Infra

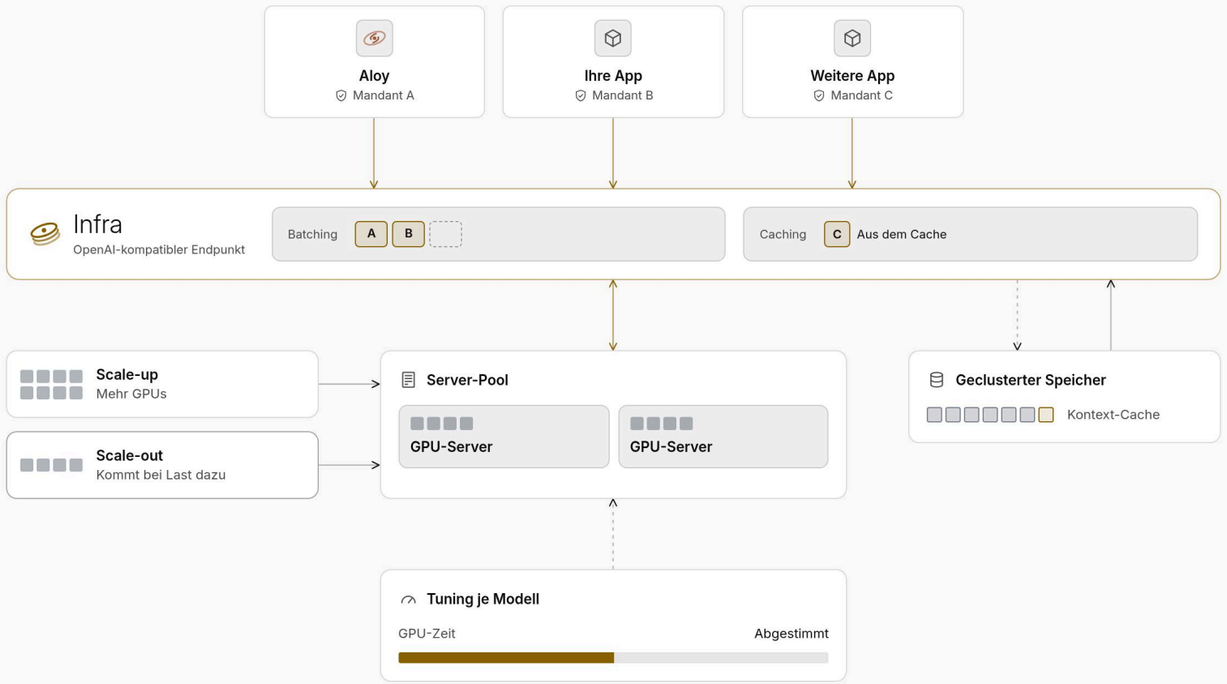
Infra ist die Inferenz-Schicht für Ihre KI-Modelle. Sie skaliert von einem einzelnen GPU-Server bis zum Cluster und stellt Ihren Anwendungen durchgehend eine einzige OpenAI-kompatible API bereit.

Diese Übersicht folgt der Seite rhenium.tech/de/products/infra/, Stand 20.11.2026.

Ein Endpunkt für alle Modelle

Anwendungen verbinden sich über einen einzigen OpenAI-kompatiblen Endpunkt mit Infra. Sowohl Aloy als auch Ihre eigenen Anwendungen nutzen dieselbe Schnittstelle.

Hinter diesem Endpunkt bündelt Infra Ihre GPU-Kapazitäten und trennt Mandanten sauber voneinander. Änderungen an der Hardware bleiben für die angebundenen Anwendungen unsichtbar.

