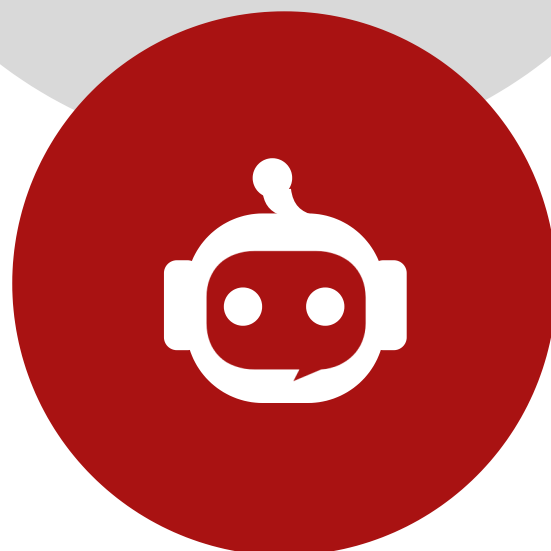


■ **brandcompetete**

AIMan

Benutzerhandbuch – alle führenden KI-Modelle
in einem Arbeitsbereich, Governance ab dem
ersten Prompt

RELEASE YOUR INNER PROMPT ENGINEER.



Inhalt

1	Willkommen bei AIMan	2
2	Erste Schritte	4
	Anmelden	4
	Sich zurechtfinden	5
	Ihr erster Prompt	6
3	Modelle	7
	Ein Modell in der Modellauswahl wählen	8
	Welches Modell für welche Aufgabe	9
4	Chats, Projekte und Tabs	10
	Mehrere Modelle nebeneinander: Tabs	10
	Mit Antworten arbeiten	11
	Projekte: einen Chat mit Kollegen teilen	12
5	Das Eingabefeld	13
	Thinking: das Modell erst nachdenken lassen	15
6	Websuche	16
7	Datenquellen: Antworten aus Ihren Dokumenten	18
	Eine Datenquelle anlegen	18
	Wissensdatenbanken: Confluence, Jira, Bitbucket	20
	Fragen an Ihre Dokumente stellen	21
	Datenquellen teilen, und was noch nicht geht	22
8	Workbench, Prompt-Bibliothek und eigene Anweisungen	23
9	Chatbots für Websites und Anwendungen	25
10	Plugins	26
11	Ihre Einstellungen	27
12	Administration	28
	Modelle, geladene Modelle und Assistenten-Anweisungen	28
	Benutzer, Rollen und Organisationseinheiten	30
	Schlüssel, Wissensdatenbanken, Branding und Anmeldung	31
13	Die API nutzen	33
14	Daten und Datenschutz: was wohin geht	35
15	Fehlerbehebung & FAQ	36
A	Referenz: Rollen, Status, Tastenkürzel, Glossar	38

1 Willkommen bei AIman

AIman bringt die führenden KI-Modelle in einen Arbeitsbereich für das ganze Unternehmen: eine Anmeldung, eine Oberfläche, Antworten auf Grundlage Ihrer Dokumente und des Webs – und Regeln, die die IT einmal festlegt.

Sie melden sich mit Ihrem Arbeitskonto an und chatten über dieselbe Oberfläche mit zwei Arten von Modellen. **Offene Modelle** wie Mistral NeMo, Qwen, Gemma, gpt-oss, Phi und DeepSeek laufen auf den dedizierten GPU-Servern von brandcompete in Deutschland – oder auf den eigenen Servern Ihres Unternehmens, wenn es AIman on-premise betreibt. **Anbietermodelle** – Anthropic Claude, OpenAI GPT, Azure OpenAI, Google Gemini, Cohere und Ollama Cloud – laufen beim jeweiligen Anbieter, mit dem Schlüssel Ihres Unternehmens. Niemand braucht für die Arbeit ein privates KI-Konto.

Antworten können auf das Wissen Ihres Unternehmens zurückgreifen: hochgeladene Dokumente, Webseiten sowie Confluence, Jira oder Bitbucket. Jede Antwort nennt, worauf sie sich stützt, mit Dokument und Seite. Mit eingeschalteter Websuche beantwortet ein Modell auch Fragen zum aktuellen Geschehen – mit Links zu den Quellen.

Womit Sie arbeiten

Begriff	Was es ist
Chat	Eine eigene Unterhaltung. Niemand sonst sieht sie.
Projekt	Ein geteilter Chat. Sie laden Kolleginnen und Kollegen, Teams, Organisationseinheiten oder das ganze Unternehmen ein und arbeiten gemeinsam darin.
Tab	Jeder Chat und jedes Projekt hat einen oder mehrere Tabs. Jeder Tab hat sein eigenes Modell, seine eigene Datenquelle und seine eigene Unterhaltung – so vergleichen Sie Modelle direkt nebeneinander.
Modell	Die KI, die antwortet. Modelle sind in Familien gruppiert (etwa <i>Anthropic</i> oder <i>qwen3.5</i>); eine Familie hat eine oder mehrere konkrete Versionen, sogenannte Tags .
Eingabefeld	Das Nachrichtenfeld unten im Chat, mit Modell, Websuche, Anhängen und Optionen.
Datenquelle	Eine Sammlung von Dokumenten, Webseiten oder eine angebundene Wissensdatenbank, in der ein Modell vor der Antwort nachschlagen kann.
Workbench	Die Leiste rechts mit Ihren gespeicherten Prompts, Ihren Datenquellen und Ihren eigenen Anweisungen.
Prompt-Bibliothek	Gespeicherte Prompts mit <code>{{Platzhaltern}}</code> für die wechselnden Teile, mit Kategorien, Tags und Freigaben (in der Oberfläche: Prompt Repository).
Eigene Anweisung	Fester Kontext, den das Modell mit jedem Prompt erhält: wer Sie sind, wie es antworten soll (Custom Instructions).
Chatbot	Ein Chat-Tab, mit seinem Modell und seiner Datenquelle als Chatbot exportiert, den Sie in eine Website oder Anwendung einbetten.

Was Sie tun können, hängt von Ihrer Rolle ab

Jedes Konto hat eine **Rolle**, etwa Staffmember, Teamlead oder Administrator. Die Rolle bestimmt, was Sie anlegen und verwalten dürfen: Chats, Datenquellen, Prompts, Chatbots – oder Benutzer und Modelle. Bereiche, die Ihre Rolle nicht umfasst, werden nicht angezeigt. Alles, was Sie anlegen, ist **privat, bis Sie es teilen** – mit einzelnen Kolleginnen und Kollegen, einem Team, einer Organisationseinheit oder dem ganzen Unternehmen, zum Lesen, zum Bearbeiten oder zum Verwalten. Die Rollen finden Sie in Anhang A.

Wohin Ihre Daten gehen, in einem Absatz

Die Modellauswahl zeigt offene Modelle und Anbietermodelle in getrennten Tabs, Sie sehen also immer, wohin ein Prompt geht. Mit einem offenen Modell bleiben Prompt, Antwort und alle Dokumentpassagen auf den GPU-Servern Ihrer AlMan-Installation. Mit einem Anbietermodell gehen der Prompt und die nötigen Passagen an diesen Anbieter, unter dem Vertrag und Schlüssel Ihres Unternehmens. Das Lesen und Indexieren Ihrer Dokumente geschieht immer auf diesen Servern, egal welches Modell antwortet. Einzelheiten stehen in Kapitel 14.

HINWEIS

Dieses Handbuch beschreibt AlMan mit Stand Oktober 2026. Die Oberfläche ist englischsprachig; Bezeichnungen auf dem Bildschirm sind hier genau so zitiert, wie sie erscheinen. Die Modelle antworten in der Sprache, in der Sie schreiben. Einige Bereiche, die Sie im Produkt sehen, sind noch nicht fertig; sie sind mit **Noch nicht verfügbar** gekennzeichnet.

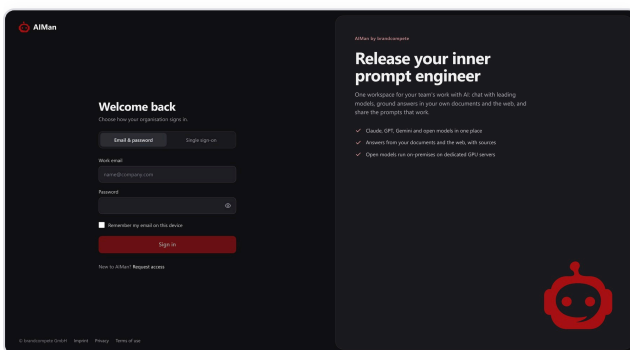
2 Erste Schritte

Anmelden, sich im Fenster zurechtfinden und den ersten Prompt senden.

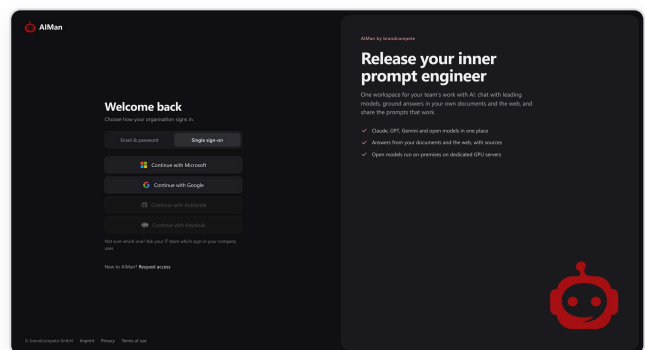
Anmelden

Öffnen Sie die AIMan-Adresse, die Ihr Unternehmen Ihnen genannt hat. Die Anmeldeseite hat zwei Tabs:

- **Email & password** – geben Sie unter **Work email** Ihre geschäftliche E-Mail-Adresse und unter **Password** Ihr Passwort ein und klicken Sie auf **Sign in**. Das Augensymbol zeigt das Passwort beim Tippen an. Mit **Remember my email on this device** ist die Adresse beim nächsten Mal schon ausgefüllt. Groß- und Kleinschreibung der Adresse spielt keine Rolle.
- **Single sign-on** – klicken Sie auf **Continue with Microsoft** oder **Continue with Google** und melden Sie sich mit dem Arbeitskonto an, das Sie ohnehin haben. AIMan unterstützt Microsoft Entra ID und Google; die Administration richtet sie pro Installation ein. Schaltflächen für nicht eingerichtete Anbieter sind ausgegraut.



