Karlsruher Institut für Technologie

# Chancen von KI zur Stärkung deliberativer Kultur

Sebastian Cacean

Institution: Karlsruher Institut für Technologie
Kaiserstr. 12
76131 Karlsruhe

Förderkennzeichen: 3924406K03

Laufzeit des Vorhabens: 01.06.2024 – 31.12.2025

Projektinternetseite: https://compphil2mmae.github.io/research/kideku/

Autor: Dr. Sebastian Cacean

DOI: 10.5445/IR/1000193096

Projektbeteiligte: Prof. Dr. Gregor Betz (Projektleitung)

Prof. Dr. Christian Seidel (Projektleitung)

Dr. Sebastian Cacean (Projektmanagement und -durchführung)

Dr. Inga Bones (Projektdurchführung)

Leonie Wahl (studentische Mitarbeiterin)

Reta Lüscher-Rieger (studentische Mitarbeiterin)

Leon Knauer (studentischer Mitarbeiter)

Das diesem Bericht zugrundeliegende Vorhaben *„Chancen von KI zur Stärkung unserer deliberativen Kultur“* wurde vom Bundesministerium für Bildung, Familie, Senioren, Frauen und Jugend im Rahmen der Richtlinie zur Förderung von Künstlicher Intelligenz für das Gemeinwohl gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt bei dem Autor.

Gefördert vom:

Bundesministerium
für Bildung, Familie, Senioren,
Frauen und Jugend

# Inhaltsverzeichnis


**1 Einleitung** **4**

**2 Einsatzszenarien von KI zur Stärkung deliberativer Kultur** **6**
2.1 Einsatzbereiche . . . . . 6
2.2 Ziele und Umsetzung . . . . . 8
2.2.1 Respektvolle Kommunikation (*Civility*) . . . . . 9
2.2.2 Rationalität des Diskurses . . . . . 11
2.2.3 Kompetenzentwicklung und Wissensmanagement . . . . . 13
2.2.4 Gleichberechtigte Teilhabe . . . . . 16
2.2.5 Weitere KI-Einsatzszenarien zur Stärkung deliberativer Kultur . 17
2.3 Umsetzungsdimensionen . . . . . 20

**3 Projektergebnisse** **22**
3.1 Toxicity-Detector . . . . . 22
3.1.1 Die Toxicity-Detector Pipeline . . . . . 24
3.1.2 Explorative Evaluierung des Toxicity-Detectors . . . . . 26
3.2 EvidenceSeeker Boilerplate . . . . . 29
3.2.1 Die EvidenceSeeker-Pipeline . . . . . 31
3.3 syncIALO Datensatz . . . . . 35
3.3.1 Erstellung von syncIALO . . . . . 35
3.3.2 Verwendungsmöglichkeiten . . . . . 37

**4 Herausforderungen und Empfehlungen** **38**
4.1 Rolle von Vertrauen . . . . . 38
4.2 Evaluierung und Optimierung von KI-Tools . . . . . 40
4.2.1 Herausforderungen bei der Optimierung von KI-Tools . . . . . 41
4.2.2 Lösungsansätze . . . . . 44
4.3 Praktische Herausforderungen . . . . . 48

**Literatur** **51**

# 1 Einleitung

Generative Sprachmodelle und die damit verbundene Zunahme von KI-generierten Inhalten verändern die Informationsproduktion und -verarbeitung rasant. Spätestens mit der Veröffentlichung von ChatGPT im Dezember 2022 ist einer breiten Öffentlichkeit klar geworden, welche herausragenden – noch vor einigen Jahren für unmöglich gehaltenen – Fähigkeiten generative Sprachmodelle (Large Language Models, LLMs) besitzen und welche gesellschaftlichen Umwälzungen diese sich rasant entwickelnde Technologie nach sich ziehen könnte.

Generative KI ist eine Dual-Use-Technologie: Manipulative Akteur:innen können sie nutzen, um liberale Demokratien zu schwächen – etwa durch die massenhaft automatisierte Erzeugung von Desinformation, Propaganda und toxischen Inhalten. Doch generative KI stellt nicht nur ein Risiko dar, sondern birgt auch Chancen, die öffentliche Diskurslandschaft und die politische Willensbildung im Sinne der Ideale liberaler Demokratien zu stärken.

Dieses Potential der Stärkung deliberativer Kultur ist vielfältig. Sprachmodelle können dazu genutzt werden, den öffentlichen Diskurs konstruktiver und inklusiver zu gestalten. Sie können die Rationalität des Diskurses stärken, indem sie Bürger:innen dabei helfen, relevante Argumente zu identifizieren, zu verstehen und zu evaluieren. KI kann auch dazu beitragen, Beteiligungsformate in einem zuvor nicht möglich gewesenen Maßstab zu skalieren und damit mehr Menschen Zugang zu politischer Partizipation zu eröffnen.

Allerdings reicht es nicht aus, sich allein auf die technischen Möglichkeiten zu konzentrieren. Eine rein technokratische Perspektive birgt die Gefahr, relevante ethische und gesellschaftliche Aspekte auszublenden. KI-gestützte Deliberation darf grundlegende deliberative Normen nicht verletzen. Vielmehr sollte sie so ausgestaltet werden, dass wir diesen Normen zu einem hohen Maß gerecht werden. Wir müssen uns auch fragen, wie KI-gestützte Deliberation ausgestaltet werden muss, damit sie von Bürger:innen akzeptiert und genutzt wird. Der Einsatz von KI-basierten Applikationen kann unsere deliberative Kultur nur dann stärken, wenn diese von Bürger:innen verwendet werden, was wiederum ein hinreichend großes Vertrauen in die Sicherheit und Verlässlichkeit solcher Anwendungen voraussetzt.

Im Projekt *„Chancen von KI zur Stärkung unserer deliberativen Kultur“* (KIdeKu) sind wir der Frage nachgegangen, wie Large Language Models eingesetzt werden können, um deliberative Kultur zu stärken. Das Projekt wurde am Karlsruher Institut für Technologie (KIT) im Arbeitsbereich *Computationale Philosophie, Philosophische Methoden, Moralphilosophie & Angewandte Ethik* (CompPhil²MMAE) durchgeführt und

vom Bundesministerium für Bildung, Familie, Senioren, Frauen und Jugend (BMBFSFJ) gefördert (Projektlaufzeit: 01.06.2024–31.12.2025).

Dabei verfolgte das Projekt drei zentrale Ziele:

1. *Entwicklung von Einsatzszenarien:* Die Identifizierung relevanter Einsatzszenarien von KI zur Stärkung deliberativer Kultur sollte einen umfassenden Überblick darüber geben, wie LLMs gemeinwohlorientiert in unserer demokratischen Praxis eingesetzt werden können.
2. *Schaffung technischer Grundlagen:* Für ausgewählte Einsatzszenarien sollten technische Grundlagen in Form von Open-Source-Prototypen und offenen Datensätzen entwickelt werden, die von der Community genutzt und weiterentwickelt werden können.
3. *Handlungsorientierung:* Aus den gewonnenen Erkenntnissen und Projektergebnissen sollten Empfehlungen für zivilgesellschaftliche und politische Akteur:innen entwickelt werden.

Die Ziele wurden mit Abschluss des Projektes im Dezember 2025 erreicht. Der vorliegende Bericht fasst die Ergebnisse in drei Kapiteln zusammen:

Kapitel 2 gibt einen Überblick über relevante Einsatzszenarien von KI zur Stärkung deliberativer Kultur. Die Szenarien werden entlang dreier Dimensionen – Einsatzbereich, Ziele und Umsetzung – charakterisiert und anhand konkreter Beispiele illustriert. Dabei wird gezeigt, wie KI-gestützte Anwendungen zur Förderung respektvoller Kommunikation, zur Stärkung der Diskursrationalität sowie zur Ermöglichung von Teilhabe und Inklusion beitragen können.

Kapitel 3 stellt die im Projekt entwickelten Prototypen und Datensätze vor: den KIdeKu *Toxicity-Detector*, einen LLM-basierten Prototyp zur Identifizierung toxischer Sprache; die *EvidenceSeeker-Boilerplate*, ein Code-Template für KI-basierte Faktenprüfung auf Grundlage eigener Wissensbestände (zum Beispiel in Form von Berichten); sowie den *syncIALO-Datensatz*, der als Trainings- und Evaluationsdatensatz für KI-gestützte Tools im Bereich der Argumentationsanalyse dienen kann.

Kapitel 4 diskutiert zentrale Herausforderungen bei der Integration von KI-basierten Tools in deliberative Prozesse und formuliert Empfehlungen für das Design KI-basierter Deliberationstools. Es wird argumentiert, dass die Verlässlichkeit solcher Tools zentral für die Erreichung deliberativer Ziele und das Vertrauen von Nutzer:innen ist, woraus sich die Notwendigkeit ihrer systematischen Evaluierung und Optimierung ergibt.

# 2 Einsatzszenarien von KI zur Stärkung deliberativer Kultur

Wie können Sprachmodelle eingesetzt werden, um unsere deliberative Kultur zu stärken, und welche Einsatzszenarien sind hierfür besonders vielversprechend? Im Folgenden werden einige Szenarien skizziert, ohne Anspruch auf Vollständigkeit zu erheben. Die Beispiele sollen vielmehr das Spektrum der vielfältigen Möglichkeiten zur Stärkung deliberativer Kultur durch KI aufzeigen.

Die Einsatzszenarien lassen sich dabei anhand der folgenden drei Dimensionen charakterisieren (siehe Abbildung 2.1):

1. **Einsatzbereich:** In welchen deliberativen Kontexten wird KI eingesetzt?
2. **Ziele:** Welche Ziele sollen durch den Einsatz von KI in diesem Einsatzbereich erreicht werden, um deliberative Kultur zu stärken?
3. **Umsetzung & Funktion:** Wie soll KI konkret genutzt werden, um diese Ziele zu erreichen?

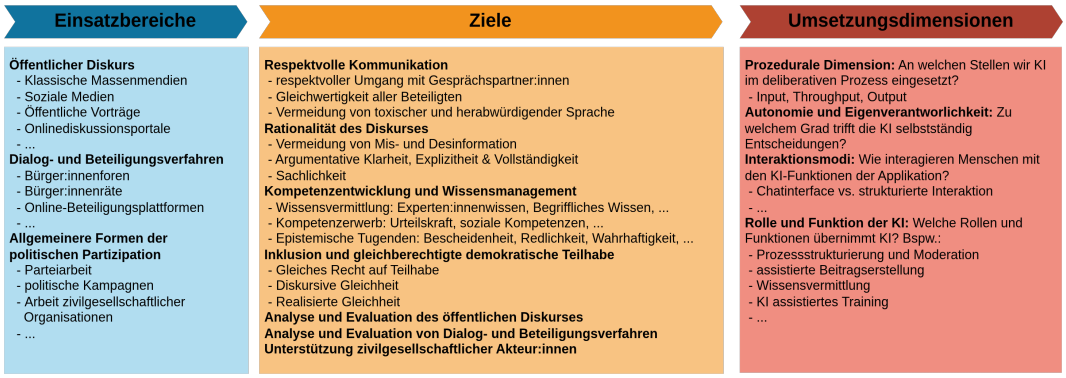


Abbildung 2.1: Überblick Einsatzszenarien

## 2.1 Einsatzbereiche

Um die relevanten Einsatzbereiche – also die deliberativen Kontexte – näher zu bestimmen, muss geklärt, was mit Deliberation eigentlich gemeint ist. Obwohl in den

Politikwissenschaften um die Details einer adäquaten Erläuterung des Deliberationsbegriffs gerungen wird,[1] lässt sich ein unproblematischer Kern wie folgt formulieren: In deliberativen Kontexten treffen Menschen in unterschiedlichen öffentlichen Rollen (beispielsweise als Bürger:innen, Politiker:innen oder Interessenvertreter:innen) zusammen, um Anliegen des öffentlichen Lebens zu diskutieren und gegebenenfalls zu entscheiden. Dabei werden unterschiedliche Ansichten ausgetauscht und im Lichte von Informationen, Gründen und Einwänden angepasst, um bestenfalls zu einer gut begründeten Entscheidung zu gelangen.

Deliberation unterscheidet sich deutlich von anderen Interaktionsformen: Der Austausch persönlicher Erlebnisse ohne Bezug zu öffentlichen Anliegen stellt keine Deliberation im hier verwendeten Sinne dar. Zudem wird Deliberation häufig von Verhandlungen oder Mediationen abgegrenzt.[2] In Verhandlungen treffen Personen mit gegensätzlichen Interessen und meist asymmetrischen Durchsetzungsmöglichkeiten aufeinander. Verhandlungen zielen auf einen Interessenausgleich ab, der zwar durch Normen reguliert ist, jedoch selten auf einer gleichberechtigten Interaktion basiert. Im Gegensatz dazu wird ideale Deliberation als Austausch von Ansichten, Perspektiven und Argumenten gleichberechtigter Gesprächsteilnehmer:innen verstanden.

Bevor wir mit der Charakterisierung idealtypischer Deliberation bereits auf mögliche Ziele eines KI-Einsatzes zu sprechen kommen, ist die vorläufige Charakterisierung ausreichend, um mögliche Einsatzbereiche, also relevante deliberative Kontexte, näher zu spezifizieren.

Der **öffentliche und teilöffentliche** Diskurs ist ein zentraler deliberativer Kontext. In sozialen Medien, Onlineforen, auf öffentlichen Vorträgen, in Plenumsdiskussionen und in parlamentarischen Debatten werden Perspektiven, Meinungen, Informationen und Argumente ausgetauscht. Auch wenn klassische Massenmedien kein Forum aktiver Deliberation darstellen, da ihnen die direkte Interaktion zwischen Bürger:innen fehlt, spielen sie für deliberative Kontexte eine zentrale Rolle: Sie spiegeln und prägen den öffentlichen Diskurs, geben Bürger:innen mindestens mittelbar die Möglichkeit, ihre eigene Meinung auszudrücken und informieren sie über soziale, politische und kulturelle Belange des öffentlichen Lebens.

Auch in vielen **Dialog- und Beteiligungsverfahren** spielt Deliberation eine entscheidende Rolle. Darunter zählen unter anderem Bürgerräte und Konsenskonferenzen.[3] In Bürgerräten kommen in etwa acht bis zwölf zufällig gewählte Bürger:innen über einen Zeitraum von zwei Tagen zusammen, die sich in einem ersten Schritt auf ein gesellschaftsrelevantes Problem als Thema einigen, um dann dieses zu diskutieren. Die vom Bürgerrat erarbeiteten Lösungsansätze werden veröffentlicht und können als Input für die politische Entscheidungsfindung dienen.

Konsenskonferenzen sind im Vergleich zu Bürgerräten stärker strukturiert und finden über einen Zeitraum von 3 Tagen mit etwa 20 zufällig gewählten Bürger:innen statt.

[1] Einen Überblick geben Neblo (2011) und Bächtiger u. a. (2010).
[2] Zusammenfassend in Goldschmidt (2014), S. 69–74.
[3] Für einen Überblick siehe Nanz und Fritsche (2012).

Im Gegensatz zum Bürgerrat wird das Thema vorgegeben, und vor Beginn der eigentlichen Konferenz werden den Teilnehmenden Informationen bereitgestellt. Während der Konferenz haben die Teilnehmenden die Möglichkeit, Expert:innen zu konsultieren und untereinander zu diskutieren. Anschließend erarbeiten sie Stellungnahmen und Empfehlungen, die in Form eines Abschlussberichts veröffentlicht und in der politischen Entscheidungsfindung berücksichtigt werden können.

Dialog- und Beteiligungsverfahren zeichnen sich dadurch aus, dass sie unter bestimmten Zielsetzungen durchgeführt werden und Teilnehmende in ihrer Rolle als Bürger:innen und Laien zu komplexen gesellschaftsrelevanten Themen Stellung nehmen (Goldschmidt 2014). In vielen Formaten spielt der Austausch von Argumenten auf Basis von Informationen und Evidenzen eine zentrale Rolle. Im Gegensatz zu reinen Abstimmungen wird im Rahmen solcher Deliberationen häufig eine gemeinsame Position entwickelt, die nicht unbedingt Konsens voraussetzt, und die dann als Ergebnis des Verfahrens im Idealfall Einfluss auf politische Entscheidungsprozesse hat.

Neben Dialog- und Beteiligungsverfahren sollen im Folgenden auch **allgemeinere Formen der Partizipation** als relevant erachtet werden, sofern sie mindestens mittelbar für gelingende Deliberation wichtig sind. Darunter zählt insbesondere die **politische Partizipation**, wie z.B. die Mitgliedschaft und Mitarbeit in Parteien, die Arbeit in zivilgesellschaftlichen Organisationen und selbstorganisiertes Bürgerengagement im Demokratiebereich.[4] Deliberation spielt sowohl in der Parteiarbeit als auch in der zivilgesellschaftlichen Arbeit eine zentrale Rolle. So diskutieren Parteimitglieder auf Ständen mit Bürger:innen oder deliberieren parteiintern auf Parteitagen und in Parteibeiräten. Auch zivilgesellschaftliche Organisationen müssen intern Positionen und Strategien entwickeln und zu gesellschaftsrelevanten Themen Stellung nehmen. Darüber hinaus tragen viele zivilgesellschaftliche Organisationen im Demokratiebereich mittelbar zur Stärkung deliberativer Kultur bei, indem sie für deliberative Werte und Normen einstehen und diese fördern.

## 2.2 Ziele und Umsetzung

Für die Bestimmung der Ziele, die durch den Einsatz von KI zur Stärkung der deliberativen Kultur verfolgt werden können, lohnt sich ein erneuter Blick auf den Deliberationsbegriff. Deliberation ist ein normativer Begriff mit impliziten Gelingensbedingungen. Diese Bedingungen formulieren deliberative Normen, anhand derer die Güte von Deliberation gemessen wird. Wenn von der Stärkung deliberativer Kultur gesprochen wird, geht es also implizit darum, Deliberation so auszurichten, dass sie diesen Normen möglichst gerecht wird. Deliberative Normen wurden prominent von Habermas (1981) und Cohen (2005) formuliert und seitdem in der Politikwissenschaft diskutiert, verfeinert und ergänzt. Die fachwissenschaftliche Debatte um diese Normen ist komplex und verzweigt.[5] Relativ unproblematisch erscheint es, die von Habermas

[4] Zum Begriff der *politischen Partizipation* vgl. Woyke (2021).
[5] Für einen Überblick vgl. Neblo (2011) und Bächtiger u. a. (2010)

und Cohen formulierten Normen als regulatives Ideal zu betrachten: Sie formulieren idealtypische Deliberation, die (mit bestimmten Ausnahmen und Bedingungen) als anzustrebende Deliberationsform gilt (Steiner u. a. 2004). Im Folgenden sollen diese Normen skizziert und es soll anhand von Beispielen illustriert werden, wie KI eingesetzt werden kann, um Deliberation im Sinne dieser Normen zu stärken.[6]

### 2.2.1 Respektvolle Kommunikation (*Civility*)

Ein respektvoller Umgang verlangt, dass die Bedürfnisse, die Rechte und die prinzipielle Gleichwertigkeit aller Beteiligten anerkannt werden (Friess u. a. 2025). Diese Arten von Anerkennung setzen voraus, dass auf toxische Sprache und damit insbesondere auf herabwürdigende und derogative Rede verzichtet wird. Dazu gehört auch, dass sich Beteiligte aufeinander beziehen und nicht etwa die Beiträge anderer ignorieren.[7]

Respektvolle Kommunikation ist eine zentrale Voraussetzung für eine Gesprächsatmosphäre, die einen konstruktiven Umgang trotz divergierender Meinungen ermöglicht. Ohne einen respektvollen Umgang sind Teilnehmende weniger bereit, die Beiträge anderer ernsthaft zu berücksichtigen (Steenbergen u. a. 2003) und gegebenenfalls ihre Überzeugungen im Lichte neuer Informationen und Argumente anzupassen.

Gerade in sozialen Medien fehlt es häufig an Respekt. So berichten in einer repräsentativen Umfrage 2/3 der Befragten zwischen 16 und 24 Jahren, Hass im Netz gesehen zu haben (Brennauer u. a. 2024). Wie können nun LLMs dazu beitragen, einen respektvollen Umgang in deliberativen Kontexten zu fördern? Eine wichtige Rolle für die KI-gestützte Förderung eines respektvollen Umgangs ist die automatisierte Identifizierung toxischer Sprache. KI-basierte Detektion toxischer Sprache ist schon länger Gegenstand der Forschung.[8]

Klassische Machine-Learning-Algorithmen haben allerdings teilweise Schwierigkeiten bei der Erkennung toxischer Sprache aufgrund der Kontextabhängigkeit dieses Phänomens (Guo u. a. 2024). Ob Äußerungen bezüglich des Kriteriums von Respekt problematisch sind, kann unter anderem davon abhängen, welche kulturellen und sozialen Normen im spezifischen Äußerungskontext gelten, welche Intentionen die Sprecher:innen verfolgen und ob indirekte Rede oder Codewörter (bspw. beim *whistleblowing*) verwendet werden. Schon die Verfügbarkeit geeigneter Trainings- und Testdatensätze ist eine Herausforderung, da die notwendigen Kontextinformationen in den Datensätzen enthalten sein sollten und die Datensätze die Heterogenität und Diversität des Toxizitätsphänomens hinreichend adäquat abbilden sollten.

