ERGOSIGN

Generative KI verstehen und sinnvoll einsetzen

Von Large Language Models bis RAG: Grundlagen moderner GenAI-Anwendungen

Hochschule Bonn-Rhein-Sieg
University of Applied Sciences

Mittelstand-Digital Zentrum Fokus Mensch

November 2026

2026 | Vertraulich | Alle Rechte vorbehalten, Ergosign

Das Webinar hat zwei Teile

Teil 1 · Heute

Generative KI verstehen

Wie KI-Anwendungen funktionieren
und was man damit bauen kann

Teil 2 · Nächstes Mal

GenAI Ops

Wie KI-Anwendungen im Live-Betrieb
zuverlässig werden

2026 | Vertraulich | Alle Rechte vorbehalten, Ergosign

Nicolas Klein

AI Software Engineer @ Ergosign GmbH

→ Ich habe bei Ergosign KI-Produkte von der Demo in den produktiven Einsatz gebracht. Was ich dabei gelernt habe, steckt in diesem Webinar.

2026 | Vertraulich | Alle Rechte vorbehalten, Ergosign

Ich erzähle Ihnen eine Geschichte ...

2026 | Vertraulich | Alle Rechte vorbehalten, Ergosign

Unser Protagonist

wohnraumplus

Ein fiktives mittelständisches Wohnungsunternehmen

40
Mitarbeitende

200+
Immobilien

Die Website
ist die wichtigste Anlaufstelle für Mietinteressenten

Das Problem

wohnraumplus

Mietinteressent, per Mail

„Welche Unterlagen brauche ich?“

zum 37. Mal diese Woche

Mietinteressentin, auf der Website

„Ich finde einfach nichts Passendes.“

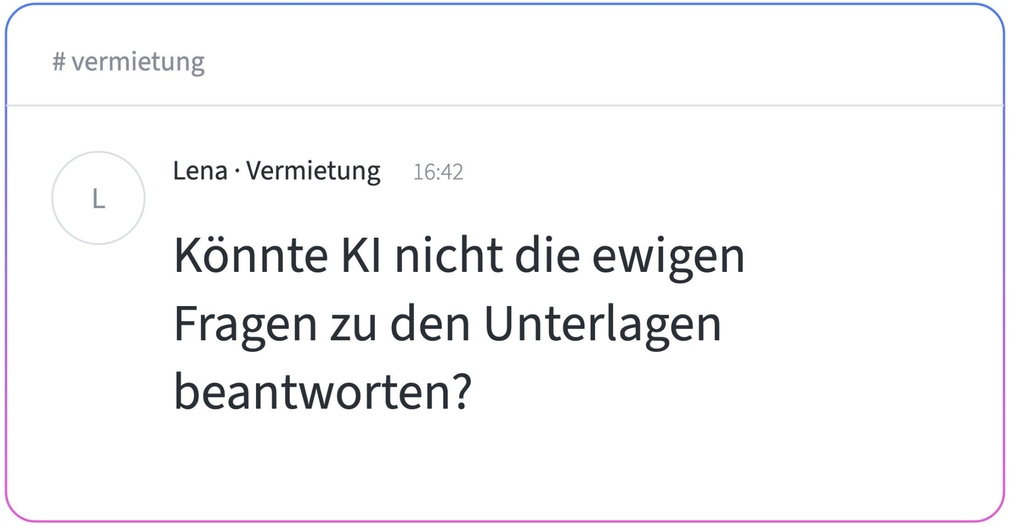
und sucht beim Nächsten weiter

Mehr Arbeit fürs Team, weniger Kunden.

Das Team fragt sich:

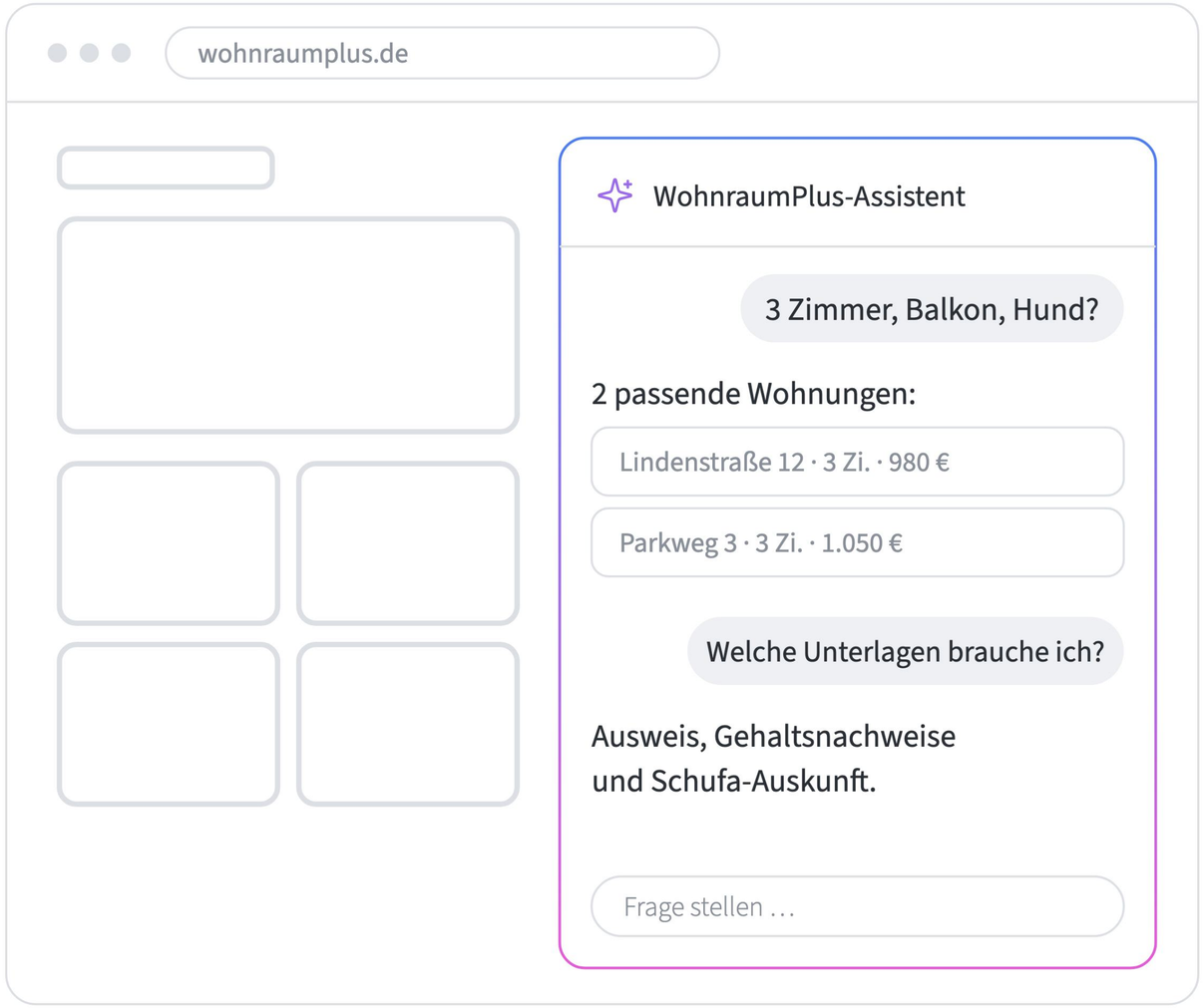
Und was machen wir mit KI?

Die Idee



FIKTIVES BEISPIEL

Ihr erster Prototyp



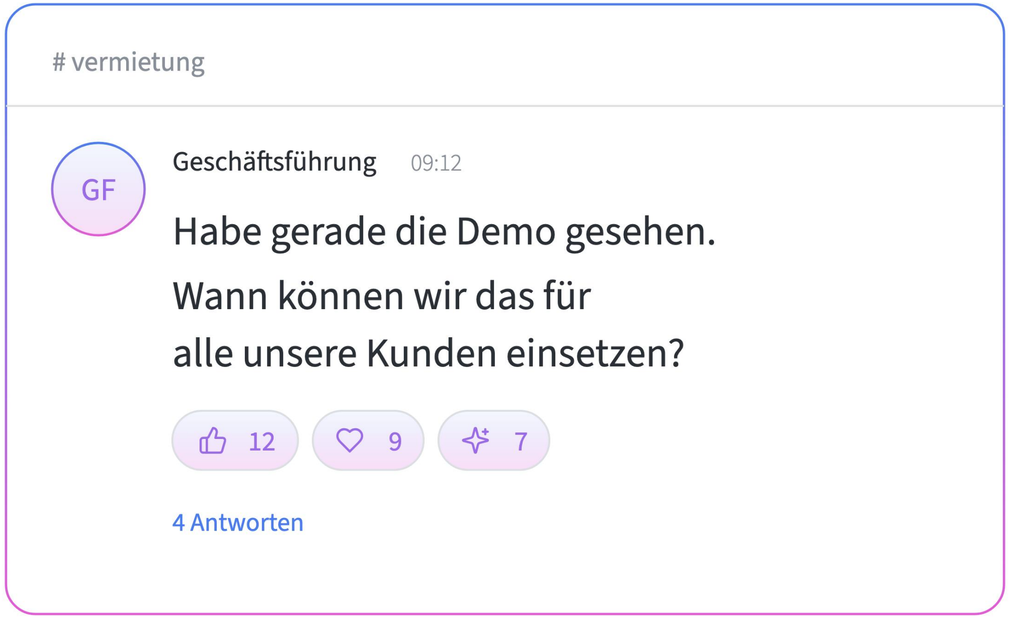
FIKTIVES BEISPIEL

Die erste Demo

Das Team nach der Demo

„Er weiß einfach alles.“

Das Team ist begeistert



FIKTIVES BEISPIEL

Weil die Ergebnisse so überzeugend waren,

haben sie ihn für alle Kunden ausgerollt ...

Weil die Ergebnisse so überzeugend waren,

haben sie ihn für alle Kunden ausgerollt ...

