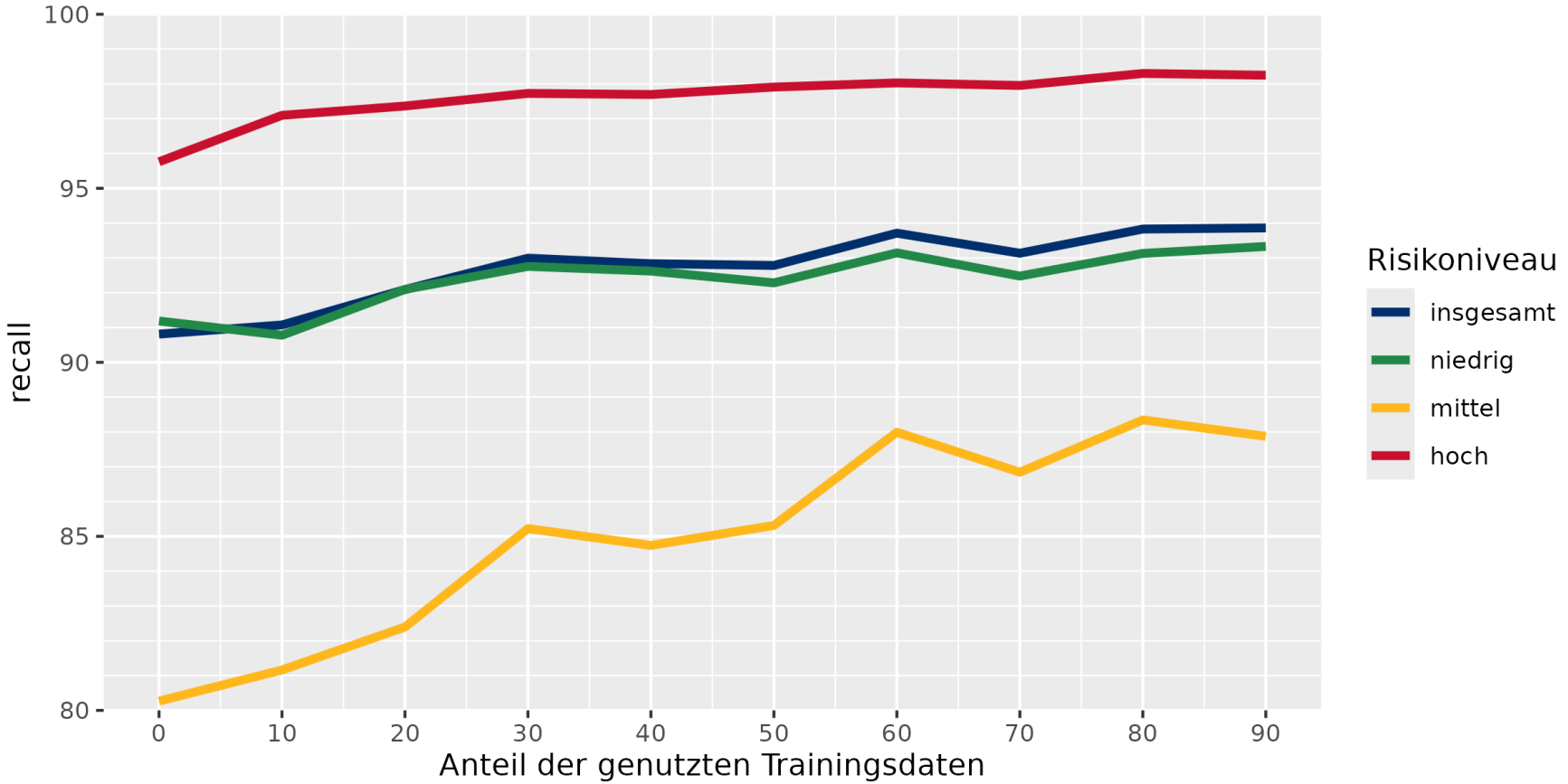
Robustheit: Domänenanpassung auf OLG → AnGer 2025

aufgeteilt nach Risikoniveau und Rechtsgebieten (+ manuelle Validierung)



Braucht man wirklich 5 M Token Goldstandard?

Lernkurven für Domänenanpassung von LeAK 2022++



Blick zurück: Evaluation von AnGer 2025 auf AG

Verringert die Domänenanpassung die Qualität des ursprünglichen AG-Modells?

