

Sehr geehrte Frau Smolko,

vielen Dank für die Möglichkeit, ein Angebot für die technische Begleitung des Digitalen Wegweisers Berlin abzugeben. Frau Julia Stertz hat mich auf das Projekt aufmerksam gemacht.

Im Anhang finden Sie mein verbindliches Angebot als PDF. Ergänzend habe ich eine Angebotsseite gestaltet, auf der Sie die wichtigsten Punkte in Ruhe ansehen können: eine Simulation des Wegweisers, kurze Erklärvideos und meinen Vorab-Test der Sprachen.

wegweiser-bewerbung.maximgrabow.de (nicht öffentlich gelistet, ohne Cookies und Tracking)

Mein Vorschlag in Kürze:

- **Geprüfte Antworten statt freier KI:** Der Wegweiser antwortet nur aus Inhalten, die Ihr Team freigegeben hat. Wo nichts Geprüftes vorliegt, verweist er an Menschen.
- **Erprobte, offene Bausteine:** Das Prinzip Ihrer IONOS-Prüfung bleibt erhalten und wird nur dort ergänzt, wo es nachweislich nötig ist. Jeder Baustein bleibt austauschbar.
- **Sprachen gemessen statt angenommen:** In einem eigenen Vorab-Test mit 11 Sprachen habe ich offene und proprietäre Modelle verglichen. Daraus folgt eine klare Empfehlung je Sprache, die endgültige Auswahl treffe ich in Phase 1 gemeinsam mit den Kursleitenden.
- **Ihr Team wird eigenständig:** Redaktion ohne Technikwissen, Schulungen und ein Übergabetest.

Ich erbringe die Leistung persönlich über alle drei Phasen. Mein Stundensatz beträgt **100,00 € netto zuzüglich 19 % Umsatzsteuer (19,00 €), also 119,00 € brutto**. Den Aufwand plane ich mit rund 380–440 Stunden, höchstens 500; davon rund 170–190 Stunden in Phase 1.

Für Rückfragen und ein persönliches Gespräch stehe ich gern zur Verfügung. Ich freue mich über eine kurze Bestätigung des Eingangs.

Mit freundlichen Grüßen

Maxim Grabow

Freiberuflicher Entwickler für KI-Anwendungen und Web

maxximgrabow@gmail.com · 0173 8241820

Anlagen: Angebot (PDF) mit technischem Vorgehen (5 Seiten), Kurzprofil, Referenzen, Preis; Anlagen A1 Modelle und Open Source, A2 Architektur und Zeitplan, A3 Anforderungs-Navigator



Angebot · Technische Begleitung

Digitaler Wegweiser Berlin

für die Johannesstift Diakonie Proclusio gGmbH

Maxim Grabow · Leistung persönlich erbracht

Stand 11.10.2026

Angebot von Maxim Grabow – keine offizielle Unterlage der Johannesstift Diakonie.

Inhalt

Kapitel	Seite
–	Auf einen Blick
3	Auf einen Blick
1–2	Angaben zur Person, Kurzprofil und Qualifikationen
3	Referenzen
4	Mein Verständnis der Aufgabe
5	Technisches Vorgehen
6	Befähigung und Nachhaltigkeit
7	Aufwand, Zeitplan und Preis
A1	Modelle, Sprachen und Open Source
A2	Architektur und Zeitplan
A3	Anforderungs-Navigator

Auf einen Blick

Sechs Kernaussagen

- 1. Geprüfte Antworten statt freier KI.** Der Wegweiser antwortet nur aus Inhalten, die Ihr Projektteam geprüft und freigegeben hat, mit Quelle und Prüfdatum. Liegt nichts Geprüftes vor, sagt er das offen und verweist an einen Menschen.
- 2. Erprobte Bausteine statt Eigenbau.** Ich übernehme das Prinzip Ihrer IONOS-Prüfung und ergänze nur, was nachweislich fehlt: Redaktion ohne SQL, eine Suche, die ihre Grenzen erkennt, Audio in der EU und eine Prüfung je Sprache. Jeder Baustein ist Open Source oder austauschbar.
- 3. Sprachen gemessen statt angenommen.** In einem eigenen Vorab-Test mit 11 Zielsprachen erreichen offene Modelle bei europäischen Anbietern in Ukrainisch, Arabisch, Farsi und Dari nach meiner Entscheidungsregel die Qualität der besten gemessenen Systeme. Wo sie das nicht schaffen, etwa bei Tigrinya, sehe ich eine abschaltbare Ausnahme in der EU vor, erst nach Ihrer Datenschutz-Folgenabschätzung.
- 4. Datenschutz durch Technik.** Anonym und ohne Konten, Verarbeitung ausschließlich in der EU. Persönliche Angaben werden automatisch entfernt. Eine Prüfschicht ohne KI sichert Notrufnummern, Fristen und Daten ab.
- 5. Ihr Team wird eigenständig.** Redaktion ohne Technikwissen, Schulungen, Videoanleitungen, Sprechstunde und ein Übergabetest. Alle Konten und Rechte liegen bei Ihnen.
- 6. Ich habe den Kern schon gebaut.** Drei meiner Referenzprojekte sind KI-Systeme, die nur aus geprüftem Wissen antworten, mit messbarer Suchqualität und Übergabe an Menschen.

Kennzahlen

Aufwand	Planwert 380–440 Stunden in 3 Phasen, höchstens 500
Phase 1	170–190 Stunden, Beta ab ca. Monat 3–4
Vorab-Test	11 Sprachen, 20 KI-Systeme
Stundensatz	100,00 € netto · 119,00 € brutto (19 % USt)
Laufende Kosten	voraussichtlich ca. 100–300 € pro Monat
Leistung	persönlich, durch alle drei Phasen

Angaben zur Person, Kurzprofil und Qualifikationen

Maxim Grabow, freiberuflicher Entwickler für KI-Anwendungen und Web, Berlin
maxximgrabow@gmail.com · 0173 8241820 · maximgrabow.de · LinkedIn

Ich entwickle KI-gestützte Wissens- und Informationssysteme, die ausschließlich aus geprüften Inhalten antworten. Dazu gehören:

- Wissensdatenbanken auf PostgreSQL/pgvector
- hybride semantische Suche
- LLM-Anbindung über offene Schnittstellen
- Redaktionsoberflächen für nicht-technische Teams

Mein Schwerpunkt: bestehende Managed Services und Open-Source-Komponenten zu einer sicheren, datenschutzgerechten und wartbaren Lösung verbinden und deren Qualität messbar machen. Ich arbeite persönlich und transparent, mit Blick auf die Übergabe: Teams sollen ihre Systeme selbst pflegen können.

Neben meiner freiberuflichen Tätigkeit verantworte ich als Head of AI bei der Evalion GmbH KI-Anwendungen und Wissenssysteme. Das Angebot erbringe ich persönlich und im eigenen Namen.

Technik und Gestaltung aus einer Hand

Seit zehn Jahren verbinde ich Web- und Softwareentwicklung mit Grafikdesign, Motion Design und Videoschnitt. In den letzten Jahren ist die Integration von KI und Sprachmodellen zu meinem Schwerpunkt geworden. Für Ihr Projekt heißt das: Ich baue die Technik und mache sie zugleich verständlich, in der Oberfläche, in Schulungsvideos und in Anleitungen.

- **Technik:** Webentwicklung, Softwareentwicklung, KI-Integration, Sprachmodelle (LLM)
- **Gestaltung:** Grafikdesign, Motion Design, Videoschnitt
- **Werkzeuge (Auswahl):** TypeScript, Next.js, Cursor, Remotion, After Effects, Premiere Pro, DaVinci Resolve, Photoshop, Illustrator, Figma

Expertise

Bereich	Was ich mitbringe
KI-Wissenssysteme (RAG)	Antworten nur aus geprüftem Wissen, Relevanz-Schwellen, Quellenprüfung, Übergabe an Menschen; Qualitätsmessung (Recall, Erfindungstest)
Datenbanken	PostgreSQL, pgvector (HNSW), hybride Suche mit Volltext und Rangfusion, Zeilen-Sicherheit, Migrationen, Sicherung mit Zeitpunkt-Wiederherstellung
KI-Integration	OpenAI-kompatible Schnittstellen, Modellketten mit Ersatzmodellen, Schutzprüfungen, eigene Modell-Benchmarks
Web und Schnittstellen	TypeScript, React, Next.js, Node.js, REST/OpenAPI, MCP (Model Context Protocol), Progressive Web Apps

Bereich	Was ich mitbringe
Audio	Spracheingabe (Whisper) und Vorlesen, auch mehrsprachig
Cloud und Betrieb	Managed Services, Container, Monitoring, Kosten- und Störungsalarme, Sicherheitsprüfungen und Härtung
Datenschutz und KI-Regulierung	technische und organisatorische Maßnahmen, Ketten der Auftragsverarbeitung, Löschkonzepte, Schwärzung personenbezogener Daten, Kennzeichnung nach EU AI Act
Barrierefreiheit und Mehrsprachigkeit	WCAG-orientierte Umsetzung, Rechts-nach-links-Schriften, mehrsprachige Oberflächen mit muttersprachlicher Prüfung
Befähigung	Anleitungen und Runbooks für Laien, Schulungen zu KI-Werkzeugen und Prompting (eigene Kursinhalte)
Gestaltung und Vermittlung	Webdesign und Motion Design: Diese Bewerbungsseite und ihre Erklärvideos habe ich selbst konzipiert und mit modernen, auch KI-gestützten Werkzeugen umgesetzt

Stationen

Rolle	Organisation	Schwerpunkt
Freiberuflicher Entwickler (KI-Anwendungen, Web)	selbstständig	KI-Anwendungen, Web, Gestaltung
Head of AI	Evalion GmbH	KI-Assistenten, interner Workspace mit Wissensdatenbank

Sprachen: Deutsch, Englisch

Weitere Projekte (Auswahl, eigene Projekte)

Projekt	Was es zeigt	Stand
KI-Assistent auf maximgrabow.de	Schutzprüfungen, Modellkette, Prüfset mit 29 von 29 bestandenen Fällen, Spracherkennung Deutsch/Englisch	Testumgebung
KI-Fitness-App	pgvector-Suche, Audio rein und raus, Kosten-Kontingente, Monitoring	vor dem Start
App mit Gastzugang per QR-bzw. Raumcode	Zugang ohne Konto, 5 Sprachen mit muttersprachlicher Prüfung, Screenreader-Optimierung	eigenes Projekt
Kurse zu KI-Werkzeugen und Prompting	Befähigung von Nicht-Techniker:innen, passend zur KI-Kompetenz nach Art. 4 AI Act	eigenes Projekt

→ Die vier Referenzprojekte mit meiner Rolle stehen im Abschnitt **Referenzen**.

Referenzen

Gefordert sind mindestens zwei einschlägige Referenzen, davon mindestens eine mit Bezug zu RAG, LLM-Integration oder Wissensdatenbanken. Die folgenden vier Projekte habe ich jeweils maßgeblich selbst konzipiert und umgesetzt. Drei davon betreffen genau den Kern des Wegweisers: eine KI, die nur aus geprüftem Wissen antwortet.

1. KI-Assistent mit Redaktionsbereich für die Website von Gutachten.org

Kontext: Gutachten.org ist die Marke, unter der die Evalion GmbH (Berlin) Immobiliengutachten anbietet. Für die Website habe ich einen KI-Assistenten gebaut, der Fragen von Kundinnen und Kunden beantwortet. Umgesetzt im Rahmen meiner Tätigkeit als Head of AI der Evalion GmbH.

Meine Rolle: Konzeption, Architektur, Umsetzung und Datenschutzkonzept (technische und organisatorische Maßnahmen, Kette der Auftragsverarbeitung)

Umgesetzt:

- KI-Chat ohne Anmeldung, der ausschließlich aus einer redaktionell gepflegten Wissensbasis antwortet (RAG mit PostgreSQL und pgvector).
- Redaktionsoberfläche mit Prüfdatum, Gültigkeit, verantwortlicher Person, Versionen und einer Liste offener Fragen.
- Feste Schutzprüfungen vor dem KI-Modell, unter anderem gegen Manipulationsversuche, dazu ein satzweiser Ausgabefilter.
- Automatische Schwärzung personenbezogener Angaben, Löschung nach 90 Tagen, Not-Aus.
- Kennzeichnung als KI nach Art. 50 EU AI Act.

Bezug zum Wegweiser: Redaktion als Haftungsgrenze, Datenschutz durch Technik, KI-Transparenz

Status: live in Produktion

2. Interner Workspace mit Wissensdatenbank und semantischer Suche

Kontext: Der Workspace ist die interne Arbeitsumgebung der Evalion GmbH, des Unternehmens hinter Gutachten.org, für ein Team von rund zehn Personen. Er enthält eine Wissensdatenbank mit KI-Assistent für die Mitarbeitenden, dazu Marketing-Dashboards und Analysen, E-Learning und weitere Module. Diese Referenz beschreibt die Wissensdatenbank. Entstanden im Rahmen meiner Tätigkeit dort.

Meine Rolle: Architektur, Umsetzung, Betrieb und Qualitätsmessung

Umgesetzt:

- Hybride Suche aus Vektorsuche (pgvector) und deutscher Volltextsuche mit anschließender Neubewertung. Im eigenen Testset fand sie bei 100 % der Testfragen den richtigen Eintrag unter den ersten fünf Treffern (Recall@5).
- Relevanz-Untergrenze: Findet die Suche nichts Passendes, kommt eine feste Antwort ohne KI-Aufruf.
- Dokumentierte REST-Schnittstelle für Importe, MCP-Server mit rund 20 Werkzeugen, API-Schlüssel, Protokoll und Missbrauchsbremse.
- Diktat und Vorlesen.

- Messskripte für Modellvergleiche, darunter ein Test, wie oft ein Modell Behauptungen erfindet, die nicht in den Quellen stehen.

Bezug zum Wegweiser: semantische Suche, messbare Qualität, Schnittstelle und MCP (§ 6.8, § 6.9), Audio

Status: live, intern produktiv im Einsatz

3. Website-Baukasten mit KI-Assistent und Redaktionsworkflow

Kontext: mandantenfähiger Website-Baukasten für Dienstleister in einem sensiblen Beratungsbereich: Websites, Livechat, Anfrageverwaltung und KI-Assistent aus einer Hand. Kundenprojekt (anonymisiert; Bildschirmfotos mit neutralen Demo-Inhalten)

Meine Rolle: alleinige Konzeption, Entwicklung und technischer Betrieb

Umgesetzt:

- KI-Chat, der nur aus gepflegtem Wissen antwortet, in drei Stufen: belegte Antwort, ehrlich benannte Lücke, direkte Übergabe an einen Menschen.
- Serverseitige Prüfung, ob die genannten Quellen wirklich existieren.
- Redaktion für Laien mit den Bereichen Wissen, Zur Prüfung, Dokumente, Lücken und Testen, dazu eine wöchentliche Zusammenfassung offener Lücken an das Team.
- Anonyme Nutzung ohne Login mit festen Aufbewahrungsfristen, Störungsalarm mit Not-Aus.
- Strikte Mandantentrennung, Vorlagen für Auftragsverarbeitung und technische und organisatorische Maßnahmen, rund 700 automatische Tests.

