

Grounding-Gate für RAG-basierte Verwaltungsassistenten

Technische Beschreibung für Website, GitHub und Hugging Face. Dieses Dokument beschreibt die Fehlerannahmen, die Grundfunktion des Prototyps und die vorgeschlagene Lösungsschicht. Es enthält bewusst keine internen Details zu weitergehenden Drift-Limiter-Mechanismen.

Kurzclaim

Ein Grounding-Gate prüft eine vom Sprachmodell entworfene Antwort vor der Ausgabe gegen die tatsächlich verwendeten Quellen. Wenn die Antwort nicht ausreichend belegt ist, wird sie nicht unverändert angezeigt: Das System fragt nach, zeigt nur den sicheren Teil, blockiert oder verweist auf eine offizielle Stelle.

Umfang	Einordnung
Prototyp	Installierbarer Python-Kern mit CLI, Testset, Claim-Prüfung, austauschbarem Prüfer und Audit-Log.
Demo	Statische Website-Demo ohne API, mit kuratierten Fällen und sichtbarer Vorher/Nachher-Erklärung.
Abgrenzung	Kein Ida-Nachbau, keine produktive Behördenauskunft, kein vollständiger Halluzinationsschutz.

1. Ausgangslage

Öffentlich beschriebene Systeme wie ida kombinieren ein Sprachmodell mit Retrieval-Augmented Generation (RAG): Das Modell erhält passende Inhalte aus einer Wissensbasis und formuliert daraus eine Antwort. Laut BRZ nutzt ida geprüfte Inhalte von oesterreich.gv.at und id-austria.gv.at, läuft auf BRZ-LLMaaS mit einem europäischen Mistral-Modell und verwendet RAG, um Halluzinationen und Falschaussagen zu vermeiden. Die offizielle ida-Seite weist zugleich darauf hin, dass fehlerhafte Ergebnisse und Halluzinationen nicht ausgeschlossen werden können.

Unser Ausgangspunkt ist daher nicht: RAG ist falsch. Unser Ausgangspunkt ist: RAG allein garantiert nicht, dass jede einzelne ausgegebene Aussage durch die abgerufenen Quellen gedeckt ist.

2. Prämissen zu den Fehlerquellen

Die folgenden Annahmen sind bewusst einfach und öffentlich erklärbar. Sie reichen, um den Prototyp zu begründen.

Fehlerquelle	Annahme	Risiko
Ungestützte Zusatzbehauptung	Das Modell kann aus echten Quellen eine Antwort formulieren und trotzdem einen nicht belegten Zusatz einfügen.	Plausible, aber falsche Auskunft.

Fehlerquelle	Annahme	Risiko
Mehrdeutige Frage	Eine kurze Frage kann mehrere legitime Behörden- oder Hilfefade meinen.	Das System wählt vorschnell den falschen Pfad.
Aktualität / Hot Facts	Manche Fragen brauchen aktuelle autoritative Quellen, nicht allgemeine FAQ-Seiten.	Veraltete oder sachfremde Antwort.
Nutzerwiderspruch	Ein Modell kann unter Nachfrage oder Druck eine korrekte Antwort in eine falsche Revision kippen.	Vertrauensverlust durch inkonsistentes Verhalten.
Scope-Grenze	Nicht jede legitime Frage gehört zum aktuellen Wissensbereich.	Das System wirkt entweder ahnungslos oder erfindet Zuständigkeit.
Zu sichere Sprache	Die Formulierung kann sicherer klingen als die Beleglage ist.	Bürger:innen überschätzen die Verbindlichkeit.

3. Kernidee: Grounding-Gate

Das Grounding-Gate ist eine ausgabeseitige Prüfschicht. Es sitzt nach Retrieval und Antwortentwurf, aber vor der Anzeige an Nutzer:innen. Es ersetzt weder RAG noch das Sprachmodell. Es prüft, ob der Entwurf tatsächlich zu den Quellen passt.

Prinzip

Nicht der Generator bewertet sich selbst. Eine getrennte Prüflogik bewertet die Antwort Aussage für Aussage gegen die Quellen.

Pipeline im Prototyp

Schritt	Beschreibung
1. Nutzerfrage	Die ursprüngliche Frage bleibt erhalten, weil kurze Antworten nur im Kontext prüfbar sind.
2. Quellen	Die tatsächlich abgerufenen Quellabschnitte werden an das Gate übergeben.
3. DraftAnswer	Das Sprachmodell erzeugt zunächst nur einen Antwortentwurf.
4. Aussageprüfung	Der Entwurf wird in prüfbare Aussagen zerlegt und jede Aussage wird gegen die Quellen bewertet.
5. Entscheidung	Je nach Ergebnis: anzeigen, nachfragen, nur sicheren Teil zeigen, blockieren oder offiziell weiterleiten.
6. Audit	Die Entscheidung wird protokolliert, damit Fehlerfälle in Tests überführt werden können.

4. Entscheidungslogik

Die Demo verwendet einfache Begriffe. Technisch kann das Gate weiterhin Scores und Bänder führen; sichtbar werden klare Handlungen.