[6] Wir folgen dabei lose den in Friess u. a. (2025) und Friess und Eilders (2015) vorgeschlagenen Kategorisierungen.

[7] Während Friess u. a. (2025) gegenseitige Bezugnahme als eine weitere Norm aufführen (*reciprocity*), subsumieren wir diese der Einfachheit halber unter die Norm des respektvollen Umgangs.

[8] Für einen Überblick siehe Schmidt und Wiegand (2017) und Fortuna und Nunes (2018).

Sprachmodelle, so die Hoffnung, könnten genutzt werden, um geeignete (synthetische) Datensätze zu erzeugen und um Toxizität akkurat zu identifizieren, wenn ihnen die entsprechenden Kontextinformationen zur Verfügung gestellt werden.[9]

Erfolgt die Detektion toxischer Sprache mittels Sprachmodelle zuverlässig, könnten diese eingesetzt werden, um in sozialen Medien und anderen Onlineplattformen automatisch Beiträge mit herabwürdigender oder derogativer Sprache zu markieren. Darauf aufbauend könnte eine KI-basierte Moderation entsprechende Beiträge kennzeichnen und gegebenenfalls entfernen. Zudem könnten Moderationstools Nutzer:innen sensibilisieren, indem sie schon während der Formulierung problematischer Äußerungen Warnungen ausgeben und alternative Formulierungen vorschlagen.[10] KI-unterstützte Beitragserstellung kann den Umgangston verbessern, ohne Inhalte zu verfälschen, und damit die Bereitschaft fördern, Gegenpositionen ernst zu nehmen (Argyle u. a. 2023). Selbst in Kontexten ohne direkte Moderationsmöglichkeiten kann KI zur Eindämmung von Toxizität beitragen, beispielsweise durch LLM-basierte Counterspeech-Bots, die Gegenrede generieren und so die Nutzerinteraktion mit toxischen Beiträgen reduzieren (Saha u. a. 2024; Podolak u. a. 2024).

Beispiel: KI-gestützte Förderung respektvoller Kommunikation

**Fragestellung:** Wie können Sprachmodelle genutzt werden, um einen respektvollen Umgang in politischen Online-Diskussionen zu fördern? In einem Experiment mit 1574 Teilnehmer:innen zum Thema Waffenkontrolle untersuchten Argyle u. a. (2023), ob KI-basierte Formulierungsassistenten zu einer Verbesserung der gegenseitigen Anerkennung und der wahrgenommenen Qualität von Online-Diskussionen beitragen können.

**KI-Ansatz:** Teilnehmer:innen wurden in Paare mit gegensätzlichen Positionen aufgeteilt, die anschließend ihre Ansichten in kurzen Dialogen austauschten. Für die Beitragserstellung durchlief der/die Nutzer:in die folgenden Schritte:

1. Der:die Nutzer:in formuliert einen Diskussionsbeitrag.
2. Das Sprachmodell analysiert den Beitrag und generiert Reformulierungsvorschläge zur Verbesserung von Höflichkeit und Respekt, die den inhaltlichen Standpunkt des Beitrags nicht verändern.
3. Der:die Nutzer:in wählt einen Vorschlag oder behält die ursprüngliche Formulierung bei, bevor der Beitrag sichtbar wird.

[9] Vorläufige Einschätzungen geben Albladi u. a. (2025), Guo u. a. (2024), Kruk u. a. (2024), Pendzel u. a. (2023) und Plaza-del-arco u. a. (2023). Im Rahmen des KIdeKU-Projekts haben wir eine explorative Evaluation unseres Toxicity-Detectors durchgeführt. Die Ergebnisse dieser explorativen Evaluation finden sich in Kapitel 3.1.2.

[10] Im IndI-Projekt wurde zum Beispiel ein KI-basierter Prototyp entwickelt, der Texteingaben bezüglich Respekt, Höflichkeit und Empathie analysiert und gegebenenfalls Reformulierungsvorschläge unterbreitet.

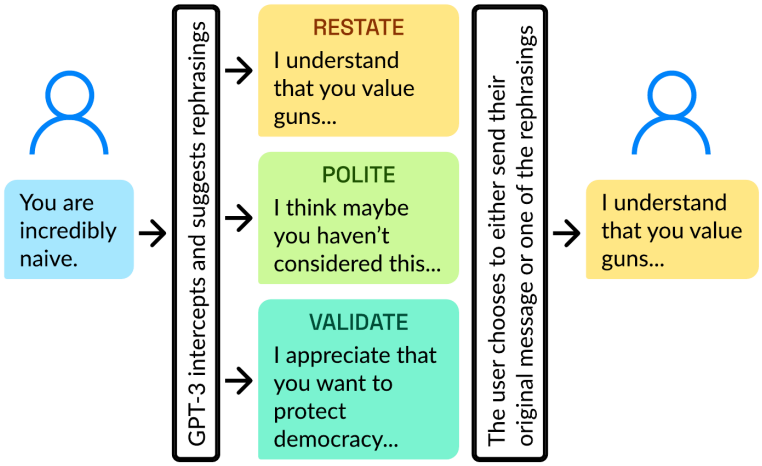


Abbildung 2.2: Ablauf KI-assistierter Reformulierungen von Beiträgen in Onlinediskussionen. Leicht angepasst von Argyle u. a. (2023).

**Ergebnis:** Die Studie ergab, dass die politischen Positionen der Teilnehmer:innen unverändert blieben. Der Einsatz von KI verbesserte jedoch die wahrgenommene Gesprächsqualität und gegenseitige Bezugnahme. Die Autor:innen schließen, dass KI-Tools genutzt werden können, um gegenseitigen Respekt und konstruktiven Dialog in digitalen Räumen zu fördern, ohne Nutzer:innen in ihren Überzeugungen zu manipulieren.

### 2.2.2 Rationalität des Diskurses

Das deliberative Modell stellt den Austausch von Ansichten, Informationen und Argumenten in den Mittelpunkt kollektiver Entscheidungsfindung. Dieser Austausch soll sachlich, konstruktiv und wahrheitsorientiert erfolgen. Deliberation ist rational, insofern Teilnehmende empirische Evidenzen und Fakten berücksichtigen und ihre Überzeugungen im Lichte rationaler Argumente anpassen (Gerber u. a. 2014). Gleichzeitig sollten Teilnehmende idealerweise nicht von Faktoren beeinflusst werden, die für die Güte von Argumenten irrelevant sind. Zu solchen Faktoren zählen beispielweise Zwang oder die Dominanz bestimmter Positionen.

Die Erfüllung dieser Rationalitätsnormen ist insbesondere für die Qualität der Entscheidungsfindung relevant. Auch bei kollektiven Entscheidungen gibt es bessere und schlechtere Entscheidungen, die sich an instrumentellen und allgemein gültigen moralischen Werten messen lassen. Wenn Teilnehmende diesen Rationalitätsnormen gerecht werden, führt das – so die Idee – zu besseren Entscheidungen.

Aber wie können KI-gestützte Anwendungen dazu beitragen, die Rationalität von Diskursen zu fördern? Es gibt verschiedene Möglichkeiten, wie KI in diesem Zusammen-

hang eingesetzt werden könnte. Zum einen könnten KI-gestützte Tools dazu genutzt werden, um Mis- und Desinformation zu erkennen und zu bekämpfen, indem sie beispielsweise Sachaussagen überprüfen und Quellen bewerten. Zum anderen könnten KI-gestützte Tools eingesetzt werden, um die Qualität von Argumenten zu bewerten und zu verbessern.

Die Verbreitung von Mis- und Desinformation kann dazu führen, dass Positionen und Entscheidungen auf falschen Aussagen beruhen. Misinformation bezeichnet die unbeabsichtigte Verbreitung falscher Aussagen, verursacht durch Fehlinformationen, Missverständnisse und unzureichende Medienkompetenz, während Desinformation die absichtliche Verbreitung falscher oder irreführender Aussagen zur Täuschung oder Manipulation bezeichnet (Fallis 2015). Desinformation wird insbesondere in sozialen Medien von einer kleinen Anzahl von Akteur:innen verbreitet (Baribi-Bartov u. a. 2024). In der politischen Kommunikation wird sie häufiger von rechtspopulistischen als von Politiker:innen anderer politischer Richtungen als Mittel gebraucht (Törnberg und Chueri 2025).

Analog zur Identifikation toxischer Sprache existieren zahlreiche KI-basierte Ansätze zur Erkennung von Mis- und Desinformation.[11] Diese nutzen Methoden wie Textmerkmalsanalyse und Faktenüberprüfung anhand bewerteter Quellen, um die Glaubwürdigkeit von Informationen zu beurteilen. KI-basierte Faktenchecker können in sozialen Medien und anderen Onlineplattformen eingesetzt werden, um die Verbreitung von Mis- und Desinformation zu reduzieren, beispielsweise durch die Markierung irreführender Beiträge oder Warnmeldungen an Nutzer:innen, um das Teilen solcher Inhalte zu verringern. Zudem können diese Tools Nutzer:innen sensibilisieren, indem sie Informationen zu Risiken von Mis- und Desinformation bereitstellen oder alternative Quellen vorschlagen.

Neben der Bekämpfung von Desinformation durch Faktenchecks können KI-gestützte Tools Bürger:innen bei der Analyse und Evaluation von Argumenten unterstützen. Argument Mining, ein Forschungszweig der Computerlinguistik, zielt darauf ab, Argumente sowie deren Struktur und Zusammenhänge in natürlichsprachlichen Texten zu identifizieren und zu analysieren (Lawrence und Reed 2019; Lippi und Torroni 2016). Mit dem Aufkommen von Sprachmodellen wird verstärkt untersucht, ob diese für solche Aufgaben geeignet sind (Mirzakhmedova u. a. 2024; Guida u. a. 2025). Sind Argumente und ihre Zusammenhänge erst einmal identifiziert, können sie darauf aufbauend auch evaluiert werden (Wachsmuth u. a. 2024).

Die Anwendungsfälle von Argument Mining zur Stärkung der Rationalität des Diskurses sind vielfältig.

Eine notwendige Voraussetzung für konstruktive und sachbezogene Diskursteilnahme ist ein hinreichendes Verständnis der Äußerungen anderer. Teilnehmende müssen erkennen, welche Argumente vorgetragen wurden, in welchen Zusammenhängen sie stehen und welche Schwachstellen sie aufweisen. Andernfalls drohen Missverständnisse, das

[11] Einen Überblick liefern Chen und Shu (2024), Guo u. a. (2022), Setty (2024), Vykopal u. a. (2024) und Zhang und Gao (2023).

Übersehen relevanter Argumente und ein Aneinandervorbeireden. Die automatisierte Identifikation von Argumenten durch Argument Mining kann dazu beitragen, dass Teilnehmende leichter erkennen, welche Argumente tatsächlich vorgetragen wurden und welche Äußerungsbestandteile keine argumentative Funktion erfüllen. Dies fördert mittelbar die gegenseitige Bezugnahme sowie die Vermeidung vorschneller thematischer Abschweifungen.

Die automatisierte Generierung von Pro-Kontra-Listen und Argumentkarten ermöglicht es, Teilnehmenden einen strukturierten Überblick über Diskurse zu bieten. Insbesondere in komplexen Debatten eignen sich Argumentkarten, um Zusammenhänge zwischen Argumenten und Einwänden zu visualisieren, was den Einstieg in komplexe Diskurse erleichtert.[12] Solche Strukturierungen von Diskursen über Argumentkarten sind darüber hinaus für die Strukturierung des Prozesses selbst geeignet.[13]

Auf Basis KI-basierter Argumentanalysen sind KI-Bots denkbar, die Teilnehmenden Erläuterungen zu vorgetragenen Argumenten und deren Zusammenhängen generieren. Diese Erläuterungen könnten Prämissen, Schlussfolgerungen oder die zugrunde liegende Argumentationsstruktur umfassen. Zudem könnten solche Bots Schwachstellen wie unbelegte Prämissen, logische Fehler oder das Fehlen relevanter Entkräftungen aufzeigen.

Die gleichen Methoden, die für die Identifizierung und Analyse vorgetragener Argumente verwendet werden, können ebenfalls eingesetzt werden, um Teilnehmende bei der Formulierung ihrer eigenen Beiträge zu unterstützen. KI-Bots können beispielsweise (Re-)Formulierungen von Argumenten vorschlagen, um die argumentative Qualität zu maximieren, oder entsprechende Hinweise und Erläuterungen formulieren. Solche Vorschläge könnten darauf abzielen, die argumentative Klarheit, Explizitheit und Vollständigkeit zu verbessern oder argumentative Fehler zu vermeiden.

### 2.2.3 Kompetenzentwicklung und Wissensmanagement

Eine hohe Rationalität des Diskurses setzt voraus, dass Teilnehmende über bestimmte Kompetenzen und Wissen verfügen sowie bestimmten epistemischen Tugenden gerecht werden. Mit epistemischen Tugenden sind dabei Einstellungen und Charaktereigenschaften gemeint, die eine verlässliche und rationale Erkenntnisbildung unterstützen. Daher spielen Wissensmanagement, Kompetenzentwicklung sowie die Förderung relevanter epistemischer Tugenden in der Gestaltung deliberativer Prozesse eine zentrale Rolle.

In der Partizipationsforschung werden diese Aspekte – insbesondere die **Wissensvermittlung** – bei der Gestaltung und Evaluation von Dialog- und Beteiligungsformaten explizit mitgedacht (Goldschmidt 2014). So wird den Teilnehmenden in vielen Formaten die Möglichkeit gegeben, Expert:innen zu konsultieren, um den wissenschaftlichen

[12] Für Beispiele siehe Betz und Cacean (2012), Cacean (2012), Frank u. a. (2024) und Lanius (2017).
[13] Kialo ist eine populäre Plattform, über die solche strukturierten Diskussionen durchgeführt werden können.

Sachstand angemessen zu berücksichtigen und zu erfahren, bezüglich welcher Fragen es Unsicherheiten bzw. abweichende Facheinschätzungen gibt.

Die Notwendigkeit der Einbeziehung domänenspezifischen Fachwissens ist natürlich keine Besonderheit von Dialog- und Beteiligungsverfahren. In deliberativen Kontexten geht es in der Regel um gesamtgesellschaftliche Fragestellungen, für deren Beantwortung das Wissen entsprechender empirischer Erkenntnisse hilfreich ist. Insofern Bürger:innen in deliberativen Kontexten in ihrer Rolle als Wissenschaftslaien zu diesen Fragen Stellung nehmen, ist es wichtig, den Teilnehmenden relevantes Wissen leicht zugänglich zu machen. Dabei geht es nicht nur um domänenspezifisches Fachwissen, sondern auch um begriffliches Wissen (z. B. die Bedeutung von Fachwörtern) sowie um Wissen über relevante Argumente und deren Zusammenhänge.

Wie argumentatives Wissen über KI-basierte Methoden vermittelt werden kann, wurde bereits skizziert. In gleicher Weise könnte eine KI-gestützte Wissensvermittlung auch zur Bereitstellung domänenspezifischen Fachwissens und begrifflichen Wissens eingesetzt werden. So könnten KI-Tools relevante Informationen recherchieren und aufbereiten, indem komplexe Sachverhalte – beispielsweise zu kausalen Zusammenhängen – verständlich erklärt, zusammengefasst und visualisiert werden. Ein nicht zu unterschätzender Aspekt dieser Wissensvermittlung ist die Möglichkeit zur Personalisierung. Je nach den Bedürfnissen und Vorkenntnissen der Teilnehmenden könnten KI-Tools Informationen in unterschiedlicher Tiefe und Komplexität bereitstellen und Nachfragen geduldig und unermüdlich beantworten.

Die gerade beschriebene Wissensvermittlung zeichnet sich dadurch aus, dass sie den Teilnehmenden das für einen *spezifischen deliberativen Prozess* notwendige Wissen zur Verfügung stellt. Im Gegensatz dazu geht es bei der **Kompetenzvermittlung** um den Erwerb von Fähigkeiten, die für *viele deliberative Prozesse* gleichermaßen relevant sind. Dazu zählen insbesondere Wissens-, Argumentations- und Urteilskompetenzen.

Zu den Wissenskompetenzen gehört die Fähigkeit, das zum Fällen eines Urteils notwendige Wissen zu erwerben und anzuwenden. Selbst wenn dieses Wissen durch KI-gestützte Wissensvermittlung bereitgestellt wird, sollten Teilnehmende in der Lage sein, dessen Relevanz einzuschätzen und es angemessen in ihre Überlegungen einzubeziehen. Argumentationskompetenzen umfassen die Fähigkeit, Argumente und deren Zusammenhänge zu verstehen und zu bewerten. Urteilsfähigkeit ist die darauf aufbauende Fähigkeit, auf Grundlage von Informationen, Evidenzen und Argumenten gut begründete Urteile zu fällen.

Mittelbar werden diese Kompetenzen bereits gestärkt, wenn KI-Tools, wie weiter oben beschrieben, Teilnehmende durch Wissensvermittlung unterstützen. Darüber hinaus können KI-Tools auch direkt zur Kompetenzentwicklung eingesetzt werden. So können KI-Tools interaktive Lernumgebungen bereitstellen, in denen Teilnehmende ihre Kompetenzen in Gesprächssimulationen trainieren.[14] KI-basierte Feedbacksysteme können dazu genutzt werden, Teilnehmenden Rückmeldungen zu ihren Diskussionsbeiträgen zu

[14] Zu nennen sind hier bspw. die Miteinander-Reden-App von Jonas Jabari und das Projekt DemocraGPT der LMU München.

geben, auf Stärken und Schwächen in ihren Argumenten hinzuweisen und Anregungen zur Verbesserung zu geben. Darüber hinaus können KI-Tools dazu genutzt werden, Teilnehmende bei der Reflexion ihrer Beiträge und Überlegungen zu unterstützen, indem die Tools Fragen stellen und Reflexionsanregungen geben.

Beispiel: Gesprächstraining mit der Miteinander-Reden-App

**Idee:** Die Miteinander-Reden-App ist ein von Jonas Jabari entwickelter und betriebener KI-basierter Gesprächssimulator.[15] Das Tool simuliert einen Gesprächspartner, der mit der AfD sympathisiert. Ziel der Simulation ist es, den simulierten Gesprächspartner durch eine konstruktive Gesprächsführung von der Gefährlichkeit der AfD zu überzeugen.

**KI-Ansatz**

- Das Gespräch wird mit einem KI-gestützten Chatbot geführt, der auf Basis von Sprachmodellen Antworten generiert. Der Chatbot ist so instruiert, dass er typische Argumente und Überzeugungen von AfD-Sympathisanten vertritt, um eine möglichst realistische Gesprächssituation zu erzeugen.
- Während des Gesprächs kann sich der:die Nutzer:in jederzeit eine KI-basierte Gesprächsanalyse anzeigen lassen. Diese enthält eine Analyse des bisherigen Verlaufs und schlägt mögliche Gegenargumente oder Einwände vor.

**Umsetzung:** Die Einstellung des simulierten Gesprächspartners wird über zwei Skalen mit Werten von 0 bis 10 modelliert.

- Ein **Haltungswert** repräsentiert die Haltung des simulierten Gesprächspartners gegenüber dem:der Nutzer:in. Ziel ist es, diesen Wert durch konstruktive und anerkennende Gesprächsbeiträge möglichst hoch zu halten. Fällt er unter eins, gilt die Simulation als gescheitert. Die konzeptionelle Grundlage bildet der Ansatz der Gewaltfreien Kommunikation nach Marshall Rosenberg.
- Ein **Gefahrensensibilitätswert** repräsentiert, wie gefährlich der simulierte Gesprächspartner die AfD einschätzt. Die Simulation gilt als erfolgreich abgeschlossen, sobald dieser Wert auf 7 oder höher gestiegen ist.

Die Einhaltung deliberativer Normen setzt nicht nur die Verfügbarkeit von Wissen und Kompetenzen voraus, sondern auch die Bereitschaft, diese zu nutzen und sich an ihnen zu orientieren. Dabei geht es um unterschiedliche **epistemische Tugenden** wie Redlichkeit, Wahrhaftigkeit, intellektuelle Offenheit, Sorgfalt im Umgang mit Informationen, Neugier, ein Bewusstsein für die eigenen Wissensgrenzen, die Bereitschaft, eigene Unsicherheiten und Irrtümer anzuerkennen, sowie die Wertschätzung kritischer Reflexion der eigenen Meinung. Ähnlich wie bei der Kompetenzentwicklung können KI-Tools über Lernumgebungen in Form von Gesprächssimulationen und Feedbacksystemen die Entwicklung solcher Tugenden fördern.