Die Krisenmomente

„Blindflug“
5 $ statt 0,50 $ pro Anfrage

„Prompt-Chaos“
Zurück nur über den Code

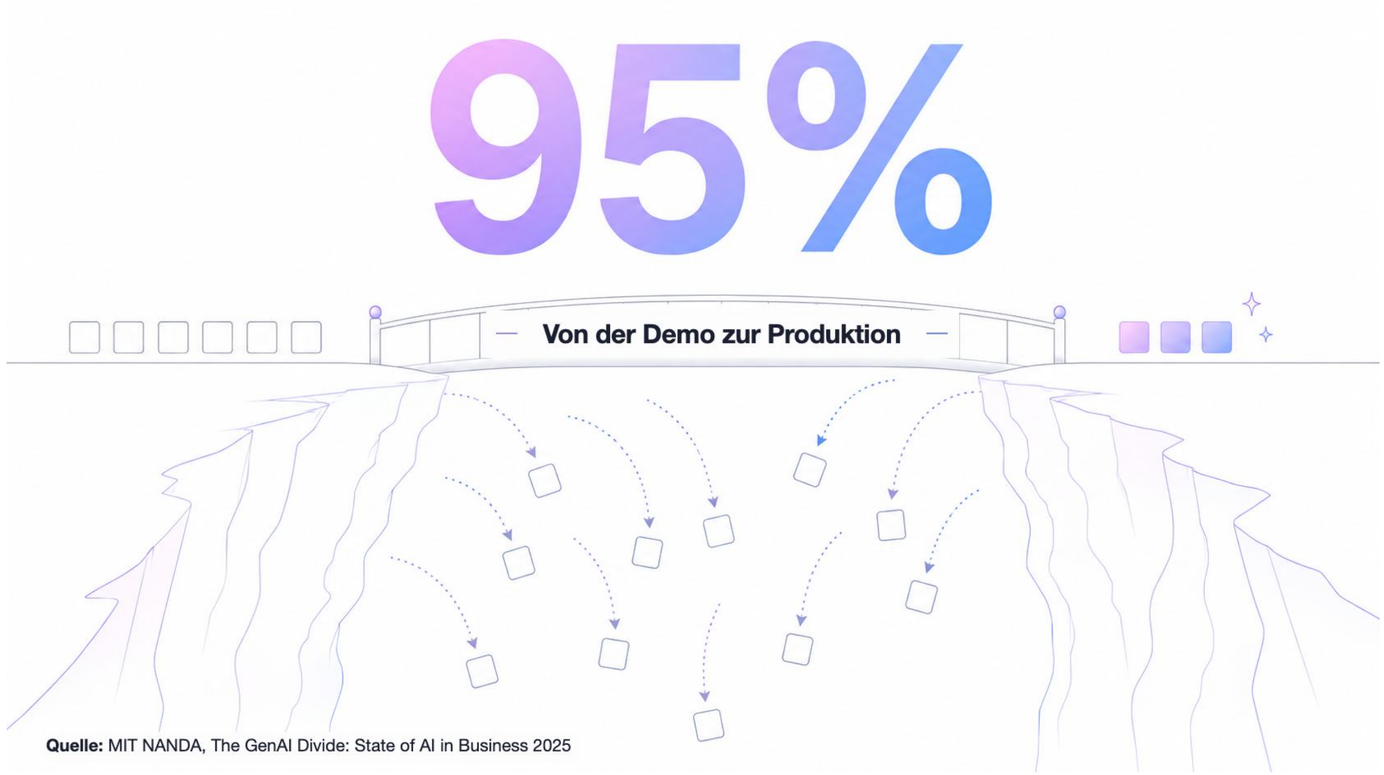
„Qualität unklar“
12 Wohnungen übersehen

„Missbrauch“
Bucht trotz Warteliste

2026 | Vertraulich | Alle Rechte vorbehalten, Ergosign

WohnraumPlus ist nicht allein

Eine Studie des MIT zeigt: 95 % der GenAI-Pilotprojekte in Unternehmen haben keinen messbaren Effekt auf den Gewinn.



Quelle: MIT NANDA, The GenAI Divide: State of AI in Business 2025

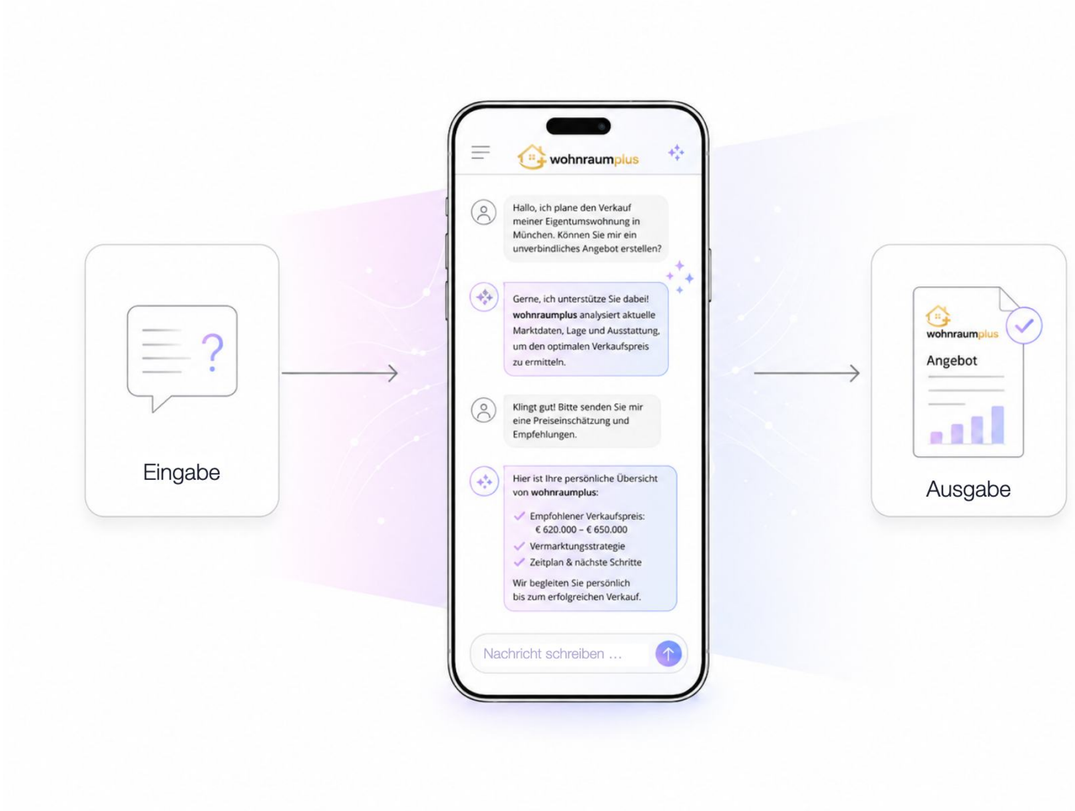
2026 | Vertraulich | Alle Rechte vorbehalten, Ergosign

Was passiert zwischen Eingabe und Ausgabe?

Von außen wirkt eine KI-Anwendung erstaunlich einfach:

Eine Frage geht hinein → Eine Antwort kommt heraus

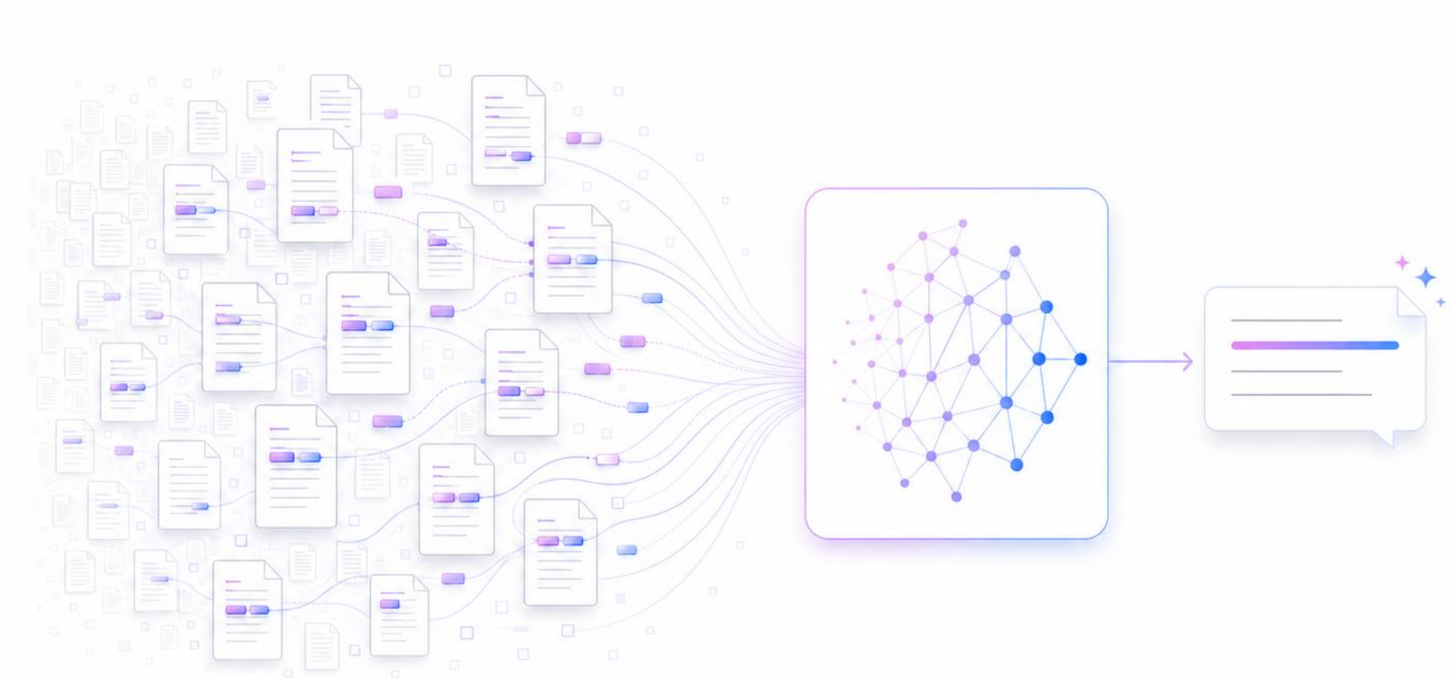
Doch was muss dazwischen alles passieren?



2026 | Vertraulich | Alle Rechte vorbehalten, Ergosign

Was ist ein Large Language Model?

Ein Large Language Model lernt Muster aus sehr großen Textsammlungen, oft mit mehreren hundert Milliarden Wörtern oder mehr.



2026 | Vertraulich | Alle Rechte vorbehalten, Ergosign

Was bedeutet ChatGPT eigentlich?

