



universität
uulm

**Fakultät für
Ingenieurwissenschaften,
Informatik und
Psychologie**
Institut für Datenbanken
und Informationssysteme (DBIS)

Identifikation von DSGVO-kritischen Aktivitäten in Business Prozessen mittels Large Language Models

Abschlussarbeit an der Universität Ulm

Vorgelegt von:

Merten Dieckmann
merten.dieckmann@uni-ulm.de
1058340

Gutachter:

Prof. Dr. Manfred Reichert
Prof. Dr. Rüdiger Pryss

Betreuer:

Magdalena von Schwerin

2025

Fassung 16. Oktober 2025

© 2025 Merten Dieckmann

Satz: PDF- \LaTeX 2 _{ϵ}

Inhaltsverzeichnis

Abkürzungen	vi
1 Einleitung	2
1.1 Problemstellung	3
1.2 Zielsetzung und Beiträge	4
1.3 Aufbau der Arbeit	5
2 Hintergrund und verwandte Arbeiten	6
2.1 Datenschutzgrundverordnung (DSGVO)	6
2.2 Business Process Model and Notation (BPMN)	7
2.3 Large Language Models (LLMs)	10
2.4 Verwandte Arbeiten	12
3 Problemdefinition und Zielkriterien	13
3.1 Aufgabenstellung	14
3.2 Qualitätsziele	14
3.3 Scope und Annahmen	17
3.4 Experimentdesign	18
4 Design und Implementierung der Klassifizierungspipeline	21
4.1 BPMN Preprocessing	21
4.2 Prompt Engineering	23
4.3 Validierung der Ausgabe	27
4.4 API-Design	29
4.5 Webapp-Sandbox	32
5 Labeling und Datensätze	34
5.1 Labeling Tool	34

5.2	Quellen und Eigenschaften der Datensätze	35
5.3	Labeling-Guide	37
6	Evaluationsframework	39
6.1	Use-Cases und Anforderungen	39
6.2	Testdaten	42
6.3	Konfiguration einer Evaluierung	42
6.4	Architektur und Komponenten	44
6.5	Evaluationsergebnisse	47
6.6	Frontend	49
6.7	Erweiterbarkeit	54
7	Modellauswahl	55
7.1	Kriterien	55
7.2	Modellvorstellung	57
8	Versuchsaufbau und Durchführung	61
8.1	Einheitliche Klassifizierungspipeline und Datensätze	62
8.2	Konfigurationen	62
8.3	Durchführung	64
9	Ergebnisse	66
9.1	Überblick	66
9.2	Analyse	70
9.3	Robustheit	75
9.4	Fallstudien	76
9.5	Antworten auf Forschungsfragen	80
10	Diskussion	81
10.1	Interpretation der Befunde	81
10.2	Hoher Recall vs. Präzision	81
10.3	EU-Modelle	82
10.4	Open-Source Modelle	82
10.5	Modellgrößen	82
10.6	Grenzen	82
11	Zusammenfassung	83

Inhaltsverzeichnis

12 Aussicht	84
A Quelltexte	85
Literatur	100

Abkürzungen

BPMN	Business Process Model and Notation
DSGVO	Datenschutz-Grundverordnung
EU	Europäische Union
FN	False Negative
FP	False Positive
KI	Künstliche Intelligenz
MoE	Mixture-of-Experts
OSI	Open Source Initiative
LM	Language Model
LLM	Large Language Model
RAG	Retrieval Augmented Generation
TN	True Negative
TP	True Positive

Abkürzungen

// TODO Überall nach einheitlichen Begriffsnutzungen schauen (Modell, Prozessmodell, LLMs, BPMN-Prozess, BPMN-Modell, Labels, Annotationen, Evaluation, Experiment)

// TODO Sprache in Bildern/der App und Diagrammen einheitlich anpassen. Entweder es soll alles auf Englisch sein oder auf Deutsch genau wie der Text.

// TODO Recall, Precision und F1-Score und Accuracy überall einheitlich entweder auf Englisch oder Deutsch.

// TODO Einheitlich die Labels von Grafiken, Tabellen und Listings drüber oder drunter machen

// TODO Mixtral 8x7B habe ich in den Diagrammen in orange eingefärbt, was für ein kleines Modell steht, aber es ist ja ein großes Modell, weil es insgesamt über 25B Parameter hat und so behandle ich es auch im Analyse Kapitel. Ggf. muss ich nochmal in Gesamtübersicht der Ergebnisse etwas im Text anpassen und auf jeden Fall die Diagramme anpassen

1 Einleitung

Geschäftsprozesse sind in nahezu allen Organisationen allgegenwärtig und bilden die Grundlage für effiziente Abläufe. Zugleich ist in Europa durch die Datenschutz-Grundverordnung (DSGVO) der Datenschutz zu einem zentralen regulatorischen Aspekt geworden [12, 16]. Unternehmen müssen sicherstellen, dass in ihren Prozessen personenbezogene Daten rechtskonform verarbeitet werden; andernfalls drohen Strafen von bis zu 20 Millionen Euro oder 4% des weltweit gesamten erzielten Jahresumsatzes [16].

Die Überprüfung von Prozessen auf Konformität in Bezug auf Datenschutz ist jedoch zeit- und kostenintensiv [48, 70]. Besonders in großen Organisationen mit hunderten parallel laufenden Prozessen ist eine manuelle Analyse kaum praktikabel und zudem fehleranfällig. Fehlerhafte Untererkennungen datenschutzkritischer Aktivitäten (False Negatives) können weitreichende Folgen haben – von Reputationsschäden bis hin zu hohen Bußgeldern [48].

Vor diesem Hintergrund rücken Large Language Models (LLMs) als aufstrebende Technologie im Bereich Künstliche Intelligenz (KI) in den Fokus. Sie sind darauf trainiert, natürliche Sprache auch in langen und komplexen Texten zu verstehen, Zusammenhänge über große Kontexte hinweg zu erkennen und Anweisungen zu befolgen. Damit erscheinen LLMs als vielversprechender Ansatz für das automatisierte Screening von Prozessmodellen. Erste Arbeiten belegen dieses Potenzial, etwa bei der Identifikation datenschutzrelevanter Verarbeitungstätigkeiten oder in der Analyse von Datenschutzerklärungen [13, 60].

Besonders interessant sind in diesem Kontext europäische Open-Source-Modelle wie die von Mistral [3]. Sie sind zum einen frei verfügbar und transparent, zum anderen wurden sie bislang kaum im Hinblick auf DSGVO-bezogene Aufgaben evaluiert. Es fehlen belastbare, reproduzierbare empirische Vergleiche, die eine fundierte Bewertung dieser Modelle erlauben würden [68].

1.1 Problemstellung

Trotz der genannten Potenziale fehlt es bisher an standardisierten, reproduzierbaren Vergleichen verschiedener Modelle für die konkrete Aufgabe Aktivitäten in Geschäftsprozessen nach „kritisch“ und „unkritisch“ zu klassifizieren. Erste Ansätze, wie z.B. der von Nake et al. [48], zeigen dass maschinelles Lernen grundsätzlich in der Lage ist DSGVO-kritische Aktivitäten in textuellen Prozessbeschreibungen zu erkennen; dennoch existieren keine einheitlichen Benchmarks, die einen systematischen vergleiche unterschiedlicher LLMs erlauben.

Auch von Schwerin et al. [68] heben hervor, dass trotz großer Fortschritte im Einsatz von LLMs für juristische Aufgaben bislang erhebliche Lücken in der Evaluation für compliance-spezifische Anwendungen bestehen und geeignete DSGVO-spezifische Benchmarks fehlen. Somit mangelt es derzeit an einer belastbaren empirischen Grundlage, um Modelle zuverlässig und vergleichbar zu bewerten.

Besonders interessant ist die Frage, wie sich Open-Source-Modelle - insbesondere mit Ursprung aus der Europäischen Union (EU) - im Vergleich zu internationalen außerhalb der EU entwickelten Modellen schlagen und welche Trade-offs dabei entstehen [68]. Diese Perspektive ist nicht nur aus Leistungs-, sondern auch aus Transparenz- und Regulierungsgründen relevant.

Eine zusätzliche Herausforderung ergibt sich aus der Natur von Business Process Model and Notation (BPMN)-Modellen: Typischerweise konzentrieren sie sich auf den Kontrollfluss und vernachlässigen die Datenebene. Datenobjekte werden oftmals gar nicht explizit modelliert oder nur implizit in den Aktivitäten referenziert. Dadurch ist die Datennutzung von Aktivitäten nicht direkt erkennbar und muss aus textuellen Beschreibungen und dem Kontext erschlossen werden [66]. Das erschwert die automatische Identifikation von DSGVO-kritischen Aktivitäten, da Algorithmen personenbezogene Datenflüsse zunächst indirekt und über den Kontext ableiten müssen.

1.2 Zielsetzung und Beiträge

Ziel der Arbeit ist es, einen methodischen Beitrag zur automatisierten Identifikation von DSGVO-kritischen Aktivitäten in Geschäftsprozessen zu leisten. Hierfür werden folgende Beiträge angestrebt:

- Entwicklung einer Klassifizierungspipeline für Geschäftsprozesse, die Aktivitäten binär in datenschutzkritisch oder unkritisch einordnet.
- Konzeption und Umsetzung eines Evaluationsframeworks, das reproduzierbare Vergleiche verschiedener LLMs und Algorithmen über eine einheitliche Schnittstelle ermöglicht.
- Entwicklung einer Labelingsoftware zur Erstellung und Annotation von Datensätzen für das Evaluationsframework.
- Aufbau eines repräsentativen Datensatzes aus gelabelten BPMN-Prozessen, inklusive klar definierter Labeling-Kriterien.
- Bereitstellung überprüfbarer empirischer Befunde, inklusive Code, Konfigurationen der Experimente und Seeds, um Nachvollziehbarkeit und Reproduzierbarkeit zu gewährleisten.

Auf dieser Grundlage ergibt sich die zentrale Forschungsfrage dieser Arbeit:

FF1 Wie zuverlässig identifizieren LLMs DSGVO-kritische Aktivitäten in BPMN-Prozessmodellen?

Um diese Frage differenziert beantworten zu können werden außerdem folgende Unterfragen betrachtet:

UF1 Wie gut schneiden europäische Modelle im Vergleich zu internationalen Modellen ab?

UF2 Wie unterscheiden sich große und kleine Modelle in ihrer Leistungsfähigkeit?

UF3 Welche Open-Source-Modelle (insbesondere aus der EU) erzielen die besten Ergebnisse?

UF4 Wie gut schneiden Open-Source-Modelle gegenüber kommerziellen Modellen wie GPT-4o ab?

Für ein initiales Screening reicht, wie in [48], eine binäre Klassifikation (kritisch vs. unkritisch). Eine tiefergehende rechtliche Prüfung kann in einem nachfolgendem Schritt durchgeführt werden und ist nicht Bestandteil dieser Arbeit.

1.3 Aufbau der Arbeit

Die Arbeit ist wie folgt gegliedert: Kapitel 2 gibt einen Überblick über den theoretischen Hintergrund, die DSGVO und BPMN sowie eine Einführung in LLMs und verwandte Arbeiten. Kapitel 3 beschreibt den Rahmen der Entwicklung der Klassifizierungspipeline, des Evaluationsframeworks und der Experimente. Kapitel 4 stellt den entwickelten Algorithmus zur Klassifikation von BPMN-Modellen und dessen einheitliche Schnittstelle vor. Kapitel 5 präsentiert die Architektur und den Funktionsumfang der Evaluationspipeline. Anschließend wird in Kapitel 6 die Labelingsoftware und die Erstellung der Datensätze erläutert. Kapitel 7 zeigt auf wie die Auswahl der LLMs erfolgte. Kapitel 8 erläutert den Versuchsaufbau, Kapitel 9 die Durchführung der Experimente und Kapitel 10 stellt die Ergebnisse vor. In Kapitel 11 werden die Ergebnisse im Kontext der Forschungsfragen diskutiert. Zum Schluss fasst Kapitel 12 die Arbeit zusammen und Kapitel 13 gibt einen Ausblick auf mögliche zukünftige Forschungsthemen.

2 Hintergrund und verwandte Arbeiten

2.1 Datenschutzgrundverordnung (DSGVO)

Die europäische Datenschutz-Grundverordnung (DSGVO) [16] bildet den zentralen rechtlichen Rahmen für den Schutz personenbezogener Daten in der EU. Sie gilt seit dem 25. Mai 2018. Durch die DSGVO werden Betroffenenrechte gestärkt und Verantwortliche zu technischen und organisatorischen Maßnahmen verpflichtet, wie z. B. *Datenschutz durch Technikgestaltung* und *datenschutzfreundliche Voreinstellungen* (Art. 25 DSGVO) [17].

Definitionen

Im Folgenden werden zentrale Begriffe der DSGVO erläutert, die für das Verständnis dieser Arbeit relevant sind:

- **Personenbezogene Daten** sind alle Informationen, die sich auf eine identifizierte oder identifizierbare natürliche Person beziehen (Art. 4 Abs. 1 DSGVO) [16]. Eine Person ist identifizierbar, wenn sie direkt oder indirekt bestimmbar ist (z. B. anhand von Name, Kennnummer, Standortdaten, Online-Kennung).
- **Verarbeitung** bezeichnet *jeden* mit personenbezogenen Daten vorgenommenen Vorgang (Art. 4 Abs. 2 DSGVO) und umfasst insbesondere das **Erheben, Speichern, Verwenden/Nutzen, Offenlegen durch Übermittlung** sowie das **Löschen/Vernichten** [16].
- Im BPMN-Kontext sind alle Aktivitäten als **datenschutzkritisch** zu betrachten, die solche Verarbeitungshandlungen an personenbezogenen Daten vor-

nehmen oder auslösen (z. B. Abruf aus einem Kundendaten-Speicher, Übergabe an externe Stellen).

Abgrenzung: Risiko-Screening vs. Rechtsberatung

Die in dieser Arbeit eingesetzten Klassifizierungsverfahren dienen einem *automatisierten Risiko-Vorscreening* von Prozessaktivitäten. Sie ersetzen keine individuelle Rechtsprüfung im Einzelfall. Insbesondere in Deutschland ist die Erbringung konkreter Rechtsdienstleistungen Personen mit entsprechender Befugnis vorbehalten [9]. Die Ergebnisse sind daher als Entscheidungshilfe zu verstehen und bedürfen - insbesondere bei Grenzfällen - der Bewertung durch qualifizierte Experten.

2.2 Business Process Model and Notation (BPMN)

BPMN ist ein Standard zur Modellierung von Geschäftsprozessen. Die Notation wurde entwickelt, um eine einheitliche Notation bereitzustellen, die sowohl von Geschäftsanalysten als auch von technischen Entwicklern verstanden wird. BPMN-Modelle bestehen aus verschiedenen Elementen wie Aktivitäten, Ereignissen, Gateways und Verbindungen, die zusammen den Ablauf eines Geschäftsprozesses darstellen [51].

Relevante BPMN-Elemente

Für die Identifikation von DSGVO-kritischen Aktivitäten sind insbesondere folgende Elemente relevant, da sie Hinweise auf den Umgang mit (personenbezogenen) Daten geben. Sie sind ebenfalls in Abbildung 2.1 dargestellt:

- **Aktivitäten** bilden die auszuführenden Arbeitsschritte eines Prozesses ab. Sie können Ein- und Ausgaben sowie Datenabhängigkeiten definieren [51]. Durch ihren Namen oder Kontext können Rückschlüsse auf die Verarbeitung personenbezogener Daten gezogen werden.



Abbildung 2.1: Die relevanten BPMN Elemente in Beziehungen zueinander

- **Sequenzflüsse** verbinden Aktivitäten, Ereignisse und Gateways und zeigen die Reihenfolge der Ausführung im Prozess an [51]. Mit ihrer Hilfe kann eine einzelne Aktivität im Kontext des gesamten Prozesses betrachtet werden, indem der Pfad zu und von der Aktivität verfolgt wird.
- **Datenobjekte und Datenspeicher** repräsentieren flüchtige oder persistente Daten, die im Prozess von z.B. Aktivitäten genutzt oder geschrieben werden können [51]. Sie können auch personenbezogene Daten enthalten.
- **Datenassoziationen** (Eingangs- und Ausgangsassoziationen) verbinden Aktivitäten mit Datenobjekten und Datenspeichern und zeigen so Ein- und Ausgaben explizit an [51]. Sie sind ein wichtiges Signal für die Verarbeitung personenbezogener Daten, da sie den direkten Bezug einer Aktivität zu bestimmten Daten verdeutlichen (z.B. Lesezugriff auf eine Kundendatenbank).
- **Pools** modellieren Organisationseinheiten oder Prozessbeteiligte, während **Lanes** Verantwortlichkeiten innerhalb eines Pools darstellen. Innerhalb eines Pools befinden sich die Aktivitäten und anderen Elemente des Prozesses [51]. Die Rollen und Verantwortlichkeiten, die durch Pools und Lanes dargestellt werden, können für die Bewertung der Datenverarbeitung relevant sein.
- **Nachrichtenflüsse** stellen den Austausch von Nachrichten zwischen verschiedenen Pools dar [51]. Sie können auf eine Übermittlung personenbezogener Daten an Dritte hinweisen (z.B. Transfer von Kundendaten an einen



Abbildung 2.2: Beispiel einer Datenassoziation als Datenschutzsinal.

externen Dienstleister).

- **Textannotationen und Assoziationen** dienen dazu, zusätzliche Informationen zu Prozessmodellen hinzuzufügen, die nicht durch die standardmäßigen BPMN-Elemente abgedeckt sind [51]. Sie können genutzt werden, um die Art der Datenverarbeitung zu präzisieren (z.B. „enthält E-Mail-Adresse“).

BPMN-XML

BPMN-Modelle werden in einer XML-Serialisierung gespeichert (BPMN 2.0 XML) [51]. Diese Darstellung enthält alle relevanten Strukturinformationen (Elementtypen, Namen, Beziehungen, Zuordnungen, Positionen der Elemente) und wird von vielen Prozess-Engines und Modellierungswerkzeugen wie Camunda [20] und BPMN.io [19] unterstützt. Für diese Arbeit dient BPMN-XML als Eingabeformat der Klassifizierungspipeline (siehe Kapitel 4).

Im Metamodell von BPMN erben fast alle Elemente von `BaseElement` und damit ein `id`-Attribut. Dieses `id` dient der eindeutigen Referenzierung und ist erforderlich [51]. Diese stabile `id` ist für die Klassifizierungspipeline wichtig, da sie eine stabile Referenzierung der Aktivitäten und anderer Elemente ermöglicht. Dies ist insbesondere dann relevant, wenn die Ergebnisse der Klassifizierung auf die ursprünglichen Prozessmodelle zurückgeführt werden müssen.

Beispiel einer Datenassoziation als Datenschutzsignal

Abbildung 2.2 zeigt ein einfaches Beispiel, wie eine Datenassoziation die DSGVO-Relevanz einer Aktivität verdeutlichen kann. In Abbildung 2.2a ist die Aktivität „Daten prüfen“ ohne Datenassoziation dargestellt, was wenig über die Art der verarbeiteten Daten aussagt. In Abbildung 2.2b hingegen zeigt die eingehende Datenassoziation von einem Datenspeicher „Kunden-DB“, dass die Aktivität personenbezogene Daten verarbeitet. Dies macht die Aktivität als potenziell datenschutzkritisch erkennbar. Dieses Beispiel unterstreicht die Notwendigkeit den gesamten Kontext einer Aktivität zu betrachten, um fundierte Rückschlüsse auf die Verarbeitung personenbezogener Daten ziehen zu können.

2.3 Large Language Models (LLMs)

LLMs sind große, vortrainierte Sprachmodelle, die auf der Transformer-Architektur basieren. Transformer, erstmals von Vaswani et al. [71] beschrieben, verarbeiten eine Eingabe nicht strikt sequenziell, sondern beachten alle Tokens einer Sequenz parallel. Über sogenannte *Self-Attention* gewichten sie, welche Token füreinander relevant sind. Als Token gelten Wörter oder Wortbestandteile, in die der Text vorab zerlegt wird. Dieser Attention-Mechanismus erfasst Abhängigkeiten über große Distanzen innerhalb der Sequenz und ermöglicht dadurch eine effiziente Kontextmodellierung - das zentrale Prinzip moderner LLMs. Die Transformer-Architektur bildet heute das Fundament moderner Sprachmodelle wie der GPT-Familie von OpenAI [5, 41, 57].

In chatbasierten Systemen wird das Verhalten des LLM über System- und User-Prompts gesteuert. Gutes Prompt Engineering kann die Leistung und Format-Treue der Ausgabe verbessern, ohne dass die Modellparameter verändert werden müssen [38]. Ein deutlicher Vorteil aktueller LLMs ist Zero-/Few-Shot Learning. Damit lassen sich Aufgaben alleine über Instruktionen und wenige Beispiele lösen, ohne dass erneutes Training benötigt wird [8, 38]. Das ist besonders nützlich für Klassifikationsaufgaben, bei denen nur wenige gelabelte Beispiele vorliegen, wie etwa die Identifikation von DSGVO-kritischen Aktivitäten in Prozessmodellen.

Um LLMs in automatisierten Pipelines zu integrieren sind schema-konforme Aus-

gaben, wie ein gültiges JSON, unerlässlich. In der Praxis gibt es dafür drei Ansätze:

1. Klare Angaben über das Ausgabeformat im System- oder User-Prompt [38].
2. API-gestützte Mechanismen wie Function Calling oder Structured-Output/JSON-Mode mit Schemaüberprüfung [4, 53, 58].
3. Constrained Decoding, das die Generierung auf eine vorgegebene Grammatik beschränkt. Ein Beispiel ist PICARD: Bei jedem Generationsschritt des Language Model (LM) werden nur zulässige Tokens ausgewählt [67].

Typische Fehlerbilder bei der Nutzung von LLMs sind Halluzinationen (plausibel wirkende, aber fehlerhafte Aussagen) und Formatfehler (wie z.B. ungültiges JSON). In [35] wird argumentiert, dass Halluzinationen bereits beim Erstellen des LLM durch die Trainings- und Evaluationsmethoden begünstigt werden, die das Modell dazu bringen, eher zu raten als Unsicherheit zuzugeben. Das Raten bei Unsicherheit verbessert die Testergebnisse. Gegenmaßnahmen gegen Halluzinationen sind u.a. präzisere Prompts, Informationserweiterung des Prompts durch Retrieval Augmented Generation (RAG) und Self-Check/Retry-Strategien als Post-Processing Methoden nach der Generierung [34].

Die meisten großen LLMs werden von Unternehmen wie OpenAI, Google oder Anthropic entwickelt und als API-Dienste angeboten. In der Industrie zählt GPT-4o aktuell zu den am weit verbreitetsten Modellen [55]. Es ist ein multimodales Modell mit starken Text-, Bild- und Audiofähigkeiten. Proprietäre Modelle wie GPT-4o sind leistungsfähig, bringen jedoch mehrere Nachteile mit sich:

- hohe Kosten,
- mangelnde Transparenz,
- serverseitige Datenverarbeitung auf Infrastruktur der Anbieter, die sich teils außerhalb der EU befindet und wo die DSGVO nicht gilt.

Für die Verarbeitung personenbezogener Daten innerhalb der EU ist das problematisch. Eine Übermittlung in Drittländer ist nur zulässig, wenn dort der Auftragsverarbeiter sämtliche Vorgaben aus Kapitel 5 (Art. 44-50) der DSGVO einhält [16].

Als Alternative zu proprietären Modellen steht eine wachsende Zahl frei verfügbarer Open-Source-LLMs zur Verfügung, die auch lokal betrieben werden können. Prominente Beispiele sind die Modelle von Mistral [3], Deepseek [2] und Qwen [62].

Der lokale Betrieb ermöglicht volle Kontrolle darüber, wo und wie Daten verarbeitet werden. Das erleichtert die Einhaltung datenschutzrechtlicher Anforderungen. Zudem bieten Open-Source-Modelle weitere Vorteile wie geringere Kosten und hohe Anpassbarkeit. In dieser Arbeit werden sowohl proprietäre als auch Open-Source-LLMs evaluiert (siehe Kapitel 7).

2.4 Verwandte Arbeiten

- Was für Ansätze gibt es bereits Prozesse automatisiert nach datenschutzkritischen Aktivitäten klassifizieren zu lassen
- LLMs zum Klassifizieren nutzen
- Prompt Engineering (Zero-Shot)
- Überblick über LLMs in Businessprozessen. Was gibt es bereits für Ansätze diese zu benutzen
- Benchmarking und Evaluierung von LLMs
- Identifizierte Forschungslücken: LLMs zur Klassifizierung, EU-Fokus, einheitliche reproduzierbare Benchmarks

3 Problemdefinition und Zielkriterien

Dieses Kapitel präzisiert die Aufgabe der Arbeit, die Qualitätsziele der Klassifikation und steckt den fachlichen Geltungsbereich ab. Außerdem wird das Experimentdesign beschrieben, um die Forschungsfragen systematisch zu beantworten. Damit schafft es die Grundlage für die in Kapitel 4 beschriebene Klassifizierungspipeline sowie für die späteren Experimente und deren Auswertung.

Abbildung ?? zeigt einen Beispielprozess, der in den folgenden Abschnitten als durchgehende Referenz dient. Er modelliert den Versand eines Statusberichts eines Onlineshops an Kunden. Hierfür werden personenbezogene Daten in den Aktivitäten „Tracking-id generieren“ und „Status Update senden“ verarbeitet.

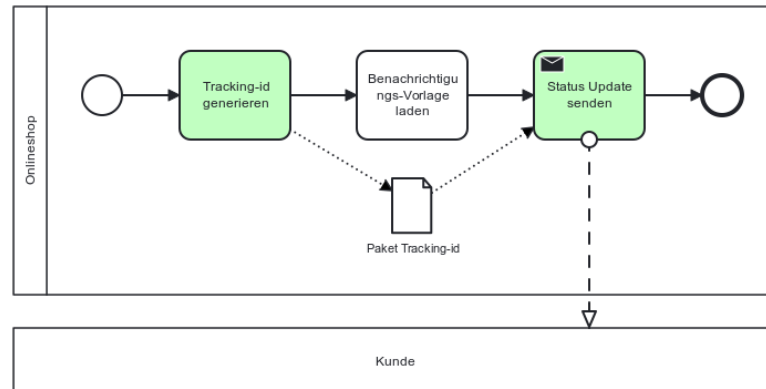


Abbildung 3.1: Beispielprozess zur Veranschaulichung der Aufgabenstellung

// TODO Abbildung im Kapitel referenzieren wo es Sinn ergibt. Aktuell ist sie noch nirgends referenziert.

3.1 Aufgabenstellung

Ziel der Arbeit ist eine *binäre Klassifikation* auf Ebene einzelner BPMN-Aktivitäten: Für jede Aktivität eines Eingabemodells im BPMN-XML-Format (Version 2.0.2) [51] soll entschieden werden, ob sie *kritisch* im Sinne des Datenschutzrechts ist oder nicht.

- **Eingabe** ist ein valides BPMN-XML mit stabilen `id`-Attributen je Aktivität [51].
- **Ausgabe** ist eine Menge von Aktivitäts-ids, die als *kritisch* klassifiziert wurden. Optional werden zusätzlich eine natürlichsprachige Begründung für einzelne Entscheidungen ausgegeben. Im Fall der Klassifizierungspipeline dieser Arbeit werden die Begründungen vom LLM generiert. Die Erklärungen dienen ausschließlich der Nachvollziehbarkeit der gewählten Klassifizierungen, werden allerdings nicht in der Evaluation berücksichtigt.

Begriffsbestimmung „kritisch“

Eine Aktivität gilt in dieser Arbeit als *kritisch*, wenn sie *personenbezogene Daten* verarbeitet. Personenbezogene Daten sind, nach Art. 4 Abs. 1 DSGVO [16], alle Informationen, die sich auf eine identifizierte oder identifizierbare natürliche Person beziehen. Gemäß Art. 4 Abs. 2 DSGVO [16] umfasst Verarbeitung jede mit personenbezogenen Daten vorgenommene Operation, wie z. B. Erheben, Speichern, Abrufen, Verwenden, Übermitteln und Löschen. Dies schließt auch die *Nutzung bereits vorhandener Daten* (z. B. Lesen/Abgleichen) ein.

