BACHELORARBEIT

EINFLUSS VON KI-UNTERSTÜTZUNG AUF TEXTPRODUKTIONSPROZESSE ZUR ERSTELLUNG VON ANFORDERUNGSSPEZIFIKATIONEN IM VERGLEICH ZU TRADITIONELLEN ANSÄTZEN MIT FOKUS AUF DEN DETAILLIERUNGSGRAD

ausgeführt an der

CAMPUS 02 GRAZ
FACHHOCHSCHULE DER WIRTSCHAFT

am Studiengang
Wirtschaftsinformatik

Von: Dipl.-Ing. Dr. techn. Petra Unger
Personenkennzeichen: 00330188

Graz, am 13.05.2026

.....
Unterschrift

EHRENWÖRTLICHE ERKLÄRUNG

Ich erkläre ehrenwörtlich, dass ich

- die vorliegende Arbeit selbständig und ohne fremde Hilfe verfasst,
- andere als die angegebenen Quellen nicht benutzt,
- die den Quellen wörtlich oder inhaltlich entnommenen Stellen als solche kenntlich
- gemacht,
- den Einsatz von generativen KI-Modellen (z.B. ChatGPT) kenntlich gemacht
- und mich sonst keiner unerlaubten Hilfsmittel bedient habe.

Die Arbeit wurde bisher in gleicher oder ähnlicher Form keiner anderen Prüfungsbehörde vorgelegt und auch noch nicht veröffentlicht. Die vorliegende Fassung entspricht der eingereichten elektronischen Version.

Graz, am 13.05.2026

Ort, Datum Unterschrift

EINSATZ VON GENERATIVEN KI-MODELLEN

Im Rahmen der vorliegenden Bachelorarbeit wurde Künstliche Intelligenz (KI) sowohl als unterstützendes Werkzeug im wissenschaftlichen Arbeitsprozess als auch als methodisches Instrument innerhalb der empirischen Untersuchung eingesetzt. Die Nutzung von KI umfasste insbesondere die Unterstützung bei der Strukturierung der Arbeit, der sprachlichen Ausarbeitung sowie der Entwicklung von Formulierungen. Sämtliche durch KI generierte Inhalte wurden dabei einer kritischen Prüfung unterzogen, eigenständig überarbeitet und in den wissenschaftlichen Kontext eingeordnet, sodass die inhaltliche Verantwortung sowie die Einhaltung der Grundsätze guter wissenschaftlicher Praxis ausschließlich bei der Autorin verbleiben.

Darüber hinaus fand KI im Rahmen der Durchführung der Studie Anwendung. Konkret wurden Large Language Models (LLMs) im Sinne eines „LLM-as-a-Judge“-Ansatzes verwendet, um Bewertungen von generierten Inhalten bzw. erhobenen Daten vorzunehmen. Die Modelle fungierten hierbei als standardisierte und nachvollziehbare Bewertungsinstanz, wodurch eine konsistente und reproduzierbare Beurteilung ermöglicht wurde.

DANKSAGUNG

Ich danke meinem Betreuer, Herrn Dipl.-Ing. Simon A. T. Jiménez, MA, für die Überlassung des spannenden Themas sowie für seine kontinuierliche Unterstützung bei der Erstellung dieser Abschlussarbeit.

Mein Dank gilt außerdem der ireo GmbH für das Sponsoring von Gutscheinen für die Teilnehmer:innen meiner empirischen Studie.

Meinen Studienkolleg:innen sowie allen Freiwilligen, die an meiner Studie mitgewirkt haben, danke ich herzlich für ihre Unterstützung. Ohne euch wären das Studium und diese abschließende Arbeit nicht dieselben gewesen.

Ganz besonders möchte ich mich bei meinem Mann Thomas bedanken. Ohne dich wäre dieses Studium nicht möglich gewesen! Danke für deine Geduld, deine fachliche und emotionale Unterstützung und deine bedingungslose Rücksichtnahme während dieser intensiven Zeit.

Meinen Kindern Alexander, Miriam, Konstantin und Valentina danke ich für ihr Verständnis und ihre Geduld in den letzten Jahren. Nun kommt wieder etwas mehr Familienzeit auf euch zu!

Ein herzliches Dankeschön geht an meine Eltern für ihre positive Einstellung zu meinem späten (weiteren) Studium und für unzählige Stunden liebevoller Kinderbetreuung, ohne die ich dieses Ziel nicht erreichen hätte können.

KURZFASSUNG

Die Qualität von Anforderungsspezifikationen stellt einen entscheidenden Erfolgsfaktor in Softwareentwicklungsprojekten dar, wobei insbesondere der Detaillierungsgrad einen wesentlichen Einfluss auf die Umsetzung und spätere Systemqualität hat. Gleichzeitig gewinnen KI-gestützte (Künstliche Intelligenz-gestützte) Werkzeuge zunehmend an Bedeutung im Requirements Engineering, da sie das Potenzial bieten, textbasierte Spezifikationen effizienter und konsistenter zu erstellen.

Vor diesem Hintergrund untersucht die vorliegende Bachelorarbeit den Einfluss von KI-Unterstützung auf den Detaillierungsgrad von Softwarespezifikationen im Vergleich zu klassischen, manuellen Ansätzen.

Zur Beantwortung der Forschungsfrage wurde eine empirische Evaluationsstudie durchgeführt, in der zwei Versuchsgruppen Anforderungen auf Basis einer identischen Aufgabenstellung erstellten. Während eine Gruppe ein KI-gestütztes Tool (Storywise) nutzte, arbeitete die Vergleichsgruppe mit einem konventionellen Textverarbeitungsprogramm (Microsoft Word). Die resultierenden Anforderungen wurden anhand eines definierten Kriterienmodells bewertet, das die Dimensionen Klarheit, Vollständigkeit, Prüfbarkeit, Kontextbezug und Granularität umfasste. Die Bewertung erfolgte mithilfe eines LLM-as-a-Judge-Ansatzes (Large Language Model as a Judge) und wurde durch eine qualitative Analyse ergänzt.

Die Ergebnisse zeigen, dass die KI-gestützte Gruppe im Durchschnitt einen höheren Detaillierungsgrad erreicht, insbesondere in den Kriterien Klarheit, Vollständigkeit und Prüfbarkeit. Gleichzeitig weisen beide Gruppen ähnliche Werte im Kontextbezug und in der Granularität auf, was darauf hindeutet, dass diese Aspekte weiterhin stark von menschlicher Expertise abhängen. Darüber hinaus zeigt sich eine geringere Streuung der Ergebnisse in der KI-gestützten Gruppe, was auf eine standardisierende Wirkung der KI hinweist.

Insgesamt verdeutlicht die Arbeit, dass KI-gestützte Werkzeuge einen positiven, jedoch selektiven Beitrag zur Verbesserung der Anforderungsqualität leisten können. Sie unterstützen insbesondere die sprachliche und formale Ausgestaltung von Anforderungen, ersetzen jedoch nicht die fachliche Kompetenz im Requirements Engineering.

ABSTRACT

The quality of requirements specifications is a crucial success factor in software development projects, with the level of detail playing a significant role in implementation and overall system quality. At the same time, AI-supported (artificial intelligence-supported) tools are gaining importance in requirements engineering, as they offer the potential to generate text-based specifications more efficiently and consistently.

Against this background, this bachelor thesis investigates the impact of AI support on the level of detail in software specifications compared to traditional, manual approaches.

To address the research question, an empirical evaluation study was conducted in which two groups created requirements based on an identical task. One group used an AI-supported tool (Storywise), while the control group worked with a conventional word processing application (Microsoft Word). The resulting requirements were evaluated using a defined criteria model covering clarity, completeness, verifiability, contextual reference, and granularity. The evaluation was carried out using an LLM-as-a-Judge (Large Language Model as a Judge) approach and complemented by qualitative analysis.

The results show that the AI-supported achieved a higher average level of detail, particularly in terms of clarity, completeness, and verifiability. At the same time, both groups showed similar results regarding contextual reference and granularity, indicating that these aspects remain strongly dependent on human expertise. Furthermore, the AI-supported group exhibited less variation in results, suggesting a standardizing effect of AI.

Overall, the findings demonstrate that AI-supported tools can provide a positive but selective contribution to improving requirements quality. They primarily support the linguistic and structural formulation of requirements, while domain knowledge and contextual understanding remain essential human responsibilities in requirements engineering.

INHALTSVERZEICHNIS

1. EINLEITUNG

„Research has repeatedly shown that high-quality requirements are essential for the success of development projects“ (Montgomery et al., 2022, S. 183). Nach Montgomery et al. (2022) spielt die Qualität von Softwarespezifikationen somit eine wesentliche Rolle für erfolgreiche Softwareentwicklungsprojekte.

Der Grad ihrer Detaillierung beeinflusst sowohl die technische Umsetzung als auch die späteren Projektkosten, weshalb Unternehmen zunehmend auf Methoden und Werkzeuge angewiesen sind, die eine strukturierte und effiziente Spezifikationserstellung unterstützen (Franch et al., 2023). Gleichzeitig gewinnen KI-gestützte (Künstliche Intelligenz-gestützte) Systeme an Bedeutung, da sie versprechen, textuelle Artefakte schneller und einheitlicher erzeugen zu können (Knuplesch, 2024).

In der Praxis werden Spezifikationen jedoch weiterhin häufig mit klassischen Textverarbeitungsprogrammen wie Microsoft Word erstellt (Franch et al., 2023). Vergleiche zwischen KI-gestützt generierten Software Requirements Specification (SRS) Documents und manuell erstellten Vergleichsdokumenten wurden in der Literatur beschrieben, beziehen sich jedoch auf den direkten Einsatz von Large Language Models (LLMs), beispielsweise GPT-4 und CodeLlama (vgl. z.B. Krishna, Gaur, Verma, Jalote, 2024).

Ziel der in der vorliegenden Arbeit genutzten Applikation storyw!se (im Folgenden zur besseren Lesbarkeit als Storywise notiert) ist es, den Prozess der Spezifikationserstellung durch KI-basierte Generierung und Strukturierung zu verbessern und zusätzlich Requirements Management, Kompatibilitäts- und Kooperationssicherstellung zur Verfügung zu stellen (storywise, n.d.). Ob die Nutzung derartiger Tools tatsächlich zu detaillierteren Spezifikationen führt, ist bislang nicht empirisch untersucht.

1.1 Zielsetzung und Forschungsfrage

Obwohl umfangreiche Forschung zur Qualität von Anforderungen existiert, konzentrieren sich bestehende Arbeiten überwiegend auf traditionelle, manuell erstellte Spezifikationen und deren Qualitätsmerkmale (Montgomery et al., 2022).

Der spezifische Einfluss von KI-gestützten Ansätzen auf die Qualität von Anforderungen, insbesondere im Hinblick auf den Detaillierungsgrad, ist bislang nur unzureichend empirisch untersucht. Zwar befassen sich aktuelle Studien mit dem Einsatz von LLMs im Requirements

Engineering, diese fokussieren jedoch primär auf die Generierung von Anforderungen und weniger auf deren qualitative Ausprägung oder strukturelle Eigenschaften (Krishna et al., 2024). Die Masterarbeit von Knuplesch (2024) liefert einen ersten Einblick in die qualitative Bewertung von KI-gestützten Tools anhand der Applikation Storywise, beinhaltete jedoch keine quantitative Bewertung der erhaltenen Anforderungen. Eine systematische Analyse der Auswirkungen von KI auf zentrale Qualitätsdimensionen, wie Klarheit, Vollständigkeit oder Granularität, fehlt somit bislang.

Darüber hinaus mangelt es an empirischen Vergleichen zwischen konkreten Werkzeugen. Obwohl in der Praxis eine Vielzahl an Tools für die Erstellung von Anforderungsspezifikationen eingesetzt wird, liegen kaum Studien vor, die klassische Ansätze und KI-gestützte Systeme hinsichtlich ihrer Wirkung auf die Ergebnisqualität direkt gegenüberstellen (Franch et al., 2023). Vor diesem Hintergrund ergibt sich eine Forschungslücke hinsichtlich eines quantitativen, toolbasierten Vergleichs der Auswirkungen von KI-Unterstützung auf den Detaillierungsgrad von Anforderungen.

Die zentrale Forschungsfrage lautet daher:

Welchen Einfluss hat KI-Unterstützung auf den Detaillierungsgrad

von Softwarespezifikationsanforderungen im Vergleich zu klassischen Methoden?

1.2 Methodik

Die vorliegende Arbeit ist der Evaluationsforschung zuzuordnen (Döring et al., 2023). Im ersten Schritt der Arbeit sollen zunächst die Evaluationskriterien entwickelt werden, die den Detaillierungsgrad von Anforderungsspezifikationen messbar abbilden. Darauf aufbauend erfolgt eine kontrollierte Vergleichsstudie mit zwei Gruppen. Eine der Gruppen setzt die KI-gestützte Applikation Storywise und die andere Gruppe das konventionelle Texterstellungsprogramm Microsoft Word zur Spezifikationserstellung ein. Die erzeugten Anforderungen der Spezifikationen werden anschließend anhand der gewählten Evaluationskriterien beurteilt und systematisch gegenübergestellt.

1.2.1 Evaluationsforschung

Nach Döring et al. (2023, S. 955) setzt Evaluationsforschung („evaluation research“) bzw. wissenschaftliche Evaluation („evaluation“) sozialwissenschaftliche Methoden ein, um einen Evaluationsgegenstand zu bewerten. Zu berücksichtigen sind hierbei die relevanten Stakeholder (Anspruchsgruppen). Die Bewertung erfolgt anhand zuvor definierter Evaluationskriterien. Diese können beispielsweise Akzeptanz, Wirksamkeit, Effizienz oder Nachhaltigkeit sein. Ebenfalls

notwendig ist es, Maßgaben zur Ausprägung der Evaluationskriterien festzulegen. In der Praxis ist die so erhaltene Bewertung dazu bestimmt, verschiedene Evaluationsfunktionen zu erfüllen.

1.2.1.1 Evaluationsgegenstand

Döring et al. (2023, S. 956) definieren den Evaluationsgegenstand bzw. das Evaluationsobjekt („evaluation object“, „evaluand“) als den Gegenstand der Untersuchung und Bewertung mittels wissenschaftlicher Methoden. Dies können beispielsweise Projekte, Produkte oder auch natürliche Personen sein.

In der vorliegenden Arbeit dienen die in der praktischen Studie (vgl. Kapitel 3.3) erhaltenen Anforderungen als Evaluationsgegenstand.

1.2.1.2 Evaluationskriterien

Evaluationskriterien („evaluation criteria“) können sowohl auf das Konzept, die Durchführung als auch die Ergebnisse einer Maßnahme oder eines Evaluationsobjektes angewandt werden. Sie können unterschiedlicher Natur sein, z.B. ökonomische, ökologische, soziale oder ethische Kriterien (Döring et al., 2023, S. 962).

Die Evaluationskriterien, anhand welcher der für die vorliegende Bachelorarbeit relevante Evaluationsgegenstand bewertet werden soll, sind in Kapitel 3.2 ausführlich beschrieben.

1.2.2 LLM-as-a-Judge

Zur Unterstützung der Bewertung der in der Studie erhobenen Anforderungen wurde ein LLM als Assistenzinstrument eingesetzt. Die Bewertung erfolgte auf Basis eines fest definierten, kriteriengeleiteten Prompts (vgl. Anhang C). Die Ergebnisse wurden manuell geprüft, sodass keine automatisierte Entscheidungsdelegation stattfand. Dieser Ansatz kann als LLM-as-a-Judge mit menschlicher Kontrolle bezeichnet werden.

1.2.3 Vorgehensweise

Für die vorliegende Arbeit wurde zur Generierung des Evaluationsgegenstands eine empirische Studie durchgeführt, im Rahmen derer Personen mit beruflicher Erfahrung im Bereich Anforderungsspezifikationen anhand eines vorgegebenen Beispiels mit bzw. ohne Nutzung eines vorgegebenen KI-gestützten Tools Anforderungen schreiben sollten. Diese Anforderungen wurden im Anschluss durch LLM-as-a-Judge (vgl. Kapitel 2.3) bewertet und diese Bewertung nachfolgend kritisch diskutiert.

1.3 Aufbau der Arbeit

Nach der Einleitung der vorliegenden Arbeit im aktuellen Kapitel folgt in Kapitel 2 ein Überblick über die theoretischen Grundlagen. Es werden Hintergründe zum Detaillierungsgrad von Softwarespezifikationen bzw. -anforderungen sowie grundlegende Definitionen zu NLP und LLM diskutiert. Nachfolgend soll auf LLM-as-a-Judge als zentrales Werkzeug für die in dieser Arbeit durchgeführte Bewertung von Anforderungen besonders eingegangen werden. Im letzten Abschnitt von Kapitel 2 wird die KI-gestützte Applikation Storywise beschrieben.

Kapitel 3 erläutert den Rahmen und die Kriterien der durchgeführten Evaluationsstudie, insbesondere Evaluationsgegenstand und -kriterien. Nach der Darstellung der praktischen Durchführung der Studie wird auf die Online-Umfrage zur Erhebung statistischer Daten der Studienteilnehmer:innen eingegangen.

Kapitel 4 stellt die Ergebnisse des praktischen Teils der vorliegenden Arbeit dar und diskutiert sie. Der Fokus liegt hierbei auf den Ergebnissen der Bewertung der einzelnen Anforderungen. Es werden jedoch auch einige statistische Informationen aus der Online-Umfrage systematisch bewertet.

Abschließend erfolgt in Kapitel 5 eine Conclusio der Ergebnisse und ein Ausblick auf weitere mögliche Forschung zu aufbauenden oder angrenzenden Themen.

2. THEORETISCHER HINTERGRUND

Definitionen in Bezug auf Requirements Engineering, User Stories, Natural Language Processing (NLP) und LLMs wurden in der Masterarbeit von Knuplesch (2024) dargestellt und in Bezug auf die Applikation Storywise diskutiert. Daher fokussiert der Theorieteil der vorliegenden Arbeit auf die Aspekte Anforderungsspezifikationen und Detaillierungsgrad sowie LLM-as-a-Judge. Einige grundlegende Definitionen zu NLP, LLM und Storywise als Applikation sollen dennoch im Sinne der Konsistenz der vorliegenden Arbeit festgehalten werden.

2.1 Anforderungsspezifikationen und Qualitätskriterien

Glinz (2005, S. 91) definiert Anforderungsspezifikation wie folgt: „Anforderungsspezifikation. Die Zusammenstellung aller Anforderungen an eine Software. Synonyme: Anforderungsdokument, Software Requirements Specification."

