C. Executive Summary

Das GSJ-Projekt kann **im Einklang mit den datenschutz- und urheberrechtlichen Anforderungen** realisiert werden.

Eine rechtskonforme Realisierung setzt insbesondere voraus: (1) den Einsatz des **Erlanger Tools**, (2) die Prüfung des Anonymisierungsstandards **älterer Gerichtsentscheidungen**, (3) eine grundsätzlich nur **justizinterne Nutzung**, wobei eine **Veröffentlichung** eines KI-Systems (i.e. nicht des Sprachmodells) **möglich** bleibt, (4) den Abschluss einer **Auftragsverarbeitungsvereinbarung** zwischen den Ministerien als gemeinsam Verantwortliche und den ausführenden Stellen, (5) die weitere **Clusterung** von Entscheidungen und Aktenauszügen sowie (6) die Prüfung von **Nutzungsvorbehalten** in den Aktenauszügen oder die Arbeit mit **abgeleiteten Textformaten**. Datenschutzrechtlich **nicht vorausgesetzt ist eine händische Anonymisierung**, sofern aufgrund ergriffener Maßnahmen (s. unter F5.a) das Sprachmodell als anonym zu bewerten ist.

Die **datenschutzrechtliche Zulässigkeit** stützt sich insbesondere auf den **fehlenden Personenbezug des Sprachmodells**, sofern hinreichende technische und organisatorische Maßnahmen einschließlich des Einsatzes des Erlanger Tools implementiert werden. Die **urheberrechtliche Zulässigkeit** ergibt sich insbesondere aus den Schranken der **§§ 44a, 44b UrhG, alternativ aus dem Rückgriff auf abgeleitete Textformate**.

F1.a: Welche Anforderungen sind an die Anonymisierung von unveröffentlichten Urteilen zu stellen, damit sie für die relevanten Akteure als nicht-personenbezogene Daten eingestuft werden können?

- Für die Frage des Vorliegens personenbezogener Daten ist grundsätzlich ein **relatives Konzept der Identifizierbarkeit** anzuwenden. Während **Kennungen** (z.B. der Name oder eine eindeutige Anschrift) stets zu einem Personenbezug führen, sind sonstige identifizierende Merkmale (z.B. Orts- und sonstige Sachverhaltsangaben) darauf zu prüfen, ob die Merkmale nach **allgemeinem Ermessen wahrscheinlich zur Identifizierung** natürlicher Personen genutzt werden. In diese Prognoseentscheidung sind z.B. die Kosten der Identifizierung, der Empfängerkreis und grundsätzlich auch die rechtliche Zulässigkeit einzustellen (s. dazu unter D.I.1).

- **Publikationswürdige Gerichtsentscheidungen** können – gestützt auf Art. 6 Abs. 1 UAbs. 1 lit. e DSGVO i.V.m. Art. 4 Abs. 1 BayDSG bzw. § 3 Abs. 1 DSG NRW, Art. 6 Abs. 1 UAbs. 1 lit. f DSGVO oder Art. 6 Abs. 4 DSGVO sowie wohl auch gestützt auf Art. 9 Abs. 2 lit. f DSGVO – selbst dann rechtskonform veröffentlicht werden, wenn sie **teilweise personenbezogene Daten** enthalten (z.B. mit Blick auf Personen der Zeitgeschichte und identifizierende mediale Berichterstattung). Die Rechtmäßigkeit einer solchen Veröffentlichung setzt insbesondere voraus, dass Kennungen durch geeignete Pseudonyme (i.e. nicht Initialen) ersetzt und sensible Daten möglichst entfernt werden (siehe dazu unter D.II.5.b). Sofern diese Voraussetzungen vorliegen (siehe dazu unter F1.b und F2.a am Beispiel von Nordrhein-Westfalen), können die anonymisierten Urteile **grundsätzlich auch für das Training des GSJ-Sprachmodells weiterverarbeitet** werden (siehe dazu insbesondere unter D.V.4.c.bb).
- Im Übrigen sind unveröffentlichte Gerichtsentscheidungen anhand unterschiedlicher Merkmale zu **clustern** und auf Wort-/Merkmalsebene zuverlässig zu **anonymisieren** (z.B. Überprüfung von Entscheidungen in Arzthaftungssachen oder Familiensachen und Entscheidungen mit einem besonders umfangreichen Tatbestand und einer Vielzahl beteiligter Personen; siehe dazu unter D.I.3). Diese Anforderungen sind **nicht zwingend zu erfüllen**, wenn und soweit das Sprachmodell und das konkrete Einsatzszenario auch ohne vollständige Anonymisierung der Gerichtsentscheidungen zulässig sind (s. hierzu auch unter F3, F5.a).