Diese Aufgabenstellung reiht sich in Arbeiten zur Kennzeichnung kritischer/unkritischer Tätigkeiten in Prozessartefakten ein und bildet die Referenz für die Qualitätsziele im nächsten Abschnitt. [48]

3.2 Qualitätsziele

Die Aufgabe der datenschutzrechtlichen Klassifikation von Prozessen ist risikosensitiv. Übersehene kritische Aktivitäten, auch False Negatives (FN) genannt, bergen

erhebliche Compliance-Risiken und können zu hohen Strafen nach der DSGVO führen. Beispielsweise erhielt Meta Platforms Ireland Limited (Meta IE) 2023 aufgrund von rechtswidriger Übermittlung von EU Nutzerdaten in die USA eine Geldbuße von 1,2 Milliarden Euro [1]. Auch Amazon wurde 2025 nach einem langjährigen Rechtsstreit wegen Datenschutzverstoßen mit 746 Millionen Euro bestraft [14, 65]. Um derartige Strafen zu vermeiden, müssen kritische Aktivitäten zuverlässig identifiziert werden. Daher ist das **Hauptziel** der Klassifikation:

Maximaler Recall bei *minimalen FN* und zugleich *akzeptabler Precision*, damit der manuelle Prüfaufwand durch False Positives (FP) begrenzt bleibt.

Konfusionsmatrix und Metriken

Zur Bewertung des Hauptziels wird eine Konfusionsmatrix verwendet. Im vorliegenden binären Kontext entspricht die positive Klasse DSGVO-kritischen Aktivitäten. Die vier Felder der Konfusionsmatrix haben folgende Bedeutung [69]:

True Positives (TP) sind korrekt als kritisch erkannte Aktivitäten. Sie bilden den unmittelbaren *Nutzen* der Klassifikation.

FP sind fälschlich als kritisch markierte Aktivitäten. Sie erhöhen den manuellen Prüfaufwand, verursachen aber *keine* unmittelbaren Compliance-Risiken.

True Negatives (TN) sind korrekt als unkritisch eingestufte Aktivitäten und reduzieren den Gesamtaufwand.

FN sind übersehene kritische Aktivitäten. Sie sind besonders problematisch [48], da sie zu ausbleibender Risikobehandlung und potenziellen Bußgeldern führen können.

Aus diesen Größen leiten sich die Evaluationsmetriken ab. Relevante Metriken für eine aussagekräftige Evaluierung sind *Accuracy*, *Precision*, *Recall*, *F1-Score* sowie die Konfusionsmatrix-Zahlen (TP, FP, TN, FN) und die Anzahl korrekt/inkorrekt klassifizierter Testfälle. Technische Fehler (z. B. Parsing-Fehler oder überschrittene Token Limits) werden separat ausgewiesen.

Diese Metriken sind in Information Retrieval und Maschinellem Lernen seit langem etabliert und bilden den De-facto-Standard zur Bewertung von Klassifikatoren [39, 69, 48]. Für das hier betrachtete binäre Problem gelten:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}, \quad \text{Precision} = \frac{TP}{TP + FP},$$

$$\text{Recall} = \frac{TP}{TP + FN}, \quad \text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}.$$

Zielwerte

Vor dem Hintergrund der oben definierten Metriken und dem Hauptziel werden im Folgenden mithilfe von vergleichbarer Literatur realistische Zielkorridore abgeleitet.

Ähnliche Arbeiten, wie von Nake et al. [48], zeigen Referenzwerte von einem maximalen *Recall* $\approx 0,83$ und *F1-Score* $\approx 0,81$ bei der Identifikation DSGVO-kritischer Aufgaben in Prozessbeschreibungen. Jüngere DSGVO-nahe LLM-Studien berichten von *Precision/Recall* im hohen 0,8x bis 0,9x-Bereich [22] und F1-Scores von $\approx 0,68$ bis zu $\approx 0,79$ [68].

Basierend darauf werden folgende Zielkorridore als *pragmatische Abnahmekriterien* gesetzt:

- **Recall** soll ein Mindestniveau von $\geq 0,80$ erreichen und ein *angestrebter* Bereich ist $\geq 0,85$.
- **Precision** soll $\geq 0,75$ als Untergrenze zur Begrenzung des Prüfaufwands erreichen.
- **F1-Score** soll $\geq 0,80$ erreichen.
- **FP je Prozess** sollen im Mittel $\leq 1,5$ betragen.

Nake et al. [48] zeigen, dass selbst ein *Recall* von 0,83 für kritische Aufgaben ohne menschliche Nachkontrolle nicht ausreicht, da die Strafen für Nichteinhaltung der DSGVO sehr hoch sind. Viel mehr eignet sich ein System mit diesem Recall-Wert für *assistierte* Prüfungen, bei denen die Ergebnisse durch qualifizierte Experten validiert werden. Für ein Screening von Geschäftsprozessen, wie es in dieser Arbeit

angestrebt wird, sind die genannten Zielwerte daher als realistisch und praxisrelevant einzuschätzen.

Zusammenfassend fixieren die Zielwerte die angestrebte Performance. Im nächsten Abschnitt wird dargelegt, dass aufgrund der nicht-deterministischen Natur von LLMs die Ergebnisstabilität über wiederholte Läufe berücksichtigt werden muss.

Stabilität über Wiederholungen

Da LLMs nicht-deterministisch sind, ist das Berichten eines einzelnen Leistungswertes nicht ausreichend für den Vergleich von Modellen. Studien wie von Reimers et al. [63] zeigen, dass die Abhängigkeit vom Seed-Wert der LLMs zu statistisch signifikanten Unterschieden in der Performance führen kann. Diese Varianz kann dazu führen, dass ein modernes, leistungsfähiges Modell von sehr gut bis mittelmäßig abschneidet. Stattdessen wird vorgeschlagen, Score-Verteilungen zu vergleichen, die auf mehreren Durchläufen basieren. Dadurch wird das Risiko reduziert, dass ein Modell nur aufgrund eines günstigen Seeds gut oder aufgrund eines ungünstigen Seeds schlecht abschneidet. In dieser Arbeit werden daher die Ergebnisse auf Basis von Wiederholungen berichtet. Es wird der Mittelwert \pm Standardabweichung (σ) je Metrik angegeben, da die Standardabweichung die Stabilität eines Modells über verschiedene Läufe hinweg darstellt. Modellvergleiche basieren am Ende auf diesen Verteilungen und nicht auf Einzelfällen, um eine fundierte Bewertung zu ermöglichen.

3.3 Scope und Annahmen

Dieser Abschnitt definiert Geltungsbereich, Annahmen und Risiken des Ansatzes. Dadurch wird eine klare Einordnung der Ergebnisse und ihrer Reproduzierbarkeit ermöglicht.

Geltungsbereich

Die folgenden Punkte definieren den Geltungsbereich der Arbeit:

- Klassifiziert werden ausschließlich Aktivitäten. Dafür wird sinnvoller Kontext (z. B. Prozessname, Pool/Lane, Datenobjekte) berücksichtigt.
- Labels und Artefakte liegen in Deutsch und Englisch vor.
- Es handelt sich um ein *Screening*, nicht um eine Rechtsprüfung. Kritisch klassifizierte Aktivitäten sind anschließend juristisch zu prüfen.

Annahmen und Risiken

Die folgenden Annahmen und potenziellen Risiken sind für die Interpretation der Ergebnisse relevant:

- Bei fehlenden Datenobjekten oder mehrdeutigen Labels kann sich die Einschätzung verschlechtern. Das ist ein bekanntes Problem in ähnlichen Studien [48].
- Optional generierte LLM-Begründungen sind als *Hilfetexte* zu verstehen, um die Entscheidung des LLMs besser einordnen zu können, bilden aber nicht zwingend die tatsächlichen Entscheidungsgründe des Modells ab.
- Ungültiges BPMN-XML oder Laufzeitfehler werden als „technischer Fehler“ erfasst und nicht in die Metrikzählung eingerechnet. Sie werden separat berichtet.

3.4 Experimentdesign

Das gesamte Kapitel definierte die binäre Klassifikation von BPMN-Aktivitäten als kritisch/unkritisch mit Fokus auf maximalen Recall bei akzeptabler Precision und legte Qualitätsziele, Metriken, Geltungsbereich sowie Annahmen fest. Darauf aufbauend beschreibt dieser Abschnitt das Experimentdesign, mit dem LLMs fair und reproduzierbar verglichen werden, um die Forschungsfrage **FF1** sowie die Unterfragen **UF1–UF4** zu beantworten. Die konkrete Ausgestaltung und Durchführung der Experimente werden in Kapitel 8 *Versuchsaufbau* erläutert. Im Folgenden werden die wesentlichen Aspekte des Experimentdesigns beschrieben:

Ziel Ziel ist ein transparenter Vergleich mehrerer LLMs, die alle dieselbe Klassifizierungspipeline durchlaufen. Sie wird in Kapitel 4 daher so entworfen, dass sich das LLM austauschen lässt. Die Auswahl der im Evaluationsframework aus Kapitel 6 zu nutzenden LLMs erfolgt zur Laufzeit anhand übergebener Identifikationsparameter (z.,B. Modellname, Basis-URL/Endpunkt).

Vergleichsgegenstand Die Experimente werden über eine deklarative Konfiguration definiert, siehe Kapitel 6.3. Sie legt fest, welche Modelle, Datensätze und weitere Parameter zum Einsatz kommen. Je nach Auswahl werden mehrere Modelle und Modellvarianten parallel im Evaluationsframework ausgeführt, darunter Open-Source und kommerzielle Modelle. Die deklarative Konfiguration sorgt für Portabilität und Wiederholbarkeit.

Datenbasis Als Datenbasis dienen die im Labeling-Tool erzeugten, gelabelten Testdatensätze, siehe Kapitel 5. Ein Testdatensatz enthält mehrere gelabelte Testfälle. Ein Testfall umfasst ein BPMN-Prozessmodell mit Labeln, die Aktivitäten als DSGVO-kritisch markieren. Die Auswahl der Datensätze für ein Experiment erfolgt in der Evaluierungskonfiguration und das Laden der Testfälle während der Laufzeit. Die Datensätze sollten idealerweise unterschiedliche Eigenschaften abdecken, damit die Forschungsfrage und die Unterfragen möglichst umfassend beantwortet werden. Unterschiede können sich etwa in der Domäne, der Größe der Prozesse, den eingesetzten Sprachen oder den verwendeten BPMN-Elementen zeigen.

Metriken und Erfolgskriterium Ausgewertet werden die in Abschnitt 3.2 beschriebenen Metriken: Accuracy, Precision, Recall und F1. Zusätzlich werden die Kennzahlen der Konfusionsmatrix betrachtet: TP, FP, TN, FN. Ein Testfall gilt als *bestanden*, wenn die vom Modell als kritisch ausgegebenen Aktivitäten exakt den gelabelten kritischen Aktivitäten entsprechen. Technische Fehler werden separat ausgewiesen.

Ablauf eines Experiments

Ein Experiment verläuft in folgenden Schritten:

1. **Konfiguration laden.** Die Konfiguration mit Modellen, Datensätzen und optionalem seed wird geladen.

2. **Ausführung.** Für jedes Modell werden alle ausgewählten Testfälle durch die Klassifizierungspipeline verarbeitet. Pro Testfall werden TP, FP, FN, TN sowie der Status „bestanden“ oder „nicht bestanden“ berechnet.
3. **Stabilität.** Die Läufe erfolgen mit temperature gleich 0¹ und festem seed, sofern das jeweilige LLM dies unterstützt. Um die Nicht-Deterministik moderner LLMs abzubilden, werden die Experimente mehrfach mit unterschiedlichen Seeds wiederholt. Die Ergebnisse werden über die Läufe gemittelt.
4. **Bericht.** Aggregierte Kennzahlen pro Modell, wie Konfusionsmatrix, die genannten Metriken sowie die Bestehensraten werden ausgegeben. Metadaten wie verwendete Modelle, Datensätze und Seeds werden dokumentiert.

Dieses Kapitel definiert, *was* verglichen wird: Modelle, Datensätze und Metriken. Es beschreibt zudem, *wie* der Vergleich erfolgt. Kapitel ?? dokumentiert später die praktische Umsetzung mit konkreten Modellen, exakten Parameterwerten, Seeds sowie den vollständigen genutzten Konfigurationen. Im nächsten Kapitel folgt das Design und die Implementierung der Klassifizierungspipeline, die für den Vergleich der LLMs verwendet wird.

¹Die temperature steuert die Zufälligkeit der Textgenerierung bei LLMs. Niedrige Werte liefern stabilere Antworten, hohe Werte vielfältigere, jedoch weniger verlässliche [50].

4 Design und Implementierung der Klassifizierungspipeline

Dieses Kapitel beschreibt die Pipeline zur Klassifikation DSGVO-kritischer Aktivitäten in BPMN-Prozessen. Ausgehend von der in Kapitel 3.1 formulierten Aufgabenstellung wird der gesamte Weg von der Eingabe eines *BPMN-XML* über die Vorverarbeitung, das Prompt Engineering bis hin zur strukturierten, schema-konformen Ausgabe aufgezeigt. Außerdem wird ein HTTP-basiertes API-Design vorgestellt, das die Einbindung in weitere Werkzeuge und das Evaluationsframework ermöglicht. Der Prozessfluss der Klassifizierungspipeline ist in Abbildung 4.1 dargestellt und wird in diesem Kapitel im Detail erläutert.

Die Klassifizierungspipeline soll eine binäre Entscheidung auf Ebene einzelner BPMN-Aktivitäten treffen: Für jede Aktivität eines Eingabemodells wird bestimmt, ob sie *kritisch* im Sinne der DSGVO ist. Die Pipeline ist so konzipiert, dass sie mit Modellen aus gängigen Modellierungswerkzeugen kompatibel ist. Dadurch kann sie in bestehende Modellierungswerkzeuge wie Camunda Modeler [20] integriert werden, um einen praktischen Einsatz in realen Prozessmodellierungs-Workflows zu ermöglichen.

4.1 BPMN Preprocessing

Ziel der Vorverarbeitung (Preprocessing) ist es, für jedes Flow-Element einen *strukturierten Kontext* zu erzeugen. Dieser Kontext umfasst die eigenen Attribute, wie *id*, *name* und *documentation*, sowie die Beziehungen zu anderen Elementen im BPMN-Diagramm. Dazu gehören vorangehende und nachfolgende Flow-Elemente, Datenobjekte, assoziierte Elemente, sowie Informationen über den Pool und die

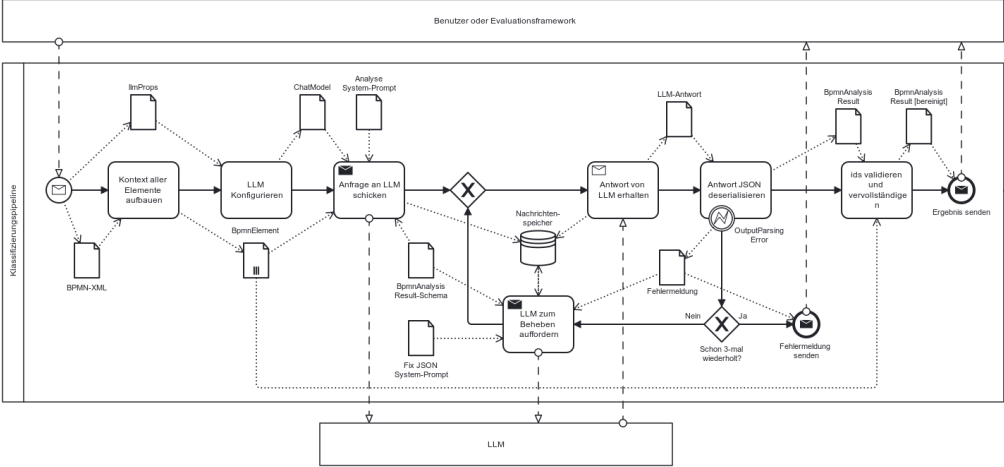


Abbildung 4.1: BPMN-Diagramm der Klassifizierungspipeline.

Lane, in denen sich das Element befindet. Das Parsen des BPMN-XML erfolgt mit der *Camunda BPMN Model API*, die das XML in ein Objektmodell überführt [10, 11]. Auf dieser Basis werden die relevanten Informationen extrahiert und in der Datenklasse `BpmnElement` strukturiert abgelegt. Die Datenklasse ist in Listing 4.1 zu sehen. Dadurch entsteht für jedes Flow-Element ein umfassender Kontext, der später im Prompt genutzt wird, um dem LLM alle notwendigen Informationen strukturiert bereitzustellen. Außerdem werden durch das Format Tokens eingespart, da irrelevante Informationen, wie die Positionen der Elemente im XML, weggelassen werden. In Abbildung 4.1 ist dieser Schritt über die Aktivität „Kontext aller Elemente aufbauen“ dargestellt.

Listing 4.1: Interne BPMN-Repräsentation je Flow-Element.

```
1 data class BpmnElement(  
2     val type: String,  
3     val id: String,  
4     val name: String? = null,  
5     val documentation: String? = null,  
6     val poolName: String? = null,  
7     val laneName: String? = null,  
8     val outgoingFlowElementIds: List<String> = emptyList(),  
9     val incomingFlowElementIds: List<String> = emptyList(),  
10    val outgoingMessageFlowsToElementIds: List<String> = emptyList(),  
11    val incomingMessageFlowsFromElementIds: List<String> = emptyList(),
```

```
12     val incomingDataFromElementIds: List<String> = emptyList(),  
13     val outgoingDataToElementIds: List<String> = emptyList(),  
14     val associatedElementIds: List<String> = emptyList()  
15 )
```

4.2 Prompt Engineering

Eine robuste Klassifikation hängt maßgeblich von sorgfältig gestalteten Prompts ab. Ziel ist es, das LLM mit klaren Anweisungen, einem konsistenten Bewertungsschema und präzisen Formatvorgaben so zu steuern, dass es die Klassifizierung zuverlässig löst und strukturierte Ausgaben liefert. Im Folgenden werden zunächst die deklarative Orchestrierung der Kommunikation mit dem LLM mithilfe von LangChain4j und anschließend die Prompt-Konzeption beschrieben.

LangChain4j: deklarative Orchestrierung

Zur Reduktion von Boilerplate und für konsistente Prompts wird *LangChain4j* [37] benutzt. Mit den *AI Services* werden Interaktionen mit dem LLM als Java/Kotlin-Interface *deklarativ* beschrieben. Zur Laufzeit erzeugt LangChain4j einen Proxy, der den System-Prompt injiziert, den User-Prompt aus den Methodenparametern generiert und die LLM-Antwort in den passenden Rückgabetypp deserialisiert [61]. Beim erstellen des AI Service werden ein `ChatModel`, die Systemnachricht und die Interface-Methoden konfiguriert. Ein `ChatModel` ist die spezifische Implementierung der Chat-Completion-Schnittstelle eines LLM von Langchain4j [36]. Die Methodenparameter der Interface-Methoden repräsentieren die Nutzereingabe. Der Rückgabetypp der Methoden definiert die erwartete Antwortstruktur des LLM. Optional kann jeder Interface-Methode noch ein eigener User-Prompt zugewiesen werden, der bei Laufzeit mit den übergebenen Parametern gefüllt wird [61].

Die Kommunikation mit dem LLM erfolgt damit über einfache Funktionsaufrufe, während LangChain4j Prompt-Erzeugung, Parameterbindung sowie die Deserialisierung der Antwort übernimmt [61]. So kann im Code ohne zusätzlichen Aufwand direkt mit typisierten Objekten gearbeitet werden.

Für die Klassifikation *DSGVO-kritisch* vs. *unkritisch* wird ein **Zero-Shot**-Ansatz verwendet. Das LLM erhält im System-Prompt eine präzise Instruktion mit Kriterien und illustrativen Beispielfällen, was als kritisch gilt. Es sind jedoch keine Beispiele mit konkreten Ein- und Ausgabe-paaren pro Prozess enthalten. Zero-Shot reduziert den Pflegeaufwand und nutzt die In-Context-Fähigkeiten moderner Modelle, nur über Instruktionen zu generalisieren [8, 38]. Wie genau die Prompts aufgebaut sind, wird im Folgenden beschrieben.

System-Prompt

Der System-Prompt definiert das Verhalten des LLM, zusätzlichen Kontext und das gewünschte Ausgabeformat. Der vollständige System-Prompt befindet sich im Anhang, siehe A.1. Im Kern legt der genutzte System-Prompt Folgendes fest:

1. **Rolle und Auftrag des Modells.** Das Modell agiert als Experte für das Analysieren von BPMN-Modellen auf DSGVO-konformität und prüft sämtliche Aktivitäten eines Prozesses auf Datenschutzrelevanz. Jede Aktivität wird berücksichtigt und die Entscheidung erfolgt auf Basis sämtlicher verfügbarer Kontextinformationen wie Name, Beschreibung, Annotationen sowie Daten- und Nachrichtenassoziationen.
2. **Rechtliche Definitionen nach DSGVO.** Der System-Prompt erläutert die Begriffe „personenbezogene Daten“ und „Verarbeitung“ gemäß Art. 4 DSGVO. Beispiele für personenbezogene Daten umfassen Identifikatoren, Kontakt- und Zahlungsdaten, Beschäftigungsdaten, Gesundheitsdaten, biometrische Merkmale, Standortinformationen und Online-Kennungen. Verarbeitung umfasst Erheben, Speichern, Abrufen, Verwenden, Übermitteln, Ausrichten, Kombinieren, Einschränken, Löschen und Vernichten.
3. **Indikatoren für Kritikalität.** Der System-Prompt enthält typische Auslöser für Datenschutzrelevanz wie Datenerfassung und Dateneingabe, Anlage und Aktualisierung von Datensätzen, Übermittlung oder Offenlegung an andere Systeme oder Dritte, Zahlungen und Finanztransaktionen und noch mehr. Diese Indikatoren sind mit Beispielen angereichert und dienen als *Entscheidungshelfer* für das Modell.

4. **Abgrenzung durch Negativbeispiele.** Der System-Prompt grenzt unkritische Fälle klar ab. Rein administrative oder logistische Schritte ohne Personenbezug werden nicht als kritisch gewertet. Ebenso gilt dies für Fälle in denen anonymisierte Daten verwendet werden und keine Identifikation einer Person mehr möglich ist.
5. **Erwartetes Ausgabeformat.** Die Antwort erfolgt als strukturierte JSON-Ausgabe mit einer Liste relevanter Aktivitäten. Für jede Aktivität wird die `id` und eine Begründung in natürlicher Sprache ausgegeben. Es werden ausschließlich Aktivitäten zurückgegeben, die nach den Kriterien als datenschutzrelevant eingestuft wurden.

Die Kombination dieser Elemente im System-Prompt stellt sicher, dass das LLM die Aufgabe versteht, die relevanten Kriterien kennt und die Ausgabe in einem maschinenlesbaren Format liefert. So entsteht die Basis für eine zuverlässige Klassifikation. Zu einer Anfrage an ein LLM gehört außerdem stets ein User-Prompt, der die eigentliche Nutzereingabe enthält. Dessen Aufbau wird im nächsten Abschnitt beschrieben.

User-Prompt

Der User-Prompt übergibt dem LLM die konkreten Eingabedaten einer Anfrage. Während der System-Prompt Regeln, Ziele und Ausgabevorgaben festlegt, liefert der User-Prompt die Fall- bzw. Kontextinformationen, auf die diese Regeln angewendet werden.

Der User-Prompt wird mithilfe der Daten aus der Vorverarbeitung aus Abschnitt 4.1 erzeugt und enthält eine Liste von `BpmnElement`-Objekten, siehe Listing 4.1. Die Interaktion mit dem LLM erfolgt deklarativ über *LangChain4j*. Dafür wird die Liste der `BpmnElement`-Objekte als Methodenparameter mit der Annotation `@UserMessage` an die Interface-Methode übergeben und dort automatisch in den User-Prompt eingebettet.

Zur Laufzeit serialisiert *LangChain4j* die `BpmnElement`-Liste zu einem JSON-Array und stellt sie als User-Prompt bereit. Der zuvor konfigurierte System-Prompt wird bei einer Anfrage an das LLM automatisch dem User-Prompt vorangestellt. Auf diese Weise wendet das LLM die im System-Prompt definierten Kriterien auf die

im User-Prompt gelieferten Informationen zum BPMN-Prozessmodell an. Dadurch wird jede Aktivität des Prozesses genau so wie im System-Prompt beschrieben klassifiziert. In Abbildung 4.1 findet dieser Schritt in der Aktivität „Anfrage an das LLM schicken“ statt.

Zusammenfassend setzt der System-Prompt typischerweise das Regelwerk, und der User-Prompt liefert die konkreten Eingabedaten. Besonders in mehrstufigen Dialogen mit dem LLM spielt dieses Muster eine größere Rolle, da der System-Prompt konstant bleibt, während der User-Prompt je nach Anfrage variiert. Im vorliegenden Szenario, wo immer nur genau eine Anfrage pro Prozessmodell gestellt wird fällt der Unterschied weniger ins Gewicht, als würden sämtliche Vorgaben direkt im User-Prompt stehen. Die Trennung erhöht dennoch die Nachvollziehbarkeit, sorgt für klare Rollen und erleichtert die Wiederverwendung.

Im folgenden Abschnitt wird beschrieben, wie auf dieser Basis strukturierte Ausgaben erzeugt werden, damit im Code direkt mit typisierten Objekten weitergearbeitet werden kann.

Strukturierte Ausgaben mit LangChain4j

Wie in Listing ?? zusehen, wird im Fall der Klassifikation ein `BpmnAnalysisResult` als Antwort erwartet, also eine Liste von Elementen mit Paaren aus `id`, `reason` und `isRelevant`. Siehe A.2 für die vollständige Definition der Datenklasse. Langchain4j inferiert auf Basis des Rückgabetyps der Interface-Methode ein JSON-Schema und fügt dieses automatisch dem User-Prompt zusammen mit der Aufforderung hinzu, die Antwort in diesem JSON-Format zu liefern [61]. Durch die explizite Angabe des gewünschten JSON-Formats im Prompt wird die Format-Treue der Antwort erhöht [38], also die Wahrscheinlichkeit, dass die Antwort tatsächlich dem gewünschten Schema entspricht.

Einige LLMs unterstützen darüber hinaus die Möglichkeit, das Antwortformat API-seitig zu erzwingen. Das ist beispielsweise bei Mistral und OpenAI der Fall [4, 58]. Falls das LLM die `response_format` Funktionalität unterstützt setzt LangChain4j dies zusätzlich auf das gewünschte Schema und erzwingt so das Ziel-JSON API-seitig [61]. Fehlt diese Fähigkeit, greift ausschließlich die Prompt-basierte Schemaanweisung.

Das vom LLM gelieferte JSON deserialisiert *LangChain4j* anschließend automatisch zu einem *BpmnAnalysisResult*. So kann im Code direkt mit einem typischen Objekt weitergearbeitet werden. In Abbildung 4.1 ist dieser Prozess über die Aktivitäten „Antwort von LLM erhalten“ und „Antwort JSON deserialisieren“ dargestellt.

4.3 Validierung der Ausgabe

Zusätzlich zu den in Kapitel 4.2 beschriebenen Maßnahmen stellt die Pipeline mehrere Validierungs- und Korrekturschritte bereit, die in Abbildung 4.1 direkt auf „Antwort JSON deserialisieren“ folgen. Diese Schritte dienen dazu, die Qualität und Korrektheit der Ausgabe des LLM zu gewährleisten. Im Folgenden werden die einzelnen Validierungsmechanismen erläutert.

Schema-Parsing und Retry-Mechanismus

Die vom LLM zurückgelieferte Antwort wird zunächst von *Langchain4j* zu einem *BpmnAnalysisResult* deserialisiert. Entspricht die Struktur dabei nicht dem erwarteten JSON-Schema, löst *Langchain4j* eine *OutputParsingException* aus. In diesem Fall greift der in Abbildung 4.1 ab dem Boundary-Error-Event „Output-ParsingError“ dargestellte Retry-Mechanismus. Dabei wird bis zu dreimal die ursprüngliche Anfrage erneut gesendet, ergänzt um die Parser-Fehlermeldung sowie eine explizite Anweisung, die Ausgabe exakt gemäß Schema zu formatieren. So bleiben sowohl der Kontext der ursprünglichen Anfrage als auch die Information über den aufgetretenen Fehler erhalten, damit das LLM die Ausgabe entsprechend anpassen kann. Schlagen alle drei Versuche fehl, wird der Fehler an die aufrufende Schnittstelle zurückgegeben und die Klassifizierung gilt als fehlgeschlagen.