GPT steht für:

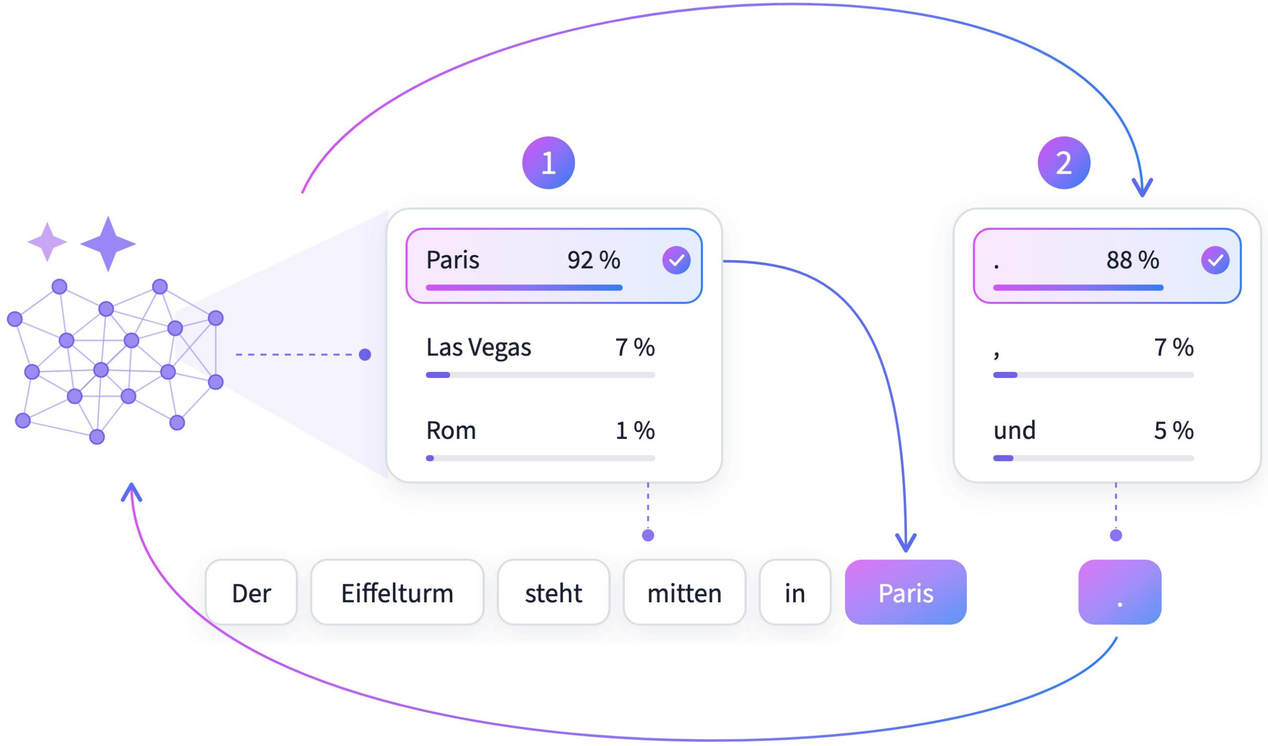
- **Generative:** Es erzeugt neue Inhalte auf Basis gelernter Muster.
- **Pretrained:** Es lernt diese Muster, bevor es in einer Anwendung eingesetzt wird.
- **Transformer:** Es nutzt eine Architektur neuronaler Netze, die Beziehungen zwischen den Teilen der Eingabe verarbeitet.

ChatGPT ist eine Anwendung von OpenAI und kein Modell. Das Sprachmodell dahinter lässt sich auswählen, zum Beispiel GPT-6.1-Sol.

2026 | Vertraulich | Alle Rechte vorbehalten, Ergosign

Was bedeutet „Generative“?

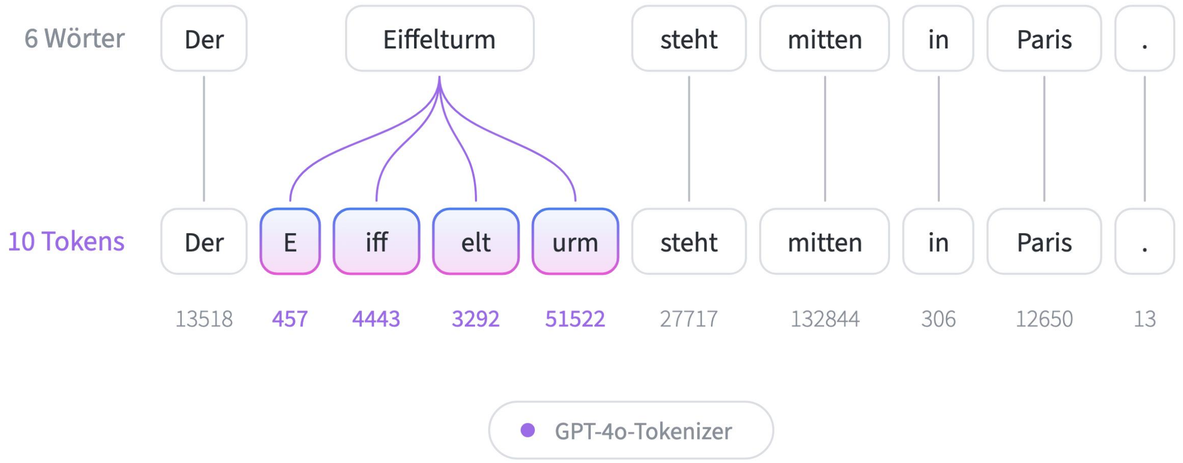
Das Modell erzeugt neuen Text, indem es Wort für Wort vorhersagt, welches Wort am wahrscheinlichsten als Nächstes kommt.



2026 | Vertraulich | Alle Rechte vorbehalten, Ergosign

Was ist ein Token?

Genau genommen arbeitet ein Sprachmodell mit Tokens. Das sind Wortstücke, und jedes bekommt eine feste Nummer. Auch die Kosten und die Textmenge, die ein Modell auf einmal lesen kann, werden in Tokens gemessen.

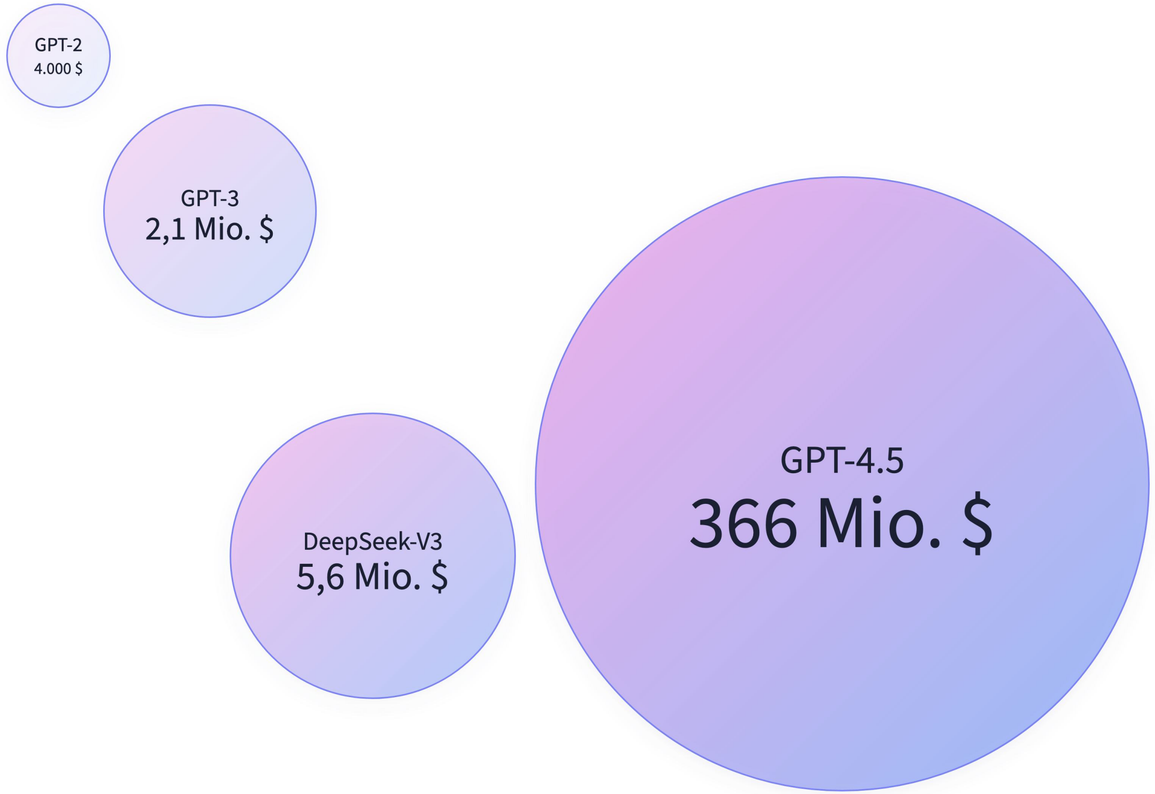


2026 | Vertraulich | Alle Rechte vorbehalten, Ergosign

Was bedeutet „Pretrained“?

Ein LLM wird trainiert, bevor es eingesetzt wird. Es wird nicht nebenbei durch Nutzernachrichten neu trainiert.

- Aktuelles Wissen lässt sich über den Prompt ergänzen
- Das Grundwissen des Modells ist auf seine Trainingsdaten begrenzt
- Das Training kostet Millionen von Dollar



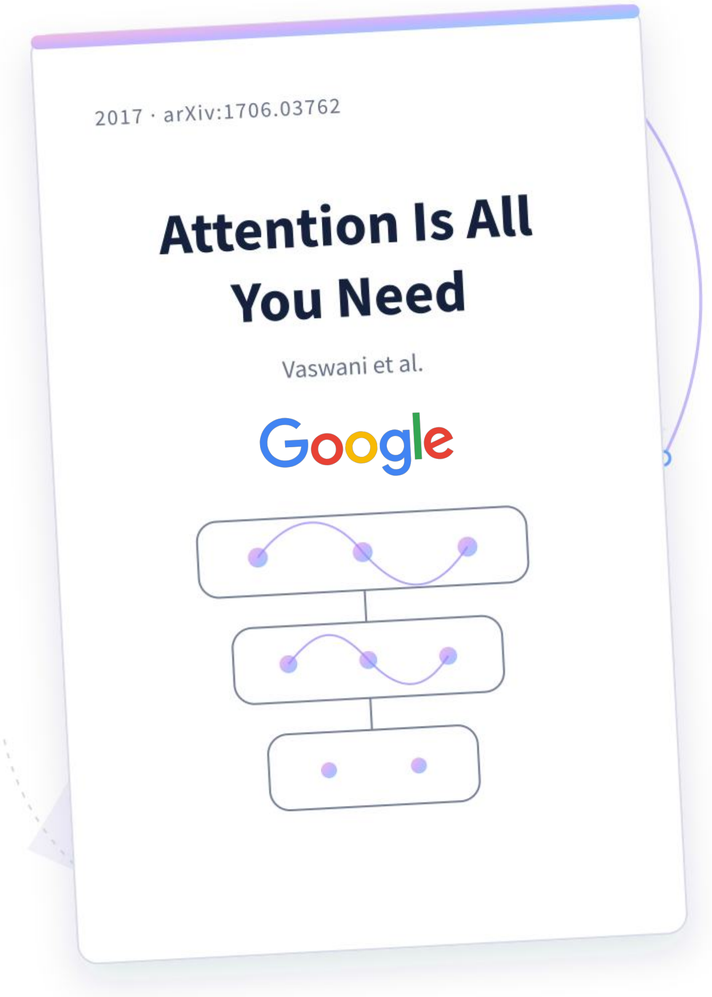
2026 | Vertraulich | Alle Rechte vorbehalten, Ergosign

Was bedeutet *Transformer*?