F1.b und F2.a: Kann dabei auf die von der Rechtsprechung entwickelten Maßstäbe zur Veröffentlichung von Gerichtsentscheidungen abgestellt werden? Sollte die zweite Frage unter F1 mit nein beantwortet werden: Gelten veröffentlichte gerichtliche Entscheidungen, die vor ihrer Veröffentlichung nach den von der Rechtsprechung entwickelten Maßstäben bearbeitet wurden als nicht-personenbezogene Daten, wenn sie für das Training eines Sprachmodells verwendet werden?

- Die Maßstäbe der **deutschen Justiz** betreffend die Entscheidungsanonymisierung sind in Ansehung des Anwendungsvorrangs der DSGVO **nicht maßgeblich**, können allerdings gleichwohl Anhaltspunkte für die Anonymisierung bieten. Denn im Hinblick auf **publikationswürdige Gerichtsentscheidungen** genügt ein geringerer Anonymisierungsstandard entsprechend den Maßstäben der deutschen Rechtsprechung und die Weiterverarbeitung zu Zwecken des GSJ-Projekts ist grundsätzlich **datenschutzrechtlich zulässig** (siehe dazu unter D.II.5.b und D.V.4.c.bb).

- Die **deutsche Praxis** der Entscheidungsanonymisierung wird in den verschiedenen Gerichtssprengeln unterschiedlich vorgenommen. Jüngere Anonymisierungsrichtlinien **genügen** bei sorgfältiger Umsetzung im Einzelfall regelmäßig insgesamt den datenschutzrechtlichen Anforderungen (z.B. die untersuchte Verwaltungsvorschrift der Justiz Nordrhein-Westfalens aus dem Jahr 2021). Demgegenüber bestehen im Hinblick auf **ältere Entscheidungen** und Richtlinien, die insbesondere Kennungen durch Initialen unzureichend anonymisieren, Bedenken (siehe dazu unter D.II.2-4). Im Übrigen besteht insoweit auch **keine Fiktionswirkung** (siehe dazu unter D.II.5).

F2.b: Falls die Frage wiederum mit nein beantwortet wird: Wirkt sich der Umstand der (Urteils-)Anonymisierung nach den gerichtlichen Maßstäben auf die Haftung oder Sanktionierung unter der DSGVO aus?

- Die (Teil-)Anonymisierung kann die **Auswirkungen** auf betroffene Personen **begrenzen**.
- Zugleich kann eine (Teil-)Anonymisierung gegebenenfalls die rechtmäßige Veröffentlichung unter der DSGVO ermöglichen und ist auch im Rahmen der Haftung sowie Sanktionierung **zu berücksichtigen** (siehe dazu unter D.II).

F3: Ist eine Anonymisierung von unveröffentlichten Urteilen durch das Erlanger Tool ausreichend, um aus diesen nicht-personenbezogene Daten zu machen? Wie ist gegebenenfalls das Restrisiko einer Verletzung des Schutzes personenbezogener Daten bei Einsatz des KI-Systems zu bewerten?

- In Ansehung des einen Wert von 99-100 % unterschreitenden Recall-Werts des Erlanger Tools im Hinblick auf Kennungen und eines deutlichen geringeren Werts mit Blick auf sonstige identifizierende Merkmale dürfte das **Erlanger Tool für sich genommen nicht** den **Anforderungen** an eine **Anonymisierung** im Einklang mit der DSGVO genügen (siehe dazu unter D.III.1).
- Es können aber die mittels Erlanger Tool bearbeiteten Gerichtsentscheidungen zum einen nach den Grundsätzen für die **Veröffentlichung** von Entscheidungen im Einzelfall gegebenenfalls **datenschutzrechtlich zulässig weiterverarbeitet** werden, zum anderen im **GSJ-Projekt durch das KI-Training als Anonymisierungsmaßnahme** und weitere technische und organisatorische Maßnahmen als **anonym anzusehen** sein (siehe hierzu und zum Restrisiko unter F5.a zum Sprachmodell).

