KI-MANAGER-AUSBILDUNG · TAG 5 · RECHERCHEERGEBNIS

KI-Assistent für Versand- und Rückgabefragen

Externe Evidenz zur 60-%-Schwelle · 6 parallele Subagenten · Stand 06.11.2026

KI-Manager Tag 5 — Recherche: KI-Assistent für Versand- und Rückgabefragen

1. Fall und Prüffrage

Vorhaben (Product Goal): Ein Onlineshop plant einen Assistenten für Versand- und Rückgabefragen. Er soll Standardfragen korrekt beantworten und Sonderfälle an Mitarbeitende übergeben. Eigene Pilotdaten fehlen.

Prüfwert vor dem Start: Der Assistent löst mindestens **60 %** der Kundenanfragen korrekt ohne Übergabe. **Metrik:** korrekt ohne Übergabe gelöste Anfragen / alle Anfragen im gleichen Testbestand. Die 60 % sind ein **vorab gesetzter Übungswert**, keine Messung.

Gegenstand dieser Recherche: ausschließlich **externe Evidenz und Übertragbarkeit**. Die tatsächliche Wirkung des eigenen, noch ungetesteten Assistenten ist mit dieser Recherche **nicht** entscheidbar — sie braucht einen Pilotversuch.

2. Ausführungsweg (belegbar)

Prüfpunkt	Befund
Subagenten	6 echte Subagenten, parallel gestartet, je eigener Kontext und eigenes Terminal
Modell	`deepseek/deepseek-v4.1-flash-1017` (Small-Model-Klasse **nicht** separat wählbar; Subagenten erben das Elternmodell)
Webzugriff	ja — Originalseiten selbst geöffnet (`web_extract` war in dieser Umgebung funktionslos, Backend search-only; Zugriff über Browser-Session, `r.jina.ai`-Reader, `curl` + `pdftotext`)
Abschluss	6/6 Aufträge abgeschlossen, 213 API-Aufrufe gesamt, jede Quelle mit Link, Herausgeber, Jahr, Stichprobe, Fundstelle
Einschränkung	Gartner, EUR-Lex, ISO/IEC 42001 blockieren automatisierte Abrufe (403/WAF/Paywall) — betroffene Zahlen über Reproduktionen gegengeprüft und als `[SEKUNDÄR]` / `[OFFEN]` markiert

Die vollständigen Rohrecherchen mit allen Einzelbelegen liegen unter `Rohrecherche/auftrag1.md` … `auftrag6.md`.

3. Einordnung der 60-%-Schwelle

> **Ergebnis: Als Normalfall für das Gesamtvolumen ist die 60-%-Hypothese durch externe Evidenz nicht gestützt. Als Best-Case für einen eng abgegrenzten Intent-Block mit Live-Datenanbindung ist sie gestützt. Für den eigenen Assistenten bleibt die Wirkung ohne Pilotdaten nicht entscheidbar.**

Begründung an vergleichbaren Messungen:

1. **Unabhängig gemessen liegt die vollständige Lösung im Self-Service bei 14 %.** Gartner, 5.728 Befragte, Feldzeit 12/2023: 14 % aller Anliegen werden vollständig im Self-Service gelöst; selbst bei „very simple issues" nur **36 %**; 43 % der Gescheiterten fanden keinen passenden Inhalt. Das ist die belastbarste unabhängige Zahl im Feld.
2. **Der größte DACH-Benchmark liegt im Median bei 16,3 %.** Chatarmin, 140 Shops, 2,6 Mio. Tickets (Jun 2025–Jun 2026): Median 16,3 %, Durchschnitt 25,2 %, bestes Viertel 61,2 %, über alle Tickets 33,2 % — **KI allein löst 4,7 %** aller Tickets vollständig. Die 60 % sind damit **Best-Case-Niveau (Top-Quartil)**, nicht Erwartungswert. Ein produktiver DACH-Shop (OBI) liegt bei 20 %.
3. **Die hohen Quoten (76–82 %) sind Anbieterangaben mit anbietereigener Metrik.** Intercom 76 % Ø (12.000+ Kunden), Zendesk „bis zu 80 %", Salesforce 76 % / 82 % (Datasite), Freshworks Retail 53 %. Keine dieser Zahlen ist von Dritten gemessen.
4. **Die Zielmetrik ist keine Standardkennzahl.** „Korrekt ohne Übergabe" = Resolution unter Containment-Bedingung. Containment allein zählt auch stille Abbrüche, Falschantworten und gescheiterte Übergaben als Erfolg — die Kennzahl **steigt, wenn der Service schlechter wird**.

