Duale Hochschule

Baden-Württemberg

Mosbach

Lohrtalweg 10

74821 Mosbach

KI-gestützte Bildverschlagwortung und Alt-Text-Generierung nach Unternehmenstaxonomien – Entwurf, Implementierung und Evaluation am Beispiel eines DAM-Systems

Bachelorarbeit

Studiengang/-richtung:	Medien/ Onlinemedien
Kurs:	ON23B
Matrikelnummer:	6411851
Name, Vorname:	Klemm, Paul
Geburtsdatum:	08.05.2003
Geburtsort:	Freudenstadt
Name, Vorname des wissenschaftlichen Betreuers:	Lammarsch, Tim
Name und Sitz des Ausbildungsbetriebes:	Hansgrohe SE Klaus-Grohe-Straße 5-9 77761 Schiltach
Abgabedatum:	30.09.2026

DANKSAGUNG

Mein besonderer Dank gilt den zwölf Fachexperten, die an der Evaluation teilgenommen haben, sowie den Mitarbeitenden der IT- und KI-Abteilungen der Hansgrohe SE, die durch ihre fachlichen Einblicke und Anregungen wesentlich zum Erfolg dieser Arbeit beigetragen haben.

INHALTSVERZEICHNIS

Danksagung ..... **II**

Abkürzungsverzeichnis ..... **VII**

Abbildungsverzeichnis ..... **IX**

Tabellenverzeichnis ..... **X**

Gender-Disclaimer ..... **XI**

1 Einleitung **1**

1.1 Problemstellung und Motivation 1

1.2 Zielsetzung 3

1.3 Forschungsfragen 4

1.4 Abgrenzung zur Vorarbeit 5

1.5 Aufbau der Arbeit 5

2 Theoretische Grundlagen **6**

2.1 Digital Asset Management und Metadatenstrukturen 6

2.2 Vision Language Models 7

2.2.1 Von Bildklassifikation zu multimodaler Textgenerierung 7

2.2.2 Architektur moderner Vision Language Models 7

2.2.3 Kontrastives Vortraining und Zero-Shot-Transfer 8

2.2.4 Instruction Tuning 9

2.2.5 Untersuchte Modelle 10

2.3 Effizienzstrategien für lokale VLM-Inferenz 11

2.3.1 Speicherbedarf und Bandbreite als limitierender Faktor 11

2.3.2 Quantisierungsverfahren 12

2.3.3 Parametereffizienz jenseits der Quantisierung 12

2.3.4 Konsequenzen für den Modellvergleich 13

2.4 Prompt Engineering für strukturierte Ausgaben 14

2.4.1 Begriffliche Einordnung ... 14
2.4.2 Prompt-Strategien ... 14
2.4.3 Strukturierte Ausgabegenerierung ... 15
2.4.4 Iterative Prompt-Optimierung ... 15
2.5 Barrierefreiheit und Alt-Text-Generierung ... 16
2.5.1 Normative Grundlagen ... 16
2.5.2 Qualitätskriterien für Alt-Texte ... 16
2.5.3 VLMs als Alt-Text-Generatoren ... 17
2.6 Evaluation generativer KI-Ausgaben ... 18
2.6.1 Grenzen automatischer Textmetriken ... 18
2.6.2 Präferenzbasierte Vergleiche ... 18
2.6.3 Präferenzquoten und statistische Einordnung ... 19
2.6.4 Begründung des Evaluationsansatzes ... 19
2.7 Themenbezogene Arbeiten ... 20
3 Forschungsmethodik ... 23
3.1 Design Science Research als Rahmen ... 23
3.1.1 Begründung der Methodenwahl ... 23
3.1.2 Zuordnung zum Sechs-Phasen-Modell ... 24
3.1.3 Artefakt-Typ und wissenschaftlicher Beitrag ... 26
3.2 Iteratives Vorgehen ... 26
3.2.1 Entwicklungsphasen ... 26
3.2.2 Rolle der KI-gestützten Entwicklung ... 27
3.3 Evaluationsdesign ... 27
3.3.1 Überblick ... 27
3.3.2 Bewertungsmethode ... 28
3.3.3 Statistische Auswertung ... 28
4 Anforderungen und Konzeption ... 29
4.1 Ableitung der Anforderungen ... 29
4.2 Funktionale Anforderungen ... 30

4.3 Nichtfunktionale Anforderungen 32
4.4 Architekturentscheidungen 33
4.5 Modellauswahl 34
4.6 Filterkriterien und Asset-Selektion 35
4.7 Rückverfolgung der Anforderungen 37
5 Implementierung 38
5.1 Systemarchitektur und Infrastruktur 38
5.2 Pipeline-Design 40
5.3 Prompt-Entwicklung 43
5.3.1 Version V1: Grundstruktur 43
5.3.2 Version V2: Oberflächenintegration 44
5.3.3 Version V3: Quellendeklaration und Regelwerk 44
5.3.4 Version V4: Konsolidierung nach Expertenfeedback 45
5.4 Produktivbetrieb und Automatisierung 46
5.5 Evaluations-Tooling 47
6 Evaluation 48
6.1 Evaluationsaufbau 48
6.1.1 Evaluationsdatensatz 48
6.1.2 Umfrage-Design 49
6.1.3 Teilnehmerauswahl 51
6.2 Ergebnisse der Präferenzerhebung 52
6.3 Objektive Modellmetriken 55
6.4 Qualitative Analyse der Expertenkommentare 56
6.5 Gegenüberstellung nach Produktkontext 60
6.6 Kostenprojektion 61
6.7 Zusammenfassung und Empfehlung 63
7 Diskussion 64

7.1 Ergebnisse des produktiven Erstlaufs 64
7.2 Beantwortung der Forschungsfragen 65
7.3 Limitationen 67
7.4 Praktische Implikationen 69

8 Fazit und Ausblick 71

8.1 Zusammenfassung 71
8.2 Ausblick 72

Literaturverzeichnis XII

Anhang XVI

Anhangsverzeichnis XVI
Anhang A – Prompts im Volltext XVI
A.1 Systemprompt (system.txt) XVII
A.2 Prompt V1 (tagging_v1.txt) XVII
A.3 Prompt V2 (tagging_v2.txt) XVIII
A.4 Prompt V3 (tagging_v3.txt) – Grundlage der vergleichenden Evaluation XIX
A.5 Prompt V4 (tagging_v4.txt) – im Produktivbetrieb eingesetzt XX
Anhang B – Evaluationsdatensatz XXII
Anhang C – Modell- und Laufzeitkonfiguration XXIV
Anhang D – Freitextkommentare mit Kategorisierung XXIV
Anhang E – Beispielausgaben der vier Modelle XXIX
Beispiel 1 – Asset 746815 (har13984.tif), Fokusprodukt hinterlegt XXX
Beispiel 2 – Asset 115303 (apu80363.tif), nur sichtbare Produkte hinterlegt XXXII
Beispiel 3 – Asset 756306 (har14555.tif), kein Produktkontext XXXIV
Anhang F – Rohdaten, Auswertung und Artefakt XXXVI

ABKÜRZUNGSVERZEICHNIS

API	Application Programming Interface
AWQ	Activation-aware Weight Quantization
AWS	Amazon Web Services
BLEU	Bilingual Evaluation Understudy
CAD	Computer-Aided Design
CLI	Command Line Interface
CMS	Content-Management-System
CSV	Comma-Separated Values
DAM	Digital Asset Management
DSR	Design Science Research
F1	Forschungsfrage 1
F2	Forschungsfrage 2
F3	Forschungsfrage 3
FA	Funktionale Anforderung
FP16	16-Bit-Gleitkommaformat (Floating Point 16)
GGUF	Dateiformat für quantisierte Modellgewichte
GPTQ	Quantisierungsverfahren für generative Transformer
HTML	Hypertext Markup Language
JSON	JavaScript Object Notation
KI	Künstliche Intelligenz
LLM	Large Language Model
MLX	Array-Framework für Apple Silicon
MoE	Mixture of Experts
NFA	Nichtfunktionale Anforderung

PIM	Product Information Management
PLE	Per-Layer Embeddings
PTQ	Post-Training Quantization
REST	Representational State Transfer
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
TTL	Time to Live
VLM	Vision Language Model
W3C	World Wide Web Consortium
WAL	Write-Ahead Logging
WCAG	Web Content Accessibility Guidelines
YAML	YAML Ain’t Markup Language

ABBILDUNGSVERZEICHNIS

Abbildung 1: Offizielle Oberflächenbezeichnung aus dem PIM und Ausgabe des Modells im Vergleich 17

Abbildung 2: Einbettung des Systems in die Systemlandschaft der Hansgrohe SE 38

Abbildung 3: Verarbeitungsstufen der Pipeline 40

Abbildung 4: Iterative Prompt-Entwicklung von Version V1 bis V4 43

Abbildung 5: Bewertungsoberfläche der Umfrage 50

Abbildung 6: Verteilung der Präferenzentscheidungen über alle fünf Antwortkategorien (je Dimension n = 360) 54

Abbildung 7: Themen der fachlichen Freitextkommentare 58

Abbildung 8: Anteil der fünften Antwortoption nach Verfügbarkeit von Produktkontext 61

TABELLENVERZEICHNIS

Tabelle 1: Spezifikation der untersuchten Modelle 10

Tabelle 2: Katalog der funktionalen Anforderungen 32

Tabelle 3: Katalog der nichtfunktionalen Anforderungen 33

Tabelle 4: Rückverfolgung der Anforderungen 37

Tabelle 5: Teilnehmende der Evaluation nach Funktionsbereich (n = 12) 52

Tabelle 6: Präferenzentscheidungen für Bildbeschreibungen (Deutsch) 53

Tabelle 7: Präferenzentscheidungen für Such-Tags (Deutsch) 53

Tabelle 8: Kennzahlen der Modellkonfigurationen (je 30 Bilder) 56

Tabelle 9: Themen der fachlichen Freitextkommentare (n = 69; Mehrfachzuordnung möglich) 57

Tabelle 10: Fünfte Antwortoption nach Verfügbarkeit von Produktkontext 60

Tabelle 11: Kostenprojektion Gemma4:31b via AWS Bedrock (107.000 Assets) 62

Tabelle 12: Kennzahlen des produktiven Erstlaufs 64

Tabelle 13: Evaluationsdatensatz mit Produktkontext (n = 30) XXII

Tabelle 14: Modell- und Laufzeitkonfiguration der evaluierten und eingesetzten Varianten XXIV

Tabelle 15: Freitextkommentare der Evaluation mit zugewiesener Kategorie (n = 77) XXV

Tabelle 16: Modellausgaben für Asset 746815 (har13984.tif), Prompt V3 XXX

Tabelle 17: Modellausgaben für Asset 115303 (apu80363.tif), Prompt V3 XXXII

Tabelle 18: Modellausgaben für Asset 756306 (har14555.tif), Prompt V3 XXXIV

GENDER-DISCLAIMER

Diese Arbeit verwendet geschlechtsneutrale Personenbezeichnungen, wo eine natürliche Form vorliegt, etwa Teilnehmende, Bewertende oder Mitarbeitende. Wo keine solche Form existiert, steht das generische Maskulinum, etwa bei Fachexperten. Die Bezeichnungen beziehen sich in beiden Fällen auf alle Geschlechter. Doppelnennungen und Schrägstrichformen werden nicht verwendet.

1 EINLEITUNG

1.1 Problemstellung und Motivation

Digital-Asset-Management-Systeme bilden in Unternehmen mit umfangreichen Medienbeständen die zentrale Infrastruktur für Verwaltung, Organisation und Bereitstellung digitaler Inhalte. Bei der Hansgrohe SE, einem international agierenden Hersteller von Sanitärprodukten, übernimmt das DAM-System Celum diese Funktion. Über Celum werden Produktbilder, Marketingmaterialien und technische Illustrationen gespeichert und an interne sowie externe Abnehmer verteilt, zum Beispiel an Webseiten, Kataloge, Handelspartner und Agenturen.

Zum Zeitpunkt dieser Untersuchung umfasst der Bestand in Celum mehr als 549.000 Assets. Gegenstand dieser Arbeit sind ausschließlich Bilddateien. Von diesen 549.000 Assets wurden im Rahmen einer Analyse mit den Fachbereichen 107.000 als relevant für die automatisierte Verschlagwortung identifiziert. Die übrigen Assets entfallen auf Dateiformate wie PDF, Video oder CAD, auf archivierte Bestände und auf Bilder in Collections, die nicht für Endkunden bestimmt sind.

Die Anzahl verdeutlicht das grundlegende Problem: Eine flächendeckende Verschlagwortung findet nicht statt. Tags und Beschreibungen entstehen projektbezogen oder werden von beauftragten Agenturen und Fotografen vergeben. Das führt zu uneinheitlichen Begriffen, fehlenden Einträgen und inkonsistenter Abdeckung. Bilder bleiben schwer auffindbar, weil Suchbegriffe fehlen oder nicht den internen Konventionen entsprechen.

Bei angenommenen zwei Minuten Bearbeitungszeit je Asset ergeben sich für die 107.000 relevanten Bilder rund 3.567 Arbeitsstunden. Bei beispielhaft 1.800 produktiven Stunden je Vollzeitjahr entspricht dies knapp zwei Arbeitsjahren. Diese Szenariorechnung verdeutlicht den Umfang einer manuellen Bearbeitung. Der tatsächliche Zeitbedarf wurde nicht erhoben. Hinzu kommt die Herausforderung, Begriffe über verschiedene Bearbeitende hinweg konsistent zu verwenden.

Für die spätere Verwendung von Bildern in barrierefrei zu gestaltenden digitalen Angeboten werden geeignete Textalternativen benötigt. Die Richtlinie (EU) 2019/882 betrifft bestimmte Produkte und Dienstleistungen, daraus folgt keine pauschale Pflicht zur Beschreibung jedes internen DAM-Assets. Ob und welcher Alternativtext erforderlich ist, hängt vom konkreten Verwendungszweck ab (Europäisches Parlament, 2019; W3C, 2024). Das System erzeugt deshalb Vorschläge, deren Eignung im jeweiligen Veröffentlichungskontext zu prüfen ist.

Mit der Weiterentwicklung von Vision Language Models (VLMs) haben sich neue Möglichkeiten ergeben. Diese Modelle können Bilder analysieren und natürlichsprachliche Ausgaben erzeugen und lassen sich über textuelle Instruktionen steuern. Ein einzelner Prompt kann sowohl Such-Tags als auch Alternativtexte generieren, in mehreren Sprachen, unter Berücksichtigung eines textuellen Kontexts.

Bei Hansgrohe bietet sich dafür die Kombination des DAM mit dem Product-Information-Management-System (PIM) Centric PXM (vormals Contentserv) an. Über Produktverknüpfungen lassen sich strukturierte Informationen wie Produktname, Linie, Marke und Oberfläche als Kontext bereitstellen. Dieser Kontext ergänzt Angaben, die aus dem Bild allein nicht zuverlässig hervorgehen. Dazu zählen etwa die genaue Produktlinie oder der offizielle Oberflächenname.

Neben der Modellwahl stellt die Prompt-Gestaltung eine eigenständige Herausforderung dar. Generative Modelle reagieren unterschiedlich auf Formulierung und Struktur der Anweisungen. Ein einmal geschriebener Prompt erzeugt nicht automatisch optimale Ergebnisse. Die iterative Verfeinerung unter Einbeziehung von Domänenexperten soll erkannte Schwächen gezielt adressieren.

Die wissenschaftliche Relevanz ergibt sich aus der Verbindung dieser Aspekte. Forschung zu VLMs, multimodaler Bildbeschreibung, Prompt Engineering und Evaluation generativer Modelle existiert jeweils für sich. Die Anwendung auf domänenspezifische Bildverschlagwortung in einem Unternehmenskontext mit PIM-Integration, iterativer Prompt-Entwicklung und präferenzbasierter Expertenevaluation ist in der gesichteten Literatur nicht dokumentiert. Abschnitt 2.7 ordnet die Arbeit in den Forschungsstand ein.

1.2 Zielsetzung

Ziel dieser Arbeit ist die Entwicklung und Evaluation eines einsatzfähigen Systems zur automatisierten Generierung von Bildmetadaten für Celum. Das System soll geeignete Bild-Assets automatisiert verarbeiten und dabei visuelle Informationen mit Produktkontext aus dem PIM-System verbinden. Dafür wird eine Verarbeitungspipeline entwickelt, die Assets aus Celum ermittelt, über Filterregeln selektiert, mit PIM-Daten anreichert, durch ein VLM analysiert und die Ergebnisse normalisiert zurückschreibt.

Der Fokus liegt auf einer vollständigen Integration in die bestehende Systemlandschaft. Das System soll nicht nur Metadaten generieren, sondern so aufgebaut sein, dass es im Unternehmenskontext dauerhaft und automatisiert betrieben werden kann. Neue Assets sollen ohne manuellen Eingriff verarbeitet werden.

Die erzeugte Ausgabe umfasst zwei Metadatentypen in jeweils zwei Sprachen: Such-Tags für die Auffindbarkeit im DAM und KI-Beschreibungen als Grundlage für Alternativtexte auf Webpräsenzen. Diese Beschränkung auf zwei Ausgabetypen und zwei Sprachen folgt den konkreten Anforderungen des Unternehmens. Weitere Sprachen lassen sich auf Grundlage der modularen Architektur ergänzen.

Die vier evaluierten Modelle sind Gemma4:31b und Gemma4:e4b (Google DeepMind, lokal via Ollama), Qwen3.8-27B-MLX (Alibaba, lokal auf Apple Silicon) und Claude Sonnet 5 (Anthropic, Cloud via AWS Bedrock). Die Auswahl deckt lokale und cloudbasierte Ausführung, unterschiedliche Parametergrößen sowie unterschiedliche Architekturen und Ausführungsumgebungen ab.

Ein weiterer Schwerpunkt liegt auf der Prompt-Entwicklung. Der Prompt wird nicht als einmalig festgelegte Anweisung verstanden, sondern in einem iterativen Prozess über vier dokumentierte Versionen (V1 bis V4) systematisch weiterentwickelt. Jede Version adressiert identifizierte Schwächen der vorherigen Iteration und stützt sich auf qualitative Rückmeldungen von Fachexperten.

1.3 Forschungsfragen

Die übergeordnete Forschungsfrage lautet: Wie kann ein einsatzfähiges System zur KI-gestützten Bildverschlagwortung und Alt-Text-Generierung nach Unternehmenstaxonomien in ein bestehendes Enterprise-DAM integriert werden, und welche der untersuchten VLM-Konfigurationen wird für deutsche Such-Tags und Bildbeschreibungen von Fachexperten bevorzugt?

Daraus leiten sich drei Teilfragen ab.

F1 adressiert die Systementwicklung: Wie lässt sich ein modular aufgebautes System zur automatisierten Generierung von Bildmetadaten in ein bestehendes Enterprise-DAM integrieren, sodass Produktkontext aus dem PIM einfließt und die Ausgaben durch ein unternehmensspezifisches Begriffsregelwerk gesteuert werden?

F2 zielt auf den Modellvergleich: Welche der untersuchten VLM-Konfigurationen wird bei Beschreibungen und Such-Tags präferiert, und wie sind diese Präferenzen zusammen mit Inferenzdauer, Token-Verbrauch und Kosten für die betriebliche Modellwahl einzuordnen?

F3 untersucht die Eingabestrategie: Wie werden Produktkontext und fachliche Anforderungen in der iterativen Prompt-Entwicklung berücksichtigt, und welche Verbesserungshinweise ergeben sich aus den Expertenrückmeldungen?

F1 wird durch Anforderungsanalyse, Konzeption und Implementierung beantwortet. F2 stützt sich auf den präferenzbasierten Vergleich und ergänzende Leistungsmetriken. F3 wird anhand der dokumentierten Prompt-Versionen und der qualitativen Expertenrückmeldungen als Entwicklungsfrage untersucht.

1.4 Abgrenzung zur Vorarbeit

Diese Bachelorarbeit baut auf einer vorangegangenen Projektarbeit auf (Klemm, 2025), die als explorative Vorstudie konzipiert war. Dort wurden funktionale Anforderungen erhoben, Umsetzungsoptionen verglichen und ein explorativer Test verschiedener Optionen durchgeführt. Die Projektarbeit empfiehlt kurzfristig den Test des angebotenen Celum-Adapters, langfristig und bei ausreichendem Budget eine Eigenentwicklung (Klemm, 2025, S. 32, 38). Eine zusätzliche Anreicherung des Prompts mit den in Celum bereits verknüpften Produktdaten regt sie als Optimierungsidee an (Klemm, 2025, S. 24).

Die vorliegende Arbeit greift die langfristige Empfehlung einer Eigenentwicklung auf und geht in mehreren Punkten über die Projektarbeit hinaus. Erstens wird ein vollständiges, produktiv eingesetztes System implementiert. Zweitens werden vier Modelle systematisch verglichen. Drittens wird der Prompt iterativ über vier Versionen weiterentwickelt. Viertens findet eine strukturierte Evaluation mit zwölf Fachexperten statt.

1.5 Aufbau der Arbeit

Die Arbeit gliedert sich in acht Kapitel. Kapitel 2 stellt die theoretischen Grundlagen dar: DAM-Systeme, Vision Language Models, Quantisierung, Prompt Engineering, Barrierefreiheit und Evaluationsmethodik. Kapitel 3 beschreibt das Forschungsdesign auf Basis von Design Science Research. Kapitel 4 leitet die Anforderungen ab, führt sie als Katalog und begründet die Architektur- und Modellentscheidungen. Kapitel 5 dokumentiert die Implementierung einschließlich der Prompt-Entwicklung. Kapitel 6 präsentiert die Evaluationsergebnisse. Kapitel 7 diskutiert die Befunde, beantwortet die Forschungsfragen und benennt Limitationen. Kapitel 8 fasst zusammen und gibt einen Ausblick.

2 THEORETISCHE GRUNDLAGEN

Dieses Kapitel stellt die theoretischen Konzepte vor, die für Entwicklung und Evaluation des Systems unmittelbar relevant sind. Die Darstellung beschränkt sich auf das zum Verständnis der technischen Entscheidungen und der Evaluationsmethodik Notwendige. Grundlagen zum Digital Asset Management wurden in der Projektarbeit (Klemm, 2025, S. 3–4) behandelt und werden hier nur insoweit aufgegriffen, als sie für die Einordnung der vorliegenden Arbeit erforderlich sind.

2.1 Digital Asset Management und Metadatenstrukturen

Digital-Asset-Management-Systeme verwalten digitale Medieninhalte über deren gesamten Lebenszyklus. Eine Darstellung der DAM-Grundlagen sowie die Ist-Analyse des Celum-Systems finden sich in der Projektarbeit (Klemm, 2025, S. 3–4, 11–12). Für diese Arbeit ist die Frage relevant, wie generierte Metadaten strukturiert und konsistent gehalten werden.

In der Literatur werden kontrollierte Vokabulare, Thesauri und Taxonomien als unterschiedliche Instrumente zur Vereinheitlichung von Metadatenwerten geführt (Baca, 2008, S. 19). Eine Taxonomie ist dabei eine meist hierarchisch geordnete Klassifikation, die die Beziehungen zwischen den klassifizierten Gegenständen ausdrücklich macht (Baca, 2008, S. 78). In der Praxis stehen diesem Ansatz flexible Alternativen gegenüber, etwa flache Tag-Systeme, Folksonomien oder hybride Lösungen, bei denen ein Regelwerk die Begriffswahl steuert, ohne eine starre Hierarchie vorzugeben.

Das in dieser Arbeit entwickelte System verfolgt einen solchen hybriden Ansatz. Statt einer formalen Taxonomie nutzt es eine prompt-gesteuerte Vokabularsteuerung. Der Prompt enthält domänenspezifische Standardbegriffe und Sprachkonventionen des Unternehmens. Diese Vorgaben lenken die Modellausgabe in Richtung konsistenter, unternehmenskonformer Begriffe. Die generierten Tags werden als flache Einträge in Celum gespeichert. Der TagResolver verknüpft sie

mit Celum-Collections. Existiert ein Begriff bereits, wird die bestehende Collection referenziert. Ist er neu, wird ein Eintrag angelegt. Der Titel der Arbeit verwendet den Begriff Unternehmenstaxonomien im weiteren Sinne als unternehmensspezifisches Begriffsregelwerk, nicht als formal-hierarchisches Klassifikationssystem. Entsprechend bezeichnet Alt-Text-Generierung die Erzeugung von Beschreibungsvorschlägen, deren Eignung als Alternativtext im jeweiligen Veröffentlichungskontext zu prüfen ist (Abschnitt 2.5.1).

2.2 Vision Language Models

2.2.1 Von Bildklassifikation zu multimodaler Textgenerierung

Bildklassifikation ordnet eine Eingabe einer von k vorgegebenen Kategorien zu (Goodfellow et al., 2016, S. 100). Für überwacht trainierte Modelle nennen Goodfellow et al. (2016, S. 20) etwa 5.000 annotierte Beispiele je Kategorie als damalige Faustregel, nicht als allgemeingültige Mindestmenge. Der Datenbedarf hängt unter anderem von Aufgabe, Vortraining und gewünschter Leistung ab. Bei variablen Beschreibungsanforderungen stößt eine feste Klassenmenge an Grenzen. Radford et al. (2021, S. 1) motivieren daher den Transfer aus natürlichsprachlicher Supervision.

Vision Language Models überwinden diese Einschränkung. Sie verbinden eine visuelle Komponente mit einem generativen Sprachmodell und überführen Bilder in natürlichsprachliche Ausgaben. Über textuelle Instruktionen lässt sich das Ausgabeformat steuern, ohne das Modell neu zu trainieren. Ein VLM kann auf dieselbe Eingabe unterschiedliche Ausgaben erzeugen, abhängig vom Prompt: eine Kurzbeschreibung, eine technische Analyse oder strukturierte Metadaten im JSON-Format.

2.2.2 Architektur moderner Vision Language Models

Viele Vision Language Models verbinden einen Bildencoder, eine Projektionsschicht und ein Sprachmodell. Vision Transformer zerlegen Bilder in Patches und verarbeiten diese als Sequenz (Dosovitskiy et al., 2021, S. 1, 3). Für LLaVA be-

schreiben H. Liu et al. (2023, S. 4) eine lineare Projektion zwischen dem vortrainierten CLIP-Bildencoder ViT-L/14 und dem Sprachmodell Vicuna. Die projizierten Bildrepräsentationen werden gemeinsam mit den Texttokens verarbeitet. Dieser Aufbau ist ein verbreitetes Beispiel, keine Definition sämtlicher VLM-Architekturen.

Für diese Arbeit bedeutet das: Bild und Produktkontext können in einem einzigen Verarbeitungsschritt berücksichtigt werden. Das Modell „sieht“ sowohl die visuellen Informationen als auch den textuellen Kontext aus dem PIM und erzeugt eine integrierte Ausgabe. Es ist kein separater Schritt zur Bildanalyse erforderlich.

2.2.3 Kontrastives Vortraining und Zero-Shot-Transfer

CLIP (Contrastive Language-Image Pre-training) bildet ein methodisches Fundament vieler aktueller VLMs (Radford et al., 2021; darauf aufbauend H. Liu et al., 2023, S. 4). Das Verfahren trainiert einen Bild- und einen Text-Encoder gemeinsam auf 400 Millionen Bild-Text-Paaren aus dem Internet (Radford et al., 2021, S. 1 f.). Durch kontrastives Lernen werden zusammengehörige Bild-Text-Paare im gemeinsamen Einbettungsraum nahe zueinander platziert, nicht zusammengehörige Paare voneinander getrennt. Die resultierende Repräsentation ermöglicht Zero-Shot-Transfer: Das Modell ordnet Bilder anhand natürlichsprachlicher Klassenbezeichnungen ein, ohne für den jeweiligen Datensatz trainiert worden zu sein (Radford et al., 2021, S. 6).

Für die Verschlagwortung heißt das: Kontrastiv vortrainierte Bildencoder ordnen Bildern textuelle Beschreibungen zu, ohne dass für jede neue Produktkategorie trainiert werden muss. Welche Encoder-Variante die hier verwendeten Modelle im Einzelnen einsetzen, ist unterschiedlich gut dokumentiert: für die Open-Weight-Modelle nur teilweise, für Claude Sonnet 5 gar nicht. Gemeinsam ist den aktuellen Architekturen, dass sie feste Klassenzuordnungen durch generative Sprachmodelle mit offener Textausgabe ersetzen.

2.2.4 Instruction Tuning

Rohe Vision-Language-Repräsentationen allein ermöglichen noch keine instruktionsgesteuerte Nutzung. Instruction Tuning bezeichnet das Feintuning auf Aufgaben-Antwort-Paare, um ein Modell an die Befolgung spezifischer Anweisungen anzupassen. Im visuellen Bereich haben H. Liu et al. (2023, S. 7) gezeigt, dass multimodale Aufgaben-Antwort-Paare die Instruktionsbefolgung bei Bildaufgaben deutlich verbessern; gemessen wurde das auf einem eigens konstruierten Benchmark mit GPT-4 als Bewerter. Die Modelle lernen, Instruktionen zu interpretieren und darauf abgestimmte Antworten zu generieren.

Für diese Arbeit ist Instruction Tuning aus zwei Gründen relevant. Es ermöglicht die Steuerung des Ausgabeformats über den Prompt, und es erlaubt die Formulierung domänenspezifischer Anforderungen in natürlicher Sprache. Die Modelle können angewiesen werden, JSON auszugeben, bestimmte Felder zu füllen oder Begriffe aus einem Kontext zu übernehmen. Instruction Tuning unterstützt damit die Befolgung der Aufgaben- und Formatvorgaben.

2.2.5 Untersuchte Modelle

Vier VLMs unterschiedlicher Herkunft und Größe werden evaluiert. Die Open-Weight-Modelle Gemma4:31b und Gemma4:e4b (Google DeepMind) sowie Qwen3.8-27B-MLX (Alibaba) laufen lokal auf einem Mac Studio M2 Ultra ohne externe Inferenzdienste. Claude Sonnet 5 (Anthropic) dient als Cloud-Referenz über AWS Bedrock. Tabelle 1 fasst die relevanten Eigenschaften zusammen und dokumentiert die jeweilige Ausführungsumgebung.