Ein Transformer ist die Architektur neuronaler Netze hinter modernen Large Language Models.

- Er ermöglicht es Modellen, Informationen innerhalb einer Eingabe zu verknüpfen und zusammenhängende Ausgaben zu erzeugen.
- Die Architektur wurde 2017 im Paper Attention Is All You Need vorgestellt.

GPT, Llama, DeepSeek und die meisten anderen modernen LLMs bauen auf dieser Grundlage auf.



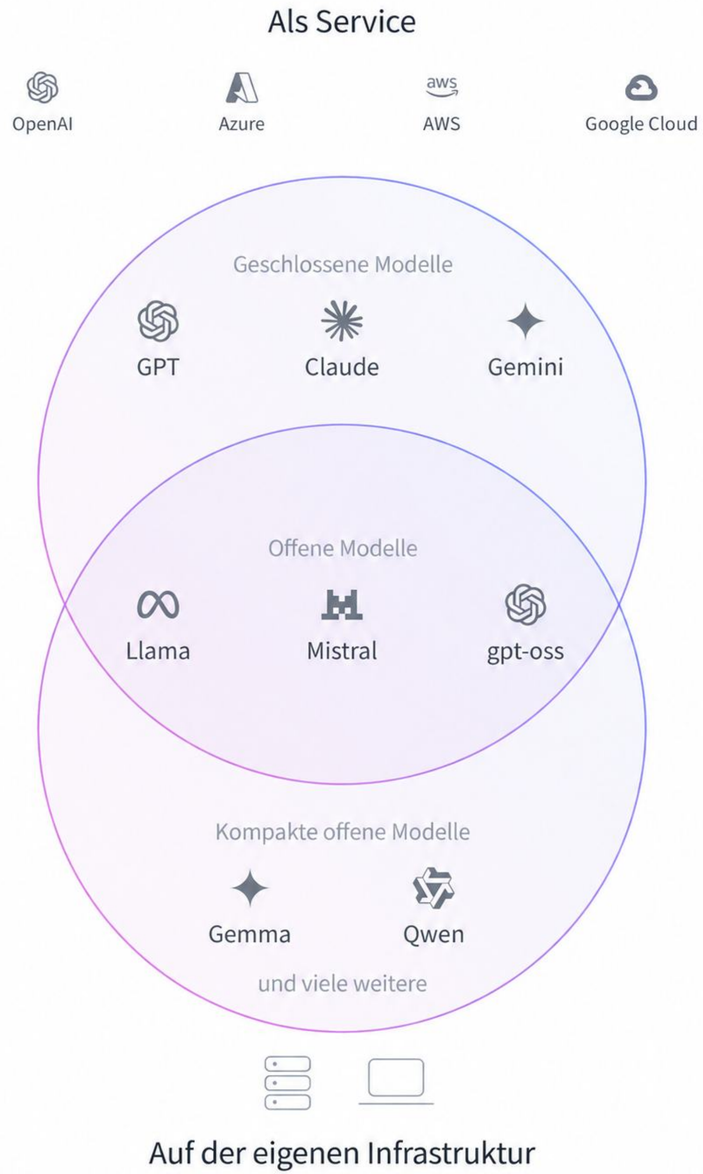
2026 | Vertraulich | Alle Rechte vorbehalten, Ergosign

Es gibt nicht das eine LLM

Es gibt viele Möglichkeiten zur Auswahl:

- Unterschiedliche Modelle und Anbieter
- Geschlossen oder offen
- Unterschiedliche Zugangswege

Das passende Setup hängt von Aufgabe, Kosten, vorhandener Infrastruktur und Compliance ab.



2026 | Vertraulich | Alle Rechte vorbehalten, Ergosign

Was weiß ein LLM nicht?

Neue Daten

Das Wissen endet zu einem festen Stichtag in den Trainingsdaten.

Ihre internen Daten

Modelle werden mit öffentlich verfügbaren Daten trainiert. Interne Dokumente Ihres Unternehmens gehören nicht dazu.

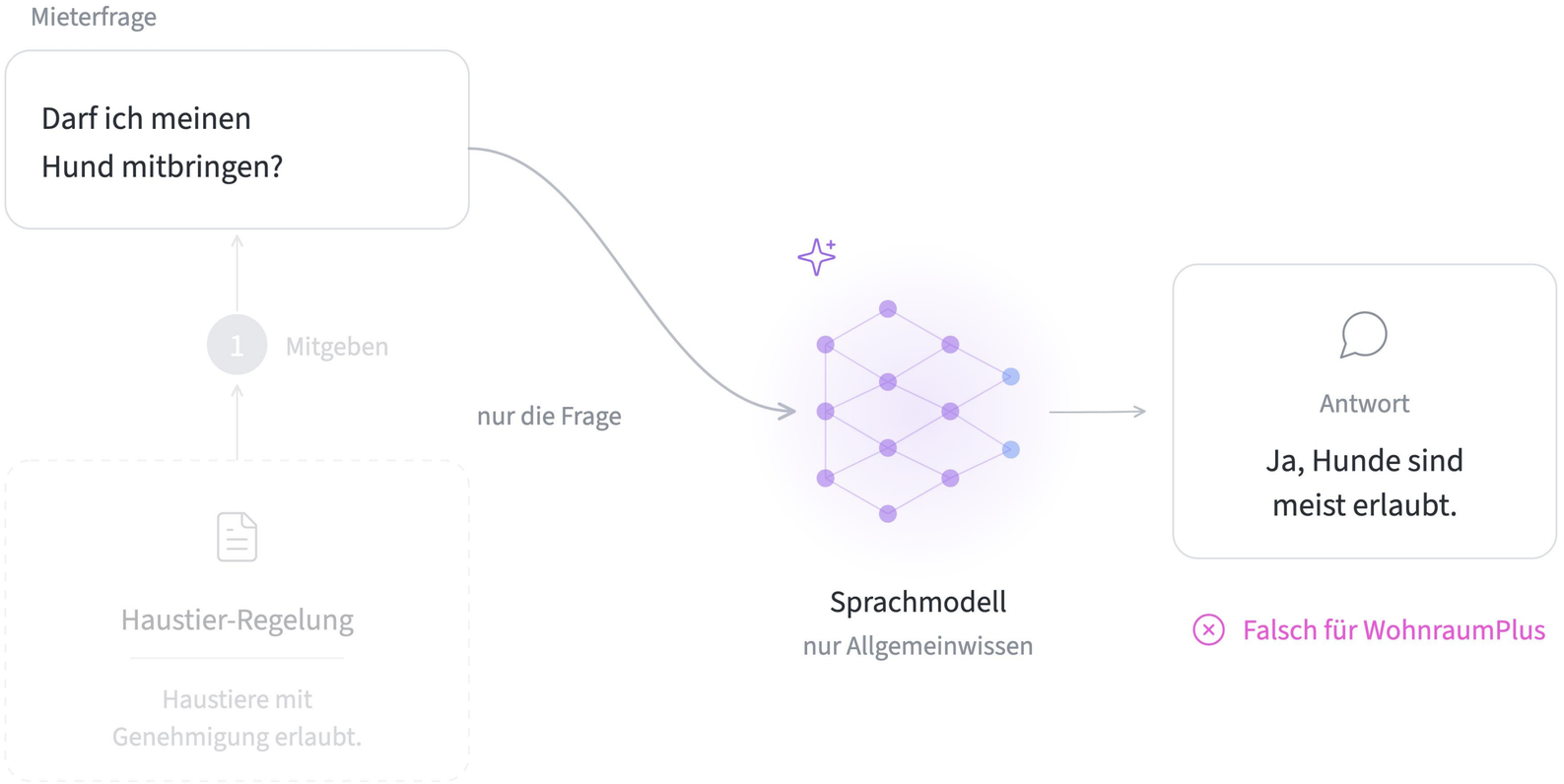
Live-Informationen

Live-Ticker, aktuell verfügbare Wohnungen: Informationen, die ständig aktualisiert werden müssen, um nützlich zu sein.

2026 | Vertraulich | Alle Rechte vorbehalten, Ergosign

Nachschlagen statt raten.

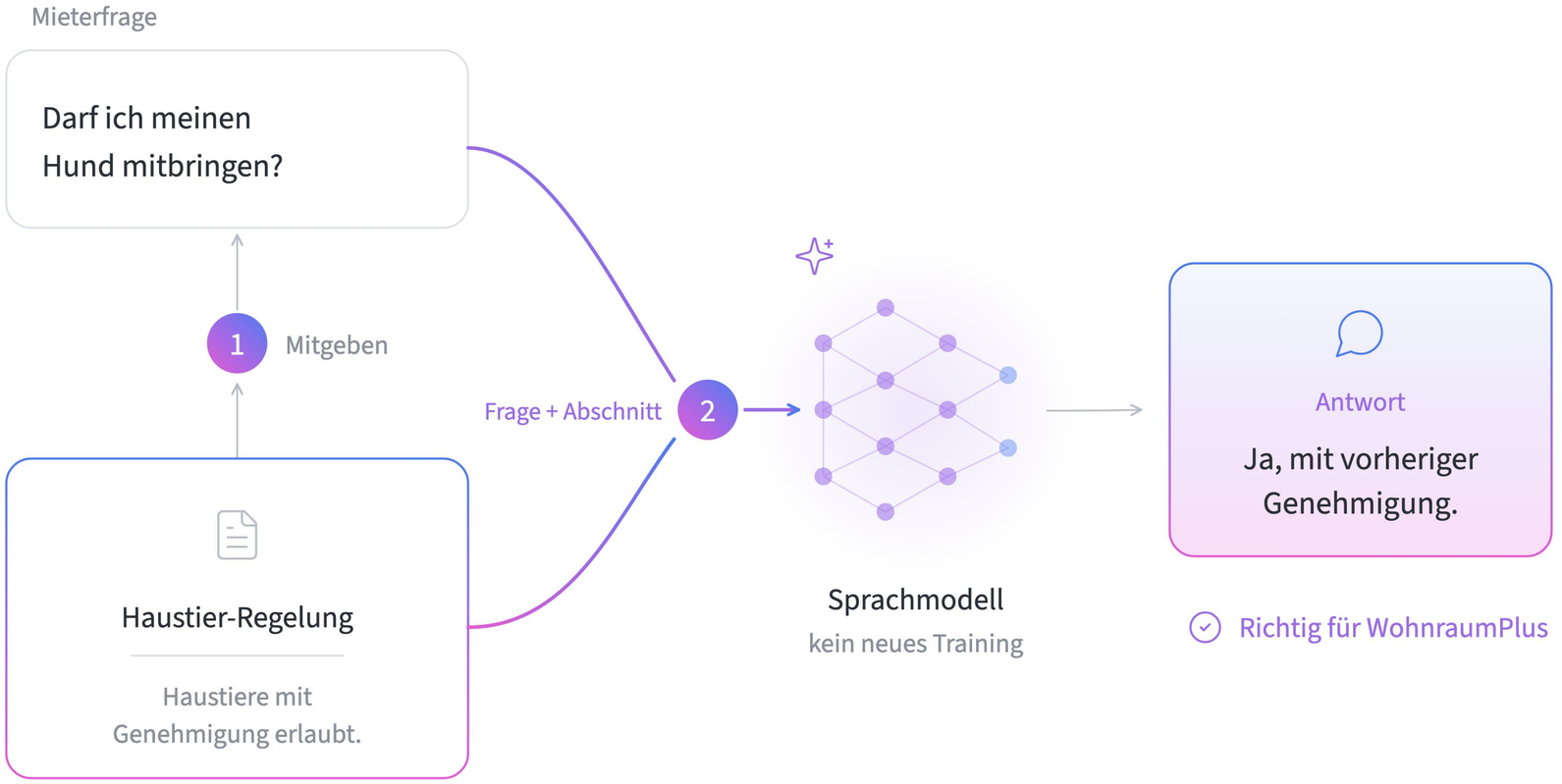
Der Assistent sucht die passende Regel von WohnraumPlus und gibt sie dem Modell mit.



