A. Sachverhalt

I. Projekt Generatives Sprachmodell für die deutsche Justiz (GSJ-Projekt)

Gegenstand des Auftrags ist die Erstellung eines Gutachtens über datenschutz- und urheberrechtliche Fragen im Kontext der Entwicklung eines generativen Sprachmodells für die deutsche Justiz (***GSJ-Projekt***) mit einem Fokus auf konkrete Use-Cases.

Von der Begutachtung umfasst sind die folgenden Use-Cases betreffend die Richterschaft bzw. der Richterschaft zuarbeitende Personen[1], wobei die Use-Cases jeweils auf einer Eingabeaufforderung und einem Eingabekontext aus einer konkreten Akte aufbauen:

(1) die Erstellung einer tabellarischen Gegenüberstellung der Sachverhaltsdarstellungen in Zivilsachen (in der Praxis zunächst Miet- und Verkehrssachen) der Parteien mit einer Klassifikation in „strittig/unstrittig". Diese tabellarische Gegenüberstellung soll auf einen entsprechenden Wunsch des Benutzers hin auch relativ zu einer gewählten Anspruchsgrundlage kontextuell erfolgen können;
(2) die Vorformulierung des textuellen Sachverhaltsabschnitts eines Urteils auf der Basis der Streitstrukturierung, gegebenenfalls mit weiterem Input der Richter zur Beweiswürdigung;
(3) die Erstellung eines Zeitstrahls in der Anwendung, der sowohl Ereignisse des Sachverhalts als auch Prozesshandlungen umfasst.

Ein zentrales Forschungsziel des zu begutachtenden GSJ-Projekts ist die Prüfung, wie sich das Training eines Sprachmodells mit großen Mengen deutscher Gerichtsurteile und Aktenauszüge (hier vornehmlich Schriftsätze von Parteien, Gerichtsbeschlüsse) auf die Performanz des Sprachmodells in bestimmten Textgenerierungsaufgaben aus dem Justizalltag auswirkt. Zum Zweck der automatisierten Anonymisierung von Urteilen und Aktenauszügen steht im GSJ-Projekt ein von der Universität Erlangen-Nürnberg entwickeltes Software-Tool zur Verfügung (***Erlanger Tool***).

1 Aus Gründen der besseren Lesbarkeit wird auf die gleichzeitige Verwendung der Sprachformen männlich, weiblich und divers (m/w/d) verzichtet. Die verwendete Sprachform orientiert sich im Folgenden grundsätzlich an der im Gesetz vorgegebenen Sprachform.

Die im Rahmen des GSJ-Projekts verwendeten Gerichtsentscheidungen und Aktenauszüge entstammen insbesondere dem Miet- und Verkehrsrecht. Demgegenüber sind von der Verwendung ausgeschlossen – insbesondere – sensible Verfahrenskategorien (z.B. Strafrecht, Arzthaftungssachen und Verfahren vor den Familiengerichten).

Das GSJ-Projekt wird durch das Bayerische Staatsministerium der Justiz und durch das Ministerium der Justiz des Landes Nordrhein-Westfalen (zusammen ***Ministerien***) geleitet. Die Professur für Legal Tech von Herrn Professor Matthias Grabmair an der TUM School of Computation, Information and Technology führt in Zusammenarbeit mit dem Lehrstuhl für Bürgerliches Recht, Handels- und Gesellschaftsrecht, Arbeitsrecht und Europäische Privatrechtsentwicklung an der Universität Köln von Frau Professorin Barbara Dauner-Lieb (zusammen: ***ausführende Stellen***) im Auftrag der Ministerien das Training des Sprachmodells durch, um den anschließenden Einsatz des trainierten Sprachmodells durch die Ministerien bzw. die Justiz zu ermöglichen. Im Rahmen des GSJ-Projekts übernehmen die ausführenden Stellen eigenständig die technische Umsetzung und führen nach eigenem Ermessen gegebenenfalls das Debugging oder Testläufe durch. Die ausführenden Stellen verwenden die erhaltenen Datensätze nur nach Weisungen der Ministerien und insbesondere nicht für eigene Forschungszwecke.