Tabelle 1: Spezifikation der untersuchten Modelle

Quelle: Eigene Darstellung; Gemma 4 nach Gemma Team, 2026, S. 2 f. und Google DeepMind, 2026, o. S.; Qwen3.8-27B nach Qwen Team, 2026, o. S. (Angaben zum Vision-Encoder aus der Konfigurationsdatei des veröffentlichten Modells); Claude Sonnet 5 nach Anthropic, 2026, o. S.

Merkmal	**Gemma4:31b**	**Gemma4:e4b**	**Qwen3.8-27B-MLX**	**Claude Sonnet 5**
Anbieter	Google DeepMind	Google DeepMind	Alibaba	Anthropic
Architektur	Dense	Dense mit Per-Layer Embeddings	Dense	nicht offengelegt
Parameter	30,7 Mrd.	4,5 Mrd. effektiv (8 Mrd. gespeichert)	27 Mrd.	nicht offengelegt
Vision-Encoder	550 Mio.	150 Mio.	27 Layer, Hidden 1.152 (Größe nicht beziffert)	nicht offengelegt
Kontextfenster	256K Token	128K Token	262K Token	1 Mio. Token
Ausführung	GGUF, 4 Bit, über Ollama	GGUF, 4 Bit, über Ollama	4 Bit laut Versuchsangabe	Cloud, AWS Bedrock
Hardware	Mac Studio M2 Ultra, 128 GB	Mac Studio M2 Ultra, 128 GB	Mac Studio M2 Ultra, 128 GB	AWS eu-central-1
Lizenz	Apache 2.0	Apache 2.0	Apache 2.0	proprietär

Die Angaben zu den Gemma-4-Varianten stammen aus dem Technical Report (Gemma Team, 2026, S. 2 f.) und der Modellkarte (Google DeepMind, 2026, o. S.). Für Qwen wird die im Evaluationslauf gespeicherte Kennung qwen3.8:27b-mlx verwendet. Die dokumentierten Ausführungsangaben sind in Anhang C zusammengefasst.

Die Modellbezeichnungen der Gemma-4-Familie bedürfen einer Erläuterung, da sie eine verbreitete Fehldeutung nahelegen. Gemma 4 umfasst die fünf Größen E2B, E4B, 12B, 26B-A4B und 31B (Gemma Team, 2026, S. 2). Das Präfix E bezeichnet die Zahl effektiver Parameter. E4B nutzt Per-Layer Embeddings (PLE): Die zugehörigen Embedding-Tabellen sind groß, werden aber nur für schnelle Nachschlagevorgänge verwendet. Deshalb liegt die effektive Parameterzahl deutlich unter der Gesamtzahl (Google DeepMind, 2026, o. S.). Bei rund acht Milliarden gespeicherten Parametern werden etwa 4,5 Milliarden effektive Parameter angegeben. Gemma4:e4b ist damit keine Quantisierung von Gemma4:31b, sondern ein eigenständiges Modell derselben Familie, das für den Betrieb auf Endgeräten ausgelegt ist.

Für die hier verglichenen Gemma-Konfigurationen ist die Dimensionierung des Bildencoders relevant. Gemma4:e4b verwendet einen Vision-Encoder mit 150 Millionen Parametern, Gemma4:31b einen mit 550 Millionen (Gemma Team, 2026, S. 2). Der Encoder des 31b-Modells besitzt somit etwa 3,7-mal so viele Parameter. Abschnitt 6.2 greift diesen Unterschied als mögliche Erklärungshypothese auf; sein eigenständiger Einfluss wird nicht gemessen.

2.3 Effizienzstrategien für lokale VLM-Inferenz

2.3.1 Speicherbedarf und Bandbreite als limitierender Faktor

Quantisierung reduziert die Bit-Breite der Modellgewichte und damit ihren Speicherbedarf (Dettmers et al., 2022, S. 4). Die Speicherangaben im Gemma-4-Bericht beziehen sich auf die dort beschriebenen Textinferenz- und Quantisierungskonfigurationen (Gemma Team, 2026, S. 3). Sie dürfen nicht als gemessener Gesamtspeicherbedarf der hier verwendeten multimodalen GGUF-Ausführung übernommen werden. Neben Gewichten benötigen unter anderem Bildverarbeitung, Aktivierungen und Key-Value-Cache Speicherplatz (RAM).

Für die Inferenzgeschwindigkeit ist der Zusammenhang weniger eindeutig. Die Generierung ist bei autoregressiven Modellen überwiegend an die Speicherbandbreite gebunden, da für jedes erzeugte Token die Gewichte einmal gelesen werden müssen. Eine geringere Bit-Breite reduziert dieses Lesevolumen und kann

die Generierung beschleunigen. Der tatsächliche Gewinn hängt jedoch von der Implementierung ab. Dettmers et al. (2022, S. 17) weisen darauf hin, dass ihre Int8-Matrixmultiplikation unterhalb einer Modelldimension von etwa 4.096 langsamer ausfällt als die Ausführung in FP16. Die in Abschnitt 6.3 berichteten Laufzeiten sind daher als Messwerte der konkreten Konfiguration zu lesen und nicht als allgemeine Eigenschaft der Bit-Breite.

Der verwendete Mac Studio M2 Ultra verfügt über 128 GB Unified Memory. Die Auswahl quantisierter Builds berücksichtigt den verfügbaren Speicher, den zusätzlichen Laufzeitbedarf und die praktische Unterstützung durch Ollama beziehungsweise MLX. Eine mögliche Beschleunigung ist implementierungsabhängig. Da keine unquantisierte Vergleichsausführung gemessen wurde, weist diese Arbeit keinen Geschwindigkeitsgewinn durch Quantisierung nach.

2.3.2 Quantisierungsverfahren

Post-Training Quantization (PTQ) quantisiert ein bereits trainiertes Modell nachträglich. GPTQ (Frantar et al., 2023, S. 1, 3) nutzt approximierte Informationen zweiter Ordnung zur schichtweisen Quantisierung. AWQ (J. Lin et al., 2024, S. 1 f.) bestimmt anhand der Aktivierungsverteilung die wenigen besonders wichtigen Gewichtskanäle und schützt sie über eine äquivalenzerhaltende Skalierung vor Quantisierungsfehlern. Das GGUF-Format bündelt quantisierte Gewichte in einer Datei und wird von Ollama verwendet. Die MLX-Quantisierung ist auf die Apple-Silicon-Architektur ausgelegt und arbeitet direkt im Unified-Memory-Modell.

In dieser Arbeit werden die beiden Gemma-Konfigurationen über Ollama in 4-Bit-GGUF-Quantisierung ausgeführt. Auch für die Qwen-Konfiguration wurde eine 4-Bit-Ausführung verwendet. Der Vergleich verwendet die von Ollama bereitgestellten Builds mit den angegebenen Namen und ist kein isolierter Vergleich der Quantisierungsverfahren.

2.3.3 Parametereffizienz jenseits der Quantisierung

Neben der Reduktion der Bit-Breite existiert eine zweite Strategie zur Senkung des Ressourcenbedarfs, nämlich die Reduktion der zur Laufzeit aktiven Parameter bei gleichbleibender Gesamtparameterzahl. Mixture-of-Experts-Architekturen

(MoE) aktivieren pro Token nur eine Teilmenge der Experten. Die Gemma-4-Variante 26B-A4B aktiviert 3,8 von insgesamt 26 Milliarden Parametern (Gemma Team, 2026, S. 2). Per-Layer Embeddings, wie sie in den Varianten E2B und E4B eingesetzt werden, verfolgen dasselbe Ziel auf anderem Weg.

Beide Verfahren senken den effektiven Ressourcenbedarf ohne die Zahlendarstellung der Gewichte zu verändern und sind damit von der Quantisierung zu unterscheiden. Für die Modellauswahl folgt daraus, dass Modellbezeichnungen wie E4B keine Quantisierungsangaben sind. Hinzu kommt eine dritte Dimension, die für multimodale Aufgaben unmittelbar relevant ist, die Dimensionierung des Vision-Encoders, die innerhalb einer Modellfamilie zwischen Größenvarianten erheblich abweichen kann (Abschnitt 2.2.5).

2.3.4 Konsequenzen für den Modellvergleich

Die untersuchten Modelle unterscheiden sich nicht in einer einzelnen Eigenschaft, sondern gleichzeitig in Parameterzahl, Architektur, Größe des Vision-Encoders, Quantisierungsverfahren und Laufzeitumgebung. Ein isolierter Quantisierungseffekt lässt sich in diesem Design nicht messen. Die Aussage, ein bestimmter Qualitätsunterschied sei auf die Bit-Breite zurückzuführen, wäre durch die vorliegenden Daten nicht gedeckt. Für sehr große Sprachmodelle berichten Frantar et al. (2023, S. 7) bei 4-Bit-Quantisierung minimale Einbußen, etwa 0,03 Perplexitätspunkte auf WikiText2 bei einem Modell mit 175 Milliarden Parametern. Dieselbe Quelle stellt jedoch fest, dass größere Modelle leichter zu quantisieren sind; bei kleinen Modellen fallen die Einbußen erheblich höher aus. Die Ergebnisse beziehen sich zudem auf unimodale Textmodelle und lassen sich nicht direkt auf multimodale Aufgaben übertragen.

Der Vergleich beantwortet stattdessen die praktisch bedeutsamere Frage, welche der real verfügbaren Konfigurationen für die konkrete Aufgabe die beste Qualität liefert und zu welchen Kosten. Das entspricht der Entscheidungssituation im Unternehmen. Dort stehen einsatzfähige Modellkonfigurationen zur Wahl und keine einzelnen Architektureigenschaften. Die Ergebnisse sind daher als Vergleich von Konfigurationen zu lesen und nicht als kontrollierte Variation einzelner Modelleigenschaften (vgl. Abschnitt 7.3).

2.4 Prompt Engineering für strukturierte Ausgaben

2.4.1 Begriffliche Einordnung

Prompt Engineering bezeichnet die systematische Gestaltung von Eingabeanweisungen zur Steuerung generativer Modelle (White et al., 2023, S. 1 f.). Identische Modelle erzeugen je nach Prompt-Formulierung qualitativ unterschiedliche Ausgaben. Die Prompt-Gestaltung ist damit eine eigenständige Entwurfsentscheidung mit erheblichem Einfluss auf die Ergebnisqualität.

2.4.2 Prompt-Strategien

Zero-Shot Prompting gibt eine Aufgabenbeschreibung ohne Beispiele. Few-Shot Prompting ergänzt Beispiele als Kontext zur Inferenzzeit, ohne die Modellgewichte anzupassen (T. B. Brown et al., 2020, S. 6). Chain-of-Thought-Prompting ergänzt ausformulierte Zwischenschritte. Wei et al. (2022, S. 1, 4) beobachten Vorteile insbesondere bei den großen Modellen ihrer damaligen Versuchsreihe. Die dort diskutierte Größenordnung von etwa 100 Milliarden Parametern ist ein Befund zu diesen Modellen und Aufgaben, keine universelle Grenze für spätere Modellgenerationen.

Die in dieser Arbeit entwickelten Prompts verwenden überwiegend Zero-Shot-Instruktion mit strukturellen Vorgaben. Die ersten drei Prompt-Versionen enthalten keine vollständigen Eingabe-Ausgabe-Beispiele. Der variable PIM-Kontext liefert pro Bild unterschiedliche Ausgangsdaten und statische Beispiele könnten die Modelle dabei irreführen. Version V4 ergänzt Beispielformulierungen für den Alt-Text-Abschnitt, um den erwarteten Stil und die Informationsdichte zu veranschaulichen. Diese Stilbeispiele bilden keine vollständigen Few-Shot-Paare aus Bildeingabe und Ausgabe. Die Anweisungen definieren darüber hinaus Regeln, Formate und Restriktionen, die modellübergreifend gelten.

2.4.3 Strukturierte Ausgabegenerierung

Für die automatisierte Weiterverarbeitung muss die Modellausgabe maschinenlesbar sein. Zwei Ansätze existieren. Prompt-basierte Steuerung instruiert textlich ein bestimmtes Format. Constrained Decoding erzwingt ein Schema auf Token-Ebene (Willard & Louf, 2023, S. 3 f.). Letzteres garantiert die Struktur der Ausgabe und arbeitet modellunabhängig mit im Mittel konstantem Zusatzaufwand je Token (Willard & Louf, 2023, S. 1 f.). Es setzt allerdings Zugriff auf den Dekodierungsschritt und einen über das Vokabular des jeweiligen Tokenizers vorberechneten Index voraus. Diese Funktion ist nicht bei allen Modellen verfügbar und würde die freie Modellwahl einschränken.

Das entwickelte System verwendet prompt-basierte Steuerung mit vorgegebenem JSON-Ausgabeformat über die unterstützten Inferenz-Backends. Die Antwort wird eingelesen und anschließend normalisiert und validiert. Fehler beim Einlesen können einen erneuten Versuch auslösen. Validierungsfehler wie fehlende Pflichtfelder führen zu einer Kennzeichnung zur manuellen Prüfung.

2.4.4 Iterative Prompt-Optimierung

Zamfirescu-Pereira et al. (2023, S. 1, 9) untersuchen, wie Personen ohne KI-Expertise Prompts entwerfen. Ihre zentralen Befunde sind: Der Prozess verläuft typischerweise opportunistisch statt systematisch, aus Einzelbeobachtungen wird übergeneralisiert und Erwartungen aus der Mensch-Mensch-Kommunikation werden auf das Modell übertragen. Die Studie belegt damit nicht die Überlegenheit iterativer gegenüber einmaliger Instruktion, sondern die Schwächen ungeleiteter Iteration.

Die vorliegende Arbeit begegnet diesen Schwächen durch einen dokumentierten Zyklus aus Generierung, Bewertung und Anpassung. Beobachtungen an realen Celum-Assets und Rückmeldungen aus dem Fachbereich begründen die Instruktionsänderungen. Abschnitt 5.3 macht diesen Prozess über vier Versionen nachvollziehbar. Vergleichend bewertet wird V3; ein quantitativer Nachweis der Qualitätssteigerung zwischen den Versionen ist damit nicht erbracht (vgl. Abschnitt 7.2).

2.5 Barrierefreiheit und Alt-Text-Generierung

2.5.1 Normative Grundlagen

Erfolgskriterium 1.1.1 der WCAG 2.2 verlangt zweckentsprechende Textalternativen für Nicht-Text-Inhalte und nennt Ausnahmen beziehungsweise besondere Behandlungen, etwa für rein dekorative Inhalte (W3C, 2024). Entscheidend ist der Einsatzkontext: Bei funktionalen Bildern ist die Funktion relevant, dekorative Bilder können eine leere Textalternative erhalten. Die Richtlinie (EU) 2019/882 legt für ihren sachlichen Anwendungsbereich funktionale Barrierefreiheitsanforderungen fest und regelt die Konformitätsvermutung über harmonisierte Normen (Art. 15). Sie begründet keine undifferenzierte Beschreibungspflicht für den gesamten internen DAM-Bestand. Die Arbeit orientiert sich an WCAG 2.2, erhebt aber keine rechtliche oder technische Konformitätsprüfung eines vollständigen Webangebots.

Ein Alternativtext soll den Zweck eines Bildes in seinem konkreten Verwendungskontext vermitteln. Für ein informatives Produktbild können Produktmerkmale relevant sein und bereits im umgebenden Text genannte Informationen können dagegen redundant sein. Da die Pipeline diesen späteren Seitenkontext nicht vollständig kennt, erzeugt sie wiederverwendbare Beschreibungstexte. Ihre Übernahme als Alternativtext erfordert eine kontextbezogene redaktionelle Entscheidung.

2.5.2 Qualitätskriterien für Alt-Texte

Die Forschung zur Barrierefreiheit von Bildbeschreibungen betont vor allem deren Kontextabhängigkeit. Welche Informationen ein Alternativtext enthalten muss, ergibt sich aus dem Verwendungszusammenhang des Bildes und nicht allein aus seinem Inhalt (Kreiss et al., 2022, S. 4685 f.). Aus dieser Argumentation und dem Erfolgskriterium 1.1.1 leitet die vorliegende Arbeit fünf Qualitätsdimensionen für Alternativtexte ab: inhaltliche Korrektheit, Informationsdichte ohne Redundanz, Kontextangemessenheit, angemessene Länge und Vermeidung dekorativer Formulierungen. Diese Zusammenstellung ist eine eigene Synthese und keine in der Literatur ausgewiesene Taxonomie. Sie ist nicht identisch mit den

fünf Bewertungsdimensionen, die Kreiss et al. (2022, S. 4688) in ihrer Erhebung verwenden.

Für die fachliche Akzeptanz sind außerdem offizielle Produkt- und Oberflächenbezeichnungen relevant. Abbildung 1 zeigt dies an deutschen Such-Tags: Die Fachpersonen beanstanden die Übersetzung der Oberflächenbezeichnung Brushed Nickel und die zusätzliche Farbangabe. Das Beispiel illustriert eine Terminologievorgabe für Tags und nicht für Beschreibungstexte. Die Ausgaberegeln beider Dimensionen sind getrennt zu beachten (Abschnitt 4.2). Die beiden Kommentare wurden zu den anonymisierten Varianten dieses Bildes abgegeben und sind keinem einzelnen Modell eindeutig zugeordnet (Anhang D, Nr. 26 und 60).



Asset 747358 (hpr10438.tif), Freisteller

Offizielle Bezeichnung im PIM hansgrohe Logis, Oberfläche *Brushed Nickel*

Ausgabe des Modells (Tags, Deutsch) … brushed nickel; *gebürstetes nickel*; *silberfarben* …

Beanstandung der Fachexperten „Brushed Nickel wird auch in Deutsch so gesucht und nicht gebürstetes Nickel, daher keine Übersetzung von Namen.“ · „‚silberfarben‘ irreführend.“

Abbildung 1: Offizielle Oberflächenbezeichnung aus dem PIM und Ausgabe des Modells im Vergleich

Quelle: Eigene Darstellung; Bild und Produktdaten: Hansgrohe SE, Celum-Asset 747358; Modellausgabe aus dem Evaluationslauf (Gemma4:31b, Prompt V3)

2.5.3 VLMs als Alt-Text-Generatoren

VLMs können gegenüber früheren Ansätzen kontextbezogene Beschreibungen produzieren, die über eine Aufzählung sichtbarer Elemente hinausgehen. Die Herausforderung liegt in der Vermeidung von Halluzinationen und der Einhaltung domänenspezifischer Konventionen. Modelle tendieren dazu, plausible, aber nicht korrekte Details zu ergänzen, wenn das Bild allein nicht genügend Information liefert. Auch die Evaluation dieser Arbeit enthält solche Fälle, etwa ein nicht vorhandenes Mikrofon oder eine vermeintlich mit Schaum gefüllte Badewanne

(Abschnitt 6.4). Die PIM-Anreicherung stellt hierfür eine ergänzende Informationsquelle bereit.

2.6 Evaluation generativer KI-Ausgaben

2.6.1 Grenzen automatischer Textmetriken

BLEU und ROUGE vergleichen n-Gramme mit Referenztexten. Ein n-Gramm ist eine zusammenhängende Folge von n Wörtern: Ein Unigramm umfasst ein einzelnes Wort, ein Bigramm zwei aufeinanderfolgende Wörter, ein Trigramm drei. BLEU verwendet modifizierte Präzision und eine Längenstrafe (Papineni et al., 2002, S. 311 f., 315), ROUGE-N einen Recall-Bezug (C.-Y. Lin, 2004, S. 74). Die lexikalische Übereinstimmung ist bei offenen Generierungsaufgaben jedoch nur ein begrenzter Stellvertreter für Qualität: Inhaltlich angemessene Formulierungen können stark voneinander abweichen (C.-W. Liu et al., 2016, S. 2122; Novikova et al., 2017, S. 2248 f.). Für die vorliegende Erhebung fehlen außerdem Referenzannotationen. Daher wurde von diesen Metriken für die Bewertung abgesehen. Zheng et al. (2023, S. 1 f.) untersuchen ergänzend Sprachmodelle als Bewerter und vergleichen deren Urteile mit menschlichen Präferenzen. Eine pauschale Unbrauchbarkeit automatischer Metriken folgt daraus nicht.

2.6.2 Präferenzbasierte Vergleiche

Menschliche Bewertungen erlauben es, die Angemessenheit generativer Ausgaben im Nutzungskontext zu erfassen (Zheng et al., 2023, S. 2). Mögliche Verfahren sind absolute Bewertungen auf Skalen, paarweise Vergleiche und die Auswahl einer bevorzugten Variante aus mehreren Ausgaben. Skalen setzen einen gemeinsamen Bewertungsmaßstab voraus und vollständige Paarvergleiche erfordern bei N Varianten $N \cdot (N-1)/2$ Entscheidungen pro Item.

In dieser Arbeit kommt ein Best-of-4-Design zum Einsatz. Pro Bild sehen die Teilnehmenden die Ausgaben aller vier Modelle und wählen die bevorzugte Variante. Das Design reduziert den Bewertungsaufwand gegenüber paarweisen Vergleichen, die bei vier Varianten sechs Paarungen pro Bild erfordern würden, liefert

dafür aber keine vollständige Rangordnung. Die Operationalisierung ist in Abschnitt 6.1 dokumentiert.

2.6.3 Präferenzquoten und statistische Einordnung

Die Präferenzquote beschreibt den Anteil der Stimmen für ein Modell an einer ausdrücklich festgelegten Bezugsmenge. Der Anteil an allen Urteilen schließt Entscheidungen ohne eindeutige Modellwahl ein. Der bedingte Anteil betrachtet nur konkrete Modellwahlen. Beide Größen beschreiben die erhobenen Präferenzen, keine absolute Qualität. Für eine Übertragung auf andere Personen oder Bilder sind zusätzlich das Auswahlverfahren und die Abhängigkeit wiederholter Urteile zu berücksichtigen.

Für unabhängige binäre Beobachtungen können Wilson-Intervalle die Unsicherheit eines Anteils beschreiben und weisen gegenüber einfachen Wald-Intervallen günstigere Überdeckungseigenschaften auf (L. D. Brown et al., 2001, S. 101 f., 107, 110). In dieser Untersuchung bewerten dieselben Personen dieselben Bilder. Die deskriptiven Präferenzanteile stehen daher im Vordergrund; ergänzende Standardtests ersetzen keine Modellierung dieser Personen- und Bildabhängigkeiten.

2.6.4 Begründung des Evaluationsansatzes

Das Best-of-4-Design ermöglicht einen kompakten relativen Vergleich ohne vorliegende Referenzannotationen. Es liefert allerdings weder absolute Qualitätswerte noch einen Nachweis, dass eine Mindestqualität erreicht wird. Absolute, kriterienspezifische Bewertungen wären hierfür ergänzend erforderlich. Eine höhere Zuverlässigkeit gegenüber solchen Bewertungen wird in dieser Arbeit nicht untersucht.

2.7 Themenbezogene Arbeiten

Die vorangegangenen Abschnitte stellen die Konzepte dar, auf denen das System aufbaut. Dieser Abschnitt ordnet die Arbeit in den Forschungsstand ein. Leitend sind zwei Fragen: Welche Arbeiten behandeln dieses oder ein verwandtes Problem? Und warum beantworten sie die hier gestellten Forschungsfragen nicht bereits? Die Darstellung folgt fünf Strängen, die für die Fragestellung zusammenlaufen müssen.

Der erste Strang betrifft Metadaten im Digital Asset Management. Baca (2008) beschreibt die Strukturierung von Metadaten über kontrollierte Vokabulare, Thesauri und Taxonomien. Automatisierte Metadatenerzeugung wird dort auf konzeptioneller Ebene behandelt, etwa als Metadata Mining, Harvesting oder Web Crawling (Baca, 2008, S. 3, 10, 72). Die generative Beschreibung von Bildinhalten nach unternehmenseigenen Begriffskonventionen liegt außerhalb dieses Rahmens. Die vorangegangene Projektarbeit schließt hier an. Sie erhebt Anforderungen und vergleicht Umsetzungsoptionen, bleibt aber in der Analysephase und implementiert kein System (Klemm, 2025, S. 2, 8).

Der zweite Strang betrifft Vision Language Models. Radford et al. (2021, S. 1 f.) verbinden Bild- und Textrepräsentationen in einem gemeinsamen Einbettungsraum. Dosovitskiy et al. (2021, S. 1) übertragen die Transformer-Architektur auf Bilder, und H. Liu et al. (2023, S. 4) machen solche Modelle über Instruction Tuning instruierbar. Diese Arbeiten optimieren auf allgemeine Bildinhalte und werden an offenen Benchmarks gemessen. Unternehmenseigene Begriffskonventionen bilden sie nicht ab. Ein Teil der in dieser Arbeit geforderten Angaben lässt sich nicht zuverlässig visuell bestimmen. Marke, genaue Linie und offizieller Oberflächenname können über das Produktdatensystem bereitgestellt werden (Abschnitt 4.2).

Der dritte Strang betrifft Alternativtexte. Kreiss et al. (2022, S. 4685 f.) zeigen, dass die Angemessenheit einer Bildbeschreibung vom Verwendungskontext abhängt und nicht allein vom Bildinhalt. Den normativen Rahmen geben die WCAG 2.2 und die Richtlinie (EU) 2019/882 vor (W3C, 2024, o. S.; Europäisches Parlament, 2019). Diese Quellen begründen, welche Eigenschaften ein Alternativtext

haben soll. Wie der nötige Kontext in einem Unternehmen technisch bereitgestellt und in die Generierung eingespeist wird, behandeln sie nicht.

Der vierte Strang betrifft die Steuerung über Instruktionen. White et al. (2023, S. 2–4) systematisieren Prompt-Muster, Wei et al. (2022, S. 1) beschreiben Chain-of-Thought-Prompting, Willard und Louf (2023, S. 3 f.) erzwingen gültige Ausgabestrukturen auf Token-Ebene. Zamfirescu-Pereira et al. (2023) untersuchen, wie Personen ohne KI-Hintergrund beim Entwurf von Prompts vorgehen. Die Literatur liefert damit Muster und Werkzeuge. Der hier untersuchte Fall verlangt aber eine schrittweise Anpassung eines Prompts an eine Fachterminologie, deren Fehlerklassen erst durch Rückmeldungen aus dem Fachbereich sichtbar werden.

Der fünfte Strang betrifft die Bewertung generativer Ausgaben. Automatische Metriken wie BLEU und ROUGE setzen Referenztexte voraus (Papineni et al., 2002, S. 311 f.; C.-Y. Lin, 2004, S. 74). Ihre Grenzen bei offenen Ausgaben sind mehrfach belegt (C.-W. Liu et al., 2016, S. 2122; Novikova et al., 2017, S. 2248 f.). Zheng et al. (2023, S. 7) untersuchen menschliche Präferenzen als Bezug für die Bewertung durch Sprachmodelle. Deren Urteile stammen teils von einem allgemeinen Publikum, teils von als Experten eingestuften Annotierenden und überwiegend von Studierenden in Master- und Promotionsprogrammen („graduate students“). Die Studie richtet sich nicht gezielt an den DAM-Kontext eines Sanitärherstellers. Für die vorliegende Aufgabe werden daher Personen mit entsprechender Nutzungserfahrung einbezogen. Ob die Tags die Suche tatsächlich verbessern, wäre ergänzend durch Suchaufgaben zu untersuchen.

Jeder Strang wurde für sich gut untersucht. Für die hier gestellte Frage müssen sie jedoch zusammenlaufen. Benötigt werden ein Modell, das den Bildinhalt erkennt, ein Produktdatensystem, das die nicht sichtbaren Angaben liefert, eine Instruktion, die beides zu unternehmenskonformen Begriffen verbindet, sowie eine Bewertung durch Personen, die den Nutzungskontext kennen. In der für diese Arbeit gesichteten Literatur beschreibt keine Untersuchung diese Verbindung an einem produktiv betriebenen System. Diese Lücke bearbeiten die Forschungsfragen F1 bis F3 (Abschnitt 1.3).

Menschliche Präferenzevaluation ist für diese Aufgabenklasse etabliert. Zheng et al. (2023, S. 4, 7) setzen dafür paarweise Vergleiche ein und aggregieren sie über Siegquoten. Bei vier Varianten erforderte dieses Verfahren sechs Paarungen pro Bild. Da Domänenexperten eine knappe Ressource sind, weicht diese Arbeit auf ein Best-of-4-Design aus. Es reduziert die Zahl der Entscheidungen pro Bild und Dimension auf eine. Der Preis dafür: Weder eine vollständige Rangordnung noch ein Abstand zwischen den Varianten lässt sich schätzen. Die Auswertung beschränkt sich deshalb auf Siegquoten.

3 FORSCHUNGSMETHODIK

Diese Arbeit folgt dem Rahmen der Design Science Research (DSR) nach Peffers et al. (2007). Das Kapitel begründet die Methodenwahl, ordnet die Arbeitsschritte dem DSR-Phasenmodell zu und erläutert das Evaluationsdesign.

3.1 Design Science Research als Rahmen

3.1.1 Begründung der Methodenwahl

DSR eignet sich für Forschungsarbeiten, deren Ziel die Konstruktion und Evaluation eines IT-Artefakts ist (Hevner et al., 2004, S. 77). Im Unterschied zu verhaltenswissenschaftlichen Ansätzen, die bestehende Phänomene erklären, zielt DSR auf die Schaffung neuer Artefakte, die ein identifiziertes Problem adressieren. Das Artefakt selbst und der Erkenntnisgewinn aus seiner Evaluation bilden den wissenschaftlichen Beitrag.

Die Arbeit orientiert sich an den sieben Guidelines von Hevner et al. (2004, S. 82 f.). Ein relevantes organisatorisches Problem existiert (fehlende Verschlagwortung bei 107.000 Assets, Guideline 2), ein IT-Artefakt adressiert es (das entwickelte System, Guideline 1), die Evaluation erfolgt systematisch (präferenzbasierter Modellvergleich mit Fachexperten, Guideline 3), und der Beitrag ist das neue Wissen über die Eignung verschiedener VLMs für domänenspezifische Bildverschlagwortung (Guideline 4). Die methodische Fundierung von Konstruktion und Evaluation (Guideline 5) dokumentiert Kapitel 3, die iterative Lösungssuche unter den Randbedingungen der Systemlandschaft (Guideline 6) die Kapitel 4 und 5, die Vermittlung an fachliches wie technisches Publikum (Guideline 7) diese Arbeit und die interne Systemdokumentation.