F4 (analog F3): Ist eine Anonymisierung von Aktenauszügen durch das Erlanger Tool ausreichend, um aus diesen nicht-personenbezogene Daten zu machen? Wie ist gegebenenfalls das Restrisiko einer Verletzung des Schutzes personenbezogener Daten bei Einsatz des KI-Systems zu bewerten?

- In Ansehung des voraussichtlich geringeren Recall-Werts bei der Anonymisierung von Aktenauszügen **genügt das Erlanger Tool** insoweit erst recht **nicht den Anforderungen an eine Anonymisierung** (siehe dazu unter D.VI.1).
- Für Aktenauszüge kann nicht auf die Standards für die Veröffentlichung wie bei gerichtlichen Entscheidungen zurückgegriffen werden. Im **GSJ-Projekt kann im Ergebnis allerdings ebenfalls eine Anonymisierung** der personenbezogenen Daten vorliegen (siehe hierzu und zum Restrisiko unter F5.a zum Sprachmodell). Das setzt voraus, dass ein **Recall-Wert von mind. 90 % mit Blick auf Kennungen** nicht signifikant unterschritten wird, die Aktenauszüge nach **Dokumentenkategorien geclustert** werden (z.B. Aussortierung oder händische Anonymisierung von beigezogenen Akten der Staatsanwaltschaft, psychologischen und gegebenenfalls ärztlichen Gutachten) und weitere technische und organisatorische Maßnahmen ergriffen werden (z.B. auch die Erkennung von OCR-Fehlern; siehe dazu unter D.VI.2).

F5.a: Enthalten Sprachmodelle, die mit personenbezogenen Daten (hier: unveröffentlichte Urteile, Aktenauszüge) trainiert wurden, personenbezogene Daten und sind daher datenschutzrechtlich von Relevanz?

- Ein Sprachmodell, das mit durch das Erlanger Tool bearbeiteten Gerichtsentscheidungen und Aktenauszügen trainiert wurde und im GSJ-Projekt eingesetzt wird, enthält **grundsätzlich keine personenbezogenen Daten** (siehe dazu unter D.IV.1). Insoweit sind datenschutzrechtliche Anforderungen, z.B. aus Art. 9 DSGVO, nicht zu beachten.
- Dieser Befund setzt voraus, dass
  - nicht im Einzelfall **Kennungen** eindeutig im Modell mathematisch repräsentiert sind. Eine solche mathematische Repräsentation wird mit Blick auf die Länge der Kennungen (z.B. eines Namens als mehrere sog. Token) und die Funktionsweise des KI-Modells als unwahrscheinlich zu bewerten sein;
  - in den Trainingsdaten **deutlich überrepräsentierte Textbestandteile** (z.B. Entscheidungsduplikate, Textbausteine) besonders sorgfältig auf enthaltene Kennungen und sonstige identifizierende Merkmale geprüft werden;