4. Belegtabelle (kondensiert)

#	Auftrag	Befund	Zahl	Messdefinition	Stichprobe	Bedingungen	Quelle / Fundstelle	Quellenart	offene Frage
D1	Erfolgsbefunde	Intercom Fin Ø-Resolution	76 %	unternehmenseigene „Resolution Rate", inkl. *assumed* (Kunde verlässt Chat)	12.000+ Kunden	Vendor-Kundenportfolio	fin.ai/help/en/articles/13975800-fin-pricing-outcomes	[VENDOR]	Anteil assumed resolutions nicht veröffentlicht
D2	Erfolgsbefunde	Zendesk AI Resolution	„bis zu 80 %"	Nenner = „AI-touched interactions within your reporting scope"	k. A.	Reporting-Scope wählbar	zendesk.com (AI-Resolution-Seite)	[VENDOR]	Bezugsgröße frei konfigurierbar
D3	Erfolgsbefunde	Salesforce/Datasite Deflection	82 % (von „low-60s")	Deflection nach erzwungener Präzisierungsfrage vor „Escalate"	1 Kunde	aktive Eskalationssperre	salesforce.com Blog (Datasite)	[VENDOR]	Effekt der Sperre nicht isoliert
D4	Erfolgsbefunde	Freshworks Benchmark 2025, Retail	53 %	Deflection, Branchenschnitt	k. A. (Branchenpanel)	Retail, höchster Branchenwert	freshworks.com Benchmark-Report S. 15–16	[VENDOR]	Grundgesamtheit/Zeitraum offen
D5	Erfolgsbefunde	Gartner: vollständig im Self-Service gelöst	**14 %** (36 % bei „very simple")	„fully resolved" ohne Mitarbeiter	5.728 Befragte	12/2023	gartner.com Newsroom 19.08.2024 (Original 403 → CX Today)	[UNABHÄNGIG], `[SEKUNDÄR]`	Erhebungsweg Survey, nicht Ticketlog
D6	Erfolgsbefunde	Chatarmin-Benchmark DACH	**16,3 %** Median / 25,2 % Ø / 61,2 % Top-Quartil / 4,7 % KI allein	„automatisch gelöste Anfragen" über alle Tickets	140 Shops, 2,6 Mio. Tickets	Jun 2025–Jun 2026	chatarmin.com/blog/kundenservice-automatisierung	[VENDOR] (Benchmark mehrerer Shops)	Methodik der Intent-Zuordnung offen
D7	Erfolgsbefunde	OBI (DACH, produktiv)	20 %	automatisierte Lösungsquote	1 Shop	k. A.	OBI-Case (über chatarmin-Benchmark referenziert)	[PRAXISBERICHT]	Zeitraum/Grundgesamtheit offen
D8	Erfolgsbefunde	ECom-Bench (2025)	43,9 % pass^1	Task-Erfolg auf echten E-Commerce-Support-Tasks	Benchmark-Set	bestes Modell (GPT-4o)	arXiv ECom-Bench	[UNABHÄNGIG]	Benchmarkset ≠ realer Ticketmix
D9	Erfolgsbefunde	τ-bench	61 % Erstversuch, 25 % bei Wiederholung	pass^k (konsistente Lösung über mehrere Versuche)	Benchmark-Set	Agenten mit Tool-Nutzung	arXiv τ-bench (S. Yao et al.)	[UNABHÄNGIG]	Präzision kommt aus Urteilszahl, nicht Item-Zahl
M1	Messdefinitionen	Zielmetrik formal	$q = \lvert\{C_i=1 \wedge H_i=0\}\rvert / N$	Resolution unter Containment-Bedingung	—	—	USPS OIG RISC-WP-21-010 (2021)	[AMT/STUDIE]	keine Standarddefinition in Normen
M2	Messdefinitionen	Prüfbarkeit der 60-%-Schwelle	n=400 @ 65 % → untere Wilson-Grenze 60,2 %	Konfidenzintervall statt Punktwert	Rechnung	zweiseitig, 95 %	Brown/Cai/DasGupta, Statistical Science 2001, doi:10.1214/ss/1009213286	[WISSENSCHAFTLICH]	DEFF-Aufschlag bei geschichteter Ziehung nötig
M3	Messdefinitionen	Kein Regelwerk definiert die Metrik	—	NIST AI RMF, ISO/IEC 42001, AI Act, DIN/VDE fordern nur Dokumentation von Metrik, Testset, Repräsentativität, Unsicherheit	—	—	NIST AI 100-1; AI Act Art. 50; ISO/IEC 42001 (Paywall, `[OFFEN]`)	[STANDARD]	ISO-Volltext nicht eingesehen