3.1.2 Zuordnung zum Sechs-Phasen-Modell

Peffers et al. (2007, S. 46, 52–56) gliedern den DSR-Prozess in sechs Aktivitäten:

(1) Problem Identification and Motivation – Das Problem wird identifiziert und seine praktische Relevanz begründet.

(2) Define the Objectives for a Solution – Quantitative oder qualitative Ziele für die angestrebte Lösung werden formuliert.

(3) Design and Development – Das Artefakt wird entworfen und implementiert.

(4) Demonstration – Das Artefakt wird in einem geeigneten Kontext eingesetzt, um seine Funktionsfähigkeit zu zeigen.

(5) Evaluation – Das Artefakt wird anhand der definierten Ziele bewertet.

(6) Communication – Die Ergebnisse werden an die relevanten Zielgruppen vermittelt.

Die Phasen sind nicht streng sequenziell. Peffers et al. (2007, S. 56) sehen ausdrücklich vor, nach der Evaluation zur Design-Phase zurückzukehren, und beschreiben vier mögliche Einstiegspunkte in den Prozess. Die vorliegende Arbeit steigt zielorientiert ein, weil die Identifizierung des Problems bereits in der vorangegangenen Projektarbeit erfolgte.

Phase 1 (Problem Identification) wurde in der vorangegangenen Projektarbeit durchgeführt (Klemm, 2025). Durch Experteninterviews und einen explorativen Test wurde der Bedarf an automatisierter Verschlagwortung identifiziert, die Dimension des Problems quantifiziert (rund 529.000 Assets, von denen laut Systemstatistik weniger als fünf Prozent vollständig verschlagwortet waren; Klemm, 2025, S. 11) und die grundsätzliche Machbarkeit KI-gestützter Bildbeschreibung bestätigt.

Phase 2 (Define Objectives) leitet aus dem identifizierten Problem konkrete Anforderungen ab. Kapitel 4 dokumentiert die funktionalen und nichtfunktionalen Anforderungen. Dazu zählen domänenspezifische Tags und Bildbeschreibungen zur Verwendung als Alternativtexte in zwei Sprachen, modulare Architektur, Vermeidung doppelter Jobeinplanung, Fehlertoleranz und die betriebliche EU-Vorgabe zur Bildverarbeitung. Als Evaluationskriterium legt die Forschungsfrage F2 einen präferenzbasierten Modellvergleich durch Fachexperten fest.

Phase 3 (Design and Development) bildet den implementierungstechnischen Kern dieser Arbeit und erstreckt sich über Kapitel 4 und 5. Die Architekturkonzeption definiert die Systemstruktur: eine KI-Abstraktionsschicht für verschiedene Backends, eine SQLite-basierte Job-Queue mit Retry-Logik, die PIM-Anbindung über Centric PXM und eine konfigurierbare Filterlogik auf Basis der mit den Fachbereichen definierten Kriterien. Die Implementierung setzt diese Konzeption in einer modularen Python-Applikation um. Die Prompt-Entwicklung über vier Versionen (V1 bis V4) folgt dabei einem eigenständigen Evaluate-Refine-Zyklus innerhalb der Design-Phase. Jede Version wird gegen reale Celum-Assets getestet. Schwächen werden durch Stichproben und Expertenrückmeldungen identifiziert und in der nächsten Iteration adressiert.

Phase 4 (Demonstration) erfolgt durch den produktiven Einsatz in der Zielumgebung. Das System verarbeitet den realen Celum-Bestand und wird für neue Uploads durch Webhooks angestoßen. Der betriebliche Erstlauf in Abschnitt 7.1 dokumentiert die Verarbeitung in großem Umfang und ergänzt eine vergleichende Evaluation.

Phase 5 (Evaluation) ist in Kapitel 6 dokumentiert. Vier Modellkonfigurationen werden anhand von 30 gezielt ausgewählten Bildern durch zwölf Fachexperten präferenzbasiert bewertet. Je Person entstehen 60 Entscheidungen zu deutschen Beschreibungen und Tags. Ergänzend werden technische Kennzahlen und 77 Freitextkommentare ausgewertet. Die Ergebnisse unterstützen die Modellwahl und liefern fachliche Hinweise für V4. Die Rückkopplung von Evaluation und Design ist damit Teil des Entwicklungsprozesses.

Phase 6 (Communication) wird durch die vorliegende Arbeit realisiert. Die Dokumentation umfasst Architekturentscheidungen, Implementierungsdetails, Evaluationsergebnisse und deren Einordnung in den Forschungskontext. Ergänzend dient die interne technische Dokumentation des Systems der Wissenssicherung im Unternehmen und ermöglicht die Weiterführung und Anpassung durch andere Entwickelnde.

3.1.3 Artefakt-Typ und wissenschaftlicher Beitrag

Nach March und Smith (1995, S. 258) ist das entwickelte Softwaresystem als Instanziierung einzuordnen. Es realisiert das Architekturkonzept in seiner betrieblichen Umgebung. Der produktive Einsatz demonstriert die Integration, und die separate Präferenzerhebung untersucht die fachliche Akzeptanz der erzeugten Ausgaben.

Der wissenschaftliche Beitrag liegt im Vergleich von vier VLM-Konfigurationen für die domänenspezifische Bildverschlagwortung sowie in der dokumentierten Entwicklung von Prompt und PIM-Integration. Die Arbeit verbindet damit einen Modellvergleich unter betrieblichen Bedingungen mit einer praktisch nutzbaren Integrationslösung.

3.2 Iteratives Vorgehen

3.2.1 Entwicklungsphasen

Die Entwicklung folgt dem DSR-Prinzip schrittweiser Verfeinerung in Zyklen aus Entwicklung und Bewertung (build-and-evaluate loop; Hevner et al., 2004, S. 78). Fünf Arbeitsschritte lassen sich unterscheiden.

In der Asset-Analyse wurde die Grundgesamtheit ermittelt. Die Filterkriterien wurden in Abstimmung mit der Celum-Administration und den Fachbereichen in mehreren Meetings definiert. Diese sind, relevante Collections, zulässige Bildtypen, Dateiformate, vorhandene Verknüpfungen und bereits vergebene Metadaten. Das Ergebnis: 107.000 von 549.000 Assets sind für die automatisierte Verschlagwortung relevant.

Die Pipeline-Implementierung umfasst den modularen Aufbau der Verarbeitungskette aus Ermittlung, Filterung, KI-Analyse mit Kontextanreicherung, Normalisierung und Rückschreiben in Celum. Jede Stufe ist als eigenständige Komponente realisiert, die unabhängig testbar ist.

Die PIM-Anbindung ergänzt visuell nicht zuverlässig bestimmbare Produktangaben. Über die Centric PXM-API werden Produktdaten zur Laufzeit abgerufen und als Prompt-Bestandteil eingefügt.

Die Prompt-Iteration stellt auch einen expliziten DSR-Evaluate-Refine-Zyklus dar. Von einer allgemeinen Instruktion (V1) wurden über vier Versionen systematisch identifizierte Schwächen adressiert. Grundlage waren jeweils die manuelle Inspektion der generierten Ausgaben und Rückmeldungen der Fachexperten.

Als Evaluation wurde der vergleichende Modelltest durchgeführt. Die Toolchain zur Umfragegenerierung und -auswertung wurde als zusätzliches Artefakt entwickelt.

3.2.2 Rolle der KI-gestützten Entwicklung

Die Implementierung erfolgte unter Einsatz generativer KI-Werkzeuge (Claude Code, GitHub Copilot) als Entwicklungsassistenten. Im Sinne wissenschaftlicher Transparenz wird dies offengelegt. Architekturentscheidungen, Prompt-Design und Evaluationsmethodik lagen beim Verfasser. Die KI-Werkzeuge beschleunigten die Umsetzung, ersetzten aber keine fachlichen Entscheidungen. Generative Werkzeuge unterstützten außerdem Quellenprüfung, Textüberarbeitung und die Überprüfung der Auswertung. Die Nutzung ist im Verzeichnis der Abfragen am Ende der Arbeit dokumentiert.

3.3 Evaluationsdesign

3.3.1 Überblick

Die Evaluation vergleicht vier VLM-Konfigurationen anhand identischer Eingabedaten: 30 Bilder, identischer Prompt V3, identische PIM-Kontextdaten. Die zentrale Ergebnisgröße ist die Präferenz der Fachexperten, ergänzt durch objektive Leistungsmetriken und qualitative Kommentare.

Die vier Varianten sind Gemma4:31b und Gemma4:e4b als zwei Größenvarianten derselben Modellfamilie, Qwen3.8-27B-MLX als alternatives Open-Weight-Modell und Claude Sonnet 5 als Cloud-Referenz. Da Prompt und Eingabedaten konstant bleiben, werden die Ausgaben unter vergleichbaren fachlichen Eingabebedingungen beurteilt. Welche einzelne Eigenschaft beobachtete Unterschiede verursacht, lässt das Design offen. Die Konfigurationen unterscheiden sich in mehreren Merkmalen gleichzeitig (Abschnitt 2.3.4).

3.3.2 Bewertungsmethode

Fachexperten wählen für jedes Bild aus vier anonymisierten deutschen Ausgaben die bevorzugte Variante für Tags und Beschreibung. Die zusätzliche Option lautet „Keine Variante ist geeignet bzw. die Varianten sind nicht sinnvoll unterscheidbar.“ Freitextkommentare ergänzen die Auswahl. Aufgrund der Sprachkenntnisse der Bewertenden beschränkt sich die Erhebung bewusst auf Deutsch. Umfang und Ablauf sind in Abschnitt 6.1 dokumentiert.

Die Teilnehmenden wurden bewusst aus einem kleinen Kreis ausgewählt: Personen, die täglich mit Celum-Assets arbeiten und die Anforderungen an Bildmetadaten aus erster Hand kennen. Domänenexpertise wurde höher gewichtet als Stichprobengröße, um Suchbegriffe und Produktterminologie im konkreten Nutzungskontext beurteilen zu lassen.

Ergänzend werden gespeicherte Inferenzzeiten, Token-Zahlen und technische Erfolgsangaben ausgewertet. Darauf aufbauende Kostenszenarien unterstützen die betriebliche Abwägung neben der fachlichen Präferenz.

3.3.3 Statistische Auswertung

Die Auswertung berichtet absolute Stimmen und Präferenzanteile mit ausgewiesener Bezugsgröße. Die beigefügten Auswertungsskripte berechnen ergänzend Wilson-Intervalle, Cochrans Q und paarweise McNemar-Tests mit Bonferroni-Korrektur (L. D. Brown et al., 2001, S. 107; Cochran, 1950, S. 259 f.; McNemar, 1947, S. 156). Diese Standardverfahren berücksichtigen die wiederholten Urteile derselben Personen und Bilder nicht vollständig. Die Schlussfolgerungen der Arbeit stützen sich deshalb auf die deskriptiven Ergebnisse. Fleiss’ Kappa beschreibt ergänzend die Übereinstimmung der Auswahlentscheidungen (Fleiss, 1971, S. 378), der Wert wird nicht als Maß absoluter Ausgabequalität interpretiert.

4 ANFORDERUNGEN UND KONZEPTION

Dieses Kapitel überführt die Problemstellung aus Kapitel 1 in ein Systemkonzept. Es entspricht der zweiten und dritten Aktivität des Forschungsdesigns, der Zieldefinition und dem Entwurf des Artefakts (Abschnitt 3.1.2). Besprochen werden die Quellen der Anforderungen, der Anforderungskatalog, die Architektur- und Modellentscheidungen und abschließend die Filterlogik für die Asset-Auswahl. Der Abschnitt ordnet jeder Anforderung ihre Umsetzung und ihren Nachweis zu. Die Ist-Analyse ist nicht Gegenstand dieser Arbeit. Sie liegt mit der vorangegangenen Projektarbeit vor (Klemm, 2025, S. 11–12).

4.1 Ableitung der Anforderungen

Die Anforderungen stammen aus drei Quellen. Erstens aus der vorangegangenen Projektarbeit, dort wurden in Experteninterviews die Erwartungen der Stakeholder erhoben und mit einem Machbarkeitstest an zehn Bildern die ersten technischen Anhaltspunkte gewonnen (Klemm, 2025, S. 2, 10). Zweitens aus den Rahmenbedingungen der bestehenden Systemlandschaft, also aus Celum, Centric PXM und den IT-Richtlinien des Unternehmens. Drittens aus den rechtlichen Vorgaben zur Barrierefreiheit, wie in Abschnitt 2.5 dargestellt.

Die Projektarbeit benennt drei wiederkehrende Probleme im Tagging-Prozess: Zeitdruck beim Upload, fehlende technische Unterstützung für semantische Vorschläge und eine eingeschränkte Suchqualität bei unvollständigen Metadaten (Klemm, 2025, S. 12). Als Erwartungen der Stakeholder hält sie die Automatisierung des Tagging-Prozesses, die Einhaltung der Cloud-Richtlinie mit EU-Hosting und eine nahtlose Integration in die bestehende Systemlandschaft fest (Klemm, 2025, S. 12). Im gewichteten Vergleich der Lösungsoptionen erreicht die Eigenentwicklung die beste Gesamtbewertung. Der verfügbare Celum-Adapter stößt nach dieser Einschätzung bei firmenspezifischen Taxonomien, Mehrsprachigkeit und Modellfreiheit an Grenzen. Die Projektarbeit verbindet diese Einschätzung

allerdings mit einer Aufwand-Nutzen-Prüfung und empfiehlt kurzfristig zunächst einen Test des fertigen Adapters (Klemm, 2025, S. 31 f.).

Die vorliegende Arbeit knüpft an diese Ergebnisse an, übernimmt die erhobenen Anforderungen aber nicht unverändert. An zwei Stellen weicht sie ab. Die Projektarbeit übernimmt aus den Stakeholder-Erwartungen eine Akzeptanzschwelle von etwa 75 Prozent F1-Score (Klemm, 2025, S. 12, 14). Ein solcher Wert setzt annotierte Sollwerte voraus, die für den Celum-Bestand nicht existieren. Abschnitt 2.6.1 begründet zudem, warum n-Gramm-basierte Metriken für offene Bildbeschreibungen nur begrenzt aussagekräftig sind. An die Stelle der Metrik tritt deshalb die präferenzbasierte Bewertung durch Fachexperten (Kapitel 6). Zweitens setzt die vorliegende Arbeit den Alternativtext produktiv um und lässt ihn durch Fachexperten bewerten. Die Projektarbeit sah ihn bereits als Ausgabe vor, erprobte ihn an zehn Bildern und nutzte ihn als Kriterium im Vergleich der Lösungsoptionen (Klemm, 2025, S. 13, 18, 30).

Die folgenden beiden Abschnitte führen die Anforderungen als Katalog mit Kennungen auf. Die Kennungen verknüpfen Anforderung, Umsetzung und Nachweis (Abschnitt 4.7). Die Priorisierung unterscheidet zwei Stufen. Muss-Anforderungen sind Bedingungen, ohne die das System seinen Zweck verfehlt. Soll-Anforderungen verbessern Betrieb oder Wartbarkeit, sind für die Zielerreichung aber nicht zwingend notwendig.

4.2 Funktionale Anforderungen

Die funktionalen Anforderungen leiten sich aus den in der Projektarbeit (Klemm, 2025, S. 12) identifizierten Limitationen und aus Gesprächen mit den Fachabteilungen ab.

Das System muss Such-Tags und Bildbeschreibungen in Deutsch und Englisch generieren. Die Tags dienen der internen Suche im DAM und müssen fachlich korrekt, suchrelevant und frei von Werbebegriffen sein. Marke und Linie werden explizit aus den Tags ausgeschlossen, da diese Informationen in Celum als eigenständige Informationsfelder gepflegt werden und über die reguläre Suche auffindbar sind. Die Beschreibungen sollen Produktinformationen in natürlicher

Sprache vermitteln und als Grundlage für Alternativtexte auf Websites und in Onlineauftritten dienen. Ihre Eignung ist im jeweiligen Verwendungskontext zu beurteilen.

Bei der Tag-Generierung muss das System visuellen Bildinhalt mit Produktkontext aus Centric PXM verbinden. Die PIM-Daten umfassen Produktnamen, Linienbezeichnungen, Marke und offizielle Oberflächennamen. Diese Daten fließen als Kontext in den Prompt ein, damit das Modell die unternehmensspezifischen Bezeichnungen korrekt verwendet. In die Tags selbst werden Oberflächen und Produkttypen übernommen, nicht jedoch Marke und Linie, da diese bereits in Celum hinterlegt sind. Eine Armatur in der Oberfläche Brushed Bronze lässt sich visuell nicht zuverlässig von Polished Gold Optic unterscheiden, weshalb die korrekte Übernahme aus dem PIM erforderlich ist.

Die erzeugten Tags werden normalisiert (Kleinschreibung, Deduplizierung, Längenbegrenzung) und über die Celum REST-API in die Metadatenfelder geschrieben. Eine Signatur mit Modell, Prompt-Version und Zeitstempel macht jede Generierung nachvollziehbar und unterstützt die Zuordnung zur verwendeten Konfiguration.

Die Asset-Auswahl muss über konfigurierbare Filterregeln gesteuert werden. Nicht alle 549.000 Assets sind relevant: Nur Bilddateien bestimmter Typen in definierten Collections sollen verarbeitet werden. Assets mit abgelaufenen Bildrechten oder bestimmten Status-Flags müssen ausgeschlossen werden.

Tabelle 2 fasst die funktionalen Anforderungen zusammen.

Tabelle 2: Katalog der funktionalen Anforderungen

Quelle: Eigene Darstellung

Kennung	Anforderung	Priorität
FA1	Such-Tags und Alternativtexte in Deutsch und Englisch erzeugen	Muss
FA2	Visuellen Bildinhalt mit Produktdaten aus dem PIM verbinden (Titel, Marke, Linie, Oberfläche)	Muss
FA3	Unternehmenseigenes Begriffsregelwerk durchsetzen: offizielle Oberflächennamen übernehmen, Marke, Linie und Werbebegriffe aus den Tags ausschließen	Muss
FA4	Tags normalisieren: Kleinschreibung, Stoppwortfilter, Synonymvereinheitlichung, Deduplizierung, Längenbegrenzung	Muss
FA5	Ergebnisse über die Celum REST-API in die vorgesehenen Informationsfelder übertragen	Muss
FA6	Jede Generierung mit Zeitstempel, Modell und Prompt-Version signieren	Soll
FA7	Zu verarbeitende Assets über konfigurierbare Filterregeln auswählen	Muss

4.3 Nichtfunktionale Anforderungen

Die modulare Architektur soll den Austausch der KI-Konfiguration erleichtern. Innerhalb der unterstützten Schnittstellen lassen sich Backend und Modell über die Konfiguration auswählen.

Zur Vermeidung mehrfacher Einplanung verwendet die Job-Queue einen eindeutigen Schlüssel aus Asset-ID, Version-ID und Prompt-Version. Ergänzend berücksichtigt die Discovery bereits vorhandene KI-Signaturen.

Fehlertoleranz stellt sicher, dass transiente Fehler (API-Timeouts, Rate Limits) nicht zum Abbruch der Gesamtverarbeitung führen. Stattdessen werden betroffene Assets nach einem Backoff-Schema erneut versucht. Permanente Fehler (keine Vorschau, fehlende Pflichtfelder) werden protokolliert und zur manuellen Prüfung markiert.

Die betriebliche Cloud-Vorgabe verlangt eine DSGVO-konforme Verarbeitung der Bilddaten innerhalb der Europäischen Union. Für AWS wird die Region eu-central-1 (Frankfurt) konfiguriert. Die Standortwahl ist ein Baustein der Datenschutzanforderungen.

Nachvollziehbarkeit verlangt, dass sich für jedes Asset rekonstruieren lässt, warum es verarbeitet oder übersprungen wurde. Jeder Ausschlussgrund wird deshalb kategorisiert protokolliert. Jede erzeugte Ausgabe trägt zusätzlich eine Signatur aus Zeitstempel, Modell und Prompt-Version.

Betriebssicherheit betrifft den Schreibzugriff auf das Produktivsystem. Celum ist das führende System für Bildmetadaten bei der Hansgrohe SE. Eine fehlerhafte Massenverarbeitung wäre aufwendig zu korrigieren. Das Rückschreiben ist deshalb an einen expliziten Schalter gebunden. Ohne ihn läuft das System im Dry-Run. Die Modellausgabe wird erzeugt und gespeichert, Celum bleibt aber unverändert.

Tabelle 3 fasst die nichtfunktionalen Anforderungen zusammen.

Tabelle 3: Katalog der nichtfunktionalen Anforderungen

Quelle: Eigene Darstellung

Kennung	Anforderung	Priorität
NFA1	Auswahl unterstützter Modelle und Backends über die Konfiguration	Soll
NFA2	Mehrfache Einplanung derselben Asset-, Versions- und Prompt-Kombination vermeiden	Muss
NFA3	Fehlertoleranz: transiente Fehler führen nicht zum Abbruch des Gesamtlaufs	Muss
NFA4	Nachvollziehbarkeit: Verarbeitungsschritte und Ausschlussgründe je Asset protokollieren	Soll
NFA5	Bildverarbeitung DSGVO-konform ausschließlich innerhalb der Europäischen Union	Muss
NFA6	Rückschreibung nach Celum nur nach expliziter Freigabe	Muss

4.4 Architekturentscheidungen

Vier Entscheidungen prägen die Architektur. Sie werden jeweils mit der verworfenen Alternative und der Begründung dokumentiert.

Erstens: PIM-Daten werden zur Laufzeit pro Asset abgerufen und als Kontextblock in den Prompt eingefügt. Alternativ wäre eine Vorab-Indexierung aller Produktdaten möglich gewesen. Die Laufzeitabfrage erleichtert die Übernahme geänderter Produktdaten; zwischengespeicherte Daten bleiben bis zum Ablauf ihrer

Gültigkeitsdauer nutzbar. Ein SQLite-basierter Cache mit konfigurierbarer TTL reduziert die Anzahl der PIM-API-Aufrufe.

Zweitens: Eine gemeinsame KI-Schnittstelle entkoppelt die Pipeline vom Backend. OpenAI-kompatible Endpunkte bedienen die lokalen Modelle und den produktiven Mantle-Zugang, eine OpenAI-kompatible Schnittstelle zu AWS Bedrock, für Claude über AWS Bedrock steht ein Backend mit Converse API bereit. Die Auswahl erfolgt über die YAML-Konfiguration.

Drittens: SQLite im WAL-Modus dient als Persistenzschicht für Job-Queue und Cache. Write-Ahead Logging ermöglicht parallele Lesezugriffe während eines schreibenden Zugriffs; Schreibzugriffe werden serialisiert. Das reicht für die anfallenden Volumina und erfordert keine externe Infrastruktur wie PostgreSQL oder Redis.

Viertens: Tags werden nicht als Freitext gespeichert, sondern als Referenzen auf Celum-Collections. Freitext wäre einfacher zu schreiben, erzeugt aber für jede Schreibweise eines Begriffs einen eigenen Suchwert. Collections bündeln gleiche Begriffe unter einer ID und machen sie im DAM filterbar. Der Preis ist ein zusätzlicher Auflösungsschritt: Für jeden Tag prüft das System, ob bereits eine Collection existiert, und legt andernfalls eine neue an. Abschnitt 7.1 berichtet die Zahl der gespeicherten Tag-Collection-Zuordnungen.

4.5 Modellauswahl

Die Modellauswahl folgt vier Dimensionen: lokale gegenüber cloudbasierter Ausführung, unterschiedliche Parametergrößen, unterschiedliche Modellfamilien und unterschiedliche Größenvarianten innerhalb einer Familie.

Gemma4:31b (Google DeepMind, 30,7 Milliarden Parameter, lokal via Ollama in 4-Bit-Quantisierung) bildet als größtes lokal ausgeführtes Modell den Ausgangspunkt des Vergleichs. Gemma4:e4b ist eine kleine vision-fähige Variante derselben Familie. Die Auswahl beantwortet die praktisch relevante Frage, ob ein auf den Betrieb auf Endgeräten ausgelegtes Modell für diese Aufgabe ausreicht. Das ist für Szenarien ohne Cloud-Anbindung und ohne dedizierte Hardware entscheidend. Qwen3.8-27B-MLX (Alibaba) repräsentiert eine alternative Modellfamilie

bei vergleichbarer Parameterzahl. Zum Zeitpunkt der Modellauswahl war Qwen3.8-27B erst zwei Tage verfügbar und wurde als aktuellstes offenes Modell dieser Größenklasse aufgenommen. Seither hat es sich in der Community als häufig bevorzugtes Modell für den lokalen Betrieb etabliert. Claude Sonnet 5 (Anthropic, Cloud via AWS Bedrock) dient als Referenz ohne lokale Hardwarebeschränkung. Quantisierung und Ausführungsform auf Serverseite sind nicht öffentlich spezifiziert. Die technischen Eigenschaften der vier Konfigurationen sind in Tabelle 1 (Abschnitt 2.2.5) dokumentiert.

Alle vier Konfigurationen erhalten dieselben Bilder, Prompt V3 und dieselben verfügbaren PIM-Daten. Der Vergleich bezieht sich auf diese jeweiligen Gesamtkonfigurationen.

4.6 Filterkriterien und Asset-Selektion

Der Eligibility-Service prüft die konfigurierten Auswahlregeln während der Discovery. Diese bestimmen, welche Assets in die Verarbeitung eingeplant werden.

Im ersten Schritt werden nur Assets aus definierten Collections berücksichtigt. Die Konfiguration umfasst Collections aus vier Bereichen: Product Assets (PIM-verknüpfte Produktbilder, Icons und Logos), Marketing (hansgrohe, AXOR, Social Media), CMS (Webseiten-Assets der verschiedenen Markenportale) und Public Relations (Events, Stock Images, Gebäude). Assets außerhalb dieser Collections werden nicht erfasst. Im zweiten Schritt erfolgt eine Filterung nach Asset-Typ. Celum unterscheidet verschiedene Typen (z. B. Produktbild, Illustration, Dokument), von denen nur sechs konfigurierte Typen für die Verschlagwortung relevant sind. Im dritten Schritt werden nur Bilddateien mit zulässigen Dateierweiterungen akzeptiert (JPG, PNG, TIFF, WebP, BMP, AVIF, HEIC). Nicht-Bild-Formate werden ausgeschlossen.

Im vierten Schritt werden Assets ausgeschlossen, bei denen bestimmte Metadatenfelder bereits belegt sind. Das betrifft Felder, die eine schon ausgeführte Verschlagwortung anzeigen, sowie Status-Flags, die eine Verarbeitung verhindern sollen (etwa ein Flag für gelöschte oder gesperrte Assets). Im fünften Schritt prüft

das System die Bildrechte zum Zeitpunkt der Discovery. Assets, deren Rechtefreigabe noch nicht begonnen hat oder bereits abgelaufen ist, werden übersprungen.

Durch diese Filterlogik reduziert sich der Bestand von 549.000 Assets auf 107.000 verarbeitungsrelevante Bilder. Die verbleibenden 442.000 Assets verteilen sich auf nicht-bildhafte Formate, archivierte Bestände, Assets ohne aktive Bildrechte und nicht relevante Collections.

4.7 Rückverfolgung der Anforderungen

Tabelle 4 verbindet die Anforderungen mit der beschriebenen Umsetzung und den zugehörigen Ergebnissen. Die qualitativen Bewertungen beziehen sich auf die deutschen Modellausgaben; Betriebskennzahlen beschreiben den technischen Ablauf.

Tabelle 4: Rückverfolgung der Anforderungen

Quelle: Eigene Darstellung

ID	Umsetzung	Bezug / Ergebnis
FA1	Abschnitt 5.2 (Stufe KI-Analyse), Abschnitt 5.3	Anhang E: Ausgaben aller vier Modelle in beiden Sprachen
FA2	Abschnitt 5.2 (Stufe PIM-Anreicherung), Abschnitte 5.3.2 und 5.3.3	Abschnitt 6.5, Anhang B: Kontextdaten der Stichprobe
FA3	Abschnitte 5.3.3 und 5.3.4	Abschnitt 6.4: fachliche Rückmeldungen zur Regelanwendung
FA4	Abschnitt 5.2 (Stufe Normalisierung)	Abschnitt 6.3: Anzahl erzeugter deutscher Tags
FA5	Abschnitt 5.2 (Write-Back)	Abschnitt 7.1: 103.568 als abgeschlossen gemeldete Jobs
FA6	Abschnitt 5.2 (Stufe Write-Back)	Abschnitt 5.2: Signatur beim Write-Back
FA7	Abschnitt 4.6, Abschnitt 5.2 (Stufe Eligibility)	Abschnitt 4.6: 107.000 von 549.000 Assets
NFA1	Abschnitte 4.4 und 5.1	Kapitel 6: vier Konfigurationen mit gemeinsamer Toolchain evaluiert
NFA2	Abschnitt 5.2 (Stufe Discovery)	Quellcode: eindeutiger Jobschlüssel
NFA3	Abschnitt 5.2 (Retry-Logik)	Abschnitt 7.1: 621 Jobs mit Wiederholung, 1.931 Ereignisse
NFA4	Abschnitte 5.1 und 5.2	Abschnitt 5.2: Status- und Ereignisprotokollierung
NFA5	Abschnitte 4.4 und 5.4	Abschnitt 5.4: konfigurierte Region eu-central-1
NFA6	Abschnitt 5.1 (Dry-Run als Standardverhalten)	Abschnitt 5.1: Dry-Run und Freigabe des Schreibzugriffs

5 IMPLEMENTIERUNG

Dieses Kapitel beschreibt die technische Umsetzung des Systems. Die Architektur wird von außen nach innen vorgestellt: Infrastruktur, Pipeline-Ablauf, Prompt-Entwicklung, Produktivbetrieb und Evaluations-Tooling.

5.1 Systemarchitektur und Infrastruktur

Das System ist als modulare Python-Applikation realisiert. Abbildung 2 zeigt seine Einbettung in die Systemlandschaft des Unternehmens. Eine zentrale YAML-Konfiguration steuert Backend-Auswahl, Collection-Zuordnung, Filterkriterien und Authentifizierung. Sensible Daten werden über Umgebungsvariablen injiziert. Die Konfiguration definiert unter anderem den Celum-API-Endpunkt und API-Key, PIM-Zugangsdaten, KI-Backend (Provider, Modell, Token-Limit), zu verarbeitende Collections, Filterregeln und Normalisierungsparameter.