Eine ausführlichere Definition findet sich bei Glinz et al. (2022, S. 10), die Anforderungsspezifikationen als systematisch dargestellte Sammlung von Anforderungen beschreiben, die vorgegebene Kriterien erfüllt.

Als Beispiel für Qualitätskriterien für Anforderungsspezifikationen sei die Publikation von Pohl und Rupp (2021, S. 115) zitiert: „Laut dem Standard [ISO/IEC/IEEE 29148:2018] sollte ein Anforderungsdokument vollständig und konsistent sein. Darüber hinaus sollte ein Anforderungsdokument durch eine klare Struktur, einen angemessenen Umfang und Nachvollziehbarkeit die Lesbarkeit unterstützen."

Glinz et al. (2022) nennen die folgenden Qualitätseigenschaften für Anforderungsspezifikationen:

- Konsistent (Anforderungen dürfen einander nicht widersprechen),
- Nicht redundant (Keine Überschneidungen zwischen Anforderungen),
- Vollständig (Alle relevanten Anforderungen enthalten),
- Modifizierbar (Modifikation ohne Qualitätsverlust möglich),
- Verfolgbar (Änderungen an der Spezifikation rückverfolgbar),
- Konform (Konformität zu verbindlichen Strukturierungs- oder Formatierungsanweisungen).

2.1.1 Abgrenzung Detaillierungsgrad – Granularität

Für die Beantwortung der in Kapitel 1.1 vorgestellten Forschungsfrage soll der Detaillierungsgrad der Anforderungen der Anforderungsspezifikationen als Qualitätskriterium untersucht werden.

Wie beispielsweise Simões und Vazquez (2017) ausführen, werden die Begriffe Detaillierungsgrad und Granularität in der Fachliteratur teilweise synonym verwendet. Daher soll für den Kontext der vorliegenden Arbeit eine abgrenzende Definition gegeben werden.

Granularität soll hier in Anlehnung an Bühne und Herrmann (2024) die Nutzung unterschiedlicher Ebenen für die Beschreibung von Anforderungen bezeichnen. Die Autoren definieren: „Epics sind Beschreibungen von Anforderungen auf einem höheren Detaillierungsgrad als User Stories. Sie fassen daher üblicherweise mehrere User Stories zusammen“ (Bühne & Herrmann, 2024, S. 214).

Hruschka et al. (2025) stellen die unterschiedlichen Granularitätsstufen wie in Abb. 1 gezeigt dar.

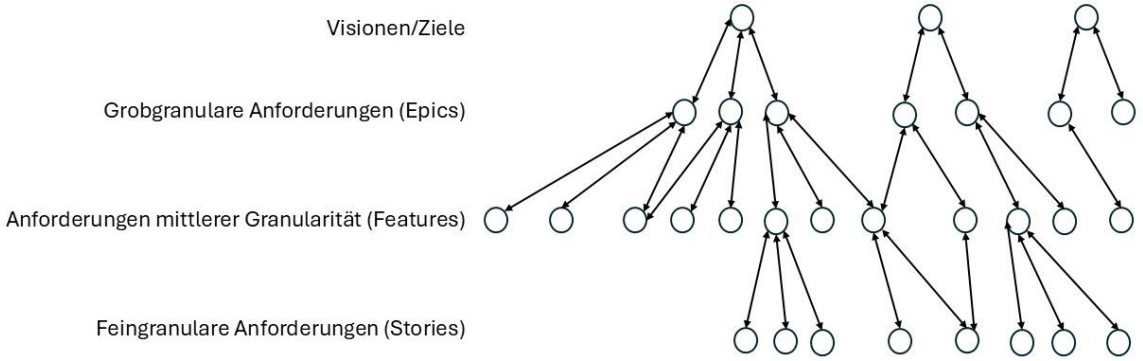

Abbildung 1: Unterschiedliche Granularitätsstufen von Anforderungen (eigene Darstellung in Anlehnung an Hruschka et al., 2025).

In Kontrast zur Granularität soll Detaillierungsgrad in der vorliegenden Arbeit das Komplexitätslevel der jeweiligen Anforderung bezeichnen. Granularität per se wird als eines der fünf zum Detaillierungsgrad beitragenden Faktoren definiert (vgl. Evaluationskriterien in Kapitel 3.2).

2.1.2 Abgrenzung Detaillierungsgrad – Detaillierungsebenen

Abzugrenzen von Detaillierungsgrad im Sinne der vorliegenden Arbeit ist auch die Definition von Detaillierungslevels nach Bühne und Herrmann (2024). In dieser Publikation wird die Detaillierungsebene (bzw. Abstraktionsebene) als Klassifizierungswerkzeug für Anforderungen beschrieben: „In vielen Projekten haben sich drei Detaillierungsebenen bewährt – auch wenn diese in der Regel immer anders benannt sind, scheint dies ein praktikables Level zu sein (z.B. Ebene der Produkt- Anforderungen, Ebene der Benutzer-Anforderungen, Ebene der System-Anforderungen)“ (Bühne & Herrmann, 2024, S. 30).

Hierbei erfolgt – analog zur Granularität – ein schematischer Zugang, der fixe Kategorien (Ebenen) nutzt, um Detailgrade zu beschreiben (Bühne & Herrmann, 2024).

Die Verwendung des Terminus Detaillierungsgrad in der vorliegenden Arbeit weicht – wie zuvor diskutiert – davon ab. Weiterführende Erläuterungen zu den Evaluationskriterien finden sich in Kapitel 3.2.

2.1.3 Detaillierungsgrad – Anforderungen und Beispiele

Der erforderliche Detaillierungsgrad ist nach Glinz et al. (2022) abhängig von zahlreichen Faktoren: Projektkontext, gemeinsames Verständnis des Problems, vorhandener Freiheitsgrad für die Designer und Programmierer, Verfügbarkeit von schnellem Stakeholder-Feedback während Konzeption und Implementierung, Kosten und Wert einer detaillierten Spezifikation, Normen und behördliche Auflagen.

Nach Glinz et al. (2022, S. 131) ist der erforderliche Detaillierungsgrad von Komplexität und Kritikalität des zu entwickelnden Systems abhängig: „Wenn ein System in Bezug auf Nutzungssicherheit (Safety) und Informationssicherheit (Security) komplex und/oder kritisch ist, muss der gewählte RE-Prozess eine detaillierte Spezifikation der kritischen Anforderungen berücksichtigen […]“.

Glinz et al. (2022, S. 38) halten ebenso fest, dass somit der eine richtige Detaillierungsgrad nicht definiert werden kann:

> Es gibt keinen allgemein „richtigen” Detaillierungsgrad für Anforderungen. Für jede Anforderung hängt der angemessene Detaillierungsgrad von vielen Faktoren ab. Je detaillierter die Anforderungen spezifiziert sind, desto geringer ist das Risiko, dass am Ende etwas Unerwartetes herauskommt, bzw. Features oder Eigenschaften fehlen. Die Kosten für die Spezifikation steigen jedoch mit zunehmender Detaillierung (Glinz et al., 2022, S. 38).

Als ein Beispiel für eine Detaillierungsgradanalyse aus der Literatur sei Absar ul Hasan und Rana (2022) genannt: Die Autoren schlagen einen NLP-basierten Klassifikationsansatz vor, der Anforderungen automatisch in drei Detaillierungsstufen einteilt. High-Level ist gekennzeichnet durch wenige Details (sehr allgemeine Ausführung), Intermediate-Level durch mehrere Aufgaben, die aber nicht vollständig beschrieben sind, Low-Level sind ausreichend detaillierte Anforderungen, die direkt implementiert werden können.

Diese Einteilung basiert auf dem Structural depth metric-Ansatz (SDM-Ansatz), der die Ausgewogenheit des Detailgrades eines Anforderungssatzes in Bezug auf die drei oben genannten Levels beschreibt und visualisiert (Laplante, 2013, zitiert nach Absar ul Hasan & Rana, 2022).

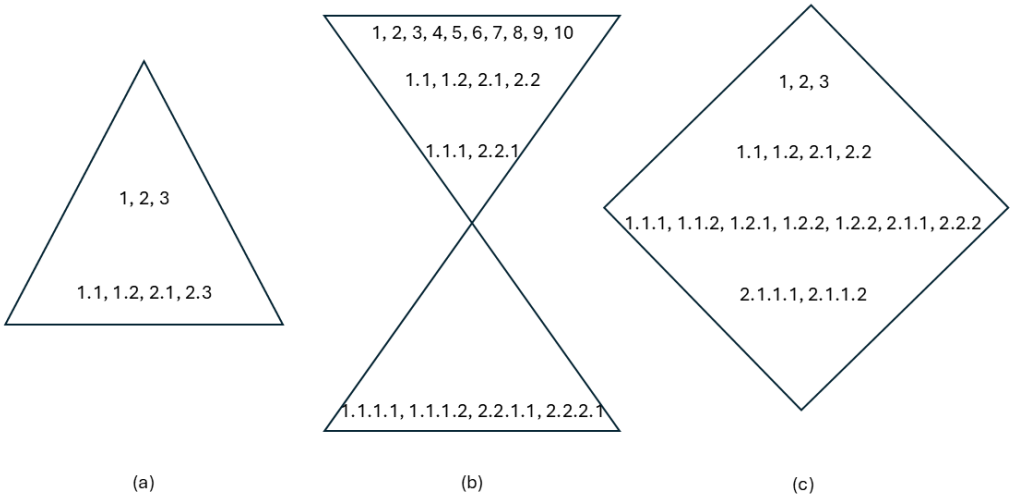

Abbildung 2: SDM-Visualisierung (eigene Darstellung in Anlehnung an Laplante, 2013, zitiert nach Absar ul Hasan & Rana, 2022).

Eine pyramidenförmige Ausprägung der Anforderungen (vgl. (a) in Abb. 2) beschreibt hierbei ein konsistentes Level an Details und entsteht durch eine adäquate Anzahl an High-Level-Anforderungen im oberen Bereich der Darstellung, gefolgt von einer höheren Anzahl an Intermediate- und Low-Level-Anforderungen im unteren Teil der Pyramide. Die Sanduhr-förmige Darstellung (vgl. (b) in Abb. 2) weist auf eine hohe Anzahl an vorhandenen administrativen Anforderungen hin. Ähnelt die Anzahl der Anforderungen in der Darstellung einem Diamanten (vgl. (c) in Abb. 2), ist die Erfassung von mehr High-Level- und Low-Level-Anforderungen indiziert (vgl. Laplante, 2013, zitiert nach Absar ul Hasan & Rana, 2022).

Die Verarbeitungsschritte der SRS nach Absar ul Hasan und Rana (2022) gliedern sich in NLP-Vorbereitungsschritte (Tokenisierung, Stop-Word-Entfernung, Lemmatisierung, POS-Tagging), Erzeugung von Features/Merkmalen (TF-IDF-Features bzw. POS-Frequency-Features), Modelltraining/Klassifikation und Modellevaluation. Anschließend erfolgt die Visualisierung über SDM (vgl. Abb. 2) und die Interpretation der Ergebnisse.

Wie beschrieben soll die vorliegende Arbeit – in Abgrenzung zu den in diesem Kapitel beschriebenen Methoden – auf der Ebene der Anforderungsbewertung durchgeführt werden.

2.2 NLP und LLM

NLP ist nach Ferrari und Ginde (2025) ein multidisziplinäres Gebiet im Bereich zwischen künstlicher Intelligenz und Linguistik. Das Ziel hinter NLP ist es, dass Computer Sprache prozessieren, verstehen, interpretieren und generieren können. Die Ergebnisse sollen hierbei für den Menschen nützlich und sinnvoll sein. NLP betrifft alle Aspekte von Sprache, beispielsweise Syntax, Semantik, Pragmatik und Phonetik.

Ferrari und Ginde (2025) beschreiben NLP4RE (Natural language processing for requirements engineering) als Einsatz von NLP-Techniken für das Requirements Engineering, beispielsweise für Klassifizierungsaufgaben, Sicherstellung der Rückverfolgbarkeit (z.B. zwischen unterschiedlichen Anforderungen) oder Fehlererkennung in Anforderungsspezifikationen.

Der Terminus LM (Language Model) beschreibt ein Rechenmodell, das die statistischen Eigenschaften natürlicher Sprache erlernt und verwendet, um die Wahrscheinlichkeit von Wortsequenzen in einem gegebenen Kontext vorherzusagen (Ferrari & Ginde, 2025).

Ferrari und Ginde (2025) definieren weiters Bedeutung und Eigenschaften von LLMs: Ein LLM stellt ihrer Publikation folgend eine weiterentwickelte und rechnerisch besonders aufwendige Form eines LM dar, das in der Regel auf Deep-Learning-Architekturen wie Transformern basiert. Zwar verfolgt es ähnliche Ziele wie klassische Sprachmodelle, unterscheidet sich jedoch durch mehrere zentrale Merkmale: Zum einen basiert es auf einer sehr großen Skalierung, da es mit enorm umfangreichen Textmengen trainiert wird, um auch feinste sprachliche Nuancen erfassen zu können. Darüber hinaus verfügt ein LLM über eine hohe Modellkapazität in Form einer großen Anzahl von Parametern, wodurch komplexe sprachliche Muster und Zusammenhänge besonders präzise gelernt werden können.

In Bezug auf die Leistungsfähigkeit erreichen solche Modelle häufig Ergebnisse auf dem aktuellen Stand der Technik in unterschiedlichen NLP-Anwendungen, beispielsweise bei Übersetzungen, Zusammenfassungen oder der Beantwortung von Fragen. Ein weiterer entscheidender Aspekt ist die Anpassungsfähigkeit: Aufgrund des umfangreichen Trainings lassen sich LLMs gezielt für spezifische Aufgaben oder Fachbereiche weiter optimieren, was ihre Einsatzmöglichkeiten erheblich erweitert.

Drei Architekturen von LLMs werden nach Ferrari und Ginde (2025) unterschieden:

- Encoder-only (z.B. BERT, RoBERTa, DistilBERT, USE)
- Decoder-only (z.B. GPT-4, GPT-5, Llama 2)
- Encoder-Decoder (z.B. T5)

Die Funktionalität der einzelnen Architekturen ist nach Ferrari und Ginde (2025) in Abbildung 3 dargestellt.

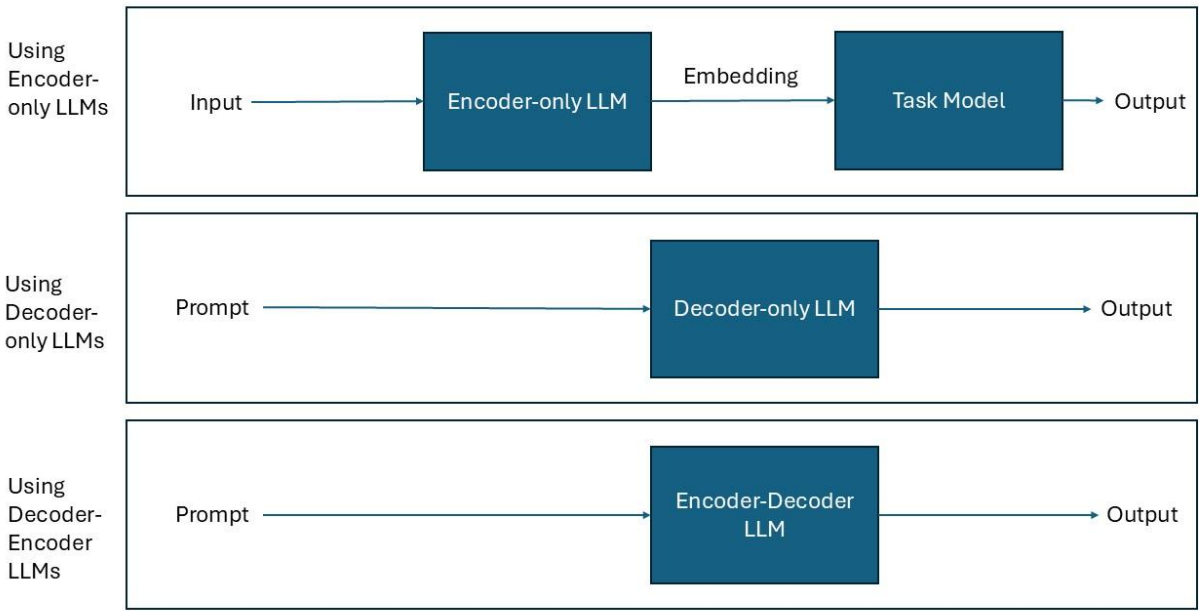

Abbildung 3: Funktionalität der unterschiedlichen LLM-Architekturen (eigene Darstellung in Anlehnung an Ferrari & Ginde, 2025).

2.3 LLM-as-a-Judge

In der vorliegenden Arbeit wird ein LLM in Form von LLM-as-a-Judge für die Bewertung von einzelnen Anforderungen genutzt. Der Einsatz von LLM-as-a-Judge wird in der aktuellen Forschung zunehmend untersucht und validiert (Zheng et al., 2023).

Gu et al. (2025) beschreiben den Einsatz großer Sprachmodelle als automatisierte Bewertungsinstanzen. Es wird erklärt, dass – vor Aufkommen von LLMs – das Gleichgewicht zwischen umfassender und skalierbarer Evaluation eine Herausforderung war. Traditionelle automatische Bewertungsmetriken aus der Computerlinguistik wie ROUGE oder BLEU, die schon lange vor dem Aufkommen von LLMs verwendet wurden, um die Qualität von maschinell erzeugten Texten zu beurteilen, werden als Werkzeuge beschrieben, die Vergleiche mit Referenztexten schnell und ohne menschliches Zutun durchführen können (Gu et al., 2025).

Nach SuperAnnotate (n.d.) sind ROUGE (Recall-Oriented Understudy for Gisting Evaluation) und BLEU (Bilingual Evaluation Understudy) zwei anerkannte Modelle für die Bewertung der Qualität maschinell erzeugter Texte. Sie ermöglichen eine Bewertung der Ähnlichkeit von maschinell erzeugten Texten mit von Menschen bereitgestellten Referenztexten und somit eine Bewertung der Leistungsfähigkeit von NLP-Systemen (SuperAnnotate, n.d.).

Nach Gu et al. (2025) basieren diese Metriken jedoch stark auf oberflächlichen lexikalischen Überlappungen (z.B. direkter Abgleich von Wörtern bzw. Phrasen), weshalb sie bei komplexen, kreativen oder stärker qualitativ geprägten Texten an ihre Grenzen stoßen.