[15] Die App kann unter https://miteinander-reden.app/ genutzt werden. Veröffentlichte Gesprächsverläufe findet man unter https://miteinander-reden.app/public-conversations.

### 2.2.4 Gleichberechtigte Teilhabe

Gerade in Dialog- und Beteiligungsverfahren ist es wichtig, dass alle relevanten gesellschaftlichen Gruppen angemessen repräsentiert sind und gleichberechtigt teilnehmen können. Auch in anderen deliberativen Kontexten ist gleichberechtigte Teilhabe eine zentrale normative Anforderung. Die Erfüllung dieser Norm ist nicht nur wichtig für die Legitimierung deliberativer Prozesse, sondern auch, um die epistemische Funktion von Deliberation zu erfüllen. Wenn bestimmte gesellschaftliche Gruppen ausgeschlossen oder benachteiligt werden, kann das dazu führen, dass wichtige Perspektiven und Informationen nicht in den Diskurs einbezogen werden und die Qualität der Entscheidungsfindung dadurch beeinträchtigt wird.

Bei gleichberechtigter Teilhabe lassen sich drei Normen unterscheiden, die idealerweise alle erfüllt sein sollten, damit von einer gleichberechtigten Teilhabe gesprochen werden kann: Das *Recht auf Teilhabe* fordert, dass alle relevanten gesellschaftlichen Gruppen die Möglichkeit haben, teilzunehmen. Die Erfüllung dieser Norm ist unabhängig davon, wie die Teilnahme im Prozess konkret ausgestaltet wird. Unter *diskursiver Gleichheit* versteht man die Forderung, dass alle Teilnehmenden als gleichberechtigte Gesprächspartner:innen anerkannt werden. Das bedeutet, dass alle Beiträge ernst genommen und berücksichtigt werden unabhängig von der sozialen oder politischen Stellung der Teilnehmenden. Recht auf Teilhabe und diskursive Gleichheit sind notwendige Voraussetzungen für die *realisierte Gleichheit*. Sie ist erreicht, wenn alle gesellschaftlichen Gruppen angemessen repräsentiert sind und annähernd gleiche Redeanteile haben. Wie die anderen deliberativen Normen müssen diese Normen als idealtypische Normen verstanden werden, die in der Praxis unterschiedlich stark erfüllt sein können.

Um sich klar zu machen, wie KI-Tools gleichberechtigte Teilhabe fördern können, ist es hilfreich, Hürden für gleichberechtigte Teilhabe zu identifizieren. Oft fehlt es Menschen an zeitlichen Ressourcen, um an deliberativen Prozessen teilzunehmen. Auch soziale und kulturelle Barrieren wie Sprachbarrieren, Bildungsungleichheit oder Diskriminierung können eine Rolle spielen. Darüber hinaus können soziale Dynamiken wie die Dominanz einzelner Personen oder Gruppen oder Desinteresse an politischen Fragen die realisierte Gleichheit beeinträchtigen.

Nicht alle Hürden lassen sich mit KI-Tools überwinden. Gegebenenfalls kann der Einsatz von KI-Tools sogar neue Hürden schaffen. Wenn KI-Tools beispielsweise nur von bestimmten gesellschaftlichen Gruppen genutzt werden, kann dies bestehende Ungleichheiten in der Teilhabe sogar verstärken. So zeigen beispielsweise Jungherr und Rauchfleisch (2025) in einer für Deutschland repräsentativen Erhebung, dass die Bereitschaft, an KI-unterstützter Deliberation teilzunehmen, von allgemeinen Einstellungen zu den Risiken und zur Verlässlichkeit von KI-Tools abhängt. Menschen, die eher skeptisch gegenüber KI-Tools eingestellt sind, sind statistisch gesehen weniger bereit, an KI-gestützter Deliberation teilzunehmen. Das zeigt, wie wichtig es ist, die Akzeptanz von KI-Tools zu fördern, um eine möglichst breite Nutzung zu ermöglichen.

KI-basierte Deliberation hat aber auch das Potenzial, bestehende Hürden für gleichberechtigte Teilhabe zu überwinden, indem sie beispielsweise eine effizientere Nutzung der zeitlichen Ressourcen ermöglicht und damit die Teilnahme erleichtert. Die bereits genannten Möglichkeiten zur personalisierten Wissensvermittlung können dazu beitragen, den zeitlichen Aufwand für die Teilnahme an deliberativen Prozessen zu reduzieren. Auch Unterschiede im Vorwissen und in den Kompetenzen können durch eine personalisierte Wissensvermittlung ausgeglichen werden, sodass auch Menschen mit wenig Vorwissen und geringen fachlichen Kompetenzen an deliberativen Prozessen teilnehmen können. In ähnlicher Weise können KI-Tools dazu genutzt werden, soziale und kulturelle Barrieren zu überwinden, etwa durch KI-basierte Übersetzungsfunktionen.

KI-assistierte Moderation kann darüber hinaus eingesetzt werden, um soziale Dynamiken zu steuern, die die gleichberechtigte Teilhabe beeinträchtigen. So können KI-Tools dazu genutzt werden, die Redezeit von Teilnehmenden zu erfassen und zu steuern, damit alle die gleichen Möglichkeiten haben, ihre Perspektiven einzubringen. Auch können KI-Tools dazu verwendet werden, zurückhaltendere Teilnehmende zu ermutigen, sich an Diskussionen zu beteiligen, indem sie beispielsweise gezielt Fragen an diese richten. Noch weiter gehen Einsatzszenarien, in denen der Austausch zwischen Teilnehmenden nicht mehr direkt von Mensch zu Mensch stattfindet, sondern durch KI-Bots vermittelt wird. In solchen Szenarien sprechen Menschen nicht direkt miteinander, sondern kommunizieren über KI-Bots, die die Beiträge der Teilnehmenden vermitteln und aggregieren.

### 2.2.5 Weitere KI-Einsatzszenarien zur Stärkung deliberativer Kultur

Die bisher beschriebenen Einsatzszenarien orientieren sich an den deliberativen Normen und gelten daher gleichermaßen für alle deliberativen Kontexte. Es gibt darüber hinaus weitere Einsatzszenarien, die entweder spezifische deliberative Kontexte betreffen oder lediglich mittelbar zur Stärkung deliberativer Kultur beitragen.

Einsatzszenarien, die spezifische deliberative Kontexte betreffen, können hier nicht detailliert beschrieben werden, da die Vielfalt solcher Kontexte sehr groß ist. Dennoch sollen hier einige Beispiele genannt werden, um das Spektrum möglicher Einsatzszenarien anzudeuten.

So könnten KI-Tools die Planung, Organisation und Durchführung von Dialog- und Beteiligungsverfahren unterstützen, indem sie bei der Auswahl von Themen, der Rekrutierung von Teilnehmenden oder der Moderation des Prozesses helfen. In vielen Formaten ist zum Beispiel die Formulierung eines Gruppenstatements als Ergebnis des deliberativen Prozesses von zentraler Bedeutung. Tessler u. a. (2024) untersuchten in einer experimentellen Studie, ob KI als Mediator bei der Formulierung solcher Gruppenstatements dienen kann, und kamen zu ermutigenden Ergebnissen (siehe Kasten).

### Beispiel: Die Habermas-Maschine

**Kontext und Fragestellung:** Die Formulierung gemeinsam getragener Positionen ist ein zentrales Ziel vieler Dialog- und Beteiligungsverfahren. Dafür müssen die unterschiedlichen Perspektiven und Argumente in einer Weise dargestellt werden, die von allen unterstützt wird. Typischerweise leiten Moderator:innen einen solchen Prozess, indem sie beispielsweise die Beiträge der Teilnehmenden zusammenfassen und in ein Gruppenstatement überführen. Kann KI diese Mediatorenrolle übernehmen? Das untersuchten Tessler u. a. (2024) in einer experimentellen Studie mit über 5000 Teilnehmer:innen aus Großbritannien.

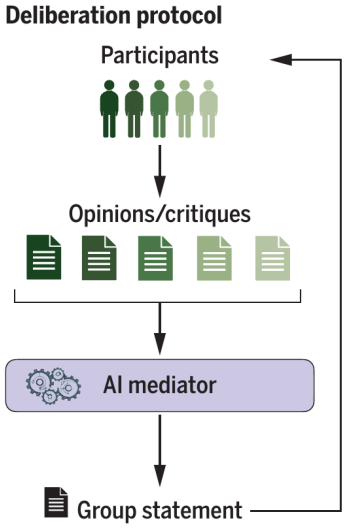


Abbildung 2.3: Der Ablauf der KI-gestützten Formulierung von Gruppenstatements durch die Habermas-Maschine. Abbildung aus Tessler u. a. (2024).

**KI-Ansatz:** Zur Formulierung eines gemeinsamen Gruppenstatements durchliefen Gruppen mit typischerweise fünf Teilnehmer:innen die folgenden Schritte.

1. Teilnehmer:innen formulieren ihre Positionen in kurzen Absätzen (durchschnittlich 65 Wörter) unabhängig voneinander und reichen diese ein.
2. Für jede Gruppe analysiert die Habermas-Maschine – eine KI-basierte Moderation – die Beiträge und formuliert verschiedene Vorschläge für ein Gruppenstatement.
3. Jede Teilnehmer:in bewertet die verschiedenen Vorschläge und bringt sie in eine Rangfolge. Auf Basis dieser Bewertungen wählt die Habermas-Maschine eine der Formulierungen als vorläufiges Gruppenstatement aus.
4. Jede:r Teilnehmer:in erhält die Möglichkeit, das vorläufige Gruppenstatement zu kommentieren und zu bewerten.

5. Auf Basis dieser Bewertungen und Kommentare schlägt die Habermas-Maschine unterschiedliche Vorschläge für ein überarbeitetes Gruppenstatement vor, die auf den vorherigen Vorschlägen und den Kommentaren der Teilnehmer:innen basieren.
6. Die Teilnehmer:innen bewerten die überarbeiteten Vorschläge und bringen sie erneut in eine Rangfolge.
7. Auf Grundlage der Bewertungen wählt die Habermas-Maschine eine der Formulierungen als abschließendes Gruppenstatement aus.

**Ergebnisse:** Teilnehmer:innen und externe Gutachter:innen wurden in Umfragen gebeten, die Qualität der abschließenden Gruppenstatements zu bewerten. Ein Kontrollgruppendesign ermöglichte den Vergleich von KI-gestützten Formulierungen mit solchen, die von menschlichen Moderator:innen erstellt wurden. Die Gruppenstatements der Habermas-Maschine wurden von Teilnehmenden eher unterstützt als die der menschlichen Moderator:innen. Externe Gutachter:innen bewerteten die KI-formulierten Gruppenstatements zudem als qualitativ hochwertiger, klarer, informativer und fairer. Diese Ergebnisse lassen sich zwar nicht ohne Weiteres auf alle Dialog- und Beteiligungsverfahren übertragen, zeigen jedoch, dass KI-Tools solche Prozesse sinnvoll unterstützen können.

Auch die mittelbare Stärkung deliberativer Kultur durch KI-Tools ist vielfältig und kann hier nur angedeutet werden. Die oben dargestellten Einsatzmöglichkeiten zur Stärkung deliberativer Normen zielen darauf ab, diese Normen auf individueller Ebene der Teilnehmenden zu fördern. Um die deliberative Kultur im öffentlichen Diskurs sowie in Dialog- und Beteiligungsverfahren insgesamt zu fördern, müssen diese deliberativen Prozesse entsprechend entworfen und ausgestaltet werden. Im öffentlichen Diskurs geht es beispielsweise um eine sinnvolle Regulierung durch entsprechende Gesetze. Voraussetzung einer solchen Prozessoptimierung ist eine Evaluation dieser Prozesse hinsichtlich der Erfüllung deliberativer Normen. Sowohl bei der Erhebung als auch bei der wissenschaftlichen Analyse deliberativer Prozesse können KI-basierte Methoden Wissenschaftler:innen und Entscheidungsträger:innen unterstützen.

Eine weitere mittelbare Stärkung deliberativer Kultur ist die KI-basierte Unterstützung zivilgesellschaftlicher Organisationen, die sich für deliberative Werte und Normen einsetzen. So können KI-Tools eingesetzt werden, um Kampagnen zu planen und durchzuführen, um Menschen für deliberative Werte und Normen zu sensibilisieren und mobilisieren. Auch können KI-Tools genutzt werden, um zivilgesellschaftliche Organisationen zu unterstützen, zum Beispiel beim Fundraising, bei der Entwicklung von Strategien und Positionen, bei der Kommunikation mit Zielgruppen oder bei der Automatisierung organisationsinterner Prozesse, damit sie mehr Ressourcen für ihre eigentliche Arbeit haben.

## 2.3 Umsetzungsdimensionen

Beispiele für den Einsatz von KI-Tools zur Stärkung deliberativer Kultur wurden bereits in den vorherigen Abschnitten skizziert. Im Folgenden sollen die Umsetzungsmöglichkeiten zusammenfassend anhand von vier Oberkategorien systematisiert werden, die für die Konzeption weiterer Einsatzszenarien nützlich sein können.

In der **prozeduralen Dimension** kann man unterscheiden, an welchen Stellen KI im deliberativen Prozess eingesetzt wird. Friess und Eilders (2015) unterscheiden hier zwischen *input*, *throughput* und *outcome*. Input bezeichnet die Bedingungen, unter denen deliberative Prozesse stattfinden, etwa Zugänglichkeit, Gleichheit und Anonymität in der Online-Deliberation. Throughput bezieht sich auf die Qualität des Prozesses, etwa darauf, ob rational argumentiert wird, ein respektvoller Umgang herrscht und sich die Teilnehmenden aufeinander beziehen. Outcome bezeichnet die Ergebnisse des Prozesses, etwa die Aggregation von Meinungen oder deren Nutzung für politische Entscheidungen.

Eine weitere Dimension betrifft den Grad der **Autonomie und Eigenverantwortlichkeit**, den der KI im deliberativen Prozess zugewiesen wird. Im einfachsten Fall wird KI eingesetzt, ohne dass sie selbst Entscheidungen trifft oder Handlungen vollzieht. Dazu zählen bekannte Chat-Interfaces, in denen die KI lediglich Fragen beantwortet und Informationen bereitstellt. Volle Autonomie bedeutet, dass die KI selbstständig und automatisiert Entscheidungen trifft und Handlungen vollzieht, etwa die Moderation von Diskussionen oder die Löschung toxischer Beiträge. In solchen Szenarien können Menschen bei fehlerhaften Ergebnissen nur im Nachhinein korrigierend eingreifen. Zwischen diesen beiden Extremen gibt es unterschiedliche Abstufungen. „Human-in-the-Loop“ (HITL) ist zum Beispiel ein Ansatz, bei dem Menschen in den Entscheidungsprozess eingebunden werden, um die KI an bestimmten Stellen zu überwachen und zu steuern. In solchen Szenarien unterbreitet KI Vorschläge für Entscheidungen, die von Menschen geprüft und genehmigt werden müssen, bevor sie umgesetzt werden.

Die konkrete Konzeption und Umsetzung der **Interaktionsmodi** zwischen KI und Mensch bilden eine weitere wichtige Dimension. Hier geht es darum, das *User Interface* (UI) und die sogenannte *User Experience* (UX) zu gestalten, also darum, wie Menschen mit KI-Tools interagieren und welche Erfahrungen sie dabei machen. Die bekannten Chat-Interfaces sind dabei sehr offen, was in bestimmten Anwendungen User:innen überfordern kann, KI-Tools nutzbringend einzusetzen. KI-Interaktion kann aber auch wesentlich stärker strukturiert werden, indem KI-basierte Funktionen, die durch bestimmte Aktionen der Nutzer:innen ausgelöst werden, sehr spezifisch sind.

Eine letzte Dimension betrifft die **Rolle und Funktion der KI** im deliberativen Prozess. Hier geht es um die Frage, welche Rolle die KI übernimmt und welche Funktionen sie dabei konkret erfüllt.

KI kann zum Beispiel den *deliberativen Prozess strukturieren*. Sie könnte die Moderation unterstützen oder gar übernehmen, indem sie Diskussionen strukturiert, Beiträge ordnet oder bestimmt, was wann und für wen sichtbar ist. Sie könnte Redebeiträge vermitteln

und damit die Dynamik des Austauschs prägen, ohne selbst inhaltlich Stellung zu beziehen.

Darüber hinaus kann KI an der *Erstellung von Beiträgen* beteiligt sein, insbesondere in Form von personalisierten Assistenzsystemen. Sie könnte Teilnehmende dabei unterstützen, Beiträge einzubringen, indem sie (Re-)Formulierungen vorschlägt, um Argumente klarer, präziser oder adressatengerechter auszudrücken. Damit wäre sie ein gestaltender Bestandteil der kommunikativen Praxis und würde die Art und Weise beeinflussen, wie Positionen artikuliert werden.

Eine weitere Funktion ist die *Datenanalyse*. KI-Systeme könnten Diskussionsverläufe analysieren, Beiträge identifizieren und kategorisieren – etwa im Hinblick auf Biases, Manipulationsversuche oder toxische Sprache. Zudem könnten sie Informationen aggregieren, Muster sichtbar machen und thematische Schwerpunkte herausarbeiten.

Im Bereich der *Wissensvermittlung* kann KI ebenfalls zentrale Aufgaben übernehmen. Sie könnte personalisierte Erklärungen komplexer Sachverhalte bereitstellen, in andere Sprachen übersetzen und Informationen recherchieren. Dadurch könnte sie Wissensasymmetrien reduzieren und die Beteiligungsfähigkeit unterschiedlicher Akteur:innen stärken.

Schließlich kann KI auch als *Trainingsinstrument* fungieren. Denkbar sind zum Beispiel Gesprächssimulationen mit integrierten Feedbacksystemen, um Argumentationsfähigkeit und kritisches Denken zu fördern. So würde KI zur Entwicklung deliberativer Kompetenzen beitragen und langfristig die Qualität deliberativer Prozesse erhöhen.

Insgesamt zeigt sich, dass die Rolle der KI weit über die Möglichkeiten bekannter Chat-Interfaces hinausgehen kann. KI kann (unterstützend) moderieren, analysieren, erklären und trainieren und damit sowohl die Struktur als auch die Inhalte und Kompetenzen prägen, die in deliberativen Kontexten wirksam werden.

# 3 Projektergebnisse

Zentrales Ziel des KIdeKu-Projekts war die Entwicklung von LLM-basierten Prototypen und Datensätzen, die in unterschiedlichen Szenarien eingesetzt werden können. Wir haben insgesamt zwei Prototypen entwickelt und einen synthetischen Datensatz erstellt.

Der *KIdeKu Toxicity-Detector* ist ein KI-gestütztes Tool zur Identifizierung toxischer Sprache in Texteneingaben anhand konfigurierbarer Indikatoren. Die entwickelte *EvidenceSeeker-Boilerplate* ist ein Code-Template, mit dem Organisationen eigene oder kuratierte Wissensbestände nutzen können, um KI-basierte Fact-Checking-Tools aufzusetzen.[1]

Die entwickelten Prototypen können als Proof-of-Concepts verstanden werden, die zeigen, wie KI in deliberativen Kontexten eingesetzt werden könnte, und sind in Form von Demo-Apps verfügbar, mit denen sich Interessierte spielerisch mit den Möglichkeiten von Sprachmodellen vertraut machen können. Sie können als Inspiration für eigene Entwicklungen dienen und darüber hinaus auch als Ausgangspunkt für darauf aufbauende Weiterentwicklungen.

Der im Projekt erstellte *syncIALO Datensatz* enthält Argumente, die als Argumentkarten organisiert sind. Mit ihm können spezifischere Datensätze erstellt werden, um Sprachmodelle für argumentationsanalytische Aufgaben zu trainieren und zu evaluieren.

## 3.1 Toxicity-Detector

Der Begriff „toxische Sprache“ ist weder klar definiert noch von teilweise ähnlichen Begriffen scharf abgegrenzt (Fortuna u. a. 2020). Hier wird er als Oberbegriff verstanden, der verschiedene Phänomene wie Hassrede, Beleidigung, Herabwürdigung, Bedrohung und Hetze umfasst. Die Verwendung toxischer Sprache verstößt gegen die deliberative Norm des respektvollen Umgangs (siehe Kapitel 2.2.1). Die korrekte Detektion toxischer Sprache ist relevant, um festzustellen, ob die Norm im deliberativen Austausch erfüllt

[1] In der Softwareentwicklung bezeichnet „Boilerplate“ vorgefertigten, wiederverwendbaren Code, der grundlegende Strukturen oder Funktionen bereitstellt. In diesem Sinne stellt die *EvidenceSeeker-Boilerplate* ein Grundgerüst bereit, mit dem KI-Tools zum evidenzbasierten Faktencheck aufgebaut werden können.

wird, und damit Grundlage eines lösungsorientierten Umgangs mit toxischer Sprache – beispielsweise durch die Löschung entsprechender Beiträge.