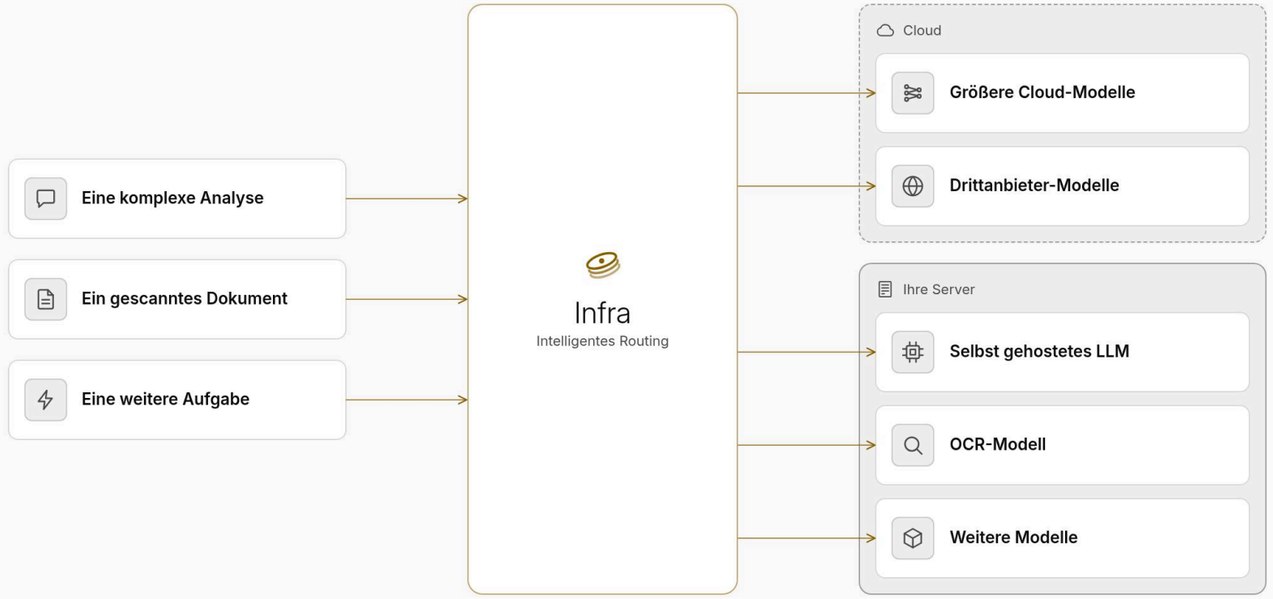


Rhenium Technologies GmbH · Sonnenstr. 31 · 85609 Aschheim
Geschäftsführung: Rastislav Kulazhanka · Amtsgericht München · HRB 218918
USt-IdNr. DE 815 572 704 · +49 173 8127829 · info@rhenium.tech

Routing zwischen eigenen Servern und Cloud

Unterschiedliche Aufgaben erfordern unterschiedliche Modelle: Komplexe Analysen nutzen größere Sprachmodelle, während gescannte Dokumente über spezialisierte OCR-Modelle laufen. Infra leitet jede Anfrage automatisch an das passende Modell weiter.

Modelle können auf Ihren eigenen Servern, in der Cloud oder über Drittanbieter laufen. Für Ihre Anwendungen bleibt es eine einheitliche Schnittstelle unter einer gemeinsamen Richtlinie.



Optimiert für geringe Latenz und niedrige GPU-Kosten

Die Betriebskosten von KI hängen direkt von der benötigten GPU-Zeit ab. Rhenium profiliert und optimiert jedes eingesetzte Modell für Ihre Ziel-Hardware, um Antwortzeiten zu verkürzen und den Durchsatz zu erhöhen.

Zusätzlich übernimmt Infra verteiltes Caching, Anfrage-Bündelung und Modelloptimierung über alle Knoten hinweg. Bereits berechneter Kontext wird wiederverwendet statt neu berechnet, was Serverlast, Netzwerkverkehr und Kosten senkt.

Modellspezifische Optimierung



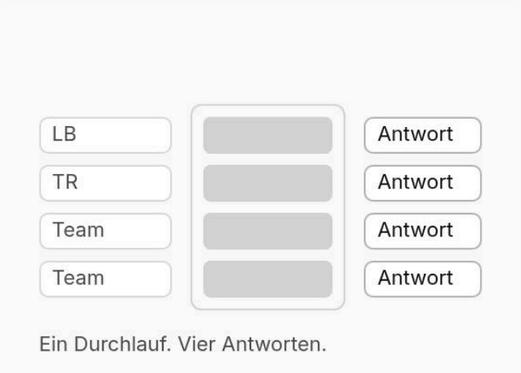
Jedes eingesetzte Modell wird für seine Architektur und Ziel-Hardware profiliert und optimiert, um Antwortzeiten bei gleichbleibender Qualität zu verkürzen.