Bezug zum Wegweiser: Redaktion durch Nicht-Techniker:innen, Übergabe an Menschen, Datenschutz in einem Umfeld mit besonders sensiblen Daten, verlässlicher Betrieb

Status: live beim Kunden

4. Sicherheitsprüfung, Härtung und Modernisierung eines Webportals

Kontext: bundesweites Vergleichs- und Vermittlungsportal für Bestattungen mit Kunden- und Partnerbereich. Kundenprojekt (anonymisiert)

Meine Rolle: Sicherheitsprüfung, Koordination der Gegenmaßnahmen, Härtung, Anleitungen für Nicht-Techniker:innen

Umgesetzt:

- Systematische Sicherheitsprüfung mit priorisierten Befunden.
- Koordinierte Gegenmaßnahmen und Einführung der Zwei-Faktor-Anmeldung.
- Verständliche Anleitungen für Nutzende auf Deutsch und Englisch.
- Modernisierungsplan für den weiteren Betrieb.

Bezug zum Wegweiser: Betrieb und Sicherheit eines öffentlich erreichbaren Systems, Befähigung von Laien

Status: laufendes Projekt

Sicherheit aus der Praxis

Ein öffentlich erreichbarer, anonymer Dienst unter dem Namen der Diakonie muss Angriffen standhalten. Diese Erfahrung bringe ich aus meinen Projekten mit (projektübergreifend und anonymisiert):

- **Ernstfall:** Nach der Kompromittierung eines Servers bei einem Kunden habe ich die Reaktion geleitet: Eindämmung, Beweissicherung, sauberer Neuaufbau, Erneuerung aller Zugangsdaten und Pflicht zur Zwei-Faktor-Anmeldung, in rund sieben Tagen.
- **Sicherheitsprüfungen:** Vier Prüfungen von Code und laufenden Systemen mit über 40 bestätigten Befunden, nach Schwere geordnet. Die Prüfungen waren KI-gestützt; jeder Befund wurde unabhängig gegengeprüft, bevor er behoben wurde.
- **Sichere Anmeldung:** Zwei-Faktor-Anmeldung und Passkeys, die gegen Phishing schützen, Wiederherstellungscodes, Abgleich mit bekannten geleakten Passwörtern sowie Anleitungen auf Deutsch und Englisch für Nicht-Techniker:innen.
- **Zero-Trust-Zugang:** Verwaltungszugänge nur über NetBird bzw. Cloudflare Access mit Mehrfaktor-Anmeldung; öffentlich erreichbare Fernwartung (SSH) geschlossen.
- **Spam, Bots und Missbrauch:** Formularschutz mit unsichtbarem Fangfeld (Honeypot) und Zeitfalle, Bot-Prüfung (Turnstile, reCAPTCHA), Web Application Firewall, Ratenbegrenzung und eine Kostenbremse für KI-Aufrufe.
- **E-Mail-Sicherheit:** SPF, DKIM und DMARC geprüft und schrittweise verschärft.
- **Überwachung und Wiederherstellung:** stündliche Integritätsprüfung mit Alarm, tägliches Lebenszeichen, nächtliche Datensicherung außerhalb des Anbieters und eine getestete Wiederherstellung (71 von 71 Zugriffsregeln).
- **Anwendungssicherheit:** in der Datenbank erzwungene Mandantentrennung mit automatischen Tests (rund 50 Lücken zwischen Mandanten geschlossen), strikte Content-Security-Policy und Schutzschranken für KI-Chats gegen Manipulationsversuche, mit Ausgabefilter und Schwärzung personenbezogener Daten.

Bildschirmfotos der Referenzsysteme zeigen ausschließlich Demo-Daten.

Mein Verständnis der Aufgabe

Wer neu in Berlin ankommt, steht vor vielen Fragen auf einmal: Was bedeutet dieser Brief? Bis wann muss ich reagieren? Wer ist zuständig, und wo bekomme ich Hilfe? Die Wegweiserkurse geben darauf in der Herkunftssprache Orientierung. Der **Digitale Wegweiser Berlin** soll dieses Wissen auch zwischen den Kursterminen erreichbar machen: anonym, per QR-Code, in der eigenen Sprache und zu jeder Zeit.

Dabei geht es um Verlässlichkeit, nicht um eine möglichst kluge KI. Eine falsche Frist oder eine falsche Zuständigkeit kann für die Nutzenden ernste Folgen haben. Deshalb verstehe ich die Aufgabe so:

1. **Eine geprüfte Wissensbasis, die Ihr Team selbst pflegt.** Die KI formuliert nur, was Ihr Projektteam geprüft und freigegeben hat. Wo nichts Geprüftes vorliegt, sagt sie das und verweist an Menschen.
2. **Einfache, erprobte Technik statt Eigenbau.** Managed Services und Open-Source-Komponenten, verbunden durch eine schlanke, gut dokumentierte Schicht. Jeder Baustein lässt sich austauschen, kein Anbieter wird unverzichtbar.
3. **Datenschutz und Barrierefreiheit von Anfang an.** Die Zielgruppe ist besonders schutzbedürftig. Deshalb gibt es keine Konten, keine Profile, Verarbeitung nur in der EU, Unterstützung für alle Schriftrichtungen, Vorlesen und große Schrift.
4. **Ihr Team wird eigenständig.** Am Ende der Laufzeit soll Ihr Team Inhalte selbst pflegen, Qualität selbst prüfen und das System bei Bedarf an einen anderen Dienstleister übergeben können.

Ich begleite das Projekt persönlich durch alle drei Phasen: von der Konzeption und Einrichtung über die Beta-Erprobung in den Unterkünften bis zur Stabilisierung und Übergabe.

1. Leitidee und Architektur

Ziel: Neu angekommene Menschen bekommen in ihrer Sprache eine verlässliche erste Orientierung, rund um die Uhr, anonym und ohne Anmeldung. Der Wegweiser antwortet ausschließlich aus Inhalten, die Ihr Projektteam geprüft und freigegeben hat. Liegt nichts Geprüftes vor, sagt er das offen und verweist an einen Menschen.

Leitidee: Die von Ihnen geprüfte IONOS-Lösung ist eine gute Grundlage. Ich übernehme ihr Prinzip: eine verwaltete PostgreSQL-Datenbank mit Vektorsuche, ein KI-Modell über eine Standardschnittstelle und eine schlanke Weboberfläche. Ergänzt wird nur, was für diesen Wegweiser nachweislich fehlt: eine Redaktionsoberfläche mit Freigabe, eine Suche, die ihre eigenen Grenzen erkennt, eine Prüfung je Sprache sowie Spracheingabe und Vorlesen innerhalb der EU. Datenbank, KI und Hosting entwickle ich nicht selbst, sondern nutze bewährte Managed Services und Open-Source-Komponenten. Eigener Code entsteht nur als schlanke Verbindungsschicht.



Systembild. Eine Frage kommt über den QR-Code in die Chat-Oberfläche und durchläuft im Wegweiser-Dienst Schutz, Suche, KI-Gateway und Prüfschicht. Die Antwort geht mit Quelle und Prüfdatum zurück. Ihr Team pflegt die Wissensdatenbank in der Redaktion; der Wegweiser liest nur Freigegebenes.

Baustein	Lösung	Warum
Datenbank	Managed PostgreSQL mit pgvector, bevorzugt STACKIT (deutscher Betreiber, C5-Testat Typ 2 inkl. Datenbank), gleichwertig IONOS	Standard-SQL, tägliche Sicherung mit Wiederherstellung auf einen Zeitpunkt, Umzug per Export jederzeit möglich
Redaktion	Payload CMS (Open Source, MIT-Lizenz)	Formulare statt SQL, Vier-Augen-Freigabe über Rollen, Versionen, Sprachfassungen inkl. Rechts-nach-links-Schrift, keine Lizenzkosten
Suche und Antwort	schlanker eigener Dienst (ca. 500 Zeilen) auf Basis offener Bibliotheken	nachvollziehbar, testbar, keine zusätzliche Plattform
KI-Modelle	Gateway mit OpenAI-kompatibler Schnittstelle (LiteLLM, Open Source); offene Modelle bei EU-Anbietern als Standard	Modellwechsel per Konfiguration, ohne Umbau
Chat-Oberfläche	Web-App (PWA) mit assistant-ui (Open Source)	per QR-Code sofort nutzbar, kein App-Store, läuft auch auf älteren Android-Geräten
Betrieb	Container bei einem EU-Anbieter (STACKIT oder Scaleway), Konfiguration als Code	keine Server- oder Kubernetes-Pflege durch Ihr Team

2. Wissensdatenbank und Redaktion

Die Wissensdatenbank ist das Herz des Wegweisers. Ihre Qualität entscheidet über die Qualität jeder Antwort, deshalb gilt: **Deutsch ist die Quelle der Wahrheit**. Jeder Eintrag wird auf Deutsch gepflegt und fachlich von Ihrem Team verantwortet.

Kritische Informationen werden nicht als Fließtext gespeichert, sondern strukturiert. Fristen, Zuständigkeiten, Adressen, Öffnungszeiten, Formulare und Verfahrensschritte haben eigene Felder. Das hat zwei Vorteile. Ändert sich eine Öffnungszeit, ändert Ihr Team sie an einer einzigen Stelle. Und die KI setzt solche Angaben nicht frei um, sondern bekommt sie wörtlich aus dem Datenfeld.

Element	Inhalt
Themen	Ankommen, Asylverfahren, Wohnen, Gesundheit, Arbeit, Familie, Alltag (abgestimmt auf Ihre Kursinhalte)
Einträge	Frage, Antwort, Kurzfassung in Einfacher Sprache, Typ (FAQ, Infotext, Verfahrensschritt, Termin)
Strukturierte Fakten	Frist (Dauer, Bezug, Folgen bei Versäumnis), Zuständigkeit (Behörde, Beratungsstelle, Notdienst), Formular, Quelle mit Abrufdatum
Aktualität	gültig ab / bis, nächste Prüfung fällig am, zuletzt geprüft am, verantwortliche Person
Status	Entwurf → geprüft → freigegeben → deaktiviert

Freigabe im Vier-Augen-Prinzip: Wer einen Eintrag prüft, gibt ihn nicht selbst frei. Die KI sieht ausschließlich freigegebene und gültige Einträge. Technisch geschieht das über eine eigene Datenbanksicht, die auch für die Schnittstelle und die spätere MCP-Anbindung der einzige Lesezugang ist. Überfällige Einträge erscheinen in der Redaktion als Warnung.

Wissenslücken werden sichtbar: Fragen, auf die der Wegweiser keine geprüfte Antwort hatte, landen anonymisiert in einer Lückenliste. So sieht Ihr Team, welche Inhalte fehlen, und kann gezielt nachpflegen.

Bestandsinhalte: Ich übernehme die vorhandenen Kursmaterialien und FAQ-Listen beratend in die neue Struktur. Die fachliche Prüfung bleibt bei Ihrem Team.

3. Suche, Antworten, Sprachen und Qualität

So entsteht eine Antwort:

1. Die Frage wird ins Deutsche übertragen, denn dort liegt das geprüfte Wissen.
2. Eine kombinierte Suche (Bedeutung über Vektoren, Wortlaut über Volltext) findet passende freigegebene Einträge. Ein zweiter Bewertungsschritt kann sie zusätzlich nach Relevanz sortieren, wenn die Messung es rechtfertigt.
3. **Liegt die Relevanz unter einer festen Schwelle, kommt eine feste Antwort ohne KI-Aufruf:** „Dazu liegt keine geprüfte Information vor“, zusammen mit dem passenden Beratungskontakt. Die Ehrlichkeit des Systems hängt damit nicht vom Verhalten der KI ab.
4. Sonst formuliert das KI-Modell eine kurze Antwort in der Sprache der Fragenden. Deutsche Behördenbegriffe wie „Aufenthaltsgestattung“ stehen mit dabei, weil sie genau so in den Briefen stehen. Jede Antwort zeigt ihre Quelle und das Prüfdatum.
5. **Eine Prüfschicht ohne KI** gleicht Zahlen, Daten, Wochentage, Fristen und Behördennamen mit den Datenfeldern ab. Notrufnummern stammen immer aus geprüften Textbausteinen, nie aus der Modellausgabe. Schlägt eine Prüfung fehl, erscheint die geprüfte deutsche Kurzfassung mit Kontakt.

Krisen: Bei Hinweisen auf Gewalt, Suizidgedanken oder akute Gefahr erscheinen sofort fest hinterlegte Hilfsangebote und Notrufnummern in der jeweiligen Sprache, ohne Umweg über die KI. Die Stichwortlisten je Sprache entstehen gemeinsam mit Ihren Kursleitenden.

Sprachen: Ich messe, statt anzunehmen. Vor der Abgabe habe ich einen eigenen Vorab-Test durchgeführt: 11 Zielsprachen, 20 KI-Systeme, je 100 Sätze Deutsch→Zielsprache und 60 Sätze Zielsprache→Deutsch, dazu ein Test auf Zahlen, Fristen und Notrufnummern.

- **Ukrainisch, Arabisch, Farsi und Dari:** Offene Modelle, die EU-Anbieter als Managed Service betreiben, erreichen dort nach meiner Entscheidungsregel die Qualität der besten gemessenen proprietären Systeme. Diese offenen Modelle sind der Standard.
- **Tigrinya und Sorani:** Hier liegen offene Modelle bei EU-Anbietern belegt deutlich zurück. Für diese und einige weitere Sprachen sehe ich eine abschaltbare Ausnahme mit einem proprietären Modell in einer EU-Region vor, erst nach Ihrer Datenschutz-Folgenabschätzung.
- **Nicht getestet:** Fachtexte von Behörden. Die Testsätze waren allgemeinsprachlich, und gemessen habe ich über einen Vermittlungsdienst statt direkt bei den EU-Anbietern.

Deshalb fällt die Endauswahl in Phase 1:

- Ihre muttersprachlichen Kursleitenden prüfen je Sprache 50–80 echte Fragen aus dem Kursalltag.
- Eine Sprache wird erst nach ihrer Freigabe eingeschaltet. Bis dahin antwortet der Wegweiser in Einfacher Sprache auf Deutsch oder Englisch und verweist an Menschen.
- Nach jedem Modellwechsel laufen dieselben Tests erneut, automatisiert und nachvollziehbar.

4. Oberfläche, Barrierefreiheit und Datenschutz

Zugang ohne Hürde:

- QR-Code scannen, Sprache wählen (Sprachnamen in eigener Schrift, keine Flaggen), Frage stellen.
- Kein Konto, keine App-Installation, geringe Datenmengen für Prepaid-Tarife.
- Eine Notfallseite mit den wichtigsten Kontakten funktioniert auch offline.
- Ein gut sichtbarer Knopf „Mit einem Menschen sprechen“ führt jederzeit zu den Beratungsangeboten.

Barrierearm (Ziel: WCAG 2.2 AA):

- Schrift stufenlos vergrößerbar (A-/A+), hohe Kontraste.
- Vollständige Unterstützung für Schriften von rechts nach links (Arabisch, Farsi, Dari, Paschtu, Sorani) und für Ge'ez (Tigrinya).
- Bedienbar mit Tastatur und Screenreader, getestet auch mit TalkBack auf einem einfachen Android-Gerät.