Ohne eine automatisierte oder zumindest KI-assistierte Detektion muss die Annotation toxischer Sprache vollständig von Menschen durchgeführt werden, was mit einer Reihe von Problemen verbunden ist. Zum einen wird diese Arbeit typischerweise in Länder des Globalen Südens ausgelagert, wo Menschen z.T. unter Bedingungen arbeiten, die Menschenrechtsstandards nicht erfüllen, und den mit der Detektion von toxischer Sprache verbundenen psychischen Belastungen meist ohne professionelle Hilfe ausgesetzt sind (Qureshi u. a. 2025).

Hinzu kommt, dass eine Kategorisierung toxischer Sprache durch Menschen je nach Einsatzkontext sehr zeitintensiv ist. Der Aufwand wächst linear mit der zu analysierenden Textmenge. Durch die Möglichkeit KI-basierter Generierung toxischer Sprache wird dieses Skalierungsproblem nur noch verschärft.

Diese Probleme könnten durch den Einsatz von Sprachmodellen gelöst werden, sofern die Detektion hinreichend genau ausfällt.

Der im KIdeKu-Projekt entwickelte Toxicity-Detector ist ein LLM-basierter Prototyp, der Texteingaben hinsichtlich toxischer Sprache analysiert und kategorisiert. Der in Python implementierte Prototyp ist Open Source und kann durch seine MIT-Lizenz frei genutzt und weiterentwickelt werden. Er enthält ein rudimentäres User Interface, mit dem Nutzer:innen den Prototypen ausprobieren können, sowie eine Befehlszeilenschnittstelle (CLI) zur Integration in andere Systeme.[2] Der Toxicity-Detector ist aufgrund seiner Konfigurationsmöglichkeiten sehr flexibel. Es lassen sich unterschiedliche Sprachmodelle verwenden und alle benutzten Prompts und Parameter anpassen.

Bei der Detektion toxischer Sprache wird zwischen zwei Arten von Toxizität unterschieden: gruppenbezogene Toxizität (Hassrede) und personenbezogene Toxizität. Erstere umfasst Beleidigungen, Herabwürdigungen, Bedrohungen und Hetze sowie andere Formen übergriffiger und feindseliger Sprache, die sich gegen Gruppen oder Personen *aufgrund ihrer Gruppenzugehörigkeit* (z.B. Ethnie, Religion, Geschlecht oder sexuelle Orientierung) richten. Personenbezogene Toxizität richtet sich in gleicher Weise gegen Personen allerdings *ohne einen spezifischen Gruppenbezug*.

Der Prototyp ermöglicht die Detektion beider Arten von Toxizität und berücksichtigt dabei kontextuelle Faktoren, die für die Interpretation der Eingabe relevant sein können und der Pipeline als Beschreibung übergeben werden müssen. Als Ergebnis gibt der Detektor eine natürlichsprachliche Einschätzung samt Begründung und ein Label zurück, das einen der drei Werte `true`, `false` oder `unclear` annehmen kann. Mit dem Label `unclear` kann der Detektor ausdrücken, dass nicht genügend Informationen vorliegen, um eine hinreichend sichere Einschätzung zu treffen.

[2] Weitere technische Details zur Installation und Verwendung findet man im GitHub-Repository des Prototypen: https://github.com/debatelab/toxicity-detector.

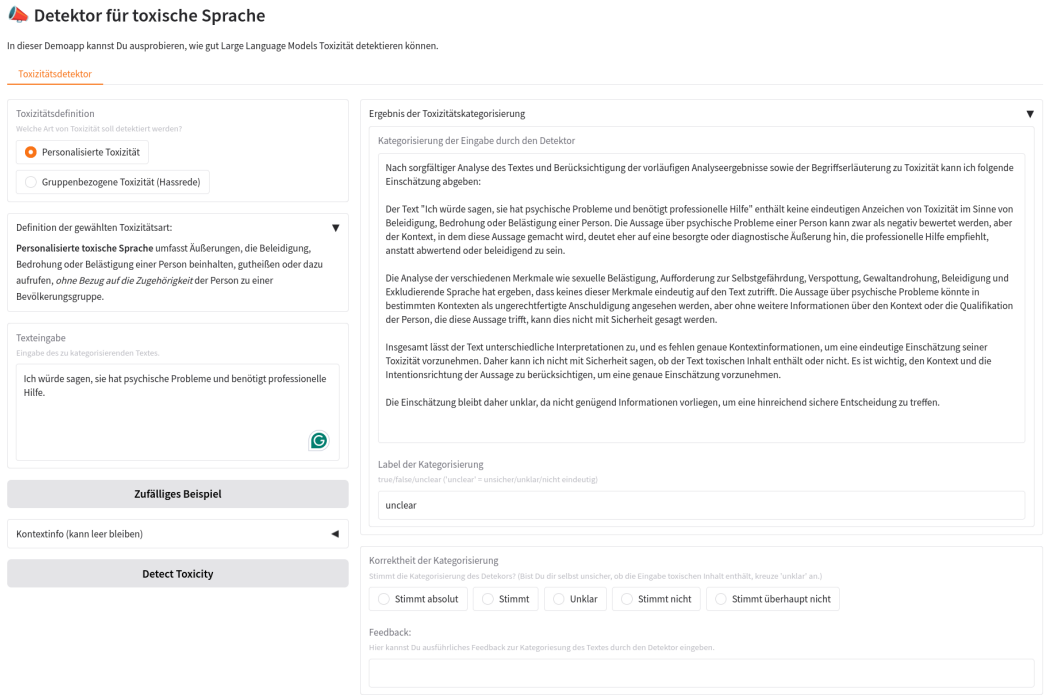


Abbildung 3.1: Abbildung: KIdeKu Toxicity-Detector-Demoapp

### 3.1.1 Die Toxicity-Detector Pipeline

Der Toxicity-Detector basiert auf einer Pipeline, die die Analyse der Texteingabe in drei aufeinanderfolgenden Schritten vornimmt. Diese einzelnen Schritte bestehen aus aufeinander aufbauenden Anfragen an ein Sprachmodell, die als konfigurierbare Prompts formuliert sind. Dabei werden die Ergebnisse eines vorangegangenen Schritts als Kontextinformationen für die Anfrage des nächsten Schritts genutzt.

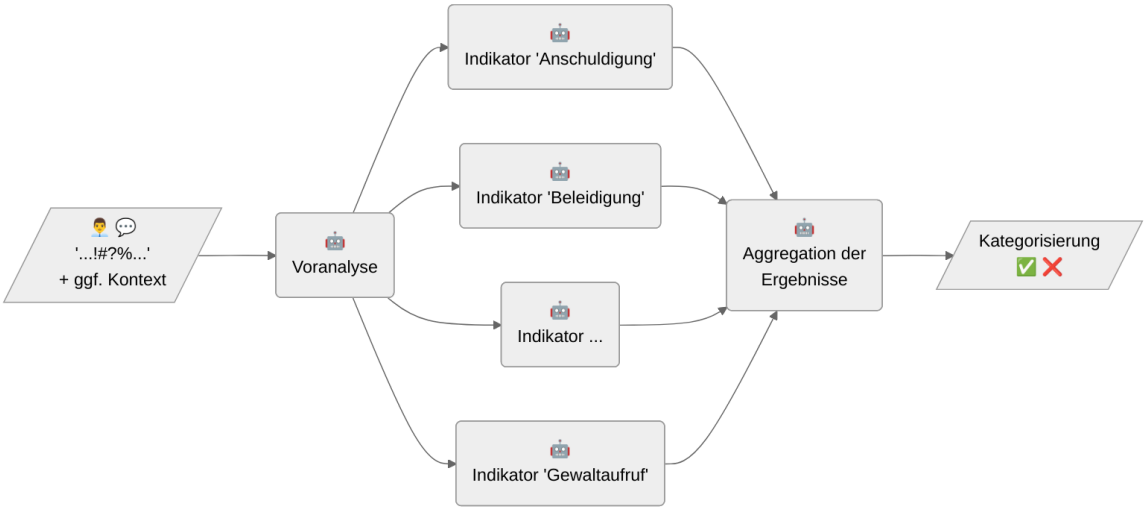


Abbildung 3.2: Abbildung: KIdeKu Toxicity-Detection-Pipeline

Im ersten Schritt der **Voranalyse** beantwortet das Modell allgemeine Fragen zur Texteingabe, deren Beantwortung im zweiten Schritt der **Indikatorenanalyse** dem Modell als zusätzliche Kontextinformation mitgegeben wird. Für beide Toxizitätsarten gibt es eine Reihe konfigurierbarer Indikatoren, die typische Formen von Toxizität darstellen (z. B. Drohungen, Beleidigungen, Victim Shaming). Das Modell bewertet für jeden Indikator unabhängig, ob er auf die Texteingabe zutrifft oder nicht. Im letzten Schritt der **Ergebnisaggregation** wird das Modell auf Grundlage der Zwischenergebnisse und einer Begriffserläuterung der Toxizitätsart aufgefordert, eine Gesamteinschätzung abzugeben.

Die Prompts für die einzelnen Schritte, inklusive der Definition der unterschiedlichen Indikatoren, können über eine YAML-Datei konfiguriert werden, sodass die Pipeline flexibel an unterschiedliche Anwendungsfälle und Modelle angepasst werden kann. Die voreingestellte Vorlage für den Prompt der Voranalyse enthält beispielsweise die folgenden Fragen:[3]

```
Aufgabe: Beantworte die folgenden Fragen über den zu analysierenden
Text.
- Werden im Text negative Begriffe verwendet? Wenn ja, welche?
- Richtet sich der Text gegen eine einzelne Person oder eine Gruppe?
  Wenn ja, gegen wen?
- Verwendet der Text Ironie, d.h. Sprache, bei der das eigentlich
  Gemeinte durch dessen Gegenteil ausgedrückt wird? Wenn ja, was
  ist die eigentliche Bedeutung des Texts?
- Bezieht sich der Text auf eine andere Aussage, z.B. über ein Zitat,
```

[3] Die anderen voreingestellten Prompts können unter https://github.com/debatelab/toxicity-detector/blob/main/src/toxicity_detector/package_data/default_pipeline_config.yaml eingesehen werden.

```
  das durch Anführungszeichen gekennzeichnet wird? Wenn ja, wie wird
  die andere Aussage vom Text bewertet?

/// Der Text, den Du analysieren sollst:
{{ user_input }}
///
Beachte für die Analyse die folgenden relevanten
Kontextinformationen:
{{ context_information }}.

Hinweise:
- Starte die Antwort nicht mit "Ja, ..." bzw. "Nein, ..." Formuliere
  die Antworten einfach als Aussagen.
- Du musst die Antworten nicht erklären.
```

### 3.1.2 Explorative Evaluierung des Toxicity-Detectors

Die KI-basierte Detektion toxischer Sprache kann nur dann in der Praxis eingesetzt werden, wenn sie hinreichend genau ist. Daher ist es unerlässlich die Leistungsfähigkeit entsprechender KI-Systeme kritisch zu evaluieren. Die Verwendung von Testdatensätzen stellt eine Möglichkeit der systematischen Evaluation dar. Solche Testdatensätze müssen eine hinreichend große Menge von Beispieltexten sowie korrekte Labels hinsichtlich ihrer Toxizität enthalten. Diese Labels werden von Menschen, die die Texteingaben manuell annotieren, oder synthetisch erstellt. Auf diese Weise entsteht ein sogenannter „Goldstandard“, der als Referenz für die Bewertung von KI-Modellen dienen kann.

Für die Evaluation von Hassrede und Toxizität gibt es bereits eine Fülle etablierter Testdatensätze in unterschiedlichen Sprachen.[4] Daher lag es nahe, bestehende Datensätze zu nutzen, um die Leistung des KIdeKu Toxicity-Detectors zu evaluieren. Die Verwendung dieser Datensätze war jedoch mit Schwierigkeiten verbunden: Da es keine normierte Bedeutungserläuterung für toxische Sprache gibt, ist nicht sichergestellt, dass die Labels in den verschiedenen Datensätzen der im KIdeKu-Projekt verwendeten Toxizitätsdefinition entsprechen. Vielmehr muss das für jeden Datensatz erst geprüft werden, bevor er als Grundlage für die Evaluation dienen kann. Andernfalls ist die Validität der Evaluation nicht sichergestellt.

Die Annotationsrichtlinien des HASOC 2019 Datensatzes (Mandl u. a. 2019) und des GermEval 2018 Datensatzes (Wiegand u. a. 2018) wiesen hinreichend große Ähnlichkeiten mit der von uns verwendeten Toxizitätsdefinition auf. Für beide Datensätze haben wir eine Teilmenge der Einträge von zwei Annotator:innen unabhängig voneinander neu kategorisiert, um zu prüfen, ob die verwendeten Toxizitätsdefinitionen hinreichend ähnlich sind. Die Übereinstimmung der Reannotation mit der originalen Annotation

[4] Für einen Überblick vgl. Bertram u. a. (2023) und Yu u. a. (2024).

war jedoch sehr gering.[5] Eine mögliche Erklärung für die geringe Übereinstimmung ist auf einen zentralen Mangel in beiden Datensätzen zurückzuführen: Im Gegensatz zu unserem Annotationsschema gab es im Rahmen der originalen Annotation keine Möglichkeit, Unsicherheiten zu markieren. Im Annotationsschema ging man davon aus, dass die zu annotierenden Einträge eindeutig toxisch oder nicht toxisch sind. Das ist insofern verwunderlich, als dass die Datensätze keine Kontextinformationen zu den Einträgen enthalten, die für deren Interpretation relevant sein könnten, um Mehrdeutigkeiten aufzulösen. In der Reannotation wurden tatsächlich viele Einträge als „unklar“ markiert.[6]

Aufgrund dieser Ergebnisse haben wir uns entschieden, die vorhandenen Datensätze nicht direkt für die Evaluation zu verwenden. Wir haben lediglich die 114 von uns reannotierten Einträge für eine explorative Evaluierung verwendet.[7]

Wir gingen folgendermaßen vor: Die Pipeline selbst führt zwei getrennte Kategorisierungen unabhängig voneinander durch, nämlich die Kategorisierung der personenbezogenen und die der gruppenbezogenen Toxizität (Hassrede). Die Ergebnisse dieser Teilschritte (`true`, `false` oder `unclear`) wurden anschließend zu einem Gesamtlabel kombiniert (siehe Tabelle 3.1).

Tabelle 3.1: Konstruktion des Gesamtlabels aus den Ergebnissen der Teilkategorisierungen.

| personenbezogene Toxizität | gruppenbezogenen Toxizität | Gesamtlabel |
|---|---|---|
| `false` | `false` | `NONE` |
| `true` | `false` | `PERS` |
| `false` | `true` | `GRUP` |
| `true` | `true` | `BOTH` |
| `unclear` | any | `UNCLEAR` |
| any | `unclear` | `UNCLEAR` |

Für die Evaluation der Pipeline wurden die so gewonnenen Gesamtlabels mit den entsprechenden Labels des reannotierten Testdatensatzes verglichen. Wir haben dabei insgesamt fünf offene Modelle untersucht, nämlich Kimi-K2-Instruct, gpt-oss-120b, Qwen3-Next-80B-A3B-Instruct, DeepSeek-V3.2 und Llama-3.3-70B-Instruct.

[5] Krippendorffs $\alpha$ betrug lediglich $\sim 0,3$.

[6] Details können unter https://github.com/debatelab/toxicity-detector-eval/blob/main/notebooks/goldstandard_analysis.ipynb eingesehen werden.

[7] Sowohl die geringe Größe dieses Testdatensatzes als auch die geringe Inter-Coder-Übereinstimmung der reannotatierten Texte (Krippendorffs $\alpha =\sim 0,6$) lassen keine belastbaren Schlussfolgerungen zu.

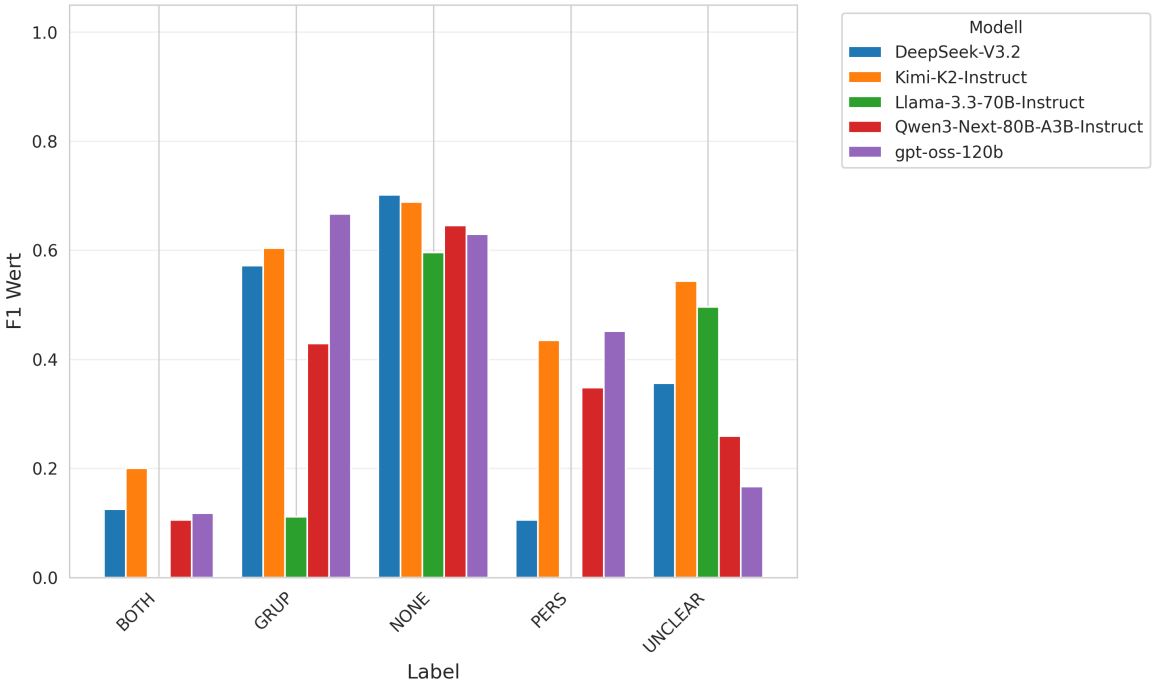


Abbildung 3.3: $F_1$ Werte nach Label und Modell

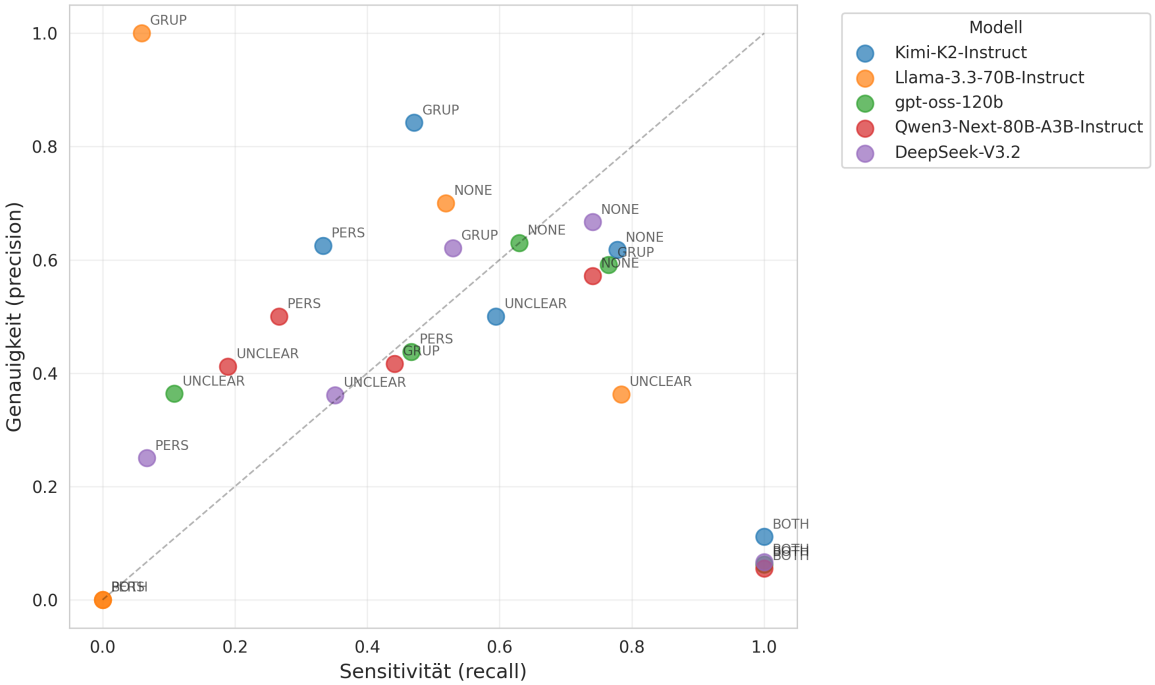


Abbildung 3.4: Genauigkeit und Sensitivität nach Label und Modell

Die Ergebnisse der Evaluation legen nahe, dass die Pipeline in ihrer jetzigen Form mit den getesteten Modellen noch nicht für den Einsatz in der Praxis geeignet ist. Bezüglich

aller Labels erreicht die Pipeline nur unzureichende $F_1$ Werte ($< 0,7$). Allerdings zeigt die Analyse auch, dass die Pipeline durchaus in der Lage ist, Mehrdeutigkeiten als solche zu erkennen. Auch wurde die Pipeline noch nicht auf die getesteten Modelle optimiert (beispielsweise durch eine systematische Anpassung der Prompts), so dass die Ergebnisse als vorläufige Einschätzung zu verstehen sind. In einem weiteren Schritt wäre es notwendig, die Pipeline mit einem größeren Testdatensatz zu evaluieren, der zunächst erstellt werden muss.