Anmelden mit E-Mail und Passwort



Der Tab für Single Sign-on mit eingerichtetem Microsoft und Google

Single Sign-on lässt nur Adressen aus den E-Mail-Domains und Verzeichnissen zu, die Ihr Unternehmen freigegeben hat. Erscheint nach der Anmeldung mit Microsoft oder Google eine Fehlermeldung, nennt sie den Grund – etwa, dass Ihre Domain nicht zugelassen ist – und der Tab für Single Sign-on öffnet sich erneut.

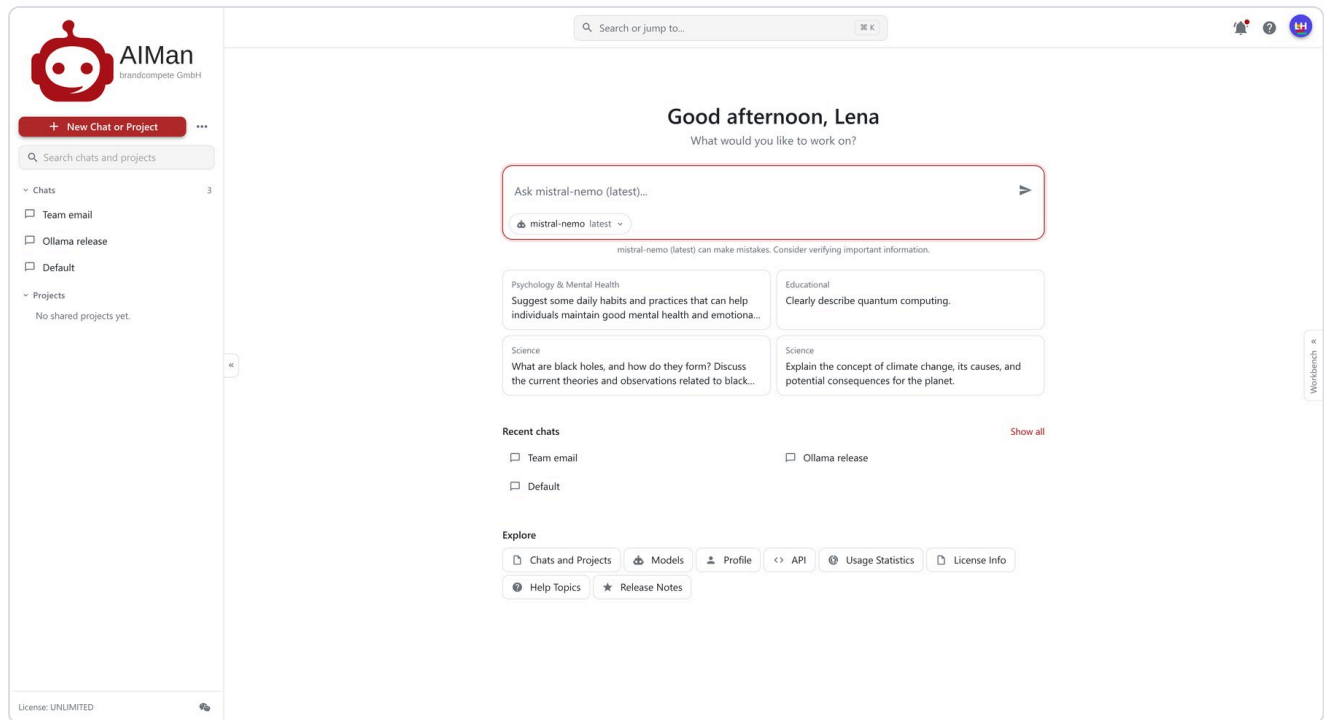
HINWEIS

Bei der ersten Anmeldung per Single Sign-on legt AIMan Ihr Konto mit der Rolle **Guest** an. Als Gast können Sie AIMan nur eingeschränkt nutzen; bitten Sie Ihre Administration um die Rolle, die Sie für Ihre Arbeit brauchen. **New to AIMan? Request access** auf der Anmeldeseite führt zu einem Kontaktformular.

WICHTIG

Passwort vergessen? Bitten Sie Ihre Administration, ein neues zu setzen. Nach der Anmeldung können Sie Ihr Passwort in Ihrem Profil selbst ändern (Kapitel 11).

Sich zurechtfinden



Die Startseite: links die Seitenleiste, in der Mitte das Eingabefeld, am rechten Rand der Griff der Workbench

- **Seitenleiste** — oben **New Chat or Project**, darunter ein Suchfeld (**Search chats and projects**) und die Abschnitte **Pinned**, **Chats** und **Projects**. Ein Klick auf eine Abschnittsüberschrift klappt sie zu. Jede Zeile hat ein ...-Menü: Chat oben anheften, sehen, mit wem er geteilt ist, oder löschen. Ein pulsierender Punkt zeigt, dass noch eine Antwort geschrieben wird. Der Griff « am Rand der Seitenleiste oder **Strg+Umschalt+S** verkleinert sie zu einer schmalen Symbolleiste.
- **Kopfzeile** — in der Mitte die Suche **Search or jump to...** (**Strg+K**, auf dem Mac **⌘K**): Sie findet jeden Chat, jedes Projekt, jeden gespeicherten Prompt, jedes Modell und jede Seite und führt Aktionen aus, etwa „neuer Chat“ oder „Fokusmodus“. Rechts: Benachrichtigungen (die Glocke), das Hilfemenü **?** (**Help Topics**, **Keyboard Shortcuts**, **Start Feature Tour**, **YouTube Tutorial Videos**, **Release Notes**, **About**) und Ihr Kontomenü (**Profile**, **Bookmarks**, **Plugins**, der **AlMan Assistant**, für Admins das **Access and Control Center**, der Hell/Dunkel-Schalter und **Sign Out**).
- **Workbench** — der Griff **Workbench** am rechten Rand öffnet die Leiste mit Ihren Prompts, Datenquellen und eigenen Anweisungen (Kapitel 8).

Die Startseite

Nach der Anmeldung landen Sie auf der Startseite. Sie begrüßt Sie mit Namen und fragt **What would you like to work on?** Unter dem Eingabefeld finden Sie vier **Prompt-Ideen** zum Ausprobieren (bei jedem Besuch neue), Ihre **Recent chats** mit **Show all** und eine Zeile **Explore** mit Abkürzungen zu **Chats and Projects**, **Models**, **Profile**, **API**, **Help Topics**, **Release Notes** und mehr. Eine andere Einstiegsseite wählen Sie in den Startoptionen (Kapitel 11).

NOCH NICHT VERFÜGBAR

Die Kachel **Usage Statistics** öffnet eine Seite, die noch entsteht: AlMan zeigt bisher keine Nutzungs- oder Kostenzahlen. Die Hilfeseite **API** im Produkt ist noch leer; die API beschreibt Kapitel 13 dieses Handbuchs.

Ihr erster Prompt

- 1 Tippen Sie Ihre Frage in das Feld auf der Startseite. Der Platzhalter verrät, welches Modell antworten wird, zum Beispiel **Ask mistral-nemo (latest)**...
- 2 Für ein anderes Modell klicken Sie auf den Modell-Chip unter dem Text (**mistral-nemo latest**) und wählen eines aus (Kapitel 3). Voreingestellt ist mistral-nemo, ein offenes Modell auf den GPU-Servern.
- 3 Drücken Sie **Enter** oder klicken Sie auf den Pfeil. AIman legt einen neuen Chat an, benennt ihn nach Ihrem Prompt, und die Antwort erscheint Wort für Wort.
- 4 Stellen Sie Rückfragen im Eingabefeld unten im Chat. Das Modell kennt den bisherigen Verlauf.

Unter jedem Eingabefeld erinnert AIman daran, dass das Modell *Fehler machen kann*. Prüfen Sie alles Wichtige – die Quellenangaben unter einer Antwort (Kapitel 6 und 7) machen das leicht.

TIPP

Jedes Modell kennt das heutige Datum und ist angewiesen, es offen zu sagen, wenn eine Frage über sein Wissen hinausgeht, statt zu raten. Für aktuelle Ereignisse schalten Sie die Websuche ein (Kapitel 6).

3 Modelle

Zwei Arten von Modellen in einer Auswahl: offene Modelle auf dedizierten GPU-Servern und die gehosteten Modelle der Anbieter mit dem Schlüssel Ihres Unternehmens.

Lokale Modelle und Anbietermodelle

Lokale Modelle (Local models) sind offene Modelle, die auf den GPU-Servern Ihrer AIMan-Installation laufen – den dedizierten Servern von brandcompete in Deutschland oder den eigenen Ihres Unternehmens. Prompts an sie verlassen diese Server nicht, und es fallen keine Gebühren pro Token an. Jedes lokale Modell verarbeitet einen Kontext von 32.000 Tokens – genug für lange Unterhaltungen und mehrere Seiten Dokumentpassagen.

Anbietermodelle (Vendor models) sind die gehosteten Modelle von Anthropic, OpenAI, Azure OpenAI, Google, Cohere und Ollama Cloud. Für schwierige Aufgaben sind sie oft die stärksten. Ihr Prompt und die nötigen Passagen gehen an den Anbieter, mit dem Schlüssel und unter dem Vertrag Ihres Unternehmens.