Die Nutzung von LLM-as-a-Judge bietet demgegenüber stark erweiterte Möglichkeiten. Hierbei kommt dem Prompting eine besondere Bedeutung zu.

Die Forschungsgruppe um Gu beschreibt vier Methoden, wie Prompts für LLM-as-a-Judge konzipiert sein können:

- Generating scores (Bewertung anhand einer Skala)
- Solving Yes/No-Questions (Bewertung einer Aussage mit Ja/Nein oder Wahr/Falsch)
- Pairwise Comparison (Paarweiser Vergleich)
- Multiple-Choice Selection (Einsatz mehrerer Methoden, Auswahl durch das LLM) (Gu et al., 2025).

Hinsichtlich Modellauswahl unterscheiden Gu et al. (2025) zwischen General LLM (z.B. GPT-4) und Fine-tuned LLM (z.B. PandaLM).

Die Publikation beschreibt ebenso die Ebenen einer Bewertung des LLM-as-a-Judge-Modells: Hierbei werden die Faktoren Agreement with Human Judgements (Übereinstimmung mit durch Menschen erfolgte Bewertungen), Überprüfung auf Bias Elicitation (systematische Verzerrungen) und Bewertung der Adversarial Robustness (Widerstandsfähigkeit gegen Manipulationsversuche) berücksichtigt. Einen vertieften Einblick in die Themenfelder Bias Elicitation und Adversarial Robustness sowie einen Ansatz für ein entsprechendes Benchmarking liefern Cantini et al. (2025).

Gu et al. (2025) betonen weiters, dass die Qualität von LLM-as-a-Judge weniger von seiner reinen Modellgröße abhängt als von der Gestaltung des Bewertungsprozesses.

Aufgrund der hohen Qualität der Ergebnisse bei Einsatz von LLM-as-a-Judge und der gleichzeitig nicht vollständig eliminierbaren Risiken dieser Beeinflussung bzw. Verzerrung wird in der Literatur auch die Nachweisbarkeit von Produkten von LLM-as-a-Judge (in Abgrenzung zu z.B. von Menschen verfassten Bewertungen) diskutiert. Li et al. schlugen beispielsweise einen von den Autoren so bezeichneten J-Detector vor, der eine derartige Analyse durchführen kann (Li et al., 2025).

2.4 Storywise

Die Applikation Storywise ist ein KI-gestütztes Werkzeug für die Erstellung von Anforderungsspezifikationen im Requirements Engineering. Dieses wurde durch das Partnerunternehmen der Autorin, die ireo GmbH, entwickelt (storywise, n.d.).

Die folgenden Funktionen basieren auf Angaben der offiziellen Website (storywise, n.d.) und auf Erfahrungen der Autorin mit dem Tool.

Ziel des Tools ist es, den traditionellen, häufig manuellen und wenig strukturierten Prozess der Spezifikationserstellung durch automatisierte Generierungs- und Strukturierungsmechanismen zu verbessern. Storywise kombiniert dabei Methoden des NLP und der LLMs. Die zentrale Funktion von Storywise besteht in der Transformation natürlichsprachiger Beschreibungen eines Systems in strukturierte Anforderungen.

Ausgangspunkt bildet in der Regel eine textuelle Problemdefinition oder ein Anwendungsfall, der durch den Anwender in das System eingebracht wird. Die erste Ansichtsmaske zur Eingabe des Anwendungsfalls bzw. der Problemdefinition ist in Abbildung 4 gezeigt.

Abbildung 4: Start eines neuen Projekts mit Storywise.

Storywise ermöglicht beim Start eines neuen Projekts sowohl die textuelle Beschreibung des Problems oder Anwendungsfalls als auch ein Hochladen eines Dokumentes mit entsprechendem Inhalt.

Auf dieser Basis analysiert das Tool den eingegebenen Text und identifiziert relevante Inhalte, wie Funktionalitäten, beteiligte Akteure oder Systemanforderungen. Diese Inhalte werden anschließend in strukturierte Artefakte des Requirements Engineering, beispielsweise User Stories oder funktionale Anforderungen, überführt. Ein wesentlicher Bestandteil der Funktionsweise von Storywise ist die automatische Strukturierung der Anforderungen. Dabei werden identifizierte Inhalte in hierarchische Elemente wie Epics und User Stories gegliedert.

In Abbildung 5 ist der Ablauf der Erstellung der Anforderungsspezifikationen in Storywise gezeigt. Die Schritte „Epic-Erstellung“ und „User Story-Erstellung“ referenzieren direkt auf zwei der zuvor beschriebenen Elemente.

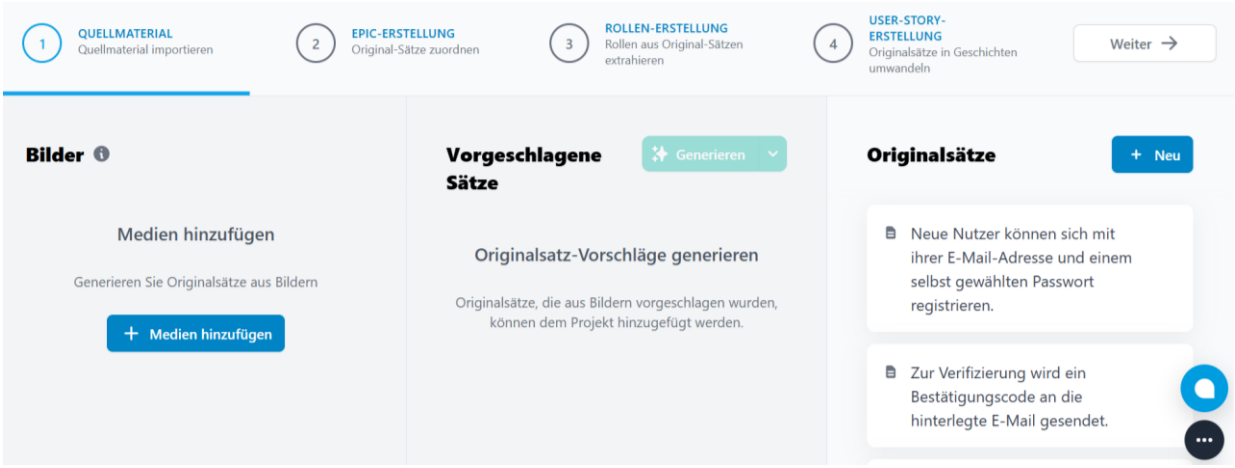

Abbildung 5: Gliederung des Ablaufs eines Projekts in Storywise.

Darüber hinaus bietet Storywise Funktionen zur automatischen Ergänzung und Verfeinerung von Anforderungen. Basierend auf trainierten Sprachmodellen kann das System implizite Inhalte explizit formulieren, fehlende Aspekte ergänzen oder Vorschläge zur Verbesserung der Formulierung liefern. Durch diese automatischen Vorschläge übernimmt das Tool eine unterstützende Rolle im Schreibprozess und wirkt als eine Art Assistenzsystem für Requirements Engineers.

Ein weiterer relevanter Aspekt ist die Unterstützung der Konsistenz und Standardisierung von Anforderungen. Durch die Verwendung vordefinierter Muster und sprachlicher Strukturen trägt Storywise dazu bei, eine einheitliche Ausdrucksweise innerhalb einer Spezifikation zu gewährleisten. Dies reduziert Mehrdeutigkeiten und erleichtert die Kommunikation zwischen unterschiedlichen Projektbeteiligten. Insbesondere in heterogenen Teams kann dies einen erheblichen Beitrag zur Qualitätssicherung leisten.

Neben der reinen Generierung von Anforderungen unterstützt Storywise auch Aspekte des Requirements Managements. Dazu gehören unter anderem die Organisation und Verwaltung von Anforderungen sowie deren Zuordnung zu übergeordneten Strukturen. Darüber hinaus kann das Tool potenziell Beziehungen zwischen Anforderungen erkennen und somit die Nachvollziehbarkeit und Abhängigkeiten innerhalb einer Spezifikation verbessern. Dies ist insbesondere für komplexe Systeme von Bedeutung, bei denen zahlreiche Anforderungen miteinander in Beziehung stehen.

Trotz der hohen Automatisierung bleibt der Mensch im Prozess eingebunden und übernimmt die Rolle eines kontrollierenden und bewertenden Akteurs. Die durch das System generierten Inhalte dienen daher als Vorschläge, die vom Benutzer überprüft, modifiziert und final freigegeben werden.

Ein weiterer funktionaler Aspekt von Storywise liegt in der Unterstützung kollaborativer Arbeitsprozesse. Da Anforderungen häufig von mehreren Personen erstellt und genutzt werden,

bietet das Tool die Möglichkeit, Anforderungen zu teilen, zu kommentieren oder gemeinsam weiterzuentwickeln. Dadurch kann Storywise nicht nur zur Erstellung, sondern auch zur Abstimmung und Kommunikation innerhalb von Projektteams beitragen.

3. EVALUATIONSSTUDIE

Um die Forschungsfrage beantworten zu können, wurde eine Evaluationsstudie durchgeführt. Die Rahmenbedingungen und Durchführungs- sowie Bewertungsparameter der Studie sind im Folgenden im Detail beschrieben.

3.1 Evaluationsgegenstand

Ziel der für die vorliegende Arbeit durchgeführten Evaluationsstudie war der Erhalt von Anforderungsspezifikationen. Die darin enthaltenen Anforderungen dienten als Evaluationsgegenstand dieser Bachelorarbeit.

3.2 Evaluationskriterien

Es wurden fünf Evaluationskriterien für die Bewertung jeder einzelnen Anforderung festgelegt.

- Klarheit ist hierbei eines der zentralen Qualitätsmerkmale: „Good requirements must be clear and verifiable because they underpin communication between multiple parties and are the basis for contracts, designs, project plans, and other product-management and engineering activities“ (IBM, n.d.).
- Vollständigkeit ist die Voraussetzung für nachvollziehbare Spezifikationen: „[…] complete: the work product covers the intended scope at the chosen level of detail“ (The Requirements Engineer, n.d.).
- Prüfbarkeit ist ebenso wie Klarheit ein zentrales Qualitätskriterium (IBM, n.d.).
- Kontextbezug wird in der Literatur ebenfalls als sehr relevant für die Qualität von Anforderungen beschrieben (The Requirements Engineer, n.d.).
- Granularität spielt ebenfalls eine wichtige Rolle, da Anforderungen weder zu grob noch zu stark technisch detailliert verfasst sein sollten. Cohn (2026) verweist beispielsweise im Kontext der Product Backlogs auf die Bedeutung dieses Kriteriums:

  When a product backlog item includes the right amount of detail, team members will feel as though they were just barely able to finish the item during the iteration. They'll feel as though enough details had been determined in advance but not so many that the creativity had been removed from the iteration (Cohn, 2026).

In ihrer Kombination erlauben diese Bewertungskriterien eine systematische, reproduzierbare und toolunabhängige Bewertung des Detaillierungsgrades von Softwarespezifikationsanforderungen.

Die Bewertung der einzelnen Anforderungen (z.B. jeder User Story) erfolgte entsprechend LLM-as-a-Judge (vgl. Kapitel 2.3) unter Nutzung der Methode Generating Scores und des General LLM GPT-5 (OpenAI, 2026). Dieses LLM ist der Gruppe der Decoder-only LLMs zuzuordnen (vgl. Kapitel 2.2). Der für die Bewertung genutzte Prompt ist in Anhang C angeführt.

Es wurde eine fünfstellige Skala für die Bewertung der einzelnen Anforderungen (z.B. User Stories) in Bezug auf die definierten Kriterien ausgewählt:

- Klarheit: Wie eindeutig ist die Anforderung formuliert?
  1 = unklar, mehrdeutig; 5 = eindeutig, ohne Interpretationsspielraum
- Vollständigkeit: Sind Zweck, Verhalten, Einschränkungen und relevante Bedingungen enthalten?
  1 = wichtige Aspekte fehlen; 5 = alle wesentlichen Bestandteile vorhanden
- Prüfbarkeit: Ist klar erkennbar, wann die Anforderung erfüllt ist?
  1 = nicht testbar; 5 = objektiv prüfbar
- Kontextbezug: Sind Abhängigkeiten, Annahmen und Umgebungsbedingungen ausreichend beschrieben?
  1 = kein Kontext; 5 = klarer, nachvollziehbarer Kontext
- Granularität: Ist die Anforderung weder zu grob noch zu technisch detailliert?
  1 = viel zu grob oder überdetailliert; 5 = passend für Planung und Umsetzung

Neben dieser Bewertung der einzelnen Kriterien je Anforderung wurde aus dem arithmetischen Mittel dieser fünf Einzelwerte ein Gesamtwert berechnet, der den Gesamtdetaillierungsgrad der jeweiligen Anforderung abbildet. Jede Anforderung liegt somit sowohl in Form einzelner Kriterienwerte als auch als aggregierter Gesamtwert vor.

Zusätzlich zur numerischen Bewertung wurde durch das LLM für jede Anforderung eine kurze textuelle Begründung ausgegeben. Diese qualitative Begründung wurde manuell auf Plausibilität geprüft, diente jedoch nicht als Grundlage für die quantitative Auswertung.

Für die Auswertung auf Personenebene wurden die Bewertungswerte aller von einer Person erstellten Anforderungen gemittelt. Dadurch ergaben sich pro Person ein Mittelwert je Kriterium sowie ein Gesamtmittelwert über alle Kriterien.

Die Gegenüberstellung der Versuchsgruppen A (KI-gestützte Erstellung mit Storywise) und B (konventionelle Erstellung) erfolgte auf Basis der gruppenweisen Mittelwerte. Hierzu wurden

zunächst die Kriterienwerte über alle in der jeweiligen Gruppe erstellten Anforderungen gemittelt. Anschließend wurde analog zur Einzelbewertung aus den fünf Kriterien ein aggregierter Gesamtwert pro Versuchsgruppe berechnet.

3.3 Praktische Durchführung der Studie

Für die Teilnahme an der praktischen Studie konnten neun Teilnehmer:innen gewonnen werden. Auswahlkriterium für die Teilnahme war ein aktives (i.S.v. Erstellen) bzw. passives (i.S.v. Nutzen als Input) Arbeiten mit Anforderungsspezifikationen im Rahmen der aktuellen beruflichen Tätigkeit.

Die Teilnehmer erhielten als Vorgabe die natürlichsprachige Beschreibung einer Zeiterfassungssoftware. Diese wurde aus Gründen der Vergleichbarkeit analog zur Masterarbeit von Knuplesch (2024) gewählt (siehe Anhang A).

Die Teilnehmer:innen wurden in zwei Gruppen eingeteilt: Gruppe A arbeitete mit der KI-gestützten Applikation Storywise, Gruppe B als Vergleichsgruppe mit der konventionellen Applikation Microsoft Word.

Die Durchführung der Studie erfolgte via Teams-Meeting. Die Dauer teilte sich auf eine allgemeine Einführung in die Studie (5 Minuten), eine Einführung in Storywise (10 Minuten), die Durchführungszeit der Erstellung der Anforderungsspezifikationen (25 Minuten) und ein Online-Feedback zu statistischen Fragen (vgl. Anhang B, 25 Minuten) auf.

Die Durchführungszeit wurde nach 25 min beendet, auch wenn das erstellte Anforderungsdokument noch nicht alle notwendigen Anforderungen enthielt. Mit den Teilnehmer:innen wurde besprochen, dass der Fokus auf der angemessenen Qualität der Anforderungen und nicht auf der Vollständigkeit des gesamten Anforderungsdokuments liegen solle. (Nota bene: Während die Vollständigkeit des gesamten Dokuments in der vorliegenden Arbeit nicht bewertet wurde, ist die Vollständigkeit jeder einzelnen Anforderungen ein wichtiges Kriterium für die Bewertung des Detaillierungsgrades!)

3.4 Erhebung statistischer Daten

Um die Teilnehmer:innen für die Interpretation der Ergebnisse besser beschreiben zu können, wurden am Ende der praktischen Studiendurchführung die Teilnehmer:innen aufgefordert, ein Online-Feedback zu statistischen Fragen zu geben.

UmfrageOnline (n.d.) bot die hierfür geeignete browserbasierte Plattform zur Erstellung von Online-Umfragen.

Der Inhalt der Umfrage ist in Anhang B nachzulesen. Die Fragen bezogen sich auf relevante berufliche Erfahrungen der Teilnehmer:innen, die Art der beruflichen Nutzung von Anforderungsspezifikationen sowie die Wahrnehmung von Storywise und die Bewertung der Einschulungsphase für die Applikation. Weiters bestand die Möglichkeit, allgemeines Feedback zu hinterlassen.

Die so erhobenen Daten wurden genutzt, um die Aussagekräftigkeit der erhaltenen Ergebnisse hinsichtlich eines möglichen Einflusses biographisch bedingter Vorerfahrungen kritisch zu hinterfragen (vgl. Kapitel 4.3).

4. ERGEBNISSE UND DISKUSSION

Dieses Kapitel stellt die Ergebnisse der im Rahmen der Evaluationsstudie erhobenen Daten dar und interpretiert diese im Hinblick auf die in Kapitel 1.1 formulierte Forschungsfrage. Ziel der Untersuchung war es, den Einfluss von KI-Unterstützung im Vergleich zu einem konventionellen Textverarbeitungswerkzeug auf den Detaillierungsgrad von Softwareanforderungen empirisch zu analysieren. Im Zentrum der Auswertung stehen dabei die anhand eines auf Evaluationskriterien basierenden Bewertungsmodells ermittelten Qualitätswerte der erstellten Anforderungen sowie ergänzend die statistischen Hintergrunddaten der Studienteilnehmer:innen.

Die Darstellung erfolgt in mehreren Schritten: Zunächst werden die Ergebnisse der Anforderungsbewertung systematisch beschrieben und vergleichend analysiert (Kapitel 4.1 und 4.2). Anschließend werden die Ergebnisse der begleitenden statistischen Erhebung herangezogen, um mögliche Einflussfaktoren besser einordnen zu können (Kapitel 4.3). Danach erfolgen Diskussion der Stichprobe (Kapitel 4.4) und der Tool-Vorerfahrung (Kapitel 4.5).

Die Diskussion der Ergebnisse erfolgt jeweils integriert in die Ergebnisdarstellung, um quantitative Befunde unmittelbar interpretativ einzuordnen.

4.1 Ergebnisse der Anforderungsbewertung

Im Folgenden sind die Ergebnisse der Bewertung der in der Studie erfassten Anforderungen sowohl als Gesamtbewertung wie auch als Bewertung des jeweiligen Evaluationskriteriums graphisch dargestellt und in Zusammenhang gesetzt.