Ein zusammenfassender Log-Auszug des Retry-Mechanismus findet sich in Listing A.5. Er zeigt exemplarisch, dass zunächst der boolesche Wert *isRelevant* fehlt und im zweiten Versuch korrekt ergänzt wird.

Relevanz-Filterung

Nach erfolgreichem Parsing werden alle Elemente mit `isRelevant = false` entfernt. Dieser Schritt geschieht bereits im Konstruktor der Datenklasse `BpmnAnalysisResult` automatisch. Dieser Mechanismus adressiert modellseitige *Überklassifizierungen*, bei denen das LLM fälschlicherweise ids von Aktivitäten ausgibt, obwohl sie nicht DSGVO-kritisch sind. Es wird sichergestellt, dass nur Aktivitäten, die als kritisch klassifiziert wurden, in der finalen Ausgabe verbleiben.

Ohne das `isRelevant`-Flag hat das LLM in der Praxis des Öfteren Aktivitäten als kritisch ausgegeben, deren Begründung jedoch ausdrücklich darlegte, *warum* sie *nicht* kritisch seien. Das Modell erkannte die Unkritikalität also korrekt, hielt sich aber nicht strikt an die Vorgabe, ausschließlich ids kritischer Aktivitäten in die Antwort aufzunehmen. Als pragmatische Absicherung wurde daher das boolesche `isRelevant`-Flag eingeführt. Das LLM muss zusätzlich neben der Ausgabe der ids auch explizit angeben, ob die jeweilige Aktivität kritisch ist oder nicht. In der Summe reduziert diese Filterung die Anzahl widersprüchlicher Ausgaben.

`isRelevant` dient ausschließlich einer internen Validierung und wird in der finalen Ausgabe der Klassifizierungs-Pipeline nicht berücksichtigt.

id-Validierung und -Vervollständigung

In der Praxis liefert das LLM mitunter unvollständige oder fehlerhafte id-Werte, die im Prozess nicht existieren. Zur Erhöhung der Robustheit werden die vom LLM ausgegebenen ids daher gegen die tatsächlich im Prozess vorhandenen Aktivitäts-ids geprüft und – wenn möglich – automatisch vervollständigt. Der Ablauf ist:

1. Ermittlung der Grundmenge aller gültigen Aktivitäts-ids aus der `BpmnElement`-Liste, die beim Preprocessing erstellt wurde.
2. Für jede vom LLM gelieferte id wird ein Präfix-Match gegen die gültigen ids durchgeführt. Ist die ausgegebene id Präfix *genau einer* gültigen id, wird sie durch diese vollständige id ersetzt.
3. Bleibt das Präfix-Match ohne eindeutiges Ergebnis, folgt ein Substring-Match: Ist die ausgegebene id Teilstring *genau einer* gültigen id, wird entsprechend vervollständigt.

4. Liefert weder Präfix- noch Substring-Match eine eindeutige Übereinstimmung, gilt die ausgegebene `id` als ungültig und wird aus der finalen Ausgabe entfernt.
5. Abschließend werden Duplikate entfernt, sodass jede kritische Aktivität höchstens einmal in der Ausgabe erscheint.

Gibt das LLM beispielsweise die `id Activity_1` aus, existiert im Prozess jedoch nur `Activity_12345`, wird die Ausgabe automatisch auf die korrekte `id` vervollständigt. Existieren hingegen sowohl `Activity_123` als auch `Activity_124` im Prozess, bleibt die Ausgabe unvollständig und wird entfernt, da keine eindeutige Zuordnung möglich ist.

Dieser Schritt fängt typische LLM-Ausgabefehler ab – etwa Halluzinationen oder abgeschnittene Bezeichner – und stellt die Konsistenz mit dem Eingabemodell sicher. Im Diagramm 4.1 ist er als „ids validieren und vervollständigen“ markiert. Ein fokussierter Code-Auszug findet sich in Listing A.3.

Nach der Validierung besteht `BpmnAnalysisResult` nur noch aus Aktivitäten, die vom LLM als kritisch eingestuft wurden (`isRelevant = true`) und deren `ids` im Prozess existieren.

Da es nun möglich ist, BPMN-Prozesse vorzuverarbeiten, zu klassifizieren und die Ausgabe zu validieren und zu beheben, folgt als nächster Schritt die Definition einer Schnittstelle zum Aufruf der Pipeline. Das nächste Kapitel beschreibt dafür das API-Design.

4.4 API-Design

// TODO Zu dem `AnalysisResponse` Schema noch `amountOfRetries` als optionale Zahl hinzufügen, die angibt, wie oft der LLM-Aufruf wiederholt wurde, falls ein Fehler aufgetreten ist. Die wird dann im Evaluationsframework ausgewertet, um zu sehen, ob manche Modelle öfter fehlschlagen.

Dieses Kapitel beschreibt das API-Design der Klassifizierungspipeline, die zur Erkennung DSGVO-kritischer Elemente in BPMN-Modellen dient. Das Ziel ist es eine standardisierte Schnittstelle zu definieren, die (1) die Einbindung in bestehen-

de Werkzeuge und das Evaluationsframework vereinfacht, (2) die Austauschbarkeit unterschiedlicher Klassifizierungsalgorithmen - insbesondere im Evaluationsframework - ermöglicht, um verschiedene Ansätze der Klassifizierung vergleichen zu können, und (3) Erweiterbarkeit fördert, sodass zukünftige Arbeiten die Schnittstelle wiederverwenden können, um ihre eigenen Klassifizierungsalgorithmen zu integrieren.

HTTP-Endpunkt

Die Klassifizierungspipeline ist über einen standardisierten HTTP-Endpunkt nutzbar, dessen Struktur und die klar definierten JSON-Schemas eine einfache Integration in bestehende Werkzeuge sowie das Evaluationsframework ermöglichen. Der POST-Endpunkt akzeptiert `multipart/form-data` mit den folgenden Teilen:

bpmnFile (Pflicht) Eine BPMN-2.0-XML-Datei (`.bpmn` oder `text/xml`), die den zu analysierenden Prozess beinhaltet.

llmProps (Optional) Ein JSON-Objekt zur Überschreibung von LLM-Eigenschaften zur Laufzeit. Siehe Listing 4.2 für das JSON-Schema. Wird nichts angegeben, nutzt die Pipeline Standardwerte.

Listing 4.2: JSON-Schema der `llmProps`.

```
1 {
2   "$schema": "https://json-schema.org/draft/2020-12/schema",
3   "title": "LlmProps",
4   "type": "object",
5   "properties": {
6     "baseUrl": { "type": "string" },
7     "modelName": { "type": "string" },
8     "apiKey": { "type": "string" },
9     "timeoutSeconds": { "type": "number" },
10    "seed": { "type": "number" },
11    "temperature": { "type": "number" },
12    "topP": { "type": "number" }
13  },
14  "required": []
```

```
15 }
```

Die `llmProps` erlauben das Überschreiben von LLM-Eigenschaften zur Laufzeit. Dadurch können unterschiedliche Modelle mit demselben Klassifizierungsalgorithmus flexibel getestet und verglichen werden, ohne die Anwendung neu starten zu müssen. Dieses Design wurde gewählt, um die Experimente wie in Kapitel 3.4 beschrieben flexibel durchführen zu können.

Die Antwort des Endpunkts wird als `application/json` geliefert und enthält eine Liste der als DSGVO-kritisch klassifizierten Elemente, einschließlich einer optionalen Begründung für jede Klassifikation und einem optionalen Namen des Elements für bessere Lesbarkeit. Das JSON-Schema der API-Antwort ist in Listing 4.3 dargestellt.

Listing 4.3: JSON-Schema der API-Antwort.

```
1 {
2   "$schema": "https://json-schema.org/draft/2020-12/schema",
3   "title": "BpmnAnalysisResult",
4   "type": "object",
5   "properties": {
6     "criticalElements": {
7       "type": "array",
8       "items": {
9         "type": "object",
10        "properties": {
11          "id": { "type": "string" },
12          "name": { "type": "string" },
13          "reason": { "type": "string" }
14        },
15        "required": ["id"]
16      }
17    },
18  },
19  "required": ["criticalElements"]
20 }
```

Im nächsten Kapitel wird die Webapp-Sandbox beschrieben, die als Beispielanwen-

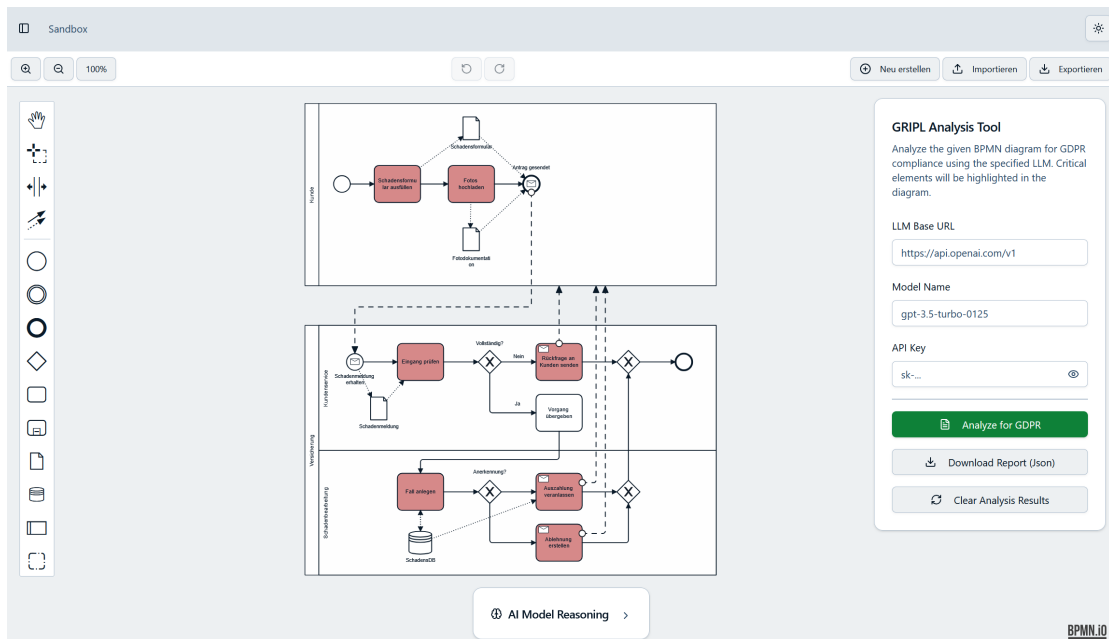


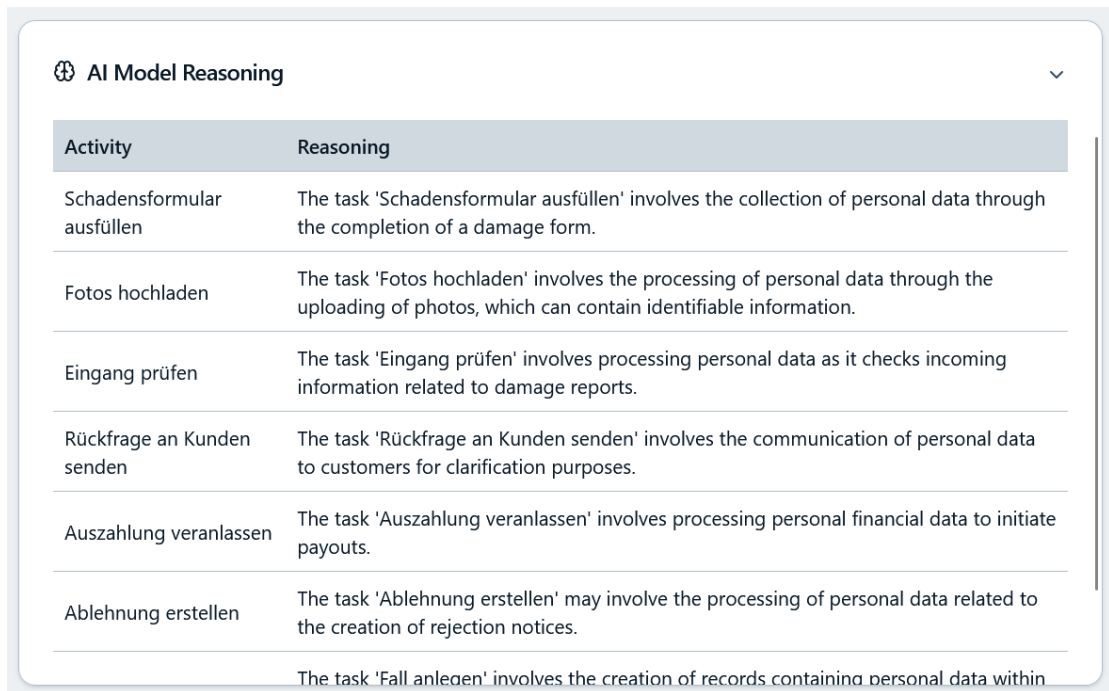
Abbildung 4.2: Sandbox im Frontend mit hervorgehobenen kritischen Aktivitäten nach Analyse.

derung dient, um die Klassifizierungspipeline intuitiv nutzen zu können. Sie verwendet das hier beschriebene API, um die Klassifizierung durchzuführen.

4.5 Webapp-Sandbox

Zur interaktiven Nutzung der Klassifizierung wurde eine *Sandbox* in Form einer Webapp entwickelt. Sie verbindet einen vollwertigen BPMN-Editor auf Basis von `BPMN.js` [18] mit der in Kapitel 4.4 beschriebenen HTTP-Schnittstelle und macht die Analyse damit intuitiv bedienbar. In der Sandbox können BPMN-Modelle erstellt, verändert, exportiert und importiert sowie auf Datenschutzrelevanz analysiert werden. Als kritisch klassifizierte Aktivitäten werden nach der Analyse direkt im Editor farblich hervorgehoben, wie in Abbildung 4.2 zu sehen ist.

Außerdem können die vom LLM generierten Begründungen zu jeder als kritisch erkannten Aktivität im Editor eingesehen werden. Diese Erläuterungen werden gesammelt in einer aufklappbaren Karte im unteren Bereich des Editors angezeigt, siehe Abbildung 4.3.



Activity	Reasoning
Schadensformular ausfüllen	The task 'Schadensformular ausfüllen' involves the collection of personal data through the completion of a damage form.
Fotos hochladen	The task 'Fotos hochladen' involves the processing of personal data through the uploading of photos, which can contain identifiable information.
Eingang prüfen	The task 'Eingang prüfen' involves processing personal data as it checks incoming information related to damage reports.
Rückfrage an Kunden senden	The task 'Rückfrage an Kunden senden' involves the communication of personal data to customers for clarification purposes.
Auszahlung veranlassen	The task 'Auszahlung veranlassen' involves processing personal financial data to initiate payouts.
Ablehnung erstellen	The task 'Ablehnung erstellen' may involve the processing of personal data related to the creation of rejection notices.
The task 'Fall anlegen' involves the creation of records containing personal data within	

Abbildung 4.3: Exemplarische Begründungen der Klassifikation durch das LLM in der Sandbox.

// TODO Die Sprache der Begründungen des LLM irgendwo thematisieren oder noch anpassen. Aktuell ist die Sprache der Begründungen immer Englisch. Vielleicht ändere ich das auf Deutsch ab, da die Mastrarbeit auf Deutsch ist oder ich passe den Code so an, dass die Begründung in der gleichen Sprache wie das Modell ist.

Um verschiedene LLMs vergleichen zu können, verfügt die Sandbox auf der rechten Seite über ein Einstellungsmenü mit konfigurierbaren LLM-Parametern (siehe Abbildung 4.2). Diese Parameter sind identisch zu den in Kapitel 4.4 beschriebenen `LlmProps` und werden beim Starten der Analyse in die API-Anfrage überführt.

5 Labeling und Datensätze

Für die Evaluation der Klassifikation ist es nötig, zuvor entsprechende Testdatensätze mit Annotationen bereitzustellen. Ein solcher Datensatz besteht dabei aus mehreren Testfällen, wobei jeder Testfall ein BPMN-Prozessmodell darstellt. Standardisierte Datensätze gewährleisten einheitliche Prüfbedingungen und ermöglichen so objektive Leistungsvergleiche.

5.1 Labeling Tool

Um die Erstellung und Verwaltung von gelabelten BPMN-Prozessmodellen zu erleichtern, wurde eine Webapplikation entwickelt. Mit dieser können BPMN-Testfälle erstellt, bearbeitet und Aktivitäten mit Labels versehen werden. Wichtige Funktionen des Labeling-Tools umfassen:

- Anlegen und Verwalten von Datensätzen.
- Erstellung beliebig vieler Testfälle pro Datensatz.
- Direkte Bearbeitung von BPMN-Modellen im Browser mittels BPMN.io [19].
- Labeling-Modus, in dem Aktivitäten als DSGVO-kritisch markiert werden können. Optional kann eine Begründung für die Markierung angegeben werden.
- Persistente Speicherung der annotierten Testfälle in einer Datenbank für die spätere Nutzung im Evaluationsframework (siehe Kapitel 6).

Abbildung 5.1 zeigt den Labeling-Editor: Hier können Anwender Prozessmodelle erstellen und Aktivitäten im Labeling-Modus direkt als kritisch markieren. Optional kann für jede markierte Aktivität eine Begründung eingegeben werden. Diese

5 Labeling und Datensätze

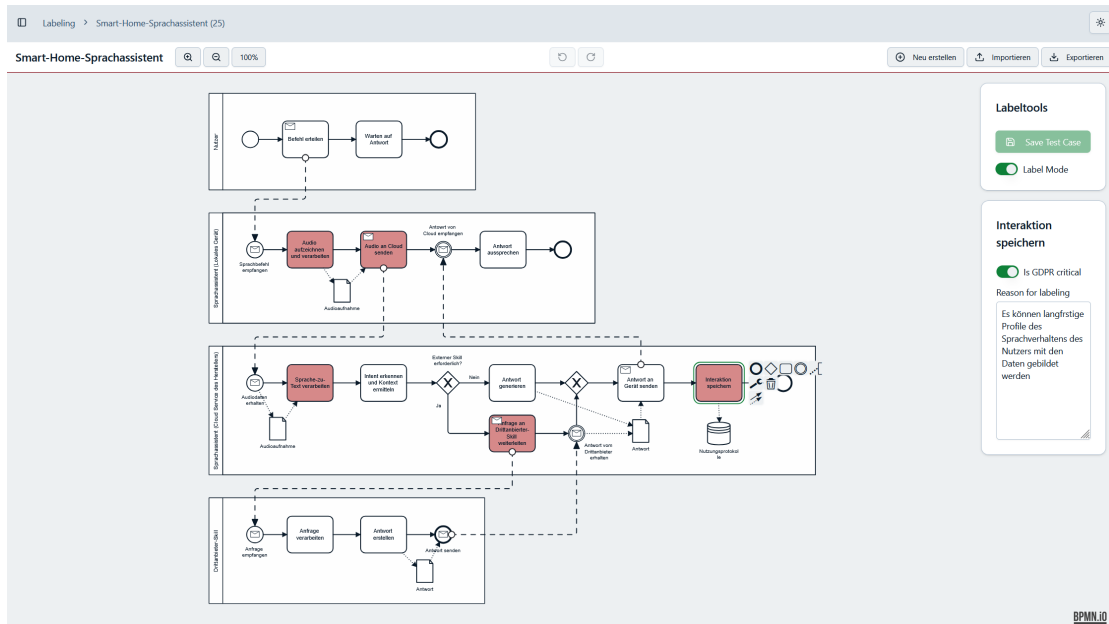


Abbildung 5.1: Labeling-Editor im Labeling-Modus mit exemplarischem Modell.

Begründung ist ausschließlich als Hilfe für den Anwender gedacht, um die eigene Entscheidung zu dokumentieren. Die Begründung wird nicht in der Evaluierung berücksichtigt.

In der Übersicht der Datensätze aus Abbildung 5.2 sind alle angelegten Datensätze und zugehörigen Testfälle aufgelistet. Von hier aus können neue Datensätze und Testfälle erstellt sowie bestehende bearbeitet werden.

5.2 Quellen und Eigenschaften der Datensätze

Für die Evaluation wurden drei Gruppen von BPMN-Datensätzen eingesetzt:

1. Prozesse, die von der Universität Ulm bereitgestellt wurden (z.B. Lehrbeispiele aus Übungsaufgaben).
2. Realistische, mittelgroße Szenarien aus verschiedenen Domänen. Diese Prozesse beinhalten Elemente wie Pools, Lanes, Datenobjekte und Gateways.
3. Kleine, reduzierte Testfälle mit maximal fünf Aktivitäten und wenigen weiteren Elementen (z.B. einfacher Sequenzfluss ohne Pools).

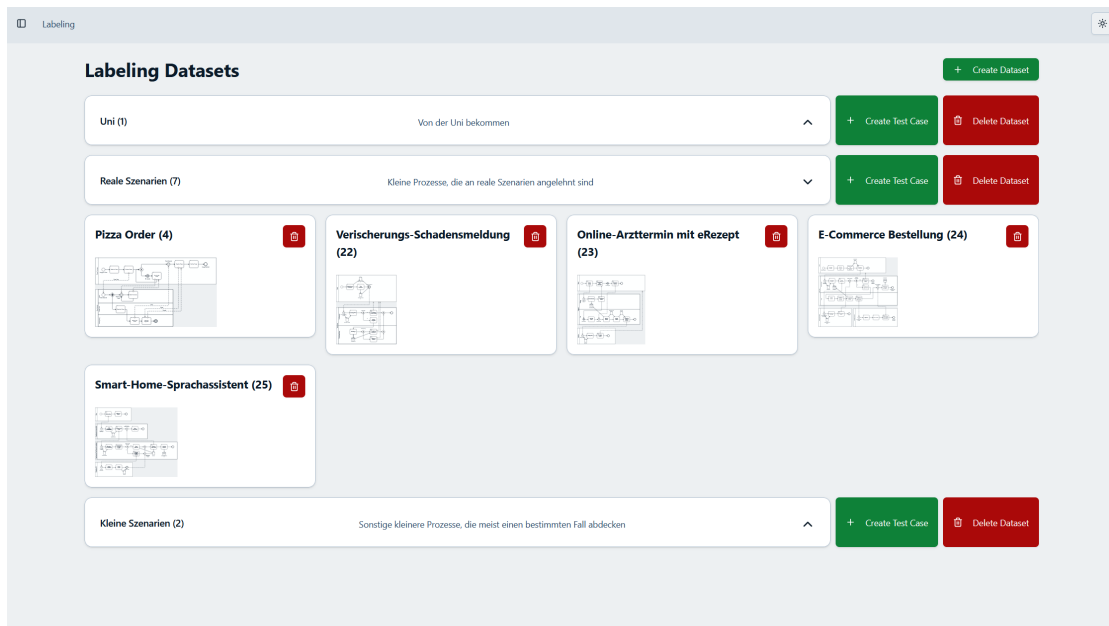


Abbildung 5.2: Übersicht der Datensätze im Labeling-Tool.

Diese heterogene Auswahl ist bewusst getroffen worden, da die Mischung aus verschiedenen Domänen und Modellkomplexitäten eine aussagekräftige Evaluation ermöglicht. In der Literatur wird betont, dass eine erhöhte Datensatzvielfalt die Robustheit der Bewertung steigert und einseitige Ergebnisse vermeidet [7]. Tabelle 5.1 zeigt die Eckdaten der Datensätze.

// TODO Noch mehr Testfälle hinzufügen - insbesondere, um auf die Datenassoziationsthematik aus dem Kapitel 2.2 einzugehen. Ich sollte also mindestens ein kleines Beispiel jeweils mit und ohne Datenassoziation haben, um den Unterschied zu verdeutlichen. Das bedeutet aber auch, dass ich die Tabelle 5.1 neu berechnen muss.

// TODO Hier noch einen Abschnitt wo ich ein paar besondere Testfälle hervorhebe, z.B. mit speziellen Datenassoziationen, oder reicht der Überblick mit den Eckdaten? Spätestens in den Fallstudien der Ergebnisse werde ich ja nochmal auf einzelne Testfälle eingehen.

Tabelle 5.1: Eckdaten der verwendeten Datensätze

	Uni-Prozesse	Reale Szenarien	Kleine Testfälle
Testfälle Gesamt	5	5	14
Testfälle (DE)	0	3	15
Testfälle (EN)	5	2	0
Ø Aktivitäten \pm SD ¹	13,4 \pm 2,6	11,6 \pm 4,2	3,9 \pm 1,4
Ø Aktivitäten (kritisch) \pm SD	8,6 \pm 3,6	6,6 \pm 1,9	2,1 \pm 1,5
Ø Datenobjekte \pm SD	1,4 \pm 1,9	3,6 \pm 2,1	0,4 \pm 0,7
Ø Datenassoziationen \pm SD	2,4 \pm 3,3	7 \pm 4	0,7 \pm 1,2
Ø Ereignisse \pm SD	21 \pm 13,8	8,2 \pm 2,8	2 \pm 0
Ø Gateways \pm SD	13 \pm 7,6	1,8 \pm 1,5	0 \pm 0
Ø Pools \pm SD	3,4 \pm 1,1	3 \pm 1	0,4 \pm 0,6
Ø Lanes \pm SD ²	3 \pm 1	4 \pm 0,7	0,3 \pm 0,5
Ø Nachrichtenflüsse \pm SD	9,4 \pm 5,3	5,2 \pm 0,8	0,1 \pm 0,3
Ø Annotationen \pm SD	1 \pm 1,7	0 \pm 0	0 \pm 0

¹ SD = Standardabweichung s der jeweiligen Anzahl pro Testfall.

² Blackbox-Pools ohne Lanes wurden nicht mitgezählt, daher kann der Durchschnittswert der Lanes geringer ausfallen als der, der Pools.

5.3 Labeling-Guide

Die Aktivitäten in den Testfällen sollen mit dem Label „kritisch“ versehen werden, wenn sie potenziell personenbezogene Daten verarbeiten und somit im Sinne der DSGVO relevant sein könnten. Die wichtigsten Begriffe der DSGVO wurden bereits in Abschnitt 2.1 definiert.

Beim Labeln einer Aktivität können Grenzfälle auftreten – etwa wenn kein Datenobjekt vorhanden ist, der Name aber auf Datenverarbeitung hindeutet (z. B. „Verträge archivieren“). Solche Verträge können personenbezogen sein (z. B. Arbeitsverträge) oder rein geschäftlich zwischen Unternehmen. In diesen Fällen wird zunächst der Kontext geprüft: Gibt es Hinweise auf personenbezogene Daten (z.,B. Pool/Lane oder angrenzende Aktivitäten im Prozess)? Fehlen eindeutige Hinweise, wird die Aktivität als unkritisch gelabelt. Deutet der Kontext hingegen auf die Verarbeitung personenbezogener Daten hin (z.,B. ein Prozessname wie „Mitarbeiterverwaltung“ oder vorangehende Aktivitäten wie „Mitarbeiterdaten erfassen“), erhält die Aktivität das Label kritisch. Im Zweifel wird kritisch gelabelt, um eine höhere Sensitivität zu gewährleisten.

Tabelle 5.2 listet beispielhaft einige Aktivitäten mit ihrer Klassifikation und einer Begründung auf.

Tabelle 5.2: Beispielhafte Aktivitäten und Label

Aktivität	Kritisch?	Kommentar
Lieferadresse eingeben	Ja	Name, Anschrift des Kunden werden aufgenommen.
Rückfrage an Kunden senden	Ja	Kontaktinformationen werden verwendet.
Fall anlegen	Ja	Aktivität befindet sich im Kundenservice-Kontext, personenbezogene Daten wahrscheinlich.
Sprache zu Text verarbeiten	Ja	Im Kontext eines Sprachassistenten werden biometrische Daten des Nutzers verarbeitet.
Produkt versenden	Nein*	Logistik und Sachvorgänge sind nicht per se Datenschutzkritisch, solange keine neue Datenverarbeitung, wie ein Labeldruck stattfindet.
Systemprotokoll auslesen	Ja	Im Kontext einer technischen Wartung können personenbezogene Daten (z.B. Nutzer-ids) enthalten sein.
Logdaten archivieren (anonym)	Nein	Keine personenbezogenen Daten enthalten.
Gerät kalibrieren	Nein	Im Kontext einer technischen Wartung werden keine personenbezogenen Daten verarbeitet.