Rechtsgebiet	Precision	Recall	Recall (PII)	AnGer 2025
Mietrecht	97.04%	96.05%	98.90%	99.23%
Verkehrsrecht	97.41%	97.38%	99.11%	99.34%
Gesamtergebnis	**97.25%**	**96.79%**	**99.03%**	**99.29%**

Robustheit

Hochrechnung des Qualitätsunterschieds für Hochrisikostellen

Teildatensatz „Allgemeine Zivilsachen“

845.015 Token, 236 Entscheidungen, 5.544 Hochrisikostellen

AnGer-Modell: erkennt **99,7 %** der Hochrisikostellen: nur **17 Stellen** werden übersehen
LeAK-Modell: erkennt **95,7 %** der Hochrisikostellen: **239 Stellen** werden übersehen

Zur Kompensation des Leistungsunterschieds müsste ein **Human in the Loop** zusätzlich **222 Hochrisikotextstellen** in 236 Entscheidungen manuell identifizieren

Hochrechnung auf 50.000 Entscheidungen

AnGer-Modell: trotz 99,7 % Recall verbleiben ca. **3.601 übersehene Hochrisikostellen**
LeAK-Modell (nicht zweckentsprechende Verwendung): ca. **50.635 übersehene Hochrisikostellen**

Erforderlicher manueller Mehraufwand: **47.033 zusätzliche Hochrisikotextstellen** müssten identifiziert und anonymisiert werden

Robustheit

Hochrechnung des Qualitätsunterschieds für alle Identifikatoren

Teildatensatz „Allgemeine Zivilsachen“

845.015 Token, 236 Entscheidungen, 19.805 Identifikatoren insgesamt

AnGer-Modell: erkennt **93 %** aller Identifikatoren: **1.387 Textstellen** werden übersehen
LeAK-Modell: erkennt **88,7 %** aller Identifikatoren: **2.238 Textstellen** werden übersehen

Zur Kompensation des Leistungsunterschieds müsste ein **Human in the Loop** zusätzlich **851 Textstellen** manuell identifizieren