4.1.1 Gesamtbewertung des Detaillierungsgrades

Abbildung 6 zeigt die aggregierte Gesamtbewertung der in der Studie erhobenen Anforderungen, getrennt nach den beiden Versuchsgruppen (Gruppe A, Nutzung von Storywise) und Gruppe B, Nutzung von Microsoft Word). Die Gesamtbewertung ergibt sich aus dem arithmetischen Mittel der fünf Einzelkriterien Klarheit, Vollständigkeit, Prüfbarkeit, Kontextbezug und Granularität und bildet damit den in der vorliegenden Arbeit definierten Detaillierungsgrad ab.

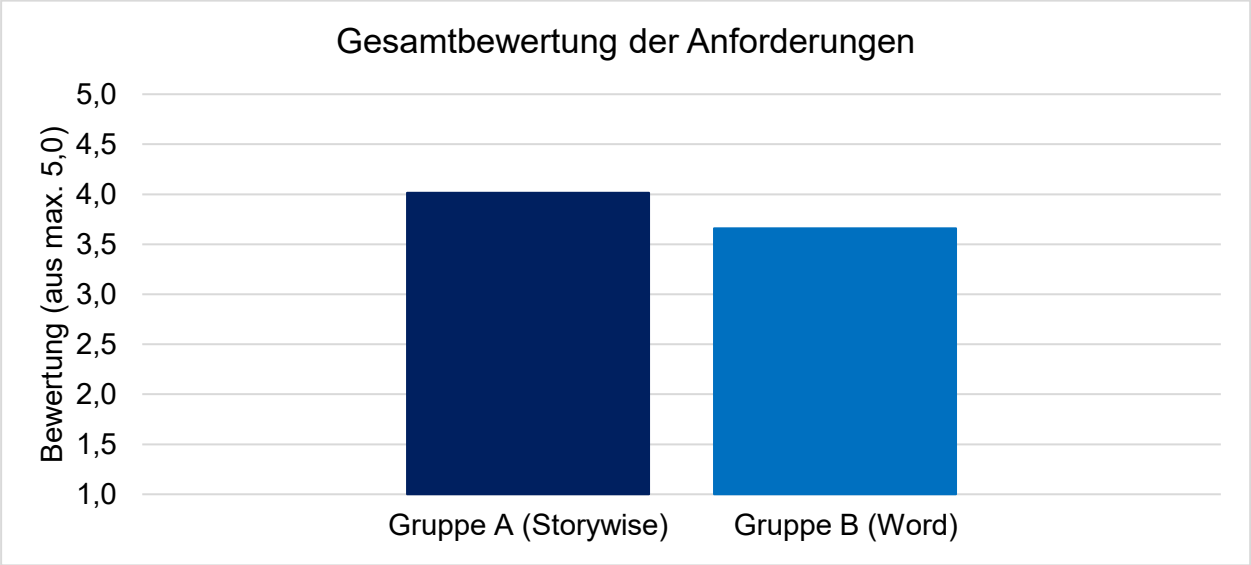

Abbildung 6: Gesamtbewertung der in der Studie erhobenen Anforderungen.

Die Ergebnisse zeigen, dass die mit der KI-gestützten Applikation Storywise erstellten Anforderungen (Gruppe A) im Durchschnitt einen höheren Gesamtwert erreichen als jene der Vergleichsgruppe, die mit Microsoft Word arbeitete (Gruppe B). Der numerische Unterschied von 0,35 Punkten (vgl. Kapitel 4.1.3) auf der fünfstufigen Skala ist zwar moderat, deutet aber einen positiven Einfluss der KI-Unterstützung auf den Detaillierungsgrad der Anforderungen an.

Im Kontext der Forschungsfrage ist dieses Ergebnis insofern bedeutsam, als dass nicht nur einzelne Qualitätsaspekte, sondern der aggregierte Detaillierungsgrad insgesamt höher ausfällt. Dies spricht dafür, dass KI-Unterstützung nicht nur isoliert einzelne Kriterien verbessert, sondern zu einer insgesamt ausgewogeneren Ausprägung mehrerer Qualitätsdimensionen beitragen kann.

Gleichzeitig ist zu berücksichtigen, dass die absolute Höhe der Gesamtwerte in beiden Gruppen im oberen Bereich der Skala liegt. Dies deutet darauf hin, dass auch die konventionelle Erstellung mit Microsoft Word grundsätzlich in der Lage ist, qualitativ hochwertige Anforderungen hervorzubringen, insbesondere wenn die Ersteller:innen über einschlägige Vorkenntnisse verfügen (vgl. Kapitel 4.3).

Abbildung 7 zeigt die individuellen Gesamtbewertungen des Detaillierungsgrades auf Ebene der einzelnen Studienteilnehmer:innen. Im Gegensatz zur aggregierten Darstellung in Abbildung 6 erlaubt diese Grafik eine differenzierte Betrachtung der Streuung innerhalb der beiden Versuchsgruppen sowie möglicher individueller Effekte.

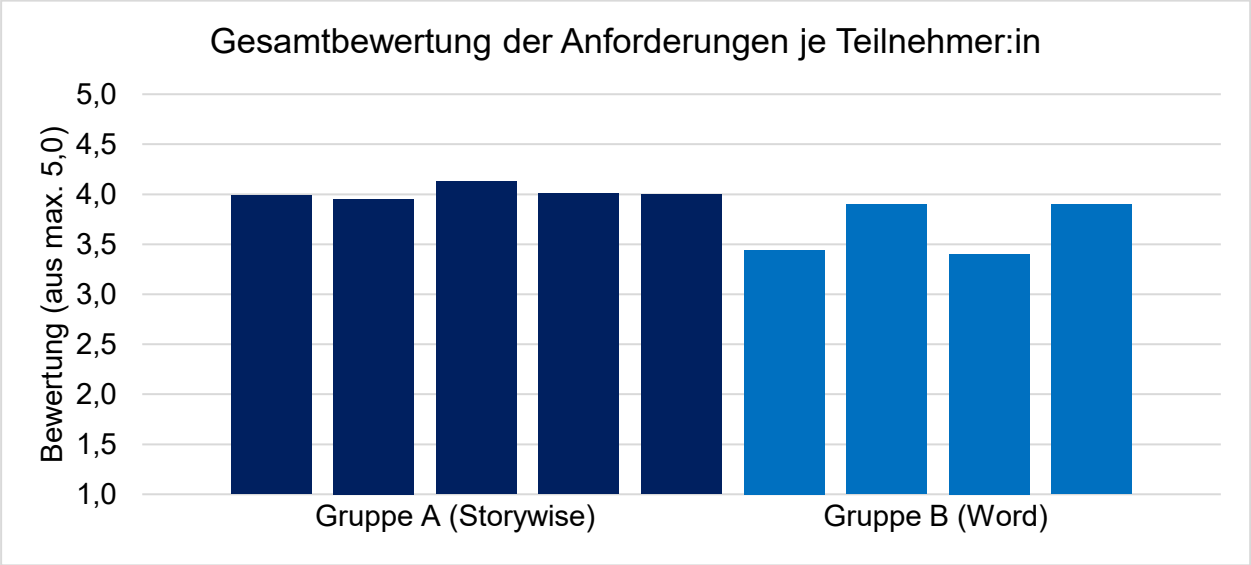

Abbildung 7: Gesamtbewertung der in der Studie erhobenen Anforderungen je Teilnehmer:in.

In Gruppe A (KI-gestützte Erstellung mit Storywise) liegen die individuellen Gesamtwerte überwiegend in einem engen Bereich zwischen etwa 3,95 und 4,13 Punkten. Diese geringe Streuung deutet auf ein vergleichsweise homogenes Qualitätsniveau der erstellten Anforderungen hin. Unabhängig von der individuellen Person erreichen die Teilnehmenden in dieser Gruppe konsistent hohe Werte im aggregierten Detaillierungsgrad.

Dieses Ergebnis kann dahingehend interpretiert werden, dass die KI-Unterstützung eine standardisierende Wirkung entfaltet. Storywise scheint den Erstellungsvorgang so zu strukturieren, dass Unterschiede in individueller Erfahrung, Schreibstil oder Vorgehensweise zumindest teilweise ausgeglichen werden. Dadurch nähern sich die Ergebnisse der einzelnen Personen qualitativ an.

Im Gegensatz dazu zeigt Gruppe B (konventionelle Erstellung mit Microsoft Word) eine deutlich größere Streuung der Gesamtbewertungen. Die individuellen Werte reichen hier von etwa 3,4 bis knapp 3,9 Punkten. Dies weist darauf hin, dass die Qualität der Anforderungen in höherem Maße von den individuellen Fähigkeiten, Vorerfahrungen und Arbeitsweisen der jeweiligen Teilnehmer:innen abhängt.

Auffällig ist zudem, dass in Gruppe B zwar einzelne Personen Werte erreichen, die nahe an jene der KI-gestützten Gruppe heranreichen, gleichzeitig jedoch auch mehrere deutlich niedrigere Bewertungen auftreten. Dieses Muster spricht dafür, dass ohne unterstützende Strukturierung durch ein spezialisiertes Tool ein höheres Risiko für weniger detaillierte Anforderungen besteht, insbesondere bei weniger erfahrenen oder weniger methodisch vorgehenden Personen.

4.1.2 Analyse der Einzelkriterien

Um die Gesamtbewertung differenzierter interpretieren zu können, werden im Folgenden die Ergebnisse der einzelnen Bewertungsdimensionen betrachtet. Ziel ist es, zu identifizieren, in welchen Aspekten sich KI-gestützte und konventionelle Erstellung besonders deutlich unterscheiden.

4.1.2.1 Klarheit

Die höchsten Mittelwerte über alle Kriterien hinweg wurden in beiden Gruppen für das Kriterium Klarheit erzielt (siehe Abbildung 8).

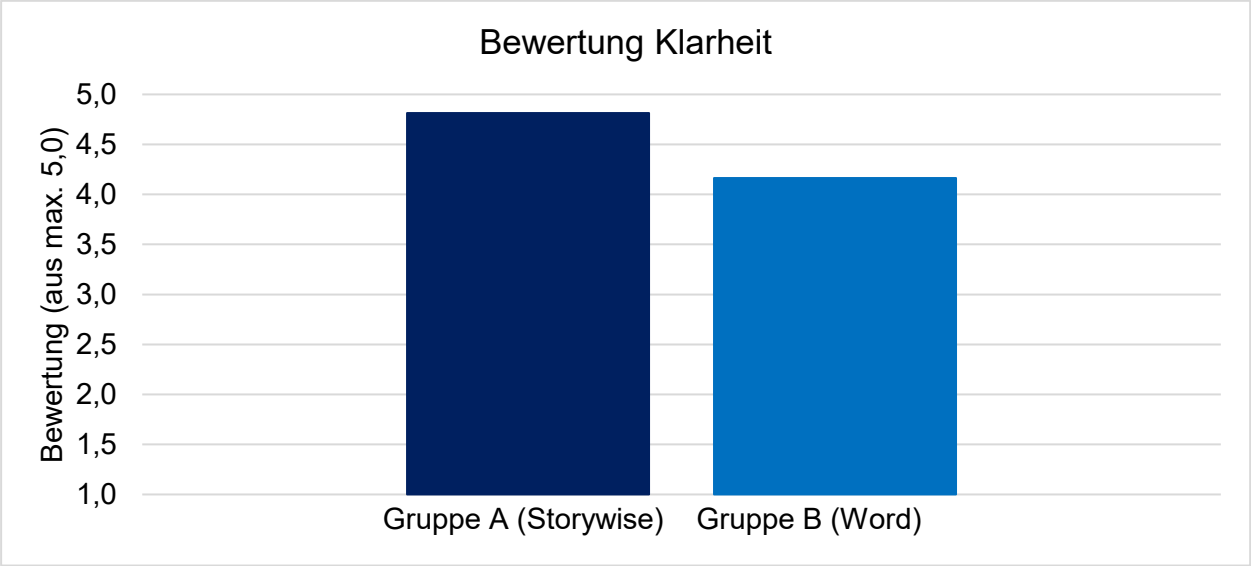

Abbildung 8: Bewertung der in der Studie erhobenen Anforderungen hinsichtlich des Kriteriums Klarheit.

Gruppe A erreichte hierbei einen deutlich höheren Wert als Gruppe B, mit einer Differenz von 0,65 Punkten (vgl. Kapitel 4.1.3). Dieser Unterschied stellt zugleich die größte Abweichung zwischen den beiden Gruppen dar.

Dieses Ergebnis legt nahe, dass KI-Unterstützung insbesondere bei der eindeutigen und präzisen Formulierung von Anforderungen wirksam ist. LLM sind darauf ausgelegt, sprachlich kohärente und strukturierte Texte zu produzieren. In der Anwendung auf Anforderungen scheint sich dieser Vorteil in klareren Satzstrukturen, einer konsistenteren Verwendung von Begriffen sowie einer geringeren Mehrdeutigkeit auszudrücken.

Im Vergleich dazu ist bei manuell erstellten Anforderungen stärker davon auszugehen, dass Formulierungsunsicherheiten, implizite Annahmen oder uneinheitliche Terminologie auftreten. Der deutliche Unterschied im Kriterium Klarheit kann daher als ein zentraler Beitrag der KI-Unterstützung zum erhöhten Detaillierungsgrad interpretiert werden.

4.1.2.2 Vollständigkeit

Auch beim Kriterium Vollständigkeit schneidet die KI-gestützte Gruppe leicht besser ab (siehe Abbildung 9).

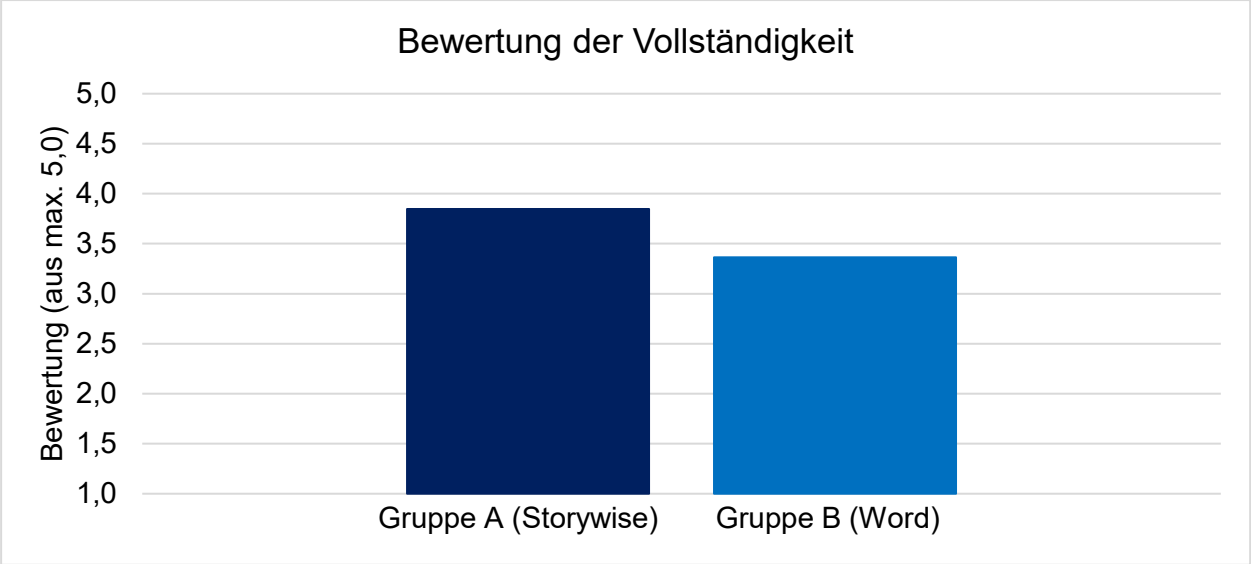

Abbildung 9: Bewertung der in der Studie erhobenen Anforderungen hinsichtlich des Kriteriums Vollständigkeit.

Die Differenz von 0,48 Punkten (vgl. Kapitel 4.1.3) weist darauf hin, dass Anforderungen aus Gruppe A häufiger Zweck, gewünschtes Verhalten sowie relevante Bedingungen explizit benennen.

Dieser Befund ist insbesondere im Lichte der Funktionsweise spezialisierter KI-Werkzeuge wie Storywise plausibel. Durch strukturierende Vorgaben, implizite Templates oder automatische Vervollständigungen werden Nutzer:innen tendenziell dazu angeleitet, mehr Aspekte einer Anforderung zu berücksichtigen. Die KI fungiert damit nicht nur als Textgenerator, sondern auch als kognitives Gerüst, das zur systematischeren Erfassung von Inhalten beiträgt.

Gleichzeitig bleibt festzuhalten, dass auch hier kein maximaler Skalenwert erreicht wird. Dies unterstreicht, dass KI-Unterstützung Vollständigkeit begünstigt, aber nicht garantiert. Nach wie vor sind Domänenwissen, Erfahrung und reflektierte Entscheidungen der Ersteller:innen notwendig, um Anforderungen inhaltlich vollständig auszuarbeiten.

4.1.2.3 Prüfbarkeit

Beim Kriterium Prüfbarkeit zeigen beide Gruppen vergleichsweise hohe Werte, wobei auch hier die KI-gestützte Gruppe leicht besser bewertet wurde (siehe Abbildung 10).

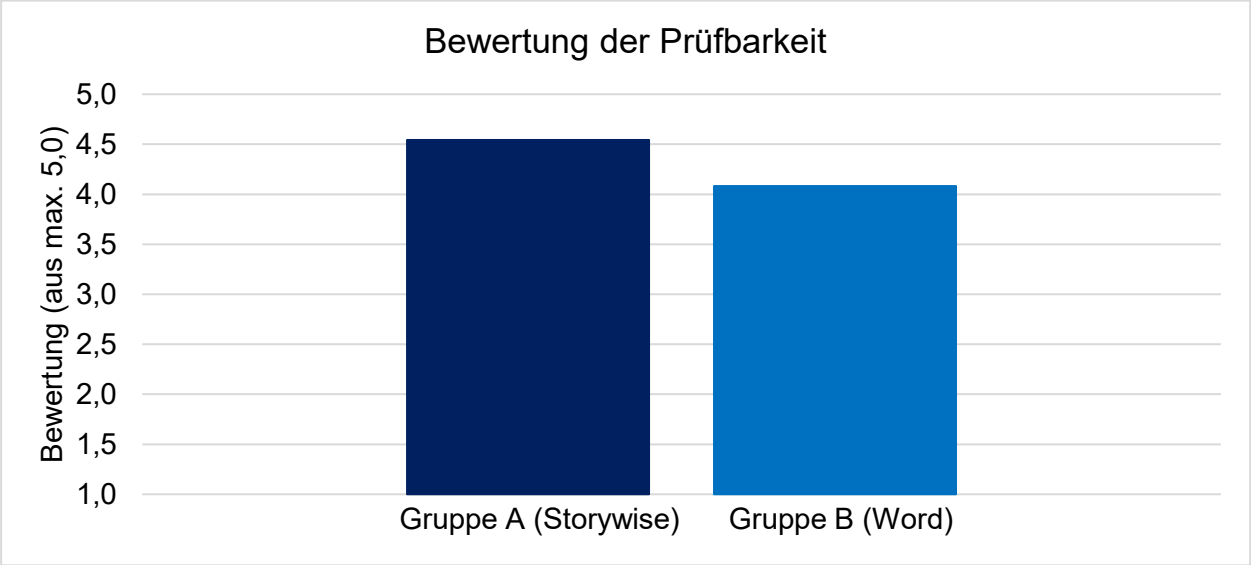

Abbildung 10: Bewertung der in der Studie erhobenen Anforderungen hinsichtlich des Kriteriums Prüfbarkeit.