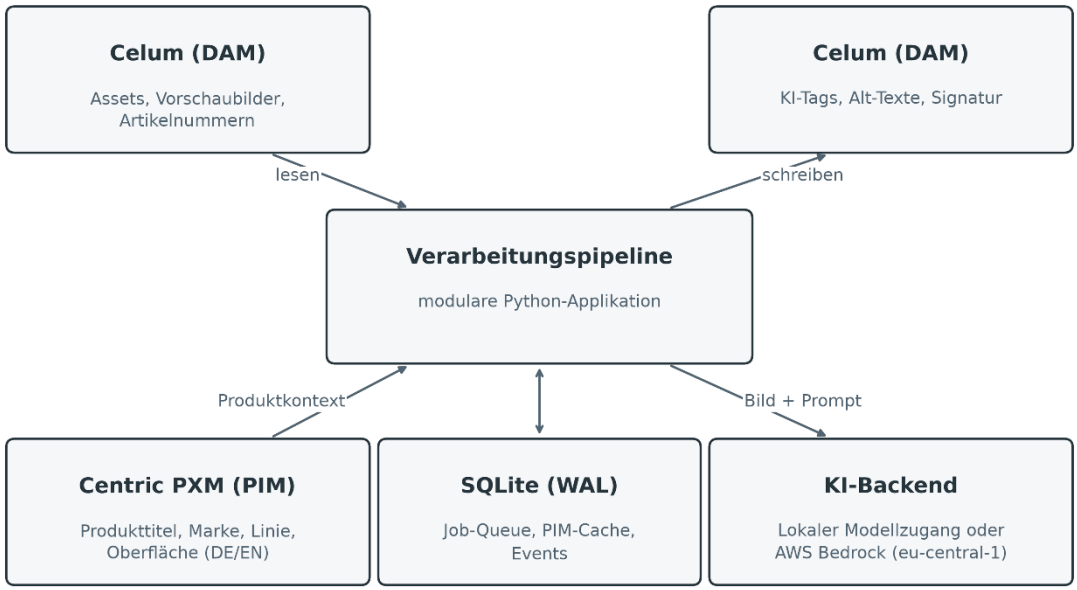


Abbildung 2: Einbettung des Systems in die Systemlandschaft der Hansgrohe SE

Quelle: Eigene Darstellung

Als Persistenzschicht dient SQLite im WAL-Modus. Die Datenbank umfasst vier Tabellen: *jobs* für die Verarbeitungs-Queue mit Status-Tracking und Retry-Logik, *pim_cache* für zwischengespeicherte Produktdaten mit konfigurierbarer TTL, *events* für die Protokollierung aller Verarbeitungsschritte und *tag_collections* für die Zuordnung generierter Tags zu Celum-Collection-IDs. WAL unterstützt gleichzeitige Lesezugriffe, aber nur einen gleichzeitigen Schreibzugriff.

Celum und Centric PXM werden über einen gemeinsamen HTTP-Client auf Basis von httpx angesprochen. Dieser führt selbst keine Wiederholungen aus, sondern klassifiziert Fehler als wiederholbar oder permanent. Die Wiederholung steuert die Job-Queue (Abschnitt 5.2). Celum nutzt einen X-API-KEY-Header, die REST-API von Centric PXM eine separate Authentifizierung. Im Quellcode trägt diese Anbindung noch die frühere Produktbezeichnung Contentserv (PXC).

Die KI-Schicht stellt zwei Backends bereit: ein OpenAI-kompatibles Chat-Completions-Backend und ein AWS-Converse-Backend. Beide erhalten Systemprompt, Aufgabenprompt und Bild und liefern Text sowie Nutzungsmetriken zurück. Lokale und cloudbasierte Modellzugänge werden damit hinter einer gemeinsamen Schnittstelle gekapselt.

Über ein CLI mit Subcommands stehen verschiedene Betriebsmodi bereit: Einzelverarbeitung zur Qualitätsprüfung (*test*), Discovery und vollständige Pipeline über alle Collections (*run*), ein Dry-Run-Modus ohne Rückschreiben sowie ein Report-Kommando (*report*) zur Übersicht über Jobstatus und Prüffälle.

Die Gemma-Modelle wurden lokal über Ollama in GGUF-Quantisierung ausgeführt. Eine zusätzlich erprobte MLX-Fassung von Gemma4:31b verarbeitete in diesem Setup keine Bildeingaben und wurde daher nicht in die Evaluation aufgenommen. Die ausgewertete Qwen-Konfiguration ist unter qwen3.8:27b-mlx protokolliert. Anhang C fasst die verfügbaren Laufzeitangaben zusammen.

5.2 Pipeline-Design

Abbildung 3 zeigt die acht fachlichen Verarbeitungsstufen. Discovery und Eligibility gehen der Verarbeitung der eingeplanten Jobs voraus.

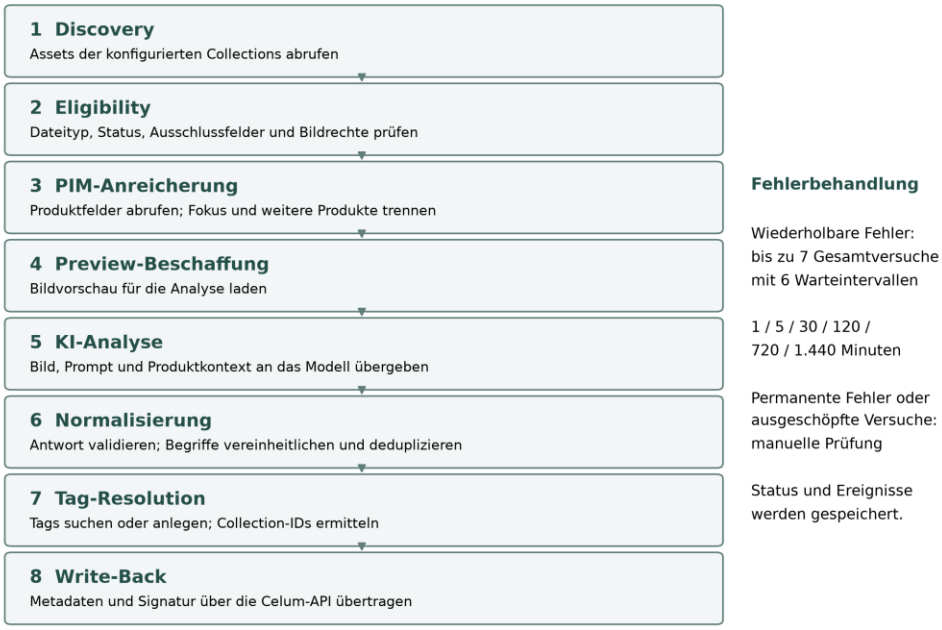


Abbildung 3: Verarbeitungsstufen der Pipeline

Quelle: Eigene Darstellung

Discovery ruft die Assets der konfigurierten Collections seitenweise ab (Limit Celum 200). Nach der Eligibility-Prüfung werden vorhandene KI-Signaturen berücksichtigt und neue Jobs gespeichert. Der Datenbankschlüssel aus Asset-ID, Version-ID und Prompt-Version verhindert eine doppelte Einplanung derselben Kombination.

Eligibility filtert anhand des konfigurierten Regelwerks: Dateierweiterung (nur Bildformate wie JPG, TIFF, PNG), Asset-Typ-ID, Ausschlussfelder (etwa „Nutzung untersagt“ oder Zeichnungstyp), Status-Prüfung gegen eine Ausschlussliste und Bildrechte-Validierung gegen Start- und Enddatum. Jeder Ausschlussgrund wird als kategorisierter Skip-Grund protokolliert, was die Nachvollziehbarkeit und das Debugging erleichtert. Die Prüfung erfolgt während der Discovery, anschließend wird der gespeicherte Asset-Stand verarbeitet.

PIM-Anreicherung extrahiert Artikelnummern aus zwei Celum-Feldern. Das sind zum einem das Fokus-Artikelfeld (das Hauptprodukt im Bild) und das andere Feld ist für alle sichtbaren Produkte. Die Felder können mehrere kommaseparierte Artikelnummern enthalten. Für jede Nummer werden über die Centric PXM-API zusätzliche Produktinformationen abgerufen, der Titel, die Marke, die Linie und die offizielle Oberflächenbezeichnung, auf Deutsch und Englisch. Die Ergebnisse werden in einem SQLite-Cache gespeichert, dessen TTL konfigurierbar ist. Fokus-Produkte werden im Prompt als solche markiert; weitere sichtbare Produkte erscheinen in einem separaten Abschnitt.

Für die Bildanalyse wird ein Celum-Vorschauderivat statt der Originaldatei verwendet. Die Pipeline fordert die Vorschau-Auflösung an und weicht bei Bedarf auf die nächstkleinere verfügbare Variante aus; die Übertragungsgröße ist konfigurierbar begrenzt. Dies reduziert das übertragene Datenvolumen. Die Tokenisierung und Verarbeitung der Bildauflösung hängen vom jeweiligen Backend ab.

Die KI-Analyse übergibt Bild, Systemprompt und versionierten Aufgabenprompt mit PIM-Kontext an das gewählte Backend. Die Antwort wird als JSON verarbeitet. Danach entfernt eine Extraktionsfunktion gegebenenfalls Markdown-Fences (```). Input- und Output-Tokens sowie die Dauer werden zusammen mit dem jeweiligen Jobergebnis gespeichert.

Normalisierung validiert und bereinigt die Ausgabe. Tags werden kleingeschrieben, gegen eine konfigurierbare Stopwort-Liste gefiltert, auf Maximallänge begrenzt (100 Zeichen pro Tag) und dedupliziert. Synonyme werden über eine Ersetzungstabelle vereinheitlicht. Beschreibungen werden auf eine Maximallänge von 500 Zeichen geprüft. Invalide Ausgaben (fehlendes JSON, fehlende Pflichtfelder, leere Tags) lösen eine ValidationError-Exception aus und werden zur manuellen Prüfung markiert. Fehler beim Einlesen der JSON-Antwort können dagegen einen erneuten Versuch auslösen.

Tag-Resolution übersetzt die generierten Tag-Strings in Celum-Collection-IDs. Der TagResolver sucht zunächst in der Celum-API nach einer existierenden Collection mit identischem Namen in der primären Sprache. Wird ein Treffer gefunden, wird die ID referenziert. Andernfalls wird eine neue Collection unter dem konfigurierten Root-Knoten angelegt. Ein Session-Cache verhindert wiederholte

API-Aufrufe für denselben Tag innerhalb eines Verarbeitungslaufs. Die ermittelten IDs werden anschließend den Assets zugeordnet.

Write-Back schreibt Tags, Beschreibungen und eine Signatur über die Celum-API zurück. Die Signatur enthält Zeitstempel, Modell, Systemversion und Prompt-Version. Tritt beim Schreibvorgang ein Fehler auf, wird dieser ebenfalls protokolliert.

Für wiederholbare Fehler sind höchstens sieben Gesamtversuche vorgesehen. Zwischen ihnen liegen Warteintervalle von 1, 5, 30, 120, 720 und 1.440 Minuten. Nach Ausschöpfung der Versuche sowie bei permanenten Fehlern erhält der Job den Status zur manuellen Prüfung.

5.3 Prompt-Entwicklung

Die Entwicklung folgt einem iterativen Verfeinerungszyklus. Vier Versionen markieren die wesentlichen Schritte. Abbildung 4 stellt die Kette aus beobachtetem Defizit, gezielter Prompt-Änderung und weiteren Beobachtungen im Überblick dar. Explorative Tests vor V1 mit allgemeinen Instruktionen wiesen keine systematische Struktur auf und dienten der Orientierung. Die Volltexte aller vier Versionen sowie der Systemprompt finden sich in Anhang A.

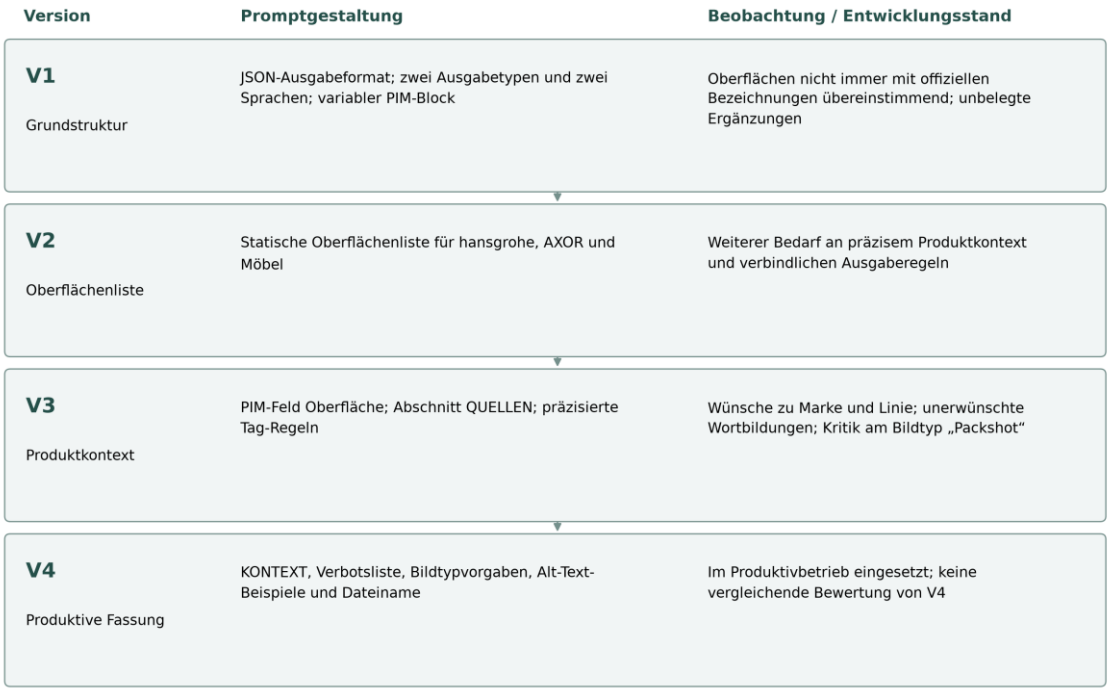


Abbildung 4: Iterative Prompt-Entwicklung von Version V1 bis V4

Quelle: Eigene Darstellung

5.3.1 Version V1: Grundstruktur

V1 definiert die grundlegenden Aufgaben. Es sollen Such-Tags für das DAM und Beschreibungen zur Nutzung als barrierefreie Alternativtexte, jeweils in Deutsch und Englisch, verfasst werden. Der Prompt spezifiziert das JSON-Format mit den Objekten *tags* und *alt_text*, jeweils mit deutschen und englischen Sprachschlüsseln und gibt Hinweise zu Stil und Inhalt. Ein variables Produkt-Kontextfeld wird

eingefügt, sofern PIM-Daten verfügbar sind. Die maximale Tag-Anzahl pro Sprache ist als Variable konfigurierbar.

Die manuelle Sichtung zeigte zwei Defizite. Erstens entsprachen Oberflächenangaben nicht immer den offiziellen Bezeichnungen: Eine Armatur in Brushed Bronze wurde als goldfarben beschrieben, eine in Brushed Black Chrome als mattschwarz. Da die PIM-Abfrage noch keine Oberflächeninformationen enthielt, fehlte hier der verlässliche Produktbezug. Zweitens wurden Angaben ergänzt, die weder im Bild noch in den verfügbaren PIM-Daten belegt waren. Die weitere Entwicklung präzisierte deshalb die Zuordnung von Informationen zu ihren Quellen.

5.3.2 Version V2: Oberflächenintegration

V2 adressiert das Oberflächenproblem mit einem statischen Ansatz. Da die PIM-Abfrage weiterhin keine Oberflächeninformationen enthielt, wird ein Abschnitt PRODUKTOBERFLÄCHEN mit Bezeichnungen für hansgrohe und AXOR in Deutsch und Englisch eingefügt. Für unklare Fälle nennt der Prompt sichtbare Farbwirkungen wie goldfarben oder silberfarben. Die allgemeinen Regeln binden offizielle Produkt- und Linienbezeichnungen an PIM-Daten.

Die manuellen Tests lieferten Hinweise auf passendere Oberflächenzuordnungen, ohne einen quantifizierten Versionsvergleich darzustellen. Allerdings ignorierten die Modelle den Kontext weiterhin selektiv, besonders bei Bildern mit mehreren sichtbaren Produkten. Zudem enthielten die erzeugten Tags regelmäßig Werbebegriffe wie hochwertig oder premium, die im DAM nicht als Suchbegriffe taugen.

5.3.3 Version V3: Quellendeklaration und Regelwerk

V3 markiert zwei parallele Veränderungen. Erstens wird die PIM-Abfrage um das Feld Oberfläche (Finish) erweitert, sodass verknüpfte Produkte ihre offizielle Oberflächenbezeichnung als Kontext liefern können. Die statische Liste aus V2 entfällt. Zweitens wird der Prompt umstrukturiert. Ein Abschnitt QUELLEN unterscheidet sichtbare Bildinformationen von Produktangaben aus dem PIM. Unsichere Merkmale sollen entfallen, offizielle Produkt-, Linien- und Oberflächenbezeichnungen sind ausschließlich aus dem PIM zu übernehmen. Damit werden

die bereits in V1 und V2 enthaltenen Aufgaben- und Quellenregeln gebündelt und präzisiert.

Der Tag-Abschnitt präzisiert das Regelwerk. Marken und Artikelnummern bleiben ausgeschlossen, Suchsynonyme wie Waschtischarmatur und Waschtischmischer sind erwünscht. Die Bildtypen Freisteller, Packshot, Lifestyle, Detailaufnahme und Ambiente werden aus V2 übernommen. Ergänzt werden unter anderem konkretere Vorgaben zu technischen Funktionen, Nutzungskontext und subjektiven Begriffen. Der Begriff Packshot erweist sich in der Evaluation als problematisch, da er der internen Terminologie des Unternehmens nicht entspricht, und wird in V4 entfernt (vgl. Abschnitt 6.4).

Der Alt-Text-Abschnitt definiert eine Reihenfolge (Produkt, Merkmale, Raum, Einbau, Handlung), verbietet den Beginn mit „Bild von“ oder „Foto von“ und fordert eigenständige Formulierungen pro Sprache statt wörtlicher Übersetzung.

V3 wird einheitlich für die vergleichende Evaluation verwendet. Dadurch bleiben Prompt und Eingabedaten zwischen den vier Konfigurationen konstant.

5.3.4 Version V4: Konsolidierung nach Expertenfeedback

V4 integriert die Erkenntnisse aus der Umfrage und aus Einzelgesprächen mit Fachexperten. Die Änderungen betreffen sechs Bereiche. Erstens wird ein Abschnitt KONTEXT eingefügt, der den Modellen den Verwendungszweck der zu erzeugenden Ausgaben erklärt. Tags werden von Fachleuten zur Bildsuche genutzt und Alt-Texte erscheinen auf hansgrohe.com und bei Handelspartnern.

Zweitens wird die Tag-Strategie verbessert. Fast alle Bilder zeigen Sanitärprodukte, deshalb helfen generische Begriffe wie Armatur oder Wasserhahn allein nicht. Der Schwerpunkt liegt auf dem, was dieses konkrete Bild von anderen unterscheidet. Für Oberflächen wird der PIM-Name der jeweiligen Sprache verwendet. Gängige Bezeichnungen werden rechtschreibkonform eingepasst, Eigennamen exakt übernommen. Technische Funktionen wie CoolStart, PowderRain oder EcoSmart dürfen nur bei PIM-Bestätigung getaggt werden.

Drittens enthält V4 eine explizite Verbotsliste für fachlich unerwünschte Wortbildungen und Formulierungen. Formulierungen wie Waschbeckenunterschale,

Marmorkonterplatte, Kopfbrausenkopf, Duschpipe oder schwebender Waschtisch werden als Negativbeispiele aufgeführt. Diese Liste entstand direkt aus den Freitextkommentaren der Evaluation.

Viertens wird die Bildtyp-Klassifikation erweitert und präzisiert. Vorgeschlagen werden Freisteller, eCom-Bild, Produktrendering, Iso-Rendering, Detailaufnahme, Lifestyle, Teilambiente mit Person und Ambiente ohne Person.

Fünftens enthält der Alt-Text-Abschnitt konkrete Beispielformulierungen, die den erwarteten Stil und die Informationsdichte veranschaulichen. Marke und Linie müssen bei erkennbarem Hauptprodukt und vorliegenden PIM-Daten vorkommen (hansgrohe immer klein, AXOR immer groß). Die Beispiele stammen aus realen Celum-Assets und wurden mit den Fachexperten abgestimmt. Sechstens wird der Dateiname des Assets in den Prompt aufgenommen. Bei Marketing-Assets enthält der Dateiname häufig Kampagneninformationen, Produktbezeichnungen oder Kontexthinweise, die für die Tag-Generierung und Beschreibung nützlich sein können.

5.4 Produktivbetrieb und Automatisierung

Die in diesem Abschnitt beschriebene Konfiguration greift dem Ergebnis der Evaluation vor. Ihre Begründung liefert Abschnitt 6.7. Nach Abschluss der Evaluation wurde das System mit Gemma4:31b und Prompt V4 in Betrieb genommen. Die Inferenz erfolgt über AWS Bedrock per Mantle API in der Region eu-central-1 (Frankfurt). Die Region wurde entsprechend der betrieblichen Vorgabe für die datenschutzkonforme Verarbeitung innerhalb der Europäischen Union gewählt.

Gegen eine lokale Verarbeitung wurde aus zwei Gründen entschieden. Erstens beträgt die durchschnittliche Inferenzdauer auf dem Mac Studio M2 Ultra 61,09 Sekunden pro Bild (Abschnitt 6.3). Die sequenzielle Verarbeitung des relevanten Bestands würde damit rund 76 Tage ununterbrochene Rechenzeit beanspruchen und die Hardware über diesen Zeitraum binden. Über AWS Bedrock lässt sich die Verarbeitung parallelisieren. Zweitens verursacht die Cloud-Ausführung unter den gewählten Betriebsannahmen nur geringe tokenbasierte Inferenzkosten und bindet keine lokale Hardware. Die Kosten liegen in derselben Größenordnung

wie die Energiekosten einer lokalen Verarbeitung und sprechen daher nicht gegen die Cloud (Abschnitt 6.6).

Zu beachten ist, dass evaluiertes und eingesetztes Modellartefakt nicht identisch sind. Die Evaluation erfolgte mit der lokal über Ollama ausgeführten, 4-Bit-quantisierten Variante von Gemma4:31b, der Produktivbetrieb nutzt dasselbe Modell über AWS Bedrock. Die serverseitige Ausführungsform ist nicht öffentlich spezifiziert und kann von der lokalen abweichen. Abschnitt 7.3 behandelt die daraus folgende Einschränkung. Außerdem wurde der Prompt von V3 auf V4 weiterentwickelt; die Präferenzbewertung bezieht sich auf V3.

Für neu hochgeladene Assets in relevanten Collections wurde die automatische Auslösung der Verarbeitung über Webhooks und AWS Lambda umgesetzt.

5.5 Evaluations-Tooling

Der Testmodus (test-Subcommand) verarbeitet einzelne Assets und erzeugt JSON-Dateien mit Modellausgaben, Eingabedaten und Leistungsmetriken. Für die Evaluation wurden die 30 Bilder des Datensatzes mit allen vier Modellen verarbeitet, sodass für jedes Bild vier Modellausgaben in insgesamt vier Ergebnisdateien vorliegen.

Ein Skript (build_survey.py) generiert aus den vier Ergebnisdateien eine offlinefähige HTML-Umfrage. Pro Bild werden die Ausgaben anonymisiert und randomisiert als Variante A bis D untereinander dargestellt. Die Randomisierung erfolgt pro Bild und Sitzung; die Modellnamen bleiben für die Bewertenden ausgeblendet. Die exportierte Variantenzuordnung ermöglicht die spätere Auswertung. Die Umfrage läuft vollständig im Browser der Teilnehmenden, ohne Server-Abhängigkeit. Ergebnisse werden als JSON-Datei exportiert.

Das Skript analyze_results.py aggregiert die exportierten Antworten und erzeugt Kennzahlen, CSV-Dateien und einen Markdown-Bericht. Das Skript statistik.py ergänzt statistische Berechnungen. Die Interpretation berücksichtigt die wiederholten Urteile derselben Personen zu denselben Bildern (Abschnitt 3.3.3). Die unveränderten Skripte und Rohdaten sind der Arbeit beigefügt.

6 EVALUATION

Dieses Kapitel dokumentiert Aufbau, Durchführung und Ergebnisse der vergleichenden Modellevaluation. Auf die Beschreibung von Datensatz, Umfrage-Design und Auswahl der Teilnehmenden folgen die Präferenzen der Fachexperten und die objektiven Modellmetriken. Die qualitative Auswertung der Freitextkommentare erklärt anschließend einen Teil der Präferenzverteilung. Danach wird die Rolle des Produktkontexts geprüft und der Aufwand für den Produktivbetrieb hochgerechnet. Der letzte Abschnitt fasst zusammen und begründet die Modellempfehlung. Die statistischen Verfahren sind in Abschnitt 3.3.3 beschrieben.

6.1 Evaluationsaufbau

6.1.1 Evaluationsdatensatz

Der Datensatz umfasst 30 gezielt ausgewählte Bilder aus dem produktiven Celum-Bestand der Hansgrohe SE. Freisteller, Ambiente-Aufnahmen, Lifestyle-Bilder, Marketingmotive und Detailaufnahmen decken unterschiedliche Nutzungsszenarien ab. Die in Anhang B dokumentierte Kategorienzuordnung beschreibt die Motivbreite. Eine repräsentative Zufallsstichprobe des Gesamtbestands wird nicht beansprucht.

Alle vier Modelle verarbeiten dieselben Bilder mit Prompt V3 und identischen, einmalig abgerufenen PIM-Kontextdaten. Verglichen werden damit Modellkonfigurationen unter gemeinsamen Eingabebedingungen. Laufzeitumgebung und Modellarchitektur unterscheiden sich.

Der Datensatz enthält bewusst beide Kontextszenarien. Für 22 der 30 Bilder liefert das PIM Produktdaten. Bei zehn dieser Bilder ist zusätzlich ein Fokusprodukt hinterlegt, das das Hauptmotiv bezeichnet. Acht Bilder sind nicht mit Artikelnummern verknüpft und werden daher rein visuell verarbeitet. Abschnitt 6.5 wertet diese Aufteilung aus. Anhang B listet alle 30 Bilder mit Dateiname, Bildkategorie, Kontextstatus und Anzahl verknüpfter Produkte auf.

6.1.2 Umfrage-Design

Pro Bild sehen die Teilnehmenden vier anonymisierte Varianten A bis D in randomisierter Reihenfolge. Für Tags und Beschreibungen wählen sie jeweils die bevorzugte deutsche Ausgabe. Zusätzlich steht die Antwort „Keine Variante ist geeignet bzw. die Varianten sind nicht sinnvoll unterscheidbar.“ zur Verfügung. Diese Option erfasst sowohl fehlende Eignung als auch fehlende Unterscheidbarkeit und wird im Folgenden als „keine eindeutige Auswahl“ bezeichnet. Ein optionales Kommentarfeld ergänzt die Präferenzentscheidung.

Aufgrund der sprachlichen Fähigkeiten der Bewertenden wurden ausschließlich deutsche Beschreibungen und Tags vorgelegt. Abbildung 5 zeigt die Bewertungsoberfläche.

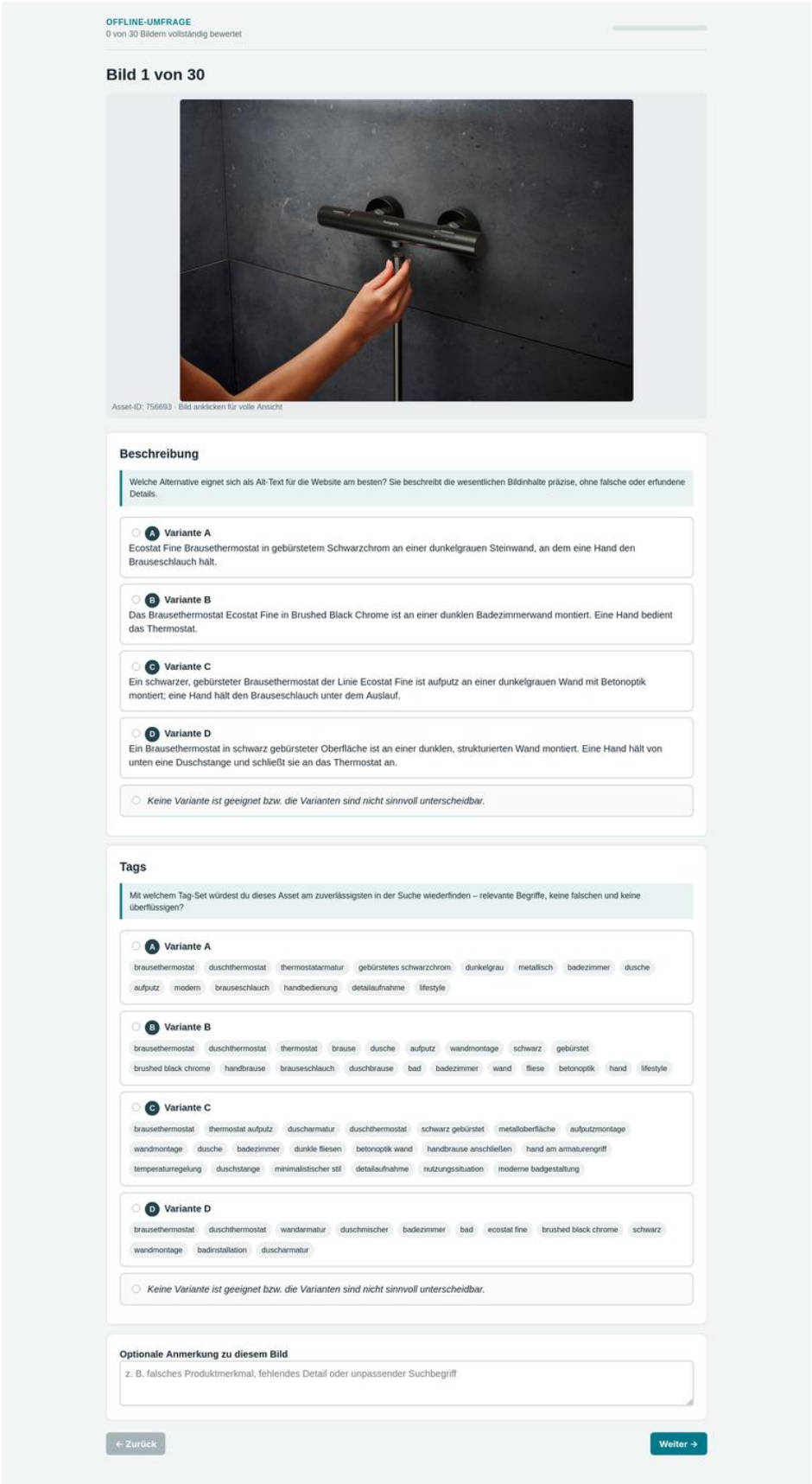


Abbildung 5: Bewertungsoberfläche der Umfrage

Quelle: Eigene Darstellung

Die Bearbeitungsdauer wurde auf zwei Ebenen erfasst. Der Median der Gesamtsitzungsdauer liegt bei 104,55 Minuten pro Person. Dieser Wert schließt Unterbrechungen ein, da die Umfrage über längere Zeit im Browser geöffnet bleiben konnte. Robuster ist der Median der personenbezogenen Median-Zeiten pro Bild: Er beträgt 71,25 Sekunden. Auf 30 Bilder hochgerechnet ergibt das eine reine Bearbeitungszeit von rund 36 Minuten. Die Gesamtsitzungsdauern streuen dagegen stark (45,2 bis 2.692,1 Minuten). Vier der zwölf Teilnehmenden hielten die Umfrage länger als acht Stunden geöffnet. Alle Teilnehmenden bewerteten die 30 Bilder vollständig. Ob die aufgewendete Zeit auch der Sorgfalt der Bewertung entspricht, lässt sich aus diesen Werten nicht ableiten.