6 Evaluationsframework

Nachdem nun Daten gelabelt werden können und der Testdatensatz für diese Arbeit erstellt wurde, wird nun in diesem Kapitel das Evaluationsframework vorgestellt. Das Framework nutzt die in Kapitel 4 entwickelte Klassifizierungspipeline, um verschiedene LLMs anhand gelabelter Testdaten systematisch, reproduzierbar und transparent zu vergleichen. Leitendes Gestaltungsprinzip ist die Entkopplung: Modelle und Klassifizierungsalgorithmen werden zur Laufzeit konfiguriert und sind dadurch austauschbar. So ermöglicht das Framework einen fairen Vergleich unterschiedlicher Modelle und Verfahren.

6.1 Use-Cases und Anforderungen

Das Evaluationsframework richtet sich an Forschende und Entwickler, die LLMs und Klassifizierungsalgorithmen für die Identifikation DSGVO-kritischer BPMN-Aktivitäten auswerten und miteinander vergleichen möchten. Es bietet eine einheitliche Ausführungs- und Auswertungsumgebung mit klar definierten Schnittstellen und standardisierten Berichten. In diesem Kapitel werden die Use-Cases und Anforderungen des Evaluationsframeworks beschrieben.

Use-Cases

Die wichtigsten Anwendungsfälle des Evaluationsframeworks sind:

- **Benchmarking von LLMs.** Systematischer Vergleich mehrerer LLMs auf denselben Datensätzen, mit identischem Algorithmus und identischen Parametern.

- **A/B-Vergleich von Algorithmen.** Gegenüberstellung verschiedener Klassifizierungspipelines, mit z.B. alternativen Prompts oder anderem Preprocessing, über eine standardisierte HTTP-Schnittstelle, die in Kapitel 4.4 definiert ist.
- **Explorative Analyse.** Detaillierte Einsicht pro Modell und Testfall (inklusive Begründungen und Visualisierungen), um Fehlklassifikationen gezielt zu untersuchen.
- **Berichterstellung.** Die Ergebnisse lassen sich als JSON oder Markdown exportieren und später wieder importieren, um sie erneut untersuchen zu können. Sie eignen sich zudem für die Publikation. Die Diagramme werden automatisch erzeugt und stehen ebenfalls zum Download bereit.

In dieser Arbeit werden keine A/B-Vergleiche unterschiedlicher Klassifizierungsalgorithmen durchgeführt, sondern lediglich verschiedene LLMs mit demselben Algorithmus verglichen. Das Framework ist jedoch so konzipiert, dass dies in zukünftigen Arbeiten möglich ist.

Funktionale Anforderungen

In der folgenden Tabelle sind die funktionalen Anforderungen an das Evaluationsframework aufgelistet, die notwendig sind um Use-Cases zu erfüllen:

ID:	FA01
Titel:	Nutzen gelabelter Testdatensätze
Beschreibung:	Das Framework kann die gelabelten Testdatensätze benutzen, die mit dem Labeling-Tool aus 5.1 erstellt worden sind.
Abhängigkeit:	

ID:	FA02
Titel:	Vergleichbarkeit von Modellen und Algorithmen
Beschreibung:	Das Framework erlaubt den direkten Vergleich verschiedener LLMs sowie unterschiedlicher Klassifizierungsalgorithmen anhand gelabelter Testdaten. Die Anbindung an Klassifizierungsalgorithmen erfolgt über die in Kapitel 4.4 definierte, standardisierte HTTP-Schnittstelle.
Abhängigkeit:	FA01

ID:	FA03
Titel:	Deklarative Konfiguration
Beschreibung:	Ein Evaluationslauf ist vollständig über eine YAML-Datei konfigurierbar. Dazu zählen Modelle, Klassifizierungsendpunkte, Testdatensätze, Seed. Experimente werden dadurch portabel und wiederholbar.
Abhängigkeit:	FA02

ID:	FA04
Titel:	Detaillierte Ergebnisaufbereitung
Beschreibung:	<p>Das Framework gibt Ergebnisse auf zwei Ebenen aus.</p> <ol style="list-style-type: none"> 1. Pro Testfall und pro Modell: Status („bestanden“/„nicht bestanden“), klassifizierte Elemente mit Begründungen, TP/FP/FN/TN und eine Visualisierung der Klassifikation im BPMN-Prozess. 2. Pro Modell als Summe über alle Testfälle: Accuracy, Precision, Recall, F1-Score und die Konfusionsmatrix. <p>Zusätzlich protokolliert das Framework Metadaten der Evaluation, z. B. Endpunkt, verwendete Modelle und den Seed.</p>
Abhängigkeit:	FA02

ID:	FA05
Titel:	Frontend
Beschreibung:	Für eine einfache Bedienung und Ansicht der Ergebnisse bietet das Evaluationsframework ein Frontend an.
Abhängigkeit:	FA02, FA03, FA04

ID:	FA06
Titel:	Visualisierung und Berichte der Gesamtergebnisse
Beschreibung:	Kennzahlen werden als Side-by-Side-Diagramme und tabellarisch dargestellt. Zusätzlich stehen Export/Import der Ergebnisse als JSON sowie ein Markdown-Report zur Verfügung.
Abhängigkeit:	FA05

6.2 Testdaten

Wie in FA01 beschrieben, kann das Evaluationsframework die mit dem Labeling-Tool erzeugten Testdatensätze unmittelbar verwenden. Da das Tool die Testdaten in einer Datenbank ablegt, lassen sie sich unkompliziert auslesen und für die Evaluierung heranziehen. In der Konfiguration des Frameworks wird festgelegt, welche Datensätze genutzt werden, wodurch sich die Auswertung gezielt auf einen bestimmten Anwendungsfall zuschneiden lässt. Wie die Konfiguration einer Evaluierung funktioniert, wird im nächsten Kapitel erläutert.

6.3 Konfiguration einer Evaluierung

Die funktionale Anforderung FA03 fordert, dass Evaluationsläufe deklarativ konfiguriert werden können. Das Framework unterstützt dies auf zwei Wegen: Erstens bietet die Weboberfläche, die in 6.6 gezeigt wird, die Möglichkeit, Evaluationsläufe interaktiv zu konfigurieren und zu starten. Zweitens lässt sich eine Evaluierung über eine YAML-Datei beschreiben, die entweder in der Weboberfläche hochgeladen oder per CLI an das Evaluationsframework übergeben wird. Auf diese Weise werden Reproduzierbarkeit und Versionierung der Evaluationsläufe sichergestellt.

Listing 6.1 zeigt ein Beispiel für eine solche YAML-Konfiguration. Ein ausführliches JSON-Schema ist im Anhang (Listing A.4) zu finden.

Listing 6.1: Beispiel einer Evaluierungskonfiguration in YAML.

```
1 defaultEvaluationEndpoint: /gdpr/analysis/prompt-engineering
2 maxConcurrent: 10
3 seed: 42
4 models:
5   - label: Mistral Medium 3.1
6     llmProps:
7       baseUrl: https://openrouter.ai/api/v1
8       modelName: mistralai/mistral-medium-3.1
9       apiKey: ${OPEN_ROUTER_API_KEY}
10  - label: Deepseek Chat v3.1
11    llmProps:
12      baseUrl: https://openrouter.ai/api/v1
13      modelName: deepseek/deepseek-chat-v3.1
14      apiKey: ${OPEN_ROUTER_API_KEY}
15  - label: GPT oss 120b
16    llmProps:
17      baseUrl: https://openrouter.ai/api/v1
18      modelName: openai/gpt-oss-120b
19      apiKey: ${OPEN_ROUTER_API_KEY}
20 datasets:
21   - 2
22   - 7
```

Die Evaluierungskonfiguration umfasst die folgenden Bausteine:

- `defaultEvaluationEndpoint` ist der Standardendpunkt für die Klassifizierung. Er wird verwendet, wenn für ein Modell kein eigener Endpunkt angegeben ist. Der Endpunkt muss die in Kapitel 4.4 beschriebene API-Spezifikation erfüllen und kann relativ (gegen die Basis-URL des Evaluationsframeworks) oder absolut (für einen externen Dienst) angegeben werden.
- `maxConcurrent` gibt die maximale Anzahl parallel auszuführender Testfälle an. So lassen sich beispielsweise Rate-Limits von LLMs einhalten.

- `seed` legt einen Startwert (Seed) für reproduzierbare Evaluationsläufe fest. Er wird bei jedem Modell an die `llmProps` weitergereicht und bei der Kommunikation mit den LLMs verwendet, sofern diese einen Seed unterstützen.
- `models` enthält die zu evaluierenden Modelle. Jedes Modell besitzt ein `label` zur Identifikation und optional spezifische `llmProps`, um die Eigenschaften des verwendeten LLMs zu definieren.
- `datasets` ist eine Liste von Datensatz-ids, die in der Evaluierung verwendet werden. Die ids referenzieren die Datensätze, die in der Datenbank verwaltet werden und jeweils eine Menge von Testfällen beinhalten.

Wie im Schema in Listing A.4 gezeigt, kann jedem Modell optional ein eigener `evaluationEndpoint` zugewiesen werden, der den in `defaultEvaluationEndpoint` definierten Standard überschreibt. Dadurch lassen sich unterschiedliche Klassifizierungsalgorithmen oder -versionen gezielt pro Modell vergleichen. Ist kein spezifischer Endpunkt angegeben, greift automatisch der Standardendpunkt.

API-Keys in den `llmProps` können optional als Umgebungsvariablen referenziert werden, wie im Beispiel in Listing 6.1 gezeigt. So lassen sich sensible Daten sicher handhaben, ohne sie direkt in der Konfigurationsdatei zu speichern. Die Umgebungsvariablen werden zur Laufzeit aufgelöst und müssen daher im Kontext des Anwendung verfügbar sein.

6.4 Architektur und Komponenten

Das Evaluationsframework ist modular aufgebaut und nutzt eine Pipeline-Architektur, um eine flexible und skalierbare Evaluierung zu ermöglichen, wie es in FA02 gefordert ist. Die Architektur ist in Abbildung 6.1 dargestellt. Sie besteht aus mehreren Hauptkomponenten, die jeweils eine klar definierte Aufgabe erfüllen. Im Folgenden werden die Komponenten und ihr Zusammenspiel beschrieben.

Einstiegspunkte

Das Framework bietet zwei Einstiegspunkte zur Ausführung einer Evaluierung:

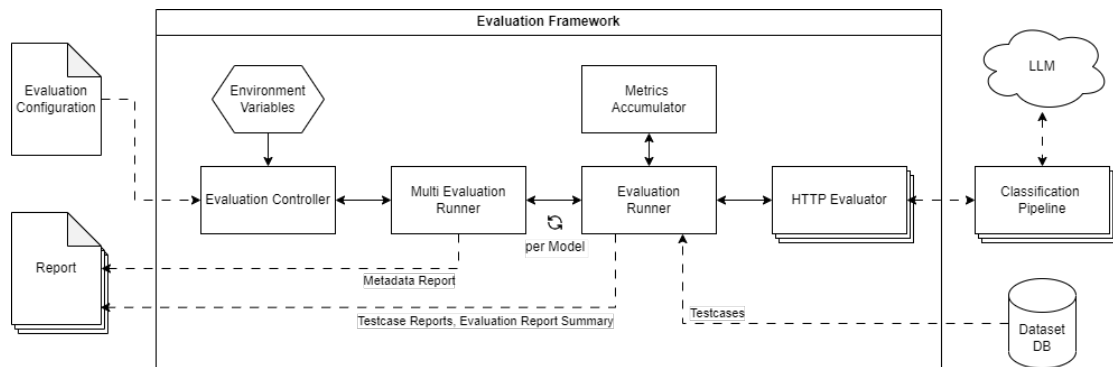


Abbildung 6.1: Architektur des Evaluationsframeworks

- **EvaluationController** als HTTP-Controller stellt REST-Endpunkte bereit, über die Evaluierungen gestartet werden können. Er akzeptiert eine YAML-Konfiguration und gibt die Ergebnisse entweder auf einmal als Markdown-Bericht oder als JSON-Stream zurück. Der Controller ermöglicht die Nutzung des Frameworks über die Weboberfläche, die in Kapitel 6.6 gezeigt wird, sowie über HTTP-APIs. Durch das Streamen der Ergebnisse können bereits abgeschlossene Testfälle sofort angezeigt werden, ohne auf das Ende der gesamten Evaluierung warten zu müssen.
- **EvaluationCommand** ist ein CLI-Befehl, der die Ausführung von Evaluierungen über die Kommandozeile erlaubt. Er liest eine YAML-Konfigurationsdatei ein, führt die Evaluierung aus und schreibt die Ergebnisse in eine Markdown-Datei. Dies eignet sich besonders für automatisierte Ausführungen, Continuous Integration oder die lokale Entwicklung.

Beide Einstiegspunkte akzeptieren die Konfiguration aus Kapitel 6.3, lösen ggf. Umgebungsvariablen auf und delegieren die Ausführung der Evaluation an den `MultiEvaluationRunner`.

Orchestrierung mit `MultiEvaluationRunner`

Der `MultiEvaluationRunner` ist für die Orchestrierung der gesamten Evaluierung verantwortlich. Er verarbeitet die Konfiguration, die mehrere Modelle und Datensätze beschreibt, und koordiniert die sequenzielle Evaluierung aller konfigurierten Modelle. Für jedes Modell ruft der `MultiEvaluationRunner` den `EvaluationRunner`

auf und übergibt diesem alle Informationen zur Ausführung der Evaluation eines einzelnen Modells.

Der `MultiEvaluationRunner` stellt zudem sicher, dass alle Modelle denselben Seed verwenden, um reproduzierbare Ergebnisse zu gewährleisten. Falls in der Konfiguration kein Seed angegeben wurde, wird an dieser Stelle einer erzeugt. Im Anschluss werden die Metadaten über die verwendeten Datensätze, Modelle, Endpunkte und den Seed gesammelt und als `MetadataReport` als Teil des Evaluationsberichts zurückgegeben. Die Teile des Berichts werden als Flow von Berichtartefakten zurückgegeben, wodurch eine streambasierte Verarbeitung ermöglicht wird. Welche Arten von Berichtartefakten es gibt, wird in Kapitel 6.5 beschrieben.

Ausführung mit `EvaluationRunner`

Der `EvaluationRunner` führt die Evaluierung für ein einzelnes Modell durch. Er lädt die Testfälle der angegebenen Testdatensätze aus der Datenbank, führt die Klassifizierung für jeden Testfall aus und sammelt die Ergebnisse. Die Testfälle werden parallel verarbeitet, wobei die Anzahl der gleichzeitigen Ausführungen durch den Parameter `maxConcurrent` in der Konfiguration gesteuert wird. Dies ermöglicht es, Rate-Limits von LLMs-Diensten einzuhalten und die Auslastung der Ressourcen zu kontrollieren.

Für jeden Testfall delegiert der `EvaluationRunner` die eigentliche Klassifizierung an den `HttpEvaluator`. Anschließend vergleicht er die erwarteten mit den tatsächlichen Ergebnissen und berechnet die Klassifikationsmetriken wie TP, FP, FN und TN. Die Ergebnisse werden in `TestCaseReport`-Objekten zusammengefasst, als Teilergebnis zurückgegeben, und an den `MetricsAccumulator` weitergeleitet.

Der `EvaluationRunner` gibt die Ergebnisse ebenfalls als Flow zurück, wodurch eine frühzeitige Rückgabe von Teilergebnissen ermöglicht wird. Dies ist besonders vorteilhaft für die Live-Ansicht in der Weboberfläche, da Testfallergebnisse sofort nach ihrer Fertigstellung angezeigt werden können.

Klassifizierung mit `HttpEvaluator`

Der `HttpEvaluator` ist für die Kommunikation mit dem Klassifizierungsendpunkt verantwortlich, der das in Kapitel 4.4 beschriebene Interface implementiert. Er nimmt das BPMN-Modell aus dem aktuellen Testfall und die `LlmProps` von dem aktuellen Modell aus der Konfiguration entgegen, baut einen HTTP-Request auf und sendet diesen an den konfigurierten Endpunkt. Nach erfolgreicher Klassifizierung extrahiert er die Liste der als kritisch identifizierten Aktivitäten aus der Antwort und gibt diese an den `EvaluationRunner` zurück.

Akkumulierung mit `MetricsAccumulator`

Der `MetricsAccumulator` sammelt die Metriken aller Testfälle eines Modells und berechnet daraus aggregierte Werte. Er ist thread-sicher implementiert und kann gleichzeitig von mehreren parallelen Evaluierungen genutzt werden. Das ist wichtig, da der `EvaluationRunner` die Testfälle parallel ausführt und somit mehrere Threads gleichzeitig auf den `MetricsAccumulator` zugreifen können.

Nach Abschluss aller Testfälle erzeugt der `MetricsAccumulator` ein `EvaluationReportSummary`-Objekt, das alle Metriken für die Evaluation eines Modells über mehrere Testfälle hinweg enthält.

Zusammenfassung

Die Architektur trennt Zuständigkeiten strikt: `MultiEvaluationRunner` koordiniert Modellläufe, `EvaluationRunner` verarbeitet Testfälle und sammelt Metriken, `HttpEvaluator` kommuniziert mit der Klassifizierungs-Pipeline, `MetricsAccumulator` aggregiert Ergebnisse pro Modell über mehrere Testfälle.

6.5 Evaluationsergebnisse

Im vorherigen Abschnitt wurde erwähnt, dass die Komponenten des Evaluationsframeworks Berichtartefakte zurückgeben. Diese sind im im Architekturbild 6.1 als

Beschriftungen über den gestrichelten Pfeilen dargestellt. Im Folgenden werden die Berichtartefakte beschrieben:

MetadataReport Berichtsartefakt, das der `MultiEvaluationRunner` zu Beginn der Evaluierung erzeugt. Es enthält Metadaten zur Evaluierung, z. B. Informationen über die Testdatensätze, die Anzahl der Testfälle sowie den verwendeten Seed. Das `MetadataReport`-Artefakt wird zuerst zurückgegeben, damit die Weboberfläche bereits Metadaten anzeigen kann, während die Evaluierung noch läuft.

TestCaseReport Berichtsartefakt, das der `EvaluationRunner` für jeden abgeschlossenen Testfall erzeugt. Es enthält u. a. die Testfall-id, das Klassifizierungsergebnis und die für diesen Testfall berechneten Metriken. `TestCaseReport`-Artefakte werden fortlaufend bereitgestellt, sobald ein Testfall abgeschlossen ist, sodass die Weboberfläche Ergebnisse unmittelbar anzeigen kann.

EvaluationReportSummary Berichtsartefakt, das der `MetricsAccumulator` am Ende der Evaluierung eines Modells erzeugt. Es fasst die aggregierten Metriken, wie z. B. *Precision*, *Recall*, *F1-Score* und *Accuracy*, sowie die Konfusionsmatrix zusammen. Das `EvaluationReportSummary`-Artefakt wird als letztes Berichtsartefakt pro Modell zurückgegeben und dient dem Modellvergleich in der Weboberfläche.

Die Informationen dieser Berichtsartefakte ermöglichen die Generierung eines ausführlichen Evaluierungsberichts, wie in FA04 gefordert. Im Folgenden ist dargestellt, welche Informationen nach Abschluss einer Evaluierung vorliegen.

Pro Testfall und Modell

Für jeden Testfall eines Modells liegen vor: die von der Klassifizierungs-Pipeline zurückgegebenen klassifizierten Aktivitäten (mit optionalen Begründungen), die gelabelten erwarteten Aktivitäten, die Zählwerte für *TP*, *FP*, *FN* und *TN* sowie eine Bild-URL zur Visualisierung des BPMN-Modells mit hervorgehobenen Aktivitäten. Aus diesen Informationen lässt sich ableiten, ob der Testfall erfolgreich war. Ein Testfall gilt als erfolgreich, wenn die klassifizierten Aktivitäten exakt den erwarteten Aktivitäten entsprechen. Technische Probleme, die während der Klassifizierung

auftreten, werden ebenfalls erfasst, z. B. Parsing-Fehler, ungültiges BPMN, Token-Limit-Überschreitungen oder Zeitüberschreitungen.

Pro Modell über alle Testfälle

Auf Modellebene stehen die Gesamtergebnisse über alle Testfälle zur Verfügung. Dazu gehören die aggregierten Kennzahlen *Precision*, *Accuracy*, *Recall* und *F1-Score* sowie eine Konfusionsmatrix mit den Gesamtwerten für *TP*, *FP*, *FN* und *TN*. Zusätzlich sind die Anzahlen der korrekt bzw. falsch klassifizierten sowie der technisch fehlgeschlagenen Testfälle aufgeführt.

Über alle Modelle

Abschließend sind die Metadaten der gesamten Evaluierung verfügbar: die verwendeten Testdatensätze, die Anzahl der Testfälle, die konfigurierten Modelle, der für die Reproduzierbarkeit verwendete Seed sowie ein Zeitstempel der Evaluierung. Zum unmittelbaren Vergleich werden die aggregierten Kennzahlen aller Modelle nebeneinander dargestellt.

6.6 Frontend

Das Frontend des Evaluationsframeworks setzt die Anforderungen FA05 und FA06 um. Es unterstützt die interaktive Konfiguration von Evaluierungen, die Live-Verfolgung des Fortschritts sowie die detaillierte Analyse der Ergebnisse bis auf Ebene einzelner Testfälle. Die Oberfläche ist so gestaltet, dass zentrale Kennzahlen wie *Accuracy*, *Precision*, *Recall* und *F1-Score*, die Konfusionsmatrix mit *TP*, *FP*, *TN*, *FN* sowie die Bestehensraten aller Modelle zunächst auf einen Blick erfasst und anschließend schrittweise vertieft werden können.

Evaluation Config
Multiple Models

Import YAML Config Download YAML Config

Default Settings

Default Evaluation Endpoint
Use preset

Preset Endpoint
Preprocessing & Prompt Engineering Analysis

Max Concurrent LLM Requests
10

Seed
Optional seed for reproducibility

Warning: Not all models support a seed, but it will be used for models that support them.

Datasets

Select Datasets
Uni Reale Scenarien Kleine Scenarien

Models Configuration

Model 1 Effective endpoint: /gdpr/analysis/prompt-engineering

Label
Deepseek Chat v3.1

Endpoint
Use default

LLM Base URL
https://openrouter.ai/api/v1

LLM Model Name
deepseek/deepseek-chat-v3.1

LLM Response Timeout (seconds)
240

API Key
\${OPEN_ROUTER_API_KEY}

Model 2 Effective endpoint: /gdpr/analysis/prompt-engineering

Label
Mistral Medium 3.1

Endpoint
Use default

LLM Base URL
https://openrouter.ai/api/v1

LLM Model Name
mistralai/mistral-medium-3.1

LLM Response Timeout (seconds)
240

API Key
\${OPEN_ROUTER_API_KEY}

Download Markdown Report Download JSON Report Upload JSON Report Start Evaluation

Abbildung 6.2: Formular zur Konfiguration einer Evaluation.

Konfigurationsansicht

Abbildung 6.2 zeigt das Formular zur Konfiguration einer Evaluierung. Sämtliche Parameter, die bereits aus der YAML-Konfiguration in Kapitel 6.3 bekannt sind, lassen sich hier setzen. Verfügbare Standardwerte, zum Beispiel der Endpunkt der in dieser Arbeit verwendeten Klassifizierungspipeline oder die in der Datenbank verfügbaren Datensätze, werden automatisch geladen.

YAML-Konfigurationen können importiert und exportiert werden, um sie zu speichern oder weiterzugeben. Unter dem Formular befinden sich Schaltflächen zum Starten der Evaluierung sowie zum Import und Export von JSON-Berichten. Auf diese Weise lassen sich Ergebnisse archivieren und später erneut laden, ohne die Evaluierung erneut ausführen zu müssen.

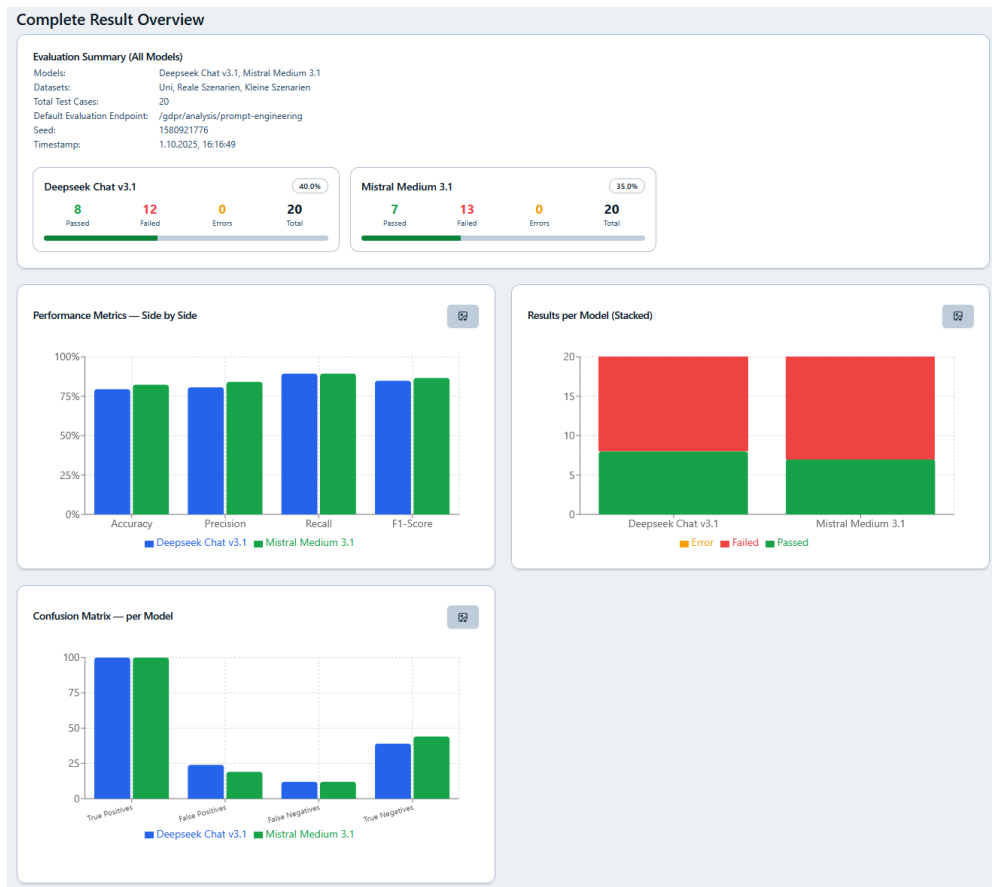


Abbildung 6.3: Gesamtübersicht einer Evaluierung mit Side-by-Side-Diagrammen.

Gesamtübersicht

Nach dem Start der Evaluierung werden die Ergebnisse pro Modell inkrementell vom Backend übermittelt und — wie in Abbildung 6.3 — angezeigt. Dadurch können Teilergebnisse bereits untersucht werden, während die Evaluierung noch läuft. Die Gesamtübersicht bietet eine kompakte Zusammenfassung pro Modell mit den Kategorien Bestanden, Nicht bestanden und Fehler sowie Side-by-Side-Diagramme aller Metriken mit allen Modellen. So lassen sich die Modelle unmittelbar nebeneinander vergleichen. Oberhalb sind die Metadaten der Evaluierung aufgeführt.