Die Differenz von 0,46 Punkten (vgl. Kapitel 4.1.3) weist darauf hin, dass Anforderungen aus Gruppe A häufiger klare Kriterien enthalten, anhand derer eine Erfüllung überprüft werden kann.

Die vergleichsweise hohen Werte in beiden Gruppen könnten darauf zurückzuführen sein, dass Prüfbarkeit ein in der Praxis gut bekanntes Qualitätsmerkmal ist, insbesondere bei Teilnehmer:innen mit Erfahrung im Requirements Engineering. Dennoch deutet der Vorsprung von Gruppe A darauf hin, dass KI-Unterstützung dabei helfen kann, implizite Testkriterien expliziter zu formulieren oder stärker ergebnisorientierte Aussagen zu treffen.

4.1.2.4 Kontextbezug

Deutlich geringer fallen die Bewertungen im Kriterium Kontextbezug aus (siehe Abbildung 11).

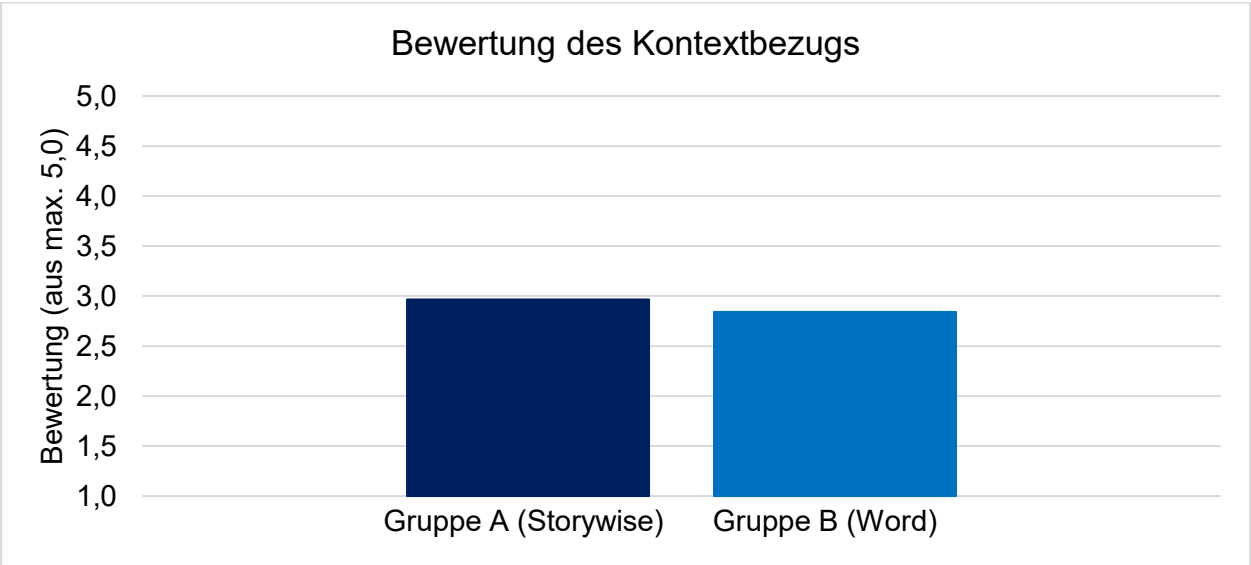

Abbildung 11: Bewertung der in der Studie erhobenen Anforderungen hinsichtlich des Kriteriums Kontextbezug.

Beide Gruppen erreichen hier die niedrigsten Mittelwerte aller Kriterien. Der Unterschied zwischen Gruppe A und Gruppe B ist mit 0,13 Punkten (vgl. Kapitel 4.1.3) vergleichsweise gering.

Dieses Ergebnis ist besonders aufschlussreich, da es auf eine generelle Schwäche in der Anforderungserstellung hinweist, unabhängig vom eingesetzten Werkzeug. Kontextinformationen wie Annahmen, Abhängigkeiten oder Umgebungsbedingungen werden offenbar häufig vernachlässigt oder nicht explizit formuliert. Eine Interpretationsmöglichkeit ist, dass implizites Kontextwissen häufig als selbstverständlich vorausgesetzt wird.

Der geringe Unterschied zwischen den Gruppen könnte darauf hindeuten, dass der Einbezug von Kontext weniger durch sprachliche Unterstützung als vielmehr durch tiefes Domänenverständnis und Erfahrungswissen beeinflusst wird. KI-gestützte Werkzeuge können hier zwar strukturierend wirken, stoßen jedoch an Grenzen, wenn kontextspezifisches Wissen nicht explizit bereitgestellt wird.

4.1.2.5 Granularität

Das Kriterium Granularität weist insgesamt sehr ähnliche Werte für beide Gruppen auf (siehe Abbildung 12).

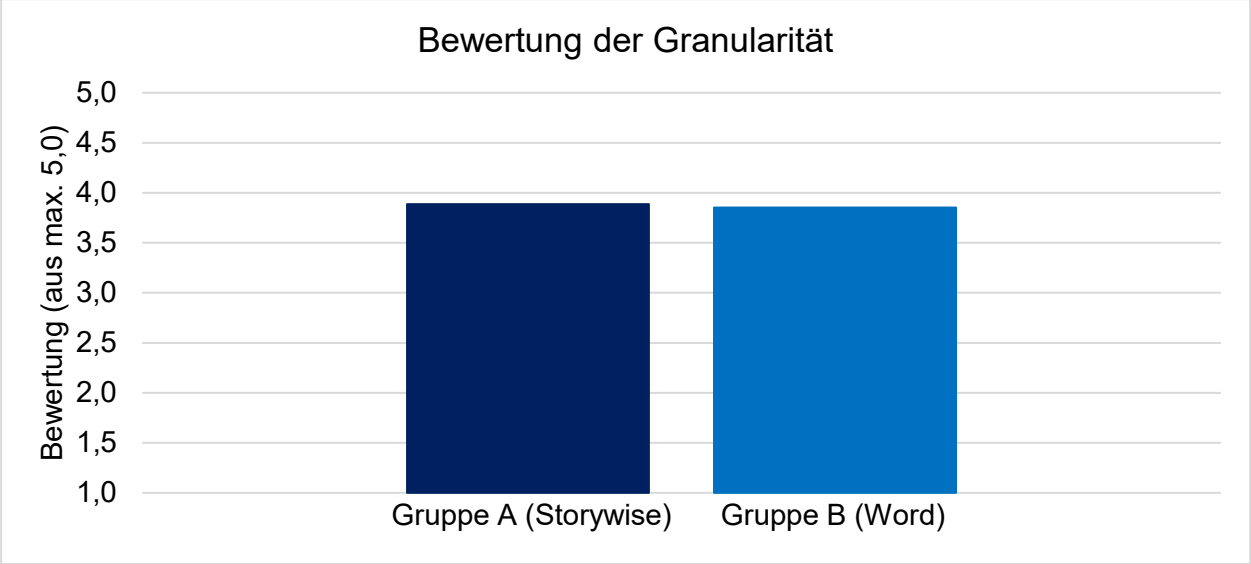

Abbildung 12: Bewertung der in der Studie erhobenen Anforderungen hinsichtlich des Kriteriums Granularität.

Die minimale Differenz von etwa 0,03 Punkten (vgl. Kapitel 4.1.3) ist praktisch vernachlässigbar und deutet darauf hin, dass die Anforderungen beider Gruppen hinsichtlich ihres Abstraktionsniveaus ähnlich angemessen formuliert wurden.

Dieses Ergebnis ist insofern relevant, als es nahelegt, dass KI-Unterstützung nicht zu einer Über- oder Unterdetaillierung im Sinne der Granularität führt. Etwaige Befürchtungen, KI-generierte Anforderungen könnten zu technisch oder zu allgemein ausfallen, bestätigen sich in dieser Studie

nicht. Vielmehr scheint die Granularität primär durch die Aufgabenstellung und das gemeinsame Verständnis der Teilnehmer:innen geprägt zu sein.

4.1.3 Interpretation der Bewertungsdifferenzen

Abbildung 13 fasst die Differenzen zwischen den beiden Gruppen über alle Kriterien hinweg zusammen.

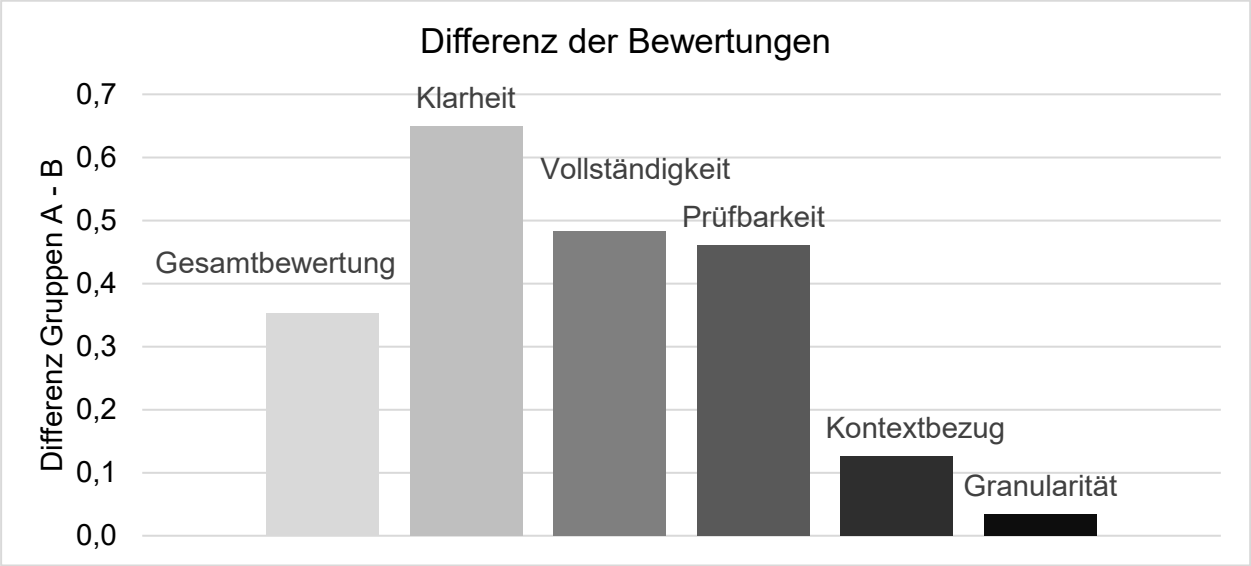

Abbildung 13: Differenzen in den Bewertungen der einzelnen Kriterien (Gruppe A - Gruppe B).

Auffällig ist, dass die größten Unterschiede in jenen Kriterien auftreten, die stark mit sprachlicher Strukturierung und Explizitheit zusammenhängen (Klarheit, Vollständigkeit, Prüfbarkeit). Demgegenüber fallen Unterschiede in kontextuellen oder strukturellen Aspekten (Kontextbezug, Granularität) deutlich geringer aus.

Dies stützt die Annahme, dass der primäre Mehrwert KI-basierter Werkzeuge in der Unterstützung der sprachlichen und formalen Qualität von Anforderungen liegt. Der Detaillierungsgrad wird somit insbesondere dort erhöht, wo explizitere Formulierungen eine Verbesserung erreichen können.

4.1.4 Gesamtinterpretation der Ergebnisse

Kapitel 4.1 hat gezeigt, dass sich zwischen der KI-gestützten Erstellung von Anforderungsspezifikationen mit Storywise und der konventionellen Erstellung mit Microsoft Word messbare Unterschiede im Detaillierungsgrad der erzeugten Anforderungen feststellen lassen. Über alle betrachteten Anforderungen hinweg weist die KI-gestützte Versuchsgruppe einen höheren aggregierten Gesamtwert auf, was auf einen insgesamt positiven Einfluss der

KI-Unterstützung schließen lässt. Der beobachtete Unterschied ist zwar moderat, tritt jedoch konsistent über mehrere Evaluationskriterien hinweg auf und gewinnt dadurch an Aussagekraft.

Eine differenzierte Betrachtung der einzelnen Kriterien verdeutlicht, dass sich die Vorteile der KI-Unterstützung nicht gleichmäßig auf alle Qualitätsaspekte verteilen. Besonders ausgeprägt sind die Unterschiede in den Kriterien Klarheit, Vollständigkeit und Prüfbarkeit. Diese Dimensionen stehen in engem Zusammenhang mit sprachlicher Präzision, expliziter Strukturierung und formaler Eindeutigkeit – Eigenschaften, bei denen LLM-gestützte Werkzeuge ihre Stärken entfalten können. Die Ergebnisse legen nahe, dass Storywise den Spezifikationsprozess dahingehend unterstützt, implizite Informationen expliziter zu machen und Anforderungen in konsistenter, prüfbarer Form zu formulieren.

Demgegenüber zeigen die Kriterien Kontextbezug und Granularität nur geringe bzw. keine relevanten Unterschiede zwischen den beiden Gruppen. Dies deutet darauf hin, dass Aspekte, die stark von Domänenwissen, situativem Verständnis und inhaltlicher Einbettung abhängen, weniger durch KI-Unterstützung beeinflusst werden. Der angemessene Abstraktionsgrad einer Anforderung sowie die Berücksichtigung kontextueller Rahmenbedingungen scheinen weiterhin primär durch menschliche Expertise geprägt zu sein.

Zusammenfassend lässt sich festhalten, dass KI-gestützte Werkzeuge wie Storywise den Detaillierungsgrad von Anforderungsspezifikationen insbesondere dort erhöhen, wo Detailtiefe durch sprachliche Explizitheit und formale Struktur erzielt wird. Kapitel 4.1 beantwortet die Forschungsfrage somit dahingehend, dass KI-Unterstützung einen positiven, jedoch selektiven Einfluss auf den Detaillierungsgrad hat. Sie wirkt weniger als Ersatz für fachliche Expertise, sondern vielmehr als qualitätsstabilisierendes und strukturierendes Hilfsmittel im Anforderungserstellungsprozess.

4.2 Beispiele qualitativer Bewertungen ausgewählter Anforderungen

Zur weiteren Vertiefung der quantitativen Ergebnisse werden im Folgenden exemplarisch zwei Sets von Anforderungen aus beiden Versuchsgruppen qualitativ analysiert. Diese Sets beziehen sich jeweils auf einen Satz des natürlichsprachigen Originaltextes (siehe Anhang A). Ziel dieser Analyse ist es, die numerischen Bewertungen aus Kapitel 4.1 anhand konkreter Textbeispiele nachvollziehbar und interpretierbar zu machen.

4.2.1 Anforderungen mit Bezug zu „Generierung von automatisierten Berichten“

In Tabelle 1 sind Anforderungen beider Gruppen aufgelistet, die sich auf den folgenden Ausschnitt aus dem natürlichsprachigen Originaltext (siehe Anhang A) beziehen: „5. Automatisierte Reports: Generierung von automatisierten Berichten über Arbeitszeiten, Überstunden, Urlaub etc. für Vorgesetzte, beispielsweise zu jedem Monatsende.“ Dieses Beispiel wurde gewählt, um zu zeigen, wie deutlich die beiden Vergleichsgruppen in manchen Bewertungsergebnissen voneinander abweichen.

Tabelle 1: Anforderungen passend zu "Generierung von automatisierten Berichten über Arbeitszeiten, Überstunden, Urlaube etc. für Vorgesetzte, beispielsweise zu jedem Monatsende." (vgl. Anhang A).

Gruppe A	**K1**	**V**	**P**	**K2**	**G**	**M**	**Qualitative Bewertung**
US17: Als Teamleiter will ich automatisierte Monatsberichte erhalten	5	4	5	3	4	4,2	Sehr gut formuliert, klar und prüfbar.
US12: Als Vorgesetzter will ich automatisierte Monatsberichte erhalten	5	4	5	3	4	4,2	Klar formuliert und prüfbar. Zweck gut erkennbar.
US13: Als Teamleiter will ich automatisierte Monatsberichte erhalten	5	4	5	3	4	4,2	Eindeutig und gut prüfbar.
US14: Als Personalmanager will ich automatisierte Monatsberichte erhalten	5	4	5	3	4	4,2	Klarer Zweck, guter Kontext.
US10: As Vorgesetzter I want receive automatically generated monthly reports	5	4	5	3	4	4,2	Eindeutig formuliert, gut prüfbar, vollständiger Zweck.
US52: Als Teamleiter will ich automatisierte Team-Reports erstellen und versenden lassen	5	4	4	3	4	4,0	Klar und testbar, aber ohne Formatvorgaben.
Gruppe B	**K1**	**V**	**P**	**K2**	**G**	**M**	**Qualitative Bewertung**
Automatische Berichtserstellung	4	4	4	3	4	3,8	Zweck und Verhalten klar erkennbar. Kontext mäßig vorhanden.
System erstellt automatisierte Reports über Arbeitszeiten	4	3	4	3	4	3,6	Klar formuliert und grundsätzlich prüfbar. Kontext jedoch begrenzt.
System erstellt automatisierte Reports über Überstunden	4	3	4	3	4	3,6	Ähnlich wie vorher: klar, aber wenig detailliert. Prüfbarkeit gut.
System erstellt automatisierte Reports über Urlaube	4	3	4	3	4	3,6	Deutlich, vollständig genug und gut prüfbar.

K1: Klarheit, V: Vollständigkeit, P: Prüfbarkeit, K2: Kontextbezug, G: Granularität, M: arithmetischer Mittelwert aller fünf Evaluationskriterien.