2026 | Vertraulich | Alle Rechte vorbehalten, Ergosign

Nachschlagen statt raten.

Der Assistent sucht die passende Regel von WohnraumPlus und gibt sie dem Modell mit.



2026 | Vertraulich | Alle Rechte vorbehalten, Ergosign

Wie kann ein LLM **Ihre** Daten kennen?

Retraining

Nutzt einen strukturierten Datensatz Ihrer Daten, um das bestehende Modell mit neuen Gewichten nachzutrainieren.

sehr teuer

braucht genug Daten

2026 | Vertraulich | Alle Rechte vorbehalten, Ergosign

Wie kann ein LLM **Ihre** Daten kennen?

In-Context-Learning

Das LLM nutzt Informationen, die direkt im Prompt bzw. Kontext stehen, um Aufgaben zu lösen, ohne Training oder Fine-Tuning

RAG (Retrieval Augmented Generation)

Passende Informationen per Suche finden, in den Kontext legen und per In-Context-Learning nutzen. So bleibt das Wissen aktuell, ohne neu zu trainieren

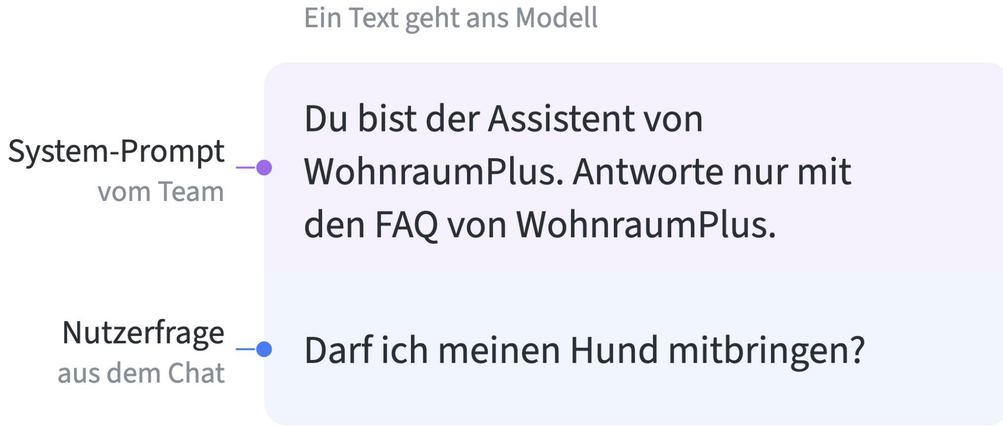
2026 | Vertraulich | Alle Rechte vorbehalten, Ergosign

Prompt und System-Prompt

Ein Prompt ist der gesamte Text, den das Sprachmodell als Eingabe erhält.

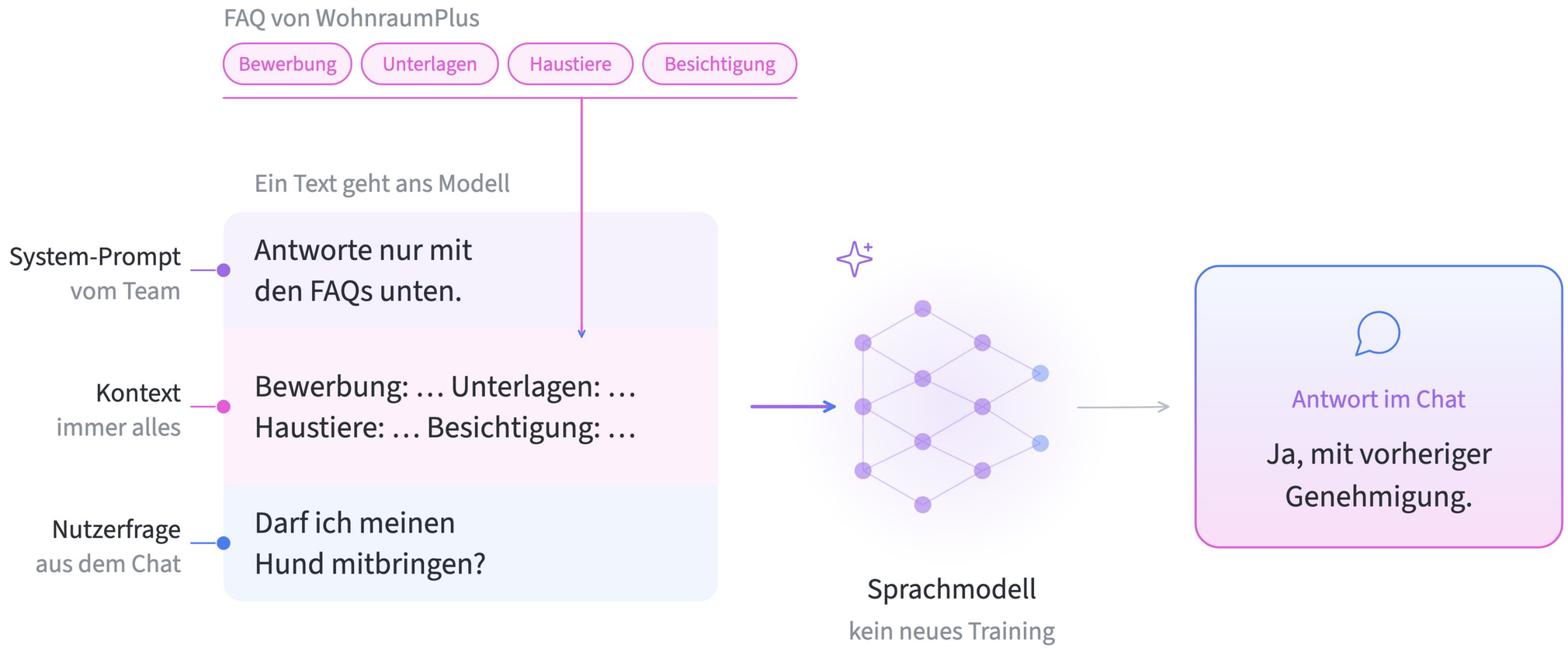
- **System-Prompt:** Anweisung des Teams hinter dem Assistenten, unsichtbar für die Mieterin
- **Nutzerfrage:** Frage der Mieterin im Chat

Widersprechen sich beide, ist das Modell darauf trainiert, dem System-Prompt zu folgen.



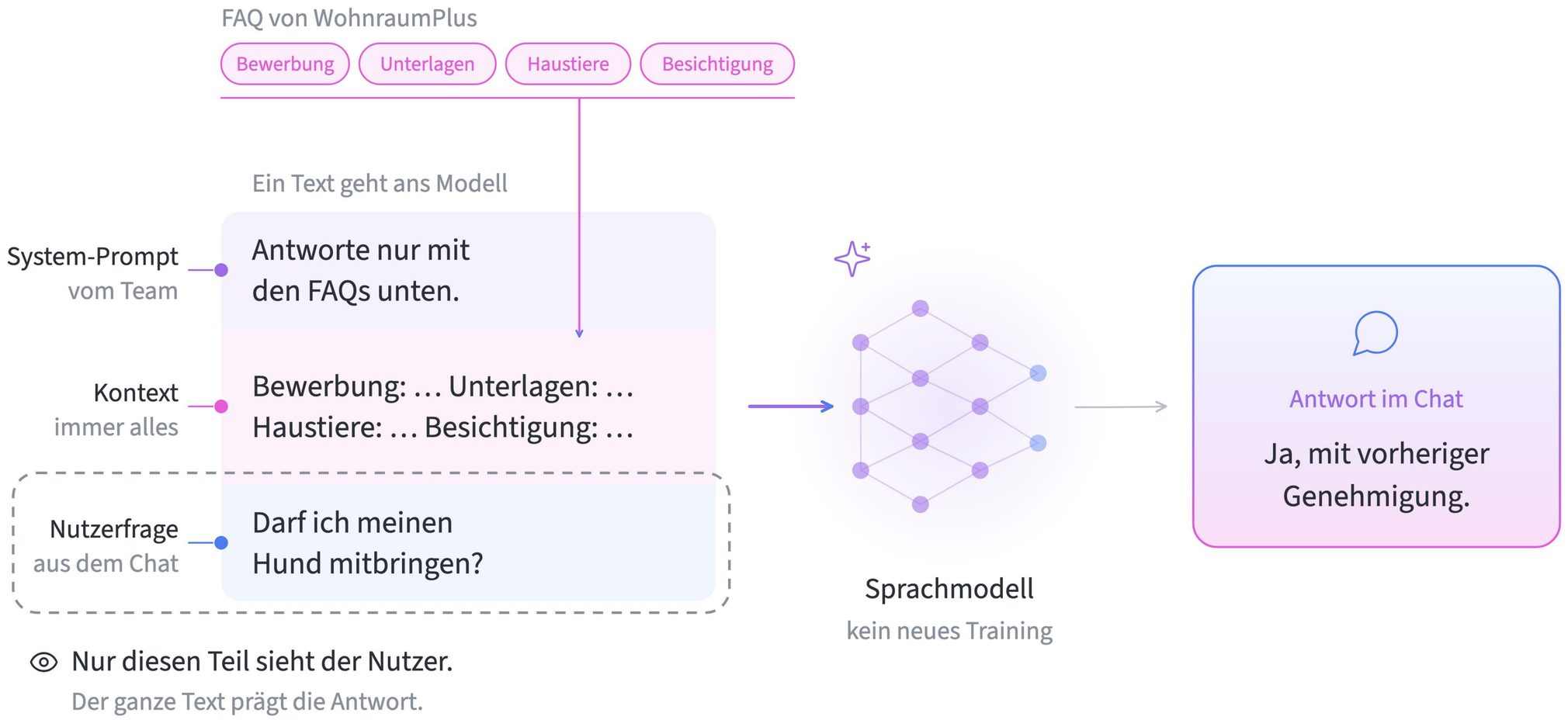
2026 | Vertraulich | Alle Rechte vorbehalten, Ergosign