- o im Hinblick auf **sonstige identifizierende Merkmale keine Wahrscheinlichkeit** der Identifizierung einer natürlichen Person besteht, weil die unter D.VIII ausführlich erörterten Maßnahmen ergriffen werden. Hierzu kommen beispielsweise – i.e., nicht zwingend oder abschließend – Maßnahmen wie die folgenden in Betracht:
  - nur ein beschränkter, **justizinterner Personenkreis** den Zugriff auf ein auf dem Sprachmodell aufbauendes KI-System erhält, wobei zudem ein sicherer Login-Mechanismus verwendet wird;
  - der Zugriff auf das KI-System Gegenstand einer separaten **Nutzungsvereinbarung, Dienstanweisung** o.ä. mit Angaben zur (zulässigen) Bedienung ist;
  - ein geeigneter **System Prompt** zum Einsatz kommt oder den Nutzern die Auswahl von Eingabeaufforderung und eines Eingabekontexts (z.B. einer konkreten E-Akte) ermöglicht wird, nicht aber eine individuelle Eingabeaufforderung verlangt wird;
  - eine **Meldefunktion** für etwaige Datenschutzverletzungen vorgesehen ist;
  - **effektive Ausgabefilter** zum Einsatz kommen (z.B. betreffend den Abgleich von Namen und sonstigen Kennungen dahingehend, ob diese auch in den Eingabedaten verwendet wurden);
  - eine **niedrige Temperatur-Einstellung** zur Vermeidung von datenschutzrechtlich relevanten Halluzinationen gewählt wird;
  - **Eingaben nicht** unmittelbar zur **Verbesserung** des Sprachmodells verwendet werden;
- Eine **Verletzung** des Schutzes personenbezogener Daten nach Art. 4 Nr. 12 DSGVO dürfte unter Beachtung dieser Vorgaben als **sehr unwahrscheinlich einzustufen sein**. Um den möglichen Schaden in einem solchen Fall zusätzlich bereits vorsorglich erheblich zu reduzieren, könnten Gerichtsentscheidungen und Aktenauszüge weiter **geclustert** und geprüft werden, z.B. mit Blick auf die Geltendmachung von Schmerzensgeld und typischerweise zu erwartenden Gesundheitsdaten (siehe dazu unter D.VIII.2).

F5.b: Gibt es Bedingungen, unter denen solche Sprachmodelle veröffentlicht werden können?

- Eine **öffentliche Zurverfügungstellung eines entwickelten KI-Systems** dürfte im Fall von beschränkten Eingabemöglichkeiten und effektiven Ausgabefiltern **datenschutzrechtlich vertretbar** sein (siehe dazu unter D.IV.3 und D.VIII.5).

- Die **Open Source-Veröffentlichung des Sprachmodells** einschließlich seiner Gewichte und gegebenenfalls des Trainingskorpus ist **datenschutzrechtlich bedenklich**. Etwaige Sicherungsmechanismen (z.B. Ausgabefilter) greifen nicht mehr ein und es dürfte ein Personenbezug des Sprachmodells anzunehmen sein. Diese Verarbeitung personenbezogener Daten kann allerdings insbesondere dann **gerechtfertigt** sein, wenn das KI-Modell nur auf Gerichtsentscheidungen trainiert wird, die den Anforderungen an eine Anonymisierung für die Veröffentlichung genügen (siehe dazu unter D.IV.3 und D.VIII.5).

F6: Soweit (a) veröffentlichte Entscheidungen (vgl. F2), Urteile und Aktenauszüge nach Anonymisierung durch das Erlanger Tool (vgl. F3 und 4) oder (b) das trainierte Sprachmodell (vgl. F5) personenbezogene Daten enthalten: Besteht für die Verarbeitung dieser personenbezogenen Daten betreffend das Training des Sprachmodells, dessen Einsatz in den einzelnen Use-Cases und die Veröffentlichung des Sprachmodells eine belastbare Rechtsgrundlage nach Art. 6 Abs. 1, Art. 9 Abs. 2 DSGVO, ohne dass ein gesetzgeberisches Tätigwerden erforderlich ist? Welche zusätzlichen datenschutzrechtlichen Verpflichtungen sind in diesem Fall zu beachten (insbesondere Informationspflichten nach der DSGVO sowie etwaige Pflichten aus der nationalen Umsetzung der JI-Richtlinie)?