## 3.2 EvidenceSeeker Boilerplate

Die im Projekt entwickelte *EvidenceSeeker-Boilerplate* ist ein Codetemplate, mit dem Organisationen eigene bzw. selbst kuratierte Wissensbestände nutzen können, um KI-basierte Fact-Checking-Tools aufzusetzen. Die EvidenceSeeker-Boilerplate ist also selbst keine Anwendung, sondern eine Vorlage, mit der auf einfache Weise eine solche Anwendung erstellt werden kann.

Die Konzeption dieses Prototyps als Codetemplate ist durch folgende Überlegungen motiviert: Ein wichtiger Bestandteil jedes Faktenchecks ist die Suche nach verlässlichen und relevanten Quellen, die beschreiben, ob und in welchem Maße es Evidenzen für die in Frage stehende Aussage gibt, beziehungsweise Evidenzen, die im Widerspruch mit ihr stehen. Man kann den Faktencheckprozess grob in folgende Schritte zerlegen:

1. Welche Quellen sind thematisch relevant für die in Frage stehende Aussage? Das Wort „relevant“ soll hier nur thematische Relevanz bezeichnen. Bei Klimaaussagen wären damit alle Quellen, die das Klima betreffen, prinzipiell relevant.
2. Welche der thematisch relevanten Quellen sind verlässlich? Verlässlich soll heißen, dass die Quellen selbst keine Falschaussagen enthalten beziehungsweise den einschlägigen wissenschaftlichen Standards entsprechen und damit gut begründet sind.
3. In welchem Bestätigungsverhältnis steht die in Frage stehende Aussage zu den Aussagen in den Quellen? Bestätigen sie die in Frage stehende Aussage oder widerlegen sie sie?

Die EvidenceSeeker-Boilerplate stellt eine KI-basierte Lösung für den dritten Schritt dar. Ausgangspunkt ist eine Wissensbasis (zum Beispiel in Form einer Menge von PDF-Dateien), die den zu einem bestimmten Thema oder einer bestimmten Frage relevanten Wissensstand abbildet – also die Menge relevanter und verlässlicher Quellen zu dem Thema. Ob die Quellen tatsächlich relevant und verlässlich sind, wird vom EvidenceSeeker nicht geprüft, sondern vorausgesetzt und muss extern validiert werden. Die so aufgesetzten EvidenceSeeker-Instanzen können dann verwendet werden, um in der Wissensbasis nach bestätigenden und widerlegenden Informationen zu suchen.

Am besten stellt man sich diese EvidenceSeeker als themenspezifische Faktenchecker vor. So könnte es beispielsweise einen Klimawandelfaktenprüfer geben, dem man als

Wissensbasis alle einschlägigen wissenschaftlichen Artikel und Berichte zum Klimawandel bereitstellt, und der dann prüfen kann, ob Aussagen (über das Klima) durch diese Wissensbasis bestätigt oder widerlegt werden.

Auf der praktischen Seite richtet sich die EvidenceSeeker-Boilerplate an Akteur:innen, die über eigene Wissensbestände verfügen oder diese kuratieren und zum Faktencheck nutzen oder anbieten möchten. Ähnlich wie der Toxicity-Detector ist die Boilerplate Open Source (unter einer MIT-Lizenz) und über viele Konfigurationsmöglichkeiten an Anforderungen und Präferenzen anpassbar. Insbesondere lassen sich EvidenceSeeker mit unterschiedlichen Modellen betreiben. Der Prototyp ist in der jetzigen Form zwar noch nicht ohne Weiteres für einen skalierbaren und verlässlichen Betrieb geeignet, kann jedoch sinnvoll als Ausgangsbasis für Optimierungen und Weiterentwicklungen dienen. Die im Projekt entwickelten Bausteine umfassen:

1. den Quellcode (https://github.com/debatelab/evidence-seeker),
2. eine ausführliche Dokumentation (https://debatelab.github.io/evidence-seeker/),
3. eine Webseite mit Beispielergebnissen (https://debatelab.github.io/evidence-seeker-results/),
4. eine Gradio DemoApp (https://huggingface.co/spaces/DebateLabKIT/evidence-seeker-demo), die lokal aufgesetzt und getestet werden kann und
5. das EvidenceSeeker Portal (https://evidence-seeker.philosophie.kit.edu/), über das Personen ohne technische Kenntnisse EvidenceSeeker aufsetzen und testen können.

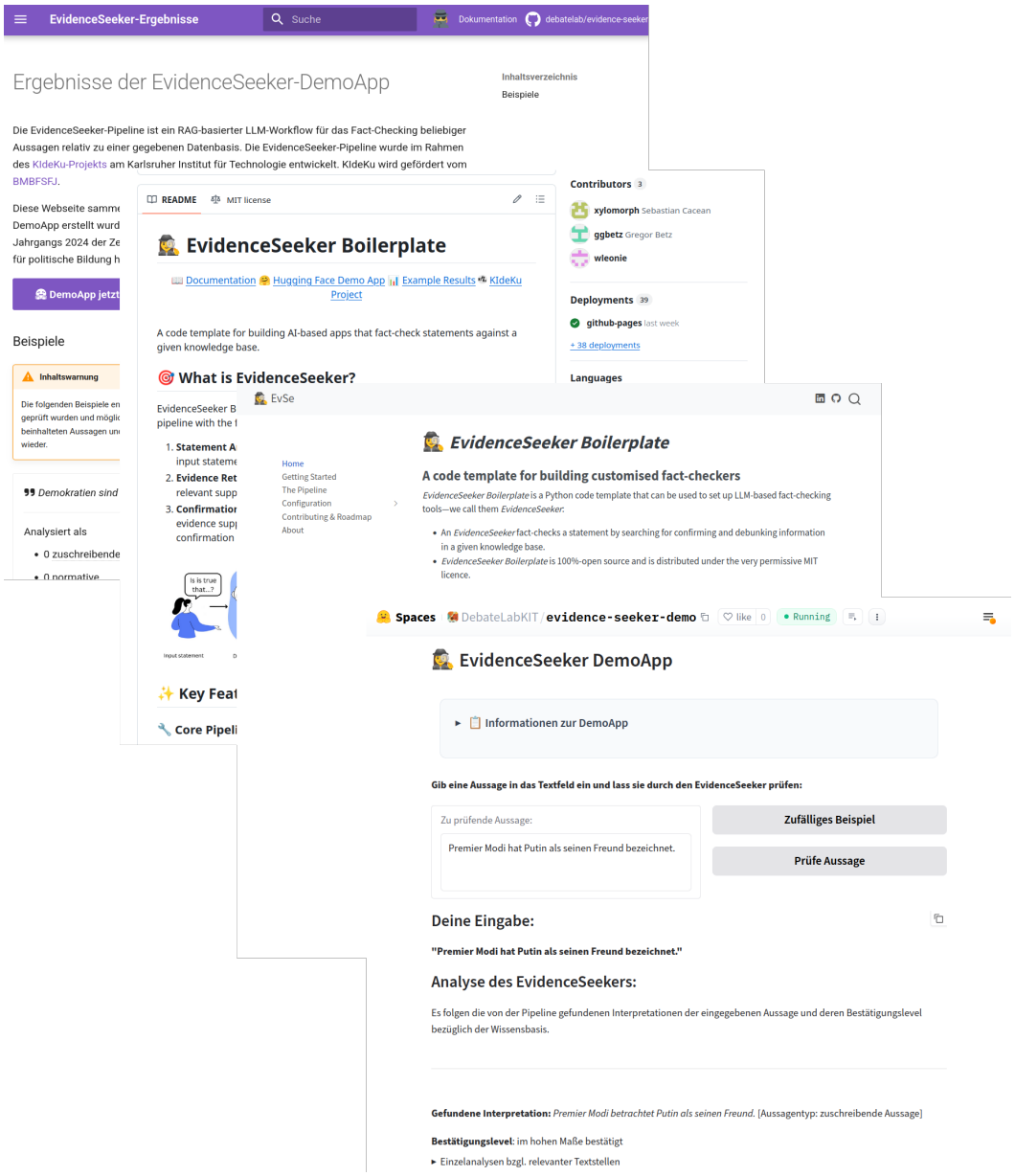


Abbildung 3.5: Projektoutputs zur EvidenceSeeker-Boilerplate

### 3.2.1 Die EvidenceSeeker-Pipeline

Die EvidenceSeeker-Boilerplate basiert auf einer Pipeline, in der eine Eingangsaussage in drei aufeinander aufbauenden Schritten geprüft wird.

Im ersten Schritt der **Disambiguierung** identifiziert die Pipeline Mehrdeutigkeiten und Vagheit, löst diese durch mögliche Interpretationen auf und unterscheidet dabei zwischen deskriptiven, zuschreibenden und normativen Aussagen. Diese Disambiguierung erfolgt durch Anfragen an ein Sprachmodell, das als Ergebnis eine Menge von Aussagen

zurückgibt. Jede dieser Aussagen entspricht einer gefundenen Interpretation mit einem eindeutigen Aussagentyp.

Im zweiten Schritt der **Extraktion** werden aus der zugrunde liegenden Wissensbasis für jede Interpretation thematisch relevante Textstellen extrahiert, die somit potentielle Evidenzen darstellen.

Im dritten Schritt der **Bestätigungsanalyse** wird dann für jede dieser Textstellen durch erneute Anfragen an ein Sprachmodell geprüft, in welchem Maße die Textstelle die Interpretation stützt oder widerlegt. Die so gewonnenen Ergebnisse werden dann für jede Interpretation zu einer Gesamtaussage aggregiert, die den User:innen mitteilt, inwiefern die Interpretation durch die Wissensbasis gestützt oder widerlegt wird.

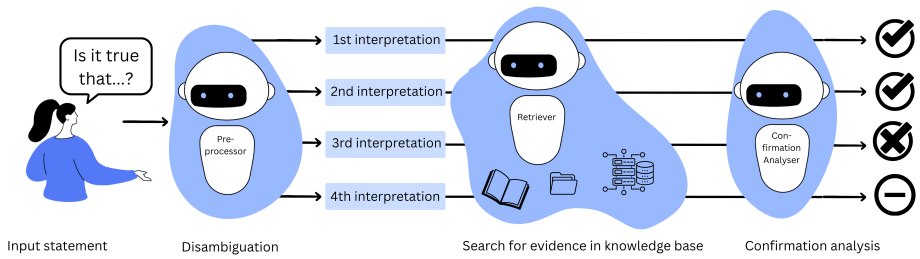


Abbildung 3.6: Pipeline der EvidenceSeeker-Boilerplate

Die einzelnen Ergebnisse für jede Interpretation werden allerdings nicht zu einer Gesamteinschätzung über die Eingangsaussage aggregiert. Als Ergebnis liefert die Pipeline also eine differenzierte Analyse. Das mag Nutzer:innen womöglich unbefriedigend erscheinen, ist jedoch bewusst intendiert: Enthält die Eingangsaussage Vagheit oder Mehrdeutigkeit, kommt es für die Wahrheitsbeurteilung der Aussage darauf an, wie man sie interpretiert. In einem solchen Fall lässt sich nichts Weiteres über den Wahrheitsgehalt der Eingangsaussage ableiten. Enthält die Eingangsaussage hingegen keine relevante Vagheit und keine relevanten Mehrdeutigkeiten, sollte die Pipeline auch nur eine Interpretation finden, die bedeutungsgleich mit der Eingangsaussage ist.

Schauen wir uns die einzelnen Schritte etwas genauer an.[8]

#### 3.2.1.1 Disambiguierung

Ziel der Voranalyse ist die Disambiguierung der zu prüfenden Aussage. Dabei sollen relevante Mehrdeutigkeiten aufgelöst und zwischen deskriptiven, normativen und zuschreibenden Aussagen unterschieden werden. Warum ist das wichtig?

Die Auflösung von Vagheit und Mehrdeutigkeit ist in vielen Fällen für einen Faktencheck notwendig, wie das folgende Beispiel illustriert: Die Aussage *„Julian ist groß“* ist ohne weiteres Kontextwissen keiner sinnvollen Wahrheitsprüfung zugänglich. Wir

[8]Weitere Details zur Pipeline unter https://debatelab.github.io/evidence-seeker/workflow.html.

müssen zunächst verstehen, was genau der:die Sprecher:in mit der Aussage meint und gegebenenfalls mehr über die betreffende Person erfahren. Was ist der implizite Vergleichskontext? Meint der:die Sprecher:in groß für ein Kind in diesem Alter, in der subjektiven Wahrnehmung oder im Vergleich zum letzten Mal, als der:die Sprecher:in die Person gesehen wurde? Usw.

Auch die Unterscheidung zwischen deskriptiven, zuschreibenden und normativen Aussagen ist für die Wahrheitsprüfung relevant.

**Deskriptive Aussagen** sind beschreibende Aussagen, die in Abhängigkeit von dem, was in der Welt der Fall ist, wahr oder falsch sind. In der Regel sind diese Aussagen in einem gewissen Maße durch empirische Beobachtungen überprüfbar. Für deskriptive Aussagen geht es daher im Faktencheck um das Auffinden relevanter empirischer Evidenzen. Vom EvidenceSeeker erwarten wir, dass er in einer Wissensbasis Aussagen findet, die solche Evidenzen beschreiben, und den entsprechenden Bestätigungs- beziehungsweise Widerlegungsgrad ermittelt.

**Zuschreibende Aussagen** sind eine besondere Klasse deskriptiver Aussagen: Sie schreiben Personen oder Gruppen Aussagen zu. Die Personen oder Gruppen können dabei unbestimmt bleiben. Die Aussage *„Manche Menschen denken, dass es keinen menschengemachten Klimawandel gibt“* sagt, dass es Menschen gibt, denen die Überzeugung zugeschrieben werden kann, dass es keinen menschengemachten Klimawandel gibt. Insofern zuschreibende Aussagen deskriptive Aussagen sind, gilt das oben bereits Gesagte für sie gleichermaßen. Allerdings muss zwischen der Zuschreibung und der zugeschriebenen Aussage unterschieden werden, da die jeweiligen Wahrheitsprüfungen unabhängig vorgenommen werden müssen und unterschiedlich ausfallen können. So ist die Beispielaussage zwar wahr, aber die zugeschriebene Aussage falsch. Darüber hinaus wird man unter Umständen unterschiedliche Wissenbasen für die Überprüfung verwenden müssen. Während für die Beispielaussage die Ergebnisse entsprechender Umfragen berücksichtigt werden können, benötigt man für die Prüfung der zugeschriebenen Aussage Erkenntnisse aus den Klimawissenschaften.

Von einem EvidenceSeeker erwarten wir, dass er zuschreibende Aussagen als solche erkennt und zwischen Zuschreibung und zugeschriebener Aussage unterscheidet.

Der dritte Typ relevanter Aussagen umfasst **normative Aussagen**, zum Beispiel Wertaussagen, Gebote, Verbote und Empfehlungen. Bei den normativen Aussagen lässt sich fragen, ob sie überhaupt einem Faktencheck unterzogen werden können. Normative Aussagen werden häufig so verstanden, dass sie nicht in derselben Weise richtig oder falsch sein können wie deskriptive Aussagen und dass sie sich auch nicht allein durch empirische Beobachtungen überprüfen lassen. Unabhängig davon lässt sich natürlich prüfen, in welchem Begründungsverhältnis solche Aussagen zu anderen normativen Aussagen stehen. Auch wenn die bisherigen Formulierungen der Evidenzsuche nicht ganz passen, wäre das Vorgehen der „Prüfung“ solcher Aussagen anhand einer „Wissensbasis“ nicht anders als bei deskriptiven Aussagen. So ließe sich beispielsweise analysieren, ob eine gegebene normative Aussage Regelungen des Grundgesetzes widerspricht oder

durch sie gestützt wird, indem als Wissensbasis das Grundgesetz selbst und die Urteile des Bundesverfassungsgesetzes herangezogen werden.

In der Konfiguration der EvidenceSeeker-Boilerplate lässt sich festlegen, ob normative Aussagen geprüft werden sollen. In der Voreinstellung werden normative Aussagen als solche erkannt, jedoch im weiteren Verlauf der Pipeline nicht weiter geprüft. So werden gefundene normative Interpretationen der Eingangsaussage zwar als Ergebnis zurückgegeben, allerdings ohne eine daran geknüpfte Bestätigungsanalyse.

Insgesamt ist die Unterscheidung zwischen deskriptiven, zuschreibenden und normativen Aussagen für Faktenchecks relevant, da für deren Wahrheitsprüfung zum Teil unterschiedliche Kriterien einschlägig sind. Es ist daher notwendig, diese Aussagen vor der eigentlichen Prüfung zu unterscheiden und entsprechende Mehrdeutigkeiten aufzulösen.

#### 3.2.1.2 Extraktion

Im Extraktionsschritt wird für jede gefundene Interpretation nach thematisch relevanten Textstellen in der Wissensbasis gesucht. Technisch wird dazu *Retrieval Augmented Generation* (RAG) herangezogen. RAG ist eine sehr verbreitete Methode, um Anfragen an Sprachmodelle mit Kontextinformationen aus einer externen Wissensbasis anzureichern. Damit soll üblicherweise die Zuverlässigkeit der Antworten von Sprachmodellen verbessert werden, um sogenannte „Halluzinationen“ zu vermeiden, also die Generierung von Antworten, die zwar plausibel klingen, aber faktisch falsch sind.

RAG wird typischerweise mithilfe von Embeddingmodellen realisiert. Mit diesen Modellen werden Textstellen als hochdimensionale reelwertige Vektoren repräsentiert, die ihre semantische Bedeutung repräsentieren. Liegen die Repräsentationen zweier Textstellen im Vektorraum nah beieinander, sind sie auch thematisch ähnlich – so die Idee.

Im Rahmen der EvidenceSeeker-Boilerplate wird die gesamte Wissensbasis in Textstellen zerlegt, die anschließend mithilfe eines Embedding-Modells in Vektoren umgewandelt werden. Der so erzeugte Index wird dann bei jeder Suche nach relevanten Textstellen für eine Interpretation durchsucht und die acht thematisch ähnlichsten Textstellen werden zurückgegeben.[9]

Wie gut eine solche Extraktion funktioniert, hängt von verschiedenen Faktoren ab, beispielsweise von der Art der Zerlegung und der Qualität des Embeddingmodells.

#### 3.2.1.3 Bestätigungsanalyse

Von den so zurückgegebenen Textstellen ist aber noch nicht klar, ob sie die Interpretation stützen oder widerlegen, oder ob sie sie weder stützen noch widerlegen. Darum wird in der Bestätigungsanalyse ein Sprachmodell befragt, den Grad der Bestätigung bzw.

[9] Diese Zahl ist ein konfigurierbarer Parameter ($top_k$).

Widerlegung anhand der Textstellen zu beurteilen. Dies wird für jede Textstelle separat gemacht. Genauer gesagt, wird das Modell über eine Multiple-Choice-Frage gebeten, zu entscheiden, ob die Textstelle hinreichende Evidenz zur Unterstützung der Interpretation liefert, ob sie Evidenz liefert, die der Interpretation widerspricht, oder ob sie die Interpretation weder unterstützt noch ihr widerspricht.

Für jede Interpretation und jede relevante Textstelle wird so ein Bestätigungsgrad berechnet, der Werte zwischen $-1$ und 1 annehmen kann, wobei $-1$ maximale Widerlegung durch die Textstelle, 1 maximale Bestätigung und 0 keine Bestätigung bedeutet.[10] Insgesamt werden damit für jede Interpretation mehrere Bestätigungsgrade berechnet, nämlich so viele wie relevante Textstellen, die anschließend zu einem aggregierten Bestätigungsgrad für jede Interpretation zusammengefasst werden.[11] Dieser aggregierte Bestätigungsgrad $DOC(I)$ für eine Interpretation $I$ wird dann in eine verbale Einschätzung überführt:

- $0.6 < DOC(I) \leq 1$: stark bestätigt
- $0.4 < DOC(I) \leq 0.6$: bestätigt
- $0.2 < DOC(I) \leq 0.4$: schwach bestätigt
- $-0.2 \leq DOC(I) \leq 0.2$: weder bestätigt noch widerlegt
- $-0.4 \leq DOC(I) < -0.2$: schwach widerlegt
- $-0.6 \leq DOC(I) < -0.4$: widerlegt
- $-1 \leq DOC(I) < -0.6$: stark widerlegt

## 3.3 syncIALO Datensatz

syncIALO ist eine Sammlung synthetischer Argumentkartendatensätze, die mithilfe von Sprachmodellen erstellt wurden.[12] Der primäre Korpus enthält über 600.000 Argumente und über 1000 Argumentkarten, die argumentative Zusammenhänge zwischen den Argumenten darstellen.