Im Rahmen des GSJ-Projekts anonymisieren die Ministerien zunächst mittels des Erlanger Tools die Gerichtsentscheidungen sowie Aktenauszüge eigenständig oder durch die ausführenden Stellen als Auftragnehmer. Die (teil-)anonymisierten Gerichtsentscheidungen und Aktenauszüge werden sodann für das weitere KI-Training verwendet.

II. Tatsächliche Annahmen – Thematische Aus- und Eingrenzungen

Der rechtlichen Begutachtung liegen die folgenden tatsächlichen Annahmen sowie Ein- und Ausgrenzungen des Untersuchungsgegenstandes zugrunde.

Nicht zum Gegenstand der Untersuchung werden insbesondere gemacht:

- die technische Prüfung der Anonymisierungsanforderungen, auch und gerade im Hinblick auf den Einsatz des Erlanger Tools. Die Untersuchung dient vielmehr der Herausarbeitung der rechtlichen Maßstäbe und der Überprüfung der Einhaltung rechtlicher Anforderungen auf der

Grundlage technischer Annahmen und zur Verfügung gestellter Sachverhaltsinformationen;
- die Prüfung der Use-Cases anhand der KI-VO.

Das Erlanger Tool erzielt:

- mit Blick auf Gerichtsentscheidungen einen Recall, d.h. einen Anteil erfolgreich anonymisierter Merkmale von allen zu anonymisierenden Merkmalen eines Texts innerhalb einer Kategorie, in einem Spektrum von 94-96 % für eindeutig identifizierende Merkmale bzw. Kennungen, wie etwa Namen und Anschriften natürlicher Personen;[2]
- mit Blick auf die übrigen Informationskategorien (z.B. Daten zum Prozessablauf, Aktenzeichen und Gerichtsort) einen Recall in einem Spektrum von 68-96 %;[3]
- mit Blick auf Aktenauszüge vergleichbare Recall-Werte, wobei sich aufgrund der abweichenden Struktur und aufgrund von Digitalisierungsdefiziten (z.B. betreffend handschriftlicher Anmerkungen) geringere Recall-Werte ergeben können.

Signifikante Abweichungen bei den jeweiligen Recall-Werten, die sich im Laufe der weiteren Evaluation des Erlanger Tools ergeben, könnten gegebenenfalls neue Beurteilungen erfordern.

Es wird unterstellt, dass:

- den Ministerien die Gerichtsentscheidungen und Aktenauszüge rechtmäßig von den Gerichten übermittelt worden sind;
- die Zusammenarbeit der Ministerien im Allgemeinen rechtlich zulässig ist.

Die Verordnung (EU) 2025/327 über den europäischen Gesundheitsdatenraum, die beispielsweise Patientenkurzakten umfasst, bleibt für die Begutachtung außer Betracht.

2 *Adrian et al.*, in: Adrian/Kohlhase/Evert/Zwickel, Manuelle und automatische Anonymisierung von Urteilen, S. 188–189, 194, 196 ff.

3 *Adrian et al.*, in: Adrian/Kohlhase/Evert/Zwickel, Manuelle und automatische Anonymisierung von Urteilen, S. 188–189, 194, 196 ff.

III. Technische Grundzüge des Sprachmodells und späteren KI-Systems

Die technischen Grundzüge eines Sprachmodells (auch: Large Language Model oder KI-Modell) auf Basis neuronaler Netze wurden bereits vielfach dargestellt.[4] Diese Untersuchung beschränkt sich daher zunächst auf eine Kurzdarstellung der wesentlichen Grundlagen und Begrifflichkeiten, die, wo dies erforderlich und angezeigt ist, im Laufe der Begutachtung vertieft wird.

Das Sprachmodell wird auf Grundlage eines Trainingskorpus an vorbearbeiteten Ausgangstexten trainiert.[5] Diese Vorbearbeitung umfasst mehrere Schritte, in deren Rahmen es zur Erstellung weiterer digitaler Kopien der Ausgangstexte und Änderungen an den Ausgangstexten und deren Format kommt.[6] Sodann werden im Rahmen des Trainings im weit verstandenen Sinne[7] allfällige Zeichenketten (sog. Token, ähnlich den Silben eines Wortes)[8] und deren typische, wahrscheinliche Beziehung zueinander ermittelt und in eine mathematische Repräsentation überführt. Mit dieser mathematischen Repräsentation des Modells werden die sog. Gewichte oder Parameter auf mehreren Ebenen (sog. layers) bezeichnet.[9] In einem Modell sind somit wahrscheinlichkeitsbasierte, mathematische Repräsentationen verschiedener Zeichenketten gespeichert.