Fachbegriff: „In-Context-Learning“



2026 | Vertraulich | Alle Rechte vorbehalten, Ergosign

Fachbegriff: „In-Context-Learning“



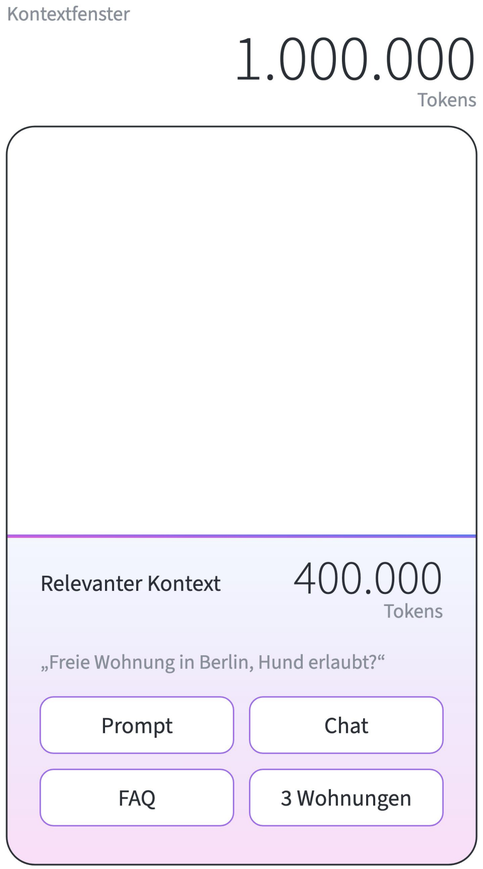
2026 | Vertraulich | Alle Rechte vorbehalten, Ergosign

Warum nicht alle Ihre Daten ins Kontextfenster legen?

Ein Kontextfenster umfasst alles, was eine KI vor ihrer Antwort berücksichtigen kann.

- Der Platz ist begrenzt.
- Anweisungen, Chatverlauf, Quellenmaterial und die Antwort selbst belegen ihn.
- Mehr Informationen können das Modell von dem ablenken, was für die aktuelle Aufgabe zählt.

Deshalb können und sollen nicht alle Informationen per In-Context-Learning eingespeist werden.



2026 | Vertraulich | Alle Rechte vorbehalten, Ergosign

Wie kann ein LLM **Ihre** Daten kennen?

In-Context-Learning

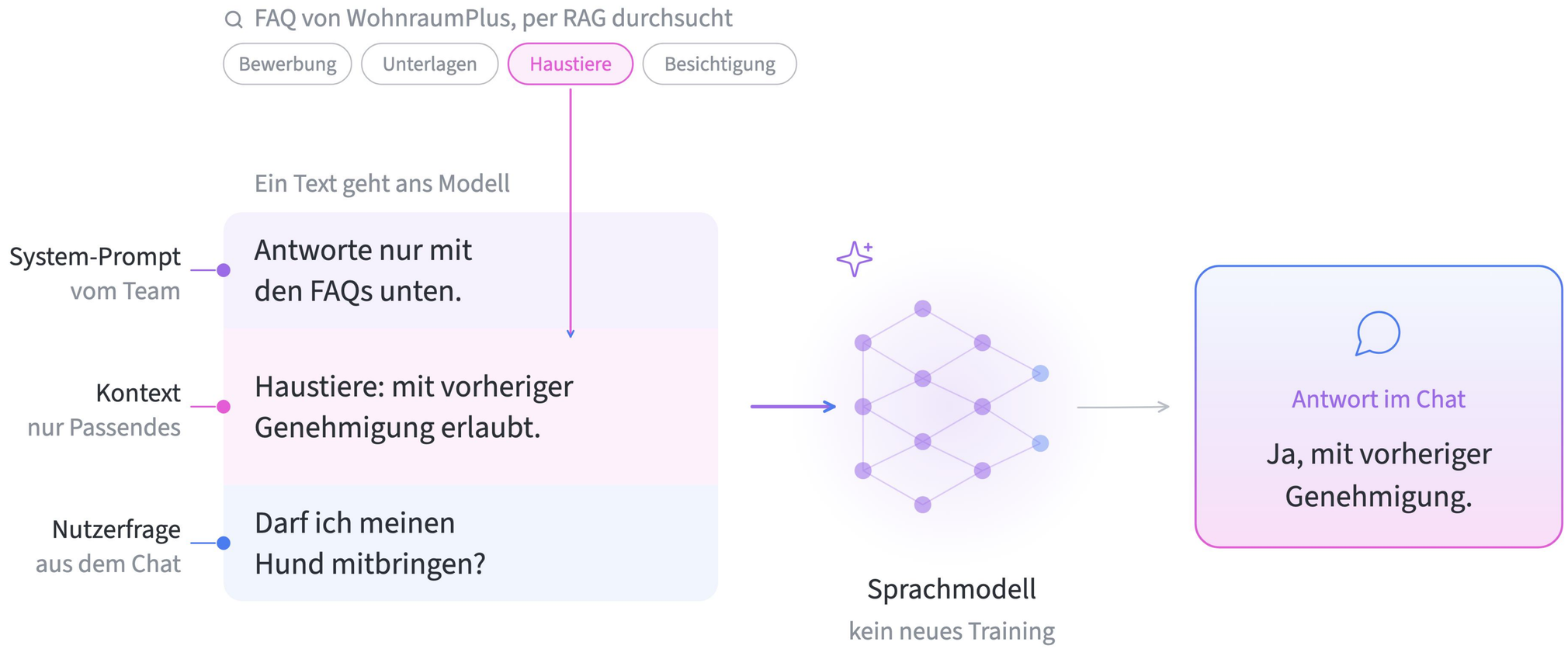
Das LLM nutzt Informationen, die direkt im Prompt bzw. Kontext stehen, um Aufgaben zu lösen, ohne Training oder Fine-Tuning

RAG (Retrieval Augmented Generation)

Passende Informationen per Suche finden, in den Kontext legen und per In-Context-Learning nutzen. So bleibt das Wissen aktuell, ohne neu zu trainieren

2026 | Vertraulich | Alle Rechte vorbehalten, Ergosign

Fachbegriff: „Retrieval Augmented Generation (RAG)“



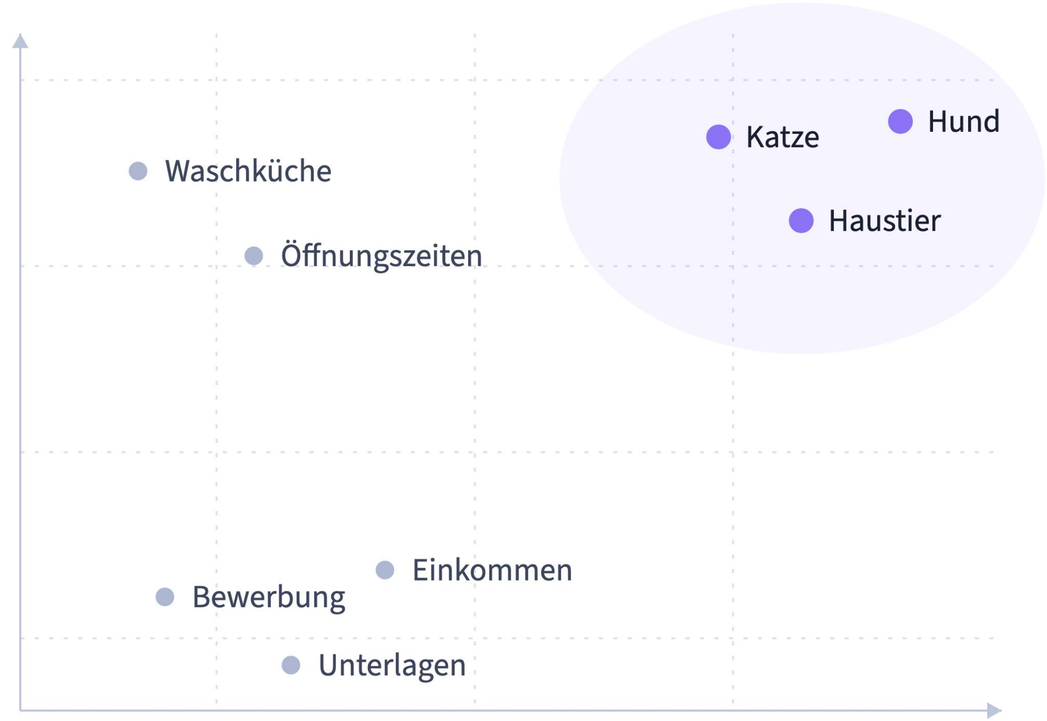
2026 | Vertraulich | Alle Rechte vorbehalten, Ergosign

Embeddings bilden Bedeutung ab

Ein Embedding stellt Text als Position auf einer Zahlenkarte dar.

- Texte mit ähnlicher Bedeutung liegen nah beieinander.
- Texte mit unterschiedlicher Bedeutung liegen weiter auseinander.
- Das Modell lernt diese Positionen aus Mustern in der Sprache.

Eine Frage zu „Hund“ kann deshalb ein Dokument über „Haustiere“ finden.



Wörter mit ähnlichem Kontext liegen nah beieinander.
Vereinfachte 2D-Ansicht. Echte Embeddings nutzen sehr viele gelernte Dimensionen.