Sprechen und Vorlesen:

- **Diktieren:** Die Nutzenden sprechen ihre Frage, sehen den erkannten Text und können ihn korrigieren. Für die großen Sprachen läuft offene Spracherkennung bei einem EU-Anbieter (Voxtral von Mistral oder Whisper, Auswahl per Messung). Aufnahmen werden nicht über die Erkennung hinaus gespeichert und nicht zum Training genutzt.
- **Vorlesen:** Die Antworten werden vorgelesen, mit Stimmen des Geräts oder eines EU-Dienstes. Dabei wird nur der geprüfte Antworttext verarbeitet, nie die Stimme der Nutzenden.
- **Sprachen mit wenig Trainingsdaten:** Für Tigrinya, Paschtu, Sorani, Kurmandschi, Dari und Somali sind gängige Modelle deutlich schwächer. Dafür starte ich in Phase 1 einen Pilot mit der spezialisierten offenen Spracherkennung Omnilingual ASR (Apache-2.0) auf einer EU-GPU, die nur bei Nutzung läuft. Eine Sprache wird erst nach Abnahmetest freigegeben. Bis dahin bleibt die Texteingabe gleichwertig, und Kerninformationen gibt es als eingesprochene Audiodateien.

KI-Transparenz: Ein dauerhaft sichtbarer Hinweis zeigt, dass eine KI antwortet und nur geprüfte Inhalte nutzt (Art. 50 EU AI Act).

Datenschutz durch Technik, nicht nur durch Erklärungen:

- **Anonym:** keine Konten, keine Profile, keine Cookies zur Wiedererkennung. Die Missbrauchsbremse arbeitet mit einem täglich wechselnden Einwegschlüssel statt mit IP-Adressen.
- **Datensparsam:** Persönliche Angaben wie Namen, Telefonnummern, Aktenzeichen oder IBAN werden vor Protokoll und KI-Aufruf automatisch entfernt. Das anonyme Frageprotokoll für die Auswertung der Beta wird nach 90 Tagen gelöscht, danach bleiben nur Zahlen je Thema.
- **Verarbeitung in der EU mit Auftragsverarbeitungsvertrag** und vertraglich ausgeschlossenem Training. Die Wissensdatenbank enthält keine personenbezogenen Daten der Nutzenden.
- **Unterlagen für Ihre Datenschutz-Folgenabschätzung** (Datenflüsse, technische und organisatorische Maßnahmen, Liste der Dienstleister) liefere ich als eigenen Meilenstein vor der Beta. Maßstab ist die DSGVO bzw. das kirchliche Datenschutzrecht (DSG-EKD), soweit es für Sie gilt.

5. Phasen, Offenheit und Betrieb

Phase	Schwerpunkte	Stunden (ca.)
1 Konzeption und Einrichtung	Workshop, Architekturentscheidung, Datenbank, Datenmodell, Redaktion, Suche, KI-Anbindung, Chat-Oberfläche, Audio, Schutz, Testgrundlage, erste Schulung	170–190
2 Test und Optimierung	Beta in ausgewählten Unterkünften, Freigabe je Sprache mit den Kursleitenden, Fehleranalyse, Optimierung, MCP-Schnittstelle (nur lesend)	120–140
3 Stabilisierung und Befähigung	Stabilisierung, Redaktionshandbuch, Videoanleitungen, Übergabe- und Exit-Test, Vorbereitung der Weiterentwicklung	90–110
Kontingent auf Abruf	Weitere Sprachen und Themen, Modellwechsel, Unvorhergesehenes	60–120

- **Abstimmung:** in Phase 1 alle zwei Wochen, danach monatlich.
- **Koordination:** Abstimmung, Stundennachweise und Zuarbeit zu Berichten weise ich offen aus (40–60 Stunden über die Laufzeit).
- **Abrechnung:** nur die tatsächlich geleisteten Stunden. Ich bringe eigene, in Produktion erprobte Bausteine mit, etwa Eingangsprüfung, Schwärzung persönlicher Angaben und Hybrid-Suche. Sie senken Aufwand und Risiko.

Kein Anbieter-Lock-in:

Komponente	Wechsel
Datenbank	Standard-PostgreSQL, Export und Umzug zu jedem Anbieter, auch zurück zu IONOS
KI-Modell	Konfiguration im Gateway
Embedding-Modell	Neuindexierung in Minuten
Redaktion	KI und Schnittstelle lesen nur über die Datenbanksicht
Oberfläche	dokumentierte Schnittstelle
Hosting	Container und Infrastruktur als Code

Konten und Rechte:

- Alle Konten laufen auf die Proclusio gGmbH und sind mit Zwei-Faktor-Anmeldung geschützt.
- Projektspezifische Entwicklungen gehen vollständig an Sie.
- An eingebrachten Bausteinen erhalten Sie ein dauerhaftes, übertragbares Nutzungsrecht. Der Weiterbetrieb ist damit auch durch Dritte jederzeit gesichert.
- Auf Wunsch kann der Projektcode unter einer Open-Source-Lizenz (EUPL) veröffentlicht werden.

Laufende Kosten: voraussichtlich **ca. 100–300 € pro Monat** im Vollausbau, je nach Nutzung. Die Spanne setzt sich gerundet so zusammen:

- Datenbank und Redaktion: 60–160 €
- KI-Modelle: 20–110 € bei 300–2.000 Gesprächen
- Hosting: 15–50 €

Kosten- und Störungsalarme sind von Anfang an eingerichtet.

Risiko	Gegenmaßnahme
Veraltete Fakten	Prüfdatum und Gültigkeit je Eintrag, Warnungen, nur gültige Inhalte im Wegweiser
Falsche Antwort in einer Sprache	Freigabe je Sprache durch Kursleitende, Prüfschicht ohne KI, „keine geprüfte Information“ statt Raten
Anbieter stellt ein Modell ein	Gateway, Neutest per Knopfdruck, Ersatzmodell je Sprache festgelegt
Missbrauch des offenen Zugangs	Missbrauchsbremse, Schutz vor Manipulationsversuchen, Kostenobergrenze, Not-Aus
Abhängigkeit von mir als Person	Dokumentation, Runbooks, Übergabetest. Ziel: Eine fremde Fachkraft setzt das System in unter zwei Stunden aus dem Repository neu auf

Ende Kapitel 5 · Technisches Vorgehen

Befähigung und Nachhaltigkeit

Ziel: Ihr Projektteam pflegt den Wegweiser nach der Einführung selbstständig. Es versteht, was die KI kann und was nicht, und ist nicht von mir abhängig.

Redaktion ohne Technikwissen

- Inhalte anlegen, ändern, prüfen, freigeben und deaktivieren erfolgt über Formulare, nicht über SQL.
- Pflichtfelder für Quelle, Prüfdatum und verantwortliche Person verhindern unvollständige Einträge.
- Ein Testbereich zeigt vor der Freigabe, wie der Wegweiser mit einem neuen Eintrag antwortet.
- Die Lückenliste zeigt, welche Fragen noch unbeantwortet sind.

Schulung und Begleitung

Angebot	Umfang	Zeitpunkt
Einführungsworkshop Redaktion	ca. 3 h, praktisch an echten Inhalten	Ende Phase 1
Workshops KI-Kompetenz (Art. 4 EU AI Act): Was kann die KI, wo sind die Grenzen, wie erkenne ich Fehler	2 × 2 h mit Teilnahmenachweis	Phase 2
Einweisung der Kursleitenden in die Sprachprüfung	ca. 1,5 h + Vorlagen	vor der Beta
Monatliche Sprechstunde (Video)	ca. 1 h	über die gesamte Laufzeit

Dokumentation, die gelesen wird

- **Redaktionshandbuch** mit Bildschirmfotos, in einfacher Sprache.
- **6–8 kurze Videoanleitungen** (je 3–5 Minuten) zu den häufigsten Aufgaben.
- **Runbooks nach Symptom** („Der Wegweiser antwortet nicht“, „Eine Antwort ist falsch“), jeweils mit klaren Schritten und Ansprechpartnern.
- **Architektur- und Betriebsdokumentation** für eine spätere IT-Fachkraft oder einen anderen Dienstleister.

Nachhaltigkeit über die Laufzeit hinaus

- **Alles gehört Ihnen:** Konten, Daten und Code liegen bei der Proclusio gGmbH, mit Zwei-Faktor-Anmeldung. Projektspezifische Entwicklungen gehen vollständig an Sie, an eingebrachten Bausteinen erhalten Sie ein dauerhaftes, übertragbares Nutzungsrecht.
- **Übergabetest:** Vor dem Ende von Phase 3 setzt eine fremde Fachkraft das System nur mit der Dokumentation aus dem Repository neu auf. Ziel sind unter zwei Stunden. Was dabei hakt, wird dokumentiert und behoben.

- **Offene Standards:** PostgreSQL, OpenAPI und MCP sowie eine OpenAI-kompatible Schnittstelle halten die Weiterentwicklung offen. Neue Kanäle oder Anwendungen können an dieselbe Wissensbasis anschließen.
- **Qualität bleibt messbar:** Die Testsets der Kursleitenden und die automatischen Prüfungen bleiben im System. Nach jeder Inhalts- oder Modelländerung lassen sie sich erneut ausführen.

Aufwand, Zeitplan und Preis

Stundensatz

	je Stunde
Netto	100,00 €
Umsatzsteuer (19 %)	19,00 €
Brutto	119,00 €

Der Satz ist fest für die gesamte Laufzeit und umfasst sämtliche Aufwendungen, auch Fahrtzeiten in Berlin, Arbeitsmittel und Software. Abgerechnet werden nur tatsächlich geleistete, nachgewiesene Stunden mit Tätigkeitsbeschreibung, passend für die Nachweise gegenüber dem Fördermittelgeber. Die laufenden Betriebskosten (Hosting, KI-Dienste, Domain) laufen über Konten der Proclusio gGmbH und sind nicht Teil des Stundensatzes. Ich schätze sie im Vollausbau auf ca. 100–300 € im Monat.

Aufwand je Phase (Obergrenze 500 Stunden)

Phase	Inhalt	Stunden (ca.)
1 Konzeption und Einrichtung	Workshop, Architektur, Datenbank, Datenmodell, Redaktion, Suche, KI-Anbindung, Oberfläche, Audio, Schutz, Testgrundlage, DSFA-Unterlagen, erste Schulung	170–190
2 Test und Optimierung	Beta in Unterkünften, Freigabe je Sprache mit den Kursleitenden, Fehleranalyse, Optimierung, MCP-Schnittstelle	120–140
3 Stabilisierung und Befähigung	Stabilisierung, Handbuch, Videos, Sprechstunde, Übergabe- und Exit-Test, Weiterentwicklung	90–110
Kontingent auf Abruf	Weitere Sprachen und Themen, Modellwechsel, Rechtsänderungen, Unvorhergesehenes	60–120
Summe		höchstens 500

Mein Planwert für die drei Phasen liegt bei rund 380–440 Stunden. Die Differenz zum Deckel von 500 Stunden ist ein Kontingent, das nur auf Ihren Abruf anfällt. Im Planwert enthalten sind 40–60 Stunden für Abstimmung, Stundennachweise und Zuarbeit zu Berichten. Eigene, erprobte Bausteine senken den Aufwand in Phase 1 um rund 40 Stunden gegenüber einem Neubau. Ein Anspruch auf Ausschöpfung der 500 Stunden besteht nicht.

Zeitplan (ab Beauftragung)

Zeitraum	Meilenstein
----------	-------------

Zeitraum	Meilenstein
Monat 1	Auftakt-Workshop, Architekturentscheidung, Konten auf die Proclusio gGmbH
Monat 1–2	Datenbank, Datenmodell und Redaktion stehen; Ihr Team beginnt mit der Pflege
Monat 2–3	Suche, KI-Anbindung, Chat-Oberfläche, Audio; erste Sprachen im internen Test
Monat 2–3	Unterlagen für die Datenschutz-Folgenabschätzung, Sicherheits- und Barrierefreiheitstests, Schulung Redaktion
ab ca. Monat 3–4	Beta in ausgewählten Unterkünften (nach Freigabe der DSFA), Freigabe je Sprache
Monat 4–18	Optimierung nach echten Fragen, weitere Sprachen, MCP-Schnittstelle, Workshops KI-Kompetenz
bis 30.06.2029	Stabilisierung, Dokumentation, Übergabetest, monatliche Sprechstunde

In Phase 1 arbeite ich mit ca. 20–30 Stunden pro Woche, danach bedarfsorientiert. Den Start der Beta bestimmen vor allem die Freigabe Ihrer Datenschutz-Folgenabschätzung und die ersten gepflegten Inhalte; ich passe den Plan an Ihren Wunschtermin an. Abstimmung: in Phase 1 alle zwei Wochen, danach monatlich. Vor-Ort-Termine bündele ich.

Abrechnung und Kostenkontrolle

- Monatliche Abrechnung mit Stundennachweis je Tätigkeit und Phase.
- Ein Stundenstand je Phase ist jederzeit einsehbar. Bei 80 % einer Phase melde ich mich, bevor weitere Stunden anfallen.
- Kosten- und Störungsalarme für die laufenden Dienste richte ich von Anfang an ein.
- **Kostenbremse:** Für die laufenden Dienste vereinbaren wir eine monatliche Obergrenze. Bei 80 % meldet ein Alarm, bevor weitere Kosten entstehen.
- **Messbudget Phase 1:** Für die Messungen in Phase 1 (Modellvergleich je Sprache, Suchqualität, Test des Schutzfilters) setze ich einmalig ca. 50–100 € an Nutzungskosten der KI-Dienste an, ebenfalls über Ihre Konten. Zum Vergleich: Mein Vorab-Benchmark mit 20 Systemen und 11 Zielsprachen kostete rund 11 US-Dollar.

Anlage A1

Modelle, Sprachen und Open Source: Was ich vorschlage und warum

So findet der Wegweiser die richtige Antwort

Der Wegweiser sucht zuerst in Ihren geprüften Inhalten, und die KI formuliert nur aus dem, was diese Suche gefunden hat.

- **Deutsch ist die Quelle der Wahrheit.** Gesucht wird nur in Einträgen, die Ihr Team auf Deutsch geprüft und freigegeben hat. Deshalb wird jede Frage zuerst ins Deutsche übertragen, wörtlich und ohne Umformulieren.
- **Zwei Suchwege.** Die Bedeutungssuche findet passende Einträge, auch wenn die Frage andere Worte benutzt. Die Wortsuche findet Behördenbegriffe, Formularnamen und Tippfehler. Beide Trefferlisten werden zu einer zusammengeführt.
- **Ein zweiter Blick, wenn er sich lohnt.** Ein spezialisiertes Modell kann die Treffer zusätzlich neu bewerten. Ob das die Trefferquote messbar verbessert, entscheidet die Messung in Phase 1.
- **Ehrlich ohne KI.** Liegt selbst der beste Treffer unter einer festen Schwelle, antwortet der Wegweiser ohne KI: „Dazu liegt keine geprüfte Information vor“, mit dem passenden Beratungskontakt.
- **Sicherheit ohne KI.** Krisenhinweise erkennen feste Stichwortlisten je Sprache. Zahlen, Daten und Fristen gleicht eine feste Prüfung ab, Notrufnummern stammen nie aus der KI. Ein KI-Schutzfilter kommt nur dazu, wenn ein Test in Phase 1 zeigt, dass er hilft, ohne harmlose Fragen zu blockieren.
- **Die Messung wählt die Modelle.** Welches Modell eine Sprache versteht und beantwortet, entscheiden die Tests mit Ihren Kursleitenden. Alle Modelle hängen an einem Gateway mit Standardschnittstelle. Ein Wechsel ist eine Konfiguration, danach laufen dieselben Tests.