- Die Verarbeitung eines Restbestands an personenbezogenen Daten im GSJ-Projekt wird **datenschutzrechtlich** – gestützt auf Art. 6 Abs. 1 UAbs. 1 lit. e DSGVO i.V.m. Art. 4 Abs. 1 BayDSG bzw. § 3 Abs. 1 DSG NRW, Art. 6 Abs. 1 UAbs. 1 lit. f DSGVO oder Art. 6 Abs. 4 DSGVO – vorbehaltlich robuster technischer und organisatorischer Maßnahmen als **zulässig** anzusehen sein (siehe dazu unter D.V). Mit Blick auf **besondere Kategorien personenbezogener Daten** empfiehlt sich eine weitgehende **Clusterung** (s.o.), da zwar im Einzelfall rechtfertigende Rechtsgrundlagen aus Art. 9 Abs. 2 DSGVO in Betracht kommen können, aber keine einheitliche Rechtsgrundlage belastbar die Verarbeitung stützen wird (siehe dazu unter D.V.5).
- Darüber hinaus sind weitere datenschutzrechtliche Anforderungen zu beachten:
  - Im GSJ-Projekt dürfte nach dem zugrunde zu legenden Sachverhalt **keine gemeinsame Verantwortlichkeit** der Ministerien und der ausführenden Stellen vorliegen; insbesondere wird für die Annahme einer gemeinsamen Verantwortlichkeit nicht ausreichen, dass die ausführenden Stellen nach eigenem Ermessen über **Debugging-Maßnahmen**

entscheiden. Es bedarf daher insoweit **keiner Rechtsgrundlage für die Übermittlung** personenbezogener Daten an die ausführenden Stellen, wenn diese als **Auftragsverarbeiter** der Ministerien handeln.
  - Die **Ministerien** werden aufgrund der Zusammenführung der Gerichtsentscheidungen als **gemeinsam Verantwortliche** anzusehen sein und daher die Anforderungen aus Art. 26 DSGVO an die Vereinbarung zu beachten haben (siehe dazu unter D.V.3.b).
  - Die **Information** der betroffenen Personen kann nach Art. 14 Abs. 5 lit. b Hs. 1 Var. 2 DSGVO **unterbleiben**. Erforderlich sind allerdings weitergehende **Maßnahmen** (z.B. die Bereitstellung von Informationen für die Öffentlichkeit, siehe dazu unter D.V.6.a).
  - Die **weiteren Betroffenenrechte** und die Einhaltung der **Datenschutzgrundsätze** erweisen sich im Übrigen **mangels Personenbezugs des Sprachmodells** als **handhabbar**, wenn entsprechende Maßnahmen z.B. gegen Halluzinationen getroffen werden (siehe dazu unter D.V.6.b).
  - Es empfiehlt sich die Durchführung einer **Datenschutzfolgenabschätzung**, gegebenenfalls aufbauend auf die Ergebnisse und Empfehlungen dieser Begutachtung. Ferner sind allgemeine **technische und organisatorische Maßnahmen** umzusetzen und regelmäßig zu überprüfen (z.B. Überprüfung des Erlanger Tools und der übrigen Software anhand des Stands der Technik, Vorsehen einer Meldefunktion für etwaige Datenlecks, Wasserzeichen und gegebenenfalls Logs, siehe dazu unter D.V.6.c).

F7.a,b: Wie ist die Verwendung von Aktenauszügen, insbesondere von anwaltlichen Schriftsätzen, zum Modelltraining urheberrechtlich zu bewerten? Gibt es Bedingungen, unter denen ein Modell mit solchen Daten für den justizinternen Gebrauch trainiert werden kann?

- Das **GSJ-Projekt** kann unter Beachtung verschiedener Maßgaben **urheberrechtskonform** im Hinblick auf Aktenauszüge realisiert werden.
- Im Einzelnen ist das Projekt urheberrechtlich wie folgt zu bewerten:
  - Die Aktenauszüge **können urheberrechtlich geschützte Bestandteile enthalten**. Das betrifft insbesondere einige, aber nicht per se alle **Schriftsätze** sowie vor allem verfasste **Gutachten** und sonstige **Anlagen** als Bestandteil der Aktenauszüge (siehe dazu unter E.I).
  - Das KI-Training und der weitere Einsatz im GSJ-Projekt betreffen maßgeblich das Verwertungsrecht der **Vervielfältigung** (§ 16 UrhG) für das Trainingskorpus sowie im Rahmen der Ausgabe und lösen

gegebenenfalls – regelmäßig im Rahmen von Text und Data Mining gerechtfertigte – **Umgestaltungen** aus (§ 23 UrhG). Mangels Reproduzierbarkeit liegt bei Beachtung der auch datenschutzrechtlich gebotenen Maßnahmen nach vorzugswürdiger, aber nicht unbestrittener Ansicht regelmäßig **keine Vervielfältigung im Sprachmodell vor** (siehe dazu unter E.II).