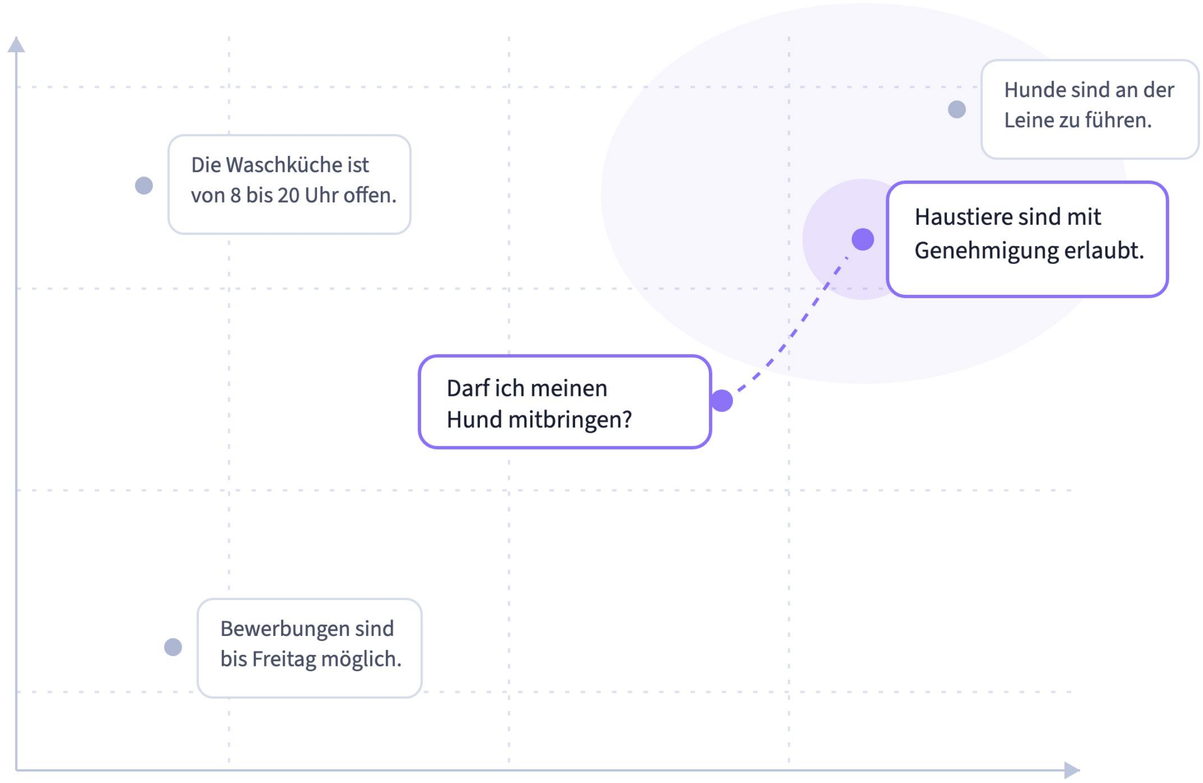
2026 | Vertraulich | Alle Rechte vorbehalten, Ergosign

Semantische Suche vergleicht Bedeutung

Die semantische Suche nutzt Embeddings, um Fragen mit Textabschnitten zu vergleichen.

- Jede Frage und jeder Abschnitt erhält eine Position auf derselben Karte.
- Ein kurzer Abstand bedeutet eine ähnliche Bedeutung.
- Die Wortwahl kann völlig unterschiedlich sein.

„Darf ich meinen Hund mitbringen?“ kann trotzdem „Haustiere sind mit Genehmigung erlaubt.“ finden.



Unterschiedliche Wörter können dasselbe bedeuten.

Die semantische Suche vergleicht ganze Fragen und Textabschnitte auf derselben Karte.

2026 | Vertraulich | Alle Rechte vorbehalten, Ergosign

RAG ergänzt Unternehmenswissen

Das Modell lernt Ihre Daten nicht. Es bekommt den passenden Auszug.

2026 | Vertraulich | Alle Rechte vorbehalten, Ergosign

RAG hängt weiterhin von der Datenqualität ab

RAG kann nur Informationen finden, die in der Wissensbasis vorhanden und nutzbar sind.

Typische Probleme bleiben:

- Veraltete Dokumente
- Widersprüchliche Dokumente
- Schlecht aufbereitete Dokumente

2026 | Vertraulich | Alle Rechte vorbehalten, Ergosign

Was, wenn sich die Daten ständig ändern?

Wie steht es bei Arsenal gegen Bayern München?

Der Spielstand kann sich jede Sekunde ändern, veraltete Daten helfen den Nutzern nicht.

Welche Wohnungen sind aktuell verfügbar?

Diese Daten von WohnraumPlus können sich jede Minute ändern. Das LLM sollte niemals bereits vergebene Wohnungen empfehlen.

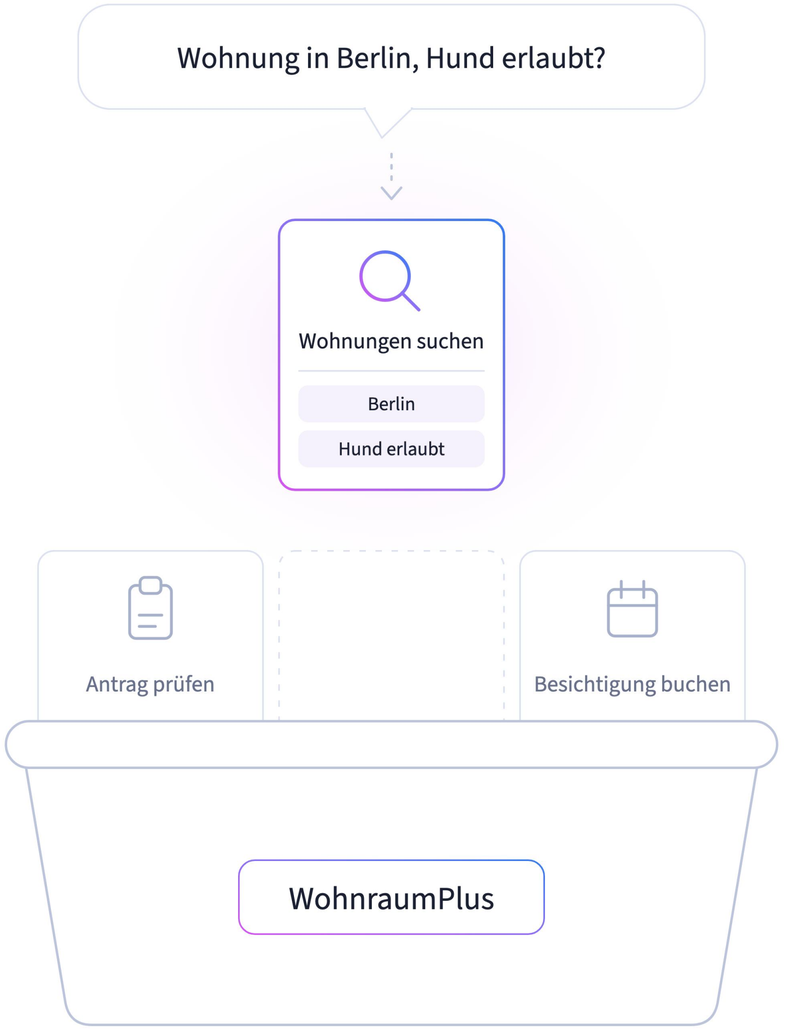
2026 | Vertraulich | Alle Rechte vorbehalten, Ergosign

Tool Calling

Die KI erhält eine Liste von Aktionen, die sie anfordern darf.

- Jedes Tool hat einen klaren Namen und eine kurze Beschreibung, was es tut.
- Jedes Tool legt fest, welche Angaben es braucht, etwa einen Ort oder ob Haustiere erlaubt sind.
- Die KI wählt ein passendes Tool und füllt die Felder aus der Frage der Nutzer aus.

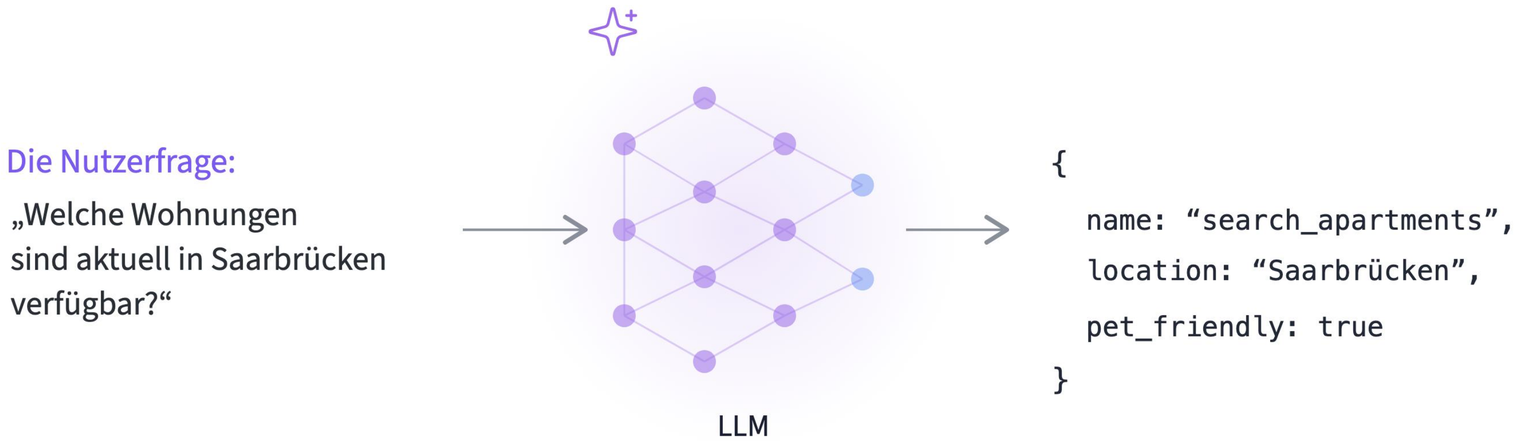
Für die KI sieht ein Tool aus wie ein kleines Formular, das sie ausfüllen muss.



2026 | Vertraulich | Alle Rechte vorbehalten, Ergosign

Tool Calling

Tool Calling bedeutet, dass die KI eine vordefinierte Funktion anfordern kann, zum Beispiel die Suche nach verfügbaren Wohnungen.



2026 | Vertraulich | Alle Rechte vorbehalten, Ergosign

Was, wenn die KI auch handeln kann?

FIKTIVES BEISPIEL

2026 | Vertraulich | Alle Rechte vorbehalten, Ergosign

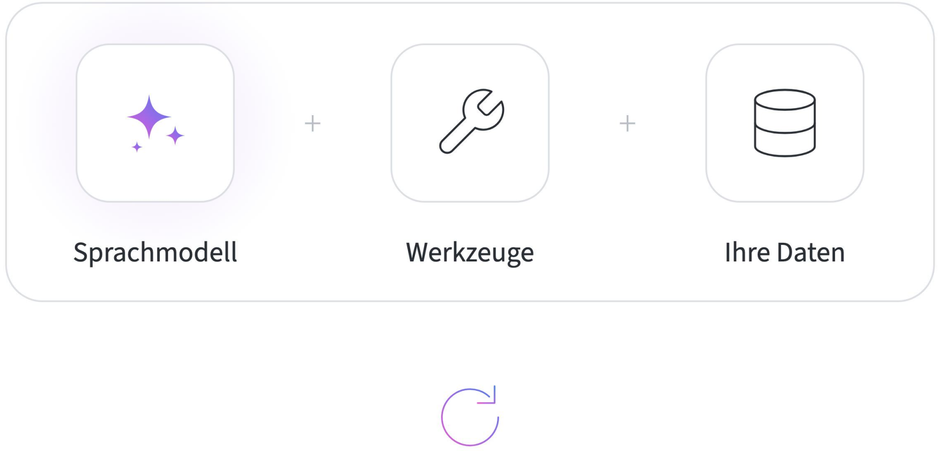
Was ist ein Agent?