Der Kern läuft bei STACKIT in Deutschland: Datenbank, Bedeutungssuche und Antwortmodell. Weitere Dienste kommen nur dazu, wenn die Messung sie rechtfertigt. Die Kandidaten sind offene Modelle unter Apache-2.0-Lizenz oder Bausteine ganz ohne KI. Ihre Herkunft nenne ich offen.

Stufe	Aufgabe	Kandidat	Ersatz oder Option	Warum
Schutz am Eingang	erkennt Krisenhinweise, Missbrauch und persönliche Angaben	feste Krisen-Stichwortlisten je Sprache und Eingangsprüfung, kein Modell	KI-Schutzfilter nur nach bestandenem Fehlalarm-Test: Shieldstral 1.0 (Mistral) oder Qwen3Guard (Alibaba, bei OVH), jeweils ohne die Kategorie „politisch heikel“	verlässlich und nachvollziehbar; ein Filter darf keine Fragen zu Verfolgung oder Religion blockieren

Stufe	Aufgabe	Kandidat	Ersatz oder Option	Warum
Übertragung	Frage wörtlich ins Deutsche	Gemma 4 31B (Google) bei STACKIT; für Hocharabisch, Farsi und Dari Mistral Large 3 (Mistral, EU-Endpunkt)	weiteres offenes Modell je Sprache; proprietäre Ausnahme nur einzeln, abschaltbar, nach DSFA	im Deutschen liegt das geprüfte Wissen; Modell je Sprache und Richtung aus meinem Vorab-Test
Bedeutungssuche (Embedding)	findet Einträge mit ähnlicher Bedeutung	Qwen3-VL-Embedding-8B (Alibaba) bei STACKIT	bge-m3 (BAAI, MIT) bei IONOS; ohne chinesisches Modell: multilingual-e5-large-instruct (Microsoft, MIT) bei Berget AI, Schweden	beim Kernanbieter verfügbar; sucht in der deutschen Fassung
Wortsuche (Volltext, Trigramm)	findet Begriffe wörtlich und trotz Tippfehlern	PostgreSQL-Volltextsuche (Deutsch) und pg_trgm, kein Modell	in jeder PostgreSQL-Datenbank, auch bei IONOS	„Aufenthaltsgestattung“ muss genau so gefunden werden; trägt allein weiter, wenn die Bedeutungssuche ausfällt
Rangfusion	führt beide Trefferlisten zusammen	Reciprocal Rank Fusion in SQL, kein Modell	–	einfach und nachvollziehbar, ohne Gewichte, die man pflegen müsste
Neubewertung (Reranker)	ordnet die besten rund 20 Treffer neu	Option, siehe unten; offen und bei einem deutschen Anbieter: Qwen3-VL-Reranker-8B (Alibaba) bei IONOS	ohne Neubewertung: Rangfusion und Schwelle	kommt nur dazu, wenn sie die Trefferquote messbar verbessert
Schwelle	antworten oder ehrlich ablehnen	feste Regel, kein Modell	–	die Ehrlichkeit hängt nicht vom Verhalten der KI ab

Stufe	Aufgabe	Kandidat	Ersatz oder Option	Warum
Antwortmodell	formuliert aus den Treffern in der Sprache der Fragenden	Gemma 4 31B (Google) bei STACKIT	Mistral Large 3 oder Qwen3.5 (IONOS) je nach Messung; proprietäre Ausnahme nur einzeln, abschaltbar, nach DSFA	Auswahl je Sprache per Messung mit den Kursleitenden
Prüfschicht	gleicht Zahlen, Daten, Fristen und Behördennamen ab; Notrufnummern nur aus geprüften Textbausteinen	eigener Code, kein Modell	schlägt die Prüfung fehl: geprüfte deutsche Kurzfassung mit Kontakt	fängt Fehler, die keine Modellwahl verhindert, etwa umgekehrte Fristen

Optionen und technische Details

Neubewertung: Optionen. Alle Modelle sind laut Hersteller mehrsprachig; der Wegweiser braucht nur Deutsch. In Phase 1 messe ich zuerst ohne Neubewertung und dann mit höchstens zwei Optionen.

Option	Herkunft	Lizenz	Verarbeitung in der EU	Einschätzung
ohne Neubewertung	–	kein Modell	keine Daten an Dritte	Ausgangspunkt der Messung
Qwen3-VL-Reranker-8B	Alibaba (China)	offen (Apache-2.0)	IONOS, Deutschland	offen, deutscher Anbieter; derzeit nur dort
bge-reranker-v2-m3	BAAI (China)	offen (Apache-2.0)	Berget AI, Schweden	offen; junger Anbieter mit öffentlichem AVV
Cohere Rerank 3.5	Cohere (Kanada)	proprietär	AWS Bedrock, nur Region Frankfurt	nicht chinesisch; US-Cloud, deshalb nur nach DSFA
Cohere Rerank 4.0 Fast	Cohere (Kanada)	proprietär	Oracle Cloud, Frankfurt	wie Rerank 3.5, neuere Version
Jina Reranker v3	Jina AI (Berlin, Teil von Elastic), auf Basis Qwen3	Gewichte nur nicht kommerziell; nutzbar über Elastic	Elastic Inference Service, EU-Regionen	braucht eine zusätzliche Plattform

Option	Herkunft	Lizenz	Verarbeitung in der EU	Einschätzung
Voyage rerank	Voyage AI (MongoDB, USA)	proprietär	MongoDB Atlas, Region Europa	Speicherung und Training vertraglich ausschließen
mxbai-rerank-v2	Mixedbread (Berlin), auf Basis Qwen2.5	offen (Apache-2.0)	kein eigener Dienst, nur in der Suchplattform des Herstellers	einzelnen nur im Eigenbetrieb, daher nicht vorgesehen

Abschnitte und Metadaten. Kurze Einträge bleiben ganz. Lange Infotexte teile ich an Zwischenüberschriften in Abschnitte, jeweils mit dem Titel des Eintrags davor. Jeder Abschnitt trägt Eintrag, Thema, Sprache, Gültigkeit, Prüfdatum, Quellen sowie Name und Version des Embedding-Modells. Fristen und Zuständigkeiten kommen aus den Datenfeldern, nicht aus dem Text.

Nur Freigegebenes. Die Suche darf über eine eigene Datenbankrolle nur die Sicht der freigegebenen und gültigen Einträge lesen. Vektoren entstehen bei der Freigabe. Läuft ein Eintrag ab oder wird er deaktiviert, fällt er sofort aus der Sicht.

Übertragen, nicht umformulieren. In meiner Referenz 2 hat ein Umformulieren der Frage durch die KI die Trefferquote gesenkt. Die Übertragung bleibt deshalb wörtlich.

Suche. Bedeutungssuche über einen HNSW-Index in pgvector mit 1.024 statt 4.096 Werten je Vektor, weil der Index höchstens 2.000 fasst. Wortsuche mit deutscher Wortstammbildung. Startwerte: Je Suchweg gehen die besten 30 Treffer in die Rangfusion ($k = 60$), die besten 20 gegebenenfalls in die Neubewertung, die besten drei bis fünf an das Antwortmodell.

Neubewertung und Schwelle. Mit Neubewertung setzt die Schwelle an deren Wert an, sonst an der Bedeutungsähnlichkeit. Ich kalibriere sie in Phase 1 mit beantwortbaren und bewusst unbeantwortbaren Testfragen. Im Zweifel lieber „keine geprüfte Information“ als eine schwache Antwort. Eine Option kommt nur dazu, wenn sie die Trefferquote messbar hebt. Jeder zusätzliche Anbieter braucht einen AVV und steht in den Unterlagen für Ihre DSFA.

Messung. Recall@5: Bei wie vielen Testfragen der Kursleitenden steht der richtige Eintrag unter den ersten fünf Treffern? Dazu die Ablehnungsquote bei unbeantwortbaren Fragen und ein Erfindungstest: Wie oft enthält eine Antwort Aussagen, die in keinem Treffer stehen? Die Tests laufen nach jeder Änderung automatisch. In Referenz 2 fand eine gleich aufgebaute Suche mit anderen Modellen im eigenen Testset bei allen Testfragen den richtigen Eintrag unter den ersten fünf. Das war ein kleiner, rein deutscher Bestand, also ein Ausgangspunkt, kein Beleg für den Wegweiser.

Modellwechsel. Ein neues Embedding-Modell baut einen zweiten Index parallel auf. Umgeschaltet wird erst, wenn die Messung mindestens gleich gut ist. Bei einigen tausend Einträgen dauert das Minuten.

Herkunft. Die Kandidaten für Bedeutungssuche und Neubewertung stammen aus China und laufen nur bei EU-Anbietern. Ohne chinesisches Modell arbeiten Wortsuche und Rangfusion; als Modelle stehen multilingual-e5 bei Berget AI und Cohere Rerank in Frankfurt bereit.

Schutzfilter. Qwen3Guard kennt eine Kategorie „politisch heikle Themen“. Für Geflüchtete sind das Kernfragen, deshalb bliebe sie abgeschaltet. Vor einem Einsatz prüfe ich mit echten Fragen zu Asyl, Verfolgung und Religion, wie oft ein Filter fälschlich anschlägt.

Kosten. Bedeutungssuche und Neubewertung sind ein kleiner Teil der laufenden Kosten (Rahmen siehe Kapitel 5, Kostenbremse und Messbudget siehe Preis).

Ausgangspunkt: die von Ihnen geprüfte IONOS-Lösung

Ihre Projektbeschreibung nennt als Beispiel IONOS Managed PostgreSQL und den IONOS AI Model Hub. Das Prinzip übernehme ich vollständig: Datenbank mit Vektorsuche, KI-Modell über eine Standardschnittstelle, schlanke Weboberfläche. IONOS bleibt für Datenbank und KI-Modelle ein gleichwertiger Anbieter. Durch das Gateway kann jedes IONOS-Modell eingesetzt werden, sobald es den Test einer Sprache besteht.

Warum ich nicht allein auf den IONOS AI Model Hub setze

- 1. Sprachen:** Für die Zielgruppe zählen gerade die Sprachen, die KI-Modelle selten gut beherrschen. In meinem Vorab-Test haben zwei Modelle des Model Hub bei schweren Sprachen typische Fehler gezeigt: Llama 3.3 und Mistral Small 3.2 lieferten Text in fremder Schrift, falsche Wochentage und eine umgekehrte Frist („nach 14 Tagen“ statt „innerhalb von 14 Tagen“). Für Ukrainisch, Arabisch, Farsi und Dari erreichen dagegen offene Modelle bei anderen EU-Anbietern (Gemma 4 bei STACKIT, Mistral Large 3 über den EU-Endpunkt von Mistral) nach meiner Entscheidungsregel die Qualität der besten gemessenen proprietären Systeme.
- 2. Modellwechsel:** Anbieter nehmen Modelle regelmäßig vom Netz, auch IONOS. Ein Gateway mit festgelegtem Ersatzmodell je Sprache und automatischem Neutest macht das zur Routineaufgabe.
- 3. Sprache und Audio:** Der Model Hub enthält keine Spracherkennung und keine Sprachausgabe. Diktieren und Vorlesen brauchen einen zusätzlichen EU-Dienst.
- 4. Betrieb der Anwendung:** Für eine Web-Anwendung mit eigener Schnittstelle bräuchte es bei IONOS einen Kubernetes-Cluster oder eine virtuelle Maschine, also Serverpflege. STACKIT bzw. Scaleway bieten Container ohne Serverpflege.

Mein Vorschlag: Den Kern betreibe ich bei **STACKIT**: Datenbank, Bedeutungssuche, Antwortmodell und Container bei einem deutschen Anbieter mit C5-Testat. Zusatzdienste kommen nur dazu, wenn die Messung sie rechtfertigt, etwa Mistral Large 3 für das Verstehen einzelner Sprachen oder eine Neubewertung der Suchtreffer; jeder hat einen eigenen Auftragsverarbeitungsvertrag. IONOS ist der gleichwertige Ersatz. Welches Modell eine Sprache bedient, entscheidet die Messung, nicht der Anbieter.

Wie Open Source eingesetzt wird

Baustein	Open Source	Lizenz
Datenbank	PostgreSQL + pgvector (als Managed Service)	PostgreSQL-Lizenz
Redaktion	Payload CMS	MIT
KI-Gateway	LiteLLM	MIT
Chat-Oberfläche	assistant-ui, AI SDK	MIT / Apache-2.0

Baustein	Open Source	Lizenz
Spracherkennung	Voxtral (Mistral) bzw. Whisper bei einem EU-Anbieter; für schwere Sprachen Omnilingual ASR (Pilot in Phase 1)	Apache-2.0 / MIT
KI-Modelle (Standard)	Gemma 4 31B, Mistral Large 3	Apache-2.0

Alle Kernbausteine sind Open Source und laufen als Managed Service bei EU-Anbietern. Ihr Team betreibt keine Server. Der Projektcode kann auf Wunsch unter der EU-Lizenz EUPL veröffentlicht werden.

Wo Open Source heute an Grenzen stößt

Vor der Abgabe habe ich einen eigenen, reproduzierbaren Vorab-Benchmark durchgeführt:

- FLORES-200 (devtest), 11 Zielsprachen, 20 Systeme
- je 100 Sätze Deutsch→Zielsprache und 60 Sätze Zielsprache→Deutsch
- chrF++ mit gepaartem Bootstrap sowie ein Fakten- und Krisennummern-Test
- Die Auswertung habe ich mit einem zweiten, unabhängig geschriebenen Skript exakt nachgerechnet.

Ergebnis: Nach der vorab festgelegten Entscheidungsregel erreicht in 11 von 22 Sprach-Richtungs-Kombinationen ein offenes, in der EU als Managed Service verfügbares Modell die Gleichwertigkeit.

- **Statistisch abgesichert** ist das für Ukrainisch, Arabisch (Levante, Irak) und Hocharabisch, jeweils in Richtung Deutsch→Zielsprache. Den Faktentest habe ich dafür offengelegt korrigiert: Zahlwort, Dual und Ordinalzahl gelten als korrekt.
- **Belegt zurück** liegt das beste EU-einsetzbare offene Modell bei Tigrinya (beide Richtungen) und Sorani→Deutsch, um mehr als 5 % hinter dem besten proprietären System. Dort sehe ich eine abschaltbare proprietäre Ausnahme in einer EU-Region mit Datenschutz-Folgenabschätzung vor.
- **Nicht statistisch bewiesen:** In fünf weiteren Kombinationen ergibt sich die Ausnahme nur aus der Regel, drei Kombinationen sind offen. Beide Gruppen prüfe ich in Phase 1 neu.
- **Deutsch→Zielsprache:** Über alle offenen Modelle war in keiner Sprache ein proprietäres System signifikant besser. Die Lücke liegt beim EU-Hosting-Angebot, nicht bei der Qualität offener Modelle.