6.1.3 Teilnehmerauswahl

An der Evaluation nahmen zwölf Fachexperten der Hansgrohe SE teil. Die Auswahl folgte dem Prinzip des Purposive Sampling. Rekrutiert wurden ausschließlich Personen, die im Arbeitsalltag regelmäßig mit Celum-Assets arbeiten und die Anforderungen an Bildmetadaten aus erster Hand kennen. Die fachliche Kompetenz der Teilnehmenden ist für eine präferenzbasierte Evaluation entscheidend, weil die Qualität von Tags und Beschreibungen nur von Personen beurteilt werden kann, die das Suchverhalten und die Nutzungskontexte im DAM kennen.

Die Gruppe umfasst fünf Personen aus dem Marketing, drei aus der Websitepflege, zwei aus Produktdatenmanagement und Celum-Administration sowie jeweils eine aus internationaler Lokalisierung und dem Bereich KI und neue Technologien (Tabelle 5). Die internationale Tätigkeit erweitert die Nutzungsperspektive, stellt aber keine Bewertung englischer Ausgaben dar.

Tabelle 5: Teilnehmende der Evaluation nach Funktionsbereich (n = 12)

Quelle: Eigene Darstellung

Kürzel	Funktionsbereich	Fachlicher Schwerpunkt
P1	Marketing	Creative Strategy und Public Relations
P2	Marketing	Unternehmenskommunikation
P3	Marketing	Content Creation
P4	Marketing	Digital Marketing
P5	Marketing	Suchmaschinenoptimierung
P6	Websitepflege	Web B2C
P7	Websitepflege	Website-Redaktion
P8	Websitepflege	Website-Daten
P9	Produktdaten und DAM	Celum
P10	Produktdaten und DAM	Produktdatenmanagement
P11	Internationale Lokalisierung	Country Adaption
P12	KI und neue Technologien	KI-Implementierung

Die Funktionsbereiche bringen unterschiedliche Anforderungen an Suchrelevanz, Beschreibung und Produktidentifikation ein. Die Befragung erfasst deren fachliche Präferenzen, prüft aber weder WCAG-Konformität im konkreten Webseitenkontext noch tatsächliche Verbesserungen der Auffindbarkeit.

Die Erhebung erfolgt pseudonymisiert und nach Information sowie Zustimmung der Teilnehmenden. Die Zuordnung der Kürzel zu Personen ist nicht Bestandteil der Arbeit.

6.2 Ergebnisse der Präferenzerhebung

Es liegen 360 Urteile je Dimension vor. Davon entfallen 301 bei Beschreibungen und 297 bei Tags auf eine konkrete Modellwahl. Die bedingte Präferenzquote bezieht sich auf diese Modellwahlen; der Anteil an allen Urteilen verwendet jeweils 360 als Nenner. Auch die fünfte Antwortoption bleibt ein gültiges Untersuchungsergebnis.

Bei den Beschreibungen erhält Gemma4:31b mit 113 Stimmen und 37,54 Prozent der konkreten Modellwahlen den höchsten Zuspruch. Claude Sonnet 5 folgt mit 89 Stimmen (29,57 Prozent), Qwen3.8-27B-MLX mit 84 (27,91 Prozent) und Gemma4:e4b mit 15 (4,98 Prozent). Tabelle 6 stellt ergänzend die Anteile an allen 360 Urteilen dar.

Tabelle 6: Präferenzentscheidungen für Bildbeschreibungen (Deutsch)

Quelle: Eigene Berechnung auf Basis der Erhebungsdaten

Modell / Antwort	Stimmen	Anteil an allen 360	Anteil an Modellwahlen
Gemma4:31b	113	31,39 %	37,54 %
Claude Sonnet 5	89	24,72 %	29,57 %
Qwen3.8-27B-MLX	84	23,33 %	27,91 %
Gemma4:e4b	15	4,17 %	4,98 %
Fünfte Option	59	16,39 %	—

Der Abstand von Gemma4:31b zu Claude Sonnet 5 und Qwen3.8-27B-MLX ist als beobachtete Präferenz zu lesen. Gemma4:e4b erhält erheblich weniger Stimmen. Eine allgemeine statistische Rangordnung wird daraus nicht abgeleitet, die methodische Einordnung folgt am Ende dieses Abschnitts.

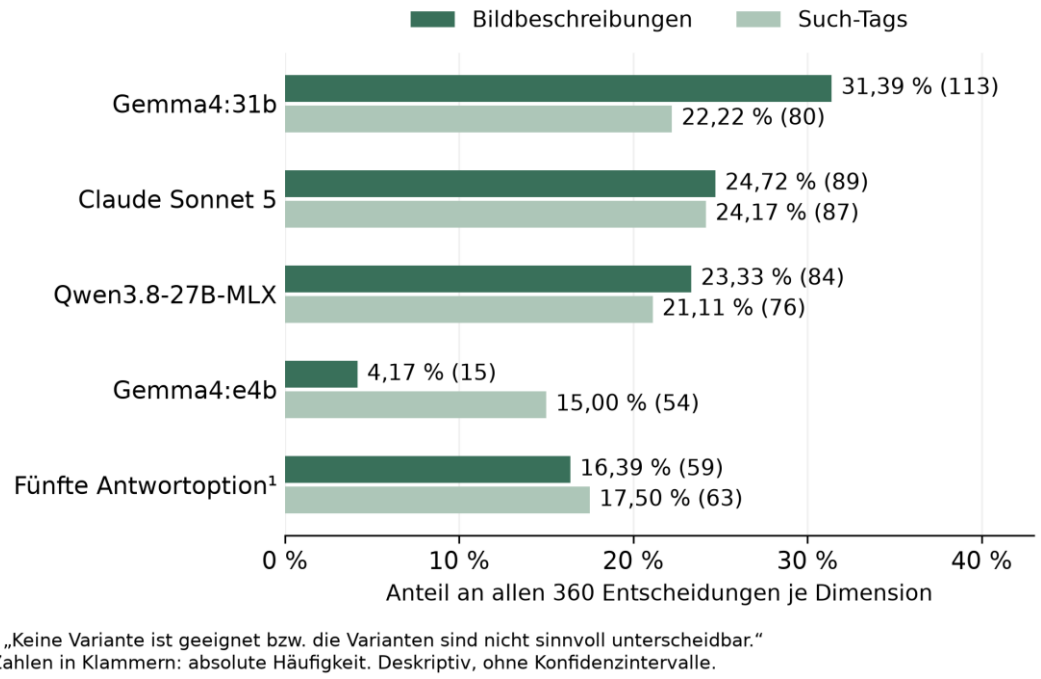
Bei den Tags liegen die Stimmenzahlen enger zusammen (Tabelle 7). Claude Sonnet 5 führt mit 87 Stimmen beziehungsweise 29,29 Prozent der konkreten Modellwahlen vor Gemma4:31b mit 80 (26,94 Prozent) und Qwen3.8-27B-MLX mit 76 (25,59 Prozent). Gemma4:e4b erreicht 54 Stimmen (18,18 Prozent).

Tabelle 7: Präferenzentscheidungen für Such-Tags (Deutsch)

Quelle: Eigene Berechnung auf Basis der Erhebungsdaten

Modell / Antwort	Stimmen	Anteil an allen 360	Anteil an Modellwahlen
Claude Sonnet 5	87	24,17 %	29,29 %
Gemma4:31b	80	22,22 %	26,94 %
Qwen3.8-27B-MLX	76	21,11 %	25,59 %
Gemma4:e4b	54	15,00 %	18,18 %
Fünfte Option	63	17,50 %	—

Die Option „keine eindeutige Auswahl“ wird 59-mal bei Beschreibungen (16,39 Prozent aller Urteile) und 63-mal bei Tags (17,50 Prozent) gewählt. Daraus lässt sich keine Ablehnungsquote ableiten, da auch als gleichwertig empfundene Varianten so beantwortet werden konnten. Abbildung 6 vergleicht die Präferenzverteilung beider Dimensionen.



¹ „Keine Variante ist geeignet bzw. die Varianten sind nicht sinnvoll unterscheidbar."
Zahlen in Klammern: absolute Häufigkeit. Deskriptiv, ohne Konfidenzintervalle.

Abbildung 6: Verteilung der Präferenzentscheidungen über alle fünf Antwortkategorien (je Dimension n = 360)

Quelle: Eigene Berechnung auf Basis der Erhebungsdaten

Die Auswertung ist deskriptiv. Dieselben zwölf Personen beurteilen dieselben 30 Bilder, sodass die 360 Urteile nicht unabhängig sind. Einfache Tests und Konfidenzintervalle auf Ebene aller Einzelurteile würden diese Struktur nicht berücksichtigen. Die beobachtete Rangfolge dient daher der betrieblichen Entscheidung, nicht dem Nachweis einer allgemeinen Überlegenheit oder Gleichwertigkeit.

Unter dieser Einschränkung lassen sich die mit den Auswertungsskripten berechneten Standardtests als ergänzende Einordnung lesen. Für die Beschreibungen weist Cochrans Q einen Gesamtunterschied zwischen den vier Varianten aus ($Q = 70{,}71$; $df = 3$; $p < 0{,}001$). In den paarweisen McNemar-Tests bleiben nach Bonferroni-Korrektur nur die drei Vergleiche gegen Gemma4:e4b signifikant ($p < 0{,}001$). Die Vergleiche zwischen den übrigen Konfigurationen sind es nicht (korrigiert: Gemma4:31b–Claude Sonnet 5 $p = 0{,}632$; Gemma4:31b–Qwen3.8-27B-MLX $p = 0{,}275$; Claude Sonnet 5–Qwen3.8-27B-MLX $p = 1{,}000$). Bei den Tags fällt der Gesamtunterschied schwächer aus ($Q = 8{,}20$; $df = 3$; $p = 0{,}042$); nach Korrektur bleibt allein Claude Sonnet 5 gegen Gemma4:e4b signifikant ($p =$

0,041). Belastbar ist damit allein der Rückstand von Gemma4:e4b bei den Beschreibungen und zwischen den drei führenden Konfigurationen stützen die Daten keine Rangordnung.

Fleiss' Kappa beträgt gerundet 0,03 für Beschreibungen und 0,04 für Tags. Werte zwischen 0,00 und 0,20 gelten als geringe Übereinstimmung (Landis & Koch, 1977, S. 165). Fleiss (1971, S. 378) entwickelt das Maß allerdings für wechselnde Gruppen von Bewertern, während hier dieselbe Gruppe alle Bilder beurteilt; der Wert ist deshalb als Näherung zu lesen. Auch die Nutzung der fünften Option variiert stark: zwischen 0 und 70 Prozent bei Beschreibungen (Median 8,33 Prozent) sowie zwischen 0 und 90 Prozent bei Tags (Median 5,00 Prozent). Eine einzelne Person verantwortet dabei 21 der 59 Nennungen bei den Beschreibungen und 27 der 63 bei den Tags; ohne sie sinken die Anteile auf 11,52 beziehungsweise 10,91 Prozent. Wegen des mehrdeutigen Wortlauts der Option lässt sich dies nicht allein auf strengere Qualitätsansprüche zurückführen. Die Rangfolge bündelt unterschiedliche Einzelpräferenzen.

Gemma4:e4b erhält bei Tags deutlich mehr Zuspruch als bei zusammenhängenden Beschreibungen. Eine mögliche Erklärung ist der unterschiedliche Anspruch an Objekterkennung, Sprachkohärenz und Fachterminologie. Die Präferenzdaten isolieren die Wirkung einzelner Architekturmerkmale jedoch nicht.

6.3 Objektive Modellmetriken

In den 30 gespeicherten Ergebnissen je Modell findet sich keine technisch ungültige Ausgabe. Bei dieser Stichprobengröße folgt daraus lediglich, dass die Quote technisch ungültiger Ausgaben mit 95 Prozent Konfidenz unter rund 11 Prozent liegt; das Wilson-Intervall für null Fehler bei 30 Beobachtungen reicht von 0 bis 11,35 Prozent. Auf den relevanten Bestand hochgerechnet entspräche die obere Schranke bis zu rund 12.000 Fehlschlägen, weshalb die Wiederholungslogik aus Abschnitt 5.2 für den Produktivbetrieb erforderlich bleibt. Der Befund beschreibt die untersuchten Bilder und ist weder eine inhaltliche Qualitätsprüfung noch eine Fehlergarantie für den Gesamtbestand. Die im Gesamtlauf beobachtete Fehleranzahl dokumentiert Abschnitt 7.1. Tabelle 8 ergänzt die Präferenzen um Laufzeit, Token-Verbrauch und Umfang der deutschen Ausgaben.

Tabelle 8: Kennzahlen der Modellkonfigurationen (je 30 Bilder)

Quelle: Eigene Berechnung auf Basis der Erhebungsdaten

Modell	Dauer Ø (s)	Dauer Median (s)	Input-Tokens	Output-Tokens	Tokens gesamt	Tags Deutsch	Alt-Text DE (Zeichen)
Gemma4:31b	61,09	56,55	1.187,8	1.197,1	2.384,9	15,20	165,1
Claude Sonnet 5	11,16	10,48	2.694,9	712,0	3.406,8	18,97	193,8
Qwen3.8-27B-MLX	34,50	34,45	1.711,0	1.105,6	2.816,5	19,77	191,0
Gemma4:e4b	17,20	16,95	1.187,8	1.326,8	2.514,5	16,73	181,8

Claude Sonnet 5 benötigt im Mittel 11,16 Sekunden je Bild, Gemma4:e4b 17,20 Sekunden, Qwen3.8-27B-MLX 34,50 Sekunden und Gemma4:31b 61,09 Sekunden. Da Cloud- und lokale Ausführung verglichen werden, beschreiben diese Werte die eingesetzten Konfigurationen und keinen isolierten Effekt der Modellgröße.

Claude Sonnet 5 weist mit durchschnittlich 3.406,8 die meisten Gesamttokens und mit 712,0 die wenigsten Output-Tokens auf. Die lokalen Konfigurationen liegen zwischen 2.384,9 und 2.816,5 Gesamttokens. Unterschiedliche Tokenisierung begrenzt die direkte Vergleichbarkeit der Token-Zahlen als Maß für Textumfang oder Rechenaufwand.

Gemma4:31b erzeugt durchschnittlich 15,20 deutsche Tags und Alt-Texte mit 165,1 Zeichen; die übrigen Modelle liegen bei 16,73 bis 19,77 Tags und 181,8 bis 193,8 Zeichen. Einzelne Kommentare bevorzugen kürzere Formulierungen. Ein allgemeiner Zusammenhang zwischen Kürze und Qualität wurde nicht gesondert gemessen.

6.4 Qualitative Analyse der Expertenkommentare

Acht der zwölf Teilnehmenden verfassten 77 freiwillige Kommentare. Sieben Kommentare nennen die fehlende Beurteilbarkeit einer Marke, ein weiterer enthält nur eine unbegründete Präferenz. Die verbleibenden 69 fachlichen Kommentare benennen konkrete Beobachtungen und Änderungswünsche. Sie ergänzen die Wahlentscheidungen und bildeten einen Ausgangspunkt für Prompt V4.

Das dazu entwickelte Kategoriensystem umfasst acht Themen. Die Zuordnung wurde KI-gestützt anhand des Wortlauts, der Ausgabedimension und der geltenden Anforderungen geprüft; dabei werden vermisste Angaben, unzutreffende Ergänzungen und Gestaltungswünsche unterschieden. Ein Kommentar zählt innerhalb einer Kategorie einmal, kann aber mehreren Kategorien angehören. So entstehen 109 Kommentar-Kategoriepaare. Eine unabhängige menschliche Zweitkodierung liegt nicht vor. Anhang D enthält alle Originalkommentare mit ihrer Zuordnung. Die dort vorangestellte Legende dokumentiert die Kodiersystematik.

Tabelle 9: Themen der fachlichen Freitextkommentare (n = 69; Mehrfachzuordnung möglich)

Quelle: Eigene Berechnung auf Basis der Erhebungsdaten

Thema	**Anzahl**	**Anteil an 69**	**mit PIM**	**ohne PIM**
Marke, Linie oder eindeutige Produktidentifikation	29	42,0 %	25	4
Bildinhalt oder bildbezogene Suchbegriffe	24	34,8 %	19	5
Terminologie oder Begriffspräferenz	15	21,7 %	14	1
Oberfläche oder Farbe	11	15,9 %	8	3
Bildtyp	10	14,5 %	10	0
Funktionales Merkmal	8	11,6 %	6	2
Alt-Text-Länge oder Formulierung	7	10,1 %	4	3
Als ungewöhnlich beanstandete Wortbildung	5	7,2 %	4	1

Die Anteile beziehen sich auf 69 fachliche Kommentare und summieren sich wegen Mehrfachkodierung nicht auf 100 Prozent. Tabelle 9 und Abbildung 7 zeigen Themenhäufigkeiten, keine allgemeinen Fehlerquoten.



Abbildung 7: Themen der fachlichen Freitextkommentare

Quelle: Eigene Berechnung auf Basis der Erhebungsdaten

Am häufigsten betreffen Kommentare Marke, Linie oder konkrete Produktidentifikation (29; 42,0 Prozent). Elf verlangen diese Angaben ausdrücklich in Tags, obwohl FA3 sie dort ausschließt, vier beziehen sich ausdrücklich auf den Alt-Text, darunter zwei zugleich auf Tags. Bei 16 Kommentaren bleibt die Dimension unklar. Der Befund zeigt neben Verbesserungswünschen einen Abstimmungsbedarf zwischen vorgegebener Tag-Regel und fachlichen Erwartungen.

Die zweithäufigste Kategorie betrifft Bildinhalt und bildbezogene Suchbegriffe (24 Nennungen, 34,8 Prozent). Dazu zählen berichtete falsche Zuordnungen wie ein nicht vorhandenes Mikrofon oder eine vermeintlich mit Schaum gefüllte Badewanne, ebenso ausgelassene bildbestimmende Elemente und fehlende gewünschte Suchbegriffe. Die Kategorie umfasst damit unterschiedliche Anliegen zur visuellen Erfassung und zur Auswahl relevanter Inhalte.

Fachterminologie bildet die dritte Kategorie (15 Nennungen, 21,7 Prozent). Beanstandet wurden technische Begriffe: Die Modelle verwendeten etwa Küchenmischer, Brausethermostat und Blende, gewünscht sind Küchenarmatur, Duschthermostat und Rosette (Anhang D, Nr. 10, 11, 16). Küchenmischer und

Brausethermostat stammen dabei aus den Produkttiteln des PIM. Der Produktkontext kann unerwünschte Fachbegriffe also selbst in die Ausgabe tragen. Eine Fachperson merkt an, dass Küchenmischer in der englischen Übersetzung zu kitchen mixer werden könne. Diese Rückmeldung betrifft eine mögliche Übersetzungsfolge, auch wenn englische Ausgaben in der Befragung nicht bewertet wurden. Die terminologischen Feinheiten sind für die Fachcommunity und das Suchverhalten im DAM relevant.

Elf Kommentare (15,9 Prozent) betreffen Oberflächen- und Farbbezeichnungen. Beanstandet werden die Ersetzung offizieller Namen durch visuelle Interpretationen sowie deren Übersetzung. Brushed Black Chrome soll beispielsweise nicht pauschal als mattschwarz bezeichnet werden. Die Hinweise zeigen den Bedarf einer mit den Produktdaten abgestimmten Terminologie.

Zehn Kommentare (14,5 Prozent) beziehen sich auf Bildtypbezeichnungen. Mehrere Fachpersonen lehnen den in Prompt V3 zugelassenen Begriff Packshot zugunsten betrieblicher Begriffe wie Freisteller, eCom-Bild oder Produktrendering ab. Dies ist ein Hinweis auf fachlichen Abstimmungsbedarf, nicht automatisch eine Verletzung der verwendeten Instruktion. Alle zehn Nennungen betreffen Bilder mit PIM-Kontext, ein Bildtypfeld wurde jedoch nicht übertragen.

Acht Kommentare (11,6 Prozent) behandeln funktionale Merkmale: Sechs vermissen eine Angabe, zwei beanstanden eine unzutreffende. Nicht jedes verlangte Merkmal ist im tatsächlich übertragenen Produktausschnitt enthalten und zwei Auslassungskommentare betreffen ein Bild ohne Produktkontext. Die Nennungen belegen daher keine einheitliche Gruppe ignorierter PIM-Werte.

Sieben Kommentare (10,1 Prozent) betreffen Länge und Formulierung der Alternativtexte. Eine Person bemängelt wiederholt zu lange Alt-Texte. Weitere Hinweise richten sich gegen Formulierungen, die eine unerwünschte Lesart nahelegen.

Fünf Kommentare (7,2 Prozent) beanstanden ungewöhnliche Wortbildungen, etwa Waschbeckenunterschale, Marmorkonterplatte oder Duschpipe. Die Einordnung folgt dem fachlichen Urteil im Kommentar und behauptet nicht, dass die Wörter grundsätzlich nicht existieren. Prompt V4 greift entsprechende Beispiele in einer Negativliste auf. Deren Wirkung wurde nicht gesondert bewertet.

Die Kommentare werden über alle vier Modellvarianten hinweg ausgewertet. Eine vollständige modellbezogene Zuordnung wurde nicht vorgenommen. Wegen der freiwilligen und selektiven Kommentierung bedeuten weder fehlende Kommentare bestätigte Fehlerfreiheit noch häufige Themen eine gemessene Fehlerquote.

6.5 Gegenüberstellung nach Produktkontext

Die 22 Bilder mit PIM-Verknüpfung ergeben 264 Urteile je Dimension, die acht ohne Verknüpfung 96. Tabelle 10 vergleicht den Anteil der fünften Antwortoption. Diese deskriptive Gegenüberstellung erfasst fehlende eindeutige Auswahl und darf nicht als Vergleich absoluter Qualität interpretiert werden.

Tabelle 10: Fünfte Antwortoption nach Verfügbarkeit von Produktkontext

Quelle: Eigene Berechnung auf Basis der Erhebungsdaten

Gruppe	Bilder	Urteile je Dimension	Fünfte Option Beschreibungen	Fünfte Option Tags
Mit PIM	22	264	42 (15,91 %)	48 (18,18 %)
Ohne PIM	8	96	17 (17,71 %)	15 (15,63 %)

Bei Beschreibungen entfallen mit PIM-Kontext 42 von 264 Urteilen (15,91 Prozent) auf die fünfte Option, ohne Kontext 17 von 96 (17,71 Prozent). Bei Tags sind es 48 von 264 (18,18 Prozent) und 15 von 96 (15,63 Prozent). Die Unterschiede betragen damit 1,80 beziehungsweise 2,56 Prozentpunkte in entgegengesetzter Richtung (Abbildung 8).

Die sieben Kommentare zur fehlenden Beurteilbarkeit betreffen jeweils die fünfte Option in beiden Dimensionen. Werden die zugehörigen Entscheidungen ergänzend ausgeklammert, verbleiben 52 von 353 (14,73 Prozent) bei Beschreibungen und 56 von 353 (15,86 Prozent) bei Tags. Die Hauptauswertung behält alle abgegebenen Urteile bei.

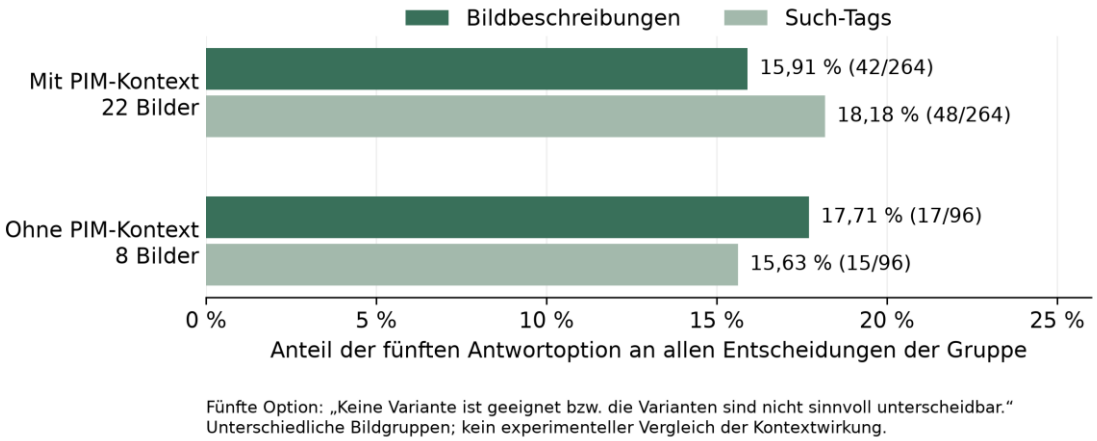


Abbildung 8: Anteil der fünften Antwortoption nach Verfügbarkeit von Produktkontext

Quelle: Eigene Berechnung auf Basis der Erhebungsdaten

Die Gruppen unterscheiden sich neben der Kontextverfügbarkeit auch in Motiv und Bildkategorie. Der Vergleich isoliert deshalb keinen Kontexteffekt. Für einen Wirkungsnachweis müssten dieselben Bilder mit und ohne Produktblock bewertet werden.

Auch die Kommentare erfordern eine feldbezogene Betrachtung. Ein Produktblock enthält nicht automatisch jede verlangte Information. Identifikationswünsche betreffen teilweise ausdrücklich ausgeschlossene Tags. Bildtypbezeichnungen sind kein übertragenes PIM-Feld. Für die Weiterentwicklung sind daher Kontextauswahl, Bezug zum sichtbaren Produkt und fachliche Ausgaberegeln gemeinsam zu betrachten.

6.6 Kostenprojektion

Für den Produktivbetrieb über AWS Bedrock fallen nutzungsabhängige Token-Kosten an. AWS stellt Gemma4:31b über das Bedrock-Modellverzeichnis bereit und berechnet 0,17 USD je eine Million Input-Tokens und 0,48 USD je eine Million Output-Tokens (Amazon Web Services, 2026; Region Europe/Frankfurt, Stand September 2026). Das Modell selbst steht unter Apache-2.0-Lizenz. Der Preis bezieht sich auf die von AWS bereitgestellte Inferenz.

Auf Basis der im Evaluationslauf gemessenen Durchschnittswerte lassen sich die Kosten für den relevanten Bestand von 107.000 Assets hochrechnen. Tabelle 11 zeigt die Kalkulation.

Tabelle 11: Kostenprojektion Gemma4:31b via AWS Bedrock (107.000 Assets)

Quelle: Eigene Berechnung; Preisangaben nach Amazon Web Services, 2026

Metrik	**Wert**
Input-Tokens pro Asset (Durchschnitt)	1.187,8
Output-Tokens pro Asset (Durchschnitt)	1.197,1
Input-Tokens gesamt	127,1 Mio.
Output-Tokens gesamt	128,1 Mio.
Input-Kosten (0,17 USD / 1 Mio. Tokens)	21,61 USD
Output-Kosten (0,48 USD / 1 Mio. Tokens)	61,49 USD
Gesamtkosten	83,10 USD

Die Projektion ergibt rund 83 USD reine Inferenzkosten für 107.000 Assets.

Die Rechnung verwendet Token-Mittelwerte des lokalen V3-Evaluationslaufs. Produktiv kommen V4 und die Bedrock-Ausführung zum Einsatz. Unterschiedliche Eingabe- und Ausgabeumfänge können die Kosten verändern. Wiederholungsversuche, Lambda, Datenübertragung und Schnittstellenlast sind nicht enthalten. Abschnitt 7.1 stellt die betrieblich dokumentierten Nutzungswerte gegenüber.

Sequenzielle lokale Verarbeitung würde bei 61,09 Sekunden je Bild für 107.000 Assets rund 1.816 Stunden beziehungsweise 75,7 Tage beanspruchen. Eine angenommene Leistungsaufnahme von 0,15 bis 0,25 kW entspräche 272 bis 454 kWh; bei 0,20 EUR je kWh wären dies rund 54 bis 91 EUR. Diese Szenariorechnung ersetzt keine gemessene Vollkostenanalyse.

Die Energiekosten der lokalen Verarbeitung liegen damit in derselben Größenordnung wie die projizierten Inferenzkosten der Cloud-Ausführung; eine reine Kostengegenüberstellung begründet die Entscheidung daher nicht. Für die Cloud-Entscheidung sprechen insbesondere parallelisierbarer Durchsatz und die Vermeidung einer langfristig gebundenen lokalen Einzelressource. Der Vergleich begründet die betriebliche Wahl, jedoch keine allgemeine wirtschaftliche Unterlegenheit lokaler Inferenz.