Diese Argumentkarten sind gerichtete Graphen, in denen die Knoten Argumente repräsentieren und die Relationen angeben, ob ein Argument ein anderes unterstützt oder angreift. Abbildung 3.7 zeigt ein Beispiel einer Argumentkarte aus dem syncIALO-Datensatz, die mit Hilfe von Argdown gerendert wurde.

### 3.3.1 Erstellung von syncIALO

Die Erstellung von syncIALO basiert auf einer dynamischen Pipeline, die einen umfassenden Argumentkartierungsprozess nachahmt. Ein LLM-basierter Agent simuliert

[10] Technisch werden diese Bestätigungsgrade aus den Tokenwahrscheinlichkeiten der Antwortoptionen aus der Multiple-Choice-Frage erzeugt.

[11] Im Wesentlichen ein gewichteter Mittelwert.

[12] Der entsprechende Code ist unter https://github.com/debatelab/syncIALO einsehbar. Die Datensätze findet man unter https://huggingface.co/datasets/DebateLabKIT/syncialo-raw.

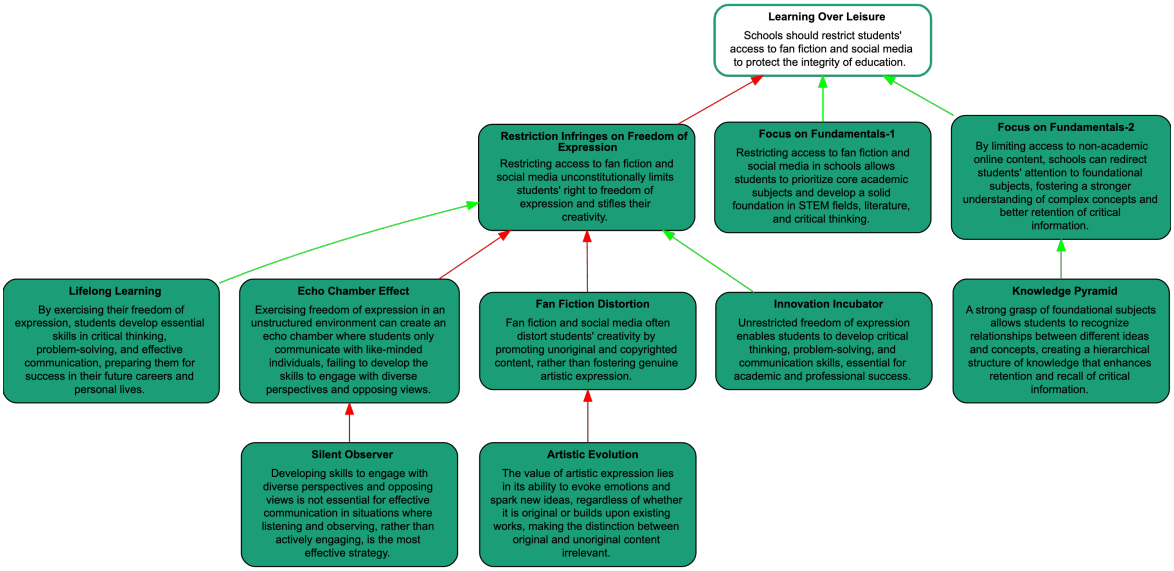


Abbildung 3.7: Beispiel einer Argumentkarte aus dem syncIALO-Datensatz

einen kritischen Denker, der nach neuen Argumenten sucht, sie bewertet und unter bestimmten Bedingungen der Argumentkarte hinzufügt.

Für eine in Frage stehende These wird die Argumentkarte rekursiv aufgebaut, indem zunächst für die These Vor- und Nachteile als Argumente und Einwände hinzugefügt werden und dieser Prozess anschließend für die bisher formulierten Argumente und Einwände fortgeführt wird, bis eine maximale Tiefe erreicht ist. Dazu identifiziert der KI-Agent die Prämissen eines Zielarguments, bevor er weitere Argumente konzipiert, die die Prämissen entweder unterstützen oder angreifen. Er wählt Kandidatenargumente hinsichtlich ihrer Relevanz und Vielfalt aus und überprüft, ob sie bereits in der Argumentkarte enthalten sind, bevor er sie der Argumentkarte hinzufügt.

Um die Themenvielfalt zu erhöhen, werden die Themen anhand einer vielfältigen Tag-Cloud gewählt. Außerdem nimmt der kritische KI-Agent bei der Generierung eines neuen Kandidatenarguments eine zufällig ausgewählte Persona an.

Der LLM-basierte Agent wird je nach Pipeline-Schritt von verschiedenen offenen Modellen angetrieben. Das Modell `meta-llama/Llama-3.1-405B` wird für die Generierung und Bewertung von Argumenten verwendet und ein fein abgestimmtes Llama-3.1-8B-Modell für weniger anspruchsvolle Aufgaben wie die Formatierung. `MoritzLaurer/deberta-v3-large-zeroshot-v2.0` dient als Mehrzweck-Klassifikator und `sentence-transformers/all-MiniLM-L6-v2`, um Satz-Embeddings zu generieren.

### 3.3.2 Verwendungsmöglichkeiten

syncIALO ist ein vielseitiger Datensatz, der für verschiedene Zwecke genutzt werden kann. Er eignet sich insbesondere für die Erstellung von Trainings- und Evaluationsdatensätzen für KI-gestützte Tools im Bereich der Argumentationsanalyse. Dazu muss aus dem rohen syncIALO-Datensatz ein spezifischer, auf die jeweilige Anwendung zugeschnittener Datensatz erstellt werden.

So ließen sich beispielsweise die Argumente einer Argumentkarte als Dialoge verbalisieren und diese Dialoge anschließend als Trainingsdaten für ein KI-Modell verwenden, das Argumente verstehen soll. Alternativ ließen sich in die Argumentkarten bestimmte Fehler einbauen und das Modell könnte dann darauf trainiert werden, diese zu erkennen und zu korrigieren. Der `deep-argmap-conversations` Datensatz ist ein Beispiel für einen spezifischeren Datensatz, der aus dem rohen syncIALO-Datensatz erstellt wurde und unterschiedliche Argumentkartierungsaufgaben enthält.

syncIALO kann nicht nur als Grundlage für die Erstellung von Trainings- und Evaluationsdatensätzen dienen, sondern auch direkt zur Entwicklung von KI-gestützten Tools im Bereich der Argumentation verwendet werden. So könnte der Datensatz verwendet werden, um Modellen Beispiele von Argumentkarten im Prompt zu übergeben, um sie zu befähigen, Argumentkarten zu verstehen oder zu erstellen (sogenanntes *few-shot prompting*).

# 4 Herausforderungen und Empfehlungen

Die Integration von LLM-basierten Tools in bestehende Prozesse stellt Beteiligte und Betroffene vor große Herausforderungen. Die Technologie ist nicht nur vergleichsweise neu und komplex, sondern entwickelt sich rasant weiter. Es ist noch nicht absehbar, welche gesellschaftlichen und wirtschaftlichen Auswirkungen Sprachmodelle und KI im Allgemeinen haben werden. Es ist daher verständlich, dass viele Menschen und Organisationen mit Unsicherheit und Skepsis gegenüber KI reagieren. Das gilt auch für die Verwendung von KI durch staatliche und zivilgesellschaftliche Akteur:innen im Demokratiebereich. In diesem abschließenden Kapitel wollen wir einige Überlegungen und Empfehlungen formulieren, die wir für wichtig halten, um die Chancen von KI zu nutzen und die Risiken zu minimieren. Dabei geht es nicht nur um technische Fragen, sondern auch um ethische, rechtliche und gesellschaftliche Aspekte. Diese Überlegungen basieren auf informellen Gesprächen mit Akteur:innen aus dem zivilgesellschaftlichen Bereich, auf der Analyse von Studien und Berichten sowie auf unseren Erfahrungen aus dem KIdeKu-Projekt. Sie sind keineswegs abschließend, sondern sollen vielmehr Anregungen und Impulse geben, um die Diskussion über KI im Demokratiebereich weiterzuführen.

## 4.1 Rolle von Vertrauen

Der Einsatz von KI-basierten Applikationen kann nur dann unsere demokratische Kultur stärken, wenn diese von Bürger:innen genutzt werden, was wiederum ein hinreichend großes Vertrauen in die Sicherheit und Zuverlässigkeit solcher Anwendungen voraussetzt. Im Folgenden werden einige allgemeine Überlegungen zur Rolle von Transparenz und Zuverlässigkeit für das Vertrauen dargestellt.[1]

Warum sind Vertrauen und Akzeptanz im Kontext der KI-gestützten Deliberation so wichtig? Deliberative Kontexte zeichnen sich durch ihre besondere Rolle für die demokratische Teilhabe aus. In ihnen üben Bürger:innen ihre politischen Rechte aus und nehmen mittelbar oder unmittelbar an der politischen Willensbildung teil, indem sie beispielsweise ihre Überzeugungen im öffentlichen Diskurs ausdrücken oder politische Selbstwirksamkeit in Beteiligungsformaten wie Bürgerräten erfahren. Daher ist zu

[1] Ähnliche und komplementäre Überlegungen finden sich im Blogbeitrag von Marvin Sieger, der aus dem Projekt „Wegweiser.UX-für-KI“ berichtet.

erwarten, dass Bürger:innen auf (wahrgenommene) Einschränkungen dieser Teilhaberechte sehr sensibel reagieren. Bei KI-gestützter Deliberation geht es also nicht nur darum, dass Nutzer:innen mit einem bestimmten KI-Tool unzufrieden sind, wenn etwas nicht richtig funktioniert, und sie als Konsequenz zum Produkt eines Konkurrenten wechseln. Viel gewichtiger ist die Gefahr, dass sie sich in der Ausübung ihrer bürgerlichen Teilhaberechte eingeschränkt fühlen. Sie könnten sich durch den Einsatz von KI ausgeschlossen, missverstanden oder anderweitig ungerecht behandelt fühlen. Im schlimmsten Fall schwächen solche Erfahrungen nicht nur das Vertrauen in KI-basierte Deliberation selbst, sondern auch das Vertrauen in Demokratie.

Eine besondere Gefahr besteht darin, dass KI-gestützte Deliberation neue Ungleichheiten in der politischen Teilhabe erzeugen kann. Wenn technische Lösungen vor allem denjenigen zugutekommen, die über digitale Kompetenzen, Vertrauen in KI oder entsprechende Ressourcen verfügen, kann politische Teilhabe insgesamt zwar wachsen, aber auch ungleicher verteilt werden.[2] Wir müssen deshalb darauf achten, dass die Einführung deliberativer KI nicht zu einem neuen „deliberative divide“ führt.[3]

Eine weitere Herausforderung könnte man als „Skalierungsversuchung“ bezeichnen.[4] Gerade Beteiligungsformate könnten durch den Einsatz von KI mit deutlich größeren Teilnehmer:innenzahlen durchgeführt werden, als es bislang möglich war.[5] Funktionierende Beteiligungsverfahren sind in der Regel sehr aufwendig: Bürger:innen müssen über teilweise komplexe Fachthemen informiert werden; ihnen wird oft die Möglichkeit gegeben, Argumente und Einwände auszutauschen; sie drücken ihre Standpunkte aus, die so aggregiert werden müssen, dass der Output der Beteiligung als Entscheidungsgrundlage für Politiker:innen dienen kann. Der Aufwand für die Organisation und Moderation solcher Formate sowie für die Analyse der Beiträge nimmt dementsprechend mit steigender Teilnehmer:innenzahl sehr schnell zu. KI, so die Hoffnung, kann viele dieser Aufgaben unterstützen oder gar übernehmen.[6] Die Skalierbarkeit eines Verfahrens ist jedoch keine Garantie für dessen demokratische Legitimität. Vielmehr gilt es zu bedenken, dass alle normativen Anforderungen an solche Prozesse weiterhin erfüllt sein müssen. Die Ergebnisse von Beteiligungsverfahren sollen in die politische Willensbildung einfließen, die nur dann legitimiert sind, wenn das Verfahren selbst hinreichend legitimiert ist. Dabei geht es zum Beispiel um gleiche Teilhabe, die Authentizität von Beiträgen, die Qualität des Outputs und die Neutralität der Moderation und Aggregation von Beiträgen. Ob die Erfüllung solcher normativen Anforderungen genauso gut skaliert wie die technischen Möglichkeiten, ist eine offene Frage.

[2] Diese Gefahr ist auch deswegen relevant, weil in der Forschung zu deliberativer KI die Anforderung gleicher Teilhabe weniger stark untersucht wird als andere Normen (Friess u. a. 2025).

[3] Dieser Ausdruck stammt von Jungherr und Rauchfleisch (2025), die in einer repräsentativen Umfrage zeigen, dass eine allgemeine Skepsis gegenüber KI mit einer skeptischen Einstellung zu den Fähigkeiten deliberativer KI korreliert.

[4] Die folgenden Punkte wurden auf dem KIdeKu-Workshop von Julian Müller und Eike Düvel vorgebracht.

[5] Im Projekt „Künstliche Intelligenz und Bürgerräte“ (KIB) werden Einsatzmöglichkeiten von KI in der Öffentlichkeitsbeteiligung untersucht.

[6] Vgl. bspw. Tessler u. a. (2024).

Um welches Vertrauen geht es hier? Neben dem bereits beschriebenen Vertrauen in die Einhaltung politischer Rechte sind vor allem Datensicherheit sowie das Vertrauen in die Korrektheit und Unvoreingenommenheit von KI-generierten Inhalten und Analysen zentral.

Diese Aspekte hängen je nach spezifischem Einsatzszenario voneinander ab. Nehmen wir als Beispiel die Detektion toxischer Sprache: Bei der Kategorisierung toxischer Sprache kann es zu unterschiedlichen Arten von Fehlern kommen.

Eine falsch positive Detektion ist die Kategorisierung von Äußerungen als toxisch, die es gar nicht sind. Dient die Detektion toxischer Sprache beispielsweise der Moderation von Online-Diskussionen, könnte eine falsch positive Detektion dazu führen, dass bestimmte Beiträge gelöscht oder Nutzer:innen gesperrt werden, obwohl diese Beiträge eigentlich nicht toxisch sind. Dies kann zu einer wahrgenommenen Einschränkung der Meinungsfreiheit und zu einer Ausgrenzung bestimmter Gruppen führen, wenn die Fehldetektionen systematisch stärker bei diesen auftreten. So zeigen zum Beispiel Giraud u. a. (2025), dass manche Modelle höhere Raten falsch positiver Toxizitätsdetektion bei Mitgliedern der afroamerikanisch-englischen Sprachgemeinschaft aufweisen. Als mögliche Ursache nennen sie einen Bias in den Trainingsdaten aufgrund fehlender Diversität unter den Annotator:innen.

Eine falsch negative Detektion ist die Nichterkennung von Äußerungen, die tatsächlich toxisch sind. In diesem Fall könnte es vorkommen, dass toxische Beiträge nicht moderiert werden und dadurch weiterhin Schaden anrichten. Im Extremfall kommt es zur Sichtbarkeit und Verbreitung von Inhalten, die Straftatbestände erfüllen. Aber auch wenn es nicht um strafrechtlich relevante Inhalte geht, können unentdeckte toxische Beiträge dazu führen, dass die Erreichung deliberativer Ideale wie Respekt, Inklusion und rationale Diskussion erschwert wird, weil sich Menschen durch solche Beiträge angegriffen, verletzt oder ausgeschlossen fühlen und infolge dessen weniger bereit zur Teilnahme an deliberativen Prozessen sind.

Eine hinreichend hohe Ergebniskorrektheit von KI-Systemen spielt damit eine zentrale Rolle für das Vertrauen in KI-gestützte Deliberation. Darüber hinaus ist diese Art von Zuverlässigkeit eine Voraussetzung für die Erreichung der genannten deliberativen Ziele (vgl. Kapitel 2.2). Im Folgenden konzentrieren wir uns daher auf die damit zusammenhängenden Herausforderungen und besprechen Lösungsansätze zur Steigerung der Zuverlässigkeit von KI-Tools.

## 4.2 Evaluierung und Optimierung von KI-Tools

Wie stellen wir nun sicher, dass KI-Tools, die in deliberativen Kontexten eingesetzt werden, hinreichend zuverlässig bezüglich der Korrektheit ihrer Ergebnisse sind? Eine wichtige Rolle spielt die *systematische* Evaluierung. Ohne zu wissen, wie zuverlässig ein KI-Tool in einem bestimmten Kontext abschneidet, können wir nicht beurteilen, ob es für diesen Kontext hinreichend zuverlässig ist. Die Ergebnisse systematischer

Evaluierungen bilden außerdem die Grundlage für eine Optimierung der Zuverlässigkeit von KI-Tools.

Die Leistungsfähigkeit eines LLM-basierten KI-Tools hängt von vielen Faktoren ab. Dazu zählen das verwendete Sprachmodell, die Modellparameter und die gesamte Pipeline, inklusive der verwendeten Prompts, in die das Modell eingebunden ist (im Engl. *scaffolding* oder *harness*). Damit gibt es auch eine Vielzahl von Faktoren, die variiert werden können, um die Zuverlässigkeit eines KI-Tools zu steigern.

Die Grundidee einer evaluationsbasierten Optimierung ist dabei sehr einfach: Schneidet die Zuverlässigkeit eines Systems nicht hinreichend zufriedenstellend ab, nimmt man so lange leistungsverbessernde Anpassungen am System vor, bis es hinreichend zuverlässig ist. Ist es möglich, einzelne Komponenten des Systems isoliert zu evaluieren, können diese Anpassungen unter Umständen sehr gezielt vorgenommen werden. Führt diese Optimierungsschleife zu keinem Erfolg oder erfordert die Optimierung zu große Abstriche bezüglich anderer relevanter Faktoren, ist es unter Umständen sinnvoll, auf den Einsatz des KI-Tools zu verzichten.

Die notwendigen Evaluierungen müssen im folgenden Sinne systematisch erfolgen: Die Ergebnisse einer Evaluation sollen belastbare Schlüsse darüber erlauben, ob ein KI-Tool für einen bestimmten Gegenstandsbereich hinreichend zuverlässig ist oder nicht. In der Regel ist es praktisch nicht möglich, den gesamten Gegenstandsbereich abzutesten. Um Schlüsse auf den gesamten Gegenstandsbereich zuzulassen, müssen die Testfälle damit so ausgewählt werden, dass sie in einer bestimmten Weise repräsentativ für den Gegenstandsbereich sind. Das könnten zum Beispiel eine hinreichend große Menge an Testfällen sein, die die Heterogenität des Gegenstandsbereichs abdecken, oder besonders schwierige Testfälle, die gezielt ausgewählt wurden, um bestimmte relevante Aspekte der Zuverlässigkeit zu testen.

Darüber hinaus muss die Evaluierung so gestaltet werden, dass sie automatisiert und einfach reproduziert werden kann, damit die oben genannten Optimierungsschleifen praktisch umsetzbar sind. Dafür können beispielsweise vorhandene Testdatensätze verwendet oder eigene durch Menschen oder automatisiert erstellt werden.

### 4.2.1 Herausforderungen bei der Optimierung von KI-Tools

Obwohl die Grundidee einer evaluationsbasierten Optimierung von KI-Tools sehr einfach ist, gibt es eine Reihe von praktischen und prinzipiellen Herausforderungen, die bei der Umsetzung zu beachten sind.

Zum einen ist die Evaluierung komplexer KI-Pipelines selbst komplex. Zwar lässt sich die Zuverlässigkeit eines KI-Tools als Ganzes evaluieren, um zu beurteilen, ob es für einen bestimmten Gegenstandsbereich hinreichend zuverlässig ist. Will man das KI-Tool jedoch optimieren, ist es hilfreich zu wissen, wo genau in der Pipeline etwas nicht korrekt funktioniert. Bei komplexen Pipelines müssen die einzelnen Komponenten damit unabhängig evaluiert werden, um gezielt Anpassungen vornehmen zu können.

Nehmen wir als Beispiel die vorgestellte Pipeline des EvidenceSeekers (siehe Kapitel 3.2). Die Pipeline besteht aus drei Komponenten, die unabhängig voneinander evaluiert werden können: die Disambiguierung, die Extraktion relevanter Textstellen und die Bestätigungsanalyse. Jede Komponente erfüllt eine andere Funktion und kann damit spezifische Fehlerquellen aufweisen.

Die Verfügbarkeit und Güte geeigneter Testdatensätze beziehungsweise Benchmarks sind eine weitere Herausforderung für die Evaluierung von KI-Tools. Testdatensätze müssen zum einen den anvisierten Gegenstandsbereich adäquat widerspiegeln, um belastbare Schlüsse über die Zuverlässigkeit eines KI-Tools zuzulassen. Zum anderen müssen sie selbst korrekte Labels enthalten.