Krisennummern setze ich unabhängig vom Modell aus geprüften Textbausteinen. Proprietäre Ausnahmen sind an Modell und Version gebunden, abschaltbar und werden bei jedem Versionswechsel neu gemessen.

Grenzen des Tests: allgemeinsprachliche Testsätze, eine Messung je System, Messung über einen API-Vermittler statt bei den EU-Hosts. Deshalb erfolgt die Endauswahl per Messung in Phase 1, gemeinsam mit den Kursleitenden.

Sprachen	Vorschlag zum Start	Status
Ukrainisch, Arabisch, Farsi, Dari	offenes Modell bei EU-Anbieter (Gemma 4 / Mistral Large 3)	gleichwertig, teils statistisch abgesichert
Kurmandschi, Somali	Ausnahme, bis ein EU-Anbieter ein gleichwertiges offenes Modell vertraglich anbietet	offen, Klärung läuft

Sprachen	Vorschlag zum Start	Status
Sorani, Paschtu	abschaltbare proprietäre Ausnahme in der EU, nach DSFA	knapp bzw. belegt
Tigrinya	proprietäre Ausnahme plus geprüfte Textbausteine, Übergabe an Menschen	belegt
Deutsch, Englisch, Französisch, Russisch, Türkisch u. a.	offenes Modell bei EU-Anbieter	Messung in Phase 1

Wichtig für Ihre Datenschutz-Folgenabschätzung: Beim Verstehen der Frage gehen Eingaben der Nutzenden an das Modell, die besonders sensible Angaben enthalten können. Jede proprietäre Ausnahme in dieser Richtung wäge ich deshalb gesondert ab und schlage sie nur mit Ihrer Freigabe vor.

Vorab-Benchmark, Stand 11.10.2026: 22 Kombinationen aus Sprache und Richtung. In 11 erreicht ein offenes, in der EU als Managed Service verfügbares Modell die Gleichwertigkeit (davon 5 statistisch abgesichert), in 8 sehe ich eine proprietäre Ausnahme vor (davon 3 statistisch abgesichert), 3 sind offen.

Sprache und Richtung	Urteil und Gewissheit	Vorschlag zum Start	Ersatz	Qualität relativ zum besten System
Ukrainisch Українська Antwort: Deutsch → Ukrainisch	Offenes EU-Modell gleichwertig ● statistisch abgesichert	Gemma 4 31B (offen, STACKIT; Scaleway nur dediziert)	Qwen3.5-397B (offen, IONOS, OVH, Scaleway), Mistral Large 3 (offen, Mistral (EU-Endpunkt); Scaleway nur dediziert)	Offen, EU-Anbieter (T1/T1d): Mistral Large 3 100,0 % Proprietär, EU-Region: Gemini 3.1 Flash-Lite 99,9 %
Ukrainisch Українська Verstehen: Ukrainisch → Deutsch	Offenes EU-Modell gleichwertig ● statistisch abgesichert	Gemma 4 31B (offen, STACKIT; Scaleway nur dediziert)	Mistral Large 3 (offen, Mistral (EU-Endpunkt); Scaleway nur dediziert), Mistral Small 3.2 24B (offen, IONOS, Scaleway)	Offen, EU-Anbieter (T1/T1d): Gemma 4 31B 97,5 % Proprietär, EU-Region: Gemini 3.1 Flash-Lite 100,0 %
Hocharabisch العربية الفصحى Antwort: Deutsch → Hocharabisch	Offenes EU-Modell gleichwertig ● statistisch ₂ abgesichert	Gemma 4 31B (offen, STACKIT; Scaleway nur dediziert)	Gemini 3.1 Flash-Lite (proprietär, Vertex AI EU) Kein offener Ersatz erfüllt hier die Regel.	Offen, EU-Anbieter (T1/T1d): Gemma 4 31B 98,3 % Proprietär, EU-Region: Gemini 3.1 Flash-Lite 100,0 %
Hocharabisch العربية الفصحى Verstehen: Hocharabisch → Deutsch	Offenes EU-Modell gleichwertig ● knapp	Mistral Large 3 (offen, Mistral (EU-Endpunkt); Scaleway nur dediziert)	Claude Sonnet 5.5 (proprietär, AWS Bedrock Frankfurt) Nur bedingt: DeepSeek V4.1 Flash über GreenPT, Vertrag zur Auftragsverarbeitung angefragt.	Offen, EU-Anbieter (T1/T1d): Mistral Large 3 96,0 % GreenPT (Vertrag angefragt): DeepSeek V4.1 Flash 96,4 % Proprietär, EU-Region: Claude Sonnet 5.5 100,0 %

Sprache und Richtung	Urteil und Gewissheit	Vorschlag zum Start	Ersatz	Qualität relativ zum besten System
Arabisch (Levante) شامي Antwort: Deutsch → Arabisch (Levante)	Offenes EU-Modell gleichwertig • statistisch ₂ abgesichert	Gemma 4 31B (offen, STACKIT; Scaleway nur dediziert)	Gemini 3.1 Flash-Lite (proprietär, Vertex AI EU), Claude Sonnet 5.5 (proprietär, AWS Bedrock Frankfurt)	Offen, EU-Anbieter (T1/T1d): Gemma 4 31B 98,2 % Proprietär, EU-Region: Gemini 3.1 Flash-Lite 100,0 %
Arabisch (Levante) شامي Verstehen: Arabisch (Levante) → Deutsch	Offenes EU-Modell gleichwertig ● knapp	Gemma 4 31B (offen, STACKIT; Scaleway nur dediziert)	Gemini 3.8 Flash (proprietär, Vertex AI EU) Nur bedingt: GLM-5.3 über GreenPT, Vertrag zur Auftragsverarbeitung angefragt.	Offen, EU-Anbieter (T1/T1d): Gemma 4 31B 97,1 % GreenPT (Vertrag angefragt): GLM-5.3 97,1 % Proprietär, EU-Region: Gemini 3.8 Flash 100,0 %
Arabisch (Irak) عراقي Antwort: Deutsch → Arabisch (Irak)	Offenes EU-Modell gleichwertig • statistisch ₂ abgesichert	Gemma 4 31B (offen, STACKIT; Scaleway nur dediziert)	Gemini 3.1 Flash-Lite (proprietär, Vertex AI EU), Claude Sonnet 5.5 (proprietär, AWS Bedrock Frankfurt)	Offen, EU-Anbieter (T1/T1d): Gemma 4 31B 99,8 % Proprietär, EU-Region: Gemini 3.1 Flash-Lite 100,0 %
Arabisch (Irak) عراقي Verstehen: Arabisch (Irak) → Deutsch	Offen / unentschieden ● knapp	Gemini 3.8 Flash (proprietär, Vertex AI EU), übergangsweise	Mistral Large 3 ist hier noch nicht vollständig gemessen; das hole ich in Phase 1 nach. Nur bedingt: GLM-5.3, DeepSeek V4.1 Flash über GreenPT, Vertrag zur Auftragsverarbeitung angefragt.	Offen, EU-Anbieter (T1/T1d): Gemma 4 31B 93,5 % GreenPT (Vertrag angefragt): GLM-5.3 96,0 % Proprietär, EU-Region: Gemini 3.8 Flash 100,0 %
Farsi فارسی Antwort: Deutsch → Farsi	Offenes EU-Modell gleichwertig ● knapp	Gemma 4 31B (offen, STACKIT; Scaleway nur dediziert)	Qwen3.5-397B (offen, IONOS, OVH, Scaleway, mit vereinheitlichten Ziffern), Gemini 3.1 Flash-Lite (proprietär, Vertex AI EU)	Offen, EU-Anbieter (T1/T1d): Gemma 4 31B 95,6 % Proprietär, EU-Region: Gemini 3.1 Flash-Lite 100,0 %
Farsi فارسی Verstehen: Farsi → Deutsch	Offenes EU-Modell gleichwertig ● knapp	Mistral Large 3 (offen, Mistral (EU-Endpunkt); Scaleway nur dediziert)	Gemini 3.8 Flash (proprietär, Vertex AI EU)	Offen, EU-Anbieter (T1/T1d): Mistral Large 3 95,5 % Proprietär, EU-Region: Gemini 3.8 Flash 100,0 %
Dari دري Antwort: Deutsch → Dari	Offenes EU-Modell gleichwertig ● knapp	Gemma 4 31B (offen, STACKIT; Scaleway nur dediziert)	Qwen3.5-397B (offen, IONOS, OVH, Scaleway), Gemini 3.8 Flash (proprietär, Vertex AI EU) Nicht vorgesehen: Llama 3.3 70B.	Offen, EU-Anbieter (T1/T1d): Llama 3.3 70B 97,7 % Proprietär, EU-Region: Gemini 3.8 Flash 97,6 %

Sprache und Richtung	Urteil und Gewissheit	Vorschlag zum Start	Ersatz	Qualität relativ zum besten System
Dari دری Verstehen: Dari → Deutsch	Offenes EU-Modell gleichwertig ● knapp	Mistral Large 3 (offen, Mistral (EU-Endpunkt); Scaleway nur dediziert)	Gemma 4 31B (offen, STACKIT; Scaleway nur dediziert), Gemini 3.8 Flash (proprietär, Vertex AI EU)	Offen, EU-Anbieter (T1/T1d): Mistral Large 3 96,7 % Proprietär, EU-Region: Gemini 3.8 Flash 100,0 %
Kurmandschi Kurmancî Antwort: Deutsch → Kurmandschi	Offen / unentschieden ● knapp	Gemini 3.1 Flash-Lite (proprietär, Vertex AI EU), übergangsweise	Nur bedingt: DeepSeek V4.1 Flash über GreenPT, Vertrag zur Auftragsverarbeitung angefragt. Nicht vorgesehen: Llama 3.3 70B, Mistral Small 3.2 24B.	Offen, EU-Anbieter (T1/T1d): Qwen3.5-397B 93,9 % GreenPT (Vertrag angefragt): DeepSeek V4.1 Flash 97,3 % Proprietär, EU-Region: Gemini 3.1 Flash-Lite 100,0 %
Kurmandschi Kurmancî Verstehen: Kurmandschi → Deutsch	Proprietäre Ausnahme ● knapp	Gemini 3.8 Flash (proprietär, Vertex AI EU)	Claude Sonnet 5.5 (proprietär, AWS Bedrock Frankfurt)	Offen, EU-Anbieter (T1/T1d): Qwen3.5-397B 90,1 % Proprietär, EU-Region: Gemini 3.8 Flash 100,0 %
Somali Soomaali Antwort: Deutsch → Somali	Offen / unentschieden ● knapp	Gemini 3.1 Flash-Lite (proprietär, Vertex AI EU), übergangsweise	Claude Sonnet 5.5 (proprietär, AWS Bedrock Frankfurt) Nur bedingt: DeepSeek V4.1 Flash über GreenPT, Vertrag zur Auftragsverarbeitung angefragt. Nicht vorgesehen: Llama 3.3 70B.	Offen, EU-Anbieter (T1/T1d): Qwen3.5-397B 94,9 % GreenPT (Vertrag angefragt): DeepSeek V4.1 Flash 96,7 % Proprietär, EU-Region: Gemini 3.1 Flash-Lite 100,0 %
Somali Soomaali Verstehen: Somali → Deutsch	Proprietäre Ausnahme ● knapp	Gemini 3.8 Flash (proprietär, Vertex AI EU)	Claude Sonnet 5.5 (proprietär, AWS Bedrock Frankfurt)	Offen, EU-Anbieter (T1/T1d): Gemma 4 31B 91,4 % Proprietär, EU-Region: Gemini 3.8 Flash 100,0 %
Paschtu پښتو Antwort: Deutsch → Paschtu	Proprietäre Ausnahme ● knapp	Gemini 3.1 Flash-Lite (proprietär, Vertex AI EU), fest an diese Version gebunden	gpt-oss-120b (offen, IONOS, STACKIT, OVH, offene Reserve, mit Glossar und Schriftfilter)	Offen, EU-Anbieter (T1/T1d): gpt-oss-120b 93,1 % Proprietär, EU-Region: Gemini 3.1 Flash-Lite 100,0 %
Paschtu پښتو Verstehen: Paschtu → Deutsch	Proprietäre Ausnahme ● knapp	Gemini 3.8 Flash (proprietär, Vertex AI EU)	Claude Sonnet 5.5 (proprietär, AWS Bedrock Frankfurt)	Offen, EU-Anbieter (T1/T1d): Gemma 4 31B 88,8 % Proprietär, EU-Region: Gemini 3.8 Flash 100,0 %

Sprache und Richtung	Urteil und Gewissheit	Vorschlag zum Start	Ersatz	Qualität relativ zum besten System
Sorani کوردی (سۆرانی) Antwort: Deutsch → Sorani	Proprietäre Ausnahme ● knapp	Gemini 3.8 Flash (proprietär, Vertex AI EU)	Gemini 3.1 Flash-Lite (proprietär, Vertex AI EU)	Offen, EU-Anbieter (T1/T1d): Qwen3.5-397B 93,9 % Proprietär, EU-Region: Gemini 3.8 Flash 99,3 %
Sorani کوردی (سۆرانی) Verstehen: Sorani → Deutsch	Proprietäre Ausnahme ● statistisch abgesichert	Gemini 3.8 Flash (proprietär, Vertex AI EU)	Claude Sonnet 5.5 (proprietär, AWS Bedrock Frankfurt)	Offen, EU-Anbieter (T1/T1d): Gemma 4 31B 85,2 % Proprietär, EU-Region: Gemini 3.8 Flash 100,0 %
Tigrinya ትግርኛ Antwort: Deutsch → Tigrinya	Proprietäre Ausnahme ● statistisch abgesichert	Gemini 3.8 Flash (proprietär, Vertex AI EU), fest an diese Version gebunden	Kein gleichwertiger proprietärer Ersatz. Deshalb zusätzlich muttersprachlich geprüfte Textbausteine und Übergabe an Menschen.	Offen, EU-Anbieter (T1/T1d): Qwen3.5-397B 83,4 % Proprietär, EU-Region: Gemini 3.8 Flash 100,0 %
Tigrinya ትግርኛ Verstehen: Tigrinya → Deutsch	Proprietäre Ausnahme ● statistisch abgesichert	Gemini 3.8 Flash (proprietär, Vertex AI EU)	Gemini 3.1 Flash-Lite (proprietär, Vertex AI EU), Claude Sonnet 5.5 (proprietär, AWS Bedrock Frankfurt)	Offen, EU-Anbieter (T1/T1d): Qwen3.5-397B 87,3 % Proprietär, EU-Region: Gemini 3.8 Flash 100,0 %

¹ Weder Gleichwertigkeit noch ein proprietärer Vorsprung > 5 % ist bewiesen; die Entscheidung hängt am GreenPT-Vertrag bzw. an der fehlenden Messung von Mistral Large 3.

² Nur mit offengelegter Korrektur des Faktentests (Zahlwort, Dual, Ordinalzahl gelten als korrekt). Bei strikter Ziffernprüfung erfüllt in Levante und Irak kein System E25.

● **statistisch abgesichert**: Auch die ungünstige Grenze der Messunsicherheit (einseitig 95 %) liegt auf derselben Seite der Schwelle von 95 %.
● **knapp**: Das Urteil folgt aus der Regel, liegt aber innerhalb der Messunsicherheit.