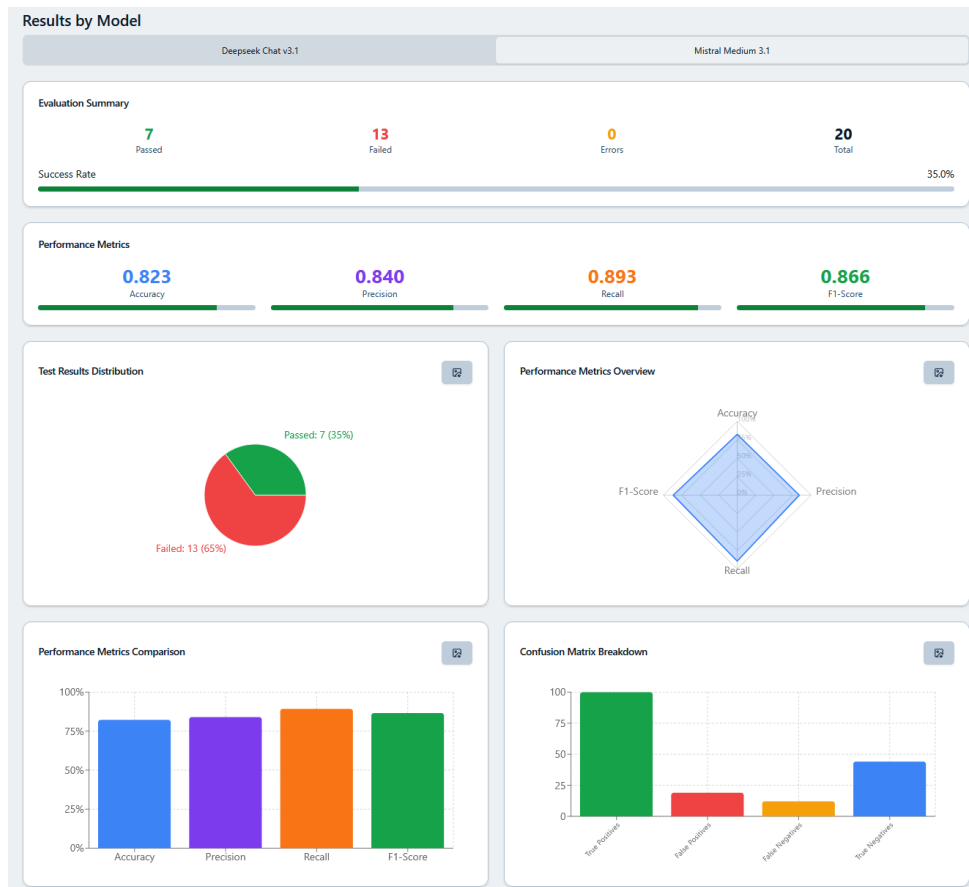


Abbildung 6.4: Modell-Detailansicht mit exemplarischen Ergebnissen.

Ergebnisse pro Modell

Für eine vertiefte Analyse stellt das Frontend für jedes Modell eine Detailansicht bereit, die alle Kennzahlen über sämtliche Testfälle aggregiert. Abbildung 6.4 zeigt diese Ansicht. Über Tabs kann zwischen den Modellen gewechselt werden, was einen schnellen Vergleich ermöglicht.

Ergebnisse pro Testfall

Neben den aggregierten Ergebnissen pro Modell lassen sich auch die Resultate einzelner Testfälle je Modell untersuchen. Abbildung 6.5 zeigt die Detailseite eines Testfalls. Sie enthält unter anderem den Status, die erwarteten gelabelten Aktivitäten und die vom Modell detektierten Aktivitäten. Zusätzlich visualisiert eine BPMN-

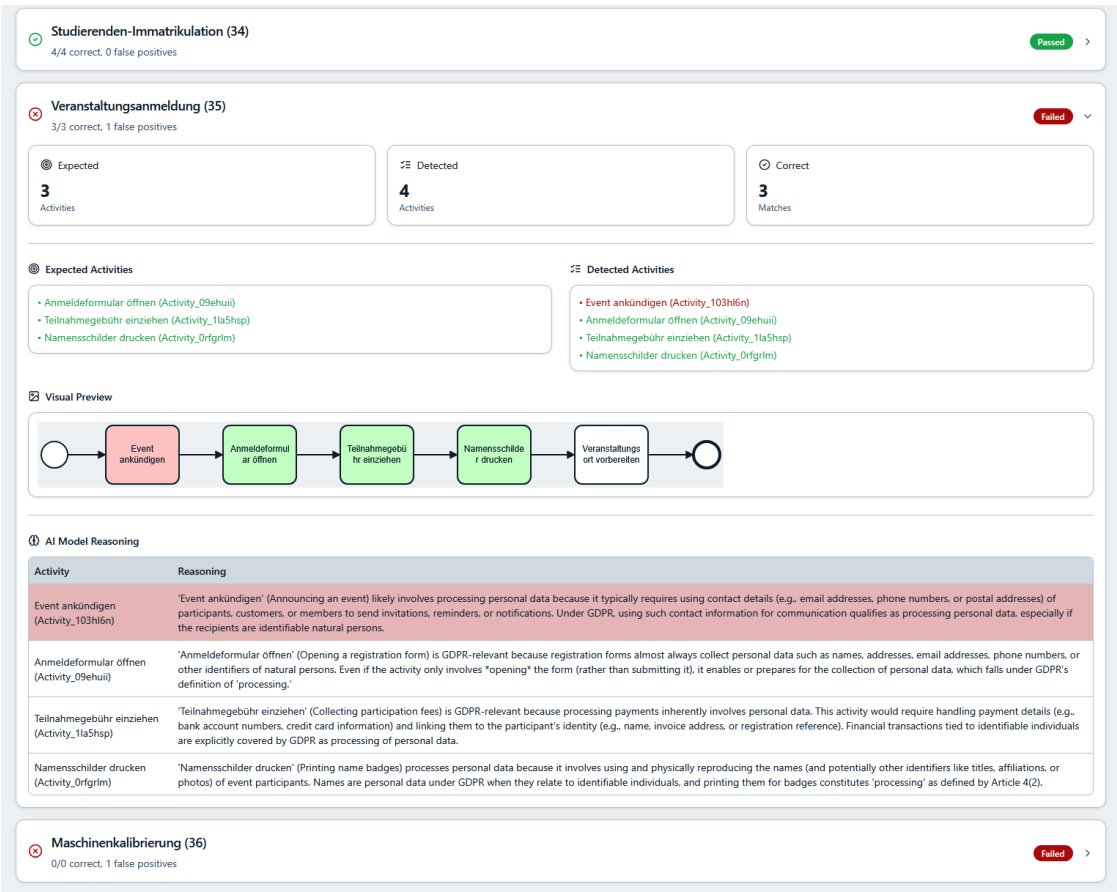


Abbildung 6.5: Detailseite eines Testfalls mit exemplarischen Ergebnissen.

Darstellung den Prozess, und die Aktivitäten sind je nach korrekter oder inkorrekt Klassifizierung farblich markiert. Falls vorhanden, wird außerdem die vom LLM gelieferte Begründung pro Aktivität angezeigt.

Abweichungen werden dadurch unmittelbar sichtbar, und typische Fehlmuster wie systematische FP bei bestimmten Aktivitätstypen lassen sich schnell erkennen. Testfälle, die aufgrund technischer Fehler nicht klassifiziert werden konnten, werden mit der entsprechenden Fehlermeldung aufgeführt.

6.7 Erweiterbarkeit

Das Evaluationsframework ist auf die Entkopplung von *Modell*, *Klassifizierungspipeline* und *Testdaten* ausgelegt. Neue Modelle werden rein konfigurationsbasiert eingebunden, indem Endpunkt, Modellname und API-Key hinterlegt werden. Dadurch ist ein Austausch ohne Codeänderungen möglich. Klassifizierungsalgorithmen bzw. Pipelines werden über einen konfigurierbaren HTTP-Endpunkt ergänzt, sofern sie das in Abschnitt 4.4 definierte Interface implementieren. Dadurch lassen sich verschiedene Varianten von Algorithmen unkompliziert unter identischen Rahmenbedingungen vergleichen. Die Testdatensätze werden zur Laufzeit aus der Datenbank geladen und pro Evaluierung können unterschiedliche Datensätze ausgewählt werden. Neue Datensätze lassen sich jederzeit hinzufügen, um weitere Domänen oder Schwierigkeitsgrade abzudecken. Diese Architektur ermöglicht es, die Umgebung schrittweise zu erweitern und aktuelle Modelle, Verfahren konsistent und reproduzierbar zu evaluieren.

7 Modellauswahl

Aufbauend auf dem im letzten Kapitel vorgestellten Evaluationsframework werden in diesem Kapitel die für die Klassifizierungsaufgabe zu vergleichenden LLMs vorgestellt. Um die Auswahl nachvollziehbar zu machen und künftige Arbeiten mit ähnlicher Zielsetzung zu unterstützen, werden zunächst die Auswahlkriterien der Modelle erläutert. Anschließend folgt eine Vorstellung der ausgewählten Modelle.

7.1 Kriterien

Dieser Abschnitt legt die Auswahlkriterien der LLMs offen, nach denen die Modelle ausgewählt und kategorisiert wurden. Tabelle 7.1 zeigt eine Übersicht der Kriterien. Diese Kriterien helfen, die Modelle systematisch zu vergleichen und ihre Eignung für die Klassifizierungsaufgabe zu bewerten.

Tabelle 7.1: Übersicht der Kriterien zur Modellauswahl

Kriterium	Beschreibung
Herkunft	Das Land in dem das Modell entwickelt wurde bzw. der Hauptsitz des Anbieters
Lizenz	Art der Lizenz, wie bspw. Open-Source oder proprietär
Größe	Anzahl der Parameter in Milliarden (B)
Kontext	Maximale Anzahl der Token, die das Modell verarbeiten kann
Letztes Update	Datum der letzten Aktualisierung des Modells bei Hugging Face [28]
Downloads	Anzahl der Downloads des Modells bei Hugging Face, sofern verfügbar

Ein wesentliches Auswahlkriterium ist die geografische **Herkunft** der Modelle. Als

EU-Modell gelten Modelle, deren Anbieter ihren Hauptsitz in der EU haben, deren Veröffentlichung in der EU erfolgt oder die schwerpunktmäßig in der EU entwickelt oder verfeinert wurden. Alle anderen Modelle werden als *international* eingeordnet. Diese Unterscheidung ist relevant, da europäische Modelle sowohl beim Training als auch beim Betrieb stärker den europäischen Datenschutzbestimmungen unterliegen und somit potenziell besser für den Einsatz in datensensiblen Bereichen wie der Klassifizierung von BPMN-Modellen geeignet sind.

Ein weiteres zentrales Kriterium ist die **Lizenzierung** der Modelle. Als *Open-Source* oder *Open-Weights* veröffentlichte Modelle bieten mehr Flexibilität und Transparenz. Sie sind häufig lizenzkostenfrei nutzbar und lassen sich eigenständig betreiben, was insbesondere ein Vorteil für Unternehmen und Organisationen mit strengen Datenschutzanforderungen ist. Proprietäre Modelle sind dagegen an spezifische Nutzungsbedingungen gebunden und bringen teils Einschränkungen bei Datenverarbeitung und -speicherung mit sich. Im engeren Sinne definiert die Open Source Initiative (OSI) Open-Source-Lizenzen über die *Open Source Definition* [52]. Viele aktuelle LLMs erscheinen als *Open-Weights*, was bedeutet, dass die Gewichte frei beziehbar sind. Die Lizenz kann jedoch restriktive Klauseln enthalten (z. B. die Meta-Llama 3 Community License mit Nutzungs- und Output-Beschränkungen [40]). In dieser Arbeit gilt ein Modell als *offen* bzw. *Open-Source-nah*, wenn

1. die Gewichte frei zugänglich sind und
2. eine *permissive* Lizenz (z. B. Apache-2.0 oder MIT) eine breite kommerzielle Nutzung erlaubt (z. B. Mistral 7B [29], GPT-OSS [54, 56] oder DeepSeek V3.1 [25]).

Modelle mit *Community*- oder *Eigennutzer*-Lizenzen (z. B. Mistral Large Instruct unter Mistral Research License [30, 43]) werden rechtlich *nicht* als OSI-Open-Source gewertet, können aber technisch als Vergleich herangezogen werden.

Die **Modellgröße** wird in *Anzahl der Parameter* angegeben. Meist in *Milliarden* (Billionen, engl. *Billion*) Parametern (B, engl. *Billion*). $1 \text{ B} = 10^9$ Parameter. Diese Zahl korreliert mit dem Ressourcenbedarf für Training und Inferenz sowie der Leistungsfähigkeit [49]. Für die Einordnung werden hier folgende Klassen verwendet:

- **Klein** ($\leq \sim 25 \text{ B}$ Parameter): z. B. Mistral 7B Instruct ($\sim 7.3 \text{ B}$) [29].
- **Groß** ($> \sim 25 \text{ B}$ Parameter): z. B. GPT-OSS 120B ($\sim 117 \text{ B}$) [54].

Die Klassifikation dient als methodische Abgrenzung für die Experimente. Kleinere Modelle lassen sich häufig lokal ausführen, größere erfordern typischerweise mehrere GPUs. Parameterzahl ist dabei ein nützlicher, wenn auch unvollständiger Indikator für Ressourcenbedarf und erwartete Leistung. Dies ermöglicht konsistente Entscheidungen zu Deployment und Kosten.

Der **Kontext** gibt an, wie viele Token ein Modell gleichzeitig verarbeiten kann. Ein Token ist dabei eine Grundeinheit von Text, die ein Wort, einen Teil eines Wortes oder sogar ein einzelnes Zeichen darstellen kann. Die Größe des Kontextfensters beeinflusst maßgeblich, wie gut ein Modell längere Texte verstehen und darauf reagieren kann [6].

Neben den genannten Hauptkriterien werden weitere Merkmale erfasst, um die Modelle umfassend zu charakterisieren. Dazu gehören das **letzte Update** des Modells, um einordnen zu können, wie aktuell das Modell ist, und wie viele **Downloads** das Modell hat, sofern verfügbar. Diese Informationen helfen, die Popularität und Akzeptanz der Modelle in der Community einzuschätzen.

Im nächsten Abschnitt werden auf Basis dieser Kriterien die ausgewählten Modelle vorgestellt.

7.2 Modellvorstellung

Tabelle 7.2 stellt die für diese Arbeit ausgewählten Modelle kurz vor. Der Stichtag der Modellauswahl war der 30. September 2025. Im Folgenden wird die Modellauswahl im Detail erläutert.

Die französische Firma Mistral AI bietet mehrere leistungsstarke Modelle an, die zum großteil offene Gewichte haben. Durch ihre Herkunft repräsentieren die Mistral Modelle in dieser Arbeit die EU-Modelle. **Mistral-7B-Instruct-v0.3** [29] ist ein 7,24 B Parameter großes Modell mit einem Kontextfenster von 32,000 Tokens. Es wurde speziell für Anweisungsfolgen (Instruct) optimiert und ist unter der Apache-2.0-Lizenz frei verfügbar. Das **Mixtral-8x7B-Instruct-v0.1** Modell [31] nutzt eine Mixture-of-Experts-Architektur mit insgesamt 46,7 B Parametern, von denen jedoch nur 12,9 B aktiv genutzt werden. Es hat ebenfalls ein Kontextfenster von 32,000 Tokens und ist unter der Apache-2.0-Lizenz lizenziert. Das **Mistral-Large-Instruct-**

Tabelle 7.2: Übersicht aller Modelle mit technischen Eckdaten (Stand 30.09.2025).

(a) Technische Eckdaten und Herkunft der Modelle.

Modell	Parameter (B)	Kontext (Tokens)	Herkunftsland
Mistral-7B-Instruct-v0.3	7.24	32,000	Frankreich [29]
Mixtral-8x7B-Instruct-v0.1	46.7 total / 12.9 aktiv ¹	32,000	Frankreich [31, 44]
Mistral-Large-Instruct-2411	123	128,000	Frankreich [30]
Mistral-Medium-3.1	n.v.	128,000	Frankreich [45]
Gemma-3-12B-it	12.2	128,000	Großbritannien ² [26]
Gemma-3-27B-it	27.4	128,000	Großbritannien [27]
Qwen2.5-7B-Instruct	7.62	131,072	China [33]
Qwen3-235B-A22B-Thinking-2507	235	256,000	China [32]
DeepSeek-R1-Distill-Qwen-14B	14.8	131,072	China [23]
DeepSeek-V3.1	671 total / 37 aktiv	128,000	China [25]
GPT-OSS-20B	20.91 total / 3.61 aktiv	131,072	USA [54]
GPT-OSS-120B	116.83 total / 5.13 aktiv	131,072	USA [54]
GPT-4o (2024-11-20)	n.v.	128,000	USA [55]

(b) Lizenz, letzte Updates und Downloads der Modelle.

Modell	Lizenz	Letztes Update	Downloads
Mistral-7B-Instruct-v0.3	Apache-2.0	24.07.2025	687,000 [29]
Mixtral-8x7B-Instruct-v0.1	Apache-2.0	24.07.2025	43,500 [31]
Mistral-Large-Instruct-2411	Mistral Research License	28.07.2025	4,200 [30, 43]
Mistral-Medium-3.1	Proprietär	25.08.2025	n.v. [45]
Gemma-3-12B-it	Gemma	21.03.2025	523,000 [21, 26]
Gemma-3-27B-it	Gemma	21.03.2025	1,180,000 [21, 27]
Qwen2.5-7B-Instruct	Apache-2.0	12.01.2025	5,100,000 [33]
Qwen3-235B-A22B-Thinking-2507	Apache-2.0	17.08.2025	52,900 [32]
DeepSeek-R1-Distill-Qwen-14B	MIT	24.02.2025	341,000 [23]
DeepSeek-V3.1	MIT	05.09.2025	447,000 [25]
GPT-OSS-20B	Apache-2.0	26.08.2025	6,450,000 [54]
GPT-OSS-120B	Apache-2.0	26.08.2025	3,600,000 [54]
GPT-4o (2024-11-20)	Proprietär	20.11.2024	n.v. [55]

¹ Mistral nutzt Mixture-of-Experts (MoE) mit 8 Experten als Architektur. Die Gesamtparameterzahl bezieht sich auf alle Experten, die aktive Parameterzahl auf den jeweils genutzten Expertenanteil pro Inferenzdurchlauf [44].

² Google DeepMind hat seinen Hauptsitz in London, gehört jedoch zu Alphabet (USA). Wo genau trainiert wurde, ist unklar.

2411 Modell [30] ist mit 123 B Parametern deutlich größer und bietet ein Kontextfenster von 128,000 Tokens. Es wird unter der Mistral Research License veröffentlicht, die die Nutzung auf nicht-kommerzielle Forschung beschränkt [43]. Das Modell **Mistral-Medium-3.1** [45] bietet ein Kontextfenster von 128,000 Tokens und gilt als aktuelles Spitzenmodell der Mistral-Modellreihe. Anders als die übrigen Mistral-Modelle ist es proprietär und wird von Mistral AI auf EU-Servern unter Beachtung der DSGVO betrieben [42, 46]. Damit ist die Verarbeitung sensibler Daten möglich. Das Modell eignet sich daher - trotz nicht veröffentlichter Gewichte – für den Einsatz in datenschutzkritischen Szenarien.

Die Gemma-3-Modelle von Google Deepmind repräsentieren eine neue Generation multimodaler LLMs mit offenen Gewichten. Die hier betrachteten Varianten sind **Gemma-3-12B-it** [26] mit 12,2 B Parametern und **Gemma-3-27B-it** [27] mit 27,4 B Parametern. Beide Modelle verfügen über ein großes Kontextfenster von 128,000 Tokens und sind unter der proprietären Gemma-Lizenz veröffentlicht, die eine breite kommerzielle Nutzung erlaubt [21]. Die genaue Herkunft der Modelle ist unklar, da Google DeepMind seinen Hauptsitz in Großbritannien hat, jedoch zu Alphabet in den USA gehört. Wo genau die Modelle trainiert wurden, ist nicht bekannt.

Die Qwen-Modelle wurden von Alibaba Cloud in China entwickelt worden. Das kleinere Modell **Qwen2.5-7B-Instruct** [33] hat 7,62 B Parameter, ein Kontextfenster von 131,072 Tokens und ist unter der Apache-2.0-Lizenz frei verfügbar. Das größere Modell **Qwen3-235B-A22B-Thinking-2507** [32] verfügt über 235 B Parameter, ein Kontextfenster von 256,000 Tokens und ist ebenfalls unter der Apache-2.0-Lizenz veröffentlicht.

Das chinesische Unternehmen DeepSeek AI hat mit **DeepSeek-R1-Distill-Qwen-14B** [23] ein Modell mit 14,8 B Parametern veröffentlicht, das auf Qwen-2.5 basiert, ein Kontextfenster von 131,072 Tokens bietet und unter der MIT-Lizenz frei verfügbar ist. Das größere Modell **DeepSeek-V3.1** [25] setzt auf eine Mixture-of-Experts-Architektur mit insgesamt 671 B Parametern, von denen pro Token 37 B aktiv sind. Es bietet ein Kontextfenster von 128,000 Tokens und ist ebenfalls MIT-lizenziert. Besonders bemerkenswert sind die DeepSeek-Modelle, weil DeepSeek im Januar 2025 mit *DeepSeek-R1-Zero* [24] eines der ersten permissiv lizenzierten Reasoning-Modelle in OpenAI-Größenordnung vorlegte und zugleich einen Trainings-Ansatz etablierte, bei dem LLM-Reasoning nahezu ausschließlich über Reinforcement Learning erlernt wird [15].

GPT-4o [55] ist ein proprietäres Modell von OpenAI mit einem Kontextfenster von 128,000 Tokens. Es wurde am 20. November 2024 veröffentlicht und ist das einzige internationale proprietäre Modell in dieser Arbeit. Die genauen Parameterzahlen sind nicht bekannt. GPT-4o wird über die OpenAI-API bereitgestellt. GPT-4o ist das erste Omni-Modell von OpenAI, das neben Text auch Bilder als Eingabe akzeptieren kann und ist der De-facto-Standard in der Industrie. *Ich weiß das von meiner Arbeit, aber brauche ich für diese Aussage auch eine Quelle?*. Das Modell dient in diesem Vergleich als Referenzpunkt für den aktuellen Stand der Technik. Mit den GPT-OSS Modellen [54] hat OpenAI zudem zwei Modelle mit offenen Gewichten und unter Apache-2.0-Lizenz veröffentlicht, die explizit für Forschung und kommerzielle Nutzung freigegeben sind. Das **GPT-OSS-20B** Modell hat 20,91 B Parameter (3,61 B aktiv) und ein Kontextfenster von 131,072 Tokens. Das größere **GPT-OSS-120B** Modell verfügt über 116,83 B Parameter (5,13 B aktiv) und das gleiche Kontextfenster. Sie sind spannende Vergleichsmodelle, da sie von einem der führenden LLM-Anbieter stammen und dennoch offen verfügbar sind.

Insgesamt deckt die Modellauswahl in dieser Arbeit eine breite Palette von Modellgrößen, Architekturen und Lizenztypen ab. Die Mistral-Modelle repräsentieren die EU-Modelle mit offenen Gewichten, während die Gemma-, Qwen- und GPT-OSS-Modelle internationale Alternativen aus verschiedenen Herkunftsländern darstellen. Die DeepSeek-Modelle bieten innovative Ansätze im Reasoning-Bereich, und GPT-4o dient als aktueller Industriestandard. Diese Vielfalt ermöglicht eine umfassende Evaluation der Modelle hinsichtlich ihrer Eignung für die Klassifizierungsaufgabe von BPMN-Modellen. Im nächsten Kapitel wird aufbauend auf den vorgestellten Modellen der Versuchsaufbau und die Durchführung der Experimente beschrieben.

8 Versuchsaufbau und Durchführung

Wie im Abschnitt 3.4 beschrieben, soll eine fairer Vergleich verschiedener LLMs erreicht werden. Dazu werden alle, der im vorherigen Kapitel beschriebenen, Modelle durch dieselbe Klassifikationspipeline geschickt und anhand der im Kapitel 3.2 definierten Metriken (Accuracy, Precision, Recall, F1 sowie die erfolgreiche klassifizierte Testfälle) bewertet. Dieses Kapitel beschreibt den konkreten Versuchsaufbau und die Durchführung der Experimente. Die hier dokumentierten Parameter und Konfigurationen sind wesentlich, um die Ergebnisse nachvollziehbar und reproduzierbar zu machen.

Um sowohl kleine als auch große Modelle testen zu können, wurde *OpenRouter* [59] als API-Anbieter genutzt. Über diese Cloud-basierte Schnittstelle lassen sich auch Modelle ausführen, die lokal aufgrund begrenzter Hardware nicht betrieben werden können. Der API-Schlüssel wird über eine Umgebungsvariable in die Konfigurationsdatei eingebunden, um sensible Daten aus den Konfigurationen fernzuhalten.

In den Experimenten wurden mehrere Modelle aus unterschiedlichen Anbieterfamilien. Für jeden Anbieter gibt es ein eigenes Experiment, in dem mehrere Modellgrößen (z. B. 7B, 8x7B, Large) gegeneinander verglichen werden. Da alle Experimente die gleiche Pipeline und die gleichen Datensätze verwenden, können auch die Ergebnisse verschiedener Anbieter untereinander verglichen werden. Diese Aufteilung in verschiedene Experimente dient lediglich der Übersichtlichkeit in der Benutzeroberfläche des Evaluationsframeworks.

8.1 Einheitliche Klassifizierungspipeline und Datensätze

Um die Vergleichbarkeit der Experimente zu gewährleisten, werden alle Modelle durch dieselbe Klassifikationspipeline geschickt. Die technische Implementierung dieser Pipeline wurde in Kapitel 4 beschrieben und kann im Evaluationsframework genutzt werden. Jeder Testfall besteht aus einem BPMN-Prozess mit Labeln für DSGVO-kritische Aktivitäten. Ein Testfall gilt als korrekt klassifiziert, wenn genau die als kritisch gelabelten Aktivitäten auch als kritisch erkannt werden - bereits ein FP oder FN führt zu einem nicht bestandenen Testfall.

Als Datenbasis kommen drei im Labeling-Tool erzeugte Testdatensätze zum Einsatz. Diese decken unterschiedliche Prozesskontexte ab und werden in den Experimenten mit den ids 1 (kleine Prozesse), 2 (Universität) und 7 (reale größere Prozesse) referenziert. Für jedes Experiment werden alle verfügbaren Datensätze verwendet. Auf diese Weise können Unterschiede zwischen den Modellen nicht auf unterschiedliche Datenquellen zurückgeführt werden.

8.2 Konfigurationen

Die Konfigurationen der Experimente sind im YAML-Format in Listings 8.1, A.6, A.7, A.8 und A.9 dargestellt. Sie enthalten die zu evaluierenden Modelle, die zu verwendenden Datensätze, den Basis-Seed sowie die Anzahl der Wiederholungen und weitere Rahmenparameter. In Listing 8.1 wird ein Experiment dargestellt, in dem vier verschiedene Mistral-Modelle über OpenRouter evaluiert werden. Die Datensätze werden jeweils fünf Mal durchlaufen. Der Basis-Seed ist auf 24523833 gesetzt.

Listing 8.1: Konfigurationsdatei des Experiments mit Mistral Modellen

```
1 defaultEvaluationEndpoint: /gdpr/analysis/prompt-engineering
2 seed: 24523833
3 maxConcurrent: 10
4 repetitions: 5
5 models:
```

```
6 - label: Mistral-7B-Instruct-v0.3
7   llmProps:
8     baseUrl: https://openrouter.ai/api/v1
9     modelName: mistralai/mistral-7b-instruct-v0.3
10    apiKey: ${OPEN_ROUTER_API_KEY}
11    temperature: 0.1
12    topP: 1
13 - label: Mistral-8x7B-Instruct-v0.1
14   llmProps:
15     baseUrl: https://openrouter.ai/api/v1
16     modelName: mistralai/mixtral-8x7b-instruct
17     apiKey: ${OPEN_ROUTER_API_KEY}
18     temperature: 0.1
19     topP: 1
20 - label: Mistral-Large-Instruct-2411
21   llmProps:
22     baseUrl: https://openrouter.ai/api/v1
23     modelName: mistralai/mistral-large-2411
24     apiKey: ${OPEN_ROUTER_API_KEY}
25     temperature: 0.1
26     topP: 1
27 - label: Mistral Medium 3.1
28   llmProps:
29     baseUrl: https://openrouter.ai/api/v1
30     modelName: mistralai/mistral-medium-3.1
31     apiKey: ${OPEN_ROUTER_API_KEY}
32     temperature: 0.1
33     topP: 1
34 datasets:
35   - 2
36   - 7
37   - 1
```

Auf Basis des Seeds aus der Konfiguration und der aktuellen Wiederholungsnummer wird in dem Evaluationsframework für jede Wiederholung deterministisch ein neuer Seed generiert. Dadurch sind die Ergebnisse reproduzierbar und dennoch wird die Stabilität der Modelle über mehrere Wiederholungen mit unterschiedlichen

Seeds abgebildet.