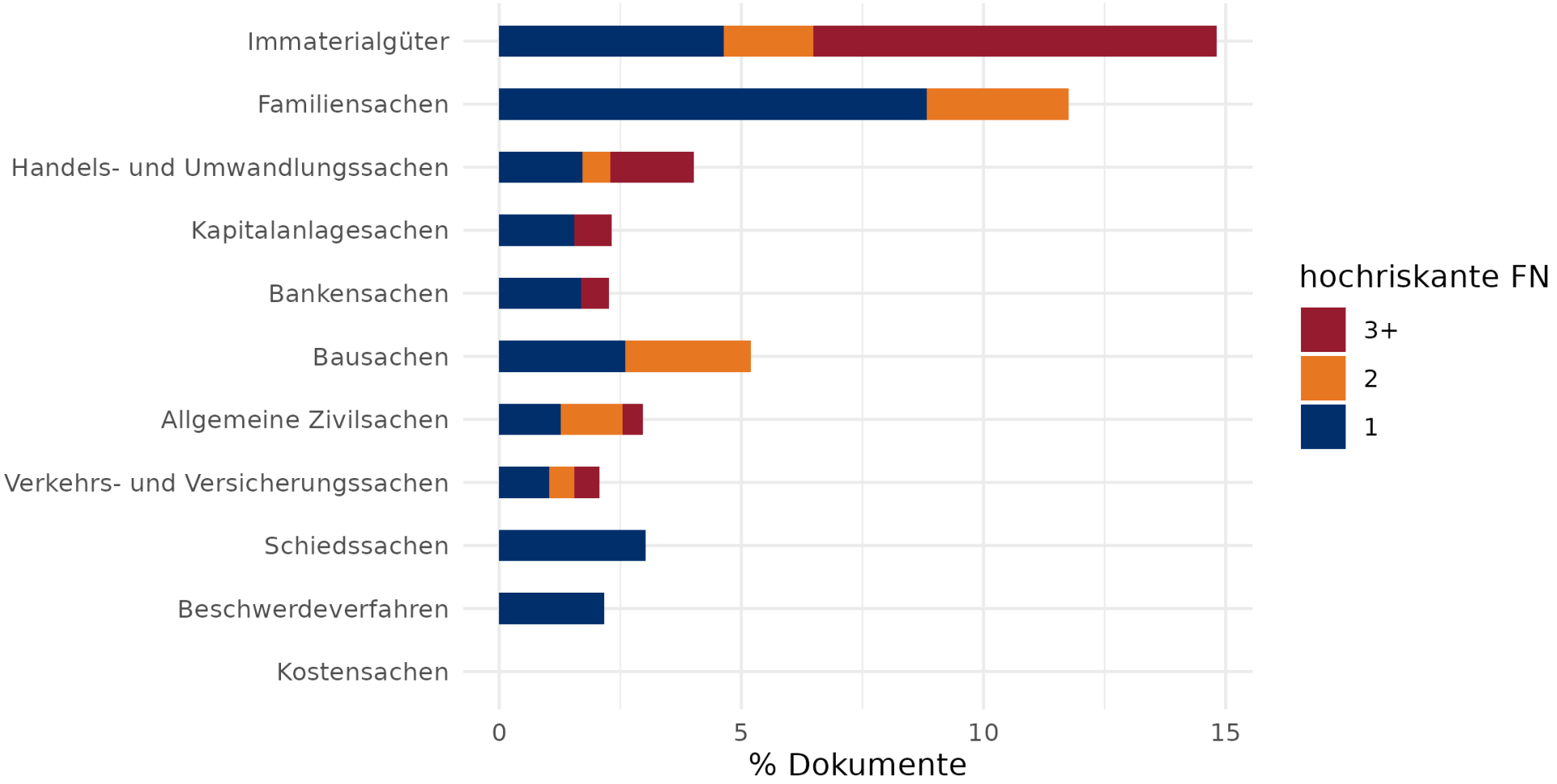
Hochrechnung auf 50.000 Entscheidungen

AnGer-Modell: bei 93 % Recall verbleiben ca. **293.719 übersehene Stellen**
LeAK-Modell (nicht zweckentsprechende Verwendung): ca. **474.153 übersehene Textstellen**

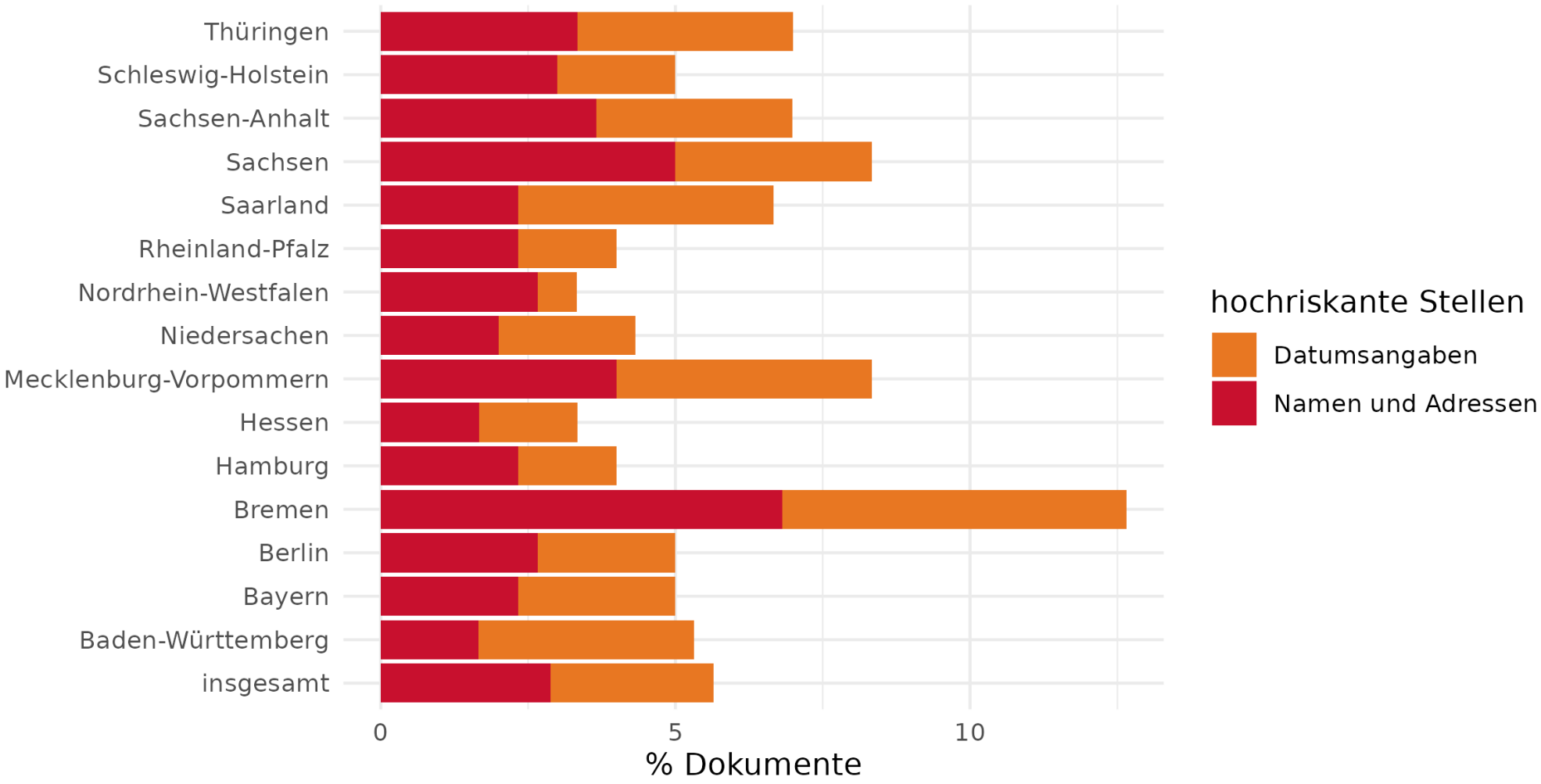
Erforderlicher manueller Mehraufwand: **180.434 zusätzliche Textstellen** müssten identifiziert und anonymisiert werden