Qualität relativ zum besten System: chrF++ des Modells geteilt durch den besten gemessenen Wert dieser Kombination (= 100 %). Schwelle der Regel: 95 %. Krisennummern setze ich unabhängig vom Modell aus geprüften Textbausteinen.

Grenzen: allgemeinsprachliche Testsätze, eine Messung je System, gemessen über einen API-Vermittler statt bei den EU-Anbietern. Die Endauswahl treffe ich per Messung in Phase 1, gemeinsam mit den Kursleitenden.

Ukrainisch Українська

System	Einordnung	Antwort (Deutsch → Ukrainisch)	Verstehen (Ukrainisch → Deutsch)
Gemma 4 31B	T1 · Offen · EU-Anbieter, Vertrag geprüft	98,5 % (Spanne 94 %–104 %)	97,5 % (Spanne 92 %–103 %)
gpt-oss-120b	T1 · Offen · EU-Anbieter, Vertrag geprüft	91,8 % (Spanne 87 %–97 %)	91,1 % (Spanne 86 %–96 %)
Llama 3.3 70B	T1 · Offen · EU-Anbieter, Vertrag geprüft	94,1 % (Spanne 89 %–100 %)	95,6 % (Spanne 90 %–101 %)
Mistral Small 3.2 24B	T1 · Offen · EU-Anbieter, Vertrag geprüft	93,3 % (Spanne 88 %–98 %)	96,8 % (Spanne 91 %–102 %)
Qwen3.5-397B	T1 · Offen · EU-Anbieter, Vertrag geprüft	97,8 % (Spanne 93 %–103 %)	93,0 % (Spanne 86 %–99 %)
Mistral Large 3	T1d · Offen · nur als dediziertes Angebot	100,0 % (Spanne 95 %–105 %)	97,1 % (Spanne 92 %–102 %)
DeepSeek V4.1 Flash	T2 · Offen · GreenPT, Vertrag angefragt	97,7 % (Spanne 93 %–103 %)	95,7 % (Spanne 90 %–100 %)
GLM-5.3	T2 · Offen · GreenPT, Vertrag angefragt	98,8 % (Spanne 94 %–105 %)	95,4 % (Spanne 90 %–100 %)

System	Einordnung	Antwort (Deutsch → Ukrainisch)	Verstehen (Ukrainisch → Deutsch)
MiniMax M3	T2 · Offen · GreenPT, Vertrag angefragt	99,3 % (Spanne 95 %–105 %)	96,0 % (Spanne 91 %–101 %)
Claude Haiku 5.5	EU · Proprietär · EU-Region (in der EU nicht verfügbar)	97,1 % (Spanne 92 %–102 %)	97,9 % (Spanne 93 %–103 %)
Claude Sonnet 5.5	EU · Proprietär · EU-Region	98,9 % (Spanne 94 %–104 %)	97,0 % (Spanne 92 %–102 %)
Gemini 3.1 Flash-Lite	EU · Proprietär · EU-Region	99,9 % (Spanne 95 %–105 %)	100,0 % (Spanne 95 %–105 %)
Gemini 3.8 Flash	EU · Proprietär · EU-Region	99,1 % (Spanne 94 %–104 %)	98,6 % (Spanne 93 %–104 %)
DeepSeek V4 Pro	Ref. · Kein EU-Host · nur Referenz	99,6 % (Spanne 95 %–105 %)	96,1 % (Spanne 91 %–101 %)
Kimi K2.6	Ref. · Kein EU-Host · nur Referenz	97,8 % (Spanne 93 %–103 %)	94,9 % (Spanne 90 %–100 %)
Tencent Hy-MT2 30B	Ref. · Kein EU-Host · nur Referenz	96,3 % (Spanne 92 %–102 %)	94,2 % (Spanne 89 %–99 %)

Hocharabisch العربية الفصحى

System	Einordnung	Antwort (Deutsch → Hocharabisch)	Verstehen (Hocharabisch → Deutsch)
Gemma 4 31B	T1 · Offen · EU-Anbieter, Vertrag geprüft	98,3 % (Spanne 94 %–103 %)	94,5 % (Spanne 90 %–99 %)
gpt-oss-120b	T1 · Offen · EU-Anbieter, Vertrag geprüft	87,5 % (Spanne 83 %–92 %)	88,3 % (Spanne 83 %–93 %)
Llama 3.3 70B	T1 · Offen · EU-Anbieter, Vertrag geprüft	84,0 % (Spanne 79 %–89 %)	89,2 % (Spanne 84 %–95 %)
Mistral Small 3.2 24B	T1 · Offen · EU-Anbieter, Vertrag geprüft	80,4 % (Spanne 75 %–85 %) · kleine Mistral	92,4 % (Spanne 87 %–98 %)
Qwen3.5-397B	T1 · Offen · EU-Anbieter, Vertrag geprüft	91,4 % (Spanne 86 %–96 %)	88,6 % (Spanne 81 %–96 %)
Mistral Large 3	T1d · Offen · nur als dediziertes Angebot	93,2 % (Spanne 89 %–97 %)	96,0 % (Spanne 90 %–102 %)
DeepSeek V4.1 Flash	T2 · Offen · GreenPT, Vertrag angefragt	93,5 % (Spanne 89 %–98 %)	96,4 % (Spanne 91 %–102 %)
GLM-5.3	T2 · Offen · GreenPT, Vertrag angefragt	92,1 % (Spanne 88 %–97 %)	93,3 % (Spanne 88 %–98 %)
MiniMax M3	T2 · Offen · GreenPT, Vertrag angefragt	92,0 % (Spanne 87 %–96 %)	91,1 % (Spanne 85 %–97 %)
Claude Haiku 5.5	EU · Proprietär · EU-Region (in der EU nicht verfügbar)	88,4 % (Spanne 84 %–92 %)	91,9 % (Spanne 87 %–97 %)
Claude Sonnet 5.5	EU · Proprietär · EU-Region	93,5 % (Spanne 89 %–98 %)	100,0 % (Spanne 94 %–105 %)
Gemini 3.1 Flash-Lite	EU · Proprietär · EU-Region	100,0 % (Spanne 95 %–105 %)	98,1 % (Spanne 92 %–103 %)
Gemini 3.8 Flash	EU · Proprietär · EU-Region	98,9 % (Spanne 94 %–104 %)	99,8 % (Spanne 93 %–106 %)
DeepSeek V4 Pro	Ref. · Kein EU-Host · nur Referenz	98,6 % (Spanne 94 %–103 %)	96,2 % (Spanne 91 %–101 %)
Kimi K2.6	Ref. · Kein EU-Host · nur Referenz	92,5 % (Spanne 88 %–96 %)	92,0 % (Spanne 85 %–97 %)
Tencent Hy-MT2 30B	Ref. · Kein EU-Host · nur Referenz	92,5 % (Spanne 88 %–97 %)	90,6 % (Spanne 85 %–96 %)

Arabisch (Levante) شامي

System	Einordnung	Antwort (Deutsch → Arabisch (Levante))	Verstehen (Arabisch (Levante) → Deutsch)
Gemma 4 31B	T1 · Offen · EU-Anbieter, Vertrag geprüft	98,2 % (Spanne 93 %–103 %)	97,1 % (Spanne 92 %–102 %)
gpt-oss-120b	T1 · Offen · EU-Anbieter, Vertrag geprüft	87,3 % (Spanne 83 %–92 %)	90,7 % (Spanne 87 %–95 %)

System	Einordnung	Antwort (Deutsch → Arabisch (Levante))	Verstehen (Arabisch (Levante) → Deutsch)
Llama 3.3 70B	T1 · Offen · EU-Anbieter, Vertrag geprüft	87,2 % (Spanne 83 %–92 %)	89,6 % (Spanne 85 %–94 %)
Mistral Small 3.2 24B	T1 · Offen · EU-Anbieter, Vertrag geprüft	80,8 % (Spanne 76 %–85 %)	92,3 % (Spanne 87 %–98 %)
Qwen3.5-397B	T1 · Offen · EU-Anbieter, Vertrag geprüft	89,3 % (Spanne 85 %–94 %)	92,8 % (Spanne 86 %–100 %)
Mistral Large 3	T1d · Offen · nur als dediziertes Angebot	93,2 % (Spanne 89 %–98 %)	nicht vollständig gemessen
DeepSeek V4.1 Flash	T2 · Offen · GreenPT, Vertrag angefragt	92,9 % (Spanne 88 %–98 %)	95,7 % (Spanne 91 %–101 %)
GLM-5.3	T2 · Offen · GreenPT, Vertrag angefragt	91,2 % (Spanne 86 %–97 %)	97,1 % (Spanne 92 %–103 %)
MiniMax M3	T2 · Offen · GreenPT, Vertrag angefragt	81,6 % (Spanne 76 %–87 %)	95,9 % (Spanne 91 %–102 %)
Claude Haiku 5.5	EU · Proprietär · EU-Region (in der EU nicht verfügbar)	88,7 % (Spanne 83 %–94 %)	93,4 % (Spanne 88 %–100 %)
Claude Sonnet 5.5	EU · Proprietär · EU-Region	95,9 % (Spanne 91 %–102 %)	99,6 % (Spanne 95 %–105 %)
Gemini 3.1 Flash-Lite	EU · Proprietär · EU-Region	100,0 % (Spanne 94 %–107 %)	98,9 % (Spanne 94 %–104 %)
Gemini 3.8 Flash	EU · Proprietär · EU-Region	94,6 % (Spanne 90 %–99 %)	100,0 % (Spanne 94 %–106 %)
DeepSeek V4 Pro	Ref. · Kein EU-Host · nur Referenz	95,2 % (Spanne 91 %–101 %)	97,8 % (Spanne 93 %–102 %)
Kimi K2.6	Ref. · Kein EU-Host · nur Referenz	94,5 % (Spanne 90 %–101 %)	96,6 % (Spanne 92 %–102 %)
Tencent Hy-MT2 30B	Ref. · Kein EU-Host · nur Referenz	89,4 % (Spanne 84 %–95 %) · einzelner Satz (Spanne 85 %–96 %)	90,5 % (Spanne 85 %–96 %)

Arabisch (Irak) عراقي

System	Einordnung	Antwort (Deutsch → Arabisch (Irak))	Verstehen (Arabisch (Irak) → Deutsch)
Gemma 4 31B	T1 · Offen · EU-Anbieter, Vertrag geprüft	99,8 % (Spanne 95 %–105 %)	93,5 % (Spanne 88 %–98 %)
gpt-oss-120b	T1 · Offen · EU-Anbieter, Vertrag geprüft	83,6 % (Spanne 78 %–88 %)	89,7 % (Spanne 85 %–95 %)
Llama 3.3 70B	T1 · Offen · EU-Anbieter, Vertrag geprüft	91,4 % (Spanne 87 %–97 %)	89,2 % (Spanne 84 %–94 %)
Mistral Small 3.2 24B	T1 · Offen · EU-Anbieter, Vertrag geprüft	87,0 % (Spanne 82 %–92 %)	nicht vollständig gemessen
Qwen3.5-397B	T1 · Offen · EU-Anbieter, Vertrag geprüft	92,3 % (Spanne 88 %–98 %)	90,5 % (Spanne 83 %–99 %)
Mistral Large 3	T1d · Offen · nur als dediziertes Angebot	93,1 % (Spanne 88 %–98 %) · kleine Mistral	nicht vollständig gemessen
DeepSeek V4.1 Flash	T2 · Offen · GreenPT, Vertrag angefragt	95,2 % (Spanne 91 %–100 %)	95,9 % (Spanne 90 %–102 %)
GLM-5.3	T2 · Offen · GreenPT, Vertrag angefragt	93,1 % (Spanne 89 %–98 %)	96,0 % (Spanne 90 %–101 %)
MiniMax M3	T2 · Offen · GreenPT, Vertrag angefragt	88,8 % (Spanne 84 %–94 %)	92,6 % (Spanne 88 %–98 %)
Claude Haiku 5.5	EU · Proprietär · EU-Region (in der EU nicht verfügbar)	88,3 % (Spanne 83 %–93 %)	94,4 % (Spanne 89 %–100 %)
Claude Sonnet 5.5	EU · Proprietär · EU-Region	98,0 % (Spanne 93 %–103 %)	98,6 % (Spanne 94 %–103 %)
Gemini 3.1 Flash-Lite	EU · Proprietär · EU-Region	100,0 % (Spanne 95 %–105 %)	98,6 % (Spanne 94 %–104 %)
Gemini 3.8 Flash	EU · Proprietär · EU-Region	94,2 % (Spanne 90 %–99 %)	100,0 % (Spanne 94 %–105 %)

System	Einordnung	Antwort (Deutsch → Arabisch (Irak))	Verstehen (Arabisch (Irak) → Deutsch)
DeepSeek V4 Pro	Ref. · Kein EU-Host · nur Referenz	99,6 % (Spanne 94 %–105 %)	97,4 % (Spanne 92 %–102 %)
Kimi K2.6	Ref. · Kein EU-Host · nur Referenz	94,1 % (Spanne 89 %–99 %)	95,4 % (Spanne 90 %–100 %)
Tencent Hy-MT2 30B	Ref. · Kein EU-Host · nur Referenz	96,3 % (Spanne 91 %–102 %)	92,3 % (Spanne 87 %–97 %)

Farsi فارسی

System	Einordnung	Antwort (Deutsch → Farsi)	Verstehen (Farsi → Deutsch)
Gemma 4 31B	T1 · Offen · EU-Anbieter, Vertrag geprüft	95,6 % (Spanne 91 %–100 %)	93,0 % (Spanne 87 %–99 %)
gpt-oss-120b	T1 · Offen · EU-Anbieter, Vertrag geprüft	89,1 % (Spanne 84 %–93 %)	89,7 % (Spanne 84 %–94 %)
Llama 3.3 70B	T1 · Offen · EU-Anbieter, Vertrag geprüft	89,7 % (Spanne 86 %–94 %)	88,8 % (Spanne 83 %–95 %)
Mistral Small 3.2 24B	T1 · Offen · EU-Anbieter, Vertrag geprüft	85,9 % (Spanne 81 %–91 %)	90,9 % (Spanne 85 %–97 %)
Qwen3.5-397B	T1 · Offen · EU-Anbieter, Vertrag geprüft	94,7 % (Spanne 90 %–99 %)	87,4 % (Spanne 80 %–96 %)
Mistral Large 3	T1d · Offen · nur als dediziertes Angebot	94,3 % (Spanne 90 %–99 %)	95,5 % (Spanne 89 %–101 %)
DeepSeek V4.1 Flash	T2 · Offen · GreenPT, Vertrag angefragt	94,3 % (Spanne 89 %–99 %)	94,5 % (Spanne 89 %–99 %)
GLM-5.3	T2 · Offen · GreenPT, Vertrag angefragt	93,7 % (Spanne 89 %–98 %)	94,8 % (Spanne 89 %–100 %)
MiniMax M3	T2 · Offen · GreenPT, Vertrag angefragt	96,5 % (Spanne 92 %–101 %)	93,2 % (Spanne 88 %–98 %)
Claude Haiku 5.5	EU · Proprietär · EU-Region (in der EU nicht verfügbar)	94,1 % (Spanne 90 %–98 %)	94,6 % (Spanne 89 %–100 %)
Claude Sonnet 5.5	EU · Proprietär · EU-Region	95,8 % (Spanne 91 %–100 %)	96,7 % (Spanne 91 %–102 %)
Gemini 3.1 Flash-Lite	EU · Proprietär · EU-Region	100,0 % (Spanne 95 %–104 %)	97,7 % (Spanne 92 %–103 %)
Gemini 3.8 Flash	EU · Proprietär · EU-Region	98,5 % (Spanne 94 %–103 %)	100,0 % (Spanne 94 %–106 %)
DeepSeek V4 Pro	Ref. · Kein EU-Host · nur Referenz	98,2 % (Spanne 93 %–103 %)	97,1 % (Spanne 91 %–103 %)
Kimi K2.6	Ref. · Kein EU-Host · nur Referenz	92,5 % (Spanne 87 %–97 %)	94,6 % (Spanne 89 %–100 %)
Tencent Hy-MT2 30B	Ref. · Kein EU-Host · nur Referenz	92,7 % (Spanne 88 %–97 %)	90,0 % (Spanne 84 %–95 %)