6.7 Zusammenfassung und Empfehlung

Gemma4:31b erhält die meisten Beschreibungsstimmen und liegt bei Tags hinter Claude Sonnet 5 auf Rang zwei. Für den vorgesehenen Einsatz werden Beschreibungen besonders gewichtet, da sie die redaktionelle Bereitstellung von Alternativtexten unterstützen. Die Präferenzen sind dabei ein Auswahlkriterium neben Kosten, Betrieb und Lizenzbedingungen.

Empfohlen wird Gemma4:31b mit dem aus den Fachrückmeldungen weiterentwickelten Prompt V4 über AWS Bedrock. Die Region eu-central-1 entspricht der vorgesehenen EU-Betriebsumgebung, die Regionswahl allein begründet keine pauschale Datenschutzkonformität. Die Empfehlung überträgt die V3-Ergebnisse als Entscheidungsgrundlage auf die Produktkonfiguration, ohne eine gesonderte V4-Qualitätsmessung vorzunehmen.

Gemma4:e4b wird wegen seines geringen Zuspruchs bei Beschreibungen nicht ausgewählt. Sein geringerer Abstand bei Tags belegt keine allgemeine Eignung für Stichwortextraktion. Da beide Gemma-Varianten gleich quantisiert wurden, lässt der Vergleich zudem offen, ob die 4-Bit-Quantisierung das kleinere Modell stärker beeinträchtigt.

7 DISKUSSION

Dieses Kapitel berichtet zunächst die Kennzahlen des produktiven Erstlaufs, beantwortet auf dieser erweiterten Grundlage die Forschungsfragen, benennt die Limitationen und leitet praktische Implikationen ab.

7.1 Ergebnisse des produktiven Erstlaufs

Nach der Evaluation wurde Gemma4:31b mit Prompt V4 über AWS Bedrock auf den relevanten Bestand angewendet. Der Erstlauf am 9. und 10. September 2026 schließt die Demonstrationsphase ab (Abschnitt 3.1.2). Die Discovery plante 108.342 Assets ein; die zuvor genannten rund 107.000 beziehen sich auf den früheren Analysestand. Tabelle 12 beruht auf der beigefügten Protokollauswertung und beschreibt den Betrieb, nicht die fachliche Qualität der Ausgaben.

Tabelle 12: Kennzahlen des produktiven Erstlaufs

Quelle: Eigene Darstellung auf Basis der beigefügten Protokollauswertung

Kennzahl	**Wert**
Eingeplante Assets	108.342
Als abgeschlossen gemeldete Jobs (COMPLETED)	103.568 (95,59 %)
Zur manuellen Prüfung markiert	4.504 (4,16 %)
Wiederholungsversuch zum Auswertungszeitpunkt ausstehend	270 (0,25 %)
Jobs mit mindestens einem Wiederholungsversuch	621
Protokollierte Retry-Ereignisse	1.931
Laufzeit des Gesamtlaufs	30,6 Stunden
Durchschnittliche Antwortzeit des Modells	4,46 s
Gespeicherte Input-Tokens	211.400.676
Gespeicherte Output-Tokens	37.495.305
Tokens je gespeichertem Jobergebnis (Ø; n = 108.002)	2.304,5
Kosten der gespeicherten Tokennutzung	53,94 USD
Gespeicherte Tag-Collection-Zuordnungen	30.118

Von 108.342 eingeplanten Jobs meldete das System 103.568 als abgeschlossen (95,59 Prozent). Zur manuellen Prüfung waren 4.504 Jobs (4,16 Prozent) markiert, für 270 (0,25 Prozent) stand ein erneuter Versuch aus. 621 Jobs hatten mindestens einen Wiederholungsversuch; insgesamt wurden 1.931 Retry-Ereignisse protokolliert.

In 30,6 Stunden wurden rund 3.400 Jobs je Stunde als abgeschlossen gemeldet. Gegenüber der Hochrechnung einer sequenziellen lokalen Verarbeitung von 107.000 Bildern mit V3 (75,7 Tage) liegt die Laufzeit des parallelen Cloud-Szenarios mit V4 etwa um den Faktor 59 niedriger. Die durchschnittliche gespeicherte Antwortdauer betrug 4,46 Sekunden. Der Vergleich beschreibt unterschiedliche Betriebsszenarien und keinen isolierten Modellgeschwindigkeitseffekt.

Die 108.002 Jobergebnisse mit Nutzungsangaben enthalten 211,4 Millionen Input- und 37,5 Millionen Output-Tokens. Mit den Preisen aus Abschnitt 6.6 ergeben sich 35,94 USD für Eingaben und 18,00 USD für Ausgaben, zusammen 53,94 USD. Die Werte umfassen die zuletzt gespeicherte Nutzung je Job und sind keine Gesamtabrechnung sämtlicher Versuche. Gegenüber der V3-Projektion stiegen die durchschnittlichen Input-Tokens von 1.187,8 auf 1.957,4, während die Output-Tokens von 1.197,1 auf 347,2 sanken. Der gesamte Asset-Bestand enthält mehr Produktbilder und weniger Lifestyle- / Ambientebilder als der Testbestand. Die daraus resultierenden kürzeren Texte sind ein Grund für die geringere Anzahl an Output-Tokens.

Der lokale Zuordnungsspeicher enthielt 30.118 Tag-Collection-Einträge. Er erfasst sowohl wiedergefundene als auch neu angelegte Collections und beschreibt damit die gespeicherten Begriffszuordnungen im Betrieb.

7.2 Beantwortung der Forschungsfragen

F1 wird durch die implementierte Pipeline beantwortet: Ein DAM-Client erschließt Assets, die PIM-Anbindung stellt verfügbaren Produktkontext bereit, die KI-Schicht erzeugt strukturierte Metadaten, und Normalisierung und TagResolver übertragen diese nach Celum. Das Begriffsregelwerk verbindet generative Ausgaben mit unternehmensspezifischen Konventionen. Die modularen Schnittstellen erlauben den Wechsel unterstützter Modellzugänge und begrenzen die Abhängigkeit von konkreten Anbietern.

Zu F2 zeigt die Befragung eine aufgabenabhängige Präferenz: Gemma4:31b führt bei deutschen Beschreibungen, Claude Sonnet 5 bei deutschen Tags. Gemma4:31b benötigt lokal die längste Inferenzzeit, erzeugt aber die knappsten

deutschen Ausgaben. Die Produktentscheidung verbindet den Beschreibungszuspruch mit Open-Weight-Lizenz und der vorgesehenen Cloud-Kostenstruktur. Sie ist eine begründete betriebliche Auswahl und kein Nachweis universeller Modellüberlegenheit.

F3 wird als Entwicklungsbeitrag beantwortet: Die dokumentierte Prompt-Iteration übersetzt fachliche Beobachtungen in konkrete Instruktionen. Die PIM-Anbindung macht unternehmensspezifische Referenzdaten für die Generierung verfügbar.

V2 adressiert Oberflächenbezeichnungen, V3 präzisiert Quellenbezug und Tag-Regeln, V4 greift Fachrückmeldungen etwa zu Terminologie und Wortbildungen auf. Nachvollziehbar ist damit, wie Anforderungen und Rückmeldungen die Gestaltung beeinflussen. Da der Modellvergleich einheitlich mit V3 durchgeführt wurde, quantifiziert er keine Verbesserung zwischen Prompt-Versionen.

Die Gegenüberstellung nach PIM-Verknüpfung isoliert ebenfalls keine kausale Wirkung des Kontexts. Die qualitative Analyse zeigt vielmehr, welche Referenzfelder und Ausgaberegeln gemeinsam abgestimmt werden müssen. Kontextdaten können die Generierung fachlich verankern, müssen aber zum sichtbaren Produkt und zum jeweiligen Ausgabezweck passen.

Die qualitative Auswertung konkretisiert diesen Entwicklungsbedarf. Von 29 Identifikationskommentaren betreffen 25 Bilder mit PIM-Kontext. Ein Teil verlangt auch Angaben in Tags, die FA3 ausschließt. Bildtypwünsche beziehen sich nicht auf ein übertragenes Feld, sondern müssen interpretiert werden. Funktionale Merkmale werden sowohl vermisst als auch als unzutreffend beanstandet. Entscheidend für die Gestaltung ist daher der Abgleich von verfügbarer Information, sichtbarem Produkt und gewünschter Ausgabe.

Die übergeordnete Forschungsfrage wird somit durch ein integriertes Produkt und einen nachvollziehbaren Auswahl- und Entwicklungsprozess beantwortet. Die Arbeit verbindet Datenintegration, Prompt-Gestaltung und fachliche Evaluation bis zum produktiven Einsatz. Ihr wissenschaftlicher Beitrag liegt in diesem dokumentierten Vorgehen und der differenzierten Bewertung der eingesetzten Konfigurationen.

7.3 Limitationen

Die Untersuchung ist auf 30 gezielt ausgewählte Bilder und zwölf Fachpersonen im Hansgrohe-Kontext begrenzt. Auswahlregeln, Zeitpunkt und auswählende Person der 30 Bilder sind nicht dokumentiert. Selektionseffekte lassen sich daher nicht ausschließen. Wiederholte Bewertungen nach Person und Bild sowie unterschiedliche Nutzungsperspektiven beschränken die Verallgemeinerbarkeit. Die Ergebnisse werden deshalb deskriptiv berichtet.

Die 360 Einzelentscheidungen je Dimension sind nach Personen und Bildern abhängig. Wie stark einfache Intervalle die Unsicherheit dadurch unterschätzen, zeigt der Vergleich mit den personenbezogenen Quoten: Für Gemma4:31b reicht deren 95-Prozent-Intervall bei den Beschreibungen von 28,50 bis 52,05 Prozent und ist damit mehr als doppelt so breit wie das Wilson-Intervall über alle Einzelurteile (32,26 bis 43,14 Prozent). Die in Abschnitt 6.2 berichteten Tests und Intervalle sind deshalb als ergänzende Einordnung und nicht als Absicherung der Rangfolge zu lesen. Auch geringe Übereinstimmungswerte zeigen heterogene Auswahlentscheidungen, ohne deren Ursachen zu erklären. Ein für Personen- und Bildeffekte geeignetes Untersuchungsdesign wäre für weitergehende Verallgemeinerungen erforderlich.

Die Auswertung zum Einfluss des Produktkontexts ist nicht vollständig aussagekräftig. Beide Bildgruppen wurden mit identischem Prompt und identischer Kontextlogik verarbeitet. Verglichen werden deshalb Bilder, die sich neben der Kontextverfügbarkeit auch in Motiv, Bildkategorie und Komplexität unterscheiden. Ein kausaler Nachweis erforderte dieselben Bilder einmal mit und einmal ohne Kontextblock. Der Testmodus des Systems erlaubt das. Ein solcher Durchlauf wurde im Rahmen dieser Arbeit nicht durchgeführt.

Evaluiert wurden die lokalen beziehungsweise cloudbasierten Modellkonfigurationen mit einem gemeinsamen Prompt V3. Produktiv nutzt Gemma4:31b dagegen die Bedrock-Ausführung und Prompt V4. Diese Kombination wurde nicht gesondert präferenzbasiert bewertet. Die Produktentscheidung stützt sich auf den Modellvergleich und die fachlich begründete Weiterentwicklung. Die Betriebskennzahlen ersetzen keinen separaten Qualitätsvergleich.

Die qualitative Auswertung beruht auf einer KI-gestützten Kodierung ohne unabhängige menschliche Zweitkodierung. Zudem wurden die Kommentare nicht vollständig einzelnen Modellen zugeordnet. Die Häufigkeiten beschreiben deshalb Themen über alle Varianten hinweg. Die Kostenprojektion beruht auf V3 und erfasst ausschließlich Inferenzkosten (Abschnitt 6.6).

Die präferenzbasierte Methode liefert relative und keine absoluten Qualitätsurteile. Die fünfte Option vermeidet zwar erzwungene Modellwahlen, trennt aber fehlende Eignung nicht von Gleichstand. Die betriebliche Einsatzentscheidung beruht deshalb auf der Gesamtschau aus Präferenzen, qualitativen Rückmeldungen und technischen Betriebsanforderungen, nicht auf einer gemessenen absoluten Abnahmequote.

Prompt V3 wurde einheitlich auf alle Modelle angewendet. Modellspezifische Prompt-Optimierung hätte die Ergebnisse einzelner Modelle verbessern können, würde aber den Prompt-Effekt vom Modell-Effekt untrennbar machen. Die bewusste Entscheidung für einen einheitlichen Prompt erhöht die Vergleichbarkeit auf Kosten potenziell höherer Einzelleistungen.

Die Ergebnisse sind an den Hansgrohe-Kontext gebunden. Produktsortiment, Bildtypen, Begriffskonventionen und Fachsprache prägen die Evaluationsergebnisse. Die Architektur und die Evaluationsmethodik sind übertragbar, die konkreten Modellvergleichswerte gelten nur für diese spezifische Domäne. Ein anderes Unternehmen mit anderen Produkten und anderen Bildtypen könnte zu abweichenden Modellpräferenzen gelangen.

Die Verfügbarkeit und Vollständigkeit der PIM-Daten bestimmt, welche zusätzlichen Referenzinformationen der Generierung zugänglich sind. Ohne Produktverknüpfung arbeitet das Modell ausschließlich mit Bild und übrigen Eingaben. Welchen Qualitätsunterschied dieser zusätzliche Kontext verursacht, wurde nicht isoliert gemessen.

Eine ergänzende Entwicklungsmöglichkeit ist die regelbasierte Prüfung eindeutig zugeordneter Referenzwerte, etwa offizieller Oberflächenbezeichnungen. Die vorhandene Normalisierung vereinheitlicht bereits Synonyme und entfernt redundante Begriffe. Ein weitergehender Abgleich müsste sicherstellen, dass der PIM-Wert tatsächlich zum sichtbaren Produkt gehört.

Für Referenzprüfungen kommen insbesondere tatsächlich übertragene und eindeutig zugeordnete Produktfelder infrage. Die Kommentarzahlen zeigen keine Zahl behebbarer Fehler. Ein Teil der Wünsche betrifft bewusst ausgeschlossene Tag-Inhalte oder nicht übertragene Merkmale. Visuelle Fehlinterpretationen und Bildtypbezeichnungen erfordern zudem andere Prüfmaßstäbe als der Abgleich eines Produktnamens.

Für einen weiterführenden Vergleich bieten sich dieselben Bilder mit unterschiedlichen Prompt-Versionen oder mit und ohne Produktblock an. Ein unabhängiger Prüfsatz und abgestimmte Kriterien für Tags und Beschreibungen würden die Wirkung einzelner Entwicklungsschritte gezielter erfassen.

7.4 Praktische Implikationen

Für Hansgrohe wurde Gemma4:31b mit Prompt V4 über AWS Bedrock in eu-central-1 umgesetzt. Die nutzungsabhängige Inferenz und die automatisierte Verarbeitung neuer Assets über AWS Lambda unterstützen den Betrieb ohne eine dauerhaft vorzuhaltende lokale Inferenzinstanz.

Der Erstlauf erschließt den relevanten Bestand. Neue Assets können fortlaufend verarbeitet werden. Die Fachredaktion behält Aufgaben bei Stichproben, Prüffällen und der Weiterentwicklung des Begriffsregelwerks. Die vorhandene Evaluations-Toolchain bietet eine Grundlage für spätere Modell- und Prompt-Vergleiche.

Übertragbar ist vor allem die Trennung von DAM-Anbindung, Kontextbeschaffung, Generierung und Normalisierung. Ein Einsatz in einem anderen Unternehmen erfordert die Anpassung der Schnittstellen und fachlichen Regeln sowie eine erneute Evaluation, denn die konkreten Präferenzwerte sind domänenspezifisch.

Eine technisch korrekt umgesetzte Anforderung kann der fachlichen Erwartung ihrer Nutzenden unmittelbar widersprechen. FA3 schließt Marke und Linie aus den Such-Tags aus, weil beide in Celum eigene Informationsfelder besitzen. Elf der fachlichen Kommentare verlangen genau diese Angaben dennoch ausdrücklich in den Tags. Eine solche Differenz ist durch Prompt-Iteration nicht zu behe-

ben, denn der Prompt setzt die Anforderung korrekt um. Sie erfordert eine Abstimmung über die Anforderung selbst. Datenintegration, Prompt-Gestaltung und Begriffsregelwerk sind deshalb gemeinsam mit den Nutzenden zu entwickeln. Technische Regelkonformität und fachliche Präferenz sind unterschiedliche Kriterien und ausdrücklich zu verhandeln.

8 FAZIT UND AUSBLICK

8.1 Zusammenfassung

Diese Arbeit entwickelt, implementiert und evaluiert ein einsatzfähiges System zur automatisierten Generierung von Bildmetadaten für das DAM-System Celum der Hansgrohe SE. Vier Vision Language Models wurden anhand von 30 Bildern durch zwölf Fachexperten in einem Best-of-4-Design verglichen.

Gemma4:31b erhält bei Beschreibungen den höchsten Zuspruch, Claude Sonnet 5 bei Tags. Gemma4:e4b liegt insbesondere bei Beschreibungen zurück. Unter Einbezug von Betrieb, Lizenz und Kosten wird Gemma4:31b für die weitere Nutzung ausgewählt. Die Befragung liefert hierfür eine praxisbezogene, deskriptive Entscheidungsgrundlage.

Die Pipeline verbindet Bildanalyse, verfügbaren Produktkontext aus Centric PXM und unternehmensspezifische Begriffsregeln. Sie führt die Ergebnisse in die bestehende DAM-Struktur zurück und verarbeitet einen durch Filter eingegrenzten Bestand von rund 107.000 relevanten Assets. Die modulare Architektur trennt fachliche Verarbeitung und technische Schnittstellen.

Die vier dokumentierten Prompt-Versionen machen die Weiterentwicklung anhand beobachteter Ausgaben und fachlicher Rückmeldungen nachvollziehbar. Die Kommentaranalyse zeigt neben inhaltlichen Beanstandungen auch unterschiedliche Erwartungen an zulässige Metadaten. Der Beitrag besteht in der systematischen Übersetzung dieser Erkenntnisse in Gestaltungsentscheidungen. Ein abschließender Wirksamkeitsvergleich der Versionen wurde nicht durchgeführt.

Gemma4:31b und Prompt V4 werden über AWS Bedrock produktiv eingesetzt. Im dokumentierten Erstlauf wurden innerhalb von 30,6 Stunden 103.568 der 108.342 eingeplanten Jobs abgeschlossen. Die Auswertung der gespeicherten Tokenwerte ergibt Kosten von 53,94 USD. Die automatisierte Weiterverarbeitung neuer Assets wurde über AWS Lambda umgesetzt.

Über den Einzelfall hinaus liefert die Arbeit ein nachvollziehbares Integrationsmuster und einen Evaluationsablauf ohne vorab erstellte Referenzbeschreibungen. Sie zeigt, wie technische Modellvergleiche mit fachlichen Anforderungen und betrieblicher Umsetzung verbunden werden können. Die Ergebnisse ergänzen den Forschungsstand durch eine konkrete Anwendung in einem Enterprise-DAM, ohne einen allgemeinen Vorrang bestimmter Modelle oder Evaluationsverfahren zu beanspruchen.

Im Verhältnis zum Forschungsstand aus Abschnitt 2.7 verdeutlicht die Anwendung damit das Zusammenspiel visueller Erkennung, Instruktionsbefolgung und fachlicher Konventionen. Die Rückmeldungen der Experten betreffen Bildinhalte, Terminologie und Ausgaberegeln. Die gewählte Expertenbefragung erschließt diese Nutzungsperspektiven. Eine Überlegenheit gegenüber automatischen Evaluationsverfahren wurde damit nicht untersucht.

8.2 Ausblick

Prompt-Weiterentwicklung bleibt eine laufende Aufgabe. Neue Modellgenerationen können die Instruktionsbefolgung verändern — bestehende Prompts sollten daher erneut evaluiert werden. Die Toolchain erleichtert die Erzeugung vergleichbarer Ausgaben und Umfragen. Neue Qualitätsvergleiche benötigen entsprechende fachliche Bewertungen.

Eine zusätzliche Qualitätskontrolle könnte fachliche Stichproben und regelbasierte Referenzprüfungen verbinden. Modellbasierte Unsicherheitssignale wären zuvor gegen tatsächlich festgestellte Fehler zu validieren. Technische Validität und inhaltliche Eignung bleiben dabei getrennte Prüfdimensionen.

Möglich ist auch eine Erweiterung auf weitere Medientypen. Video-Assets erfordern Frame-Sampling und eine Aggregation über mehrere Einzelbilder. PDF-Dokumente erfordern eine Kombination aus Textextraktion und visueller Analyse. Die modulare Architektur des Systems bietet Ansatzpunkte für solche Erweiterungen, weil die fachliche Verarbeitung von den Schnittstellen getrennt ist. Zusätzliche Eingabeformate erfordern passende Aufbereitung und Validierung.

Eine Ausweitung auf weitere Sprachen erfordert entsprechende Ausgabe- und Verarbeitungsvorgaben sowie sprachkundige Bewertungen. Bereits die vorhandenen englischen Ausgaben bilden hierfür einen Ausgangspunkt. Ihre Qualität wurde in der deutschsprachigen Präferenzbefragung nicht untersucht.

Eine Studie zur Nutzung der generierten Metadaten könnte den betrieblichen Mehrwert weiter untersuchen. Dazu gehören Suchaufgaben zur Auffindbarkeit der Assets und die Prüfung der Alternativtexte im tatsächlichen Webseitenkontext. Auch Konsistenz und Pflege der verwendeten Begriffe sind relevant, denn umfangreiche Metadaten helfen nur, wenn sie zu den Suchgewohnheiten und fachlichen Konventionen passen.

Auf der bestehenden Datenbasis sind auch Auswertungen nach Personen und Bildern oder eine Rekonstruktion eindeutig modellbezogener Kommentare möglich. Ein Vergleich von V3 und V4 oder derselben Bilder mit und ohne Produktkontext erfordert neue Modellausgaben und erneute Bewertungen. Die vorhandene Toolchain bildet dafür die technische Grundlage.

LITERATURVERZEICHNIS

Amazon Web Services. (2026). Amazon Bedrock Pricing. https://aws.amazon.com/bedrock/pricing/ Abrufdatum: 14.09.2026

Anthropic. (2026). Models overview. https://platform.claude.com/docs/en/models/overview Abrufdatum: 18.09.2026

Baca, M. (Hrsg.). (2008). Introduction to Metadata (2. Aufl., Version 3.0). Getty Research Institute.

Brown, L. D., Cai, T. T. & DasGupta, A. (2001). Interval Estimation for a Binomial Proportion. Statistical Science, 16(2), 101–133. https://doi.org/10.1214/ss/1009213286

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language Models are Few-Shot Learners. Advances in Neural Information Processing Systems, 33, 1877–1901. https://doi.org/10.48550/arXiv.2005.14165

Cochran, W. G. (1950). The Comparison of Percentages in Matched Samples. Biometrika, 37(3/4), 256–266. https://doi.org/10.2307/2332378

Dettmers, T., Lewis, M., Belkada, Y. & Zettlemoyer, L. (2022). LLM.int8(): 8-bit Matrix Multiplication for Transformers at Scale. Advances in Neural Information Processing Systems, 35. https://doi.org/10.48550/arXiv.2208.07339

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J. & Houlsby, N. (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. International Conference on Learning Representations. https://doi.org/10.48550/arXiv.2010.11929

Europäisches Parlament. (2019). Richtlinie (EU) 2019/882 über die Barrierefreiheitsanforderungen für Produkte und Dienstleistungen. Amtsblatt der Europäischen Union, L 151/70.

Fleiss, J. L. (1971). Measuring Nominal Scale Agreement Among Many Raters. Psychological Bulletin, 76(5), 378–382. https://doi.org/10.1037/h0031619

Frantar, E., Ashkboos, S., Hoefler, T. & Alistarh, D. (2023). GPTQ: Accurate Post-Training Quantization for Generative Pre-Trained Transformers. International Conference on Learning Representations. https://doi.org/10.48550/arXiv.2210.17323

Gemma Team. (2026). Gemma 4 Technical Report. arXiv. https://doi.org/10.48550/arXiv.2607.02770

Goodfellow, I., Bengio, Y. & Courville, A. (2016). Deep Learning. MIT Press.

Google DeepMind. (2026). Gemma 4 model card. https://ai.google.dev/gemma/docs/core/model_card_4 Abrufdatum: 14.09.2026

Hevner, A. R., March, S. T., Park, J. & Ram, S. (2004). Design Science in Information Systems Research. MIS Quarterly, 28(1), 75–105. https://doi.org/10.2307/25148625

Klemm, P. (2025). Asset-Verschlagwortung in einem DAM (Digital Asset Management) & MAM (Media Asset Management) System mit Hilfe von KI [Projektarbeit 2]. Duale Hochschule Baden-Württemberg Mosbach.

Kreiss, E., Bennett, C., Hooshmand, S., Zelikman, E., Ringel Morris, M. & Potts, C. (2022). Context Matters for Image Descriptions for Accessibility: Challenges for Referenceless Evaluation Metrics. Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 4685–4697. https://doi.org/10.18653/v1/2022.emnlp-main.309

Landis, J. R. & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. Biometrics, 33(1), 159–174. https://doi.org/10.2307/2529310

Lin, C.-Y. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. Text Summarization Branches Out: Proceedings of the ACL-04 Workshop, 74–81.

Lin, J., Tang, J., Tang, H., Yang, S., Chen, W.-M., Wang, W.-C., Xiao, G., Dang, X., Gan, C. & Han, S. (2024). AWQ: Activation-aware Weight Quantization for On-Device LLM Compression and Acceleration. Proceedings of Machine Learning and Systems, 6. https://doi.org/10.48550/arXiv.2306.00978

Liu, C.-W., Lowe, R., Serban, I. V., Noseworthy, M., Charlin, L. & Pineau, J. (2016). How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2122–2132. https://doi.org/10.18653/v1/D16-1230

Liu, H., Li, C., Wu, Q. & Lee, Y. J. (2023). Visual Instruction Tuning. Advances in Neural Information Processing Systems, 36. https://doi.org/10.48550/arXiv.2304.08485

March, S. T. & Smith, G. F. (1995). Design and Natural Science Research on Information Technology. Decision Support Systems, 15(4), 251–266. https://doi.org/10.1016/0167-9236(94)00041-2

McNemar, Q. (1947). Note on the Sampling Error of the Difference Between Correlated Proportions or Percentages. Psychometrika, 12(2), 153–157. https://doi.org/10.1007/BF02295996

Novikova, J., Dušek, O., Cercas Curry, A. & Rieser, V. (2017). Why We Need New Evaluation Metrics for NLG. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2241–2252. https://doi.org/10.18653/v1/D17-1238

Papineni, K., Roukos, S., Ward, T. & Zhu, W.-J. (2002). BLEU: A Method for Automatic Evaluation of Machine Translation. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 311–318. https://doi.org/10.3115/1073083.1073135

Peffers, K., Tuunanen, T., Rothenberger, M. A. & Chatterjee, S. (2007). A Design Science Research Methodology for Information Systems Research. Journal of Management Information Systems, 24(3), 45–77. https://doi.org/10.2753/MIS0742-1222240302

Qwen Team. (2026). Qwen3.8-27B model card. Hugging Face. https://hugging-face.co/Qwen/Qwen3.8-27B Abrufdatum: 14.09.2026

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G. & Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. Proceedings of the 38th International Conference on Machine Learning, 8748–8763. https://doi.org/10.48550/arXiv.2103.00020

W3C. (2024). Web Content Accessibility Guidelines (WCAG) 2.2. W3C Recommendation vom 12.12.2024. https://www.w3.org/TR/WCAG22/ Abrufdatum: 14.09.2026

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q. V. & Zhou, D. (2022). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. Advances in Neural Information Processing Systems, 35. https://doi.org/10.48550/arXiv.2201.11903

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J. & Schmidt, D. C. (2023). A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT. arXiv preprint. https://doi.org/10.48550/arXiv.2302.11382

Willard, B. T. & Louf, R. (2023). Efficient Guided Generation for Large Language Models. arXiv preprint. https://doi.org/10.48550/arXiv.2307.09702

Zamfirescu-Pereira, J. D., Wong, R. Y., Hartmann, B. & Yang, Q. (2023). Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts. Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. https://doi.org/10.1145/3544548.3581388

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. Advances in Neural Information Processing Systems, 36. https://doi.org/10.48550/arXiv.2306.05685

ANHANG

Der Anhang dokumentiert die Artefakte, die zur Reproduktion der Ergebnisse erforderlich sind.

Anhangsverzeichnis

Anhang A – Prompts im Volltext

Anhang B – Evaluationsdatensatz

Anhang C – Modell- und Laufzeitkonfiguration

Anhang D – Freitextkommentare mit Kategorisierung

Anhang E – Beispielausgaben der vier Modelle

Anhang F – Rohdaten, Auswertung und Artefakt

Die Seitenzahlen der einzelnen Anhänge sind dem Inhaltsverzeichnis zu entnehmen. Die in Anhang F genannten Dateien liegen als elektronische Anlage bei.

Anhang A – Prompts im Volltext

Die Prompts liegen als versionierte Textdateien im Quellcode-Verzeichnis. Die geschweiften Klammern in doppelter Form sind Escape-Sequenzen der Template-Engine und erscheinen im an das Modell übergebenen Prompt als einfache Klammern. Die Platzhalter in einfachen geschweiften Klammern werden zur Laufzeit ersetzt: {language_text} durch die konfigurierten Sprachen, {product_section} durch den aus dem PIM erzeugten Kontextblock, {max_tags} durch die maximale Tag-Anzahl je Sprache und {filename} durch den Dateinamen des Assets.