Um den Gegenstandsbereich adäquat abzubilden, müssen Testdatensätze so gestaltet sein, dass sie die Heterogenität des Gegenstandsbereiches abdecken. Das bedeutet zum Beispiel, dass sie eine hinreichend große Anzahl an Testfällen enthalten müssen, die die sprachliche und kulturelle Diversität abbilden. Damit dürfen die Testdatensätze auch nicht zu „sauber“ und artifiziell sein. Außerdem sollten sie einen ausgewogenen Anteil an allen relevanten Labels enthalten, damit die Evaluierung nicht durch eine unausgewogene Verteilung der Testfälle verzerrt wird.

Testdatensätze können von Menschen oder automatisiert erstellt werden. Die menschliche Annotation von Testdatensätzen ist oft sehr aufwendig, weshalb die Verwendung automatisiert erstellter Testdatensätze eine attraktive Alternative darstellt. Je nach Anwendungsfall gibt es verschiedene Möglichkeiten, solche synthetischen Testdatensätze zu erstellen, die teils von hohem Ideenreichtum zeugen. So kann man auf Grundlage eines vorhandenen kleineren Testdatensatzes neue Testfälle generieren, indem man Bestandteile in Frage-Antwort-Paaren substituiert, von denen man weiß, dass sie für die Korrektheit der Labels nicht relevant sind. In manchen Fällen ist es auch möglich, Testdatensätze von starken Sprachmodellen generieren zu lassen (sogenanntes *LLM-as-a-Generator*). Das setzt allerdings voraus, dass die Zuverlässigkeit der Sprachmodelle, die für die Erstellung von Testdatensätzen verwendet werden, bereits hinreichend evaluiert und optimiert wurde.[7]

Synthetische Testdatensätze sind jedoch nicht für alle Anwendungsfälle geeignet. In bestimmten Fällen bedarf es nach wie vor Menschen für die Erstellung von Testdatensätzen. Das ist allerdings mit eigenen Herausforderungen verbunden. Neben dem hohen Ressourcenaufwand muss für so erstellte Testdatensätze in gleicher Weise sichergestellt werden, dass sie korrekte Labels enthalten. Dafür werden in der Regel Methoden aus der Inhaltsanalyse verwendet.[8] So werden typischerweise mehrere Annotator:innen eingesetzt, um die Testdatensätze zu annotieren, und die Inter-Annotator:innen-Übereinstimmung wird berechnet, um die Güte der Annotationen zu bewerten. Das liefert zumindest indirekte Hinweise darauf, ob die Testdatensätze korrekte Labels enthalten.

[7] Der syncAILO-Datensatz ist ein Beispiel für einen synthetischen Datensatz, der mit einer LLM-Pipeline erstellt wurde und auf dessen Grundlage unterschiedliche Test- und Trainingsdatensätze erstellt werden können. Vergleiche Kapitel 3.3.

[8] Wie bspw. in Krippendorff (2019) dargestellt.

Ein weiteres praktisches Problem ist die sogenannte *Training-Test Contamination*. Wenn KI-Modelle bereits während des Trainings Zugang zu den Testdatensätzen haben, sind ihre Testergebnisse auf diesen Datensätzen kein Indikator für ihre Zuverlässigkeit, sondern eher eine Überprüfung, ob sie die Testdatensätze bereits „gelernt“ haben. Das kann zum Beispiel passieren, wenn die Testdatensätze öffentlich zugänglich sind und damit in den Trainingsdaten enthalten sein könnten. Das kann unter Umständen zu einer systematischen Überschätzung der Zuverlässigkeit von KI-Tools führen.

#### 4.2.1.1 Mehrdeutigkeit und Kontextabhängigkeit

Die bisher dargestellten Herausforderungen werden in der Wissenschaft und Praxis der Entwicklung und Evaluierung von KI-basierten Systemen bereits breit diskutiert. Eine prinzipielle Herausforderung, die weniger Beachtung findet, soll hier ausführlicher dargestellt werden. Das bisherige Bild der Optimierung von KI-Tools durch systematische Evaluation suggeriert unter Umständen, dass es bei Fragen zur Korrektheit von KI-generierten Resultaten immer eine eindeutige Antwort gibt. In vielen Fällen, insbesondere in deliberativen Kontexten, fehlt es an dieser Eindeutigkeit.

Das Beispiel zur Detektion toxischer Sprache kann wieder als einfache Illustration dienen. Toxische Sprache ist ein soziokulturelles Phänomen, das in mehrfacher Hinsicht von Mehrdeutigkeit geprägt ist.

Zum einen ist die Detektion toxischer Sprache häufig kontextabhängig. Das heißt, ob eine Äußerung toxische Sprache darstellt, kann vom kulturellen und situativen Kontext abhängen. Dazu zählen kulturelle Normen, der verwendete Dialekt, die Intentionen der Sprecher:innen, die Frage, ob es sich um indirekte Rede, Metaphern oder Satire handelt, und ob Codes oder sogenannte Geusenwörter (im Engl. *reclaimed speech*) verwendet werden. Diese Informationen sind der Äußerung nicht immer selbst ablesbar, sodass zusätzliche Kontextinformationen verfügbar sein müssen. Das gilt für die Trainings- und Testdatensätze sowie für den Einsatz von KI zur Detektion toxischer Sprache. Nur wenige der vorhandenen Testdatensätze verfügen allerdings über solche Kontextinformationen.[9]

Eine weitere Herausforderung ist die definitorische Vielfalt beim Begriff toxischer Sprache. Toxische Sprache kann auf unterschiedliche Weise definiert werden (Fortuna u. a. 2020). Hinzu kommt, dass es weitere Begriffe gibt, die zwar Überschneidungen, aber eben auch Unterschiede zu toxischer Sprache aufweisen, wie zum Beispiel „hate speech“, „offensive speech“ oder „uncivil speech“. Auch wenn diese Begriffe und die dazugehörigen Definitionen Überschneidungen aufweisen, gibt es Unterschiede, die in vielen Einzelfällen für die Toxizitätskategorisierung von Äußerungen relevant sind. Diese definitorische Vielfalt spiegelt sich auch in den Datensätzen wider, sodass es teilweise schwierig ist, passende Datensätze für die Evaluierung von KI-Tools zu finden.

[9] Vgl. auch Kapitel 3.1.2.

Die beiden bisher genannten Probleme sind keine prinzipiellen Hindernisse für die Evaluierung und Optimierung von KI-Tools, sondern eher praktische Herausforderungen, die mit entsprechenden Ressourcen und Aufwand überwunden werden können. So muss dafür gesorgt werden, dass die notwendigen Kontextinformationen verfügbar sind, und es muss immer klar sein, um welches sprachliche Phänomen es geht, beziehungsweise welche Definitionen als Grundlage für eine Kategorisierung dienen. Das löst die Probleme der Mehrdeutigkeit und der Kontextabhängigkeit allerdings nicht vollständig: Selbst wenn alle notwendigen Kontextinformationen verfügbar sind und es eine klare Definition gibt, kann es immer noch einen Graubereich geben, in dem die Kategorisierung von Äußerungen als toxisch oder nicht toxisch interpretationsoffen ist. Das liegt unter anderem daran, dass toxische Sprache ein graduelles Phänomen ist und Äußerungen also mehr oder weniger toxisch sein können. Selbst die Festlegung einer Grenze kann die Interpretationsoffenheit nicht vollständig beseitigen.

Die Herausforderungen begrifflicher Vielfalt, Kontextabhängigkeit und nicht verschwindender Graubereiche betreffen viele Anwendungen KI-gestützter Deliberation, weil es häufig um die Generierung und Analyse natürlicher Sprache geht, die in einem soziokulturellen und situativen Kontext eingebunden und damit im Einzelfall interpretationsoffen sein kann.

### 4.2.2 Lösungsansätze

Diese Herausforderungen bei der Evaluierung und Optimierung KI-basierter Deliberation sprechen nicht grundsätzlich gegen den Einsatz von KI – zumal sie gleichermaßen berücksichtigt werden müssen, wenn Menschen diese Aufgaben übernehmen. Gerade in deliberativen Kontexten sind Aufgaben, die KI unterstützen soll, häufig normativ aufgeladen und nur selten vollständig objektivierbar. Daher müssen KI-Systeme hier besonders vorsichtig, transparent und korrigierbar eingesetzt werden. Die folgenden Empfehlungen können dazu beitragen, durch Evaluation und Transparenz Vertrauen in KI-gestützte Deliberation zu schaffen.

Durch die zentrale Rolle der Zuverlässigkeit und Korrektheit von deliberationsunterstützenden KI-Tools empfehlen sich **unabhängige und transparente Evaluationen**. Die Evaluierung von KI-Tools sollte von unabhängigen Dritten durchgeführt werden, die in keinem direkten Interessenkonflikt stehen. Die Ergebnisse solcher Evaluationen sollten transparent und vollständig veröffentlicht werden, damit sie von der Öffentlichkeit nachvollzogen und kritisch bewertet werden können. Solche unabhängigen Evaluationen und die Veröffentlichung der Ergebnisse bilden eine wichtige Grundlage für das Vertrauen von Nutzer:innen. Die Qualität der Testdatensätze hinsichtlich der oben genannten Kriterien muss ebenfalls unabhängig überprüfbar sein. Eine Möglichkeit besteht darin, die Datensätze frei zugänglich zu machen, damit ihre Qualität unabhängig bewertet werden kann, was jedoch auch mit einer erhöhten Gefahr der genannten *Training-Test-Contamination* einhergeht.

Ein weiteres Mittel zur Steigerung des Vertrauens in KI-Tools ist **technische Transparenz**. Das bedeutet, dass die technischen Details der KI-Pipelines offengelegt werden sollten, damit unabhängige Dritte nachvollziehen können, wie Ergebnisse entstehen und ob grundlegende Anforderungen wie z.B. Datensicherheit erfüllt sind. Technische Transparenz umfasst zum Beispiel die Offenlegung der verwendeten Modelle, des Pipelinedesigns, der Prompts sowie der IT-Infrastruktur. Ein hohes Maß an Offenheit erfordert auch die Veröffentlichung des Programmcodes unter Open-Source-Lizenzen und die Nutzung von Open-Weight-Modellen, die im Gegensatz zu proprietären Modellen wie ChatGPT frei verfügbar und selbst betreibbar sind. Bei technischer Offenheit geht es nicht primär darum, dass Nutzer:innen in die Lage versetzt werden, Ergebnisse selbst zu reproduzieren, sondern darum, dass unabhängige Dritte die KI-Tools kritisch analysieren und bewerten können. Durch solche Analysen kann nicht nur die Zuverlässigkeit der KI-Tools überprüft, sondern auch die Vertrauenswürdigkeit der Anbieter:innen bewertet werden, was wiederum das Vertrauen von Nutzer:innen in die KI-Tools stärken kann.

Die bisher genannten Empfehlungen reichen allerdings unter Umständen nicht aus, um genug Vertrauen in KI-gestützte Deliberation zu schaffen. Sie sorgen im besten Fall für hinreichende Zuverlässigkeit und Transparenz bei der Verlässlichkeitsprüfung. Die durch systematische Evaluation optimierte Zuverlässigkeit von KI-Tools bleibt jedoch begrenzt. Damit können und werden diese Tools im Einzelfall fehleranfällig sein. Selbst wenn die Fehlerquote gering ist, kann dies insgesamt zu einem Vertrauensverlust führen, wenn die Fehleranfälligkeit der Systeme nicht adäquat abgefangen wird. Hierfür sind weitere Maßnahmen notwendig, um Vertrauen in KI-gestützte Deliberation zu schaffen.

So sollte es eine hinreichende **Transparenz und Erklärbarkeit der Ergebnisse** geben. Nutzer:innen sollten nachvollziehen können, wie und warum bestimmte Ergebnisse entstehen. Hierbei geht es nicht unbedingt um kausale Erklärungen auf technischer Ebene, die für die meisten Nutzer:innen ohnehin selten hilfreich wären, sondern um nachvollziehbare Erklärungen für Techniklaien. Je nach Aufbau der KI-Anwendung gibt es dafür verschiedene Möglichkeiten: Wenn die KI Informationen aus Dokumenten extrahiert, sollten diese Quellen und ihr Zusammenhang mit dem Ergebnis transparent gemacht werden; wenn die KI ihr Ergebnis in mehreren Schritten erarbeitet, sollten auch die Zwischenergebnisse einsehbar sein. Eine besondere Rolle kommt generativen Sprachmodellen zu: Von diesen kann man sich Ergebnisse durch Rückfragen erklären lassen und Rechtfertigungen einfordern. Bei den neueren Reasoning-Modellen, einer Weiterentwicklung des Chain-of-Thought-Ansatzes (Betz u. a. 2021), können die der Antwort vorgeschalteten „Überlegungsschritte“ (engl. *reasoning trace*) für Nutzer:innen einsehbar gemacht werden. Unabhängig davon, ob diese natürlichsprachlichen Erklärungen und Rechtfertigungen tatsächlich kausal relevant für die Ergebnisse sind oder nur ex post erzeugt werden, können sie für Nutzer:innen hilfreich sein, um die Ergebnisse besser zu verstehen und einzuordnen. Sie können auch als Grundlage für eine mögliche Prüfung der Ergebnislegitimität dienen.

Damit kommen wir zu einer weiteren Anforderung. Die Ergebnisse algorithmischer Entscheidungen sollten **anfechtbar und revidierbar** sein. Nutzer:innen sollten nicht nur die Möglichkeit haben, Rechtfertigungen und Erklärungen einzufordern, sondern darüber hinaus ihr Recht ausüben können, die Ergebnisse anzufechten und Korrekturen zu verlangen. Dafür muss es in den entsprechenden deliberativen Kontexten Prozesse und Anlaufstellen geben. Die aktive Einbindung der Nutzer:innen hilft nicht nur, KI-Anwendungen zu verbessern, sondern zeigt Nutzer:innen auch, dass sie fehlerhaften Ergebnissen einer KI nicht ausgeliefert sind.

Für den **Umgang mit Mehrdeutigkeit und Kontextabhängigkeit** gibt es je nach konkretem Einsatzszenario unterschiedliche Möglichkeiten, die sich kombinieren lassen. So kann in einem ersten Schritt, Mehrdeutigkeit und Kontextabhängigkeit durch die Bereitstellung von Kontextinformationen sowie durch die Festlegung klarer Definitionen reduziert werden. Auf der Evaluationsseite setzt dies voraus, dass Test- und ggf. Trainingsdatensätze mit entsprechenden Kontextinformationen und klaren Definitionen erstellt werden. Auf der Seite der KI-Tools muss dafür gesorgt werden, dass die notwendigen Kontextinformationen verfügbar sind und die KI-Tools so ausgestaltet sind, dass sie diese Informationen auch nutzen können.

Kann durch solche Maßnahmen trotzdem keine Ergebniseindeutigkeit sichergestellt werden, muss mit der verbleibenden Mehrdeutigkeit transparent umgegangen werden. Die beiden vorgestellten Prototypen (siehe Kapitel 3) zeigen zwei grundlegende und komplementäre Wege auf:

Eine Möglichkeit besteht darin, die Mehrdeutigkeit so aufzulösen, dass die KI-Anwendung in einem ersten Schritt aufgefordert wird, den Interpretationsspielraum durch die Formulierung unterschiedlicher Interpretationen auszuloten. Anschließend werden dann für jede der identifizierten Interpretationen die weiteren Schritte der Pipeline durchlaufen. Dieser Weg wird durch die EvidenceSeeker-Boilerplate illustriert (Kapitel 3.2). Die Details der Disambiguierung müssen in Abhängigkeit der konkreten Anwendung gestaltet werden. Im EvidenceSeeker ging es vor allem darum, deskriptive und normative Aussagen zu unterscheiden, weil diese Differenzierung wichtig für die weiteren Schritte des Faktencheckprozesses ist. Disambiguierung stellt auch separate Anforderungen an die Evaluierung, weil es hier nicht nur darum geht, ob die KI-Tools die richtigen Ergebnisse liefern, sondern auch, ob sie Mehrdeutigkeiten korrekt erkennen und auflösen können.

Eine andere Möglichkeit besteht darin, die Mehrdeutigkeit nicht aufzulösen, sondern die damit verbundene Unsicherheit in den Ergebnissen transparent zu machen. Statt eindeutiger Antworten könnten Unsicherheiten qualitativ oder quantitativ angegeben werden. Es ist auch möglich, dass KI-Pipelines sich bei unzureichender Informationslage einer Antwort explizit enthalten, wie der Toxicity-Detector illustriert (Kapitel 3.1). Auch dieser Weg stellt besondere Anforderungen an die Evaluierung. So müssen mögliche Unsicherheiten in den Testdatensätzen dargestellt werden, um zu evaluieren, ob die KI-Tools diese korrekt angeben können.

**Zusammenfassung der Empfehlungen**

1. **Unabhängige und transparente Evaluationen:** Die Evaluierung von KI-Tools sollte von unabhängigen Dritten durchgeführt werden, und die Ergebnisse solcher Evaluationen sollten transparent und vollständig veröffentlicht werden.
2. **Technische Transparenz:** Die technischen Details der KI-Pipelines sollten offengelegt werden, damit unabhängige Dritte nachvollziehen können, wie Ergebnisse entstehen und ob grundlegende Anforderungen wie z.B. Datensicherheit erfüllt sind.
3. **Transparenz und Erklärbarkeit von Ergebnissen:** Nutzer:innen sollten nachvollziehen können, wie und warum bestimmte Ergebnisse entstehen. Insbesondere sollten Erklärungen und Rechtfertigungen für Ergebnisse Nutzer:innen zugänglich sein.
4. **Revidierbarkeit algorithmischer Entscheidungen:** Nutzer:innen sollten nicht nur die Möglichkeit haben, Rechtfertigungen und Erklärungen einzufordern, sondern darüber hinaus ihr Recht ausüben können, die Ergebnisse anzufechten und Korrekturen einzufordern.
5. **Umgang mit Mehrdeutigkeit und Kontextabhängigkeit:** Mehrdeutigkeit und Kontextabhängigkeit sollten durch die Bereitstellung von Kontextinformationen sowie die Festlegung klarer Definitionen reduziert werden. Wenn trotzdem keine Ergebniseindeutigkeit sichergestellt werden kann, sollte mit der verbleibenden Mehrdeutigkeit transparent umgegangen werden, zum Beispiel durch die explizite Darstellung von Unsicherheiten oder die Formulierung unterschiedlicher Interpretationen.

Diese Anforderungen formulieren Ideale, die in der Praxis mehr oder weniger stark umgesetzt werden können. Wie wichtig sie für einen bestimmten Anwendungsfall sind, lässt sich nicht pauschal beantworten, sondern hängt davon ab, wie relevant die genannten Herausforderungen im konkreten Fall sind. Darüber hinaus müssen die formulierten Gründe womöglich mit anderen Überlegungen abgewogen werden. So könnten sie beispielsweise mit berechtigten Geschäftsinteressen in Konflikt stehen. Hier gilt es, gesamtgesellschaftlich passende regulatorische Rahmenbedingungen für deliberative KI zu schaffen, um solche Konflikte möglichst gering zu halten.

KI hat großes Potenzial zur Unterstützung und Verbesserung deliberativer Prozesse. Ob dieses Potenzial ausgeschöpft wird, entscheidet sich jedoch nicht allein anhand der technischen Leistungsfähigkeit der KI-Anwendungen, sondern an ihrer Einbettung in transparente, überprüfbare und kontrollierbare Verfahren.

## 4.3 Praktische Herausforderungen

Neben den bereits genannten Herausforderungen, die sich vor allem auf die Evaluierung und Optimierung von KI-Tools sowie deren Zusammenhang mit dem Vertrauen in KI-basierte Deliberation beziehen, gibt es eine ganze Reihe praktischer Herausforderungen, die zum Teil technologiespezifisch und zum Teil spezifisch für den zivilgesellschaftlichen Bereich sind. Diese Herausforderungen können hier weder vollständig noch abschließend behandelt werden. Wir wollen zumindest Herausforderungen aufgreifen, die uns im Laufe des Projekts begegnet sind, und vorläufige Überlegungen darstellen.[10]

Eine praktische Herausforderung bei der Entwicklung KI-basierter Applikationen ist die **hohe Geschwindigkeit, mit der sich die Technologie weiterentwickelt.** Das betrifft nicht nur die Sprachmodelle selbst, sondern auch die Frameworks und Möglichkeiten, KI-basierte Arbeitsabläufe zu gestalten. Während zu Beginn des KIdeKu-Projekts die Entwicklung von KI-Tools hauptsächlich auf der Grundlage statischer Pipelines erfolgte, hat sie sich im Laufe des Projekts zunehmend in Richtung flexibler Pipelines in Form von agentenbasierten Systemen entwickelt. Das macht statische KI-Workflows nicht überflüssig, aber es zeigt, dass die Entwicklung von KI-Tools in ständigem Wandel begriffen ist und es leicht passieren kann, dass eine konkrete Implementierung nach kurzer Zeit schon wieder veraltet ist.