#	Auftrag	Befund	Zahl	Messdefinition	Stichprobe	Bedingungen	Quelle / Fundstelle	Quellenart	offene Frage
P1	DACH-Praxis	Juskys (Shop)	1,8 % → ~60 %	automatisierte Lösung nach Anliegen-Zählung	1 Shop	nach Anbindung plentyONE/Shopify	cleanercode.de/usecase/juskys	[PRAXISBERICHT]	Zählung durch Zendesk, nicht unabhängig auditiert
P2	DACH-Praxis	Edel Optics	25 % → ~75 % „über Nacht"	Lösungsquote nach Live-Bestelldaten	1 Shop	User Data API angebunden	Edel-Optics-Case	[PRAXISBERICHT]	Vorher/Nachher nicht kontrolliert
P3	DACH-Praxis	Lascana	„99,4 % Resolution Rate"	**technische Abbruchrate**, keine Lösungsquote	1 Shop	—	Lascana-Case	[PRAXISBERICHT]	Metrik irreführend benannt
P4	DACH-Praxis	Mister Spex	„88 %", „92 %" (Walbusch: Intent-Erkennung)	nur Rückversand-Labels bzw. Intent-Erkennung	1 Shop je	—	Cases	[PRAXISBERICHT]	Case widerspricht sich intern
I1	Intent-Verteilung	Post-Purchase dominiert	56,5 % After-Sales-Chats; 38,9 % WISMO+Retouren (DACH); 18 % WISMO allein	Anteil an allen Service-Chats	0,4 Mio. Chats/Tag (Taobao); 131 Shops/2,9 Mio. (DACH)	gemessen	Ni et al., arXiv:2603.29888v2, Abschn. 3.2 S. 8; chatarmin; Gorgias	[WISSENSCHAFTLICH]/[VENDOR]	isolierter Retourenanteil für westliche Shops offen
I2	Intent-Verteilung	Rückerstattung/Retoure ist größte Einzelkategorie (~77 % des Blocks)	—	Kategorienverteilung	s. o.	—	arXiv:2603.29888v2, Tab. A9	[WISSENSCHAFTLICH]	Kanalspezifische Intent-Matrix existiert nicht
I3	Intent-Verteilung	„Zendesk CX Trends 2025: 35–45 % WISMO"	**nicht verifizierbar**	—	—	—	Newsroom + vollständiges Deck geprüft	Negativbefund	kursierende Zahl unbelegt
G1	Gegenbelege	Alibaba/Taobao-Feldexperiment	Bot-only 64,6 % schneller, aber **−0,858** Bewertungspunkte; emotional aufgestaute Eskalation **+40,8 %** Zeit, **+6 pp** Wiederkontakte, **−0,928** Punkte	Bearbeitungszeit, Kundenzufriedenheit, Wiederkontakt	647 Agenten, 680.676 Chats	08/2024, Feldversuch	arXiv:2605.14830v2	[WISSENSCHAFTLICH]	einzige kausale Trennung von Tempo und Qualität
G2	Gegenbelege	Messartefakte	Containment 85 % / Automation 70 % / echte Resolution 55 % gleichzeitig möglich	Abbrecher, Close-on-silence 72 h, Reopen = neues Ticket	Praxisberichte	—	Platforma (2026) u. a.	[PRAXISBERICHT]	Short-Circuit-Quote nicht unabhängig gemessen
G3	Gegenbelege	Klarna	2024 „zwei Drittel der Chats" → 2025 nur „−40 % Kosten/Transaktion"	Kommunikationswechsel	—	—	Klarna-Presse + BI-Korrektur	[VENDOR] + Korrektur	Ursachenzuordnung „KI" von Klarna bestritten
U1	Wissensbasis	Häufigster Self-Service-Fehler	43 % fanden keinen passenden Inhalt	Abbruchgrund	5.728	12/2023	Gartner (s. D5)	[UNABHÄNGIG]	—
U2	Wissensbasis	DSK-Anforderungen an Referenzdokumente	Qualität, Aktualität, Vollständigkeit; deutschsprachige Embeddings	—	—	06.05.2024	DSK Orientierungshilfe KI und Datenschutz, Rz. 12–13, 22	[BEHÖRDE]	—

erstellt/modifiziert von Max Headroom digitaler Assistent von Lars Rißmann-Kaplusch