Dari دری

System	Einordnung	Antwort (Deutsch → Dari)	Verstehen (Dari → Deutsch)
Gemma 4 31B	T1 · Offen · EU-Anbieter, Vertrag geprüft	96,9 % (Spanne 92 %–102 %)	94,6 % (Spanne 88 %–101 %)
gpt-oss-120b	T1 · Offen · EU-Anbieter, Vertrag geprüft	90,6 % (Spanne 86 %–95 %)	81,8 % (Spanne 77 %–88 %)
Llama 3.3 70B	T1 · Offen · EU-Anbieter, Vertrag geprüft	97,7 % (Spanne 92 %–103 %)	88,1 % (Spanne 83 %–94 %)
Mistral Small 3.2 24B	T1 · Offen · EU-Anbieter, Vertrag geprüft	88,0 % (Spanne 83 %–92 %)	89,7 % (Spanne 84 %–96 %)
Qwen3.5-397B	T1 · Offen · EU-Anbieter, Vertrag geprüft	95,8 % (Spanne 90 %–100 %)	92,9 % (Spanne 86 %–100 %)
Mistral Large 3	T1d · Offen · nur als dediziertes Angebot	96,1 % (Spanne 91 %–100 %)	96,7 % (Spanne 91 %–103 %)

System	Einordnung	Antwort (Deutsch → Dari)	Verstehen (Dari → Deutsch)
DeepSeek V4.1 Flash	T2 · Offen · GreenPT, Vertrag angefragt	94,8 % (Spanne 89 %–100 %)	95,8 % (Spanne 90 %–102 %)
GLM-5.3	T2 · Offen · GreenPT, Vertrag angefragt	92,3 % (Spanne 87 %–97 %)	95,5 % (Spanne 89 %–102 %)
MiniMax M3	T2 · Offen · GreenPT, Vertrag angefragt	96,1 % (Spanne 91 %–101 %)	96,0 % (Spanne 90 %–102 %)
Claude Haiku 5.5	EU · Proprietär · EU-Region (in der EU nicht verfügbar)	92,6 % (Spanne 87 %–97 %)	92,5 % (Spanne 86 %–99 %)
Claude Sonnet 5.5	EU · Proprietär · EU-Region	94,0 % (Spanne 89 %–98 %)	99,0 % (Spanne 94 %–105 %)
Gemini 3.1 Flash-Lite	EU · Proprietär · EU-Region	97,5 % (Spanne 92 %–102 %)	96,4 % (Spanne 91 %–103 %)
Gemini 3.8 Flash	EU · Proprietär · EU-Region	97,6 % (Spanne 93 %–102 %)	100,0 % (Spanne 95 %–106 %)
DeepSeek V4 Pro	Ref. · Kein EU-Host · nur Referenz	98,3 % (Spanne 93 %–104 %)	99,2 % (Spanne 94 %–106 %)
Kimi K2.6	Ref. · Kein EU-Host · nur Referenz	95,1 % (Spanne 90 %–100 %)	92,0 % (Spanne 86 %–99 %)
Tencent Hy-MT2 30B	Ref. · Kein EU-Host · nur Referenz	91,3 % (Spanne 86 %–96 %)	87,9 % (Spanne 82 %–94 %)

Kurmandschi Kurmancî

System	Einordnung	Antwort (Deutsch → Kurmandschi)	Verstehen (Kurmandschi → Deutsch)
Gemma 4 31B	T1 · Offen · EU-Anbieter, Vertrag geprüft	88,5 % (Spanne 84 %–92 %)	86,7 % (Spanne 81 %–92 %)
gpt-oss-120b	T1 · Offen · EU-Anbieter, Vertrag geprüft	67,5 % (Spanne 47 %–83 %)	71,6 % (Spanne 67 %–77 %)
Llama 3.3 70B	T1 · Offen · EU-Anbieter, Vertrag geprüft	77,9 % (Spanne 74 %–82 %)	82,3 % (Spanne 77 %–89 %)
Mistral Small 3.2 24B	T1 · Offen · EU-Anbieter, Vertrag geprüft	nicht vollständig gemessen	82,4 % (Spanne 76 %–89 %)
Qwen3.5-397B	T1 · Offen · EU-Anbieter, Vertrag geprüft	93,9 % (Spanne 89 %–98 %)	90,1 % (Spanne 82 %–98 %)
Mistral Large 3	T1d · Offen · nur als dediziertes Angebot	91,6 % (Spanne 87 %–95 %)	89,2 % (Spanne 83 %–96 %)
DeepSeek V4.1 Flash	T2 · Offen · GreenPT, Vertrag angefragt	97,3 % (Spanne 93 %–101 %)	89,2 % (Spanne 83 %–96 %)
GLM-5.3	T2 · Offen · GreenPT, Vertrag angefragt	92,2 % (Spanne 88 %–96 %)	88,6 % (Spanne 82 %–95 %)
MiniMax M3	T2 · Offen · GreenPT, Vertrag angefragt	92,6 % (Spanne 88 %–96 %)	89,0 % (Spanne 82 %–96 %)
Claude Haiku 5.5	EU · Proprietär · EU-Region (in der EU nicht verfügbar)	93,7 % (Spanne 89 %–98 %)	86,9 % (Spanne 81 %–93 %)
Claude Sonnet 5.5	EU · Proprietär · EU-Region	97,8 % (Spanne 93 %–102 %)	98,4 % (Spanne 92 %–105 %)
Gemini 3.1 Flash-Lite	EU · Proprietär · EU-Region	100,0 % (Spanne 95 %–104 %)	92,0 % (Spanne 86 %–98 %)
Gemini 3.8 Flash	EU · Proprietär · EU-Region	99,3 % (Spanne 95 %–103 %)	100,0 % (Spanne 94 %–107 %)
DeepSeek V4 Pro	Ref. · Kein EU-Host · nur Referenz	99,6 % (Spanne 95 %–104 %)	95,2 % (Spanne 89 %–102 %)
Kimi K2.6	Ref. · Kein EU-Host · nur Referenz	92,5 % (Spanne 88 %–96 %)	89,7 % (Spanne 84 %–95 %)
Tencent Hy-MT2 30B	Ref. · Kein EU-Host · nur Referenz	69,3 % (Spanne 66 %–72 %)	74,1 % (Spanne 68 %–81 %)

Somali Soomaali

System	Einordnung	Antwort (Deutsch → Somali)	Verstehen (Somali → Deutsch)
Gemma 4 31B	T1 · Offen · EU-Anbieter, Vertrag geprüft	93,7 % (Spanne 90 %–98 %)	91,4 % (Spanne 84 %–99 %)
gpt-oss-120b	T1 · Offen · EU-Anbieter, Vertrag geprüft	94,0 % (Spanne 90 %–98 %)	82,9 % (Spanne 76 %–90 %)
Llama 3.3 70B	T1 · Offen · EU-Anbieter, Vertrag geprüft	72,6 % (Spanne 69 %–76 %)	74,8 % (Spanne 68 %–82 %)
Mistral Small 3.2 24B	T1 · Offen · EU-Anbieter, Vertrag geprüft	nicht vollständig gemessen	nicht vollständig gemessen
Qwen3.5-397B	T1 · Offen · EU-Anbieter, Vertrag geprüft	94,9 % (Spanne 91 %–99 %)	88,7 % (Spanne 80 %–98 %)
Mistral Large 3	T1d · Offen · nur als dediziertes Angebot	85,2 % (Spanne 81 %–89 %)	79,2 % (Spanne 73 %–86 %)
DeepSeek V4.1 Flash	T2 · Offen · GreenPT, Vertrag angefragt	96,7 % (Spanne 93 %–101 %)	89,7 % (Spanne 83 %–97 %)
GLM-5.3	T2 · Offen · GreenPT, Vertrag angefragt	88,7 % (Spanne 85 %–93 %)	90,8 % (Spanne 84 %–98 %)
MiniMax M3	T2 · Offen · GreenPT, Vertrag angefragt	94,2 % (Spanne 90 %–98 %)	87,8 % (Spanne 81 %–94 %)
Claude Haiku 5.5	EU · Proprietär · EU-Region (in der EU nicht verfügbar)	91,6 % (Spanne 87 %–96 %)	85,7 % (Spanne 80 %–93 %)
Claude Sonnet 5.5	EU · Proprietär · EU-Region	95,5 % (Spanne 92 %–100 %)	98,3 % (Spanne 92 %–107 %)
Gemini 3.1 Flash-Lite	EU · Proprietär · EU-Region	100,0 % (Spanne 96 %–104 %)	95,5 % (Spanne 88 %–103 %)
Gemini 3.8 Flash	EU · Proprietär · EU-Region	98,8 % (Spanne 95 %–104 %)	100,0 % (Spanne 94 %–107 %)
DeepSeek V4 Pro	Ref. · Kein EU-Host · nur Referenz	99,3 % (Spanne 95 %–104 %)	93,0 % (Spanne 87 %–99 %)
Kimi K2.6	Ref. · Kein EU-Host · nur Referenz	88,1 % (Spanne 85 %–92 %)	87,4 % (Spanne 81 %–94 %)
Tencent Hy-MT2 30B	Ref. · Kein EU-Host · nur Referenz	80,1 % (Spanne 77 %–84 %)	78,4 % (Spanne 72 %–86 %)

Paschtu پښتو

System	Einordnung	Antwort (Deutsch → Paschtu)	Verstehen (Paschtu → Deutsch)
Gemma 4 31B	T1 · Offen · EU-Anbieter, Vertrag geprüft	88,3 % (Spanne 84 %–92 %)	88,8 % (Spanne 83 %–94 %)
gpt-oss-120b	T1 · Offen · EU-Anbieter, Vertrag geprüft	93,1 % (Spanne 88 %–98 %)	81,0 % (Spanne 76 %–86 %)
Llama 3.3 70B	T1 · Offen · EU-Anbieter, Vertrag geprüft	76,9 % (Spanne 73 %–82 %)	85,1 % (Spanne 80 %–90 %)
Mistral Small 3.2 24B	T1 · Offen · EU-Anbieter, Vertrag geprüft	nicht vollständig gemessen	84,6 % (Spanne 79 %–90 %)
Qwen3.5-397B	T1 · Offen · EU-Anbieter, Vertrag geprüft	78,8 % (Spanne 72 %–85 %)	88,7 % (Spanne 82 %–96 %)
Mistral Large 3	T1d · Offen · nur als dediziertes Angebot	nicht vollständig gemessen	nicht vollständig gemessen
DeepSeek V4.1 Flash	T2 · Offen · GreenPT, Vertrag angefragt	92,8 % (Spanne 88 %–98 %)	91,0 % (Spanne 86 %–96 %)
GLM-5.3	T2 · Offen · GreenPT, Vertrag angefragt	88,3 % (Spanne 84 %–93 %)	91,2 % (Spanne 86 %–96 %)
MiniMax M3	T2 · Offen · GreenPT, Vertrag angefragt	91,8 % (Spanne 87 %–97 %)	89,0 % (Spanne 84 %–94 %)
Claude Haiku 5.5	EU · Proprietär · EU-Region (in der EU nicht verfügbar)	90,7 % (Spanne 87 %–95 %)	89,4 % (Spanne 84 %–95 %)

System	Einordnung	Antwort (Deutsch → Paschtu)	Verstehen (Paschtu → Deutsch)
Claude Sonnet 5.5	EU · Proprietär · EU-Region	93,5 % (Spanne 89 %–99 %)	99,8 % (Spanne 94 %–106 %)
Gemini 3.1 Flash-Lite	EU · Proprietär · EU-Region	100,0 % (Spanne 96 %–105 %)	95,9 % (Spanne 90 %–102 %)
Gemini 3.8 Flash	EU · Proprietär · EU-Region	89,8 % (Spanne 86 %–94 %)	100,0 % (Spanne 95 %–105 %)
DeepSeek V4 Pro	Ref. · Kein EU-Host · nur Referenz	99,8 % (Spanne 95 %–105 %)	95,4 % (Spanne 90 %–101 %)
Kimi K2.6	Ref. · Kein EU-Host · nur Referenz	88,0 % (Spanne 84 %–92 %)	88,8 % (Spanne 84 %–94 %)
Tencent Hy-MT2 30B	Ref. · Kein EU-Host · nur Referenz	64,7 % (Spanne 61 %–68 %)	80,0 % (Spanne 75 %–86 %)

Sorani (سۆرانی) کوردی

System	Einordnung	Antwort (Deutsch → Sorani)	Verstehen (Sorani → Deutsch)
Gemma 4 31B	T1 · Offen · EU-Anbieter, Vertrag geprüft	86,3 % (Spanne 82 %–90 %)	85,2 % (Spanne 79 %–92 %)
gpt-oss-120b	T1 · Offen · EU-Anbieter, Vertrag geprüft	66,0 % (Spanne 62 %–70 %) · einzelner Satz (Spanne 46 %–72 %)	67,2 % (Spanne 63 %–72 %)
Llama 3.3 70B	T1 · Offen · EU-Anbieter, Vertrag geprüft	73,4 % (Spanne 70 %–77 %)	80,4 % (Spanne 76 %–86 %)
Mistral Small 3.2 24B	T1 · Offen · EU-Anbieter, Vertrag geprüft	nicht vollständig gemessen	53,8 % (Spanne 50 %–58 %)
Qwen3.5-397B	T1 · Offen · EU-Anbieter, Vertrag geprüft	93,9 % (Spanne 90 %–98 %)	84,8 % (Spanne 77 %–92 %)
Mistral Large 3	T1d · Offen · nur als dediziertes Angebot	71,6 % (Spanne 69 %–74 %)	nicht vollständig gemessen
DeepSeek V4.1 Flash	T2 · Offen · GreenPT, Vertrag angefragt	94,8 % (Spanne 90 %–98 %)	90,6 % (Spanne 84 %–97 %)
GLM-5.3	T2 · Offen · GreenPT, Vertrag angefragt	90,6 % (Spanne 86 %–95 %)	88,9 % (Spanne 84 %–94 %)
MiniMax M3	T2 · Offen · GreenPT, Vertrag angefragt	89,7 % (Spanne 85 %–94 %)	86,5 % (Spanne 81 %–92 %)
Claude Haiku 5.5	EU · Proprietär · EU-Region (in der EU nicht verfügbar)	89,7 % (Spanne 85 %–93 %)	87,2 % (Spanne 82 %–94 %)
Claude Sonnet 5.5	EU · Proprietär · EU-Region	91,7 % (Spanne 88 %–95 %)	96,8 % (Spanne 90 %–104 %)
Gemini 3.1 Flash-Lite	EU · Proprietär · EU-Region	98,2 % (Spanne 94 %–102 %)	95,7 % (Spanne 89 %–102 %)
Gemini 3.8 Flash	EU · Proprietär · EU-Region	99,3 % (Spanne 94 %–104 %)	100,0 % (Spanne 93 %–106 %)
DeepSeek V4 Pro	Ref. · Kein EU-Host · nur Referenz	100,0 % (Spanne 95 %–104 %)	93,1 % (Spanne 88 %–99 %)
Kimi K2.6	Ref. · Kein EU-Host · nur Referenz	86,1 % (Spanne 82 %–90 %)	87,1 % (Spanne 81 %–93 %)
Tencent Hy-MT2 30B	Ref. · Kein EU-Host · nur Referenz	36,3 % (Spanne 33 %–39 %)	66,3 % (Spanne 61 %–72 %)