Die Entwicklung modellagnostischer Anwendungen stellt damit lediglich die grundlegendste Anforderung an KI-Tools dar. KI-Tools sollten darüber hinaus so gestaltet werden, dass es möglichst einfach ist, die konkrete technische Implementierung zu erweitern bzw. vollständig auszuwechseln. Dafür gibt es verschiedene Möglichkeiten: So sollten Anwendungen mindestens stark modular aufgebaut sein, damit einzelne Komponenten unabhängig voneinander angepasst und ausgetauscht werden können. Die immer weiter voranschreitende KI-getriebene Softwareentwicklung schafft selbst neue Möglichkeiten mit diesem Problem umzugehen. So könnte man die Ziele und Umsetzungskonzepte natürlichsprachlich als generische Beschreibungen formulieren und die Implementierung darauf aufbauend weitgehend KI-getrieben umsetzen. Diese generischen Beschreibungen sollten möglichst explizit und vollständig sein und unter anderem das Design, die Anforderungen, die Arbeitsabläufe, die Prompts sowie den Technologie-Stack umfassen. Entwickeln sich die Frameworks dann so stark weiter, dass eine Neuimplementierung notwendig wird, müsste man – etwas vereinfacht ausgedrückt – lediglich die Beschreibung des Technologie-Stacks anpassen, um die neue Implementierung zu generieren.

Ein konkreterer Vorschlag ist die Verwendung sogenannter *Agent Skills* (Engl. für „Agentenfähigkeiten“), die in der agentenbasierten KI-Entwicklung zunehmend an Popularität gewinnen.[11] Skills sind im Grunde genommen natürlichsprachliche Anleitungen,

[10] Für einen Überblick aus der Perspektive von NROs sind die Resultate des KINiro-Projekts der OTH Regensburg von großer Relevanz. Dort wurde in qualitativen und quantitativen Studien untersucht, welche Herausforderungen NROs bei der Entwicklung und Nutzung von KI-Tools sehen.

[11] Skills wurden ursprünglich von Anthropic für Claude eingeführt, werden aber mittlerweile von den meisten Frameworks für agentenbasierte KI-Tools unterstützt.

die beschreiben, wie bestimmte Aufgaben oder Funktionen erfüllt werden sollen. Sie können in agentenbasierten Systemen dynamisch aufgerufen und ausgeführt werden, um bestimmte Funktionen zu erfüllen, und werden unabhängig von der konkreten technischen Implementierung formuliert.

Am Beispiel der vorgestellten Pipeline des EvidenceSeekers: Die Prompts für die drei unterschiedlichen Schritte der Pipeline sind bereits technologieunabhängig formuliert und liegen in Konfigurationsdateien vor, sodass sie relativ einfach angepasst und in alternativen Implementierungen wiederverwendet werden können. Weitergehen könnte man, indem man die unterschiedlichen Schritte der Pipeline und das Design der darauf aufbauenden Pipeline als Skills formuliert. Die so beschriebene Fähigkeit, Evidenzen in einer vorhandenen Wissensbasis zu identifizieren und den Grad ihrer Bestätigung zu qualifizieren, ließe sich damit in agentenbasierten Abläufen technologieunabhängig einbinden.

Sofern das Ziel darin besteht, KI-Anwendungen so zu konzeptionieren und umzusetzen, dass sie über das Stadium von Prototypen hinausgehen und tatsächlich in der Praxis eingesetzt werden, ergeben sich eine Reihe weiterer praktischer Herausforderungen.

Zunächst stellt sich die grundlegende strategische Frage nach der **Wahl des Modell-Ökosystems**. Open-Weight-Modelle, wie sie über Hugging Face angeboten werden, bieten potenziell größere Kontrolle, Transparenz und Unabhängigkeit. Sie können außerdem über eigene IT-Infrastrukturen betrieben werden, was die Anpassung an spezifische Anforderungen erleichtert, Abhängigkeiten vermeidet und insbesondere die Einhaltung sowie das Monitoring von Datenschutz- und Sicherheitsanforderungen ermöglicht.

Demgegenüber müssen die Anforderungen an Infrastruktur, Wartung, Sicherheit und Modellpflege selbst erfüllt werden. Sofern KI-basierte Deliberationstools im zivilgesellschaftlichen Bereich entwickelt und eingesetzt werden, kann dies eine erhebliche Hürde darstellen, da diese Organisationen häufig klein sind und nur über stark begrenzte Ressourcen verfügen. Das ist umso bedenklicher, als dass Datenschutz, Transparenz, Zuverlässigkeit und Sicherheit gerade in deliberativen Kontexten besonders relevant sind (siehe oben).

Eine mögliche Lösung könnte darin bestehen, dass Bund oder Länder zivilgesellschaftlichen Organisationen KI-Infrastruktur zur Nutzung bereitstellen, die über entsprechende API-Schnittstellen zugänglich ist.

Ein weiterer Aspekt bei der Abwägung zwischen Open-Weight- und proprietären Modellen betrifft die Performance. Nach wie vor besteht eine Lücke zwischen der Leistungsfähigkeit von Open-Weight-Modellen und der von proprietären Modellen, die von großen Plattformanbietern bereitgestellt werden. Aus der Perspektive von User:innen reicht es unter Umständen nicht aus, dass KI-Tools hinreichend zuverlässig sind, sondern sie müssen mit der Leistungsfähigkeit von ChatGPT und Co. mithalten können, um breite Akzeptanz zu erreichen. Das gilt insbesondere für Anwendungen, die in direkter Konkurrenz zu proprietären Modellen stehen, etwa KI-gestützte Chatbots oder Recherchetools.

Erst einmal lässt sich festhalten, dass sich diese Performance-Lücke in den letzten zwei Jahren verkleinert hat. Geht dieser Trend weiter, könnte diese Herausforderung in Zukunft weniger relevant sein. Darüber hinaus können verschiedene Maßnahmen ergriffen werden, um die Akzeptanz von Applikationen, die auf Open-Weight-Modellen basieren, zu steigern. So könnten die genannten Vorteile stärker kommuniziert werden. Vielleicht sollte auch darauf verzicht werden, mit generisch ausgelegten KI-Tools zu konkurrieren. Stattdessen sollte der Fokus auf die Entwicklung von KI-Anwendungen liegen, die einen sichtbaren Mehrwert generieren, indem sie auf spezifische organisatorische Prozesse, Fachlogiken oder regulatorische Anforderungen zugeschnitten werden und domänenspezifisches und institutionelles Wissen strukturiert einbinden.

Ein weiterer zentraler Aspekt betrifft die **Nachhaltigkeit von KI-Projekten** im zivilgesellschaftlichen Bereich. Gerade im Kontext öffentlich geförderter oder zivilgesellschaftlicher Vorhaben tritt häufig das Problem auf, dass Anwendungen nach Projektende nicht weiterentwickelt und in der Folge auch nicht genutzt werden. Mit dem Auslaufen der Finanzierung enden oft Wartung, Weiterentwicklung und Hosting. Zurück bleiben veraltete Systeme, die Sicherheitsrisiken mit sich bringen, oder schlicht abgeschaltete Dienste.[12]

Dieses Problem ist sicherlich komplex und hängt mit strukturellen Rahmenbedingungen zusammen. Eine Möglichkeit, dem entgegenzuwirken, ist, die Anwendungen so zu entwickeln und aufzusetzen, dass sie möglichst reibungslos von Dritten aufgegriffen und weiterentwickelt werden können. Das könnte zum Beispiel durch die Veröffentlichung von Quellcode unter geeigneten liberalen Open-Source-Lizenzen, die Bereitstellung ausführlicher Dokumentation und die Einbindung von Nutzer:innen in die Entwicklung geschehen.

[12] Ein Phänomen, das sich exemplarisch im „Civic Tech Graveyard“ dokumentiert findet.

# Literatur


Albladi, Aish, Minarul Islam, Amit Das, u. a. 2025. „Hate Speech Detection Using Large Language Models: A Comprehensive Review“. *IEEE Access* 13: 20871–92. https://doi.org/10.1109/ACCESS.2025.3532397.

Argyle, Lisa P., Christopher A. Bail, Ethan C. Busby, u. a. 2023. „Leveraging AI for Democratic Discourse: Chat Interventions Can Improve Online Political Conversations at Scale“. *Proceedings of the National Academy of Sciences* 120 (41): e2311627120. https://doi.org/10.1073/pnas.2311627120.

Bächtiger, André, Simon Niemeyer, Michael Neblo, Marco R. Steenbergen, und Jürg Steiner. 2010. „Disentangling Diversity in Deliberative Democracy: Competing Theories, Their Blind Spots and Complementarities*“. *Journal of Political Philosophy* 18 (1): 32–63. https://doi.org/10.1111/j.1467-9760.2009.00342.x.

Baribi-Bartov, Sahar, Briony Swire-Thompson, und Nir Grinberg. 2024. „Supersharers of Fake News on Twitter“. *Science* 384 (6699): 979–82. https://doi.org/10.1126/science.adl4435.

Bertram, Markus, Johannes Schäfer, und Thomas Mandl. 2023. „Comparative Survey of German Hate Speech Datasets: Background, Characteristics and Biases“. *CEUR Workshop Proceedings*, Oktober.

Betz, Gregor, und Sebastian Cacean. 2012. *Ethical Aspects of Climate Engineering*. KIT Scientific Publishing.

Betz, Gregor, Kyle Richardson, und Christian Voigt. 2021. „Thinking Aloud: Dynamic Context Generation Improves Zero-Shot Reasoning Performance of GPT-2“. *arXiv preprint arXiv:2103.13033*. https://arxiv.org/abs/2103.13033.

Brennauer, Jutta, Valentin Dander, Corinna Dolezalek, und Kompetenznetzwerk gegen Hass im Netz, Hrsg. 2024. *Lauter Hass - leiser Rückzug: wie Hass im Netz den demokratischen Diskurs bedroht: Ergebnisse einer repräsentativen Befragung*. Februar 2024. Kompetenznetzwerk Hass im Netz.

Cacean, Sebastian. 2012. „Ethische Aspekte von Cognitive Enhancement“. In *Sport, Doping und Enhancement Ergebnisse und Denkanstöße*, herausgegeben von Giselher Spitzer und Elk Franke. Sportverlag Strauß.

Chen, Canyu, und Kai Shu. 2024. *Can LLM-generated Misinformation Be Detected?*

Cohen, Joshua. 2005. „Deliberation and Democratic Legitimacy“. In *Debates in Contemporary Political Philosophy*. Routledge.

Fallis, Don. 2015. „What Is Disinformation¿‘ *Library Trends* 63 (3): 401–26.

Fortuna, Paula, und Sérgio Nunes. 2018. „A Survey on Automatic Detection of Hate Speech in Text“. *ACM Comput. Surv.* 51 (4): 85:1–30. https://doi.org/10.1145/3232676.

Fortuna, Paula, Juan Soler, und Leo Wanner. 2020. „Toxic, Hateful, Offensive or Abusive? What Are We Really Classifying? An Empirical Analysis of Hate Speech Datasets“. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, herausgegeben von Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, u. a. European Language Resources Association.

Frank, David, Viktoria Scheidler, und Eva Schmid. 2024. *Argumentkartierung von politischen Debatten: Identifikation von Zielkonflikten und Lösungsstrategien in transdisziplinären Forschungsprojekten*. {Working Paper} SP III 2024-601. WZB Discussion Paper.

Friess, Dennis, und Christiane Eilders. 2015. „A Systematic Review of Online Deliberation Research“. *Policy & Internet* 7 (3): 319–39. https://doi.org/10.1002/poi3.95.

Friess, Dennis, Carina Weinmann, und Mira Warné. 2025. „AI and Deliberation: Normative Ideals in the Light of Current AI Research - A Review“. *Journal of Deliberative Democracy* 21 (1). https://doi.org/10.16997/jdd.1805.

Gerber, Marlène, André Bächtiger, Irena Fiket, Marco Steenbergen, und Jürg Steiner. 2014. „Deliberative and Non-Deliberative Persuasion: Mechanisms of Opinion Formation in EuroPolis“. *European Union Politics* 15 (3): 410–29. https://doi.org/10.1177/1465116514528757.

Giraud, Eva Haifa, Elizabeth Poole, Ed de Quincey, und John E. Richardson. 2025. „Learning from Online Hate Speech and Digital Racism: From Automated to Diffractive Methods in Social Media Analysis“. *The Sociological Review* 73 (6): 1388–407. https://doi.org/10.1177/00380261241305260.

Goldschmidt, Rüdiger. 2014. *Kriterien Zur Evaluation von Dialog- Und Beteiligungsverfahren*. VS Verlag für Sozialwissenschaften.

Guida, Matteo, Yulia Otmakhova, Eduard Hovy, und Lea Frermann. 2025. „LLMs for Argument Mining: Detection, Extraction, and Relationship Classification of Pre-Defined Arguments in Online Comments“. In *Proceedings of the 23rd Annual*

*Workshop of the Australasian Language Technology Association*, herausgegeben von Jonathan K. Kummerfeld, Aditya Joshi, und Mark Dras. Association for Computational Linguistics.

Guo, Keyan, Alexander Hu, Jaden Mu, u. a. 2024. *An Investigation of Large Language Models for Real-World Hate Speech Detection*. arXiv:2401.03346. arXiv. https://doi.org/10.48550/arXiv.2401.03346.

Guo, Zhijiang, Michael Schlichtkrull, und Andreas Vlachos. 2022. „A Survey on Automated Fact-Checking“. *Transactions of the Association for Computational Linguistics* 10 (Februar): 178–206. https://doi.org/10.1162/tacl_a_00454.

Habermas, Jürgen. 1981. *Theorie Des Kommunikativen Handelns.* Suhrkamp.

Jungherr, Andreas, und Adrian Rauchfleisch. 2025. „Artificial Intelligence in Deliberation: The AI Penalty and the Emergence of a New Deliberative Divide“. *Government Information Quarterly* 42 (4): 102079. https://doi.org/10.1016/j.giq.2025.102079.

Krippendorff, Klaus. 2019. *Content Analysis an Introduction to Its Methodology*. Fourth edition. SAGE.

Kruk, Julia, Michela Marchini, Rijul Magu, Caleb Ziems, David Muchlinski, und Diyi Yang. 2024. „Silent Signals, Loud Impact: LLMs for Word-Sense Disambiguation of Coded Dog Whistles“. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, herausgegeben von Lun-Wei Ku, Andre Martins, und Vivek Srikumar. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.675.

Lanius, David. 2017. *Wie argumentieren Rechtspopulisten? Eine Argumentationsanalyse des AfD-Wahlprogramms*. (Karl), Online-Vorab-Publikation. https://doi.org/10.5445/IR/1000074060.

Lawrence, John, und Chris Reed. 2019. „Argument Mining: A Survey“. *Computational Linguistics*, Oktober, 1–55. https://doi.org/10.1162/COLI_a_00364.

Lippi, Marco, und Paolo Torroni. 2016. „Argumentation Mining: State of the Art and Emerging Trends“. *ACM Trans. Internet Technol.* 16 (2): 10:1–25. https://doi.org/10.1145/2850417.

Mandl, Thomas, Sandip Modha, Prasenjit Majumder, u. a. 2019. „Overview of the HASOC Track at FIRE 2019: Hate Speech and Offensive Content Identification in Indo-European Languages“. *Proceedings of the 11th Annual Meeting of the Forum for Information Retrieval Evaluation* (New York, NY, USA), Fire '19, 14–17. https://doi.org/10.1145/3368567.3368584.

Mirzakhmedova, Nailia, Marcel Gohsen, Chia Hao Chang, und Benno Stein. 2024. „Are Large Language Models Reliable Argument Quality Annotators¿‘ In *Robust Argumentation Machines*, herausgegeben von Philipp Cimiano, Anette Frank, Michael Kohlhase, und Benno Stein. Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-63536-6_8.

Nanz, Patrizia, und Miriam Fritsche. 2012. *Handbuch Bürgerbeteiligung - Verfahren Und Akteure, Chancen Und Grenzen*. Bd. 1200. Bpb Schriftenreihe. Bundeszentrale für politische Bildung.

Neblo, Michael A. 2011. „Family Disputes: Diversity in Defining and Measuring Deliberation“. *Swiss Political Science Review* 13 (4): 527–57. https://doi.org/10.1002/j.1662-6370.2007.tb00088.x.

Pendzel, Sagi, Tomer Wullach, Amir Adler, und Einat Minkov. 2023. *Generative AI for Hate Speech Detection: Evaluation and Findings*. arXiv:2311.09993. arXiv. https://doi.org/10.48550/arXiv.2311.09993.

Plaza-del-arco, Flor Miriam, Debora Nozza, und Dirk Hovy. 2023. „Respectful or Toxic? Using Zero-Shot Learning with Language Models to Detect Hate Speech“. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, herausgegeben von Yi-ling Chung, Paul Röttger, Debora Nozza, Zeerak Talat, und Aida Mostafazadeh Davani. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.woah-1.6.

Podolak, Jakub, Szymon Łukasik, Paweł Balawender, u. a. 2024. „LLM Generated Responses to Mitigate the Impact of Hate Speech“. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, herausgegeben von Yaser Al-Onaizan, Mohit Bansal, und Yun-Nung Chen. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-emnlp.931.

Qureshi, Muhammad Deedahwar Mazhar, M. Atif Qureshi, und Wael Rashwan. 2025. *Explainable AI for Hate Speech Moderation: A Stakeholder-Centered and Socially Grounded Review*. TechRxiv. https://doi.org/10.36227/techrxiv.175440435.54783623/v1.

Saha, Punyajoy, Aalok Agrawal, Abhik Jana, Chris Biemann, und Animesh Mukherjee. 2024. „On Zero-Shot Counterspeech Generation by LLMs“. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, herausgegeben von Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, und Nianwen Xue. ELRA and ICCL.

Schmidt, Anna, und Michael Wiegand. 2017. „A Survey on Hate Speech Detection Using Natural Language Processing“. In *Proceedings of the Fifth International*

*Workshop on Natural Language Processing for Social Media*, herausgegeben von Lun-Wei Ku und Cheng-Te Li. Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-1101.

Setty, Vinay. 2024. „Surprising Efficacy of Fine-Tuned Transformers for Fact-Checking over Larger Language Models“. *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval* (New York, NY, USA), Sigir '24, 2842–46. https://doi.org/10.1145/3626772.3661361.

Steenbergen, Marco R., André Bächtiger, Markus Spörndli, und Jürg Steiner. 2003. „Measuring Political Deliberation: A Discourse Quality Index“. *Comparative European Politics* 1 (1): 21–48. https://doi.org/10.1057/palgrave.cep.6110002.

Steiner, Jürg, André Bächtiger, Marco Steenbergen, und Markus Sprörndli. 2004. *Deliberative Politics in Action: Analyzing Parliamentary Discourse*. Theories of Institutional Design. Cambridge University Press.

Tessler, Michael Henry, Michiel A. Bakker, Daniel Jarrett, u. a. 2024. „AI Can Help Humans Find Common Ground in Democratic Deliberation“. *Science* 386 (6719). https://doi.org/10.1126/science.adq2852.

Törnberg, Petter, und Juliana Chueri. 2025. „When Do Parties Lie? Misinformation and Radical-Right Populism Across 26 Countries“. *The International Journal of Press/Politics*, Januar, 19401612241311886. https://doi.org/10.1177/19401612241311886.

Vykopal, Ivan, Matúš Pikuliak, Simon Ostermann, und Marián Šimko. 2024. *Generative Large Language Models in Automated Fact-Checking: A Survey*. arXiv. https://doi.org/10.48550/arXiv.2407.02351.

Wachsmuth, Henning, Gabriella Lapesa, Elena Cabrio, u. a. 2024. „Argument Quality Assessment in the Age of Instruction-Following Large Language Models“. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, herausgegeben von Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, und Nianwen Xue. ELRA and ICCL.

Wiegand, Michael, Melanie Siegel, und Josef Ruppenhofer. 2018. „Overview of the GermEval 2018 Shared Task on the Identification of Offensive Language“. *ISBN*, 1–10.

Woyke, Wichard. 2021. „Politische Partizipation“. In *Handwörterbuch des politischen Systems der Bundesrepublik Deutschland*, herausgegeben von Uwe Andersen, Jörg Bogumil, Stefan Marschall, und Wichard Woyke. Springer Fachmedien. https://doi.org/10.1007/978-3-658-23666-3_112.

Yu, Zehui, Indira Sen, Dennis Assenmacher, u. a. 2024. „The Unseen Targets of Hate: A Systematic Review of Hateful Communication Datasets“. *Social Science Computer Review*, Juni, 08944393241258771. https://doi.org/10.1177/08944393241258771.

Zhang, Xuan, und Wei Gao. 2023. „Towards LLM-based Fact Verification on News Claims with a Hierarchical Step-by-Step Prompting Method“. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, herausgegeben von Jong C. Park, Yuki Arase, Baotian Hu, u. a. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.ijcnlp-main.64.