Alle Datensätze, Konfigurationen und die daraus resultierenden Ergebnisse sind außerdem im GitLab-Repository verfügbar¹.

Um bei der Zero-Shot-Klassifikation deterministische und formatkonsistente Ergebnisse zu erzielen, wurden die Inferenz-Hyperparameter `temperature` und `topP` bewusst konservativ gewählt. Der Parameter `temperature` steuert die Zufälligkeit der Modellausgabe: niedrige Werte priorisieren die wahrscheinlichsten Tokens und machen die Ausgabe deterministischer. Für Aufgaben, die faktische Genauigkeit und Präzision erfordern, empfehlen aktuelle Arbeiten daher sehr niedrige Temperaturen; höhere Werte (z. B. `temperature=0,8` oder `2`) verschlechtern hingegen die Klassifikationsleistung und führen zu nicht-reproduzierbaren Ausgaben [64, 47]. In dieser Arbeit wird daher konsequent `temperature=0,1` verwendet. Dieser Wert reduziert Zufallseffekte erheblich, ohne den Output zu stark einzuschränken, und entspricht den Empfehlungen vergleichbarer Studien zur Zero-Shot-Klassifikation [47].

Der Parameter `topP` legt fest, bis zu welcher kumulierten Wahrscheinlichkeit Tokens für die nächste Auswahl herangezogen werden. Durch die Wahl `topP=1` werden keine Tokens vorab ausgeschlossen, sodass alle möglichen Tokens des Vokabulars berücksichtigt werden und allein die `temperature` den Grad der Stochastik bestimmt [64]. In Kombination mit einer sehr niedrigen `temperature` ermöglicht `topP=1` einen fokussierten und weitgehend deterministischen Output bei gleichzeitig maximalem Stichprobenraum [47].

8.3 Durchführung

Die Durchführung der Experimente erfolgt automatisiert über das Evaluationsframework. Für jede in der Konfigurationsdatei angegebene Modellvariante werden alle Testfälle aus den ausgewählten Datensätzen an die Klassifikationspipeline übergeben. Während der Ausführung werden für jeden Testfall die Einzelergebnisse der Konfusionsmatrix sowie der Status „bestanden“ oder „nicht bestanden“ bestimmt.

¹ Siehe das GitLab Repository: // TODO

Diese Kennzahlen werden pro Modell aggregiert und anschließend genutzt, um die aus Kapitel 3.2 bekannten Metriken zu berechnen.

Für jedes Modell werden außerdem über alle Wiederholungen hinweg sowohl die Durchschnittswerte als auch dessen Standardabweichung für die Metriken berechnet. Die Standardabweichung gibt an, wie stark die Ergebnisse der einzelnen Läufe um den Mittelwert streuen. Ein niedriger Wert deutet auf eine hohe Stabilität des Modells hin, während ein hoher Wert auf eine größere Variabilität in den Ergebnissen hinweist. Diese Information ist besonders wichtig, um die Zuverlässigkeit der Modelle zu bewerten, da einige LLMs aufgrund ihrer nicht-deterministischen Natur unterschiedliche Ergebnisse bei wiederholten Ausführungen desselben Testfalls liefern können.

9 Ergebnisse

In diesem Kapitel werden die Resultate der durchgeführten Experimente präsentiert und im Kontext der Forschungsfragen aus Abschnitt 1.2 analysiert. Alle aufgeführten Kennzahlen beziehen sich auf den Durchschnitt über fünf unabhängige Läufe mit unterschiedlichen Seeds. Die Standardabweichungen dienen der Abschätzung der Robustheit der LLMs und sind in allen Tabellen und Abbildungen angegeben.

// TODO: Ich muss noch auf die Zeilwerte eingehen und aufzeigen welches Modell die Zielwerte erreicht hat.

// TODO: Noch hinzufügen, wie viele Testfälle pro Modell im durchschnitt + Standardabweichung bestanden wurden.

9.1 Überblick

Abbildung 9.1 zeigt für jedes der untersuchten Modelle die durchschnittlichen Metrik-Werte über alle fünf Wiederholungen hinweg inklusive Standardabweichung. Für einen besseren Vergleich sind die proprietären Modelle rot, die kleineren Modelle orange und die größeren Modelle blau dargestellt. Diese Einteilung entspricht der in Kapitel 7 beschriebenen Kategorisierung. Die Diagramme aus diesem Kapitel stammen nicht direkt aus dem Evaluationsframework, sondern wurden mit den zusammengeführten Ergebnissen aller Experimente erstellt, damit alle Modelle auf einen Blick verglichen werden können. Im Folgenden wird zuerst ein Überblick über die Ergebnisse gegeben, bevor im Anschluss eine detaillierte Analyse erfolgt.

Positiv aufgefallen ist, dass neun von dreizehn Modellen das Qualitätsziel eines F1-Scores von $\geq 0,80$ erreichen konnten - darunter auch kleine Modelle. Über alle Metriken hinweg sind Qwen3-235B-A22B-Thinking-2507, GPT-0SS-120B und GPT-0SS-20B ganz vorne mit dabei. Auffällig ist das schwache Abschneiden des

9 Ergebnisse

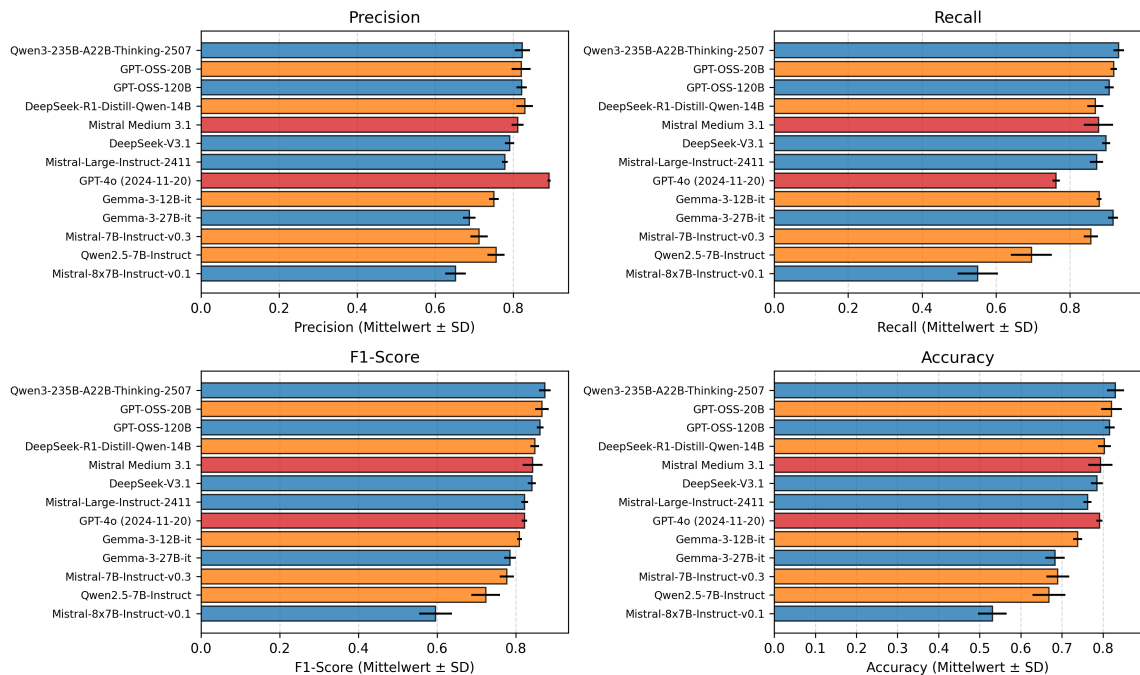


Abbildung 9.1: Durchschnittliche Metrik-Werte der untersuchten Modelle über alle Wiederholungen hinweg inklusive Standardabweichung.

europäischen Mixtral-8x7B-Instruct-v0.1-Modells, das sowohl in Precision, als auch Recall weit zurückfällt und als einziges Modell einen F1-Score von $\leq 0,60$ erreicht.

Die Abbildung 9.1 macht deutlich, dass einige Modelle – etwa GPT-4o oder Qwen-2.5-7B-Instruct – hohe Präzisionswerte aufweisen, aber im Recall zurückliegen. Anders sieht es bei Gemma-3-27B-it aus, das einen sehr guten Recall erreicht, aber bei der Präzision auf dem vorletzten Platz liegt. Modelle wie Qwen3-235B-A22B und GPT-OSS-20B erreichen einen ausgezeichneten Recall bei gleichzeitig hoher Präzision.

Abbildung 9.2 zeigt die Robustheit der Modelle gemessen an der Standardabweichung des F1-Scores über alle Wiederholungen hinweg. Hier zeigt sich eine große Varianz: Während einige Modelle wie Gemma-3-12B.it und Mistral-Large-Instruct-2411 eine sehr geringe Standardabweichung von $\leq 0,01$ aufweisen, zeigen andere Modelle wie Mixtral-8x7B-Instruct-v0.1 und Qwen2.5-7B-Instruct eine hohe Varianz von $\geq 0,03$ bis $\geq 0,04$ im F1-Score. Dies deutet darauf hin, dass die Leistung dieser Modelle stark von der Wahl des

9 Ergebnisse

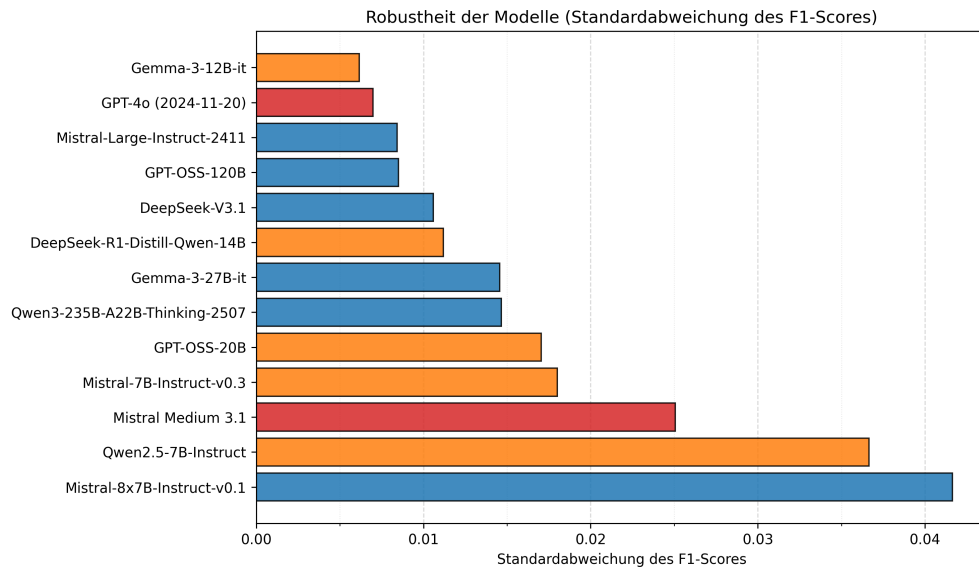


Abbildung 9.2: Robustheit der Modelle gemessen an der Standardabweichung des F1-Scores über alle Wiederholungen hinweg.

Seeds abhängen und ihre Leistung weniger stabil ist.

Neben den Diagrammen 9.1 und 9.2 liefert Tabelle 9.1 eine tabellarische Übersicht der konkreten durchschnittlichen Metrik-Werte inklusive Standardabweichungen für alle Modelle über alle fünf Wiederholungen hinweg. Diese Werte werden genutzt um im Folgenden die Leistungen der Modelle detailliert aufzuzeigen.

Die proprietären Modelle GPT-4o und Mistral Medium 3.1 erreichen mit 0,822 bzw. 0,843 im Vergleichsfeld einen mittleren F1-Score. Mistral Medium 3.1 liegt mit einem F1-Wert von 0,843 leicht über GPT-4o und weist zugleich einen höheren Recall von 0,877 auf. GPT-4o fällt vor allem durch eine sehr hohe Precision von 0,892 auf, die jedoch einem vergleichsweise niedrigem Recall von 0,762 gegenübersteht. Dies deutet darauf hin, dass GPT-4o eher konservativ kritische Aktivitäten klassifiziert und somit weniger FP, aber auch mehr FN produziert.

Die Mehrzahl der großen Modelle steht in den meisten Metriken vor den kleineren Modellen, auch wenn teilweise nur knapp. Ein bemerkenswerter Befund ist jedoch die starke Leistung von GPT-OSS-20B, das mit Blick auf die F1-, Recall- und Accuracy-Werten von 0,866, 0,918 bzw. 0,821 fast mit den besten Modellen mithalten kann und dort sogar besser abschneidet als das größere GPT-OSS-120B. Nur in der Precision liegt GPT-OSS-120B mit 0,821 knapp vor GPT-OSS-20B. Allerdings

9 Ergebnisse

Tabelle 9.1: Aggregierte Mittelwerte und Standardabweichungen der Evaluationsmetriken über alle fünf Wiederholungen hinweg.

Modell	Precision	Recall	F1-Score	Accuracy
DeepSeek-V3.1	0.791 ± 0.012	0.897 ± 0.011	0.841 ± 0.011	0.785 ± 0.015
DeepSeek-R1-Distill-Qwen-14B	0.829 ± 0.021	0.868 ± 0.022	0.848 ± 0.011	0.803 ± 0.016
Gemma-3-12B-it	0.751 ± 0.013	0.879 ± 0.006	0.810 ± 0.006	0.738 ± 0.011
Gemma-3-27B-it	0.687 ± 0.016	0.916 ± 0.014	0.785 ± 0.015	0.683 ± 0.023
Mistral-7B-Instruct-v0.3	0.712 ± 0.022	0.856 ± 0.019	0.777 ± 0.018	0.690 ± 0.028
Mixtral-8x7B-Instruct-v0.1	0.652 ± 0.027	0.550 ± 0.054	0.596 ± 0.042	0.531 ± 0.035
Mistral-Large-Instruct-2411	0.779 ± 0.008	0.872 ± 0.018	0.823 ± 0.008	0.762 ± 0.010
Mistral Medium 3.1	0.811 ± 0.015	0.877 ± 0.040	0.843 ± 0.025	0.794 ± 0.029
GPT-OSS-20B	0.820 ± 0.024	0.918 ± 0.009	0.866 ± 0.017	0.821 ± 0.025
GPT-OSS-120B	0.822 ± 0.013	0.906 ± 0.012	0.862 ± 0.009	0.816 ± 0.012
GPT-4o	0.892 ± 0.004	0.762 ± 0.010	0.822 ± 0.007	0.791 ± 0.007
Qwen2.5-7B-Instruct	0.756 ± 0.022	0.696 ± 0.055	0.724 ± 0.037	0.668 ± 0.040
Qwen3-235B-A22B-Thinking-2507	0.824 ± 0.019	0.932 ± 0.014	0.874 ± 0.015	0.830 ± 0.021

deuten die Standardabweichungen auf eine geringere Stabilität von GPT-OSS-20B hin. Ein weiteres starkes klines Open-Source-Modell ist DeepSeek-R1-Distill-Qwen-14B, das in F1-Score und Accuracy die beiden proprietären Modelle übertrifft und in der Precision sogar über Qwen3-235B-A22B-Thinking-2507 und die GPT-OSS-Modelle hinausgeht.

Die Ergebnisse unterstreichen, dass Qwen3-235B-A22B-Thinking-2507, Gemma-3-27B-it und GPT-OSS-20B mit $\geq 0,91$ die höchsten Recall Werte erzielen. Dies bedeutet, dass diese Modelle die meisten kritischen Aktivitäten korrekt identifizieren und somit das Risiko von FN minimieren. Besonders hervorzuheben ist Qwen3-235B-A22B-Thinking-2507, das mit einem Recall von 0,932 und einer Precision von 0,824 eine ausgezeichnete Balance zwischen Sensitivität und Genauigkeit bietet. Mit Abstand am schwächsten im Recall schneiden Mixtral-8x7B-Instruct-v0.1 und Qwen2.5-7B-Instruct ab, die mit 0,550 bzw. 0,696 weit hinter den anderen Modellen zurückbleiben.

Mixtral-8x7B-Instruct-v0.1 schnitt insgesamt am schlechtesten ab. Sowohl die Precision von 0,652 als auch der Recall von 0,55 und die Accuracy von 0,531 sind klar unter den Werten der anderen Modelle. Besonders bemerkenswert ist, dass das kleinere Mistral-7B-Instruct-v0.3 – mit achtmal weniger Parametern – deutlich bessere Ergebnisse erzielt. Auch gegenüber dem ähnlich großen Qwen2.5-7B-Instruct ist es im F1-Score und in der Accuracy leicht voraus und im Recall mit einer Differenz von 0,106 schneidet Mistral-7B-Instruct-v0.3

deutlich besser ab. Dies deutet darauf hin, dass `Mistral-8x7B-Instruct-v0.1` für die hier betrachtete Klassifikationsaufgabe ungeeignet ist.

Das große europäische Modell `Mistral-Large-Instruct-2411` und das große chinesische Modell `DeepSeek-V3.1` erreichen mit 0,823 bzw. 0,841 im F1-Score einen mittleren Platz im Vergleichsfeld. Beide Modelle liegen in allen Metriken auf einem ähnlichen Niveau, wobei `DeepSeek-V3.1` in jeder Metrik etwas bessere Werte erzielt. Mit Blick auf die Standardabweichungen sind beide Modelle robust und zeigen eine geringe Varianz über die Wiederholungen hinweg. Obwohl `Mistral-Large-Instruct-2411` das schwächste der großen Modelle ist, liegt sein F1-Score immer noch über dem des Referenzmodells `GPT-4o`.

Die beiden Varianten der Gemma-Reihe erreichen F1-Scores knapp unterhalb der Benchmark-Modelle `GPT-4o` und `Mistral Medium 3.1`, sind mit 12B beziehungsweise 27B Parametern jedoch deutlich kleiner. `Gemma-3-12B-it` erreicht mit 0,810 einen soliden F1-Score, während `Gemma-3-27B-it` mit 0,785 etwas darunter liegt. Beide Modelle zeigen eine gute Balance zwischen Precision und Recall, wobei `Gemma-3-27B-it` mit einem Recall von 0,916 besonders stark in der Identifikation kritischer Aktivitäten ist. Allerdings leidet die Precision mit 0,687 darunter, was auf eine höhere Rate an FP hindeutet.

Zusammenfassend zeigen die Ergebnisse, dass sowohl große als auch kleinere LLMs in der Lage sind, die datenschutzrechtliche Klassifikation von Prozessaktivitäten mit hoher Genauigkeit durchzuführen. Besonders hervorzuheben sind die Modelle `Qwen3-235B-A22B-Thinking-2507`, `GPT-05S-20B` und `DeepSeek-R1-Distill-Qwen-14B`, die in mehreren Metriken Spitzenwerte erzielen. Allerdings gibt es auch Modelle wie `Mistral-8x7B-Instruct-v0.1`, die für diese Aufgabe ungeeignet sind.

9.2 Analyse

Im Folgenden werden die Ergebnisse nach Modellkategorien aufgeschlüsselt und analysiert. Zunächst werden die Leistungen der Open-Weight-Modelle mit denen der proprietären Modelle verglichen. Anschließend werden die Modelle nach ihrer Größe in kleine und große Modelle unterteilt und deren Leistungen gegenüberge-

stellt. Abschließend wird die Leistung der Modelle mit europäischem Bezug mit der der internationalen Modelle verglichen.

Proprietäre vs. Open-Weight Modelle

Die beiden proprietären Modelle GPT-4o und Mistral Medium 3.1 zeigen im direkten Vergleich mittelmäßige Leistungen. Mistral Medium 3.1 erreicht einen F1-Score von 0,843 und übertrifft damit das Referenzmodell GPT-4o mit 0,822 leicht. Der höhere Recall von Mistral Medium 3.1 von 0,877 deutet darauf hin, dass dieses Modell sensiblen Aktivitäten eher als kritisch klassifiziert, während GPT-4o durch eine sehr hohe Precision von 0,892 bei gleichzeitig niedrigem Recall von 0,762 eher konservativ klassifiziert.

Die Open-Weight-Modelle präsentieren ein heterogenes Bild, wobei mehrere Modelle die proprietären Vertreter deutlich übertreffen. So erzielt Qwen3-235B-A22B-Thinking-2507 als bestes Modell einen F1-Score von 0,874 bei gleichzeitig höchstem Recall von 0,932 und einer Precision von 0,824. Auch weitere kleinere offene Modelle wie GPT-05S-20B mit einem F1-Score von 0,866 und einem Recall von 0,918 DeepSeek-R1-Distill-Qwen-14B mit einem F1-Score von 0,848 liegen beim F1-Score vor den proprietären Modellen. Die hohe Precision von DeepSeek-R1-Distill-Qwen-14B von 0,829 und der hohe Recall von Gemma-3-27B-it von 0,879 zeigen außerdem, dass Open-Weight-Modelle sowohl bei Precision als auch bei Recall überzeugen können.

Allerdings zeigt sich bei den offenen Modellen eine starke Varianz. Das europäische Mixtral-8x7B-Instruct-v0.1 weist mit einem F1-Score von 0,596, einer Precision von 0,652 und einem Recall von lediglich 0,550 die bei weitem schwächsten Ergebnisse auf. Insgesamt lässt sich festhalten, dass mehrere offene Modelle die proprietären Lösungen klar übertreffen, wobei die Leistungsunterschiede innerhalb der Open-Source-Kategorie sehr groß sind. Im nächsten Abschnitt werden die Modelle nach ihrer Größe kategorisiert und verglichen, um mögliche Zusammenhänge zwischen Modellgröße und Klassifizierungsleistung zu untersuchen.

Kleine vs. große Modelle

Die Gegenüberstellung kleiner Modelle mit $\leq 25B$ und großer Modelle mit $> 25B$ Parametern in Tabelle 9.2 zeigt nahezu keinen Unterschied im durchschnittlichen F1-Score mit 0,805 vs. 0,806. Wird jedoch der durchschnittliche F1-Score der großen Modelle ohne das Ausreißermodell Mixtral-8x7B-Instruct-v0.1 betrachtet, ergibt sich mit 0,836 ein deutlich höherer Wert. Dies deutet darauf hin, dass größere Modelle tendenziell bessere Leistungen erbringen können, aber auch anfälliger für Leistungseinbußen sind, wenn sie nicht optimal auf die Aufgabe abgestimmt sind. Auch Precision und Accuracy sind vergleichbar mit ein wenig besseren Werten bei den großen Modellen. Beim Recall zeigen die kleinen Modelle mit 0,843 sogar einen leicht höheren Wert als die großen Modelle mit 0,839. Allerdings ist die Standardabweichung bei den großen Modellen mit 0,089 deutlich höher als bei den kleinen Modellen mit 0,057, was auf eine größere Varianz in der Leistung der großen Modelle hinweist. Diese Varianz wird ebenfalls vor allem durch das Ausreißermodell Mixtral-8x7B-Instruct-v0.1 verursacht, das mit einem Recall von nur 0,550 deutlich schlechter abschneidet als die anderen großen Modelle.

Tabelle 9.2: Kleine vs. große Modelle: Mittelwerte je Gruppe und bestes Modell.

Metrik	Klein ($\leq 25B$)	Groß ($> 25B$)
Anzahl Modelle ¹	5	8
Ø F1-Score \pm SD ²	0.805 \pm 0.057	0.806 \pm 0.089
Ø Precision \pm SD	0.774 \pm 0.050	0.779 \pm 0.085
Ø Recall \pm SD	0.843 \pm 0.086	0.839 \pm 0.128
Ø Accuracy \pm SD	0.744 \pm 0.067	0.749 \pm 0.099
Bester F1-Score	0.866	0.874
Bestes Modell (F1-Score)	GPT-OSS-20B	Qwen3-235B-A22B-Thinking-2507
Bester Precision	0.829	0.892
Bestes Modell (Precision)	DeepSeek-R1-Distill-Qwen-14B	GPT-4o
Bester Recall	0.918	0.932
Bestes Modell (Recall)	GPT-OSS-20B	Qwen3-235B-A22B-Thinking-2507
Beste Accuracy	0.821	0.830
Bestes Modell (Accuracy)	GPT-OSS-20B	Qwen3-235B-A22B-Thinking-2507

¹ Einteilung nach gesamten Milliarden Parametern bei MoE. Die Proprietären Modelle GPT-4o und Mistral Medium 3.1 wurden trotz fehlender Parameterangabe als große Modelle eingeordnet.

² Ohne Mixtral-8x7B-Instruct-v0.1 beträgt der Durchschnitt der großen Modelle \pm SD 0.836 \pm 0.029.

Das beste kleine Modell GPT-OSS-20B erreicht mit 0,866 einen F1-Score, der nur geringfügig unter dem besten großen Modell Qwen3-235B-A22B-Thinking-2507 mit 0,874 liegt. Außerdem hat GPT-OSS-20B mit einem Recall von 0,918 eben-

falls nur einen minimal kleineren Wert als Qwen3-235B-A22B-Thinking-2507 mit 0,932 Insgesamt zeigen die Ergebnisse, dass die Modellgröße allein kein ausschlaggebender Faktor für die Klassifizierungsleistung ist. Vielmehr spielen die Trainingsdaten, die Feinabstimmung und die Architektur eine entscheidende Rolle. Die richtigen kleinen Modelle können mit den großen Modellen durchaus mithalten, was insbesondere für den praktischen Einsatz relevant ist, da kleinere Modelle oft ressourcenschonender und kostengünstiger betrieben werden können.

Europäische versus internationale Modelle

Die Betrachtung der Herkunft zeigt, dass die europäischen Mistral-Modelle unterschiedlich stark abschneiden. Mistral Medium 3.1 erreicht als bestes EU-Modell einen F1-Score von 0,843 und einen Recall von 0,877 und hat damit sogar besser abgeschnitten als das internationale Benchmark-Modell GPT-4o, das einen F1-Score von 0,822 und einen Recall von 0,762 erreichte. Das größte europäische Open-Weight-Modell Mistral-Large-Instruct-2411 liegt mit einem F1-Score von 0,823 und einer sehr geringen Standardabweichung ebenfalls im Mittelfeld. Deutlich schwächer fallen Mistral-7B-Instruct-v0.3 mit einem F1-Score von 0,777 und insbesondere Mixtral-8x7B-Instruct-v0.1 mit einem F1-Score von 0,596 aus. Unter Berücksichtigung der Größenklassen sind die europäischen Modelle überwiegend in der unteren Ranglistenhälfte und bestenfalls im Mittelfeld vertreten. Die große Streuung der Ergebnisse verdeutlicht, dass die EU-Modelle noch nicht durchgängig zur Spitzengruppe aufschließen und teilweise erhebliches Optimierungspotenzial aufweisen.

Die internationalen Modelle – hierzu zählen insbesondere die chinesischen Qwen- und DeepSeek-Modelle sowie die US-amerikanischen GPT-Modelle und die Gemma-Modelle – liegen auf der Rangliste tendenziell weiter vorne als die europäischen Mistral Modelle. Qwen3-235B-A22B-Thinking-2507 erzielt mit einem F1-Score von 0,874 und einem Recall von 0,932 die besten Werte im gesamten Vergleich. Auch das kleinere DeepSeek-R1-Distill-Qwen-14B mit einem F1-Score von 0,848 und GPT-05S-20B mit einem F1-Score von 0,866 liefern Ergebnisse, die über den europäischen Modellen liegen. Lediglich Qwen2.5-7B-instruct mit einem F1-Score von 0,724 schneidet schlechter ab, als das gleich große europäische Mistral-7B-Instruct-v0.3 mit einem F1-Score von 0,777. Die starke Perfor-

mance dieser internationalen Open-Weight-Modelle legt nahe, dass sie umfangreiche und vielfältige Trainingsdaten sowie ausgereifte Modellarchitekturen nutzen.