A.1 Systemprompt (system.txt)

```
Du bist ein KI-Modell, das Metadaten für ein Digital Asset Management (DAM) System
generiert. Du analysierst Bilder aus Sicht eines Unternehmens der Sanitär- und Bad-
branche (Hansgrohe). Gib deine Antwort ausschließlich als JSON-Objekt zurück.
```

A.2 Prompt V1 (tagging_v1.txt)

```
Analysiere das Bild und erzeuge zwei Arten von Metadaten:

1. SUCH-TAGS für das DAM-System (Freitextsuche)
2. ALT-TEXT für die Website (WCAG 2.1 SC 1.1.1, Barrierefreiheit)

Sprachen:
{language_text}
{product_section}
=== SUCH-TAGS (tags) ===
Zweck: Nutzer im DAM finden Bilder über Freitextsuche. Tags müssen alle relevanten
Suchbegriffe abdecken.
Regeln:
- {max_tags} Tags pro Sprache, präzise und suchrelevant. Für alle möglichen Suchme-
thoden z.B.: Farbe, Raum, Umgebung...
- Synonyme einschließen wenn sinnvoll (z.B. "Waschtischarmatur" + "Waschtischmi-
scher").
- Keine generischen Füllwörter (Foto, Bild, hochwertig, premium, professionell).
- Keine Artikelnummern, Markennamen oder Linien.

=== ALT-TEXT (alt_text) ===
Zweck: WCAG-konformer Alternativtext für Screenreader. Ermöglicht blinden/sehbehin-
derten Nutzern ein gleichwertiges, beschreibendes Erlebnis.
Regeln:
- **So kurz wie möglich, so ausführlich wie nötig.**
- NIEMALS mit "Bild von", "Foto von", "Image of" beginnen — Screenreader kündigen
die Rolle bereits an.
- Die Linie ist der wichtigste Produktbezeichner (nicht der volle technische Ti-
tel).
- Sperrige Produktnamen kürzen: "Handbrause 120 3jet Powderrain" → "Handbrause Pow-
derrain"
- Nur beschreiben was sichtbar und entscheidungsrelevant ist. Auch Gefühl, Anmutung
(z.B.: Farbe...) des Bildes beschreiben, wenn nötig.
- Bei Freistellern: Produkt, Ausführung und ggf. Farbe/Material beschreiben.
- Bei Milieu-/Lifestyle-Bildern: Situation, Raum und Anwendungskontext beschreiben

=== ALLGEMEINE REGELN ===
- Nur klar sichtbare Inhalte beschreiben.
- Produktnamen, Marken und Linien NUR aus den PIM-Daten verwenden, niemals erfin-
den.
- Wenn kein Produkt identifizierbar: rein visuell beschreiben ohne Markennamen.
- Schreibweisen strikt einhalten: "hansgrohe" immer klein, "AXOR" immer groß.

Antwort als JSON:
{{
  "tags": {{"de": ["..."], "en": ["..."]}},
  "alt_text": {{"de": "...", "en": "..."}}
}}
Kein Markdown, kein Text außerhalb des JSON.
```

A.3 Prompt V2 (tagging_v2.txt)

```
Analysiere das Bild und erzeuge SUCH-TAGS (DAM) und ALT-TEXT (Website/Barrierefrei-
heit).

Sprachen: {language_text}
{product_section}

PRODUKTOBERFLÄCHEN

hansgrohe:
Chrom | Mattschwarz | Mattweiß | Polished Gold Optic | Brushed Bronze | Brushed
Black Chrome | Stainless Steel Optic

AXOR:
Chrom | Brushed Brass | Polished Gold Optic | Brushed Gold Optic | Brushed Bronze |
Polished Red Gold | Brushed Red Gold | Brushed Nickel | Stainless Steel Optic |
Polished Black Chrome | Brushed Black Chrome | Mattschwarz | Mattweiß

Möbel:
Schiefergrau Metallic | Schiefergrau Matt | Diamantgrau Matt | Textil Dunkelgrau |
Eiche Dunkel | Nussbaum Dunkel | Eiche Kaschmir | Eiche Natur | Bronze Struktur |
Sandbeige Metallic | Sandbeige Matt | Mattweiß | Weiß Metallic | Weiß Hochglanz

EN:
Chrome | Matt Black | Matt White | Brushed Brass | Polished Gold Optic | Brushed
Gold Optic | Brushed Bronze | Polished Red Gold | Brushed Red Gold | Brushed Nickel
| Stainless Steel Optic | Polished Black Chrome | Brushed Black Chrome | Slate Grey
Metallic | Slate Grey Matt | Diamond Grey Matt | Textile Dark Grey | Dark Oak |
Dark Walnut | Cashmere Oak | Natural Oak | Bronze Structure | Sand Beige Metallic |
Sand Beige Matt | White Metallic | White High Gloss

Nicht eindeutig → nur sichtbare Farbwirkung beschreiben (z.B. goldfarben, silber-
farben, schwarz matt).

TAGS

- Max. {max_tags} Tags je Sprache.
- Keine Dubletten.
- Suchrelevante Synonyme verwenden (z.B. Waschtischarmatur/Waschtischmischer, Kopf-
brause/Regendusche).
- Tags nur aus sichtbaren Inhalten ableiten.
- Abdecken, soweit sichtbar: Produkttyp, Farbe, Oberfläche, Material, Raum, Stil,
Einbausituation, Nutzungskontext, Bildtyp.
- Bildtyp erlaubt: Freisteller, Packshot, Lifestyle, Detailaufnahme, Ambiente.
- Offizielle Oberflächenbezeichnungen nur verwenden, wenn eindeutig sichtbar oder
in PIM-Daten vorhanden.
- Keine Marken, Liniennamen, Artikelnummern, Dateinamen oder Füllwörter.

ALT_TEXT

- WCAG-konform.
- So kurz wie möglich, so ausführlich wie nötig.
- Vollständige Sätze.
- Niemals mit „Bild von", „Foto von" oder „Image of" beginnen.
- Reihenfolge: Produkt → Raum/Material/Licht → Einbausituation → sichtbare Nutzung.
- Freisteller kurz, Ambiente mit sichtbarem Kontext.
- Linienname nur verwenden, wenn in PIM-Daten vorhanden.
- Sonst Produkttyp verwenden.
- Produkttitel bei Bedarf kürzen.
- Werbefrei formulieren.
- Personen nur über sichtbare Handlung beschreiben.
- Jede Sprache natürlich formulieren, nicht wörtlich übersetzen.

ALLGEMEIN

- Nur Sichtbares beschreiben.
- Produktnamen, Liniennamen und Oberflächen nur aus PIM-Daten verwenden.
- Nichts ergänzen oder vermuten.
- Ist kein Produkt identifizierbar, rein visuell beschreiben.
```

```
- Schreibweise: hansgrohe klein, AXOR groß.

Antwort ausschließlich als JSON:

{{
  "tags": {{"de": ["..."], "en": ["..."]}},
  "alt_text": {{"de": "...", "en": "..."}}
}}
```

A.4 Prompt V3 (tagging_v3.txt) – Grundlage der vergleichenden Evaluation

```
Analysiere das Bild und erstelle SUCH-TAGS für das DAM sowie WCAG-konforme ALT-
TEXTE für die Website.

Sprachen: {language_text}
PIM-DATEN: {product_section}

QUELLEN
- Bild: nur sichtbare Produkte, Farbe, Materialwirkung, Raum, Stil, Einbau,
  Nutzung, Personen und Bildtyp.
- PIM: Produktname, Linie, Produkttyp, offizielle Oberfläche, technische Merkmale.
- Nichts raten. Unsichere Merkmale weglassen.
- Offizielle Produkt-, Linien- und Oberflächenbezeichnungen ausschließlich aus PIM
verwenden.
- Ohne PIM-Oberfläche nur sichtbare Farbwirkung nennen, z. B. silberfarben,
  goldfarben, bronzefarben, schwarz matt, weiß matt oder holzfarben.

TAGS
- Maximal {max_tags} Tags pro Sprache, nach Relevanz sortiert.
- Keine Dubletten, Widersprüche, Marken, Liniennamen, Artikelnummern,
  Dateinamen, Logos, Bildtexte, Füllwörter oder Werbeaussagen.
- Sinnvolle Suchsynonyme ergänzen, z. B. Waschtischarmatur/Waschtischmischer,
  Kopfbrause/Regendusche.
- Soweit sichtbar oder durch PIM bestätigt: Produkttyp, Farbe/Oberfläche,
  Material, Raum, Stil, Einbausituation, Nutzungskontext, Bildtyp.
- Technische Funktionen nur bei PIM-Bestätigung oder eindeutiger Sichtbarkeit.
- Raum und Stil nur bei sichtbarem Kontext; bei Freistellern nur aus PIM
  bestätigten Nutzungskontext verwenden.
- Erlaubte Bildtypen: Freisteller, Packshot, Lifestyle, Detailaufnahme, Ambiente.
- Keine subjektiven Begriffe wie hochwertig, premium, Designer, Wellness,
  Entspannung oder Präzisionsarbeit.
- Tags je Sprache natürlich formulieren, nicht wortwörtlich übersetzen.

ALT-TEXT
- Vollständige, kurze, natürliche und werbefreie Sätze.
- Beschreibe nur sichtbare und relevante Inhalte.
- Reihenfolge: Produkt/Situation → sichtbare Merkmale → Raum/Material →
  Einbausituation → Handlung.
- Nie mit „Bild von", „Foto von" oder „Image of" beginnen.
- Keine Präsentationsdetails wie Freisteller, Packshot, Produktfoto,
  Studioaufnahme oder „vor weißem Hintergrund" nennen.
- Linie nur aus PIM verwenden; sonst Produkttyp. Lange Produkttitel sinnvoll kür-
zen.
- Personen nur anhand sichtbarer Handlung, Kleidung oder Position beschreiben.
- Natürlich in jeder angeforderten Sprache formulieren, nicht übersetzen.

AUSGABE
Nur gültiges JSON, kein Markdown oder zusätzlicher Text:
{{
  "tags": {{"de": ["..."], "en": ["..."]}},
  "alt_text": {{"de": "...", "en": "..."}}
}}
```

A.5 Prompt V4 (tagging_v4.txt) – im Produktivbetrieb eingesetzt

```
Erstelle SUCH-TAGS und eine BESCHREIBUNG für das abgebildete Bild.

SPRACHEN: {language_text}
DATEINAME: {filename}

PIM-DATEN:
{product_section}

---

KONTEXT

Tags werden intern von Fachleuten genutzt, die gezielt nach passenden Bildern
für Kataloge, Webseiten oder Präsentationen suchen. Tags müssen präzise,
fachlich korrekt und suchrelevant sein.

Die Beschreibung (alt_text) erscheint auf hansgrohe.com und bei Handelspartnern
als WCAG-konformer Alternativtext für Screenreader. Sie muss Menschen mit
Sehbehinderung ein gleichwertiges Verständnis des Bildinhalts ermöglichen.
Die Sprache repräsentiert hansgrohe gegenüber Kunden und Partnern und muss
professionell, fachlich korrekt und markenkonform sein.

---

TAGS ({max_tags} pro Sprache)

Stell dir vor, jemand sucht genau dieses Bild mit Suchbegriffen.
Fast alle unsere Bilder zeigen Sanitärprodukte — Begriffe wie Armatur oder
Wasserhahn allein helfen nicht. Tagge was dieses Bild wiedererkennbar macht.
Sanitärbegriffe dürfen vorkommen, aber maximal 2-3 pro Sprache. Der Schwerpunkt
liegt auf dem, was dieses konkrete Bild von anderen unterscheidet.
Nutze die volle Anzahl aus, lieber ein relevanter Tag zu viel als einer zu wenig.

Inhalte (alles taggen was sichtbar oder durch PIM bestätigt):
- Produkttyp, max. 2 Begriffe (z.B. Küchenarmatur, Waschtischarmatur)
- Oberfläche/Finish: den PIM-Namen für die jeweilige Sprache verwenden.
  Gängige Oberflächen rechtschreibkonform einpassen (DE gross: Chrom, Mattschwarz;
  EN klein: chrome, matte black). Eigennamen exakt aus PIM übernehmen
  (Polished Gold Optic, Brushed Black Chrome, Eiche Dunkel).
- Funktionen und Strahlarten aus dem PIM-Titel (CoolStart, EcoSmart, PowderRain,
  UniversalFit, Select, QuickClean — nur wenn im PIM genannt)
- Sichtbare Handlungen (Hände waschen, Glas spülen, Entkalken, Reinigen)
- Personen: Anzahl und Tätigkeit, nicht Geschlecht oder Beruf
- Materialien im Umfeld (Holz, Marmor, Beton, Fliesen, Glas, Keramik, Messing)
- Gegenstände und Details (Blumen, Handtücher, Kerzen, Seife, Spiegel)
- Raumtyp (Kücheninsel, Duschnische, freistehende Wanne, Gästebad, Hotelbad)
- Einrichtungsstil (minimalistisch, industriell, skandinavisch, modern, klassisch)
- Architekturelemente (Rundbogen, Dachschräge, bodentiefe Fenster)
- Besondere Lichtstimmung oder Farbwirkung (nur wenn prägend)
- Wasser (fliessendes Wasser, Wasserstrahl, Tropfen — nur wenn sichtbar)
- Bildtyp: genau einen verwenden:
  Freisteller, eCom-Bild, Produktrendering, Iso-Rendering,
  Detailaufnahme, Lifestyle, Teilambiente mit Person, Ambiente ohne Person

Regeln:
- Nur echte deutsche Wörter und etablierte Fachbegriffe verwenden.
  VERBOTEN: erfundene Zusammensetzungen wie unterputzig, Waschbeckenunterschale,
  Marmorkonterplatte, Keramiktoilette, Kopfbrausenkopf, Duschpipe,
  schwebender Waschtisch.
- Keine Marken, Linien oder Artikelnummern (sind in eigenen CELUM-Feldern).
- Keine Dubletten, Füllwörter oder Werbebegriffe.
- Oberflächen und Farben nur aus PIM, nie visuell schätzen.
- Nicht Packshot verwenden, stattdessen Freisteller oder Produktrendering.
- Bei Freistellern keine Raum-, Stil- oder Materialtags ergänzen.
- Fachbegriffe korrekt unterscheiden:
  Aufputz ≠ Unterputz ≠ Wandmontage.
  Handbrause, Kopfbrause, Brausestange ≠ Duschsystem/Showerpipe.
  Waschtisch ≠ Waschbecken ≠ Aufsatzbecken.
  Rosette, nicht Blende. Duschthermostat, nicht Brausethermostat.
```

```
---

ALT-TEXT (genau ein Satz pro Sprache, nicht länger)

Wenn das Hauptprodukt erkennbar ist UND PIM-Daten vorliegen:
- Marke und Linie MÜSSEN vorkommen.
- Schreibweisen: "hansgrohe" immer klein, "AXOR" immer gross.
- Oberfläche: den PIM-Namen für die jeweilige Sprache verwenden.
  Gängige Oberflächen rechtschreibkonform (DE: "in Chrom", EN: "in chrome").
  Eigennamen exakt aus PIM (Polished Gold Optic, Brushed Black Chrome).

Aufbau:
- Freisteller: [Marke Linie] Produkttyp in [Oberfläche].
- Milieu: [Marke Linie] Produkttyp in Einbausituation, prägendes Szenenmerkmal.

Beispiele:
- "hansgrohe Raindance Select S Handbrause in Chrom wird in einer Glasschüssel
   mit Wasser und Zitrone entkalkt"
- "AXOR Citterio Waschtischarmatur in Brushed Nickel auf einem weissen Aufsatzbe-
cken"
- "hansgrohe Vernis Blend Küchenarmatur in Mattschwarz an einer Spüle, eine Person
   füllt eine Glasvase mit Wasser"

Sprache:
- Endkundenfreundliche Begriffe bevorzugen: Küchenarmatur statt Küchenmischer,
  Waschtischarmatur statt Waschtischmischer.
- Nur reale, im Duden verzeichnete Wörter. Nichts erfinden.
- PIM-Titel kürzen: "Einhebel-Küchenmischer 260 1jet" → "Küchenarmatur".
- "für eine Funktion" statt "für einen Verbraucher".
- "Unterputz-Thermostat" statt "Thermostat Unterputz".
- Jede Sprache eigenständig formulieren, nicht wörtlich übersetzen.

Inhalt:
- Nur Sichtbares beschreiben. Nichts vermuten.
- Nie mit "Bild von", "Foto von" beginnen.
- Keine Bildtyp-Begriffe, Artikelnummern oder Dateinamen.
- Ohne PIM: rein visuell beschreiben, keine Markennamen.

---

AUSGABE (nur gültiges JSON, kein weiterer Text):
{{
  "tags": {{"de": ["..."], "en": ["..."]}},
  "alt_text": {{"de": "...", "en": "..."}}
}}
```

Anhang B – Evaluationsdatensatz

Die 30 Bilder des Evaluationsdatensatzes stammen aus dem produktiven Celum-Bestand. Die Spalte Produktkontext gibt an, ob zum Asset Artikelnummern hinterlegt sind: „Fokusprodukt“ bedeutet, dass zusätzlich das Hauptmotiv über das Fokus-Artikelfeld bezeichnet ist, „nur sichtbare Produkte“, dass ausschließlich das Feld für alle sichtbaren Produkte belegt ist, „kein Kontext“, dass keine Verknüpfung besteht. Die Bildkategorie wurde nachträglich anhand des Motivs vergeben.

Tabelle 13: Evaluationsdatensatz mit Produktkontext (n = 30)

Quelle: Eigene Darstellung

Asset-ID	Dateiname	Bildkategorie	Produktkontext	Anzahl Produkte	Marke
2330	apb00961.tif	Freisteller	nur sichtbare Produkte	9	AXOR
2606	Werkplaats-113.jpg	Ambiente	kein Kontext	–	–
115303	apu80363.tif	Freisteller	nur sichtbare Produkte	1	AXOR
430933	apr03896.tif	Freisteller	nur sichtbare Produkte	1	AXOR
434008	har01695.tif	Ambiente	Fokusprodukt	10	hansgrohe
511135	har03822.tif	Lifestyle	nur sichtbare Produkte	2	hansgrohe
516703	hnr01333.jpg	Marketingmotiv	nur sichtbare Produkte	3	hansgrohe
526265	_DSC1594.jpg	Marketingmotiv	kein Kontext	–	–
665693	apr11609.tif	Freisteller	nur sichtbare Produkte	1	AXOR
667111	axw_novelties_axor_citterio_c_cold-start_1500x2000.png	Marketingmotiv	kein Kontext	–	–
693998	har11132.tif	Lifestyle	Fokusprodukt	3	hansgrohe
697824	har11428.tif	Detailaufnahme	Fokusprodukt	6	hansgrohe
700346	har11748.jpg	Detailaufnahme	nur sichtbare Produkte	1	hansgrohe
702350	har11852.jpg	Detailaufnahme	nur sichtbare Produkte	2	hansgrohe
710768	har12048.jpg	Lifestyle	nur sichtbare Produkte	10	hansgrohe
720714	pma_axor_novelties_2025_axor_bidette_showers_2160x1620_02.jpg	Ambiente	kein Kontext	–	–

Asset-ID	Dateiname	Bildkategorie	Produktkontext	Anzahl Produkte	Marke
730100	har12998.tif	Ambiente	Fokusprodukt	1	hansgrohe
737002	hpr09898.jpg	Marketingmotiv	nur sichtbare Produkte	1	hansgrohe
746815	har13984.tif	Lifestyle	Fokusprodukt	2	hansgrohe
747358	hpr10438.tif	Freisteller	nur sichtbare Produkte	1	hansgrohe
748938	har14105.tif	Ambiente	kein Kontext	–	–
750937	har14171.tif	Ambiente	Fokusprodukt	4	hansgrohe
752487	har14305.tif	Ambiente	Fokusprodukt	8	hansgrohe
756306	har14555.tif	Lifestyle	kein Kontext	–	–
756693	har14583.tif	Detailaufnahme	Fokusprodukt	1	hansgrohe
757371	aad02119.jpg	Ambiente	kein Kontext	–	–
757374	aad02113.jpg	Detailaufnahme	kein Kontext	–	–
762027	har14641.tif	Detailaufnahme	Fokusprodukt	2	hansgrohe
764702	har14752.tif	Ambiente	Fokusprodukt	10	hansgrohe
770102	hpr10966.tif	Freisteller	nur sichtbare Produkte	1	–

Anhang C – Modell- und Laufzeitkonfiguration

Die folgende Tabelle dokumentiert die verwendeten Modellkennungen und die verfügbaren Angaben zur Ausführung. Die exakten Quantisierungsvarianten konnten aufgrund fehlender Daten nicht durchgängig protokolliert werden.

Tabelle 14: Modell- und Laufzeitkonfiguration der evaluierten und eingesetzten Varianten

Quelle: Eigene Darstellung

Konfiguration	**Repository / Modell-ID**	**Quantisierung**	**Runtime und Version**
Gemma4:31b	gemma4:31b (Ollama-Registry)	GGUF, 4 Bit Variante nicht protokolliert	Ollama v0.32.15
Gemma4:e4b	gemma4:e4b (Ollama-Registry)	GGUF, 4 Bit Variante nicht protokolliert	Ollama v0.32.15
Qwen3.8-27B-MLX	qwen3.8:27b-mlx (Ollama-Registry; Modellkennung)	4 Bit genauer Build nicht dokumentiert	Ollama v0.32.15
Claude Sonnet 5	AWS Bedrock Inference Profile (eu-central-1)	nicht offengelegt	AWS Converse API
Produktivbetrieb	google.gemma-4-31b (AWS Bedrock Mantle, eu-central-1)	nicht offengelegt	OpenAI-kompatible API Bedrock Mantle

Die lokalen Modelle wurden auf einem Mac Studio M2 Ultra mit 128 GB Unified Memory ausgeführt. Für die Cloud-Ausführung wurde die Region eu-central-1 verwendet.

Anhang D – Freitextkommentare mit Kategorisierung

Die Tabelle enthält alle 77 abgegebenen Freitextkommentare im Wortlaut. Die Buchstaben der kommentierenden Personen sind unabhängig von den Funktionskürzeln in Tabelle 5. Die Zuordnung folgt dem in Abschnitt 6.4 beschriebenen

Kodiersystem. Sieben Kommentare zur fehlenden Beurteilbarkeit und eine unbegründete Präferenz gehen nicht in die Auswertung der 69 fachlichen Kommentare ein. Die Originalschreibweisen bleiben erhalten.

Tabelle 15: Freitextkommentare der Evaluation mit zugewiesener Kategorie (n = 77)

Quelle: Eigene Darstellung; Kommentare im Wortlaut aus der Erhebung

Kodierlegende: Jede Beanstandung ist als Buchstabe mit Ausgabedimension und Art notiert, etwa I (Tags: fehlt); mehrere Zuordnungen je Kommentar sind durch Komma getrennt. Buchstaben: I = Identifikation (Marke, Linie, Produkt); B = Bildinhalt/Suchbegriffe; T = Terminologie; O = Oberfläche/Farbe; Y = Bildtyp; F = Funktionales Merkmal; S = Alt-Text-Länge/Formulierung; K = ungewöhnliche Wortbildung. Dimension: Alt = Bildbeschreibung, Tags = Such-Tags, ? = Bezug im Wortlaut nicht eindeutig. Art: „fehlt“ = erwartete Angabe fehlt; „falsch“ = vorhandene Angabe wird als unzutreffend beanstandet; „beanstandet“ = Formulierung wird zurückgewiesen; „Wunsch“ = zusätzliche Angabe wird angeregt. Alle Angaben geben das Urteil der Fachperson aus der Umfrage wieder, keine unabhängige Fehlerverifikation. * = Identifikationswunsch für Tags im Konflikt mit FA3. Tabelle 9 zählt jeden Kommentar je Buchstabe einmal, auch wenn er dort mehrfach vorkommt. Rohkommentare bleiben unverändert; Kommentar 22 endet bereits im Export unvollständig.

Nr.	P.	Asset-ID	Kategorie	Kommentar
1	A	2330	fehlende Beurteilbarkeit	Axor kann ich nicht beurteilen
2	A	757374	fehlende Beurteilbarkeit	Axor kann ich nicht beurteilen
3	A	115303	fehlende Beurteilbarkeit	Axor kann ich nicht beurteilen
4	A	667111	fehlende Beurteilbarkeit	Axor kann ich nicht beurteilen
5	A	720714	fehlende Beurteilbarkeit	Axor kann ich nicht beurteilen
6	A	430933	fehlende Beurteilbarkeit	Axor kann ich nicht beurteilen
7	A	434008	fehlende Beurteilbarkeit	Axor kann ich nicht beurteilen
8	B	747358	T (Alt: Wunsch)	Beim Alt-Text am ehesten eine Mischung aus Variante C und D. D wegen der detaillierten Beschreibung und C wegen dem Wort Waschbecken. Tischmontage sagt niemand. Anstatt Waschtischmischer sollten wir auch eher Waschtischarmatur nehmen, weil die wenigsten Endkunden solche technischen Begriffe benutzen.

Nr.	P.	Asset-ID	Kategorie	Kommentar
9	B	770102	T (Alt: falsch)	Ab- und Umstellventil ist der völlig falsche Begriff für dieses Produkt und somit auch für den Alt-Text. Wenn ein Screenreader das einer blinden Person vorliest, dann ist im worst Case genauso schlau wie davor.
10	B	756693	T (?: Wunsch)	Duschthermostat anstatt Brausethermostat
11	B	746815	T (Alt: Wunsch)	Küchenarmatur wäre besser für den Alt-Text. Küchenmischer wird nicht nachgefragt und ist zu technisch. In der Übersetzung wurde das in der Vergangenheit sogar zu kitchen mixer, was ja offensichtlich was ganz anderes ist.
12	B	516703	K (?: beanstandet); B (Alt: fehlt)	Kopfbrausenkopf und Duschpipe sind keine richtigen Begriffe. Außerdem fehlt bei allen Alt-Texten die Texte auf dem Bild.
13	B	430933	T (?: Wunsch)	Waschtischmischer ist ein suboptimaler Begriff
14	B	737002	T (?: fehlt)	Wasserhahn oder Waschtischarmatur fehlt überall als zentraler Begriff
15	C	748938	I (?: fehlt); O (?: fehlt); T (?: falsch)	es fehlen die konkreten Produktlinien und auch bei den Farben die spezifischen Farben (unsere Namen - z.B. sandbeige matt), um wirklich perfekt zu sein. Außerdem ist es Avalegra keine Wandarmatur.
16	C	770102	T (?: Wunsch); Y (?: fehlt)	Auch hier wäre es am besten, wenn es unser Wording verwenden würde. Rosette statt Blende. Iso-Rendering.
17	C	746815	I (Tags: fehlt)*; Y (Tags: fehlt); O (?: falsch)	Im Text wird die Linie genannt, bei den Tags nicht mehr. Es fehtl der Bildsubtyp im Tag - Teilambiente mit Person. Die Spüle ist nicht mattschwarz - inhaltlich also falsch
18	C	697824	B (Tags: fehlt); B (Alt: fehlt)	Bei den Tags fehlen Begriffe zur Reinigung. Dafür wird das Bild verwendet und sollte im Fließtext besser zum Ausdruck kommen.
19	C	693998	I (?: fehlt); T (?: fehlt)	Variante A qäre am besten, aber Raindance Alive als Linie fehlt. Auch was wie Aufputzshowerpipe.
20	C	434008	O (?: fehlt); O (?: falsch)	hier würde das wichtigste Merkmal brushed black chrome überall fehlen, teilw. wird es fehlerhaft als mattschwarz ausgeleitet
21	C	752487	O (?: fehlt); I (?: fehlt); B (Tags: fehlt)	richtige Oberflächenbezeichnung, Linienzuordnung fehlt. Waschplatz als Tag fehlt. Inneneinrichtungsstil: Minimalisitsch, Skandi fehlt.
22	C	737002	I (?: fehlt); Y (Tags: fehlt); Y (?: falsch)	Linie fehlt bei D - das ist keine echte Verpackung, sondern ein eCom Bild. Daher fehlt eCom hier in den Tags komplett und es wird irreführend. Packshot ist nicht eindeutig von den Tags her - wir nutzen bei hansgrohe eher Freisteller, Produktrendering, diese Tags müssten kommen. Das ist ein e
23	C	756693	F (?: fehlt)	Universal Fit fehlt als Merkmal. Alle schläuche passen zum THM.
24	C	511135	I (?: fehlt); B (?: fehlt)	Linie fehlt. Stil des Bades fehlt.
25	C	667111	F (?: fehlt)	CoolStart ist hier zwingend nötig.
26	C	747358	O (?: Wunsch); I (Tags: fehlt)*; Y (?: fehlt)	Brushed Nickel wird auch in Deutsch so gesucht und nicht gebürstetes Nickel, daher keine Übersetzung von Namen. Linie fehtl in den Tags. Bildart Freisteller fehlt.
27	C	756306	B (?: falsch)	A ist noch am nächsten. Das ist ein Designstudio - Atelier 522. Der Kollege ist allerdings ein Mann.