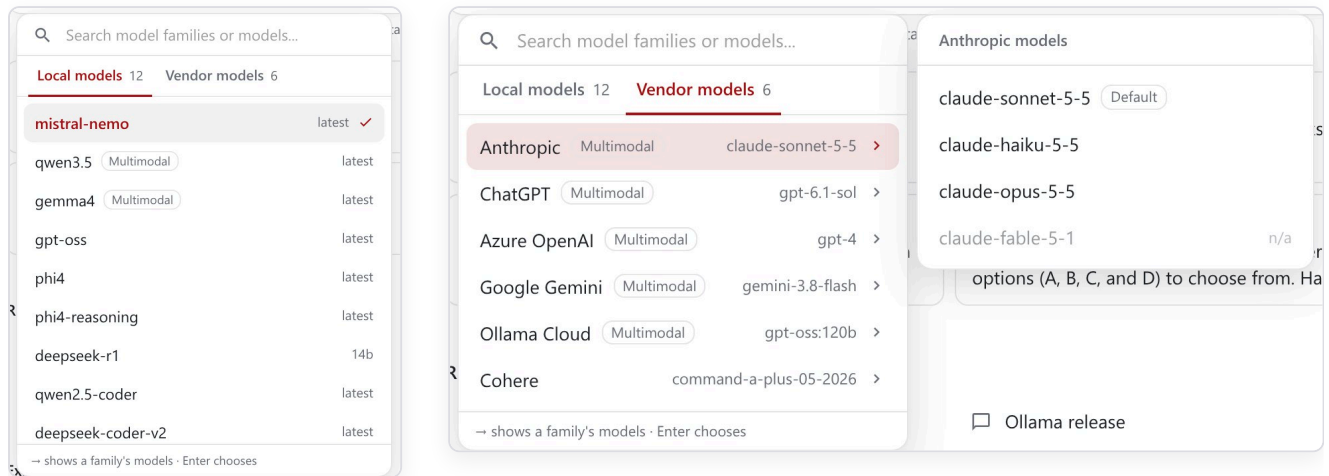
Welche Familien und Versionen Sie sehen, entscheidet Ihre Administration. Die Tabellen zeigen den Katalog mit Stand Oktober 2026.

Lokales Modell	Stark in
mistral-nemo	Der Standard für den Alltag. Das 12B-Modell von Mistral AI und NVIDIA: schnell, gut auf Deutsch, kann Werkzeuge nutzen (etwa die Websuche).
qwen3.5	Qwen 3.5 von Alibaba: versteht Bilder, nutzt Werkzeuge, kann Schritt für Schritt nachdenken.
gemma4	Gemma 4 von Google DeepMind: schnell, stark für seine Größe, mehrsprachig, versteht Bilder.
gpt-oss	Das Open-Weight-Reasoning-Modell von OpenAI (20B), mit Werkzeugaufrufen.
phi4	Phi-4 von Microsoft (14B): stark in Mathematik und Logik.
phi4-reasoning	Phi-4, auf schrittweises Schlussfolgern trainiert.
deepseek-r1	Das offene Reasoning-Modell von DeepSeek (14B-Destillat).
qwen2.5-coder	Code: schreiben, erklären und prüfen.
deepseek-coder-v2	Code, das Mixture-of-Experts-Code-Modell von DeepSeek.
minicpm-v	Bildverständnis: Bilder beschreiben und lesen.
llama3.1	Llama 3.1 von Meta (8B), ein kompaktes Allzweckmodell.

Anbieterfamilie	Standardversion	Weitere Versionen
Anthropic	claude-sonnet-5-5	claude-opus-5-5, claude-haiku-5-5
ChatGPT (OpenAI)	gpt-6.1-sol	gpt-6-sol, gpt-6-astra, gpt-6-luna, gpt-5.4-mini
Azure OpenAI	gpt-4	gpt-4o, gpt-4o-mini
Google Gemini	gemini-3.8-flash	gemini-3.1-pro-preview, gemini-3.5-flash-lite
Ollama Cloud	gpt-oss:120b	nemotron-3-super, gemma4:31b, nemotron-3-nano:30b
Cohere	command-a-plus-05-2026	command-a-reasoning-08-2025, command-a-vision-07-2025, command-r7b-12-2024

Ein Modell in der Modellauswahl wählen

Klicken Sie auf den Modell-Chip im Eingabefeld – oder auf das Feld **Model** im Dialog **New Chat or Project** –, um die Auswahl zu öffnen.



Der Tab **Local models**

Der Tab **Vendor models** mit den Versionen der Familie Anthropic

- Zwei Tabs, **Local models** und **Vendor models**, jeweils mit Anzahl. Die Auswahl öffnet sich auf dem Tab des Modells, das Sie gerade nutzen – ein Chat mit Claude öffnet also den Anbieter-Tab.
- Tippen Sie in **Search model families or models...**, um beide Tabs zugleich zu durchsuchen.
- Eine Familie mit mehreren Versionen hat einen Pfeil. Ein Klick darauf oder **→** zeigt rechts ihre Versionen; die Standardversion der Familie ist mit **Default** markiert. Eine Version mit **n/a** ist nicht bereitgestellt und nicht wählbar. Eine Familie mit nur einer Version (die meisten lokalen Modelle) wählen Sie mit einem Klick.
- Das Abzeichen **Multimodal** kennzeichnet Familien, die Bilder verstehen.
- Tastatur: Pfeiltasten zum Bewegen, **Enter** zum Wählen, **←** oder **Esc** zurück.

Das gewählte Modell gilt nur für den aktuellen Tab. Zum Vergleichen geben Sie jedem Tab sein eigenes (Kapitel 4). Die Hilfeseite **Available Models** (Kachel **Models** auf der Startseite) listet jedes Modell mit einer Kurzbeschreibung und dem Hinweis, ob es bereitgestellt ist.

Welches Modell für welche Aufgabe

Aufgabe	Gute Wahl
Alltägliches Schreiben, E-Mails, Zusammenfassungen, deutsche Texte	mistral-nemo (Standard); gemma4 für andere Sprachen
Vertrauliches Material	Jedes lokale Modell: Der Prompt bleibt auf den GPU-Servern
Schwierige Analysen, langes Schlussfolgern, beste Qualität	Claude (claude-opus-5-5, claude-sonnet-5-5), GPT, Gemini; lokal qwen3.5 oder gpt-oss mit eingeschaltetem Thinking
Mathematik und Logikrätsel	phi4, phi4-reasoning, deepseek-r1
Programmieren	qwen2.5-coder, deepseek-coder-v2; Claude und GPT für größere Aufgaben
Bilder: beschreiben, ein Diagramm oder Foto lesen	qwen3.5, gemma4, minicpm-v; die Anbieterfamilien mit Multimodal
Aktuelles (mit Websuche)	Claude, Gemini und GPT nutzen ihre eigene Suche; lokal qwen3.5 oder mistral-nemo
Antworten aus Ihren Dokumenten	lokal mistral-nemo oder qwen3.5; jedes Anbietermodell, sofern Ihr Unternehmen erlaubt, dass Passagen dorthin gehen

Bilder

Modelle mit **Multimodal** können Bilder ansehen. Hängen Sie ein Bild mit der Büroklammer im Eingabefeld an (JPEG, PNG, GIF, WebP oder SVG) und fragen Sie danach: beschreiben lassen, den Text in einem Screenshot lesen, zwei Diagramme vergleichen. Bilder erzeugt AIman nicht.

4 Chats, Projekte und Tabs

Ein Chat gehört Ihnen allein; ein Projekt ist ein Chat, den Sie teilen. Beide können mehrere Tabs enthalten, jeder mit seinem eigenen Modell.

Einen Chat oder ein Projekt anlegen

Am schnellsten tippen Sie auf der Startseite: Der erste Prompt legt einen Chat an. Mehr Kontrolle haben Sie mit **New Chat or Project** in der Seitenleiste (**Strg+Umschalt+P**):

- 1 Wählen Sie unter **Model** das Modell für den ersten Tab.
- 2 Geben Sie einen **Chat / Project Name** ein (Pflichtfeld).
- 3 Für einen geteilten Chat haken Sie **This is a Project (Shared Chat)** an und wählen unter **Share With...**, mit wem Sie ihn teilen – Kolleginnen und Kollegen, Teams, Organisationseinheiten oder das Unternehmen.
- 4 Optional öffnen Sie **Advanced Settings**: eine **Description**, eine **Data Source** für den Chat und Ihre **Custom Instructions**.
- 5 Klicken Sie auf **Create**.

Der Dialog **New Chat or Project**

Über einem geöffneten Chat steht sein Name als **Chat: ...** oder **Project: ...**. Mit den Symbolen oben rechts bearbeiten Sie ihn (Name, Beschreibung, Freigabe), öffnen **Prompt All Tabs** (siehe unten) und löschen ihn. Löschen lässt sich nicht rückgängig machen. Die Seite **Chats and Projects** (Startseite, **Explore**) zeigt alle Ihre Chats und Projekte als Karten.

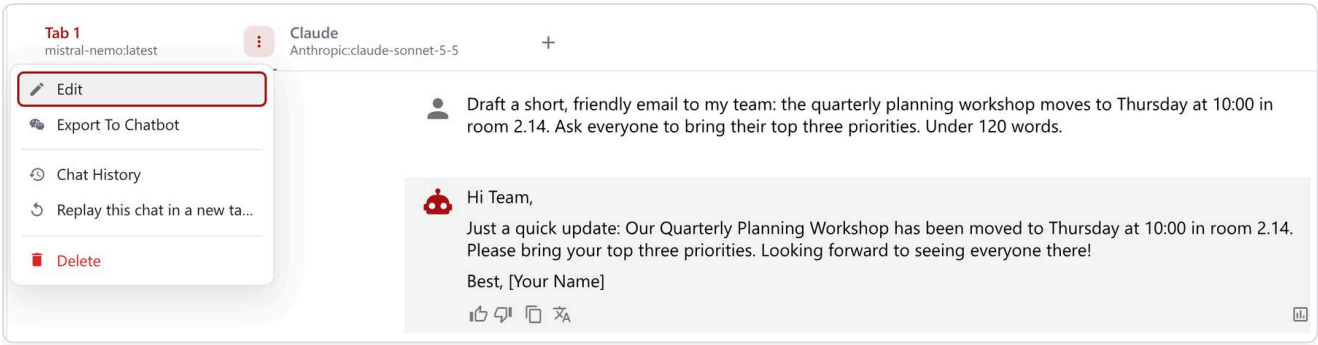
Mehrere Modelle nebeneinander: Tabs

Ein Chat mit zwei Tabs: „Tab 1“ mit mistral-nemo und „Claude“ mit claude-sonnet-5-5 beantworten denselben Prompt

Jeder Tab zeigt seinen Namen und darunter sein Modell (etwa **Anthropic:claude-sonnet-5-5**). Jeder Tab hat seine eigene Unterhaltung, sein Modell, seine Datenquelle und seine Einstellungen. Einen Tab fügen Sie mit **+** neben den Tabs hinzu (**Add New Tab**, **Strg+Umschalt+,**) und wählen sein Modell.