4.2.2 Anforderungen mit Bezug zu „Gruppierung von Mitarbeitern in Teams“

In Tabelle 2 sind Anforderungen beider Gruppen aufgelistet, die sich auf den Ausschnitt aus dem Originaltext (siehe Anhang A) „Mitarbeiter können in Teams gruppiert werden“ beziehen. Dieses Beispiel wurde ausgewählt, um zu zeigen, dass es auch Sets an Anforderungen gibt, die keine eindeutige Bewertungsdifferenz zwischen Gruppe A und Gruppe B aufweisen.

Tabelle 2: Anforderungen passend zu "Mitarbeiter können in Teams gruppiert werden" (vgl. Anhang A).

Gruppe A	**K1**	**V**	**P**	**K2**	**G**	**M**	**Qualitative Bewertung**
US2: Als Mitarbeiter will ich in Teams gruppiert werden können	5	3	4	3	4	3,8	Zweck klar erkennbar, aber ohne nähere Bedingungen. Prüfbarkeit gut.
US1: As Vorgesetzter I want manage a team of employees	5	4	5	3	4	4,2	Klar formulierte Story mit erkennbarer Intention. Testbar, ausreichend granular.
US1: Als Mitarbeiter will ich in Teams gruppiert werden um einer passenden Teamstruktur zugeordnet zu sein	5	3	4	3	4	3,8	Die Story ist klar, aber der Zweck bleibt allgemein. Prüfbarkeit gegeben, Kontext mäßig beschrieben.
US1: Als Mitarbeiter will ich in Teams gruppiert werden	5	3	4	3	4	3,8	Klarer Zweck, einfache Struktur, prüfbar.
Gruppe B	**K1**	**V**	**P**	**K2**	**G**	**M**	**Qualitative Bewertung**
Teamleiter kann Mitarbeiter einem Team hinzufügen oder entfernen	4	3	4	3	4	3,6	Die Anforderung ist klar formuliert und prüfbar. Vollständigkeit ist mittel, da Einschränkungen oder Bedingungen fehlen. Kontext ist erkennbar, aber nicht umfassend.
Zuweisung von Mitarbeitern zu Teams	4	4	4	3	4	3,8	Solide formuliert und prüfbar. Kontext vorhanden, aber nicht vollständig.
Als Teamleiter möchte ich meine Mitarbeiter gruppieren, in eigene Teams hinzufügen und diese Teams verwalten.	4	3	4	3	4	3,6	Die Story ist verständlich und beschreibt klar das gewünschte Verhalten. Die Prüfbarkeit ist möglich, da konkrete Aktionen ableitbar sind. Kontext ist nur teilweise ausgeführt, Granularität angemessen.
Als Administrator möchte ich Teams erstellen können, denen ich Mitarbeiter und exakt einen Teamleiter zuweisen kann.	5	4	5	4	4	4,4	Die Story ist präzise formuliert und enthält klare fachliche Regeln. Die Prüfbarkeit ist vollständig gegeben, da alle Kernpunkte objektiv überprüfbar sind. Der Kontext ist eindeutig und unterstützend.

K1: Klarheit, V: Vollständigkeit, P: Prüfbarkeit, K2: Kontextbezug, G: Granularität, M: arithmetischer Mittelwert aller fünf Evaluationskriterien.

4.2.3 Kritische Diskussion der ausgewählten Anforderungen und ihrer Bewertungen

Die beiden untersuchten Beispiele – „Generierung von automatisierten Berichten“ (4.2.1) und „Gruppierung von Mitarbeitern in Teams“ (4.2.2) – zeigen, dass die Wirkung von Storywise gegenüber klassischen Werkzeugen wie Microsoft Word nicht einheitlich positiv ausfällt, sondern je nach Anforderungsart und Komplexität variiert. Es gibt sowohl Bereiche, in denen die KI-gestützte Gruppe A ein klar höheres Qualitätsniveau erreicht, als auch solche, in denen beide Gruppen ähnliche Ergebnisse erzielen.

Im ersten Beispiel („Generierung von automatisierten Berichten“) sind die Unterschiede besonders deutlich. Die Anforderungen von Gruppe A sind sprachlich präziser, konsistenter strukturiert und inhaltlich besser nachvollziehbar. Ihre Formulierungen folgen einem einheitlichen Aufbau und benennen Zweck, Handlung und Ergebnis klar. Das spiegelt sich auch in den Einzelbewertungen wider: In den Kriterien Klarheit, Vollständigkeit und Prüfbarkeit schneidet die KI-gestützte Gruppe A deutlich besser ab als die manuelle Vergleichsgruppe B. Hier zeigt sich, dass Storywise – vermutlich durch vordefinierte Strukturen und sprachliche Hilfestellungen – eine Standardisierung begünstigt, welche die Formulierungsqualität und Verständlichkeit der Anforderungen merklich verbessert. Auch das Ergebnis der Gesamtbewertung (Gruppe A: 4,17, Gruppe B: 3,65) spiegelt diese Interpretation wider.

Das zweite Beispiel („Gruppierung von Mitarbeitern in Teams“) verdeutlicht dagegen, dass die Vorteile der KI-Unterstützung nicht in allen Fällen zum Tragen kommen. Beide Gruppen erreichten hier insgesamt ähnliche Mittelwerte, teils sogar leicht höhere Bewertungen in der Word-Gruppe. Während die Vollständigkeit der Anforderungen von Gruppe A höher ist als die der Anforderungen von Gruppe B, liegen die beiden Gruppen bei der Bewertung der Prüfbarkeit und der Granularität gleichauf. Die Vollständigkeit sowie der Kontextbezug wurden in diesem Beispiel bei Gruppe B (Word) höher bewertet als bei Gruppe A (Storywise). Die Gesamtbewertung der Anforderungen der beiden Gruppen lässt keinen eindeutigen Schluss bzgl. des höheren Detaillierungsgrades der Anforderungen einer Gruppe zu (Gruppe A: 3,90, Gruppe B: 3,85).

Im Gesamtbild wird somit deutlich, dass die KI-Unterstützung vor allem dann wirksam ist, wenn der Qualitätsaspekt mit sprachlich-struktureller Präzision zusammenhängt. Besonders Klarheit und teilweise auch Vollständigkeit sowie Prüfbarkeit profitieren vom Einsatz von Storywise, weil das Tool offenbar die Satzlogik und Textkohärenz stärkt. In diesen Dimensionen agiert die KI als eine Art formale Qualitätskontrolle: Sie reduziert Mehrdeutigkeiten, fördert verständliche Satzstrukturen und lenkt den Fokus der Nutzenden auf konkrete, testbare Aussagen. Weniger Effekte zeigen sich dagegen bei Merkmalen, die auf inhaltliches Verständnis und kontextspezifisches Wissen angewiesen sind – hier stößt KI an ihre Grenzen. Die gleichbleibenden Werte bei Kontextbezug und Granularität deuten darauf hin, dass die inhaltliche

Tiefe und fachliche Angemessenheit einer Anforderung weiterhin maßgeblich von menschlicher Expertise geprägt werden.

Ein weiterer wichtiger Gesichtspunkt betrifft die Zuverlässigkeit der Bewertung selbst, die auf dem LLM-as-a-Judge-Ansatz basiert. Der Einsatz eines LLM (in der vorliegenden Arbeit GPT-5) als Bewertungsinstanz gewährleistet eine konsistente und nachvollziehbare Analyse, insbesondere durch die einheitliche Anwendung der Bewertungsrubriken. Diese Methode reduziert subjektive Verzerrungen und schafft Transparenz, da alle Anforderungen nach denselben Kriterien beurteilt werden. Allerdings ist auch diese Form der Bewertung auf textanalytische Aspekte beschränkt. Das Modell erkennt formale Qualität – also sprachliche Klarheit, logischen Aufbau oder explizite Prüfbedingungen – doch es vermag keine impliziten Zusammenhänge, logischen Lücken oder fachlichen Fehlinterpretationen zu erfassen. Eine Anforderung kann folglich aus Sicht des Modells qualitativ hochwertig erscheinen, obwohl sie inhaltlich – etwa im Hinblick auf organisatorische Realisierbarkeit oder technische Abhängigkeiten – mangelhaft ist.

Daraus ergibt sich, dass die qualitative KI-Beurteilung allein kein vollständiges Bild der Anforderungsqualität liefern kann. Sie eignet sich hervorragend als objektive Grundlage zur Bewertung formaler Güte, sollte aber stets durch menschliche Interpretation und domänenspezifisches Wissen ergänzt werden. Dies ist insbesondere wichtig, da nicht alle der LLM-generierten qualitativen Beschreibungen vollständig und voll aussagekräftig sind. Als Beispiel sei die Anforderung „System erstellt automatisierte Reports über Urlaube“ aus Tabelle 2 genannt. Warum die qualitative Bewertung durch LLM-as-a-Judge mit „deutlich, vollständig genug und gut prüfbar.“ erfolgt ist, lässt sich – vor allem in Bezug auf die qualitative Bewertung der analogen Anforderungen „System erstellt automatisierte Reports über Arbeitszeiten“ und „System erstellt automatisierte Reports über Überstunden.“ nicht intuitiv erschließen. Die numerische Bewertung (die bei allen drei genannten Anforderungen denselben Wert ergibt), bildet die Qualität der Anforderungen konsistenter ab.

Ein Ansatz, um eine Verbesserung hinsichtlich der Bewertung durch LLM-as-a-Judge zu erreichen, wäre es, ein vortrainiertes LLM anstelle eines allgemeinen Modells einzusetzen. Zusätzlich sollte das im Prompt definierte Bewertungsmodell verfeinert werden, um eine tiefere Interpretation der in der Studie erhaltenen Anforderungen zu ermöglichen.

Zusammenfassend zeigt Kapitel 4.2, dass KI-gestützte Werkzeuge wie Storywise das Niveau sprachlicher und formaler Qualität in der Anforderungserstellung messbar erhöhen können. Sie fördern Einheitlichkeit und Verständlichkeit, standardisieren den Schreibprozess und wirken sich positiv auf die Klarheit sowie teilweise die Vollständigkeit und Prüfbarkeit von Anforderungen aus. Gleichzeitig machen die Ergebnisse deutlich, dass diese Verbesserungen nicht automatisch zu einer höheren inhaltlichen Detailtiefe führen. Für kontextbezogene und fachlich komplexe

Anforderungen sowie hinsichtlich der optimalen Granularität bleibt die Expertise menschlicher Requirements Engineer:innen weiterhin unverzichtbar.

Die Ergebnisse aus Kapitel 4.2 ergänzen somit die Erkenntnisse aus Kapitel 4.1. und runden sie auf der Ebene des Evaluationsgegenstandes (in der Studie erhaltene Anforderungen) ab.

4.3 Ergebnisse der Erhebung statistischer Daten

Zur besseren Einordnung der Bewertungsergebnisse wurden ergänzend statistische Daten zu den Teilnehmer:innen erhoben (vgl. Anhang B). Diese liefern wichtige Kontextinformationen für die Interpretation der Ergebnisse.

4.3.1 Berufliche Erfahrung mit Anforderungsspezifikationen

Die Auswertung der Antworten zur bisherigen Dauer der Arbeit mit Anforderungsspezifikationen (Abbildung 14) zeigt, dass beide Gruppen aus Personen mit mehrjähriger Erfahrung bestehen.

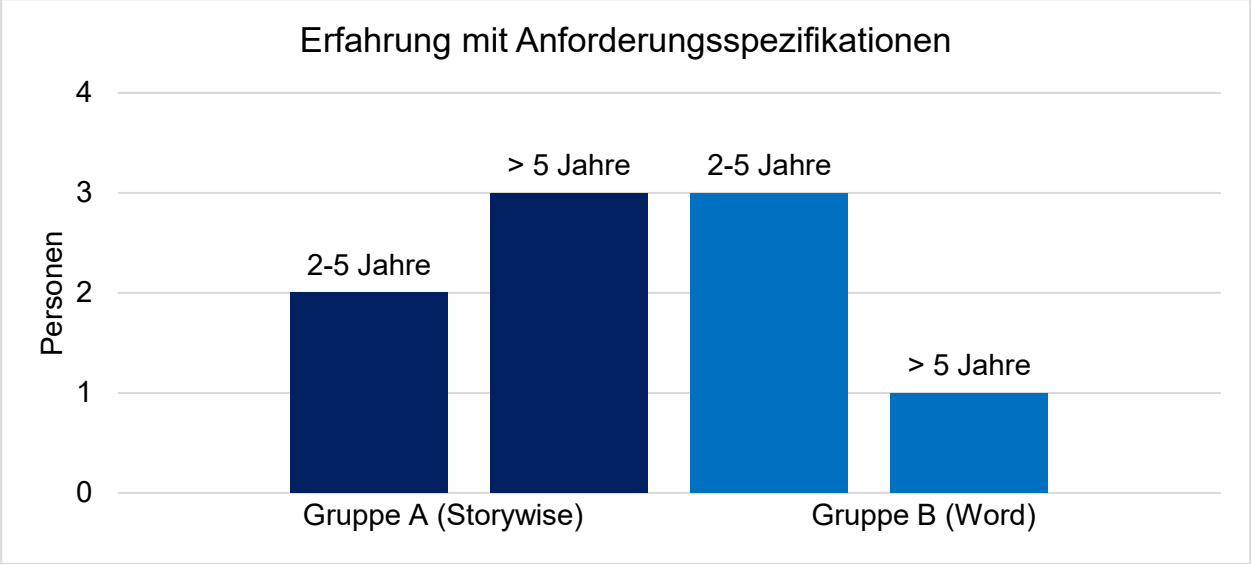

Abbildung 14: Antworten auf die Frage „Wie lange arbeitest du schon mit Anforderungsspezifikationen?“

In Gruppe A verfügen drei von fünf Teilnehmenden über mehr als fünf Jahre Erfahrung, während in Gruppe B die Teilnehmer:innen mit einer Erfahrung mit Anforderungsspezifikationen von zwei bis fünf Jahren überwiegen.

Um zu erfahren, ob ein Effekt der längeren Berufserfahrung von Gruppe A mit Anforderungsspezifikationen in Bezug auf die höheren Bewertungen der Evaluationskriterien vorliegt, ist dieser Zusammenhang in Abbildung 15 graphisch aufbereitet.

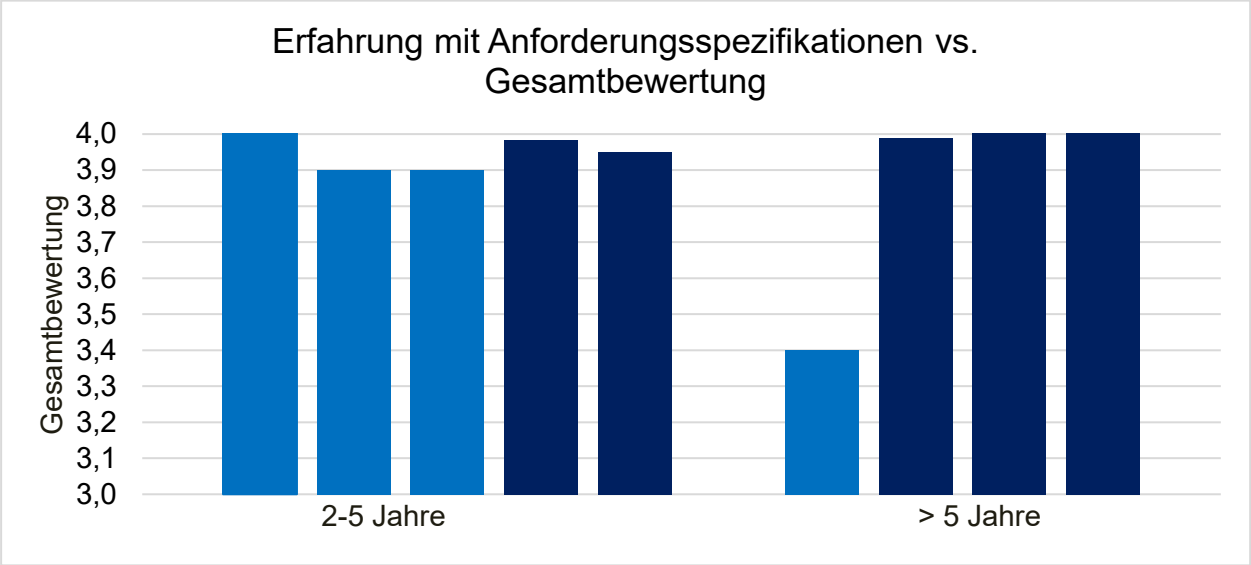

Abbildung 15: Zusammenhang zwischen beruflicher Erfahrung mit Anforderungsspezifikationen und der Gesamtbewertung der Anforderungen je Person.

In Gruppe A liegen alle erreichten Gesamtwerte auf einem sehr hohen Niveau. Ein eindeutiger Einfluss der Berufserfahrung mit Anforderungsspezifikationen auf die Gesamtbewertung der Anforderungen der jeweiligen Person konnte nicht erkannt werden.

Für Gruppe B kann ein Zusammenhang zwischen beruflicher Erfahrung in der Arbeit mit Anforderungsspezifikationen ausgeschlossen werden. Die einzige Person, die in dieser Gruppe > 5 Jahre Berufserfahrung angegeben hat, erreichte im Gesamtvergleich die niedrigste Bewertung aller Anforderungen (Bewertung mit 3,40).

4.3.2 Berufliche Erfahrung mit User Stories

Die statistische Auswertung der Dauer der beruflichen Beschäftigung mit User Stories (Abbildung 16) zeigt keine systematischen Unterschiede, die eine Verzerrung nahelegen würden. Beide Gruppen umfassen Teilnehmer:innen mit unterschiedlichem Erfahrungsniveau, einschließlich Personen ohne vorherige Erfahrung.