#	Auftrag	Befund	Zahl	Messdefinition	Stichprobe	Bedingungen	Quelle / Fundstelle	Quellenart	offene Frage
U3	Übergabe/QA	Übergabe-Designprinzipien	u. a. Übergabe auf Anforderung, proaktiver Trigger nach >3 erfolglosen Vorschlägen, Kontext-/Verlaufsübergabe, Vorwarnung	Designprinzipien, **keine Effektstärken**	Experteninterviews + Prototyp	HICSS 2021	Poser/Singh/Bittner, HICSS 2021, DP1–DP5	[WISSENSCHAFTLICH]	Schwellenwert nicht empirisch optimiert
U4	Übergabe/QA	Übergabe ist die schwächste Stelle	10 % „immer reibungslos"; 59 % müssen erneut erklären; 46 %: Vorgeschichte unbekannt	Verbraucherbefragung	Methodik nicht einsehbar	—	Platforma (2026), B24	[PRAXISBERICHT]	Stichprobe offen
U5	Recht	Letztentscheidung löst Art. 22 DSGVO aus	formelle Mensch-Beteiligung genügt nicht	—	—	DSK 06.05.2024	DSK-OH, Rz. 12–13	[BEHÖRDE]	—
U6	Recht	§ 356a BGB (ab 2026)	dauerhaft verfügbare, hervorgehobene Widerrufsfunktion + Eingangsbestätigung	—	—	seit 19.06.2026	gesetze-im-internet.de/bgb/__356a.html	[RECHT]	ob ein Chat sie ersetzt: Auslegung `[OFFEN]`
U7	Recht	Art. 50 KI-VO	Transparenzpflicht für KI-Interaktion	—	—	gilt seit 02.08.2026	Amtsblatt-Volltext (Browser geprüft)	[RECHT]	Kommissions-Leitlinien noch nicht final

erstellt/modifiziert von Max Headroom digitaler Assistent von Lars Rißmann Karlovsch

5. Doppelte und abhängige Quellen

- **Gartner 14 % / 36 % / 43 %** erscheint in den Aufträgen 1, 5 und 6 — es ist **eine** Erhebung (n = 5.728, 12/2023), dreifach zitiert. Nicht als drei unabhängige Belege zählen.
- **Die Gartner-Pressemitteilung liefert HTTP 403**; der Inhalt wurde über CX Today und weitere Reproduktionen gegengeprüft. Zwei Reproduktionen derselben Zahl sind **kein** zweiter Beleg.
- **Chatarmin-Benchmark** (Median 16,3 %) und der OBI-/Edel-Optics-/Juskys-Case stammen aus demselben Anbieter-Ökosystem — Fallberichte sind dort Marketing-Assets. Getrennt behandelt, aber nicht als unabhängige Messungen zu werten.
- **Zendesk/Intercom/Salesforce/Freshworks** sind vier Wettbewerber mit vier eigenen Definitionen — eine gemeinsame „Branchenquote" entsteht daraus nicht.
- **Die vielzitierte Gartner-80-%-Zahl ist eine Prognose für 2029** und gilt ausdrücklich nur für „**common**" Anliegen. Wer daraus 60 % für das Gesamtvolumen ableitet, tauscht die Bezugsgröße.

erstellt/modifiziert von Max Headroom digitaler Assistent von Lars Rißmann-Kaplusch

6. Widersprüche

1. **Vendor 76–82 % vs. unabhängig 14 %.** Auflösbar nur über den Nenner: Vendor zählt nur „AI-touched interactions" im eigenen Reporting-Scope, teils inklusive stiller Abbrüche; Gartner fragt Kundinnen und Kunden, ob ihr Anliegen vollständig gelöst wurde. Beide Zahlen können gleichzeitig „stimmen" — und messen Verschiedenes.
2. **Datasite 82 %** entstand laut Salesforce-Blog durch ein „deterministic gate", das eine Präzisierungsfrage erzwingt, bevor „Escalate" auslöst („filtering out premature handoffs"). Die Zahl ist damit teils eine **UI-Konstante**, kein Qualitätsnachweis.
3. **Chatarmin: Ø 25,2 % vs. „KI allein löst 4,7 %".** Zwei verschiedene Erfolgsbegriffe im gleichen Report (mit Agenten-Unterstützung vs. rein KI). Wer den höheren Wert zitiert, muss den Unterschied nennen.
4. **Lascana „99,4 % Resolution Rate"** ist laut Quelle selbst die **technische Abbruchrate** — nicht die Lösungsquote. Mister Spex „88 %" gilt nur für Rückversand-Labels; der Case nennt die Label-Automatisierung gleichzeitig als Zukunft.
5. **Klarna:** 2024 „zwei Drittel der Chats", 2025 nur noch „−40 % Kosten pro Transaktion" — mit ausdrücklicher Korrektur, dass die CEO-Aussage Kostensenkung allgemein betraf, nicht KI.
6. **τ-bench 61 % vs. 25 %:** derselbe Benchmark, zwei Werte — Erstversuch vs. konsistente Wiederholung. Ein Einzelwert ohne k-Angabe ist nicht interpretierbar.