Prompt All Tabs schickt einen Prompt an alle Tabs zugleich. Klicken Sie auf das Symbol oben rechts im Chat, tippen Sie den Prompt in das Feld, das sich öffnet, und senden Sie ihn ab: Jeder Tab antwortet mit seinem eigenen Modell. So sehen Sie am schnellsten, wie Claude, GPT und ein lokales Modell dieselbe Aufgabe lösen.

Das Tab-Menü



Eintrag	Was er tut
Edit	Tab umbenennen und Modell, Datenquelle und eigene Anweisungen ändern. Ein Klick auf den Modell-Chip im Eingabefeld öffnet denselben Dialog.
Export To Chatbot	Macht aus dem Tab einen einbettbaren Chatbot (Kapitel 9).
Share Chat	Erzeugt einen Link auf eine schreibgeschützte Kopie der Unterhaltung. Nur sichtbar, wo Link-Freigaben für Ihre Installation eingeschaltet sind (siehe „Freigabe-Links, Export und Druck“ unten).
Chat History	Listet die in diesem Tab gesendeten Prompts zur Wiederverwendung: ins Eingabefeld übernehmen oder mit Send directly to chat gleich senden.
Replay this chat in a new tab with another model	Schickt alle Prompts der Unterhaltung in einem neuen Tab noch einmal an ein Modell Ihrer Wahl – so sehen Sie am einfachsten, wie ein anderes Modell geantwortet hätte.
Delete	Löscht den Tab mitsamt Unterhaltung.

Mit Antworten arbeiten

Antworten erscheinen, während sie geschrieben werden – mit Formatierung, Tabellen und Syntaxhervorhebung; Codeblöcke haben eine eigene Kopierschaltfläche. Unter jeder Antwort:

Element	Was es tut
Daumen hoch / runter	Good Response / Bad Response : bewertet die Antwort. Die Bewertung wird mit der Antwort gespeichert.
Kopieren	Kopiert die Antwort in die Zwischenablage (Copy Response).
Übersetzen	Übersetzt die Antwort mit DeepL in eine andere Sprache.
Statistik (rechts)	Stats For Nerds : Tokens pro Sekunde und Antwortzeit dieser Antwort.
Sources	Mit Websuche oder Datenquelle: die nummerierten Quellen, auf die sich die Antwort stützt (Kapitel 6 und 7).

- **Antworten laufen weiter.** Während ein Modell schreibt, können Sie in einen anderen Chat wechseln, einen Tab hinzufügen oder einen neuen Chat beginnen; die Antwort wird im Hintergrund fertig geschrieben und gespeichert. Ein Fortschrittsbalken am Tab und ein Punkt in der Seitenleiste zeigen, dass sie noch läuft. Ein Neuladen der Seite kostet nur die Live-Ansicht, nicht die Antwort.
- **Stopp** – während ein Modell schreibt, wird die Senden-Schaltfläche zu einem roten Quadrat. Ein Klick darauf bricht die Antwort ab.
- **Lange Unterhaltungen** – wird eine Unterhaltung für das Modell zu lang, ersetzt AIMan die ältesten Beiträge durch eine kurze Zusammenfassung, statt sie stillschweigend wegzulassen. So weiß das Modell weiterhin, was zu Beginn vereinbart wurde.

Projekte: einen Chat mit Kollegen teilen

Ein Projekt teilen Sie beim Anlegen oder Bearbeiten unter **Share With...** mit Personen, Teams, Organisationseinheiten oder dem Unternehmen. Jede Freigabe hat eine Stufe:

- **Lesen** (Read) – das Projekt öffnen und die Unterhaltung verfolgen;
- **Schreiben** (Write) – zusätzlich Prompts senden und in den Tabs arbeiten;
- **Verwalten** (Admin) – zusätzlich das Projekt und seine Freigaben ändern.

Mit Ihnen geteilte Projekte erscheinen in der Seitenleiste unter **Projects**, gekennzeichnet mit einem Symbol:

This project has been shared with you by another person. Click here to view details. Die Details zeigen, wer sonst Zugriff hat und mit welchen Rechten. Brauchen Sie ein Projekt nicht mehr, entfernt **Leave Project** es aus Ihrer Liste; die anderen behalten es. Eine Freigabe an ein Team erreicht alle Mitglieder des Teams, auch spätere.

Freigabe-Links, Export und Druck

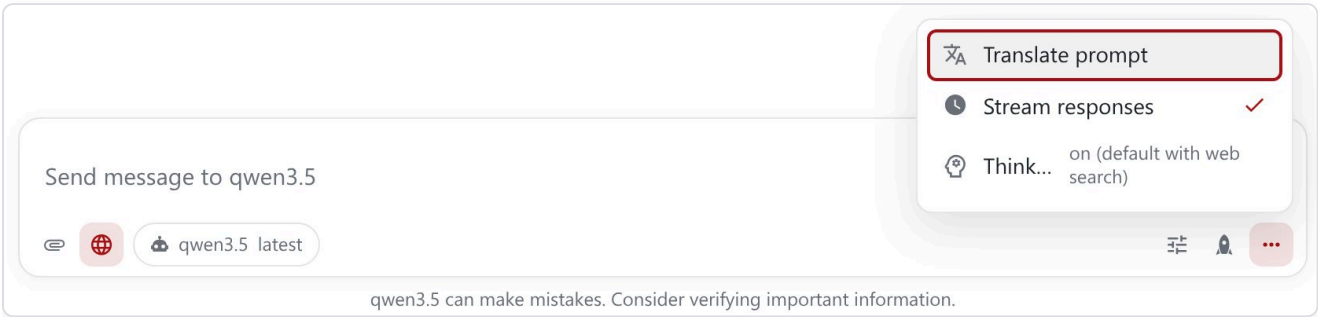
Wo Ihre Installation Link-Freigaben erlaubt, bietet das Tab-Menü **Share Chat** und die Projektkopfzeile **Share Project**. **Copy URL** kopiert einen Link auf eine schreibgeschützte Seite der Unterhaltung. Dort laden Sie die Unterhaltung als Text, HTML, Bild (PNG) oder JSON herunter oder drucken sie mit **Print Chat**. Ohne Link-Freigabe nutzen Sie **Copy Response** für einzelne Antworten oder teilen das Projekt selbst.

Anheften, Lesezeichen und Suche

Heften Sie Chats an, die Sie täglich brauchen: ... in der Zeile → anheften. Angeheftete Chats bleiben oben in der Seitenleiste und in der schmalen Symbolleiste. **Bookmarks** (Kontomenü) speichert Links auf beliebige Seiten in AIMan: **Bookmark This Page**. Und **Strg+K** findet jeden Chat und jedes Projekt über den Namen.

5 Das Eingabefeld

Alles, was die nächste Antwort beeinflusst, sitzt im Nachrichtenfeld unten im Chat.



Das Eingabefeld mit eingeschalteter Websuche und geöffnetem ...-Menü

Element	Was es tut
Büroklammer	Hängt Dateien an den Prompt an: Bilder für multimodale Modelle oder ein Dokument, das das Modell für diese eine Frage lesen soll.
Globus	Websuche für diesen Tab ein oder aus (Kapitel 6). Rot hinterlegt, wenn eingeschaltet.
Modell-Chip	Zeigt das Modell des Tabs, etwa qwen3.5 latest . Ein Klick ändert Modell oder Datenquelle.
Schieberegler-Symbol	Prompt Options: Temperatur und Antwortlänge (siehe unten).
Raketen-Symbol	Öffnet und schließt die Workbench (Kapitel 8).
...	Weitere Optionen: Translate prompt , Stream responses , Thinking und mit Datenquelle Amount of Chunks und Only from the documents . Ein kleiner Punkt an der Schaltfläche zeigt, dass etwas vom Standard abweicht.
Pfeil / rotes Quadrat	Sendet den Prompt (Enter); während eine Antwort geschrieben wird, stoppt es sie.

Der Platzhalter im Feld nennt das Modell, das antworten wird: **Send message to qwen3.5**. **Umschalt+Esc** setzt den Cursor von überall im Chat ins Eingabefeld.

Dateien und Bilder anhängen

Klicken Sie auf die Büroklammer, wählen Sie Dateien und senden Sie Ihren Prompt. Ein Bild wird direkt an das Modell übergeben; wählen Sie also ein Modell mit **Multimodal**. Ein so angehängtes Dokument ist für die aktuelle Frage gedacht. Für Dokumente, zu denen Sie immer wieder fragen oder die Sie mit Kollegen teilen wollen, legen Sie besser eine Datenquelle an (Kapitel 7): Sie wird einmal indexiert und bei jedem Prompt durchsucht.

Prompt-Optionen

Das Schieberegler-Symbol öffnet die **Prompt Options** für den aktuellen Chat (**Fine-tune your prompt options to get the best responses.**):

- **temperature** – niedrige Werte machen Antworten fokussierter und wiederholbarer, hohe vielfältiger und kreativer;
- **num predict** – die maximale Länge der Antwort in Tokens.

Lassen Sie ein Feld leer, gilt der Standardwert des Modells; das empfiehlt sich für die meisten Chats, denn jedes Modell ist anders abgestimmt. Gesendet wird nur, was Sie ausfüllen.

Defaults leert die Felder, **Apply** speichert sie. Einige aktuelle Anbietermodelle akzeptieren keine Temperatur mehr; AIman lässt sie für diese weg.