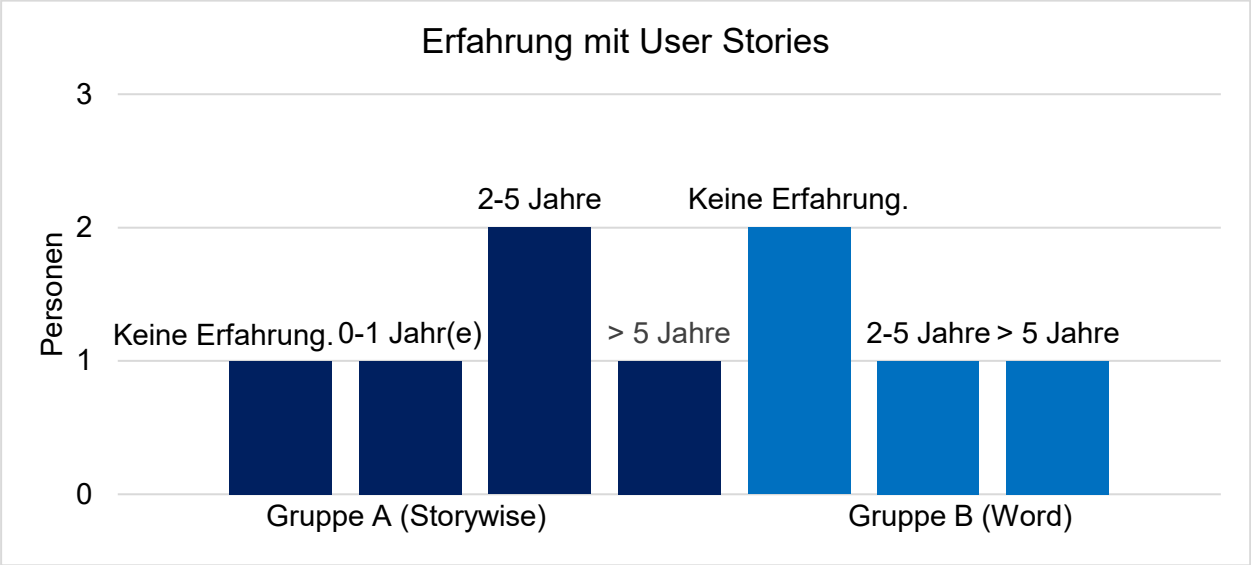

Abbildung 16: Antworten auf die Frage „Wie lange arbeitest du schon mit User Stories?"

Zu betonen ist hierbei, dass selbst Teilnehmende mit geringer Erfahrung in der KI-gestützten Gruppe vergleichsweise hohe Bewertungswerte erreichen. Dies deutet darauf hin, dass KI-Unterstützung insbesondere für weniger erfahrene Verfasser:innen von Anforderungen eine kompensierende Wirkung haben kann.

4.3.3 Beruflicher Nutzungskontext von Anforderungsspezifikationen

Abbildung 17 zeigt die Verteilung der Antworten auf die Frage, ob die Studienteilnehmer:innen Anforderungsspezifikationen primär selbst verfassen oder diese überwiegend als Input für ihre eigene Arbeit nutzen.

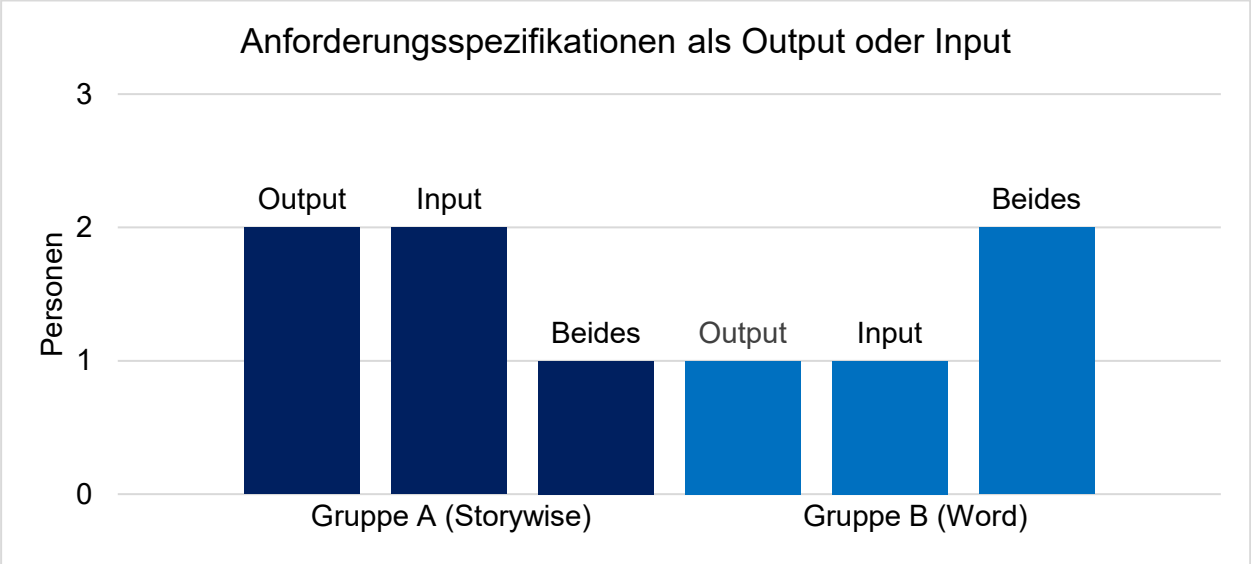

Abbildung 17: Antworten auf die Frage "Schreibst du Anforderungsspezifikationen selbst oder nutzt du sie als Input für deine Arbeit?"

Die Ergebnisse verdeutlichen, dass ein Großteil der Befragten im beruflichen Kontext aktiv in die Erstellung von Anforderungsspezifikationen eingebunden ist, während ein kleinerer Anteil Anforderungen hauptsächlich als Arbeitsgrundlage nutzt. Diese Rollenverteilung spricht für eine heterogene Stichprobe, die unterschiedliche Perspektiven innerhalb des Requirements Engineering-Prozesses abbildet.

Für die Interpretation der Bewertungsergebnisse ist dieses Ergebnis insofern relevant, als dass Personen, die Anforderungen selbst verfassen, in der Regel ein stärker ausgeprägtes Bewusstsein für Qualitätskriterien wie Klarheit, Vollständigkeit oder Prüfbarkeit besitzen. Gleichzeitig zeigt die Präsenz von Teilnehmenden, die Anforderungen primär konsumieren, dass auch die Sicht von nachgelagerten Rollen – etwa Entwicklung oder Test – in die Studie einfließt.

Im Kontext der vorliegenden Bachelorarbeit deutet die Verteilung darauf hin, dass die beobachteten Unterschiede im Detaillierungsgrad nicht ausschließlich auf eine bestimmte Nutzerrolle zurückzuführen sind.

Darüber hinaus legt die Abbildung nahe, dass KI-gestützte Werkzeuge potenziell eine Brückenfunktion zwischen schreibenden und nutzenden Rollen einnehmen können. Durch eine klarere und strukturiertere Formulierung von Anforderungen profitieren nicht nur die Ersteller:innen selbst, sondern auch jene, die Anforderungen als Input für weitere Arbeitsschritte verwenden.

Abbildung 18 stellt die jeweilige Nutzungsfunktion von Anforderungen im beruflichen Kontext in Zusammenhang mit der Gesamtbewertung je Person.

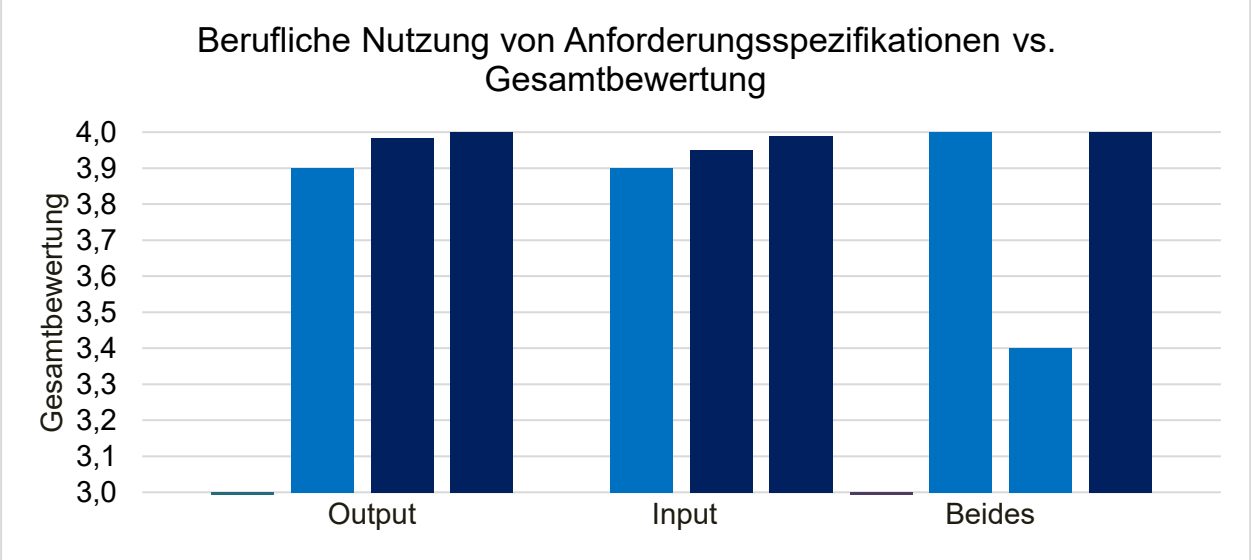

Abbildung 18: Nutzungsverhalten von Anforderungsspezifikationen (Nutzung als Output, Input oder beides) vs. Gesamtbewertung.

Abbildung 18 lässt schließen, dass die Art der beruflichen Nutzung von Anforderungsspezifikationen (als Input, Output oder in beiden Formen) keine Auswirkung auf den Detaillierungsgrad der in der Studie erstellen Anforderungen hat.

4.3.4 Weitere Erkenntnisse der statistischen Datenerhebung

Die Teilnehmer:innen aus Gruppe A wurden im Zuge der Online-Befragung auch gebeten, zu bewerten, ob die Einschulungsphase in Storywise (Dauer: 10 Minuten) ausreichend gewesen sei. Vier von fünf Teilnehmer:innen beantworteten diese Frage positiv.

Ebenfalls die Teilnehmer:innen aus Gruppe A wurden auch bezüglich ihrer Zufriedenheit mit der Unterstützung durch Storywise befragt, wobei die Bewertung mit 1 „nicht zufrieden“ und die Bewertung mit 5 „sehr zufrieden“ bedeutete. Die Antworten auf diese Frage waren sehr heterogen: Jeweils eine Person beantwortete diese Frage mit 2, 3 oder 4. Zwei Teilnehmer:innen gaben eine Bewertung von 5.

Die natürlichsprachigen Antworten auf die Fragen „Inwieweit kannst du dir vorstellen, Storywise in deinen Arbeitsalltag zu integrieren?“ und „Hast du weiteres Feedback für uns?“ sind in Anhang D abgelegt. Da sie keinen für die vorliegende Bachelorarbeit relevanten Input geben, wurden sie aus Gründen der Vollständigkeit dokumentiert, fließen jedoch nicht in die Interpretation der Ergebnisse ein.

4.4 Kritische Diskussion der Stichprobe

Ein wesentlicher limitierender Faktor der vorliegenden Evaluationsstudie ist die geringe Stichprobengröße von neun Teilnehmer:innen. Diese wurde bewusst in Kauf genommen, da es sich um eine explorative Vergleichsstudie im Rahmen einer Bachelorarbeit handelt und die praktische Durchführung mit einem spezialisierten Werkzeug unter kontrollierten Bedingungen erfolgte. Dennoch hat die geringe Anzahl an Teilnehmer:innen Auswirkungen auf die Aussagekraft und Generalisierbarkeit der Ergebnisse.

Zunächst ist festzuhalten, dass mit der vorliegenden Stichprobengröße keine statistisch abgesicherten Schlussfolgerungen getroffen werden können. Die Ergebnisse erlauben daher keine Verallgemeinerung auf eine größere Grundgesamtheit von Anwender:innen. Vielmehr liefern sie indikative, richtungsweisende Erkenntnisse über mögliche Effekte von KI-Unterstützung auf den Detaillierungsgrad von Anforderungsspezifikationen.

Gleichzeitig zeigt sich jedoch, dass die beobachteten Unterschiede zwischen den beiden Versuchsgruppen über mehrere Bewertungsdimensionen hinweg konsistent auftreten. Insbesondere die wiederkehrend höheren Werte der KI-gestützten Gruppe in den Kriterien

Klarheit, Vollständigkeit und Prüfbarkeit sprechen dafür, dass es sich nicht ausschließlich um zufällige Effekte einzelner Personen handelt.

Die kleine Stichprobe erlaubt darüber hinaus eine engmaschige qualitative Analyse der erzeugten Anforderungen. Dadurch konnten einzelne Ausprägungen detailliert diskutiert und beispielhaft gegenübergestellt werden. In diesem Sinne unterstützt die geringe Stichprobengröße eine tiefere inhaltliche Durchdringung der Ergebnisse, auch wenn sie gleichzeitig deren statistische Tragweite einschränkt.

Nichtsdestotrotz bleibt festzuhalten, dass individuelle Faktoren – wie persönliche Erfahrung im Requirements Engineering oder Schreibstil – bei einer so kleinen Teilnehmerzahl einen vergleichsweise hohen Einfluss auf die Ergebnisse haben können. Auch wenn ergänzende statistische Auswertungen (Kapitel 4.3) keine offensichtlichen Beeinflussungen aufzeigten, können derartige Effekte nicht ausgeschlossen werden.

Für zukünftige Forschung wäre daher eine Ausweitung der Studie mit einer größeren und heterogeneren Stichprobe wünschenswert. Eine höhere Teilnehmerzahl würde es ermöglichen, statistische Signifikanztests durchzuführen, Subgruppen (z. B. nach Erfahrung oder Rolle) gezielt zu vergleichen und die Stabilität der hier beobachteten Effekte zu überprüfen.

Zusammenfassend ist die geringe Stichprobengröße als strukturelle Einschränkung der vorliegenden Arbeit zu bewerten. Sie begrenzt die Generalisierbarkeit der Ergebnisse, schmälert jedoch nicht deren Relevanz als explorativer Beitrag. Die Studie liefert belastbare Hinweise auf potenzielle Effekte der KI-Unterstützung im Anforderungserstellungsprozess und bildet damit eine valide Grundlage für weiterführende, umfangreichere Untersuchungen.

4.5 Einfluss der Tool-Vorerfahrung

Ein potenzieller Einflussfaktor auf die Ergebnisse der vorliegenden Evaluationsstudie ist die Vorerfahrung der Studienteilnehmer:innen mit den eingesetzten Applikationen Storywise und Microsoft Word.

Hinsichtlich der Nutzung von Microsoft Word ist davon auszugehen, dass alle Teilnehmer:innen von Gruppe B über eine grundlegende Vertrautheit mit dem Textverarbeitungsprogramm verfügten, da es sich um ein weit verbreitetes Standardwerkzeug handelt.

Für die Teilnehmer:innen der KI-gestützten Gruppe A stellte Storywise hingegen ein neues Werkzeug dar. Um diesen Umstand zu berücksichtigen, wurde zu Beginn der Studie eine Einschulungsphase von zehn Minuten durchgeführt. Die Rückmeldungen der Teilnehmer:innen zeigen, dass diese Einschulungszeit von der Mehrheit als ausreichend eingeschätzt wurde (vgl. Kapitel 4.3.4). Dennoch weisen einzelne qualitative Rückmeldungen darauf hin, dass bestimmte

Funktionen des Tools als nicht vollständig intuitiv wahrgenommen wurden und teilweise Unsicherheiten im Umgang mit der Applikation bestanden.

Vor diesem Hintergrund ist kritisch zu reflektieren, dass eine längere Einarbeitungsphase in Storywise potenziell zu noch höheren Qualitätsgewinnen hätte führen können. Gleichzeitig zeigt gerade der Umstand, dass trotz begrenzter Einschulung ein höherer durchschnittlicher Detaillierungsgrad erreicht wurde, die Wirksamkeit der KI-Unterstützung.

Für zukünftige Studien wäre es dennoch sinnvoll, längere Trainingsphasen oder wiederholte Nutzungsszenarien zu berücksichtigen, um Lerneffekte systematisch zu untersuchen.

5. CONCLUSIO UND AUSBLICK

Die vorliegende Bachelorarbeit hatte das Ziel, den Einfluss von KI-Unterstützung auf den Detaillierungsgrad von Softwarespezifikationsanforderungen im Vergleich zu traditionellen Ansätzen empirisch zu untersuchen. Ausgangspunkt war die Beobachtung, dass zwar zunehmend KI-basierte Werkzeuge im Requirements Engineering eingesetzt werden, jedoch bislang nur begrenzte quantitative Erkenntnisse zu ihrem tatsächlichen Nutzen vorliegen. Die zentrale Forschungsfrage lautete daher, welchen Einfluss KI-gestützte Werkzeuge auf die Qualität – insbesondere auf den Detaillierungsgrad – von Anforderungen ausüben.

Auf Basis einer kontrollierten Evaluationsstudie mit zwei Versuchsgruppen konnte gezeigt werden, dass die Nutzung des KI-gestützten Tools Storywise im Vergleich zur konventionellen Erstellung von Anforderungen mittels Microsoft Word zu einem insgesamt höheren Detaillierungsgrad der Anforderungen führt.

Eine differenzierte Betrachtung der Ergebnisse zeigt, dass sich die Stärken der KI-Unterstützung insbesondere in den Kriterien Klarheit, Vollständigkeit und Prüfbarkeit erkennen lassen. In diesen Dimensionen trägt die KI maßgeblich dazu bei, Anforderungen präziser zu formulieren, implizite Inhalte expliziter auszudrücken und den Erstellungsprozess zu standardisieren.

Demgegenüber zeigen die Ergebnisse auch klare Grenzen der KI-Unterstützung: Die Kriterien Kontextbezug und Granularität (Aspekte, die stark von Domänenwissen, Erfahrung und situativem Verständnis abhängig sind) weisen kaum Unterschiede zwischen den beiden Gruppen auf, was auf die auch zukünftig hohe Relevanz von menschlichen Requirements Engineers hindeutet.

Ein weiterer relevanter Befund betrifft die Streuung der Ergebnisse: Während die KI-gestützte Gruppe ein homogeneres Qualitätsniveau aufweist, zeigen sich in der konventionellen Gruppe deutlich größere Unterschiede zwischen den einzelnen Teilnehmer:innen. Insbesondere weniger erfahrene Anwender:innen profitieren von der strukturierten KI-Unterstützung und erreichen ein Qualitätsniveau, das näher an jenes einer erfahreneren Person heranreicht.

Es sind jedoch auch relevante Einschränkungen der Arbeit zu berücksichtigen: Die geringe Stichprobengröße von neun Teilnehmer:innen stellt eine zentrale Limitation dar und erlaubt keine statistisch abgesicherten Verallgemeinerungen. Zudem war die Durchführungszeit der Studie mit 25 Minuten bewusst kurz gehalten, um eine vergleichbare Ausgangssituation der beiden Vergleichsgruppen zu gewährleisten. Diese zeitliche Einschränkung könnte jedoch dazu geführt haben, dass komplexere Anforderungen nicht vollständig ausgearbeitet wurden.