Gleichzeitig haben die EU-Modelle den Vorteil, dass sie von Anbietern stammen, die strenger den europäischen Datenschutzbestimmungen unterliegen und sich daher besser für datenschutzsensible Szenarien eignen. Modelle wie Mistral Medium 3.1 zeigen, dass leistungsfähige europäische Alternativen existieren, doch im Vergleich zu internationalen Spitzenmodellen besteht weiterhin eine Lücke. Eine tabellarische Übersicht der aggregierten Kennzahlen für beide Gruppen findet sich in Tabelle 9.3.

Tabelle 9.3: Europäische vs. internationale Modelle: Mittelwerte je Gruppe und bestes Modell.

Metrik	EU-Modelle	Internationale Modelle
Anzahl Modelle	4	9
Ø F1-Score \pm SD	0,760 \pm 0,098	0,826 \pm 0,045
Ø Precision \pm SD	0,738 \pm 0,061	0,797 \pm 0,056
Ø Recall \pm SD	0,789 \pm 0,138	0,864 \pm 0,076
Ø Accuracy \pm SD	0,694 \pm 0,101	0,771 \pm 0,057
Bester F1-Score	0,843	0,874
Bestes Modell (F1-Score)	Mistral Medium 3.1	Qwen3-235B-A22B-Thinking-2507
Bester Precision	0,811	0,892
Bestes Modell (Precision)	Mistral Medium 3.1	GPT-4o
Bester Recall	0,877	0,932
Bestes Modell (Recall)	Mistral Medium 3.1	Qwen3-235B-A22B-Thinking-2507
Beste Accuracy	0,794	0,830
Bestes Modell (Accuracy)	Mistral Medium 3.1	Qwen3-235B-A22B-Thinking-2507

Die EU-Modelle umfassen Mistral-7B-Instruct-v0.3, Mixtral-8x7B-Instruct-v0.1, Mistral-Large-Instruct-2411 und Mistral Medium 3.1. Die internationalen Modelle sind die übrigen in Kapitel 9.1 betrachteten Modelle.

Die vorangegangene Analyse hat die Leistungsunterschiede zwischen proprietären und offenen Modellen sowie zwischen europäischen und internationalen Modellen aufgezeigt. Besonders Modelle wie Qwen3-235B-A22B-Thinking-2507 und GPT-05S-20B haben hervorragend abgeschnitten, während andere Modelle klare Schwächen offenbaren. Im nächsten Abschnitt werden die Modelle auf ihre Robustheit hin untersucht, um deren Stabilität und Zuverlässigkeit im praktischen Einsatz zu bewerten.

9.3 Robustheit

Die Robustheit der Modelle wurde anhand zweier Kriterien bewertet: der Varianz der F1-Scores über unterschiedliche Seeds und der Anzahl der Retries, die nötig waren, um eine formatkorrekte JSON-Antwort von den LLMs in der Klassifizierungspipeline zu erhalten. Beide Größen geben Aufschluss darüber, wie stabil ein Modell im produktiven Einsatz ist.

Abbildung 9.2 zeigt die Standardabweichungen der F1-Scores über fünf unabhängige Läufe mit unterschiedlichen Seeds. Die Mehrzahl der Modelle weisen Werte von deutlich unter 0,02 auf. Sie liefern damit weitgehend gleiche Ergebnisse, unabhängig vom gewählten Seed, und gelten als stabil. Höhere Streuungen finden sich hingegen bei Mistral Medium 3.1, Qwen2.5-7B-Instruct und insbesondere Mixtral-8x7B-Instruct-v0.1. Mit Standardabweichungen zwischen etwa 0,025 und knapp über 0,04 reagieren diese Modelle spürbar sensibler auf eine Änderung vom Seed. Ihre Leistung kann daher zwischen zwei Wiederholungen stärker schwanken.

Neben der Varianz des F1-Scores wurde die Zuverlässigkeit beim Einhalten des vorgegebenen Ausgabeformats betrachtet. Abbildung 9.3 stellt die durchschnittliche Anzahl der Retries über alle Testfälle hinweg dar, die notwendig waren, bis eine gültige JSON-Antwort von den Modellen zurückgegeben wurde. Hier zeigen sich deutliche Unterschiede: Die meisten Modelle lieferten schon beim ersten Aufruf oder nach maximal einem weiteren Versuch eine gültige JSON-Struktur. Besonders positiv fallen DeepSeek-R1-Distill-Qwen-14B, GPT-4o, Gemma-3-12B-it und Mistral-Large-Instruct-2411 auf, die über alle Testfälle hinweg keinen einzigen Retry benötigten. Am anderen Ende des Spektrums steht Mistral-7B-Instruct-v0.3, das im Mittel rund 15,8 zusätzliche Anfragen brauchte, bis für alle 25 Testfälle eine formatkorrekte Antwort vorlag; das entspricht im Durchschnitt 0,632 Retries pro Testfall. Das deutet darauf hin, dass dieses Modell Schwierigkeiten hat, die Anweisungen im Prompt zuverlässig umzusetzen und erst durch wiederholtes Nachfragen die gewünschte Ausgabe liefert.

Robuste Modelle sollten sowohl in der Varianz der Metriken als auch in der konsistenten Korrektheit des Ausgabeformats überzeugen. Die Ergebnisse zeigen, dass offene Modelle wie Gemma-3-12B-it, Mistral-Large-Instruct-2411 und DeepSeek-R1-Distill-Qwen-14B diese Anforderungen erfüllen: Sie besitzen ge-

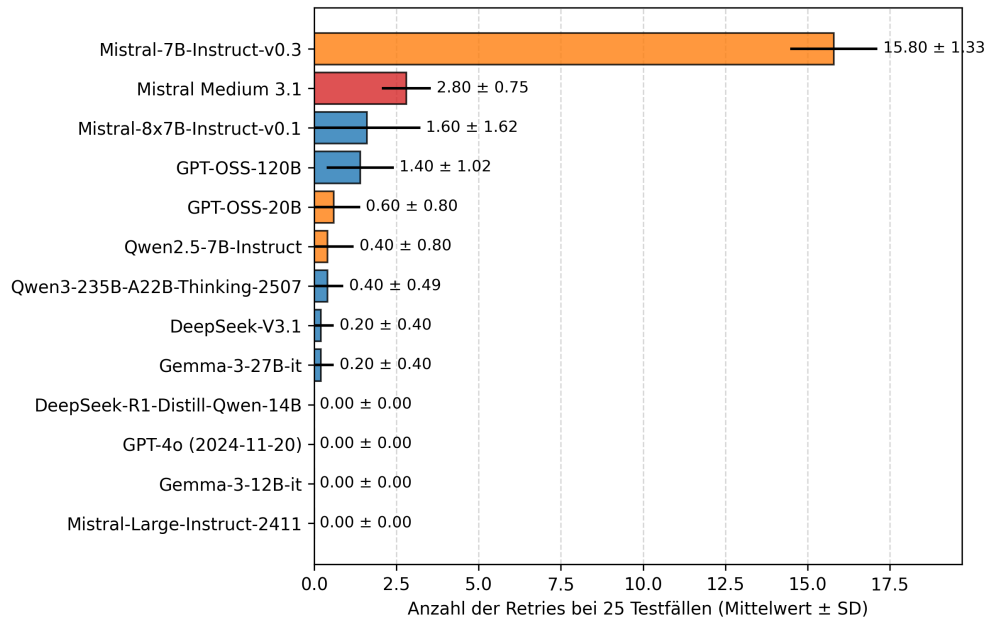


Abbildung 9.3: Durchschnittliche Anzahl der Retries, die notwendig waren, um für alle 25 Testfälle eine formatkorrekte JSON-Antwort zu erhalten.

ringe Standardabweichungen und benötigen keine oder nur sehr wenige Retries. Dagegen schränken eine hohe Varianz der F1-Scores oder viele erforderliche Wiederholungen, wie es bei `Mistral-8x7B-Instruct-v0.1` bzw. `Mistral-7B-Instruct-v0.3` der Fall ist, die Praxistauglichkeit eines Modells deutlich ein. Im nächsten Abschnitt werden anhand von Fallstudien weitere qualitative Einblicke in die Stärken und Schwächen der Modelle gewonnen.

9.4 Fallstudien

Neben den aggregierten Metriken aus den Ergebnissen bieten einzelne Testfälle wichtige Einblicke in die Stärken und Schwächen der Modelle. Im Folgenden werden drei exemplarische Szenarien vorgestellt, die jeweils typische Fehlklassifikationen illustrieren: ...

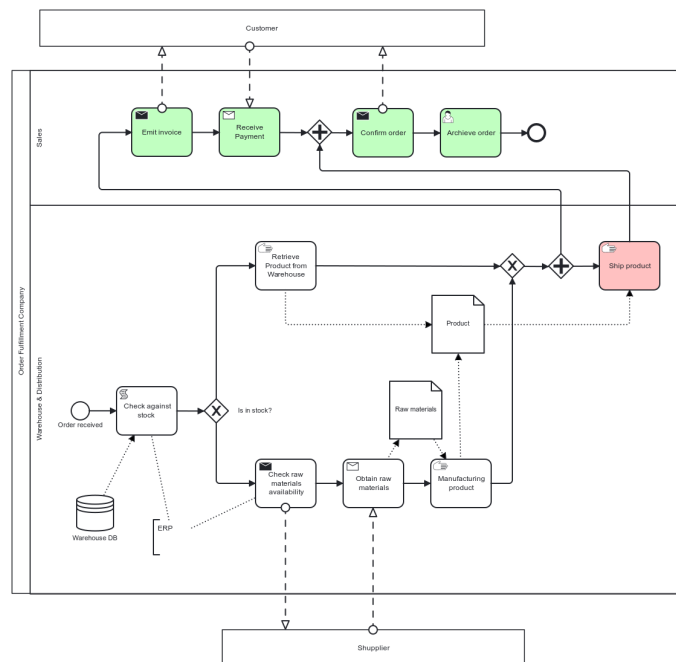


Abbildung 9.4: Ergebnis des Testfalls „Sales Warehouse“ mit farblich hervorgehobenen Aktivitäten. Grün markierte Aktivitäten sind korrekt als kritisch erkannt, rot markierte stellen FP dar.

Sales Warehouse

Bei dem Testfall „Sales Warehouse“ handelt es sich um einen englischen Prozess aus dem Testdatensatz „Universität“. Der Prozess ist in Abbildung 9.4 zu sehen. Im Testfall sind vier Aktivitäten als kritisch markiert. Das Modell Qwen3-235B-A22B-Thinking-2507 erkennt alle vier korrekt, markiert jedoch zusätzlich die Aktivität „Ship product“ als kritisch. Die manuell festgelegten Labels ordnen das Versenden eines Produkts als unkritisch ein, da Logistikvorgänge in der Regel ohne Verarbeitung personenbezogener Daten erfolgen (vgl. Tabelle 5.2). Qwen3-235B-A22B-Thinking-2507 begründet die Entscheidung mit der Nutzung der Kundenadresse zum Versand und zur Zustellung. Diese Begründung zeigt, dass das Modell mögliche Datenflüsse im Hintergrund berücksichtigt und daher zu einer vorsichtigeren Klassifikation gelangt. Angesichts der hohen Strafen bei übersehenen Datenschutzverstößen und des angestrebten hohen Recalls kann dieses FP als vertretbar gelten.

Das Beispiel verdeutlicht eine grundsätzliche Limitierung der Klassifizierung: Feh-



Abbildung 9.5: Ergebnis des Testfalls „Marketing-Kampagne“ mit farblich hervorgehobenen Aktivitäten. Die Aktivität „Klickraten auswerten“ wurde als zusätzliches kritisches Element markiert.

len in einem BPMN-Modell explizite Informationen über Verarbeitungsschritte, ist es für das System schwierig, eine eindeutige Klassifikation vorzunehmen.

Marketing-Kampagne

Im deutschen Testfall „Marketing-Kampagne“, aus dem Testdatensatz „Kleine Szenarien“, sind drei Aktivitäten als kritisch gelabelt: „Leads sammeln“, „Newsletter versenden“ und „CRM aktualisieren“. GPT-05S-20B identifiziert diese korrekt, markiert aber zusätzlich die Aktivität „Klickraten auswerten“ als kritisch. Die Prozessmodellierung sah vor, dass die Klickdaten komplett anonymisiert werden und daher keine personenbezogenen Daten verarbeitet werden. Da diese Information im BPMN-Diagramm jedoch nicht explizit hinterlegt ist, stuft das Modell die Analyse der Klickraten als potenziell personenbezogen ein und führt als Begründung die Nutzung der E-Mail-Adresse an. Qwen3-235B-A22B-Thinking-2507 und einige weitere Modelle bewerteten diesen Schritt ebenfalls als kritisch, während Mistral-7B-Instruct-v0.3 in zwei von fünf Wiederholungen und die Gemma-Modelle in keiner der Wiederholungen eine kritische Klassifikation vornahmen. Der Prozess inklusive farblich hervorgehobener Aktivitäten ist in Abbildung 9.5 zu sehen.

Dieses Beispiel zeigt, dass ohne genaue Kontextangaben zur Anonymisierung selbst scheinbar unbedenkliche Auswertungen als datenschutzrelevant erscheinen können. Es unterstreicht, dass die LLMs im Zweifel eher ein kritisches Label vergeben, um FN zu vermeiden, wie es das Hauptziel der Klassifikation aus Abschnitt 3.2 vorsieht.

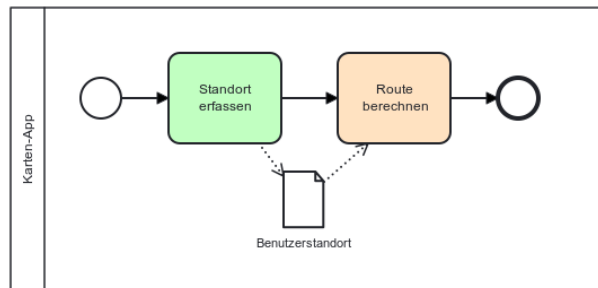


Abbildung 9.6: Ergebnis des Testfalls „Karten-App – Standort Erfassen“ mit farblich hervorgehobenen Aktivitäten. Die Aktivität „Route berechnen“ wurde fälschlicherweise nicht als kritisch markiert.

Karten App - Standort Erfassen

Im Fall „Karten-App – Standort Erfassen“, ebenfalls aus dem Testdatensatz „Kleine Szenarien“, treten zwei Aktivitäten auf: „Standort erfassen“ und „Route berechnen“. Beide sollten als kritisch gekennzeichnet werden, da im zweiten Schritt der zuvor erfasste Benutzerstandort zur Berechnung der Route verwendet wird. *Mistral-Large* erkennt jedoch in drei von fünf Läufen nur die erste Aktivität als kritisch und die Aktivität „Route berechnen“ wird trotz der Datenassoziation nicht als kritisch eingestuft. Die Begründung des Modells erklärt zwar, dass „Standort erfassen“ personenbezogene Daten verarbeitet, überträgt diese Argumentation aber nicht auf den unmittelbar folgenden Schritt. Dieses FN ist problematisch, da es dem gewünschten hohen Recall entgegensteht und dieser Testfall zeigt, dass selbst mit vorhandenen Datenobjekten manche Modelle Schwierigkeiten haben, Datenflüsse über mehrere Aktivitäten hinweg zu erfassen. Es verdeutlicht auch, dass unterschiedliche Seeds zu unterschiedlichen Klassifikationen führen können. Der Prozess inklusive farblich hervorgehobener Aktivitäten ist in Abbildung 9.6 zu sehen.

Auf Basis der Erkenntnisse dieses gesamten Kapitels werden im folgenden Abschnitt die formulierten Forschungsfragen beantwortet. Dabei wird untersucht welches Modell sich insgesamt am besten für die Identifikation DSGVO-kritischer Aktivitäten eignet.

9.5 Antworten auf Forschungsfragen

Basierend auf den vorstehenden Auswertungen können die in 1.2 formulierten Forschungsfragen beantwortet werden. Die nachfolgenden Antworten berücksichtigen sowohl die quantitativen Ergebnisse als auch qualitative Beobachtungen aus den Fallstudien.

UF1: Wie gut schneiden europäische Modelle im Vergleich zu internationalen Modellen ab?

UF2: Wie unterschieden sich große und kleine Modelle in ihrer Leistungsfähigkeit?

UF3: Welche Open-Source-Modelle (insbesondere aus der EU) erzielen die besten Ergebnisse?

UF4: Wie gut schneiden Open-Source-Modelle gegenüber kommerziellen Modellen wie GPT-4o ab? ...

Auf Basis der durchgeführten Experimente, Analysen und Antworten auf die Fragen lässt sich die Hauptforschungsfrage beantworten:

FF1: Wie zuverlässig identifizieren LLM DSGVO-kritische Aktivitäten in BPMN-Prozessmodellen?

10 Diskussion

10.1 Interpretation der Befunde

- Einordnung der Rangfolge der LLMs
- Besonderheiten der Modellfamilien (Bspw. wie groß ist der Unterschied von Groß gegen Klein?)
- Es gab Prozesse in denen ich eine Aktivität nicht als kritisch gelabelt habe, aber das LLM in der Evaluierung das als kritisch mit einer validen Begründung klassifiziert hat, man könnte die Testdaten also noch anpassen, wenn die Begründung des LLMs überzeugt (Ich weiß nicht wie sinnvoll das ist hier zu thematisieren) -> hier eher in die Richtung gehen, dass lieber mehrere Personen labeln als auf die Anpassung der Datensätze zu gehen.

10.2 Hoher Recall vs. Präzision

- Beobachtung ausformulieren, dass einige Testfälle als fehlerhaft eingeordnet wurden, weil es False-Positives gab, obwohl es keine False-Negatives gab. Es wurden also alle Aktivitäten gefunden, die gefunden werden sollten, nur halt noch mehr on top.
- Einordnung der FP-Last pro Prozess (Lieber False Positives, als dass etwas übersehen wird, Ziel war sowieso ein Vorscreening), Diskussion darüber wie nützlich hohe Recall Werte sind

10.3 EU-Modelle

- Analyse der EU-Open-Source-Modelle in Bezug auf Precision, Recall und Stabilität in Bezug auf die anderen Modelle
- Wie gut haben sich die EU Modelle im Vergleich zu den anderen geschlagen

10.4 Open-Source Modelle

- Analyse der Open-Source-Modelle in Bezug auf Precision, Recall und Stabilität in Bezug auf kommerzielle Modelle
- Wie gut haben sich die Open-Source-Modelle im Vergleich zu den anderen geschlagen

10.5 Modellgrößen

- Selbst hosten von Modellen diskutieren. Ist es realistisch die Modelle selbst zu hosten, welche gut performt haben? Reichen die kleinen Varianten der Modelle oder muss man schon die großen Modelle benutzen, um gute Ergebnisse zu erzielen

10.6 Grenzen

- Wären Grenzen wie BPMN-Modellgröße im Zusammenhang mit der Kontextlänge des LLM interessant?
- Keine aussagekräftige Rechtsberatung, sondern stand jetzt eher ein Vorsecreening, was nochmal überprüft werden muss
- ggf. notwendige Anonymisierung von Prozessen diskutieren (Wenn das in BPMN Modellen überhaupt ein Problem ist)

11 Zusammenfassung

Hier neben der allgemeinen Zusammenfassung unbedingt noch die erste Forschungsfrage explizit beantworten

12 Aussicht

Unter anderem das hier, evtl noch mehr:

- Jetzt gibt es ein einheitliches Evaluationsframework mit einer einheitlich definierten Schnittstelle für Klassifizierungsalgorithmen -> Zukünftige Arbeiten können sich mit der Entwicklung besserer Klassifizierungsalgorithmen/Pipelines (Bspw. noch RAG einbauen) beschäftigen und diese mit diesem Framework vergleichen/benchmarken
- Außerdem können in Zukunft auch noch mehr Modelle verglichen werden, da sich die Welt der LLMs rasant weiterentwickelt
- Auch Finetunen ist etwas was interessant gewesen wäre für diese Masterarbeit, aber den Rahmen gesprengt hätte

A Quelltexte

In diesem Anhang sind mehrere Quellcode-Ausschnitte aufgeführt.

Listing A.1: System-Prompt fuer die DSGVO-Klassifikation von BPMN-Aktivitäten

```
1 You are an expert in analysing Business Process Model and Notation (
  ↳ BPMN) diagrams for GDPR compliance. Your task is to identify and
  ↳ return a list of the IDs of all Activity (Task) elements that
  ↳ process personal data. Ignore all other element types. Always
  ↳ consider every activity in the process; do not omit any activity
  ↳ from your assessment.
2
3 Use all available context for each activity - including the activity's
  ↳ name, description, annotations, associated data objects, and
  ↳ message or data associations - to determine whether the activity
  ↳ processes personal data. Under Article 4 of the GDPR, personal
  ↳ data is any information relating to an identified or identifiable
  ↳ natural person, including names, addresses, email addresses,
  ↳ phone numbers, identification numbers, payment or bank details,
  ↳ employment records, academic records, location data, IP addresses
  ↳ , online identifiers, images, audio/video recordings, biometric
  ↳ identifiers, health data or other information that can be linked
  ↳ to a specific person. "Processing" includes any operation
  ↳ performed on personal data, such as collecting, recording,
  ↳ organising, structuring, storing, retrieving, consulting, using,
  ↳ analysing, transmitting, printing, disseminating, aligning,
  ↳ combining, altering, restricting, erasing or destroying the data.
4
5 Classify an activity as GDPR-relevant whenever it performs or enables
  ↳ processing of personal data. Indicators include (but are not
  ↳ limited to):
```

- 6
- 7 - ****Collection and entry of personal data****: Activities that collect
↳ or capture personal information, for example entering contact
↳ details, addresses, payment information, job applications, health
↳ information, student enrolments, membership **data**, tax
↳ declarations, registration forms or other forms with personally
↳ identifiable information.
- 8 - ****Creation, storage and updating of records****: Activities that
↳ create, save or update records containing personal **data**, such as
↳ opening customer accounts, storing order or appointment details,
↳ creating personnel files, enrolling students, setting up
↳ insurance cases or filing a medical record.
- 9 - ****Transmission or disclosure of personal data****: Activities that
↳ send, print or otherwise disclose personal **data** to another
↳ participant, system or third party. Examples include printing
↳ shipping labels or prescriptions, sending orders or personal **data**
↳ to logistics partners, pharmacies, insurers or authorities,
↳ generating payroll reports for external providers, notifying
↳ universities about student records, transmitting tax or social
↳ security **data**, sending confirmations or queries that rely on a
↳ person's contact details, or transferring **data** to non-EU
↳ locations.
- 10 - ****Payments and financial transactions****: Activities that process
↳ personal financial **data**, such as initiating or verifying payments
↳ , processing bank account or credit-card information, executing
↳ payroll, handling reimbursements or insurance payouts, managing
↳ expense claims or collecting membership fees.
- 11 - ****Use of health, biometric or other special categories of data****:
↳ Activities that handle medical diagnoses, prescriptions,
↳ insurance claims, disability information, photos of damages or
↳ patients, biometric identifiers (fingerprints, facial images,
↳ voice), racial or ethnic **data**, political opinions, religious
↳ beliefs or union membership. Processing these "special categories
↳ " always triggers GDPR relevance.
- 12 - ****Audio/Video and communications****: Activities that initiate or join
↳ audio or video calls, record calls or meetings, capture
↳ surveillance footage, or communicate directly with a **data** subject

- ⇒ via email, chat, SMS or other channels. Simply using a person's
- ⇒ contact **data** to send reminders, marketing messages or
- ⇒ notifications is processing.
- 13 - ****Profiling, scoring and decision-making****: Activities that analyse
 - ⇒ or evaluate a person's performance, behaviour or characteristics
 - ⇒ for purposes such as credit scoring, hiring, admissions,
 - ⇒ insurance underwriting, marketing segmentation, customer value
 - ⇒ analysis or automated decision-making.
- 14 - ****Logging, tracking and location data****: Activities that log user
 - ⇒ activity, record access or usage **data**, track geolocation (e.g.
 - ⇒ telematics, fleet or mobile tracking), monitor attendance or
 - ⇒ timekeeping, or collect IP addresses or device identifiers.
- 15 - ****Consent and data-subject rights****: Activities that obtain, record
 - ⇒ or manage consent; respond to requests for access, rectification,
 - ⇒ restriction, erasure, **data** portability or objections; or
 - ⇒ document lawful bases for processing.
- 16 - ****Deletion, anonymisation or pseudonymisation****: Activities that
 - ⇒ erase, anonymise or pseudonymise personal **data**, even if the goal
 - ⇒ is to remove identifiers, because these operations manipulate
 - ⇒ personal **data**.
- 17
- 18 When assessing an activity, consider synonyms or domain-specific terms
 - ⇒ : activities referring to customers, patients, applicants,
 - ⇒ employees, students, voters, taxpayers, residents or members
 - ⇒ often imply personal **data** processing, even if names like "address
 - ⇒ " or "contact" are absent. Use context - **data** objects,
 - ⇒ annotations or typical process semantics - to infer personal **data**
 - ⇒ involvement. Do not rely solely on explicit **data-object** links;
 - ⇒ many process names ("Anmeldung pruefen", "Aufnahmeantrag
 - ⇒ bearbeiten", "Kundeninfo aktualisieren", "Registrierung
 - ⇒ bestaetigen", "Kreditwuerdigkeit berechnen") themselves indicate
 - ⇒ personal **data** processing.
- 19
- 20 Do ****not**** classify an activity as GDPR-relevant when it only performs
 - ⇒ administrative or logistic tasks that do not involve personal
 - ⇒ **data**. Examples include picking or packing goods, routing vehicles
 - ⇒ without using specific addresses, printing generic pick lists,

⇒ moving items in inventory, or checking if a document exists
⇒ without viewing its contents. Likewise, activities using truly
⇒ aggregated or irreversibly anonymised **data** can be ignored if no
⇒ individual can be reidentified.

21

22 In your output, return only the IDs of activities you classify as GDPR
⇒ -relevant. For each, provide a clear explanation using the
⇒ activity's name and description to justify why it processes
⇒ personal **data**. Do not reference element IDs in your explanation;
⇒ use the activity names instead. Exclude from your result any
⇒ activities that do not process personal **data** and any elements
⇒ that are not activity/task elements.