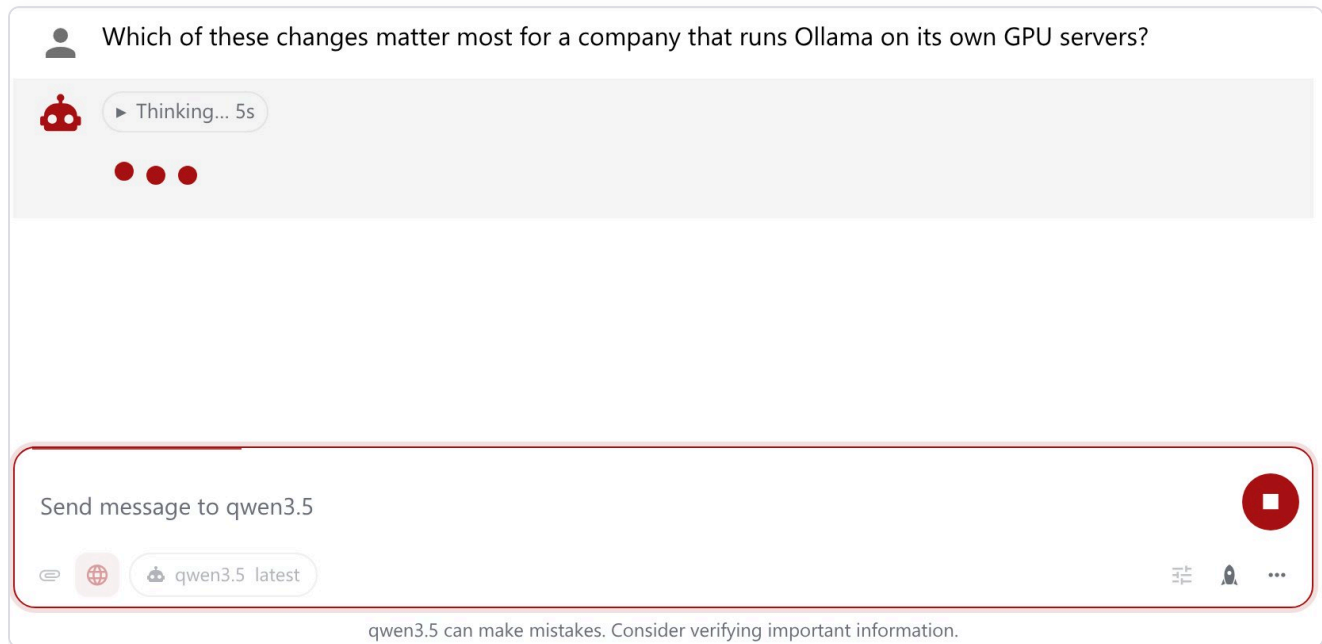
Der Dialog **Prompt Options**

Das Menü „...“

Eintrag	Was er tut
Translate prompt	Übersetzt Ihren Prompt vor dem Senden mit DeepL – nützlich, weil viele Modelle auf Englisch am besten antworten. Sie wählen die Zielsprache; die Wahl gilt für die aktuelle Sitzung, sofern Sie sie nicht speichern.
Stream responses	Ein (Häkchen): Die Antwort erscheint Wort für Wort. Aus: Sie erscheint vollständig, wenn sie fertig ist.
Thinking	Nur bei Modellen, die nachdenken können. Der Hinweis zeigt den aktuellen Zustand, etwa off - faster (default) oder on (default with web search) . Ein Klick schaltet um.
Amount of Chunks	Mit Datenquelle: wie viele Passagen pro Frage nachgeschlagen werden, 1 bis 10, Standard 5 (Kapitel 7).
Only from the documents	Mit Datenquelle: angehakt (Standard) nur aus den Dokumenten antworten; ohne Haken darf das Modell Allgemeinwissen ergänzen, als solches gekennzeichnet (Kapitel 7).

Websuche, Thinking und **Only from the documents** merkt sich AIman pro Tab, auch nach dem Neuladen.

Thinking: das Modell erst nachdenken lassen



Ein Modell denkt nach, bevor es antwortet; das rote Quadrat stoppt es

Reasoning-Modelle können ein Problem Schritt für Schritt durchdenken, bevor sie antworten. Das dauert länger, hilft aber bei Rechnungen, Logik, Planung und Fragen mit mehreren Bedingungen. Bei eingeschaltetem **Thinking** erscheinen die Überlegungen des Modells live in einem zugeklappten Block mit Zeitanzeige (**Thinking... 5s**); die Antwort folgt darunter. Ein Klick auf den Block zeigt die Überlegungen. Sie sind sichtbar, solange Sie auf der Seite bleiben, werden aber nicht mit der Antwort gespeichert.

- Thinking ist **standardmäßig aus**, weil Antworten dann schneller kommen.
- Mit eingeschalteter Websuche ist es bei lokalen Modellen **standardmäßig an**: Mehrere Webseiten abzuwägen profitiert vom Nachdenken.
- Den Eintrag **Thinking** gibt es nur bei Modellen, die nachdenken können, etwa qwen3.5 und gemma4; mistral-nemo, phi4 und llama3.1 haben keinen.
- Denkt ein Modell nach, liefert aber keine Antwort, fragt AIman automatisch noch einmal.

6 Websuche

Modelle wissen nur, was sie bis zu ihrem Trainingsstand gelernt haben. Die Websuche stützt eine Antwort auf heutige Seiten und listet die Quellen auf.

Schalten Sie die Websuche mit dem **Globus** im Eingabefeld ein (**Web search: use current web results for the answers**). Sie bleibt für diesen Tab an, bis Sie sie ausschalten. Nutzen Sie sie für Nachrichten, Preise, neue Versionen, Öffnungszeiten – für alles, was sich ändert.

What is the latest Ollama release and what are its most important changes? Three bullet points, with sources.

Based on the latest information available, here are the details about the most recent Ollama release:

Latest Release: Ollama v0.34.3 [5]

Release Date: September 19, 2026 [5]

Most Important Changes/Features:

- **Enhanced Apple Silicon Support** – Preview support for Apple silicon was announced in March 2026, optimizing performance on Mac hardware using the MLX framework [5].
- **Integrated Cloud & Local Model Features** – Continued expansion of cloud model hosting alongside local inference capabilities, with improved tool and coding-agent integrations [5].
- **Expanded Ecosystem Integrations** – Added support for integration with AI coding tools such as Claude Code, Codex, OpenCode, Copilot CLI, and OpenClaw [5].

Sources

- [5] [Ollama - Wikipedia](#) en.wikipedia.org **excerpt**

Eine Antwort aus dem Web: Aussagen mit [5] nummeriert, darunter die Quelle mit Website und dem Link **excerpt**

Was beim Fragen passiert

AIMan sucht im Web nach Ihrer Frage – Rückfragen wie „und in Deutschland?“ berücksichtigen die vorherigen Nachrichten –, liest die besten Seiten und übergibt dem Modell die Passagen, die die Frage beantworten. Währenddessen zeigt der Chat **Searching the web and reading the results...**; das dauert einige Sekunden, bevor die Antwort beginnt. Wie gesucht wird, hängt vom Modell ab:

Modell	So wird gesucht
Claude, Gemini, GPT (OpenAI)	Mit der eigenen Websuche des Anbieters, durchgeführt beim Anbieter.
Lokale Modelle, die Werkzeuge nutzen können (etwa mistral-nemo, qwen3.5, gemma4)	Das Modell entscheidet selbst, ob es sucht, wonach und welche Seiten es liest. Eine Frage wie „Was ist 12 × 12?“ beantwortet es ohne Suche.
Alle anderen Modelle (etwa phi4, Azure OpenAI, Cohere, Ollama Cloud)	Mit der Suche von AIMan: Sie formuliert eine oder zwei Suchanfragen, liest die besten Seiten und bewertet deren Passagen, bevor das Modell antwortet.

Quellen

Die Antwort zitiert Webseiten wie Dokumente, mit Nummern wie [3]. Unter **Sources** verlinkt jede Nummer auf ihre Seite, mit dem Namen der Website; **excerpt** zeigt die verwendete Passage. Die Quellen werden mit der Antwort gespeichert. Die Websuche funktioniert auch zusammen mit einer Datenquelle: Die Antwort kann dann beides nutzen, und Dokumente und Seiten werden ohne Überschneidung nummeriert.

Gut zu wissen

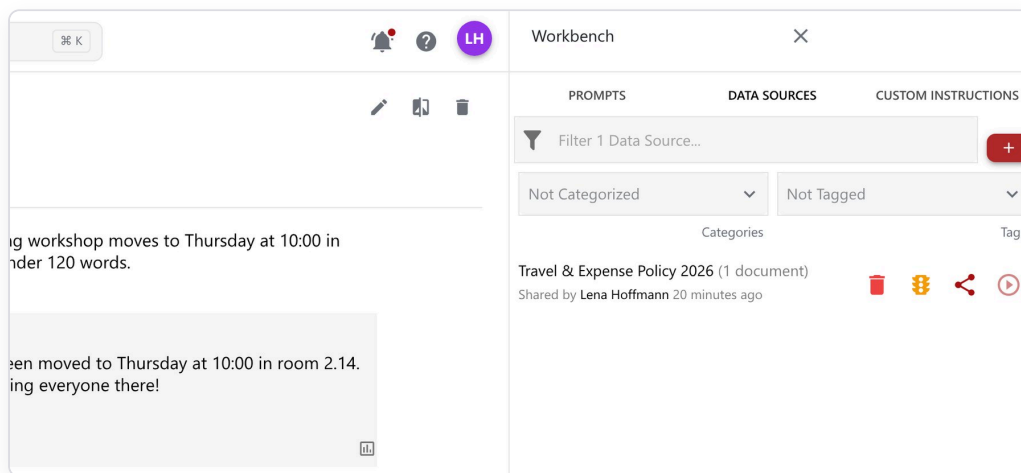
- **Datenschutz** – Ihre Frage oder eine daraus abgeleitete kurze Suchanfrage geht an einen Suchdienst: bei Claude, Gemini und GPT an die Suche des Anbieters, sonst über den Suchdienst von AIMan an öffentliche Suchmaschinen oder an eine kostenpflichtige Such-API, die Ihr Unternehmen eingerichtet hat. Schalten Sie die Websuche nicht für Fragen mit vertraulichen Details ein.
- **Webseiten können falsch oder veraltet sein.** Prüfen Sie bei Wichtigem die verlinkten Quellen.
- **Webinhalte sind Information, niemals Anweisung.** Eine Seite kann dem Modell nichts vorschreiben; AIMan liest nur öffentliche Adressen.
- Findet die Suche nichts Brauchbares, sagt das Modell das und antwortet ohne aktuelle Ergebnisse.
- Die Administration kann die Websuche für die ganze Installation abschalten; dann verschwindet der Globus. In Websites eingebettete Chatbots bieten keine Websuche.