Tigrinya ትግርኛ

System	Einordnung	Antwort (Deutsch → Tigrinya)	Verstehen (Tigrinya → Deutsch)
Gemma 4 31B	T1 · Offen · EU-Anbieter, Vertrag geprüft	63,0 % (Spanne 58 %–68 %)	79,1 % (Spanne 73 %–85 %)
gpt-oss-120b	T1 · Offen · EU-Anbieter, Vertrag geprüft	14,0 % (Spanne 10 %–18 %) · einzelner Satz (Spanne 4 %–54 %)	51,5 % (Spanne 48 %–54 %)
Llama 3.3 70B	T1 · Offen · EU-Anbieter, Vertrag geprüft	nicht vollständig gemessen	49,6 % (Spanne 46 %–55 %)
Mistral Small 3.2 24B	T1 · Offen · EU-Anbieter, Vertrag geprüft	nicht vollständig gemessen	39,3 % (Spanne 37 %–42 %)

System	Einordnung	Antwort (Deutsch → Tigrinya)	Verstehen (Tigrinya → Deutsch)
Qwen3.5-397B	T1 · Offen · EU-Anbieter, Vertrag geprüft	83,4 % (Spanne 77 %–90 %)	87,3 % (Spanne 80 %–94 %)
Mistral Large 3	T1d · Offen · nur als dediziertes Angebot	67,8 % (Spanne 63 %–73 %)	65,6 % (Spanne 61 %–71 %)
DeepSeek V4.1 Flash	T2 · Offen · GreenPT, Vertrag angefragt	87,4 % (Spanne 81 %–94 %)	83,9 % (Spanne 78 %–90 %)
GLM-5.3	T2 · Offen · GreenPT, Vertrag angefragt	65,2 % (Spanne 61 %–71 %)	74,2 % (Spanne 69 %–80 %)
MiniMax M3	T2 · Offen · GreenPT, Vertrag angefragt	53,0 % (Spanne 47 %–58 %) · einzelner Satz (Spanne 47 %–58 %)	79,9 % (Spanne 74 %–85 %)
Claude Haiku 5.5	EU · Proprietär · EU-Region (in der EU nicht verfügbar)	80,8 % (Spanne 75 %–87 %)	82,2 % (Spanne 76 %–88 %)
Claude Sonnet 5.5	EU · Proprietär · EU-Region	83,1 % (Spanne 78 %–90 %)	95,3 % (Spanne 89 %–102 %)
Gemini 3.1 Flash-Lite	EU · Proprietär · EU-Region	90,8 % (Spanne 83 %–99 %)	95,9 % (Spanne 89 %–104 %)
Gemini 3.8 Flash	EU · Proprietär · EU-Region	100,0 % (Spanne 92 %–108 %)	100,0 % (Spanne 93 %–108 %)
DeepSeek V4 Pro	Ref. · Kein EU-Host · nur Referenz	95,8 % (Spanne 90 %–102 %)	90,3 % (Spanne 84 %–96 %)
Kimi K2.6	Ref. · Kein EU-Host · nur Referenz	61,2 % (Spanne 56 %–66 %)	80,5 % (Spanne 73 %–87 %)
Tencent Hy-MT2 30B	Ref. · Kein EU-Host · nur Referenz	12,8 % (Spanne 10 %–16 %) · einzelner Satz (Spanne 10 %–16 %)	40,8 % (Spanne 39 %–43 %)

Spanne: 95%-Intervall des chrF++-Werts aus dem Bootstrap, geteilt durch den Bestwert (Näherung). „Nicht vollständig gemessen“ zählt weder für noch gegen ein Modell. Modelle ohne EU-Host sind nur zum Vergleich aufgeführt und nicht zum Einsatz vorgesehen.

Architektur und Zeitplan

Systembild



Systembild. Eine Frage kommt über den QR-Code in die Chat-Oberfläche und durchläuft im Wegweiser-Dienst Schutz, Suche, KI-Gateway und Prüfschicht. Die Antwort geht mit Quelle und Prüfdatum zurück. Ihr Team pflegt die Wissensdatenbank in der Redaktion; der Wegweiser liest nur Freigegebenes.

Bausteine

Baustein	Lösung	Warum
Datenbank	Managed PostgreSQL mit pgvector, bevorzugt STACKIT (deutscher Betreiber, C5-Testat Typ 2 inkl. Datenbank), gleichwertig IONOS	Standard-SQL, tägliche Sicherung mit Wiederherstellung auf einen Zeitpunkt, Umzug per Export jederzeit möglich
Redaktion	Payload CMS (Open Source, MIT-Lizenz)	Formulare statt SQL, Vier-Augen-Freigabe über Rollen, Versionen, Sprachfassungen inkl. Rechts-nach-links-Schrift, keine Lizenzkosten
Suche und Antwort	schlanker eigener Dienst (ca. 500 Zeilen) auf Basis offener Bibliotheken	nachvollziehbar, testbar, keine zusätzliche Plattform
KI-Modelle	Gateway mit OpenAI-kompatibler Schnittstelle (LiteLLM, Open Source); offene Modelle bei EU-Anbietern als Standard	Modellwechsel per Konfiguration, ohne Umbau
Chat-Oberfläche	Web-App (PWA) mit assistant-ui (Open Source)	per QR-Code sofort nutzbar, kein App-Store, läuft auch auf älteren Android-Geräten

Baustein	Lösung	Warum
Betrieb	Container bei einem EU-Anbieter (STACKIT oder Scaleway), Konfiguration als Code	keine Server- oder Kubernetes-Pflege durch Ihr Team

Zeitplan (ab Beauftragung)

Zeitraum	Meilenstein
Monat 1	Auftakt-Workshop, Architekturentscheidung, Konten auf die Proclusio gGmbH
Monat 1–2	Datenbank, Datenmodell und Redaktion stehen; Ihr Team beginnt mit der Pflege
Monat 2–3	Suche, KI-Anbindung, Chat-Oberfläche, Audio; erste Sprachen im internen Test
Monat 2–3	Unterlagen für die Datenschutz-Folgenabschätzung, Sicherheits- und Barrierefreiheitstests, Schulung Redaktion
ab ca. Monat 3–4	Beta in ausgewählten Unterkünften (nach Freigabe der DSFA), Freigabe je Sprache
Monat 4–18	Optimierung nach echten Fragen, weitere Sprachen, MCP-Schnittstelle, Workshops KI-Kompetenz
bis 30.06.2029	Stabilisierung, Dokumentation, Übergabetest, monatliche Sprechstunde

In Phase 1 arbeite ich mit ca. 20–30 Stunden pro Woche, danach bedarfsorientiert. Den Start der Beta bestimmen vor allem die Freigabe Ihrer Datenschutz-Folgenabschätzung und die ersten gepflegten Inhalte; ich passe den Plan an Ihren Wunschtermin an. Abstimmung: in Phase 1 alle zwei Wochen, danach monatlich. Vor-Ort-Termine bündele ich.

Anforderungs-Navigator: Ihre Anforderungen, meine Antworten

Diese Übersicht ordnet jede Anforderung Ihrer Leistungsbeschreibung meiner Antwort zu und verweist auf den Abschnitt, in dem sie ausführlich steht.

§ 6 Leistungsumfang

§	Ihre Anforderung	Meine Antwort in einem Satz	Abschnitt
6.1	Technische Konzeption, Bewertung von Managed Services und Open Source, Trennung von Wissensbasis und Anwendung	Ich übernehme das Prinzip Ihrer IONOS-Prüfung, ergänze nur Fehlendes und habe die Modellwahl vorab gemessen; die Wissensbasis ist über eine Datenbanksicht von der Anwendung getrennt.	Architektur → Modelle →
6.2	Managed-Datenbank mit Vektordaten, Rechten und Datensicherung	Managed PostgreSQL mit pgvector bei STACKIT (gleichwertig IONOS), Rollenkonzept, Zeitpunkt-Wiederherstellung und getestete Rücksicherung.	Architektur →
6.3	Datenmodell mit FAQs, Themen, Quellen, Zuständigkeiten, Fristen, Formularen und Aktualisierungsständen	Kritische Fakten liegen strukturiert in eigenen Feldern, jeder Eintrag hat Status, Gültigkeit, Prüfdatum und verantwortliche Person.	Redaktion →
6.4	Redaktionsoberfläche ohne SQL	Payload CMS mit Vier-Augen-Freigabe, Versionen, Sprachfassungen, Pflichtfeldern und Lückenliste.	Redaktion →
6.5	Semantische Suche, RAG, Erkennen fehlender Grundlage, Weiterleitung	Kombinierte Suche nur in Freigegebenem; unterhalb einer Relevanzschwelle eine feste Antwort ohne KI mit Beratungskontakt.	So funktioniert's → Architektur →
6.6	LLM per Standard-Schnittstelle, EU, keine Speicherung, kein Training, austauschbar	Gateway mit OpenAI-kompatibler Schnittstelle, offene Modelle bei EU-Anbietern als Standard, Modellwechsel per Konfiguration.	Modelle →

§	Ihre Anforderung	Meine Antwort in einem Satz	Abschnitt
6.7	Chat ohne Login, mobil, QR, Muttersprache, KI-Transparenz, Quellen, barrierearm, Audio, Kontakt zu Menschen	Web-App per QR-Code, alle Schriftrichtungen, A-/A+, Diktieren und Vorlesen in der EU, KI-Hinweis, Quellen mit Stand, „Mit einem Menschen sprechen“.	Demo → Qualität und Datenschutz →
6.8	Dokumentierte Schnittstelle, Entkopplung	Eigener Dienst mit OpenAPI-Dokumentation; Oberfläche und weitere Anwendungen greifen nur darüber zu.	Architektur →
6.9	Vorbereitung auf MCP, keine autonomen Agenten	Nur lesender MCP-Server über dieselbe Schnittstelle in Phase 2.	Architektur → Vorgehen →
6.10	Anonymität, Datensparsamkeit, kein Training, Zugriffskontrolle	Keine Konten und Profile, keine IP-Speicherung, automatische Entfernung persönlicher Angaben, Löschung nach 90 Tagen, Zwei-Faktor-Anmeldung für das Team.	Qualität und Datenschutz →
6.11	Beta-Begleitung: Inbetriebnahme, Fehleranalyse, Auswertung	Begleitung in den Unterkünften, Monitoring über den echten Zugang, Freigabe je Sprache mit den Kursleitenden, anonyme Auswertung.	Vorgehen →
6.12	Technische Optimierung	Messbar: Trefferqualität, Quellentreue, Sprachrichtigkeit, Antwortzeit; Neutest nach jedem Modellwechsel.	Vorgehen → Modelle →
6.13	Dokumentation und Wissenstransfer für Nicht-Techniker:innen	Redaktionshandbuch, 6–8 Videoanleitungen, Runbooks nach Symptom, Sprechstunde, Übergabetest.	Befähigung →

§ 7 Grundsätze

§	Ihr Grundsatz	Meine Antwort	Abschnitt
7.1	Bestehende Komponenten statt Eigenentwicklung	Managed Datenbank, Payload, assistant-ui, LiteLLM, Whisper; eigener Code nur als schlanke Verbindungsschicht (ca. 500 Zeilen Kern).	Architektur →

§	Ihr Grundsatz	Meine Antwort	Abschnitt
7.2	Offene Standards, Entkopplung	Standard-SQL, OpenAPI, MCP, OpenAI-kompatible Schnittstelle, Container, Infrastruktur als Code.	Architektur →
7.3	Keine Anbieterabhängigkeit; Open Source erwünscht, soweit geeignet	Jeder Baustein einzeln austauschbar, auch zurück zu IONOS; Open Source als Standard, wo der Test Gleichwertigkeit zeigt.	Modelle →
7.4	Niedrige Einstiegshürde für das Team	Formulare statt Code, Pflichtfelder, Testbereich, Schulungen und Videos.	Befähigung →
7.5	Angemessene Komplexität	Kein Kubernetes, keine RAG-Plattform, wenige Anbieter; Phase 1 mit rund 170–190 Stunden.	Vorgehen →

§§ 8, 9, 12 Person, Referenzen, Angebotsinhalte

Anforderung	Wo im Angebot	Abschnitt
Kenntnisse PostgreSQL/pgvector, LLM/RAG, APIs, Cloud, Datenschutz, Befähigung	Expertise-Tabelle und Referenzen	Person → Referenzen →
Mindestens zwei Referenzen mit eigener Rolle, mindestens eine mit RAG/LLM	vier Referenzen, drei davon KI-Wissenssysteme	Referenzen →
Angaben zur Person, Kurzprofil, Qualifikationen	Zur Person	Person →
Verständnis der Aufgabe	Mein Verständnis	Verständnis →
Technisches Vorgehen (höchstens 5 Seiten), Architektur und Komponenten	PDF-Kapitel 5; auf der Website Architektur, Modelle, Redaktion, Qualität	Architektur →
Befähigung des Projektteams	Befähigung und Nachhaltigkeit	Befähigung →
Netto-Stundensatz, USt-Satz, Brutto-Stundensatz	100,00 € · 19 % · 119,00 €	Preis →

§ 15 Vertragliche Punkte, die ich vorschlage

Thema	Mein Vorschlag	Abschnitt
-------	----------------	-----------

Thema	Mein Vorschlag	Abschnitt
Nutzungsrechte	Projektspezifische Entwicklungen gehen vollständig an Sie. An eingebrachten Bausteinen erhalten Sie ein dauerhaftes, übertragbares Nutzungsrecht. Auf Wunsch wird der Code unter EUPL veröffentlicht.	Befähigung →
Zugänge und Zugangsdaten	Alle Konten laufen auf die Proclusio gGmbH, mit Zwei-Faktor-Anmeldung.	Befähigung →
Beendigung und weitere Nutzbarkeit	Dokumentation, Runbooks und ein Übergabetest mit einer fremden Fachkraft.	Befähigung →
Zusammenarbeit	In Phase 1 alle zwei Wochen, danach monatlich; Stundennachweis je Tätigkeit.	Vorgehen → Preis →