- Diese Nutzungshandlungen sind grundsätzlich einschließlich der Einbindung Dritter (unter E.II.3) zulässig und können:
  - gegebenenfalls teilweise unterbleiben, indem auf **abgeleitete Textformate** und ein Training an der Datenquelle gesetzt wird (siehe dazu unter E.II.1.a);
  - auf **§ 44b Abs. 2 S. 1 UrhG** (**Text und Data Mining**) gestützt werden, soweit nicht im Einzelfall wirksam ein **Vorbehalt** i.S.d. § 44b Abs. 3 UrhG erklärt wurde (siehe dazu unter E.II.2.c);
  - auf **§ 44a Nr. 2 UrhG** gestützt werden, soweit es nur zu **vorübergehenden** Vervielfältigungen kommt (siehe dazu unter E.II.2.d);
  - **nicht auf § 45 UrhG** mangels eines konkreten **Verfahrensbezugs** gestützt werden (siehe dazu unter E.II.2.b);
  - **nicht auf § 60d UrhG** für **wissenschaftliche Forschungszwecke** gestützt werden, da der Forschungszweck nur Nebenzweck ist und im GSJ-Projekt der operative Einsatz in der Justiz im Vordergrund steht (siehe dazu unter E.II.2.c.dd).
- Die **Voraussetzung** für die urheberrechtliche Zulässigkeit ist, vorbehaltlich des Einsatzes abgeleiteter Textformate zur Vermeidung von Vervielfältigungen, **maßgeblich** die Beachtung von **Vorbehalten** im Rahmen des § 44b UrhG (zu einzelnen Vorbehalten siehe unter E.V.1).
- Vereinzelte **Vervielfältigungen** im Rahmen der **Ausgabe** sind durch geeignete Maßnahmen zu **verhindern** (siehe dazu unter E.V), insbesondere durch:
  - **datenschutzrechtlich** gebotene **Anonymisierungsmaßnahmen**, die es insbesondere erschweren, unter Bezugnahme auf Kennungen urheberrechtlich geschützte Ausgaben aus den Trainingsdaten zu extrahieren (z.B. „Schreibe im Stil von x“);
  - **Nutzungsvereinbarungen, Dienstanweisungen** o.ä. über die Benutzung des KI-Systems;
  - gegebenenfalls **System Prompts und Filtermaßnahmen**, soweit diese effektiv die Wahrscheinlichkeit von Wiedergaben aus den Trainingsdaten reduzieren (können).

- o Die Wahrscheinlichkeit einer solchen Vervielfältigung ist ferner zu verringern, indem **Dokumentenkategorien aus dem KI-Training ausgeschlossen** oder nachbearbeitet werden, die voraussichtlich urheberrechtlich geschützt sind, aber gegebenenfalls von geringer Bedeutung für den Einsatz im GSJ-Projekt sind (z.B. ausgewählte Gutachtenkategorien, siehe dazu unter E.V.1).

F7.c: Gibt es Bedingungen, unter denen es veröffentlicht werden kann?

- Die **Veröffentlichung** eines auf dem Sprachmodell aufbauenden **KI-Systems** ist **urheberrechtlich grundsätzlich zulässig**. Empfehlenswert sind Einschränkungen bei der Eingabeaufforderung anstelle einer freien Eingabe (siehe dazu unter E.V.3).
- Die **Open Source-Veröffentlichung des Sprachmodells** einschließlich seiner Gewichte und gegebenenfalls des Trainingskorpus ist **urheberrechtlich bedenklich**. In diesen Fällen dürften regelmäßig Vervielfältigungen vorliegen, die nicht auf eine Schrankenbestimmung gestützt werden können (siehe dazu unter E.II.1.a.bb unter E.V.3).