7 Datenquellen: Antworten aus Ihren Dokumenten

Hängen Sie eigene Dokumente, Webseiten oder eine Wissensdatenbank Ihres Unternehmens an einen Chat – jede Antwort entsteht dann aus den passendsten Passagen, mit Dokument und Seite als Quelle.

Ein Modell kennt normalerweise nur, womit es trainiert wurde. Ist eine **Datenquelle** angehängt, durchsucht AIman sie zuerst nach den Passagen, die am besten zu Ihrer Frage passen, übergibt sie dem Modell zusammen mit der Frage und weist es an, sie zu zitieren. Sie müssen nichts weiter tun: Datenquelle wählen und wie gewohnt chatten.

Datenquellen liegen in der **Workbench**, Tab **DATA SOURCES**. Die Liste zeigt jede Datenquelle mit der Anzahl ihrer Dokumente und wer sie geteilt hat, dazu vier Symbole: Löschen, Status, Freigabe-Info und **Use Data Source**.

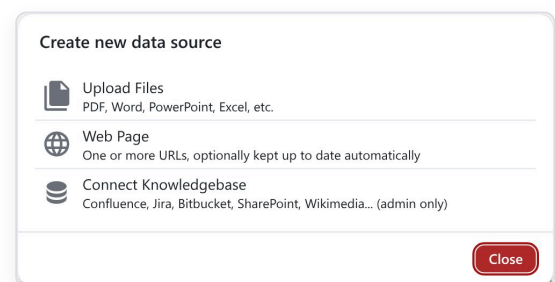


Eine Datenquelle in der Workbench: Löschen, Status (gelb: wird indexiert), Freigabe-Info und **Use Data Source**. Links die Symbole der Chat-Kopfzeile: Bearbeiten, **Prompt All Tabs**, Löschen

Eine Datenquelle anlegen

Klicken Sie im Datenquellen-Tab der Workbench auf **+**. **Create new data source** bietet drei Arten:

- **Upload Files** – PDF, Word, PowerPoint, Excel usw.
- **Web Page** – eine oder mehrere URLs, auf Wunsch automatisch aktuell gehalten.
- **Connect Knowledgebase** – Confluence, Jira oder Bitbucket; nur für Admins.



Die Art der Datenquelle wählen

Dateien hochladen

- 1 Geben Sie einen **Name** (Pflichtfeld) und unter **Summary** kurz an, was die Dokumente enthalten.
- 2 Ziehen Sie Dateien auf die rote Fläche oder klicken Sie auf **Browse**. Jede Datei erscheint als Karte; das x entfernt sie wieder.
- 3 Wählen Sie optional, mit wem Sie die Datenquelle teilen (**Share Data Source With...**), und vergeben Sie **TAGS** und **CATEGORIES**, um sie wiederzufinden.
- 4 Klicken Sie auf **Create**. AIMan liest und indexiert die Dokumente; der Status wechselt von **Pending** und **Indexing** zu **Ready**, und eine Benachrichtigung meldet, wenn alles fertig ist.

Create Data Source

Name

Travel & Expense Policy 2026

Summary

Northwind travel and expense rules: approval, rail, flights, hotels, per diems, claims.

Browse you computer or drag and drop document files here...

PDF

Northwind-Travel-Pol...

52.4 KB

Advanced Settings

Share Data Source With...

Share Data Source With...

TAGS

Tags...

CATEGORIES

Categories...

Close

Create

Eine Datenquelle mit einem PDF, bereit zum Anlegen

Format	Dateitypen
PDF	.pdf (mit Textebene)
Word	.doc, .docx
PowerPoint	.ppt, .pptx
Excel und CSV	.xls, .xlsx, .csv
Text und Markdown	.txt, .md
Webseite	.html

Tabellen bleiben ganz: Eine Tabelle – aus einer Tabellenkalkulation oder in einem anderen Dokument – wird als eine Passage zusammengehalten, ihre Kopfzeile also nie von den Daten getrennt.

Erweiterte Einstellungen

Unter **Advanced Settings** finden Sie **Chunk Size** und **Chunk Overlap** (wie lang die Passagen sind und wie stark sie sich überlappen) sowie **Use Contextual Retrieval Augmented Generation**. Bei dieser kontextuellen Indexierung schreibt ein Modell zu jeder Passage zunächst eine kurze Notiz, die sie im Gesamtdokument verortet („dies stammt aus dem Preisteil des Q3-Berichts ...“). Fragen finden dann in langen, dichten oder mehrdeutigen Dokumenten wie Verträgen, Berichten und Handbüchern zuverlässiger die richtigen Passagen. Das Indexieren dauert länger. Für kurze, in sich geschlossene Dokumente lassen Sie es aus.

Webseiten

Connect Web Page erwartet einen **Name** und eine oder mehrere URLs: jede in **Add New URI Data Source...** eintippen und auf **+** klicken. AIman ruft jede Seite ab, wandelt sie in lesbaren Text um und indexiert sie wie ein Dokument; Quellenangaben verlinken auf die Seite.

Schalten Sie **Keep this up to date automatically** ein und stellen Sie **Recrawl Every (Days)** ein (mindestens ein Tag), damit die Seiten regelmäßig neu gelesen werden. Die Datenquelle bleibt während der Aktualisierung nutzbar.

Der Dialog **Connect Web Page**

Wissensdatenbanken: Confluence, Jira, Bitbucket

Eine Wissensdatenbank verbindet AIman direkt mit einem System, das Ihr Unternehmen ohnehin betreibt, statt Exporte hochzuladen. Admins richten sie einmal mit einem Zugriffstoken ein; danach wird sie wie jede andere Datenquelle geteilt – meist mit dem ganzen Unternehmen. Solange sie nicht geteilt ist, sieht sie nur der Admin, der sie angelegt hat.

System	Editionen	Was indexiert wird
Confluence	Cloud sowie Server / Data Center ab 7.9	Die Seiten eines oder mehrerer Bereiche (Space Key). Anhänge und Bilder werden nicht gelesen.
Jira	Server / Data Center ab 8.14	Die Vorgänge eines oder mehrerer Projekte (Project Key(s)).
Bitbucket	Server / Data Center ab 5.5	Ein oder mehrere Repositorys (Project/Repository Key(s) , etwa PROJ/my - repo).

Der Admin trägt die **Base URL** ein, wählt die **Authentication** (**Personal Access Token**, bei Confluence Cloud eine **Account Email** mit API-Token) und klickt auf **Connect**; der erste Abruf startet sofort. Mit **Keep this up to date automatically** werden die Inhalte im eingestellten Abstand von Tagen neu gelesen. SharePoint und Wikimedia erscheinen im Dialog, sind aber mit **Coming soon** markiert.

HINWEIS

Antworten aus einer Wissensdatenbank spiegeln den Stand des letzten Abrufs. Das gespeicherte Token wird nie wieder angezeigt; lassen Sie das Feld beim Bearbeiten leer, um es zu behalten. Jira Cloud und Bitbucket Cloud werden nicht unterstützt.

Status

Status	Symbol	Bedeutung
Pending	grau	Wartet auf die Verarbeitung.
Indexing	gelb	Die Dokumente werden gelesen, in Passagen zerlegt und indexiert.
Ready	grün	In Chats nutzbar. Erst jetzt ist Use Data Source aktiv.
Failed	rot	Die Verarbeitung ist fehlgeschlagen, etwa bei einer unlesbaren oder passwortgeschützten Datei. Zeigen Sie auf das Symbol, um den Grund zu sehen; zusätzlich erhalten Sie eine Benachrichtigung.

Fragen an Ihre Dokumente stellen

Eine Datenquelle hängen Sie auf drei Wegen an:

- **In der Workbench – Use Data Source** bei einer fertigen Datenquelle. Wählen Sie **Set as default Data Source for the current project or chat** oder **Set to individual Tab(s) from the current project or chat** und haken Sie die Tabs an. Der Projektstandard wird den Tabs angeboten; um ihn in einem Tab zu nutzen, wenden Sie ihn auf diesen Tab an.
- **Im Eingabefeld** – auf den Modell-Chip klicken und die **Data Source** wählen.
- **Beim Anlegen** eines Chats oder Projekts – **Advanced Settings** → **Data Source**.

Dann fragen Sie wie gewohnt. Die Antwort kennzeichnet, was sie aus den Dokumenten übernommen hat, mit Nummern wie [1], [2], und listet die Quellen darunter auf: Dokument und – bei PDFs und anderen Formaten mit Seiten – die Seite. Mehrere Passagen derselben Seite teilen sich eine Nummer.

Screenshot folgt: Antwort aus einer Datenquelle mit Quellenangaben

Eine Antwort auf Grundlage eines hochgeladenen Dokuments, mit nummerierten Verweisen im Text und Dokument und Seite unter
Sources

Nur aus den Dokumenten – oder auch mit Allgemeinwissen

Standardmäßig kommen Antworten **nur aus den angehängten Dokumenten**. Soll das Modell ergänzen, was es allgemein weiß – etwa einen Begriff erklären, den die Dokumente verwenden, aber nicht definieren –, entfernen Sie im ...-Menü des Eingabefelds den Haken bei **Only from the documents**. Das Modell nutzt dann zuerst die Dokumente und kennzeichnet alles andere als Allgemeinwissen (*From general knowledge: ...*). Die Einstellung gilt pro Tab.

Wenn nichts Passendes gefunden wird

Beantwortet keine der Passagen Ihre Frage wirklich, sagt AIman das, statt das Modell raten zu lassen: **No relevant content was found in the attached data source for this question**. Formulieren Sie die Frage um, hängen Sie eine andere Datenquelle an, entfernen Sie den Haken bei **Only from the documents** – oder nehmen Sie die Datenquelle heraus, wenn Sie eine allgemeine Frage stellen wollten.