Verteiltes Caching



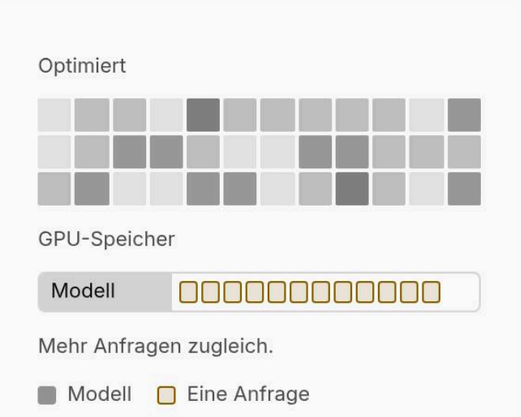
Gemeinsamer Kontext und wiederkehrende Eingaben werden über Inferenz-Knoten hinweg zwischengespeichert, sodass gleiche Arbeit nicht doppelt berechnet wird.

Dynamisches Batching



Gleichzeitige Anfragen aus dem Unternehmen werden effizient gebündelt verarbeitet, damit der Durchsatz auch bei Lastspitzen stabil bleibt.

Speicheroptimierung



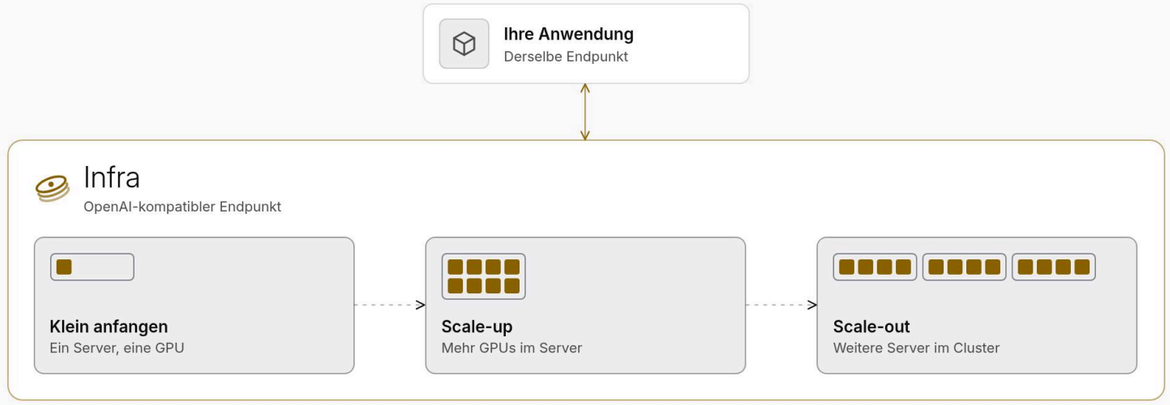
Modelle werden so kalibriert, dass sie weniger Speicher pro Anfrage benötigen. Das ermöglicht höhere Parallelität oder kompaktere Hardware bei geprüfter Genauigkeit.

Rhenium Technologies GmbH · Sonnenstr. 31 · 85609 Aschheim
Geschäftsführung: Rastislav Kulazhanka · Amtsgericht München · HRB 218918
USt-IdNr. DE 815 572 704 · +49 173 8127829 · info@rhenium.tech

Vom einzelnen GPU-Server zum Cluster

Starten Sie mit Infra auf einem einzelnen Server mit einer GPU. Steigt die Nutzung, erweitern Sie den Server um zusätzliche GPUs (Scale-up) oder verbinden mehrere Server zu einem Cluster (Scale-out).

Dieselbe Software läuft vom Einzelserver bis zum Rechenzentrum, während Endpunkt und Integration Ihrer Anwendungen unverändert bleiben.



Betriebsarten

Infra arbeitet im Verbund mit Aloy oder eigenständig als Inferenz-Schicht für Ihre Anwendungen.

Einsatz	Was Sie bekommen
Mit Aloy	Setzt das Berechtigungs- und Mandantenmodell von Aloy durchgehend um.
Eigenständig	Stellt Ihren eigenen Anwendungen einen nativen OpenAI-kompatiblen Endpunkt bereit.
Für Systemintegratoren	Fügt sich in die Architektur Ihrer Kunden ein, vom Einzelserver bis zum vollständigen Cluster.
Offline	Im Verbund mit Aloy und einer Offline-Richtlinie läuft Infra vollständig abgeschottet, ohne Netzwerkverkehr nach außen.

Sprechen Sie mit uns

Jedes Projekt beginnt mit einem Gespräch. Erzählen Sie uns, woran Sie arbeiten und wo es hakt. Wir antworten Ihnen per E-Mail.

E-Mail info@rhenium.tech

Web rhenium.tech/de/products/infra/