7. Gegenbelege (bewusst gesucht)

- **Eskalation ist der teuerste Fehler, nicht die Fehlantwort.** Bot-only-Chats sind schneller, aber messbar schlechter bewertet; nach aufgestautem Frust („emotionale Eskalation") steigen Bearbeitungszeit und Wiederkontakte, und die Bewertung fällt weiter — der Mensch repariert es dann nicht mehr. Technische Eskalationen sind dagegen reparierbar. Das Argument gegen die Hypothese lautet also nicht „KI antwortet falsch", sondern **„KI eskaliert zu spät"**.
- **Containment misst die Abwesenheit von Menschen, nicht die Lösung.** Dokumentierte Artefakte: Abbrecher/Timeouts zählen als „contained"; Close-on-silence nach 72 h; Reopen als neues Ticket („one failure becomes two successful resolutions"); reiner Chat-Nenner; erste Wochen mit handverlesenen Intents.
- **Nicht automatisierbar** sind nachweislich Anfragearten mit Einzelfallprüfung: Reklamationen mit Beweiswürdigung, Kulanz, Zahlungsstörungen, Rechtsstreit.
- **Gefundene Gegenbelege, die sich nicht halten:** „40 % WISMO (Forrester)" und „80 % repetitive Tickets" — in den Originalen nicht auffindbar; als unbelegt verworfen.

erstellt/modifiziert von Max Headroom digitaler Assistent von Lars Rißmann-Kaplusch

8. Übertragbarkeit: Bedingungen

Die belegten Erfolgsfaktoren sind **inhaltlich und prozessual**, nicht modelltechnisch:

Hebel	Beleg	Wirkung
Systemanbindung (OMS/Tracking) statt nur Wissensbasis	Juskys 1,8 % → ~60 %; Edel Optics 25 % → ~75 %; Tonerpartner −36 % Tickets	größter dokumentierter Sprung
Wissensbasis: Qualität, Aktualität, Vollständigkeit	Gartner: 43 % scheitern an fehlendem Inhalt; DSK: Referenzdokumente + deutschsprachige Embeddings	notwendige, nicht hinreichende Bedingung
Intent-Zuschnitt	Intent-Verteilung: 38,9 % WISMO+Retouren (DACH)	bestimmt die Obergrenze
Übergabefähigkeit	HICSS-Designprinzipien; DSK	Teil der Kennzahl, nicht Störfaktor
Testbestand-Design	Wilson/VCC; τ-bench	entscheidet, ob die 60 % überhaupt prüfbar sind

Rechtlicher Rahmen (keine Rechtsberatung): Nicht die Antwort kippt, sondern die **Entscheidung**. Eine automatisierte Letztentscheidung über Rücknahme/Erstattung löst Art. 22 DSGVO aus; formelle Mensch-Beteiligung genügt laut DSK nicht. § 356a BGB verlangt seit 2026 eine dauerhaft verfügbare, hervorgehobene Widerrufsfunktion mit Eingangsbestätigung — ein Chat ersetzt sie nicht (Auslegung offen). Art. 50 KI-VO (Transparenz) gilt seit 02.08.2026.

Konsequenz für die Quote: Eine Containment-Zahl aus einem Betrieb mit 24/7-Verfügbarkeit und dedizierten Übergabe-Slots ist auf einen Betrieb mit Sprechzeiten und Nebenaufgaben **nicht** übertragbar. Je höher die Quote, desto größer die Beweislast, dass keine Letztentscheidung durch das System erfolgt.