Aus der **Zusammenschau von Datenschutzrecht und Urheberrecht** lassen sich zudem die folgenden Befunde mit Relevanz für das GSJ-Projekt festhalten:

- Der Begriff der **wissenschaftlichen Forschungszwecke** i.S.d. DSGVO und des UrhG ist in seinem jeweiligen Kern unter beiden Rechtsakten ähnlich auszulegen, wobei zugleich Abweichungen in den Details in Betracht kommen. Die Verfolgung wissenschaftlicher Forschungszwecke ist jeweils abzulehnen, wenn Handlungen mit Forschungsbezug (z.B. Verarbeitungen oder Vervielfältigungen zur Entwicklung neuer Produkte) unmittelbar und vorrangig auf die Entwicklung eines Produkts für den operativen Einsatz abzielen. Wenn und soweit **urheberrechtlich** (privilegierte) wissenschaftliche Forschungszwecke i.S.d. § 60d UrhG angenommen werden, würden sich regelmäßig **datenschutzrechtlich** mit Blick auf Art. 89 DSGVO und nationale Vorschriften **strengere Anforderungen** ergeben – insbesondere – für die Implementierung technischer und organisatorischer Maßnahmen einschließlich der Anonymisierung.
- Sowohl im Datenschutzrecht als auch im Urheberrecht ist für die Bewertung des Sprachmodells die **Möglichkeit der Extraktion** von Daten bzw. Werkbestandteilen aus den **Ausgaben** zu **berücksichtigen**, auch wenn sich die Anforderungen im Detail unterscheiden. Daher erfordern beide Rechtsgebiete jeweils Clusterungen und Maßnahmen im Hinblick auf das konkrete KI-System.

- Die **Entfernung von Kennungen aus den Trainingsdaten** ist zwar primär datenschutzrechtlich geboten, kann sich aber auch urheberrechtlich als vorteilhaft erweisen.

Sofern man hypothetisch für eine Zusammenarbeit zwischen Ministerien und ausführenden Stellen die **Verfolgung eigener Forschungszwecke** unterstellt, ergeben sich folgende Besonderheiten:

- In **datenschutzrechtlicher** Hinsicht wären die Ministerien und die ausführenden Stellen **gemeinsam Verantwortliche** für die Übermittlung und Weiterverarbeitung der in den Gerichtsentscheidungen und Aktenauszügen enthaltenen personenbezogenen Daten (Art. 26 DSGVO). Das Ministerium bedürfte daher in Abweichung vom Sachverhalt des GSJ-Projekts einer **Rechtsgrundlage** nach Art. 6, 9 DSGVO für die Übermittlung personenbezogener Daten zu Forschungszwecken der ausführenden Stellen. Eine solche Rechtsgrundlage würde sich angesichts der zahlreichen betroffenen Personen und in Abhängigkeit von den konkret übermittelten Datenkategorien **nicht ohne Weiteres aus den landesrechtlichen oder sonstigen Vorschriften** ergeben. Insbesondere verlangen die landesdatenschutzrechtlichen Forschungsgeneralklauseln regelmäßig eine Anonymisierung, mithin einen Anonymisierungsgrad, der **über den durch das Erlanger Tool gewährleisteten Anonymisierungsgrad** hinausgeht. Darüber hinaus ergäben sich weitere Anforderungen (z.B. mit Blick auf eine Datenschutzfolgenabschätzung; siehe dazu unter D.VII).
- In **urheberrechtlicher** Hinsicht könnte die Notwendigkeit entfallen, allfällige **Vorbehalte** für ein Text und Data Mining nach § 44b Abs. 3 UrhG zu berücksichtigen, wenn einzelne Vervielfältigungen abweichend vom Sachverhalt des GSJ-Projekts primär dem forschungsbezogenen Text und Data Mining der ausführenden Stellen dienten (§ 60d UrhG). Eine solche Ausnahme nach § 60d UrhG von der Verpflichtung zur Beachtung von Text und Data Mining-Vorbehalten gilt auch in diesem hypothetischen Szenario nicht für Vervielfältigungen, die für ein Text und Data Mining unmittelbar zur Entwicklung des Sprachmodells im operativen Justizeinsatz durchgeführt werden (siehe dazu unter E.IV).

Es wird ferner hingewiesen auf **die Kataloge der Handlungsempfehlungen** aus **datenschutzrechtlicher** Perspektive (siehe dazu unter D.VIII) und **urheberrechtlicher Perspektive** (siehe dazu unter E.V).