So werden die richtigen Passagen gefunden

- **Hybride Suche** – jede Frage wird sowohl über Stichwörter (exakte Begriffe, Namen, Kennungen) als auch über die Bedeutung (Umschreibungen, Ihre eigenen Worte) abgeglichen, und beides wird kombiniert.
- **Reranking** – ein zweites Modell liest Ihre Frage zusammen mit jeder in Frage kommenden Passage und behält nur die besten. Wie die gesamte Dokumentverarbeitung läuft das auf den Servern Ihrer AIman-Installation.
- **Rückfragen** – „Und wer ist dafür zuständig?“ sagt für sich allein wenig. Deshalb formuliert AIman die Frage vor der Suche mithilfe der letzten Nachrichten zu einer vollständigen Frage um. Das Modell beantwortet trotzdem Ihren eigenen Wortlaut.
- **Amount of Chunks** (Eingabefeld, ...) legt fest, wie viele Passagen weitergegeben werden: 1 bis 10, Standard 5. Mehr Passagen helfen bei breiten Fragen („Fasse ... zusammen“), weniger halten Antworten fokussiert.

Datenquellen teilen, und was noch nicht geht

Eine Datenquelle teilen Sie beim Anlegen oder Bearbeiten unter **Share Data Source With...** mit Kollegen, Teams, Organisationseinheiten oder dem Unternehmen, zum Lesen, Schreiben oder Verwalten. Mit Lesezugriff können andere sie in ihren Chats nutzen. Mit Schreibzugriff können sie zusätzlich Dokumente hinzufügen oder entfernen und die Neuindexierung anstoßen, aber weder Namen noch Zusammenfassung noch Freigaben ändern (**You have read and write access to this data source...**). Löschen dürfen nur die Verwalter einer Datenquelle; dabei gehen alle Dokumente und Freigaben verloren.

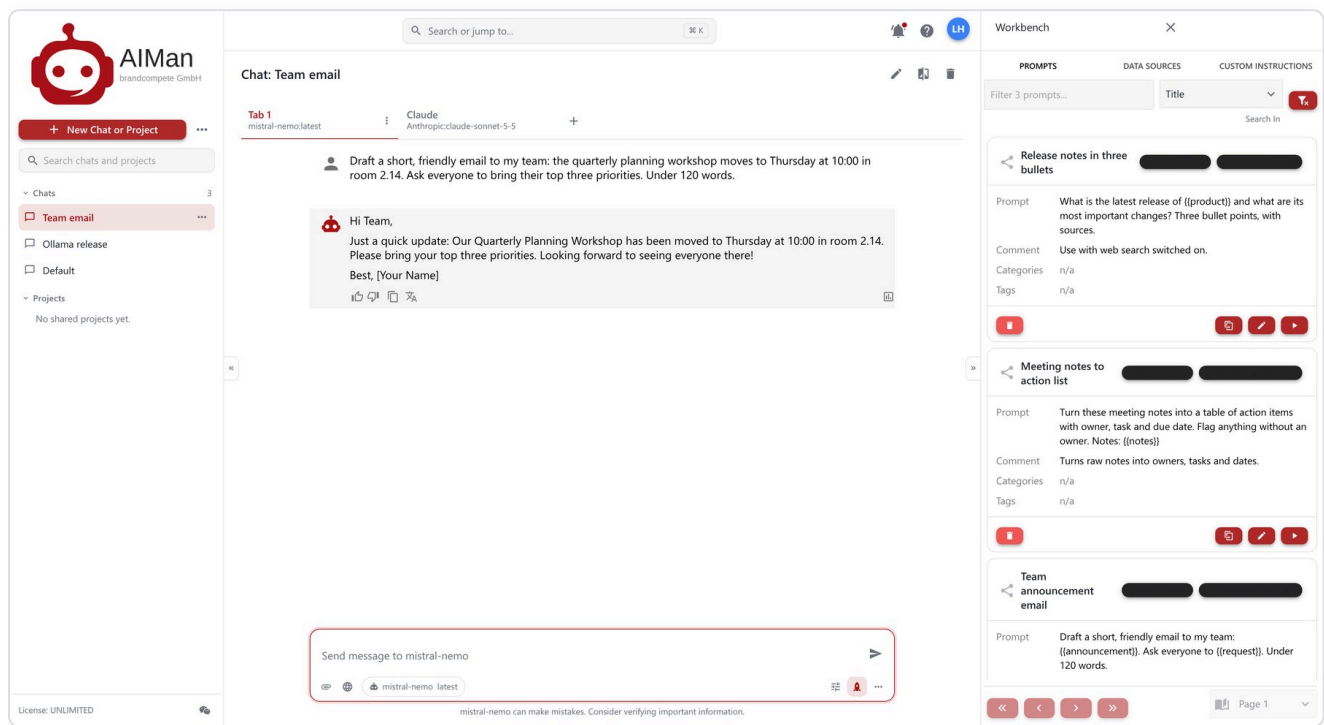
NOCH NICHT VERFÜGBAR

Gescannte Dokumente werden noch nicht gelesen. AIman liest die Textebene eines Dokuments; eine Texterkennung (OCR) gibt es noch nicht. Ein gescanntes PDF ohne Textebene oder ein Foto eines Dokuments ist nicht durchsuchbar; die Datenquelle meldet dann, dass OCR nicht unterstützt wird. Auch Bilder in Dokumenten und Confluence-Anhänge werden nicht gelesen. Texterkennung steht auf der Roadmap.

8 Workbench, Prompt-Bibliothek und eigene Anweisungen

Die Workbench hält neben jedem Chat bereit, was Sie wiederverwenden: gespeicherte Prompts, Datenquellen und feste Anweisungen.

Sie öffnen die Workbench mit dem Griff am rechten Fensterrand, mit dem Raketen-Symbol im Eingabefeld oder über **Strg+K** → „Show/hide prompt repository“. Auf breiten Bildschirmen steht sie neben dem Chat, auf kleineren schiebt sie sich darüber. Sie hat drei Tabs: **PROMPTS**, **DATA SOURCES** (Kapitel 7) und **CUSTOM INSTRUCTIONS**.



Die Prompt-Bibliothek in der Workbench, mit Platzhaltern in den Prompts

Die Prompt-Bibliothek

Ein guter Prompt ist es wert, aufgehoben zu werden. Jeder gespeicherte Prompt hat einen Titel, den Prompt-Text, einen Kommentar, Kategorien und Tags sowie seine Freigaben. Filtern Sie die Liste per Text (**Filter ... prompts...**) und wählen Sie, worin gesucht wird (Titel, Prompt, Kommentar ...). Die Schaltflächen auf jeder Karte löschen, kopieren und bearbeiten den Prompt; ► führt ihn im aktuellen Chat aus.

Einen Prompt speichern

Speichern Sie einen gesendeten Prompt oder legen Sie in der Bibliothek einen neuen an. Geben Sie ihm in **Save Prompt to Repository** einen **Title** (Pflichtfeld), optional einen **Comment** – etwa „Use with web search switched on.“ –, Kategorien, Tags und Freigaben (**Share With...**).

Platzhalter

Schreiben Sie die wechselnden Teile als Platzhalter in doppelte geschweifte Klammern: `What is the latest release of {{product}}...`. AIman listet sie unter **PROMPT PLACEHOLDERS** auf. Wenn Sie den Prompt ausführen, füllen Sie jeden Platzhalter vor dem Senden aus. Ein Prompt dient so für jedes Produkt, jeden Kunden, jedes Dokument.

Einen Prompt mit dem Platzhalter `{{product}}` speichern

TIPP

Teilen Sie Ihre besten Prompts mit Ihrem Team. Die Kollegen starten dann mit dem, was schon funktioniert, statt es neu zu erfinden – und alle Prompts werden gemeinsam besser.

Eigene Anweisungen

Eigene Anweisungen (**Custom Instructions**) sind fester Kontext, den das Modell mit jedem Prompt erhält, damit Sie ihn nicht wiederholen müssen. Es gibt zwei Arten (**Instruction Type**):

- **What would you like AIman to know about you?** – Ihre Rolle, Ihre Abteilung, Ihre Kunden, Begriffe, die Sie verwenden;
- **How would you like AIman to respond?** – Ton, Länge, Format, Sprache, etwa „Antworte auf Deutsch, in kurzen Absätzen, mit einer Zusammenfassung vorweg“.

Legen Sie eine Anweisung in der Workbench an (**CUSTOM INSTRUCTIONS** → **+**, **Create New Custom Instruction**) – mit **Instruction Name**, Beschreibung und **Instruction Content**. Nutzen Sie sie dann für ein Projekt oder einen Tab: im Dialog **New Chat or Project** unter **Advanced Settings** oder beim Bearbeiten eines Tabs. Wie Prompts und Datenquellen lassen sich auch eigene Anweisungen mit Kollegen, Teams oder dem Unternehmen teilen.

Ihre eigenen Anweisungen kommen nach den **Assistenten-Anweisungen** des Unternehmens, die Ihre Administration für jedes Modell festlegt (Kapitel 12).

9 Chatbots für Websites und Anwendungen

Jeder Chat-Tab – mit Modell, Datenquelle und Anweisungen – kann zum Chatbot auf einer Intranetseite oder in einer Anwendung werden.

Einen Tab als Chatbot exportieren

- 1 Richten Sie einen Tab mit Modell, Datenquelle und den eigenen Anweisungen ein, die der Chatbot nutzen soll, und testen Sie ihn mit ein paar Fragen.
- 2 Öffnen Sie das Tab-Menü (⋮) und wählen Sie **Export To Chatbot**. Bestätigen Sie mit **Export**.
- 3 Richten Sie den Chatbot im Editor ein (Tabelle unten) und klicken Sie auf **Get Code**.
- 4 Kopieren Sie den Code mit **Copy Embed Code** und fügen Sie ihn am Ende Ihrer Seite ein, direkt vor dem schließenden `</body>` -Tag.

Einstellung	Was sie tut
Title, Description	Der Text in der Titelleiste des Chatbots; die Beschreibung sehen nur Sie, im Chatbots Management.
Color, Background Color, Color Presets, Dark Theme	Farben passend zu Ihrer Website.
Chatbot Width, Messages Area Height, Position, Full Screen	Größe in Pixeln und Position links oder rechts – oder die ganze Seite, für Seiten, die nur den Chatbot zeigen.
Start Minimized, Minimize in Icon	Startet zugeklappt, als Schaltfläche oder Symbol, bis ein Besucher ihn öffnet.
Realtime Streaming, Sample Prompts	Antworten erscheinen Wort für Wort; Beispielfragen, die Besucher vor dem Start anklicken können.
Clearable, Downloadable, Play Sound	Besucher dürfen den Chat leeren oder als HTML-Seite herunterladen; bei einer neuen Nachricht ertönt ein Signal.
Language of the UI	Die Sprache der Schaltflächen und Beschriftungen des Chatbots selbst.