9. Übergabe und Qualitätsprüfung

- **Design-Muster (belegt als Design-Wissen, nicht als Effektstärke):** Übergabe auf jederzeitige Anforderung; proaktiver Trigger bei Zeitgrenze oder mehr als drei erfolglosen Lösungsvorschlägen; Kontext- und Verlaufsübergabe inklusive Bestell-/Kundendaten; Vorwarnung der übernehmenden Person.
- **Der belastbarste Befund ist negativ:** Der Service scheitert am sichtbarsten **an der Übergabe** (Sackgassen, erzwungene Wiederholung, verlorener Kontext, stille Warteschlangen) — und **keiner dieser Fehler senkt die Containment-Quote, mehrere erhöhen sie sogar**.
- **Metriken, die zwingend mitgemessen werden müssen:** Resolution, **False Containment**, Recontact Rate (48/72 h), **Trigger-Precision und -Recall** der Übergabe. Ohne Trigger-Recall schönt jede Quote sich selbst.

10. Zwei Quellen zur menschlichen Prüfung

1. **Intercom-Definition von „Resolution"** — der entscheidende Hebel, weil dort offensteht, dass eine „Resolution" auch als *assumed* zählt, wenn der Kunde das Gespräch ohne Nachfrage verlässt, und genau das abgerechnet wird: https://fin.ai/help/en/articles/13975800-fin-pricing-outcomes
2. **Gartner, 14 % vollständig im Self-Service gelöst** (5.728 Kunden, 08/2024) — die einzige unabhängige Zahl, die die 60 % als Normalfall infrage stellt: https://www.gartner.com/en/newsroom/press-releases/2024-08-19-gartner-survey-finds-only-14-percent-of-customer-service-issues-are-fully-resolved-in-self-service Ersatzzugang (Original liefert 403): https://www.cxtoday.com/contact-center/only-1-in-7-customer-service-queries-resolved-with-self-service-gartner-study-finds/

11. Nächster prüfbarer Schritt

1. **Metrik vor dem Test einfrieren:** q = korrekt ohne Übergabe gelöste Anfragen / alle Anfragen im **gleichen** Testbestand; zusätzlich False Containment, Recontact (72 h), Trigger-Precision/-Recall.
2. **Testbestand ziehen:** stratifizierte Zufallsstichprobe über echte Anfragen, **≥ 400 Fälle** als Untergrenze, 600–1.000 für einen echten Nachweis; bei 65 % beobachteter Quote liegt die untere Wilson-Grenze bei n = 400 gerade bei 60,2 %.

erstellt/modifiziert von Max Headroom digitaler Assistent von Lars Rißmann-Kaplusch

3. **Urteile absichern:** Doppelcodierung, Cohens κ berichten, LLM-Judge gegen Mensch kalibrieren.
4. **Voraussetzung vor dem Test klären:** Zugriff auf Bestell-/Tracking-Daten. Ohne Systemanbindung liegt die realistische Erwartung nach der DACH-Evidenz bei ~20–25 %, nicht bei 60 %.
5. **Entscheidung trifft der Product Owner** — diese Recherche formuliert keine Entscheidung.

12. Offene Punkte

- Keine unabhängig auditierte Containment-Messung für einen produktiven E-Commerce-Bot gefunden (nur Vendor-Angaben mit eigener Metrik).
- Anteil *assumed resolutions* an den Intercom-76 % nicht veröffentlicht.
- Ein shop-übergreifender Intent-Benchmark mit offenem Nenner existiert nicht — eigene Tagging-Messung bleibt unvermeidlich.
- Kanalspezifische Intent-Verteilung (E-Mail vs. Chat vs. Telefon) öffentlich nicht belegt.
- Kein empirischer Schwellenwert für Konfidenz-gesteuerte Übergabe-Trigger.
- ISO/IEC 42001, DIN EN 15838, Kommissions-Leitlinien zu Art. 50: Volltexte nicht eingesehen (Paywall/Cloudflare/nicht final).
- Ob ein Chat die Widerrufsfunktion nach § 356a BGB erfüllen kann: Auslegung offen.

13. Rohrecherche

- `Rohrecherche/auftrag1.md` — Erfolgsbefunde / Containment-Benchmark (15 Datenpunkte D1–D15)
- `Rohrecherche/auftrag2.md` — Messdefinition und Evaluation (25 Quellen)
- `Rohrecherche/auftrag3.md` — DACH-Praxis (14 Fälle)
- `Rohrecherche/auftrag4.md` — Intent-Verteilung (A/B/C/R-Belege)
- `Rohrecherche/auftrag5.md` — Gegenbelege und Messartefakte
- `Rohrecherche/auftrag6.md` — Übertragbarkeit, Wissensbasis, Recht, Übergabe und Qualitätsprüfung (B1–B35)