Der Einsatz eines Sprachmodells[10] erfordert eine Schnittstelle (z.B. in Verbindung mit einer Benutzeroberfläche), die ein Teil des sog. KI-Systems[11] ist. Das KI-System nimmt Eingabeaufforderungen (sog. Prompts) einschließlich eines Eingabekontexts (z.B. einer Verfahrensakte) entgegen, stellt einen einleitenden Befehl voran (den sog. System Prompt), bringt die

4 *Dornis/Stober*, Urheberrecht und Training generativer KI-Modelle, S. 23 ff.; *Pesch/Böhme*, MMR 2023, 917 (918 f.); *T. Radtke*, ZGE 17 (2025), 1 (5 ff.).

5 Im Überblick zum Training *International Working Group on Data Protection in Technology*, Working Paper on Large Language Models (LLMs), S. 11 ff.; etwa *Becher/Berkhin/Freeman*, in: Ramakrishnan/Stolfo/Bayardo/Parsa, S. 424.

6 *Shalev-Shwartz/Ben-David*, Understanding machine learning, S. 228 ff.

7 D.h. einschließlich eines Pre-Trainings und Fine-Tunings, s. z.B. *Devlin et al.*, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

8 Z.B. „B-uch-staben-ket-ten" nach dem Tokenizer für ChatGPT-4o, https://platform.openai.com/tokenizer.

9 *Naveed et al.*, A Comprehensive Overview of Large Language Models, S. 2.

10 S. auch grundlegend zu Mechanismen der sog. Transformer-Architektur *Vaswani et al.*, Attention Is All You Need.

11 S. hierzu auch die Definition eines KI-Systems in Art. 3 Nr. 1 KI-VO sowie eines KI-Modells in Art. 3 Nr. 63 KI-VO. Danach zeichnet ein Modell insbesondere aus, dass es „in eine Vielzahl nachgelagerter Systeme oder Anwendungen integriert werden kann".

gesamte Eingabe in ein geeignetes Format und übergibt die umgewandelten Daten an das Sprachmodell.

Das Sprachmodell führt auf Grundlage der Eingabe und der trainierten Gewichte als Multiplikatoren einzelner, numerisch repräsentierter Aspekte der Eingabe verschiedene Berechnungen durch und liefert sodann über das KI-System eine wahrscheinlichkeitsbasierte Textausgabe zurück. Abhängig von verschiedenen Faktoren einschließlich der sog. Temperatur-Einstellung kann die Ausgabe kreativer ausgestaltet werden, indem für die generierte Ausgabe nicht nur die jeweils wahrscheinlichsten Zeichenketten aneinandergereiht, sondern zufällig Zeichenketten mit geringerer Wahrscheinlichkeit genutzt werden. Diese Ausgaben können teilweise in Abhängigkeit von der Eingabeaufforderung mit Auszügen aus dem Trainingskorpus übereinstimmen. Ferner können die Ausgaben auf Basis der Trainings- und Eingabedaten auch zu unzutreffenden Aussagen führen (sog. Halluzinationen).[12]

Sowohl auf Ebene der Eingabe als auch der Ausgabe können durch das KI-System zahlreiche Beschränkungen vorgesehen werden. Beispielsweise kann die Eingabe derart beschränkt werden, dass der Nutzer des KI-Systems nur eine Auswahl unter mehreren Eingabeaufforderungen treffen kann. Ein- und Ausgabe können außerdem durch einfache oder fortschrittliche Filter (vor-)bearbeitet werden,[13] bevor die Eingabe sodann an das Sprachmodell bzw. die Ausgabe an den Nutzer übergeben wird.

12 Etwa *Wachter/Mittelstadt/Russell*, R Soc Open Sci 11 (2024), 240197; *Seemann*, Künstliche Intelligenz, Large Language Models, ChatGPT und die Arbeitswelt der Zukunft, S. 28.

13 Z.B., indem (reguläre) Ausdrücke ersetzt werden oder die Ausgabe durch ein weiteres KI-System auf das Vorliegen identifizierender Merkmale überprüft wird, hierzu *Moos*, CR 2024, 442 (450).