Ein Agent ist eine Schleife. In jeder Runde fragt er sich:

Kann ich die Frage jetzt beantworten?

- Wenn ja, antwortet er und hört auf.
- Wenn nein, wählt er das Tool, das ihn weiterbringt.

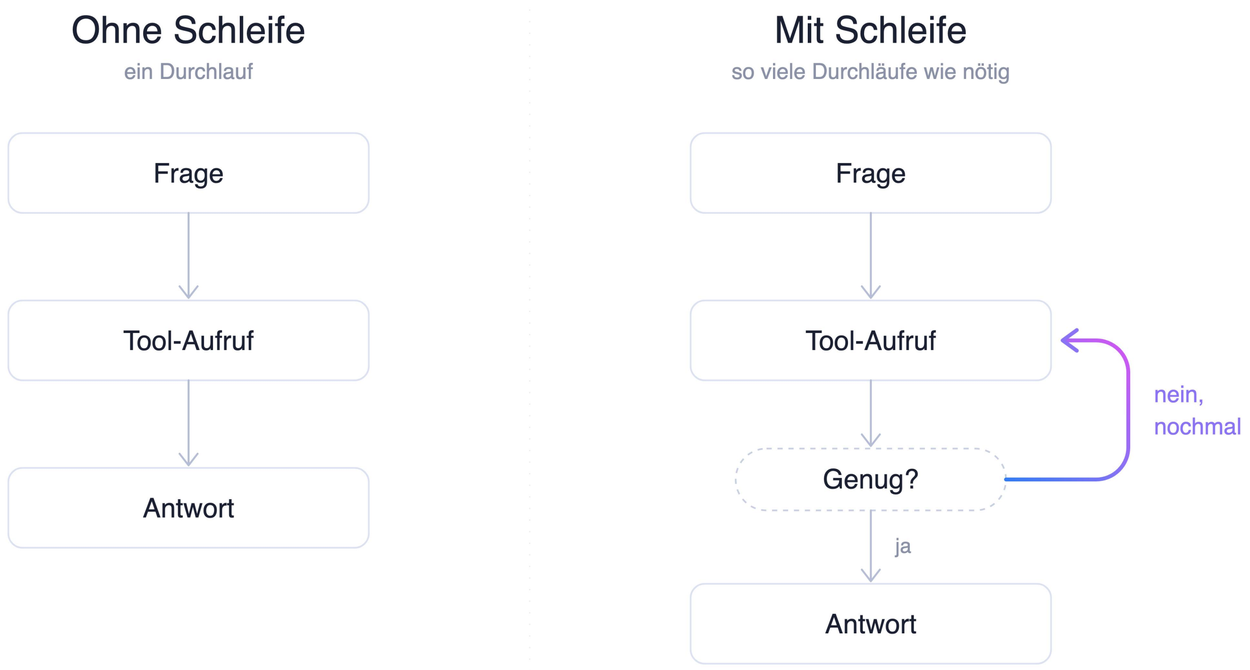
Ihr System führt dieses Tool aus und das Ergebnis geht zurück in den Kontext. Dann beginnt die Schleife mit derselben Frage von vorn.



2026 | Vertraulich | Alle Rechte vorbehalten, Ergosign

Der Unterschied ist die Schleife

Ein Agent muss nicht sofort antworten. Er kann sich ansehen, was zurückkam, und entscheiden, noch eine Runde zu drehen.



2026 | Vertraulich | Alle Rechte vorbehalten, Ergosign

Ihre Daten sind weiterhin der wichtigste Faktor.

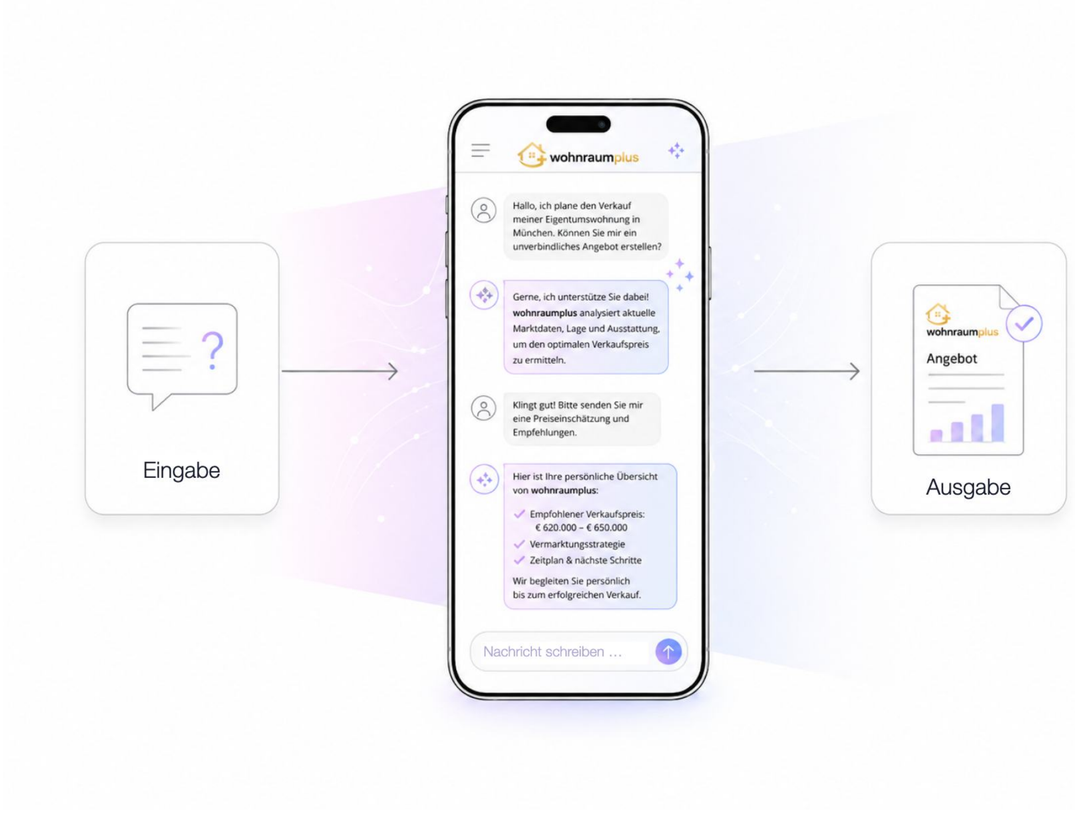
2026 | Vertraulich | Alle Rechte vorbehalten, Ergosign

Was passiert zwischen Eingabe und Ausgabe?

Von außen wirkt eine KI-Anwendung erstaunlich einfach:

Eine Frage geht hinein → Eine Antwort kommt heraus

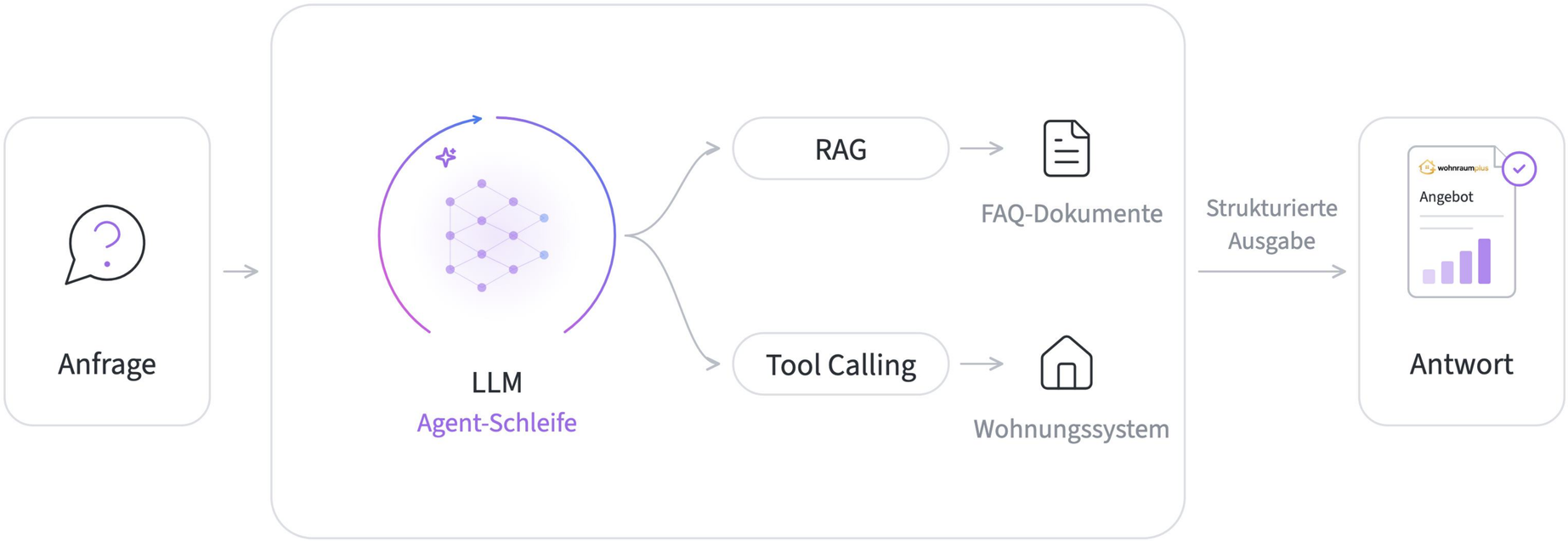
Aber was muss dazwischen alles passieren?



2026 | Vertraulich | Alle Rechte vorbehalten, Ergosign

Die aktuelle Lösung von WohnraumPlus

Das hat WohnraumPlus tatsächlich gebaut. Jeder Teil davon kam in den letzten Folien vor.



2026 | Vertraulich | Alle Rechte vorbehalten, Ergosign

Das Fundament steht

Teil 1 · Heute

Generative KI verstehen

Vom LLM bis zum Agenten

Teil 2 · Nächstes Mal

GenAI Ops

Wie WohnraumPlus im Live-Betrieb zuverlässig wird

2026 | Vertraulich | Alle Rechte vorbehalten, Ergosign

Was wir gelernt haben und was noch kommt

Teil 1 · Heute gelernt

Generative KI verstehen

- LLM ✓
- Prompts ✓
- RAG ✓
- Agenten ✓
- Beispiel WohnraumPlus ✓

Teil 2 · Nächstes Mal

GenAI Ops

- Tracing
- Prompt-Management
- Evals
- Guardrails
- Praxisfall Fichtel-Tagesplaner

2026 | Vertraulich | Alle Rechte vorbehalten, Ergosign

Vielen Dank!
Haben Sie Fragen?

Nicolas Klein

nicolas.klein@ergosign.de

Hochschule
Bonn-Rhein-Sieg
University of Applied Sciences

Mittelstand-Digital
Zentrum
Fokus Mensch

ERGOSIGN