Demo-Wort	Technische Bedeutung	Nutzerfreundliche Antwort
Antwort zeigen	Alle wesentlichen Aussagen sind durch Quellen belegt.	Antwort mit sichtbarer Quelle anzeigen.
Kurz nachfragen	Die Frage ist mehrdeutig oder der richtige Pfad ist unklar.	Eine knappe Rückfrage stellen.
Nur sicheren Teil zeigen	Ein Teil ist belegt, ein Zusatz nicht.	Nur den belegten Teil ausgeben.
Nicht zeigen	Die Antwort passt nicht zu den Quellen.	Keine falsche Antwort anzeigen.
Weiterleiten	Frage liegt außerhalb des aktuellen Wissensbereichs oder braucht eine autoritative Stelle.	Offiziellen Link oder passende Anlaufstelle anbieten.

Hinweis zu CalibrateState

Der in der Demo erwähnte CalibrateState ist als vorgelagerte Schärfung der Anfrage zu verstehen: Welche Art von Frage liegt vor, wie klar ist der Auftrag, braucht es eine kurze Rückfrage und welcher Wissensbereich ist zuständig? Der Begriff gehört zu einer übergeordneten Idee namens Drift Limiter (DOI: <https://doi.org/10.5281/zenodo.20633292>). Dieses öffentliche Briefing zeigt nur die daraus abgeleitete Nutzfunktion; interne Drift-Limiter-Mechanismen bleiben bewusst außerhalb des Dokuments.

5. Was der Prototyp heute kann

- Installierbarer Python-Kern ohne verpflichtende externe Abhängigkeiten.
- CLI-Kommandos: check, harvest und build-testset.
- Austauschbares Judge-Interface: einfacher Offline-Prüfer für Entwicklung, Mistral-kompatibler Prüfer für echte Semantiktests.
- Fail-closed-Verhalten: Ohne Quellen oder bei Prüffehlern wird die Antwort nicht einfach durchgelassen.
- Testset mit gedeckten FAQ-Fragen und kuratierten Negativfällen.
- Audit-Log als Grundlage für Fehleranalyse und Regressionstests.
- Statische Website-Demo ohne API, die das Verhalten sichtbar macht.

Implementierungsstand

Der aktuelle Prototyp bildet den Grounding-Gate-Kern ab. Nicht alle in diesem Briefing beschriebenen Erweiterungen sind bereits vollständig implementiert. Teile der vorgeschlagenen Weiterentwicklung - insbesondere CalibrateState- und Drift-Limiter-nahe Funktionen - sind derzeit Forschungs- und Roadmap-Komponenten und nicht Bestandteil des öffentlichen Grundprototyps.

6. Was bewusst nicht behauptet wird

- Der Prototyp ist kein Nachbau von ida.
- Er ist keine produktive Behördenauskunft.
- Er beweist keine vollständige Fehlerfreiheit.
- Er ersetzt keine fachliche Abnahme und keine produktionsnahe Kalibrierung am echten Infrastruktur-Endpoint.
- Er veröffentlicht keine internen Drift-Limiter-Details; der öffentliche Kern ist das Grounding-Gate.

7. Warum das für Open Source attraktiv ist

Der Prototyp ist klein genug, um verstanden, geforkt und getestet zu werden. Gleichzeitig adressiert er ein reales Problem öffentlicher KI-Systeme: Wie verhindern wir, dass eine plausibel formulierte Antwort ohne ausreichende Quellenbasis ausgeliefert wird?

Zielgruppe	Nutzen
GitHub	Klarer, testbarer Kern mit CLI und überschaubarem Code. Gute Grundlage für Issues, PRs und Erweiterungen.
Hugging Face	Gut demonstrierbare Safety-Komponente für RAG-Systeme: Dataset-Fälle, Demo und reproduzierbare Evaluation.
Verwaltung / Civic Tech	Sprache und Entscheidungslogik sind erklärbar: anzeigen, nachfragen, nur sicheren Teil zeigen, blockieren, weiterleiten.
Accessibility-Interessierte	Die Demo zeigt optional, wie Quellen, Status und nächste Schritte auch assistiv erreichbar gemacht werden können.

8. Öffentliche Kurzbeschreibung

Kurzbeschreibung

Grounding-Gate ist ein kompakter Open-Source-Prototyp für eine unabhängige Quellenprüfung von RAG-Antworten. Er zeigt, wie ein Verwaltungsassistent vor der Ausgabe prüfen kann, ob einzelne Aussagen wirklich durch die verwendeten Quellen belegt sind.

Quellen und Bezugspunkte

- BRZ: Technischer Hintergrund zu ida, der intelligenten digitalen Assistenz, 22. Juni 2026.
<https://www.brz.gv.at/presse/technischer-hintergrund-ida.html>
- oesterreich.gv.at: Chatbot ida. <https://www.oesterreich.gv.at/de/landingpages/chatbot-ida>
- Bundeskanzleramt: Bundesministerinnen und Bundesminister, Staatssekretärinnen und Staatssekretäre.
<https://www.bundeskanzleramt.gv.at/bundeskanzleramt/die-bundesregierung/bundesministerinnen-und-bundesminister.html>
- Methodischer Bezugspunkt: Drift-Limiter / CalibrateState-Konzept. DOI: <https://doi.org/10.5281/zenodo.20633292>
- Projekt-/Attributionsangabe: Verein "AI Forschungsinstitut für Menschen", 1100 Wien, Österreich, office@aifim.at, ZVR-Zahl 1547230392; Lead Engineer: Dejan Nikolic.
- Prototyp-Unterlagen: ida-gate, Demo und Testset im Repository-Paket.