Risikoabschätzung: Wie viele Urteile enthalten kritische Fehler?

Anzahl Urteile mit high-risk FN nach manueller Validierung



Vergleich mit der gängigen Praxis



Fazit

Fazit

Bevor man ein subsymbolisches KI-System für die Justiz auswählt, anschafft und einsetzt, sollte man aufgrund der KI-VO jedenfalls zumindest klären:

- Zu welchem Zweck wurde das System entwickelt? Was steht im Benutzerhandbuch?
- Welche Trainings-, Validierungs- und Testdatensätze wurden benutzt?
- Wurden diese aus Datenschutzgründen anonymisiert, realistisch maskiert, etc.?
- Waren diese frei von Verzerrungen (Bias) bzw. wurden bestehende Verzerrungen erkannt? Wie wurde das gemessen?
- Wie hoch ist die Genauigkeit? Welcher Goldstandard wurde wie erstellt? Recall? Precision?
- Wie robust ist das System, wenn es nicht auf typischen Anwendungsfällen eingesetzt wird? Wie verändern sich dann Recall und Precision?

Weiterführende Literatur

- Auslegung des KI-VO-E zur Evaluation von symbolischen Deduktionsverfahren der künstlichen Intelligenz, Axel Adrian, Michael Keuchen, Max Rapp, Alexander Steen, in Sprachmodelle: Juristische Papageien oder mehr?, Tagungsband des 27. Internationalen Rechtsinformatik Symposiums IRIS 2024, Hrsg. Erich Schweighofer/Stefan Eder/Federico Costatini/Felix Schmautzer/Jonas Pfister, Bern 2024, Seite 85 ff.

- Auslegung des KI-VO-E zur Evaluation von Verfahren der künstlichen Intelligenz am Beispiel der automatischen Anonymisierung von Gerichtsentscheidungen, Axel Adrian, Stephanie Evert, Philipp Heinrich, Michael Keuchen, in Sprachmodelle: Juristische Papageien oder mehr?, Tagungsband des 27. Internationalen Rechtsinformatik Symposiums IRIS 2024, Hrsg. Erich Schweighofer/Stefan Eder/Federico Costatini/Felix Schmautzer/Jonas Pfister, Bern 2024, Seite 205 ff.

- XAI – Erklärbare Künstliche Intelligenz in der Rechtswissenschaft, Axel Adrian, in Zeitschrift für die notarielle Beratungs- und Beurkundungspraxis, Heft 6/2024, Seite 201 ff.

- Robustheit und Domänenanpassung bei der automatischen Anonymisierung von Gerichtsentscheidungen, Axel Adrian/Stephanie Evert/Bao Minh Doan Dang/Philipp Heinrich/Mahdi Mantash/Daniel Odorfer/Melanie Rosa/Pei-Yu Shen/Julian Werner, in Künstliche Intelligenz und Recht, Heft 2/2025, S. 60 ff.

Vielen Dank
für Ihre Aufmerksamkeit!