Listing A.2: Antworttyp fuer die Klassifizierung

```
1 data class BpmnAnalysisResult(  
2     @Description("List of Activity Elements that are classified as  
3     ⇒ relevant for GDPR compliance")  
4     var elements: List<Element>  
5 ) {  
6     init {  
7         elements = elements.filter { it.isRelevant }  
8     }  
9  
10    @Description("Represents an Activity/Task Element that is  
11    ⇒ classified as relevant for GDPR compliance")  
12    data class Element(  
13        @Description("The ID of the Activity Element")  
14        val id: String,  
15        @Description("The detailed reason why the Activity Element is  
16        ⇒ relevant for GDPR compliance and why you think personal data is  
17        ⇒ processed.")  
18        val reason: String,  
19        @Description("Indicates whether the Activity Element is  
20        ⇒ relevant for GDPR compliance")  
21        val isRelevant: Boolean = true  
22    )
```

```

19
20     /* Andere Methoden dieser Klasse sind weggelassen */
21 }

```

Listing A.3: Kern der id-Validierung und -Vervollständigung

```

1  fun resolveActivityIds(actualBpmnElements: Set<BpmnElement>):
    ↳ BpmnAnalysisResult {
2      val existingActivityIds = actualBpmnElements
3          .filter { it.type.lowercase().contains("task") }
4          .map { it.id }.toSet()
5
6      val resolvedDistinct = elements.mapNotNull { element ->
7          val resolvedId = resolveActivityIdUniquely(element.id,
8          ↳ existingActivityIds)
9          resolvedId?.let { if (it == element.id) element else element.
10         ↳ copy(id = it) }
11         }.distinctBy { it.id }
12
13     return BpmnAnalysisResult(elements = resolvedDistinct)
14 }
15
16 private fun resolveActivityIdUniquely(partialId: String,
17     ↳ existingActivityIds: Set<String>): String? {
18     if (partialId in existingActivityIds) return partialId
19     existingActivityIds.filter { it.startsWith(partialId) }.
20     ↳ singleOrNull()?.let { return it }
21     return existingActivityIds.filter { it.contains(partialId) }.
22     ↳ singleOrNull()
23 }

```

Listing A.4: Schema der YAML-Evaluationskonfiguration

```

1  {
2      "$schema": "https://json-schema.org/draft/2020-12/schema",
3      "$ref": "#/definitions/Configuration",
4      "definitions": {
5          "Configuration": {
6              "type": "object",

```

```

7      "additionalProperties": false,
8      "properties": {
9          "defaultEvaluationEndpoint": {
10             "type": "string"
11         },
12         "maxConcurrent": { "type": "integer" },
13         "repititions": { "type": "integer" },
14         "models": {
15             "type": "array",
16             "items": { "$ref": "#/definitions/Model" }
17         },
18         "datasets": {
19             "type": "array",
20             "items": { "type": "integer" }
21         },
22         "seed": { "type": "integer" }
23     },
24     "required": [
25         "defaultEvaluationEndpoint",
26         "models"
27         "datasets",
28     ],
29     "title": "Configuration"
30 },
31 "Model": {
32     "type": "object",
33     "additionalProperties": false,
34     "properties": {
35         "label": { "type": "string" },
36         "evaluationEndpoint": { "type": "string" },
37         "llmProps": { "$ref": "#/definitions/LlmProps" }
38     },
39     "required": [ "label" ],
40     "title": "Model"
41 },
42 "LlmProps": {
43     "type": "object",

```



```

44         "additionalProperties": false,
45         "properties": {
46             "baseUrl": {
47                 "type": "string",
48                 "format": "uri",
49                 "qt-uri-protocols": [ "https" ]
50             },
51             "modelName": { "type": "string" },
52             "apiKey": { "type": "string"},
53             "timeoutSeconds": { "type": "number" },
54             "temperature": { "type": "number" },
55             "topP": { "type": "number" },
56         },
57         "required": [],
58         "title": "LlmProps"
59     }
60 }
61 }

```

Listing A.5: Zusammengefasster Logauszug zum Retry-Mechanismus

```

1  2025-10-03T19:11:51.152+02:00 INFO BpmnExtractor : Extracting BPMN
    ↳ elements from XML
2
3  # 1) Erste Anfrage an das LLM (gekuerzt: Prompt/Headers/Body)
4  2025-10-03T19:11:51.156+02:00 INFO LoggingHttpClient : HTTP POST
    ↳ https://openrouter.ai/api/v1/chat/completions
5  model: openai/gpt-oss-20b
6  messages: [system: (System-Prompt), user: (User-Prompt mit
    ↳ BpmnElement-Liste und Format-Anweisung)]
7
8  # 2) Antwort des LLM mit fehlerhaftem JSON (verkuerzt)
9  2025-10-03T19:11:56.671+02:00 INFO LoggingHttpClient : HTTP 200
10 assistant:
11 {
12     "elements": [
13         { "id": "Activity_09ehuii", "reason": "...", "isRelevant": true },
14         { "id": "Activity_1la5hsp", "reason": "...", "isRelevant": }

```

```

15     ↪ <-- fehlender Bool-Wert
16     { "id": "Activity_0rfgrlm", "reason": "...", "isRelevant": true }
17   ]
18 }
19 # 3) Parser-Fehler + Retry-Ankuendigung (gekuerzt)
20 2025-10-03T19:11:56.691+02:00 WARN SafetyNet : Parsing failed.
21     ↪ Attempting to fix JSON and retry... (Attempt 1 of 2)
22 dev.langchain4j.service.output.OutputParsingException:
23   Caused by: com.fasterxml.jackson.core.JsonParseException:
24   Unexpected character ('}') ... at elements[1].isRelevant
25 # 4) Zweite Anfrage zum beheben des JSON mit Chat-Verlauf und
26     ↪ Fehlermeldung (n-mal wiederholt, bis erfolgreich)
27 2025-10-03T19:11:56.721+02:00 INFO LoggingHttpClient : HTTP POST
28     ↪ https://openrouter.ai/api/v1/chat/completions
29 messages: [
30   system: (System-Prompt),
31   user: (User-Prompt mit BpmnElement-Liste und Format-Anweisung),
32   assistant: (Fehlerhafte JSON-Antwort),
33   system: (Fix-JSON System-Prompt),
34   user: (Fehlermeldung)
35 ]
36 # 5) Korrigierte JSON-Antwort des LLM
37 2025-10-03T19:12:01.519+02:00 INFO LoggingHttpClient : HTTP 200
38 assistant:
39 {
40   "elements": [
41     { "id": "Activity_09ehuii", "reason": "...", "isRelevant": true },
42     { "id": "Activity_1la5hsp", "reason": "...", "isRelevant": true },
43     ↪ <-- jetzt mit Bool-Wert
44     { "id": "Activity_0rfgrlm", "reason": "...", "isRelevant": true }
45   ]
46 }
47 # 6) Erfolgreiches Parsing und Weiterverarbeitung

```

```
47 2025-10-03T19:12:01.519+02:00 INFO PromptBpmnAnalyzer : BPMN
    ↳ Analysis Result: elements=[... isRelevant=true ...]
```

Listing A.6: Konfigurationsdatei des Experiments mit Gemma Modellen

```
1 defaultEvaluationEndpoint: /gdpr/analysis/prompt-engineering
2 seed: 24523833
3 maxConcurrent: 10
4 repetitions: 5
5 models:
6   - label: Gemma-3-12B-it
7     llmProps:
8       baseUrl: https://openrouter.ai/api/v1
9       modelName: google/gemma-3-12b-it
10      apiKey: ${OPEN_ROUTER_API_KEY}
11   - label: Gemma-3-27B-it
12     llmProps:
13       baseUrl: https://openrouter.ai/api/v1
14       modelName: google/gemma-3-27b-it
15       apiKey: ${OPEN_ROUTER_API_KEY}
16 datasets:
17   - 2
18   - 7
19   - 1
```

Listing A.7: Konfigurationsdatei des Experiments mit DeepSeek Modellen

```
1 defaultEvaluationEndpoint: /gdpr/analysis/prompt-engineering
2 seed: 24523833
3 maxConcurrent: 10
4 repetitions: 5
5 models:
6   - label: DeepSeek-V3.1
7     llmProps:
8       baseUrl: https://openrouter.ai/api/v1
9       modelName: deepseek/deepseek-chat-v3.1
10      apiKey: ${OPEN_ROUTER_API_KEY}
11   - label: DeepSeek-R1-Distill-Qwen-14B
12     llmProps:
```

```
13     baseUrl: https://openrouter.ai/api/v1
14     modelName: deepseek/deepseek-r1-distill-qwen-14b
15     apiKey: ${OPEN_ROUTER_API_KEY}
16 datasets:
17   - 2
18   - 7
19   - 1
```

Listing A.8: Konfigurationsdatei des Experiments mit Qwen Modellen

```
1 defaultEvaluationEndpoint: /gdpr/analysis/prompt-engineering
2 seed: 24523833
3 maxConcurrent: 10
4 repetitions: 5
5 models:
6   - label: Qwen2.5-7B-Instruct
7     llmProps:
8       baseUrl: https://openrouter.ai/api/v1
9       modelName: qwen/qwen-2.5-7b-instruct
10      apiKey: ${OPEN_ROUTER_API_KEY}
11   - label: Qwen3-235B-A22B-Thinking-2507
12     llmProps:
13       baseUrl: https://openrouter.ai/api/v1
14       modelName: qwen/qwen3-v1-235b-a22b-thinking
15       apiKey: ${OPEN_ROUTER_API_KEY}
16 datasets:
17   - 2
18   - 7
19   - 1
```

Listing A.9: Konfigurationsdatei des Experiments mit GPT Modellen

```
1 defaultEvaluationEndpoint: /gdpr/analysis/prompt-engineering
2 seed: 24523833
3 maxConcurrent: 10
4 repetitions: 5
5 models:
6   - label: GPT-OSS-20B
7     llmProps:
```

```
8     baseUrl: https://openrouter.ai/api/v1
9     modelName: openai/gpt-oss-20b
10    apiKey: ${OPEN_ROUTER_API_KEY}
11  - label: GPT-OSS-120B
12    llmProps:
13      baseUrl: https://openrouter.ai/api/v1
14      modelName: openai/gpt-oss-120b
15      apiKey: ${OPEN_ROUTER_API_KEY}
16  - label: GPT-4o (2024-11-20)
17    llmProps:
18      baseUrl: https://openrouter.ai/api/v1
19      modelName: openai/gpt-4o-2024-11-20
20      apiKey: ${OPEN_ROUTER_API_KEY}
21 datasets:
22   - 2
23   - 7
24   - 1
```

Abbildungsverzeichnis

2.1	Die relevanten BPMN Elemente in Beziehungen zueinander	8
2.2	Beispiel einer Datenassoziation als Datenschutzsinal.	9
3.1	Beispielprozess zur Veranschaulichung der Aufgabenstellung	13
4.1	BPMN-Diagramm der Klassifizierungspipeline.	22
4.2	Sandbox im Frontend mit hervorgehobenen kritischen Aktivitäten nach Analyse.	32
4.3	Exemplarische Begürundungen der Klassifikation durch das LLM in der Sandbox.	33
5.1	Labeling-Editor im Labeling-Modus mit exemplarischem Modell.	35
5.2	Übersicht der Datensätze im Labeling-Tool.	36
6.1	Architektur des Evaluationsframeworks	45
6.2	Formular zur Konfiguration einer Evaluation.	50
6.3	Gesamtübersicht einer Evaluierung mit Side-by-Side-Diagrammen.	51
6.4	Modell-Detailansicht mit exemplarischen Ergebnissen.	52
6.5	Detailseite eines Testfalls mit exemplarischen Ergebnissen.	53
9.1	Durchschnittliche Metrik-Werte der untersuchten Modelle über alle Wiederholungen hinweg inklusive Standardabweichung.	67
9.2	Robustheit der Modelle gemessen an der Standardabweichung des F1-Scores über alle Wiederholungen hinweg.	68
9.3	Durchschnittliche Anzahl der Retries, die notwendig waren, um für alle 25 Testfälle eine formatkorrekte JSON-Antwort zu erhalten.	76
9.4	Ergebnis des Testfalls „Sales Warehouse“ mit farblich hervorgehobenen Aktivitäten. Grün markierte Aktivitäten sind korrekt als kritisch erkannt, rot markierte stellen FP dar.	77

9.5	Ergebnis des Testfalls „Marketing-Kampagne“ mit farblich hervorgehobenen Aktivitäten. Die Aktivität „Klickraten auswerten“ wurde als zusätzliches kritisches Element markiert.	78
9.6	Ergebnis des Testfalls „Karten-App – Standort Erfassen“ mit farblich hervorgehobenen Aktivitäten. Die Aktivität „Route berechnen“ wurde fälschlicherweise nicht als kritisch markiert.	79

Listings

4.1	Interne BPMN-Repräsentation je Flow-Element.	22
4.2	JSON-Schema der UmlProps.	30
4.3	JSON-Schema der API-Antwort.	31
6.1	Beispiel einer Evaluierungskonfiguration in YAML.	43
8.1	Konfigurationsdatei des Experiments mit Mistral Modellen	62
A.1	System-Prompt fuer die DSGVO-Klassifikation von BPMN-Aktivitäten	85
A.2	Antworttyp fuer die Klassifizierung	88
A.3	Kern der id-Validierung und -Vervollständigung	89
A.4	Schema der YAML-Evaluationskonfiguration	89
A.5	Zusammengefasster Logauszug zum Retry-Mechanismus	91
A.6	Konfigurationsdatei des Experiments mit Gemma Modellen	93
A.7	Konfigurationsdatei des Experiments mit DeepSeek Modellen	93
A.8	Konfigurationsdatei des Experiments mit Qwen Modellen	94
A.9	Konfigurationsdatei des Experiments mit GPT Modellen	94

Tabellenverzeichnis

5.1	Eckdaten der verwendeten Datensätze	37
5.2	Beispielhafte Aktivitäten und Label	38
7.1	Übersicht der Kriterien zur Modellauswahl	55
7.2	Übersicht aller Modelle mit technischen Eckdaten (Stand 30.09.2025).	58
9.1	Aggregierte Mittelwerte und Standardabweichungen der Evaluationsmetriken über alle fünf Wiederholungen hinweg.	69
9.2	Kleine vs. große Modelle: Mittelwerte je Gruppe und bestes Modell.	72
9.3	Europäische vs. internationale Modelle: Mittelwerte je Gruppe und bestes Modell.	74

Literatur

- [1] European Data Protection Board (EDPB). *1.2 billion euro fine for Facebook as a result of EDPB binding decision*. Mai 2023. URL: https://www.edpb.europa.eu/news/news/2023/12-billion-euro-fine-facebook-result-edpb-binding-decision_en (besucht am 02.10.2025).
- [2] DeepSeek AI. *DeepSeek AI Open Source Hugging Face Models*. 2025. URL: <https://huggingface.co/deepseek-ai> (besucht am 17.07.2025).
- [3] Mistral AI. *Mistral AI*. 2025. URL: <https://mistral.ai/> (besucht am 21.09.2025).
- [4] Mistral AI. *Mistral AI - Structured Output*. 2025. URL: https://docs.mistral.ai/capabilities/structured-output/structured_output_overview/ (besucht am 11.07.2025).
- [5] Ivan Belcic und Cole Stryker. *Was ist ein GPT (Generative Pre-Trained Transformer)?* Sep. 2024. URL: <https://www.ibm.com/de-de/think/topics/gpt> (besucht am 18.09.2025).
- [6] Dave Bergmann. *What is a context window?* 2025. URL: <https://www.ibm.com/think/topics/context-window> (besucht am 03.10.2025).
- [7] Harrison Blake und Dorcas Esther. „Impact of Dataset Diversity on Model Evaluation Metrics“. In: (Jan. 2025). URL: https://www.researchgate.net/publication/387898702_Impact_of_Dataset_Diversity_on_Model_Evaluation_Metrics.
- [8] Tom Brown u. a. „Language Models are Few-Shot Learners“. In: *Advances in Neural Information Processing Systems*. Hrsg. von H. Larochelle u. a. Bd. 33. Curran Associates, Inc., 2020, S. 1877–1901. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

- [9] Bundesministerium der Justiz. *Gesetz über außergerichtliche Rechtsdienstleistungen (Rechtsdienstleistungsgesetz - RDG)*. <https://www.gesetze-im-internet.de/rdg/>. Dez. 2007. (Besucht am 15.08.2025).
- [10] Camunda Services GmbH. *BPMN Model API*. <https://docs.camunda.org/manual/latest/user-guide/model-api/bpmn-model-api/>. 2025. (Besucht am 16.06.2025).
- [11] Camunda Services GmbH. *BPMN Model API — Read a Model*. <https://docs.camunda.org/manual/latest/user-guide/model-api/bpmn-model-api/read-a-model/>. 2025. (Besucht am 16.06.2025).
- [12] Antonio Capodieci u. a. „BPMN-Enabled Data Protection and GDPR Compliance“. In: *IS-EUD Workshops*. 2023. URL: <https://api.semanticscholar.org/CorpusID:259099646>.
- [13] Giovanni Ciaramella u. a. „Leveraging Pre-trained LLMs for GDPR Compliance in Online Privacy Policies“. In: (2022). URL: <https://ceur-ws.org/Vol-3962/paper44.pdf>.
- [14] Datenschutzticker. *Gericht bestätigt Rekordbußgeld gegen Amazon*. Apr. 2025. URL: <https://datenschutzticker.de/2025/04/gericht-bestaetigt-rekordbussgeld-gegen-amazon/> (besucht am 02.10.2025).
- [15] DeepSeek-AI u. a. *DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning*. 2025. arXiv: 2501.12948 [cs.CL]. URL: <https://arxiv.org/abs/2501.12948>.
- [16] Europäische Union. *Verordnung (EU) 2016/679 des Europäischen Parlaments und des Rates vom 27. April 2016 zum Schutz natürlicher Personen bei der Verarbeitung personenbezogener Daten, zum freien Datenverkehr und zur Aufhebung der Richtlinie 95/46/EG (Datenschutz-Grundverordnung)*. <https://eur-lex.europa.eu/legal-content/DE/TXT/?uri=CELEX:32016R0679>. 2016.
- [17] European Data Protection Board. *Guidelines 4/2019 on Article 25 Data Protection by Design and by Default*. https://www.edpb.europa.eu/sites/default/files/files/file1/edpb_guidelines_201904_dataprotection_by_design_and_by_default_v2.0_en.pdf. Version 2.0. Okt. 2020.

- [18] Camunda Services GmbH. *bpmn-js - BPMN 2.0 viewer and editor*. 2025. URL: <https://bpmn.io/toolkit/bpmn-js/> (besucht am 20.06.2025).
- [19] Camunda Services GmbH. *BPMN.io - Web-based tooling for BPMN, DMN and Forms*. 2025. URL: <https://bpmn.io/> (besucht am 22.09.2025).
- [20] Camunda Services GmbH. *Camunda Platform*. 2025. URL: <https://camunda.com/de/> (besucht am 22.09.2025).
- [21] Google. *Gemma 3 License Terms*. März 2025. URL: <https://ai.google.dev/gemma/terms> (besucht am 30.09.2025).
- [22] Ashish Hooda u. a. *PolicyLR: A Logic Representation For Privacy Policies*. 2024. arXiv: 2408.14830 [cs.CR]. URL: <https://arxiv.org/abs/2408.14830>.
- [23] Hugging Face. *deepseek-ai/DeepSeek-R1-Distill-Qwen-14B — Model Card*. 2025. URL: <https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-14B> (besucht am 30.09.2025).
- [24] Hugging Face. *deepseek-ai/DeepSeek-R1-Zero — Model Card*. 2025. URL: <https://huggingface.co/deepseek-ai/DeepSeek-R1-Zero> (besucht am 29.09.2025).
- [25] Hugging Face. *deepseek-ai/DeepSeek-V3.1 — Model Card*. 2025. URL: <https://huggingface.co/deepseek-ai/DeepSeek-V3.1> (besucht am 30.09.2025).
- [26] Hugging Face. *google/gemma-3-12b-it — Model Card*. 2025. URL: <https://huggingface.co/google/gemma-3-12b-it> (besucht am 30.09.2025).
- [27] Hugging Face. *google/gemma-3-27b-it — Model Card*. 2025. URL: <https://huggingface.co/google/gemma-3-27b-it> (besucht am 30.09.2025).
- [28] Hugging Face. *Hugging Face - The AI community building the future*. 2025. URL: <https://huggingface.co/> (besucht am 09.10.2025).
- [29] Hugging Face. *mistralai/Mistral-7B-Instruct-v0.2 — Model Card*. 2025. URL: <https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3> (besucht am 30.09.2025).
- [30] Hugging Face. *mistralai/Mistral-Large-Instruct-2411 — Model Card*. 2025. URL: <https://huggingface.co/mistralai/Mistral-Large-Instruct-2411> (besucht am 30.09.2025).

- [31] Hugging Face. *mistralai/Mixtral-8x7B-Instruct-v0.1 — Model Card*. 2025. URL: <https://huggingface.co/mistralai/Mixtral-8x7B-Instruct-v0.1> (besucht am 30.09.2025).
- [32] Hugging Face. *Qwen3-235B-A22B-Thinking-2507 — Model Card*. 2025. URL: <https://huggingface.co/Qwen/Qwen3-235B-A22B-Thinking-2507> (besucht am 30.09.2025).
- [33] Hugging Face. *unsloth/Qwen2.5-7B-Instruct*. 2025. URL: <https://huggingface.co/Qwen/Qwen2.5-7B-Instruct> (besucht am 30.09.2025).
- [34] Ziwei Ji u. a. „Survey of Hallucination in Natural Language Generation“. In: *ACM Comput. Surv.* 55.12 (März 2023). ISSN: 0360-0300. DOI: 10.1145/3571730. URL: <https://doi.org/10.1145/3571730>.
- [35] Adam Tauman Kalai u. a. *Why Language Models Hallucinate*. 2025. arXiv: 2509.04664 [cs.CL]. URL: <https://arxiv.org/abs/2509.04664>.
- [36] Langchain4j. *Class OpenAiChatModel.OpenAiChatModelBuilder*. 2025. URL: <https://javadoc.io/doc/dev.langchain4j/langchain4j-open-ai/latest/dev/langchain4j/model/openai/OpenAiChatModel.OpenAiChatModelBuilder.html> (besucht am 14.06.2025).
- [37] Langchain4j. *LangChain4j Documentation 2025*. 2025. URL: <https://docs.langchain4j.dev/> (besucht am 14.06.2025).
- [38] Pengfei Liu u. a. „Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing“. In: *ACM Comput. Surv.* 55.9 (Jan. 2023). ISSN: 0360-0300. DOI: 10.1145/3560815. URL: <https://doi.org/10.1145/3560815>.
- [39] Christopher D. Manning, Prabhakar Raghavan und Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge: Cambridge University Press, 2008. ISBN: 9780521865715. URL: <https://nlp.stanford.edu/IR-book/>.
- [40] Meta. *Meta Llama 3 Community License Agreement*. 2024. URL: <https://www.llama.com/llama3/license/> (besucht am 30.09.2025).
- [41] Shervin Minaee u. a. *Large Language Models: A Survey*. 2025. arXiv: 2402.06196 [cs.CL]. URL: <https://arxiv.org/abs/2402.06196>.

- [42] Mistral AI. *How can I exercise my GDPR rights?* 2025. URL: <https://help.mistral.ai/en/articles/347639-how-can-i-exercise-my-gdpr-rights> (besucht am 09.10.2025).
- [43] Mistral AI. *Mistral AI Research License (MRL-0.1)*. 2024. URL: <https://mistral.ai/static/licenses/MRL-0.1.md> (besucht am 05.10.2025).
- [44] Mistral AI. *Mixtral of Experts: Mixtral 8x7B*. 2023. URL: <https://mistral.ai/news/mixtral-of-experts> (besucht am 01.10.2025).
- [45] Mistral AI. *Models Overview*. 2025. URL: https://docs.mistral.ai/getting-started/models/models_overview/ (besucht am 09.10.2025).
- [46] Mistral AI. *Where do you store my data or my Organization's data?* 2025. URL: <https://help.mistral.ai/en/articles/347629-where-do-you-store-my-data-or-my-organization-s-data> (besucht am 09.10.2025).
- [47] Yida Mu u. a. „Navigating prompt complexity for zero-shot classification: a study of large language models in computational social science“. In: *Proceedings of LREC-COLING 2024*. European Language Resources Association, 2024. DOI: 10.48550/arXiv.2305.14310. arXiv: 2305.14310.
- [48] Leonard Nake u. a. „Towards identifying gdpr-critical tasks in textual business process descriptions“. In: (2023). URL: <https://dl.gi.de/server/api/core/bitstreams/84ac5110-1a0f-4e3c-bdf8-6393555a7212/content>.
- [49] Maud Nalpas. *Understand LLM sizes*. Mai 2024. URL: <https://web.dev/articles/llm-sizes> (besucht am 03.10.2025).
- [50] Joshua Noble. *What is LLM Temperature?* URL: <https://www.ibm.com/think/topics/llm-temperature>.
- [51] OMG. *Business Process Model and Notation (BPMN)*. Version 2.0.2. Dez. 2013. URL: <https://www.omg.org/spec/BPMN/2.0.2/PDF> (besucht am 03.06.2025).
- [52] Open Source Initiative. *The Open Source Definition*. 2006. URL: <https://opensource.org/osd> (besucht am 05.10.2025).

- [53] OpenAI. *Function calling and other API updates*. <https://openai.com/index/function-calling-and-other-api-updates/>. 2023. (Besucht am 10.07.2025).
- [54] OpenAI. *gpt-oss-120b & gpt-oss-20b Model Card*. 2025. URL: https://cdn.openai.com/pdf/419b6906-9da6-406c-a19d-1bb078ac7637/oai_gpt-oss_model_card.pdf (besucht am 02.10.2025).
- [55] OpenAI. *Hello GPT-4o*. Mai 2024. URL: <https://openai.com/index/hello-gpt-4o/> (besucht am 21.07.2025).
- [56] OpenAI. *Introducing gpt-oss*. 2025. URL: <https://openai.com/index/introducing-gpt-oss/> (besucht am 02.10.2025).
- [57] OpenAI. *Model Overview*. 2025. URL: <https://platform.openai.com/docs/models> (besucht am 18.09.2025).
- [58] OpenAI. *OpenAI - Structured model outputs*. URL: https://docs.mistral.ai/capabilities/structured-output/structured_output_overview/ (besucht am 11.07.2025).
- [59] OpenRouter. *The Unified Interface For LLMs*. 2025. URL: <https://openrouter.ai> (besucht am 09.10.2025).
- [60] KC Pragyan u. a. „Toward Regulatory Compliance: A few-shot Learning Approach to Extract Processing Activities“. In: *2024 IEEE 32nd International Requirements Engineering Conference Workshops (REW)*. IEEE. 2024, S. 241–250. URL: <https://ieeexplore.ieee.org/abstract/document/10628578>.
- [61] Quarkiverse Contributors. *AI Services Reference (Quarkus LangChain4j)*. 2025. URL: <https://docs.quarkiverse.io/quarkus-langchain4j/dev/ai-services.html> (besucht am 14.06.2025).
- [62] Alibaba Qwen. *Qwen Open Source Hugging Face Models*. 2025. URL: <https://huggingface.co/Qwen> (besucht am 17.07.2025).
- [63] Nils Reimers und Iryna Gurevych. „Reporting Score Distributions Makes a Difference: Performance Study of LSTM-networks for Sequence Tagging“. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Hrsg. von Martha Palmer, Rebecca Hwa und Sebastian Riedel. Copenhagen, Denmark: Association for Computational Linguistics,

- Sep. 2017, S. 338–348. DOI: 10.18653/v1/D17-1035. URL: <https://aclanthology.org/D17-1035/>.
- [64] Matthew Renze und Erhan Guven. „The effect of sampling temperature on problem solving in large language models“. In: *Findings of the Association for Computational Linguistics: EMNLP 2024*. Association for Computational Linguistics, 2024, S. 7346–7356. DOI: 10.48550/arXiv.2402.05201. arXiv: 2402.05201.
- [65] Reuters. *Amazon hit with record EU data privacy fine*. Juli 2021. URL: https://www.reuters.com/business/retail-consumer/amazon-hit-with-886-million-eu-data-privacy-fine-2021-07-30/?utm_source=chatgpt.com (besucht am 02.10.2025).
- [66] Konrad Schneid u. a. „Uncovering data-flow anomalies in BPMN-based process-driven applications“. In: *Proceedings of the 36th Annual ACM Symposium on Applied Computing*. 2021, S. 1504–1512. URL: <https://dl.acm.org/doi/abs/10.1145/3412841.3442025>.
- [67] Torsten Scholak, Nathan Schucher und Dzmitry Bahdanau. „PICARD: Parsing Incrementally for Constrained Auto-Regressive Decoding from Language Models“. In: *CoRR* abs/2109.05093 (2021). arXiv: 2109.05093. URL: <https://arxiv.org/abs/2109.05093>.
- [68] Magdalena von Schwerin und Manfred Reichert. „A systematic comparison between open-and closed-source large language models in the context of generating gdpr-compliant data categories for processing activity records“. In: *Future Internet* 16.12 (2024), S. 459. URL: <https://www.mdpi.com/1999-5903/16/12/459>.
- [69] Marina Sokolova und Guy Lapalme. „A systematic analysis of performance measures for classification tasks“. In: *Information processing & management* 45.4 (2009), S. 427–437. URL: <https://www.sciencedirect.com/science/article/pii/S0306457309000259>.
- [70] Ángel Jesús Varela-Vaca u. a. „Business process models and simulation to enable GDPR compliance“. In: *International Journal of Information Security* 24.1 (2025), S. 41. URL: <https://link.springer.com/article/10.1007/s10207-024-00952-7>.

- [71] Ashish Vaswani u. a. „Attention Is All You Need“. In: *CoRR* abs/1706.03762 (2017). arXiv: 1706.03762. URL: <http://arxiv.org/abs/1706.03762>.

Name: Merten Dieckmann

Matrikelnummer: 1058340

Erklärung

Ich erkläre, dass ich die Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Ulm, den

Merten Dieckmann