Nr.	P.	Asset-ID	Kategorie	Kommentar
28	C	764702	I (?: fehlt); Y (?: fehlt)	Linie fehlt. Bildart fehlt: Badambiente ohne Person.
29	C	516703	Y (?: fehlt)	es fehlt Angabe eCom Bild
30	C	2606	B (?: fehlt)	Kartusche fehlt auf jeden Fall
31	C	702350	F (?: fehlt)	Im Idealfall würde die Strahlart: PowderRain noch genannt werden.
32	C	710768	I (?: fehlt); F (?: fehlt); Y (?: fehlt)	Linie fehlt. Strahlart fehlt. Bildart fehtl.
33	D	516703	B (?: fehlt)	die USPs auf dem Bild werden nicht berücksichtigt
34	D	115303	I (Alt: fehlt)	Text in Variante C mit Produktname wäre besser
35	D	700346	B (Tags: fehlt)	bei den Tags wäre die Form "rund" noch gut
36	D	746815	B (Alt: fehlt)	Text der Variante A evtl. etwas ausführlicher (Blumen liegen auf der Ablage)
37	E	757374	S (Alt: beanstandet)	Die Alt-Text Vorschläge sind zu lang.
38	E	720714	S (Alt: beanstandet)	Auch hier wieder - viel zu lange Alt-Texte!!!
39	E	667111	O (Alt: falsch)	Zum Alt-Tag: Ich hätte im ersten Schritt die Option A abgelehnt, weil die Oberfläche nicht richtig beschrieben ist. Das ist kein Silber. Ansonsten die einzig passende Option.
40	E	757371	S (Alt: beanstandet)	Alt-Text wieder viel zu lang. Inhaltlich am ehesten Variante D
41	E	697824	B (Alt: fehlt)	Es geht um das Entkalken. Das fehlt bei den Alt-Text Vorschlägen.
42	E	702350	B (?: falsch)	Bezeichnung der Form ist falsch.
43	E	756306	B (Alt: falsch)	Zum Alt-Text: Am ehesten noch Variante C - allerdings erkenne ich keine Beine. Für mich eindeutig keine Frau.
44	E	434008	T (?: Wunsch); S (Alt: beanstandet)	Industriestil? Benutzt man diese Bezeichnung? Am ehesten noch Variante D. Aber alles viel zu lang.
45	E	693998	Präferenz ohne fachliche Begründung	Am ehesten noch B.
46	E	516703	S (Alt: beanstandet)	Alt-Text: Am ehesten noch A . Irritierend finde ich "wobei aus der Kopfbrause Wasser fließt." Vor allem das "wobei" suggeriert hier, dass etwas pessiert, was nicht passieren sollte.
47	E	430933	B (?: fehlt)	Rundbogen als Begriff fehlt mir.
48	E	511135	S (Alt: beanstandet); O (?: Wunsch)	Alt-Text sind viel zu lang. Ich frage mich, ob man von silberfarbenen Armaturen spricht. Bringt das nicht jeder mit Chrom in Verbindung?
49	F	516703	Y (Tags: fehlt)	Hätte hier Schlagwörter zu der Art des Bildes erwartet. Also dass es informativ ist, mit Icons drauf, etc.
50	F	115303	I (?: fehlt); B (?: fehlt)	Kombination aus AXOR, 1-loch und oval fehlt mir
51	F	762027	B (Tags: fehlt)	wenn Produkte zu sehen sind, würde ich immer erwarten, dass ich sie auch in den Schlagwörtern sehen würde
52	F	430933	I (Tags: fehlt)*	Hier muss meiner Meinung nach zwingend der Artikelname oder Nummer in den Schlagworten sein
53	G	757374	O (?: falsch); I (?: fehlt)	Falsche Widergabe der Oberfläche. Ist weder silber noch chrom. Kann man hier den Namen der Kollektion und im besten Fall auch die Marke irgendwie integrieren?
54	G	516703	B (?: fehlt)	Es fehlt die Wiedergabe der Layout-Elemente und Logos

Nr.	P.	Asset-ID	Kategorie	Kommentar
55	H	700346	I (Tags: fehlt)*; T (Alt: Wunsch); I (Alt: fehlt); T (Alt: falsch)	Mir fehlt bei den Tags der Linienname und die Marke. Bei der Beschreibung ist Variante D schon relativ gut. Allerdings würde ich nicht von "1 Verbraucher" sprechen, sondern stattdessen "für 1 Funktion" verwenden wollen. Auch hansgrohe als Marke könnte ergänzt werden. Und statt "Ein Thermostat Unterputz" müsste es richtig lauten: "Ein Unterputz-Thermostat". Außerdem ist der Begriff "Duschsystem" im gezeigten Kontext nicht zutreffend - es handelt sich nicht um eine Showerpipe, sondern um eine Handbrause (und die ist kein Duschsystem).
56	H	737002	F (?: falsch); I (?: fehlt); T (?: falsch)	Variante A, D: CoolStart ist nicht richtig? Variante B sollte noch Marke und Linie enthalten. Die Begriffe "Waschbecken", "Badmöbel", "Waschtisch" sind hier irreführend.
57	H	693998	I (Alt: fehlt); F (Tags: falsch); B (Tags: falsch); I (Tags: fehlt)*	Bei der Beschreibung fehlen Marke und Linie. Variante B wäre noch die beste Option. Bei den Tags sind "EcoSmart", "3jet", "Duschwand" nicht zutreffend. Es fehlen ebenfalls Marke und Linie.
58	H	750937	K (Alt: beanstandet); B (Tags: fehlt)	Beschreibung D wäre gut, aber den Begriff "unterputzig" gibt es schlicht nicht. Tags: "Hotelbad" oder "GP"/"Global Projects" wären noch wichtige Begriffe zur Beschreibung dieses Bildes.
59	H	746815	I (Tags: fehlt)*	Bei den Tags fehlen mir Marke und Linie. Variante B käme aber einer Lösung am Nächsten.
60	H	747358	Y (?: falsch); O (?: falsch); Y (Tags: fehlt)	"Packshot" ist inhaltlich falsch. "silberfarben" irreführend. Zu Variante C (Tags) würde ich noch den Hinweis "Freisteller" oder "Cut-out" ergänzen.
61	H	526265	B (?: falsch)	"Mikrofon" ist falsch identifiziert.
62	H	702350	I (Tags: fehlt)*; S (Alt: beanstandet); B (Alt: beanstandet)	Bei Variante D (Tags) fehlen mir noch Marke und Linie. Die Beschreibungen sind alle nicht wirklich brauchbar, da z. T. stilistisch oder inhaltlich ungeeignet.
63	H	430933	O (Tags: falsch); I (Tags: fehlt)*	Tags: Messing und goldfarben sind zwei verschiedene Farbtöne - nicht mischen. Marke und Linie fehlen.
64	H	757374	K (?: beanstandet)	"marmorkonterplatte" ist ein unbekannter Begriff.
65	H	730100	K (Tags: beanstandet)	Variante B (Tags): "keramiktoilette" ist ein ungewöhnlicher Begriff.
66	H	115303	I (Alt: fehlt); K (Tags: beanstandet)	Variante C (Beschreibung": Marke und Linie fehlen. Variante C (Tags): "waschbeckenunterschale" gibt es nicht.
67	H	770102	T (Tags: falsch)	Die Tags enthalten z. T. falsche und irreführende Begriffe.
68	H	762027	T (?: falsch); I (Tags: fehlt)*	"Auszug" und "Auszugsschlauch" sind inhaltlich falsch. Bei den Tags fehlen Marke und Linie.
69	H	516703	I (?: fehlt); Y (Tags: fehlt)	Marke und Linie fehlen. Zudem sollte ein Hinweis in die Tags aufgenommen werden, dass es sich um ein eComm-Bild handelt.
70	H	511135	B (Alt: falsch); I (Tags: fehlt)*	Variante D (Beschreibung): Die Wanne wird nicht mit Schaum gefüllt. :D Tags: Marke und Linie fehlen.
71	H	720714	I (?: fehlt)	Marke und Linie fehlen.
72	H	667111	F (Alt: fehlt); F (Tags: fehlt); I (?: fehlt)	Beim Motiv geht es um die Funktionsweise von CoolStart. Der Begriff sollte in Beschreibung und Tag vorkommen. Zudem fehlen Marke und Linie.

Nr.	P.	Asset-ID	Kategorie	Kommentar
73	H	752487	B (Alt: fehlt); O (Alt: falsch); T (Tags: beanstandet); I (Tags: fehlt)*	Beschreibung: Der minimalistische Charakter kommt nicht zum Ausdruck. Die Konsolenplatte ist weiß, nicht schwarz (schwarz ist der Griff). Tags: "schwebender Waschtisch" ist Unsinn. Das Möbel ist an der Wand befestigt. Es fehlen Marke und Linie.
74	H	665693	I (?: fehlt)	Es fehlen Marke und Linie.
75	H	2606	B (?: falsch)	"Qualitätskontrolle" erscheint mir nicht zutreffend.
76	H	756693	F (Tags: fehlt); I (?: fehlt)	Bei den Tags müsste auf die Installation bzw. die universelle Kompatibilität zu allen gängigen Brauseschläcuhen ("Universal-Fit") hingewiesen werden. Es fehlen Marke und Linie.
77	H	710768	I (?: fehlt); B (Tags: beanstandet)	Es fehlen Marke und Linie. Die Tags sind zu unspezifisch, um das Bild in Celum zu finden (da gäbe es zu viele Treffer).

Anhang E – Beispielausgaben der vier Modelle

Die folgenden drei Beispiele zeigen die tatsächlichen Ausgaben aller vier Modelle für Bilder mit unterschiedlichem Produktkontext. Wiedergegeben sind die deutschen Such-Tags in normalisierter Form sowie die Alternativtexte in beiden Sprachen. Die englischen Tags und die vollständigen Roh-Antworten liegen im Umfrage-Verzeichnis (Anhang F).

Beispiel 1 – Asset 746815 (har13984.tif), Fokusprodukt hinterlegt

PIM-Kontext: Fokus: hansgrohe Vernis Blend M35, Einhebel-Kuechenmischer 260, Mattschwarz



Tabelle 16: Modellausgaben für Asset 746815 (har13984.tif), Prompt V3

Quelle: Eigene Darstellung; Ausgaben der evaluierten Modelle

Modell	Such-Tags (Deutsch, normalisiert)	Alternativtext (Deutsch)	Alternativtext (Englisch)
Gemma4:31b	einhebel-küchenmischer; küchenarmatur; küchenmischer; mattschwarz; schwarz; küche; lifestyle; einbauspüle; holzarbeitsplatte; glasvase; wasser; blumen; armatur; spüle; einhebelmischer	Einhebel-Küchenmischer Vernis Blend M35 in Mattschwarz an einer schwarzen Spüle auf einer Holzarbeitsplatte, während eine Person eine Glasvase mit Wasser füllt.	Vernis Blend M35 single lever kitchen mixer in matte black on a black sink and wooden countertop, with a person filling a glass vase with water.

Modell	Such-Tags (Deutsch, normalisiert)	Alternativtext (Deutsch)	Alternativtext (Englisch)
Claude Sonnet 5	küchenarmatur; einhebelmischer; küchenmischer; mattschwarz; spültischarmatur; küche; spüle; arbeitsplatte holz; wasserhahn küche; person; glas spülen; fließendes wasser; ambiente; küchenzeile; blumenstrauß; küchenutensilien; hände; alltagsszene; nutzung; wohnküche	Ein Einhebel-Küchenmischer in Mattschwarz ist an einer dunklen Spüle mit heller Holzarbeitsplatte montiert. Eine Person spült unter dem laufenden Wasser ein geripptes Glas, im Vordergrund liegt ein bunter Blumenstrauß auf der Arbeitsplatte.	A matte black single-lever kitchen mixer is mounted at a dark sink with a light wooden countertop. A person rinses a ribbed glass under running water, with a colorful bouquet of flowers lying on the counter in the foreground.
Qwen3.8-27B-MLX	küchenarmatur; küchenmischer; einhebelmischer; waschtischarmatur; spülbeckenarmatur; matt schwarz; küche; holz-arbeitsplatte; spüle; aufsatzarmatur; geschirr spülen; glas spülen; wasser zapfen; person; handwäsche; keramik; blumen; moderner stil; natürliche deko; lifestyle	Eine Person in einem hellen Hemd spült ein Glas unter einem mattschwarzen Einhebel-Küchenmischer, der auf einer Holz-Arbeitsplatte über einem dunklen Spülbecken sitzt; im Vordergrund liegen Blumen, daneben stehen cremefarbene Keramik-Flaschen.	A person in a light shirt rinses a glass under a matte black single-lever kitchen mixer mounted on a wooden countertop above a dark sink; in the foreground lie flowers, with cream-colored ceramic bottles beside them.
Gemma4:e4b	waschtischarmatur; einhebelmischer; badarmatur; mattschwarz; badezimmer; waschtisch; sanitär; bad; ambiente; mischer; badmöbel; armaturen; badgestaltung; wasser; lifestyle	Ein Einhebel-Mischer in Mattschwarz ist an einem Waschtisch montiert. Eine Person benutzt die Armatur, um Hände zu waschen. Die Szene zeigt ein Badezimmerambiente mit Blumen und Accessoires.	A matte black single-lever mixer is mounted on a washbasin. A person is using the faucet to wash hands. The scene shows a bathroom environment with flowers and accessories.

Beispiel 2 – Asset 115303 (apu80363.tif), nur sichtbare Produkte hinterlegt

PIM-Kontext: Weitere Produkte: AXOR Massaud, Einbauwaschtisch 585/469 mit 1-Lochbohrung, Weiss



Tabelle 17: Modellausgaben für Asset 115303 (apu80363.tif), Prompt V3

Quelle: Eigene Darstellung; Ausgaben der evaluierten Modelle

Modell	**Such-Tags (Deutsch, normalisiert)**	**Alternativtext (Deutsch)**	**Alternativtext (Englisch)**
Gemma4:31b	waschtisch; waschbecken; einbauwaschtisch; weiß; badezimmer; bad; 1-lochbohrung; oval	Weißer Einbauwaschtisch AXOR Massaud mit einer Lochbohrung für das Badezimmer.	White AXOR Massaud built-in washbasin with one tap hole for the bathroom.

Modell	Such-Tags (Deutsch, normalisiert)	Alternativtext (Deutsch)	Alternativtext (Englisch)
Claude Sonnet 5	einbauwaschtisch; waschbecken; waschtisch; weiß; keramik; 1-lochbohrung; ovale form; badezimmer; einbauwaschbecken; aufsatzwaschbecken; sanitärkeramik; asymmetrische form; glänzende oberfläche	Einbauwaschtisch in Weiß mit asymmetrischer, ovaler Form und einer Lochbohrung für die Armatur.	Built-in washbasin in white with an asymmetric oval shape and a single hole for the faucet.
Qwen3.8-27B-MLX	einbauwaschtisch; waschtisch; waschbecken; 1-lochbohrung; armaturbohrung; weiß; glänzend; keramik; oval; organische form; bad; badezimmer; einbau; sanitär; modern; minimalistisch; produktfoto; waschtischbecken; badmöbel	Weißer Einbauwaschtisch mit 1-Lochbohrung in ovaler, organischer Form und glänzender Oberfläche.	White built-in washbasin with single-hole drilling in an oval, organic shape and glossy surface.
Gemma4:e4b	waschbecken; waschtisch; einbauwaschtisch; axor massaud; weiß; bad; sanitär; waschbeckenarmatur; waschtischmischer; keramik; badezimmer; waschbeckenunterschale	Das Waschbecken der Linie AXOR Massaud ist ein weißes, rechteckiges Einbauwaschtisch. Es ist aus Keramik gefertigt und für den Einsatz im Badezimmer konzipiert.	The AXOR Massaud washbasin is a white, rectangular built-in sink. It is made of ceramic and designed for bathroom use.

Beispiel 3 – Asset 756306 (har14555.tif), kein Produktkontext

PIM-Kontext: kein Produktkontext



Tabelle 18: Modellausgaben für Asset 756306 (har14555.tif), Prompt V3

Quelle: Eigene Darstellung; Ausgaben der evaluierten Modelle

Modell	**Such-Tags (Deutsch, normalisiert)**	**Alternativtext (Deutsch)**	**Alternativtext (Englisch)**
Gemma4:31b	lifestyle; ambiente; büro; frau; planungsphase; moodboard; designprozess; glaswand; innenraum; modern; arbeitszimmer; inspiration; fotowand; gestaltung; graues hemd; jeans	Eine Frau in grauem Hemd und Jeans arrangiert kleine Bilder an einer Glaswand in einem hellen, modernen Büro.	A woman in a grey shirt and jeans arranges small images on a glass wall in a bright, modern office.
Claude Sonnet 5	büroraum; glaswand; person; moodboard; bilder anpinnen; arbeitsplatz; vorhang; aktenordner; schreibtisch; bürostuhl; fenster; tageslicht; innenraum; kreativarbeit; konzeption; regal; modernes büro; ambiente; arbeitsalltag; bildersammlung	Eine Person steht in einem Büro an einer Glaswand und befestigt mehrere Bildausdrucke daran. Im Hintergrund sind ein Schreibtisch, Aktenordner und ein großes Fenster mit Vorhang zu sehen.	A person stands at a glass wall in an office, attaching several printed images to it. In the background, a desk, binders, and a large window with a curtain are visible.

Modell	Such-Tags (Deutsch, normalisiert)	Alternativtext (Deutsch)	Alternativtext (Englisch)
Qwen3.8-27B-MLX	büro; arbeitsraum; designstudio; kreativbüro; person; frau; moodboard; glaswand; innenarchitektur; raumplanung; fotografie; aufhängen; tageslicht; fenster; schreibtisch; monitor; lampe; aktenschränke; planung; kollaboration	Eine Person in Jacke und Jeans hängt im Büro Fotos an einer Glaswand und gestaltet ein Moodboard, im Vordergrund ruhen entspannt Beine, rechts steht ein Schreibtisch mit Monitor bei Tageslicht.	A person in a jacket and jeans pins photos to a glass wall in an office to create a mood board, with legs resting in the foreground and a desk with a monitor lit by daylight on the right.
Gemma4:e4b	badezimmer; badgestaltung; showroom; produktberatung; badmöbel; armaturen; badzimmer; inneneinrichtung; materialauswahl; modernes bad; badrenovierung; badaccessoires; badbedarf; beratung; interieur; bad; ausstellung	Ein Mann betrachtet Produktmuster und Ausstellungsboards in einem hellen, modernen Showroom. Der Raum zeigt minimalistisches Design und verschiedene Materialoptionen für Badarmaturen und Badmöbel.	A man examines product swatches and display boards in a bright, modern showroom. The room features minimalist design and various material options for bathroom fittings and furniture.

Anhang F – Rohdaten, Auswertung und Artefakt

Der digitale Anhang enthält den eingesetzten Quellcode, die Prompt-Dateien, die Umfrage mit Ergebnisexporten sowie die zitierte Literatur. Die Originalartefakte bleiben unverändert.

Quellcode/celum_ai_image_tagging enthält die eingesetzte Python-Pipeline mit Konfiguration, Prompt-Dateien und begleitenden Skripten. Der Aufbau und die Verarbeitungsschritte sind in Kapitel 5 beschrieben.

Umfrage enthält die Evaluationsbilder, die Ausgaben der vier Modelle, die ausgelieferte HTML-Umfrage, die zwölf Antwortexporte sowie die ursprünglichen Auswertungsskripte samt ihren Ausgaben.

Die Protokollauswertung des produktiven Erstlaufs dokumentiert die betrieblichen Kennzahlen. Abschnitt 7.1 ordnet die Jobstatus, gespeicherten Nutzungsdaten und Tag-Collection-Zuordnungen in den Untersuchungszusammenhang ein.

Zitierte Quellen enthält sämtliche im Literaturverzeichnis aufgeführten Referenzen als PDF- oder Textdatei. Der Dateiname entspricht dem Kurzbeleg ohne Sonderzeichen, bestehend aus dem Familiennamen der erstgenannten Person und dem Erscheinungsjahr, etwa Radford2021.pdf. Seitenangaben in den Belegen beziehen sich auf die hier abgelegte Fassung der jeweiligen Quelle.

Selbstständigkeitserklärung und Erklärung zum Nachweis der Verwendung von KI-Systemen bei Erstellung einer wissenschaftlichen Arbeit

Klemm, Paul	6411851
Nachname, Vorname	Matrikelnummer

Ich versichere hiermit, dass ich die vorliegende Arbeit mit dem Thema „KI-gestützte Bildverschlagwortung und Alt-Text-Generierung nach Unternehmenstaxonomien – Entwurf, Implementierung und Evaluation am Beispiel eines DAM-Systems“ selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Ich versichere zudem, dass die eingereichte elektronische Fassung mit der gedruckten Fassung übereinstimmt.

Ich erkläre zudem, dass ich *(Zutreffendes ankreuzen*)

☐ Generative Systeme (Text-, Bild-, Codeerstellung etc.), wie beispielsweise den textbasierten Chatbot ChatGPT von OpenAI, bei Erstellung der vorliegenden wissenschaftlichen Arbeit **nicht** verwendet habe.

☒ Generative Systeme (Text-, Bild-, Codeerstellung etc.) bei Erstellung der vorliegenden wissenschaftlichen Arbeit in der nachfolgend beschriebenen Art und dem nachfolgend beschriebenen Umfang verwendet habe. Auf Wunsch des Betreuers wurden die verwendeten Prompt-Anfragen im nachfolgenden Verzeichnis zusammengefasst aufgeführt.

Alpirsbach, 30.09.2026	
Ort, Datum	Unterschrift der/des Studierenden

Art und Umfang der Verwendung generativer Systeme

Generative Systeme unterstützten Strukturierung, Formulierung, Quellenprüfung, Softwareentwicklung und die Durchsicht der Arbeit. Bei der abschließenden Überarbeitung wurden sie außerdem für die Prüfung vorhandener Auswertungen und die Zuordnung der Freitextkommentare verwendet. Die Erhebungsdaten und die eingesetzten Software- und Prompt-Artefakte wurden dabei nicht verändert. Die fachliche Verantwortung verbleibt beim Verfasser.

Die folgende Übersicht beschreibt die Art der Verwendung anhand sinngemäß zusammengefasster Prompt-Typen. Sie ist kein vollständiges Gesprächs- oder Aufrufprotokoll. Überlieferte Produkt-, Modell- und Häufigkeitsangaben werden nur im dokumentierten Umfang angegeben. Nicht vollständig rekonstruierbare Details werden nicht ergänzt. HansAI bezeichnet eine unternehmensinterne Plattform ohne öffentlichen Zugang. Die zusätzliche KI-gestützte Überarbeitung ist gesondert aufgeführt.

Verzeichnis der Abfragen

Gleichartige Abfragen sind sinngemäß zusammengefasst; […] steht für eingefügte Texte, Daten oder Dokumentteile. Häufigkeiten entsprechen der vorhandenen Dokumentation und sind keine lückenlos verifizierte Nutzungsstatistik. Bei fortlaufend verwendeten Werkzeugen fehlen Einzelprotokolle. Doppelte gleichlautende Einträge wurden als ein Prompt-Typ zusammengeführt.

Produktname	URL	Modell/Einstellung	Datum der Abfrage	Zusammengefasster Prompt	Häufigkeit
ChatGPT	https://chatgpt.com	GPT-4o	06/2026–07/2026	„Erstelle ein Gerüst für meine Bachelorarbeit im Studiengang Onlinemedien und sammle unter jedem Kapitel stichpunktartig die vorgesehenen Inhalte: […]“	5
ChatGPT	https://chatgpt.com	GPT-4o	06/2026–07/2026	„Erstelle aus den folgenden Angaben ein Exposé zur Arbeit und überarbeite es anschließend anhand des Betreuerfeedbacks: […]“	5
ChatGPT	https://chatgpt.com	GPT-4o	07/2026	„Formuliere Kapitel […] auf Basis der folgenden Stichpunkte wissenschaftlich aus und beachte den vorgegebenen Seitenumfang: […]“	4
ChatGPT	https://chatgpt.com	GPT-5.6, Modus Sol	09/2026	„Bewerte diese Bachelorarbeit als unparteiischer Gutachter anhand der Anforderungen des Betreuers und der Richtlinien der Hochschule, prüfe Aussagen, Logik, Zitate und Belege und erstelle ein Gutachten mit Note: […]“	2

Produktname	URL	Modell/Einstellung	Datum der Abfrage	Zusammengefasster Prompt	Häufigkeit
HansAI	kein öffentlicher Zugang	GPT-5.6, Modus Luna	07/2026	„Welche Varianten der Gemma-4-Familie eignen sich für einen Modellvergleich und wofür steht das Präfix E in der Modellbezeichnung?“	3
HansAI	kein öffentlicher Zugang	Claude 5 Opus, Reasoning high	07/2026–08/2026	„Bewerte den folgenden Prompt zur Bildverschlagwortung hinsichtlich Vollständigkeit und Token-Effizienz und vergleiche ihn mit der folgenden Alternative: […]“	3
HansAI	kein öffentlicher Zugang	Claude 5 Opus, Reasoning high	07/2026–08/2026	„Recherchiere, welche Vision Language Models sich für Verschlagwortung und Alt-Text-Generierung bei begrenzter lokaler Rechenleistung vergleichen lassen: […]“	3
HansAI	kein öffentlicher Zugang	GPT-5.6, Modus Terra	08/2026	„Erstelle eine offline nutzbare HTML-Umfrage, die je Bild vier anonymisierte Varianten zur Auswahl stellt und die Antworten als JSON-Datei ausgibt, sowie das zugehörige Auswertungsskript: […]“	6
HansAI	kein öffentlicher Zugang	GPT-5.6, Modus Terra	08/2026	„Prüfe, ob das beschriebene Erhebungsdesign für eine wissenschaftliche Auswertung ausreicht: […]“	1
HansAI	kein öffentlicher Zugang	GPT-5.6, Modus Terra	08/2026	„Formuliere die Frage zur Präferenzentscheidung so kurz wie möglich und so ausführlich wie nötig: […]“	1
Microsoft Copilot	https://copilot.microsoft.com	GPT-5.6, Tiefere Analyse	07/2026	„Recherchiere im Firmenwissen, welche Anträge und Freigaben für das Vorhaben einzuholen sind.“	1
Microsoft Copilot	https://copilot.microsoft.com	GPT-5.6, Tiefere Analyse	07/2026	„Analysiere die folgende Fehlermeldung der lokalen Inferenzumgebung und nenne mögliche Ursachen: […]“	2
Microsoft Copilot	https://copilot.microsoft.com	GPT-5.6, Tiefere Analyse	08/2026	„Formuliere die Frage zur Präferenzentscheidung so kurz wie möglich und so ausführlich wie nötig: […]“	1
Microsoft Copilot	https://copilot.microsoft.com	GPT-5.6, Tiefere Analyse	08/2026	„Formuliere die Einladung zur Evaluationsumfrage an die Fachexperten auf Basis der folgenden Angaben: […]“	4
Claude Code	https://claude.com/product/claude-code	Modell nicht vollständig dokumentiert	07/2026–08/2026	„Implementiere beziehungsweise überarbeite das folgende Modul der Verarbeitungspipeline: […]“ (Codeassistenz bei der Entwicklung des Artefakts, vgl. Abschnitt 3.2.2)	nicht einzeln protokolliert
GitHub Copilot	https://github.com/features/copilot	Copilot, IDE-Integration	07/2026–08/2026	Kontextbasierte Codevervollständigung ohne eigenständig formulierte Abfragen (vgl. Abschnitt 3.2.2)	entfällt
Claude	https://claude.ai	Sonnet 5, Reasoning medium	08/2026	„Strukturiere die Bachelorarbeit unter Berücksichtigung des folgenden Betreuerfeedbacks und bewerte die vorliegende Gliederung: […]“	1
Claude	https://claude.ai	Opus 4.6, Reasoning high	08/2026	„Strukturiere die Bachelorarbeit unter Berücksichtigung des folgenden Betreuerfeedbacks, schlage eine Kapitelgewichtung vor und prüfe die Abgrenzung zur Vorarbeit: […]“	7
Claude	https://claude.ai	Opus 4.6, Reasoning high	08/2026–09/2026	„Formuliere Kapitel […] auf Basis der folgenden fachlichen Angaben aus und recherchiere ergänzende Quellen: […]“	9
Claude	https://claude.ai	Opus 4.6, Reasoning high	08/2026–09/2026	„Überarbeite den bestehenden Text wissenschaftlich, ohne Ergebnisse oder Quellen zu erfinden, und kennzeichne offene Stellen als Platzhalter: […]“	10
Claude	https://claude.ai	Opus 4.6, Reasoning high	08/2026–09/2026	„Korrigiere die folgenden fachlichen Angaben im Text und ergänze die genannten Informationen: […]“	3

Produktname	URL	Modell/Einstellung	Datum der Abfrage	Zusammengefasster Prompt	Häufigkeit
Claude	https://claude.ai	Opus 4.6, Reasoning high	09/2026	„Prüfe die im Text verwendeten Quellen auf Korrektheit und Nachweisbarkeit: […]“	1
Claude	https://claude.ai	Opus 4.6, Reasoning high	09/2026	„Sichte die folgenden Freitextkommentare der Evaluation und schlage wiederkehrende Beanstandungskategorien vor: […]“ (ursprüngliche Kategorienbildung; überarbeitete Zuordnung siehe Abschnitt 6.4)	2
Claude	https://claude.ai	Opus 5, Reasoning xhigh	09/2026	„Setze die Kritikpunkte des folgenden Gutachtens im Dokument um und bewerte das Ergebnis anschließend erneut unabhängig: […]“	5
Claude	https://claude.ai	Opus 5, Reasoning xhigh	09/2026	„Prüfe das Dokument gegen die formalen Gestaltungsrichtlinien und korrigiere Abbildungs- und Tabellenbeschriftungen: […]“	3
Claude	https://claude.ai	Opus 5, Reasoning xhigh	09/2026	„Suche eine belastbare Quelle für die Aussage in Abschnitt […] und gib den Beleg im Zitationsformat aus.“	3
Claude	https://claude.ai	Opus 5, Reasoning xhigh	09/2026	„Formuliere die Antwort auf die folgende Rückmeldung des Betreuers: […]“	1
Claude	https://claude.ai	Opus 5, Reasoning xhigh	09/2026	„Bewerte diese Bachelorarbeit als unparteiischer Gutachter anhand der beigefügten Dokumente, prüfe Aussagen, Logik, Zitate und Belege, erstelle ein Gutachten mit Note und beschreibe zu jedem Befund die Korrektur: […]“	3
Claude	https://claude.ai	Opus 5, Reasoning xhigh	09/2026	„Bewerte diese Bachelorarbeit als unparteiischer Gutachter anhand der beigefügten Checkliste, prüfe Aussagen, Logik, Zitate und Belege, erstelle ein Gutachten mit Note, beschreibe zu jedem Befund die Korrektur und setze diese anschließend systematisch im Dokument um: […]“	nicht vollständig rekonstruierbar
Claude	https://claude.ai	Opus 5, Reasoning xhigh	09/2026	„Prüfe, wo und wozu statistik.py im Artefakt verwendet wird, und entferne die nicht auffindbare Quelle […] durchgängig aus Text und Literaturverzeichnis.“	nicht vollständig rekonstruierbar
DeepL	https://www.deepl.com	DeepL Write und Übersetzer	07/2026–09/2026	kein Prompt nötig	nicht einzeln protokolliert
OpenAI Codex	https://openai.com/codex/	GPT-6 low; mit Agenten	09/2026	„Prüfe die Arbeit anhand der Anforderungen und korrigiere belegbare Fehler in Text, Quellen und Auswertung. Belasse Code, Umfrage und Prompts unverändert, dokumentiere die KI-Unterstützung und speichere eine neue Fassung.“	fortlaufend; nicht einzeln erfasst