Ein weiterer Einflussfaktor ist die Vertrautheit der Teilnehmer:innen mit dem jeweils verwendeten Tool: Während Microsoft Word als Standard gesehen werden kann, der allen Teilnehmer:innen aus der Praxis vertraut ist, war die Einschulungszeit von Gruppe A in Storywise auf zehn Minuten beschränkt. Die vorliegende Arbeit kann nicht abschließend beurteilen, ob eine längere Einschulungs- oder Trainingsphase die Ergebnisse von Gruppe A weiter verbessert hätte.

Darüber hinaus ist auch der Einsatz von LLM-as-a-Judge kritisch zu reflektieren. Obwohl dieser Ansatz eine konsistente und nachvollziehbare Bewertung ermöglicht, liegt sein Fokus primär auf der Analyse sprachlicher und formaler Aspekte. Tiefere fachliche Bewertungen oder das Erkennen semantischer Inkonsistenzen können dadurch nur eingeschränkt erfolgen. Eine Kombination aus KI-gestützter und menschlicher Bewertung erscheint daher auch für zukünftige Anwendungen sinnvoll.

Aus den genannten Einschränkungen ergeben sich mehrere Ansatzpunkte für weiterführende Forschung: Zukünftige Studien sollten insbesondere mit größeren und heterogeneren Stichproben durchgeführt werden, um statistische Signifikanz zu erreichen und differenziertere Analysen zu ermöglichen. Ebenso wäre ein Vergleich unterschiedlicher KI-Werkzeuge sinnvoll, um die Generalisierbarkeit der Ergebnisse zu überprüfen und potenzielle Unterschiede zwischen verschiedenen Systemen zu identifizieren. Darüber hinaus könnte die Untersuchung auf andere Arten von Dokumenten, wie beispielsweise nicht-funktionale Anforderungen oder technische Spezifikationen, ausgeweitet werden. Durch eine Verlängerung der Versuchsdauer könnte erforscht werden, ob ein insgesamt höherer Komplexitätsgrad der Anforderungspezifikation und. ein deutlicherer Beitrag der KI-Unterstützung zustande kommt. Ein weiterer Ansatz wäre die wirtschaftliche Betrachtung des Einsatzes KI-gestützter Applikationen für die Erfassung von Softwarespezifikationen. Durch den höheren Detaillierungsgrad könnten Nachbearbeitungskosten bzw. Rückfragen bei der Umsetzung reduziert werden, was in weiterer Folge zu einem finanziellen Benefit – selbst in Anbetracht der Lizenzkosten – führen könnte.

Zusammenfassend zeigt die vorliegende Arbeit, dass KI-gestützte Werkzeuge ein erhebliches Potenzial zur Verbesserung der formalen Qualität von Anforderungsspezifikationen besitzen. Sie tragen insbesondere zur Erhöhung von Klarheit, Vollständigkeit und Prüfbarkeit bei und wirken qualitätsstabilisierend auf den Erstellungsprozess. Gleichzeitig ersetzen sie jedoch keine fachliche Expertise, sondern ergänzen diese als unterstützendes Werkzeug. Der größte Mehrwert entsteht somit in der Kombination aus menschlichem Domänenwissen und KI-basierter Strukturierungs- und Formulierungshilfe.

Die Ergebnisse liefern damit einen wichtigen empirischen Beitrag zur Diskussion über den Einsatz von KI im Requirements Engineering und bilden eine fundierte Grundlage für zukünftige Forschungs- und Praxisentwicklungen in diesem Bereich.

ANHANG A - 1. Anhang

Projektbeschreibung Timewi.se: Entwicklung einer Personal-Zeiterfassungssoftware für österreichische Betriebe

Projektziel: Entwicklung einer benutzerfreundlichen und rechtssicheren Personal-Zeiterfassungssoftware, die österreichischen Betrieben dabei hilft, die Arbeitszeiten ihrer Mitarbeiter effizient zu verwalten, einschließlich Remote-Arbeitszeiten. Die Software soll sämtliche Anforderungen des österreichischen Arbeitsrechts erfüllen und Web- sowie Cloud-Technologien nutzen, um Remote-Arbeitsmöglichkeiten zu unterstützen.

Funktionalitäten:

1. Benutzer- und Teamverwaltung:

Mitarbeiter können in Teams gruppiert werden und die Teams sollen von einem Teamleiter verwaltet werden können. Administratoren können neue Mitarbeiter hinzufügen, löschen und bearbeiten.

2. Arbeitszeitverwaltung:

Alle Mitarbeiter sollen ihre Anwesenheitszeiten und Remote-Arbeitszeiten selbst in Echtzeit erfassen können. Die Software soll Unterstützung für Gleitzeit und Teilzeitarbeit, inklusive der Möglichkeit, geplante Arbeitszeiten festzulegen, bieten. Außerdem ist eine automatische Erkennung von Überstunden und Minusstunden im Rahmen einer Ad-Hoc-Auswertung gewünscht.

3. Urlaubsmanagement:

Mitarbeiter können Urlaubsanträge stellen und den Status ihrer Anträge überprüfen. Teamleiter bzw. Vorgesetzte können diese Urlaubsanträge genehmigen oder ablehnen.

4. Rechtssicherheit:

Es wird Wert auf die Einhaltung des österreichischen Arbeitsrechts, einschließlich Ruhezeiten, Feiertagen und Vermeidung von Sonntagsarbeit gelegt. Dies muss bei der Entwicklung berücksichtigt werden.

5. Automatisierte Reports:

Generierung von automatisierten Berichten über Arbeitszeiten, Überstunden, Urlaube etc. für Vorgesetzte, beispielsweise zu jedem Monatsende.

6. Sicherheit und Datenschutz:

Unterstützung durch Multi-Faktor-Authentifizierung für erhöhte Sicherheit. Außerdem müssen datenschutzkonforme Datenverarbeitung und -speicherung gewährleistet werden.

ANHANG B - 2. Anhang

Der Fragebogen zu statistischen Fragestellungen enthielt die folgenden Fragen:

1) Frage: Wie lange arbeitest du schon mit Anforderungsspezifikationen?

Antwortmöglichkeiten: 0 – 1 Jahr / 2 – 5 Jahre / > 5 Jahre / Ich arbeite nicht mit Anforderungsspezifikationen.

2) Frage: Schreibst du Anforderungsspezifikationen selbst oder nutzt du sie als Input für deine Arbeit?

Antwortmöglichkeiten: Ich schreibe selbst Anforderungen./Ich nutze sie als Input für meine Arbeit./Sonstiges: _____

3) Frage: Waren die 10 Minuten Einschulungsphase für Storywise ausreichend?

Antwortmöglichkeiten: Ja/Nein/Ich habe in Word gearbeitet.

4) Frage: Wie zufrieden bist du mit der Unterstützung durch Storywise?

Antwortmöglichkeiten: Skala 0 – 10. / Ich habe in Word gearbeitet.

5) Inwieweit kannst du dir vorstellen, Storywise in deinen Arbeitsalltag zu integrieren?

Antwortmöglichkeit: Freitext.

6) Hast du weiteres Feedback für uns?

Antwortmöglichkeit: Freitext

ANHANG C - 3. Anhang

Der nachfolgende Prompt wurde für die Bewertung durch LLM-as-a-Judge verwendet:

Du agierst als neutraler Evaluator im Bereich Requirements Engineering. Deine Aufgabe ist die strukturierte, konsistente und vollständige Bewertung aller Anforderungen im hochgeladenen Dokument, mit der Einschränkung, dass Originalsätze vollständig ignoriert werden. Bewertet werden soll der Detaillierungsgrad der Anforderungen.
Bewerte ausschließlich User Stories, funktionale Anforderungen oder andere explizite Anforderungsformate, die im Dokument enthalten sind. Originalsätze dienen nicht als Bewertungsgrundlage und werden nicht berücksichtigt.
Es dürfen keine Annahmen, kein externes Wissen und keine projektspezifischen Interpretationen verwendet werden.

Nutze das folgende Bewertungsmodell:
Bewertet werden fünf Kriterien auf einer Skala von 1 (sehr gering) bis 5 (sehr hoch).

Kriterien:

- Klarheit
  Wie eindeutig ist die Anforderung formuliert?
  1 = unklar, mehrdeutig
  5 = eindeutig, ohne Interpretationsspielraum
- Vollständigkeit
  Sind Zweck, Verhalten, Einschränkungen und relevante Bedingungen enthalten?
  1 = wichtige Aspekte fehlen
  5 = alle wesentlichen Bestandteile vorhanden
- Prüfbarkeit
  Ist klar erkennbar, wann die Anforderung erfüllt ist?
  1 = nicht testbar
  5 = objektiv prüfbar
- Kontextbezug
  Sind Abhängigkeiten, Annahmen und Umgebungsbedingungen ausreichend beschrieben?
  1 = kein Kontext
  5 = klarer, nachvollziehbarer Kontext

- Granularität
  Ist die Anforderung weder zu grob noch zu technisch detailliert?
  1 = viel zu grob oder überdetailliert
  5 = passend für Planung und Umsetzung
-

Gesamtbewertung:
Mittelwert der fünf Kriterien

Vorgehen

1. Identifiziere alle Anforderungen im Dokument, jedoch ausschließlich:
   - User Stories (z. B. „Als … möchte ich … um …“)
   - explizite funktionale Anforderungen
   - andere klar formulierte Anforderungen
2. Ignoriere vollständig: alle „Originalsätze“, alle beschreibenden Textpassagen, die keine Anforderungen darstellen
3. Bewerte jede Anforderung einzeln nach Kriterium 1-5.
4. Berechne die Gesamtpunktzahl.

- Leite den Detaillierungsgrad ab:
- Gib für jede Anforderung eine kurze qualitative Begründung (2–3 Sätze).
- Ausgabeformat (für jede Anforderung):

Code
Anforderung: <Originaltext der Anforderung>

Kriterium 1: _
Kriterium 2: _
Kriterium 3: _
Kriterium 4: _
Kriterium 5: _
Detaillierungsgrad: _
Qualitative Begründung: _
Alle Anforderungen werden nacheinander in diesem Format als csv-Datei ausgegeben.

ANHANG D - 4. Anhang

Antworten aus der Online-Umfrage auf die Frage „Inwieweit kannst du dir vorstellen, Storywise in deinen Arbeitsalltag zu integrieren?“

Gute Idee - muss aber noch besser umgesetzt werden. Im moment sehe ich es nicht als vollständige lösung.
Sinnvolle Ergänzung aufgrund der Rahmenbeschreibung eines Projekts für den Start.
Ich finde es sehr gut und ich habe Interesse Storywise in meinem Arbeitsalltag zu integrieren.
Es wäre definitiv hilfreich für die schnelle Klassifizierung und Erstellung von Stories. Das Tool müsste jedoch Funktionalitäten bieten um sie in eine bestehende Toollandschaft zu integrieren.
ich finde es ein super tool, mit welchem man gut die Anforderung zusammenfassen kann.
Es klingt wie eine gute Idee, jedoch macht Jira derzeit die stories selbst wenn man es nur kurz zusammenfasst was man will. Ich werde es höchstwahrscheinlich beim derzeitigen Arbeitgeber nicht integrieren
Ich bevorzuge standardisierte Anforderungen und schätze dadurch die Umsetzung und Unterstützung durch KI.
Denke ja

Antworten aus der Online-Umfrage auf die Frage „Hast du weiteres Feedback für uns?“

Zu viel Automatismus, ohne Möglichkeit händisch einzugreifen. zB. die "Sentences" kann man nur trennen oder verbinden, aber nicht bearbeiten oder umformulieren. Keine Möglichkeit Nicht-funktionale Anforderungen (NFRs) zu erfassen. Kein "Undo" - ich habe mehrmals neustarten müssen. Bedienung ist nicht ganz intuitiv. Vorschlag: sentences zu Epics per Drag&Drop hinzufügen. Copy&Paste für Epics und Stories hat gefehlt.
Tolles Tool, an manchen Stellen nicht sehr intuitiv.
Die kurze Einschulungszeit hat ausgereicht, jedoch war das Tool selbst nicht ganz intuitiv. Ich wollte eigenständig neue Stories hinzufügen, jedoch passierte beim Drücken auf Hinzufügen nichts. Das Tool hat mir bei der ersten initialen Übersicht angezeigt, dass ich 0 Stories zu den Epics zugeordnet habe. Nachdem ich nochmals auf die Startseite ging und aktualisiert habe waren die User Stories ersichtlich
Word ist definitiv nicht mein bevorzugtes Tool für IT-Anforderungen.
Tool wirkt von der EInführung ganz brauchbar.

ABKÜRZUNGSVERZEICHNIS

AI	Artificial Intelligence
BLEU	Bilingual Evaluation Understudy
KI	Künstliche Intelligenz
LLM	Large Language Model
LM	Language Model
NLP	Natural Language Processing
NLP4RE	Natural Language Processing for Requirements Engineering
ROUGE	Recall-oriented Understudy for Gisting Evaluation
SDM	Structural depth metric
SRS	Software Requirements Specification

ABBILDUNGSVERZEICHNIS

TABELLENVERZEICHNIS

LITERATURVERZEICHNIS

Absar ul Hasan, S., & Rana, Z. A. (2022). Determining the level of detail of software requirements. *International Conference on Frontiers of Information Technology (FIT),* 13–17. https://doi.org/10.1109/FIT57066.2022.00013

Bühne, S., & Herrmann, A. (2024). *CPRE Requirements management handbook.* International Requirements Engineering Board (IREB). https://cockpit-v1.ireb.org/media/pages/downloads/cpre-requirements-management-handbook/6cdb531889-1733311674/ireb-cpre-handbook-for-requirements-management-de-v2.1.pdf

Cantini, R., Orsino, A., Ruggiero, M., & Talia, D. (2025). Benchmarking adversarial robustness to bias elicitation in large language models: scalable automated assessment with LLM-as-a-judge. *Machine Learning*. https://doi.org/10.1007/s10994-025-06862-6

Cohn, M. (2026). How detailed should a user story be? Mountain Goat Software. https://www.mountaingoatsoftware.com/blog/what-level-of-detail-should-be-captured-in-a-user-story

Döring, N., Gäde, J. C., & Schermelleh-Engel, K. (2023). *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften* (6. Aufl.). Springer. https://doi.org/10.1007/978-3-662-64762-2

Ferrari, A., & Ginde, G. (Eds.). (2025). *Handbook on natural language processing for requirements engineering.* Springer. https://doi.org/10.1007/978-3-031-73143-3

Franch, X., Palomares, C., Quer, C., Chatzipetrou, P., & Groschek, T. (2023). The state-of-practice in requirements specification: an extended interview study at 12 companies. *Requirements Engineering, 28*, 377-409. https://doi.org/10.1007/s00766-023-00399-7

Glinz, M. (2005). *Software Engineering: Eine Einführung*. Universität Zürich. https://files.ifi.uzh.ch/rerg/amadeus/teaching/courses/software_engineering_hs08/skript/Kapitel_07.pdf

Glinz, M., van Loenhoud, H., Staal, S., & Bühne, S. (2022). *CPRE – Certified professional for requirements engineering foundation Level*. International Requirements Engineering Board (IREB). https://cockpit-v1.ireb.org/media/pages/downloads/cpre-foundation-level-handbook/a46a1b6077-1776936817/cpre_foundationlevel_handbook_de_v1.3.2.pdf

Hruschka, P., Lauenroth, K., Meuten, M., Rogers, G., Gärtner, S., & Steffe, H.-J. (2025). *CPRE RE@Agile handbook* (Version 2.2). International Requirements Engineering Board (IREB). https://cpre.ireb.org/de/downloads-and-resources/downloads#cpre-re-agile-handbook

IBM. (n.d.). Guidelines for good requirements. https://www.ibm.com/docs/en/erqa?topic=assistant-guidelines-good-requirements

Knuplesch, S. (2024). Mit künstlicher Intelligenz auf dem Weg zu effektiven und effizienten Softwareangeboten (Masterarbeit). FH Joanneum. https://opus.campus02.at/frontdoor/index/index/searchtype/collection/id/16222/rows/10/start/0/yearfq/2024/facetNumber_author_facet/all/author_facetfq/Knuplesch%2C+Stefan/docId/981

Krishna, M., Gaur, B., Verma, A., & Jalote, P. (2024). Using LLMs in software requirements specifications: an empirical evaluation. arXiv. https://arxiv.org/abs/2404.17842

Laplante, P. A. (2013). *Requirements engineering for software and systems* (2nd ed.). Auerbach Publications. https://doi.org/10.1201/b15939

Li, D., Tan, Z., Zhao, C., Jiang, B., Huang, B., Ma, P., Alnaibari, A., Shu, K., & Liu, H. (2025). Who’s your judge? On the detectability of LLM-generated judgements. arXiv. https://arxiv.org/abs/2509.25154

Montgomery, L., Fucci, D., Bourafa, A., Scholz, L., & Maalej, W. (2022). Empirical research on requirements quality: A systematic mapping study. *Requirements Engineering, 27*, 183–209. https://doi.org/10.1007/s00766-021-00367-z

OpenAI. (2026). GPT-5 (Large Language Model). OpenAI.

Pohl, K. & Rupp, C. (2021). *Basiswissen Requirements Engineering. Aus- und Weiterbildung nach IREB-Standard zum Certified Professional for Requirements Engineering Foundation Level* (5. Auflage). dpunkt.verlag.

PURE Consultant. (n.d.). Anforderungsspezifikation: Definition, Techniken und Kritik. https://www.pureconsultant.de/de/requirements-engineering/anforderungsspezifikation/

Simões, G. S., & Vazquez, C. E. (2017). Functional requirements and their levels of granularity. Requirements Engineering Magazine. https://re-magazine.ireb.org/articles/functional-requirements-and-their-levels-of-granularity

storywi.se. (n.d.). https://storywi.se/

SuperAnnotate. (n.d.). LLM Evaluation: BLEU – ROUGE. https://doc.superannotate.com/docs/guide-bleu-rouge

The Requirements Engineer. (n.d.). Quality criteria. https://the-requirements-engineer.com/glossary/quality-criteria/

Wake, B. (2003). INVEST in good stories, and SMART tasks. https://xp123.com/invest-in-good-stories-and-smart-tasks/

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., & Stoica, I. (2023). Judging LLM-as-a-judge with MT-bench and chatbot Arena. arXiv. https://arxiv.org/abs/2306.05685