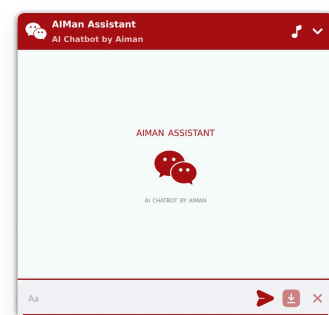
Chatbots Management

Das **Chatbots Management** (aus Ihrem Profil, unter Ihren Abkürzungen und Werkzeugen) listet die Chatbots, die Sie angelegt haben: **View, update or delete the custom chatbots you have created**. Geänderte Einstellungen ändern nicht das Aussehen auf Websites, in die der Chatbot schon eingebettet ist; ein gelöschter Chatbot funktioniert nirgends mehr.

Der AIMan Assistant

Der **AIMan Assistant** im Kontomenü öffnet unten rechts ein Chatbot-Fenster in AIMan – so sieht auch ein exportierter Chatbot auf Ihrer Website aus: Titelleiste, Unterhaltung und ein Nachrichtenfeld (**Aa**).

Chatbots bieten keine Websuche. Sie antworten mit Modell, Datenquelle und Anweisungen des exportierten Tabs; ihre Unterhaltungen werden nach einer festgelegten Frist archiviert.

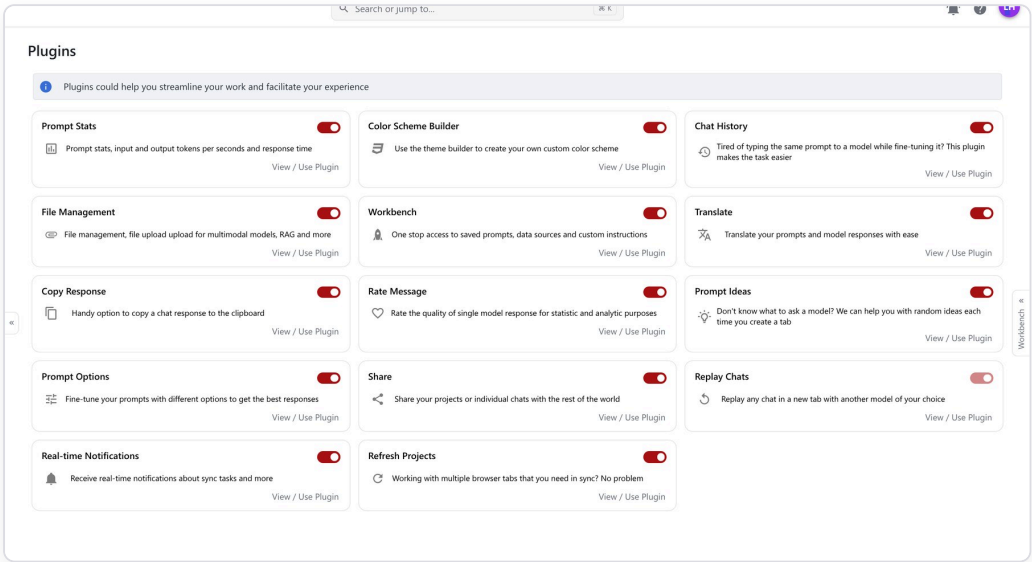


Der AIMan Assistant, ein eingebetteter Chatbot

10 Plugins

Viele Funktionen sind Plugins. Schalten Sie ein, was Sie nutzen – dann bleibt die Oberfläche so schlicht wie Ihre Arbeit.

Öffnen Sie **Plugins** aus Ihrem Kontomenü. Jede Karte hat einen Schalter und **View / Use Plugin**. Schalten Sie ein Plugin aus, verschwinden seine Schaltflächen und Menüeinträge für Sie; Ihre Kollegen betrifft das nicht.



Die Seite **Plugins**

Plugin	Was es ergänzt
Prompt Stats	Ein- und Ausgabe-Tokens pro Sekunde und Antwortzeit unter jeder Antwort (Stats For Nerds).
Color Scheme Builder	Ein eigenes Farbschema für AIMan (Kapitel 11).
Chat History	In einem Tab gesendete Prompts wiederverwenden.
File Management	Datei-Upload für multimodale Modelle, Datenquellen und mehr.
Workbench	Die Leiste mit gespeicherten Prompts, Datenquellen und eigenen Anweisungen.
Translate	Übersetzung von Prompts und Antworten mit DeepL.
Copy Response	Kopiert eine Antwort in die Zwischenablage.
Rate Message	Daumen hoch und runter für jede Antwort.
Prompt Ideas	Zufällige Prompt-Vorschläge auf der Startseite und in neuen Tabs.
Prompt Options	Temperatur und Antwortlänge pro Chat.
Share	Freigabe-Links für Chats und Projekte, wo eingeschaltet.
Replay Chats	Einen Chat in einem neuen Tab mit einem anderen Modell wiederholen.
Real-time Notifications	Benachrichtigungen in Echtzeit, etwa wenn eine Datenquelle fertig oder fehlgeschlagen ist.
Refresh Projects	Hält mehrere Browser-Tabs mit AIMan synchron.

11 Ihre Einstellungen

Ihr Profil, Ihre Startseite, Ihre Farben und die Tastenkürzel, die Zeit sparen.

Profil

Kontomenü → **Profile**. Unter **Account Details** ändern Sie Name und Foto (**Upload Photo**) sowie Ihr Passwort (**Change Password**). **Your AIman Shortcuts and Tools** führt zu Ihren Startoptionen, Lesezeichen, Chatbots und persönlichen Service-Tokens.

Startoptionen

Startup Options – Choose how to start AIman.

Option	Wo Sie nach der Anmeldung landen
Start Page	Gleich einen neuen Chat beginnen, die letzten Chats sehen und zu Modellen, Profil, Hilfe und mehr springen (Standard).
Your Chats and Projects	Die Übersicht Ihrer Chats und Projekte.
New Project	Ein neuer Chat mit einem Modell Ihrer Wahl, automatisch angelegt, wenn es keinen Standard gibt.
Your Profile	Direkt in Ihr Profil.
Custom URI	Eine beliebige Seite in AIman, etwa ein bestimmtes Projekt. Löschen Sie dieses Projekt später, entfernen Sie es auch hier.

Darstellung und Farben

AIman hat eine helle und eine dunkle Darstellung; Sie wechseln im Kontomenü (**Set dark or light theme**). Farbe und Logo Ihres Unternehmens legt Ihre Administration fest. Mit dem Plugin **Color Scheme Builder** erstellen Sie ein eigenes Farbschema (**Theme Builder**, mit Vorschau für hell und dunkel); es überschreibt die Unternehmensfarben nur für Sie. **Set Default Theme** kehrt zu denen des Unternehmens zurück.

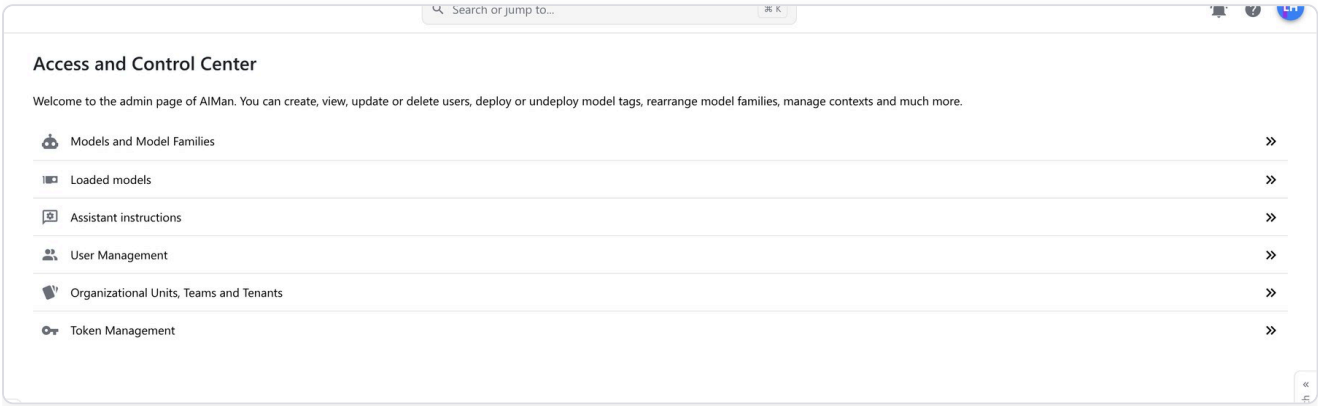
Schneller arbeiten

- **Befehlspalette** – **Strg+K** (**⌘K**) durchsucht Chats, Projekte, gespeicherte Prompts, Modelle und Seiten und führt Aktionen aus. Der schnellste Weg zu allem.
- **Fokusmodus** – **Strg+Umschalt+.** blendet Kopfzeile, Seitenleiste und Workbench aus, für lange Chats, in denen Sie nur die Unterhaltung brauchen. Verlassen mit **Esc**, demselben Kürzel oder **Exit focus mode** oben rechts. Danach ist alles wieder wie vorher.
- **Symbolleiste** – **Strg+Umschalt+S** klappt die Seitenleiste zu einer schmalen Leiste ein, in der „neuer Chat“, Suche und angeheftete Chats erreichbar bleiben.
- **Tastenkürzel** – **Strg+J** zeigt alle; die vollständige Liste steht in Anhang A.
- **Funktionstour** – **Start Feature Tour** im Hilfemenü führt Sie durch das Fenster.
- **Benachrichtigungen** – die Glocke in der Kopfzeile sammelt Meldungen, etwa wenn eine Datenquelle fertig ist.

12 Administration

Im Access and Control Center legt die Administration fest, welche Modelle die Mitarbeitenden sehen, wer was darf und mit welchen Schlüsseln jeder Aufruf läuft.

Sie öffnen es über das Kontomenü: **Access and Control Center**. Sie sehen nur die Kacheln, die Ihre Rolle erlaubt:



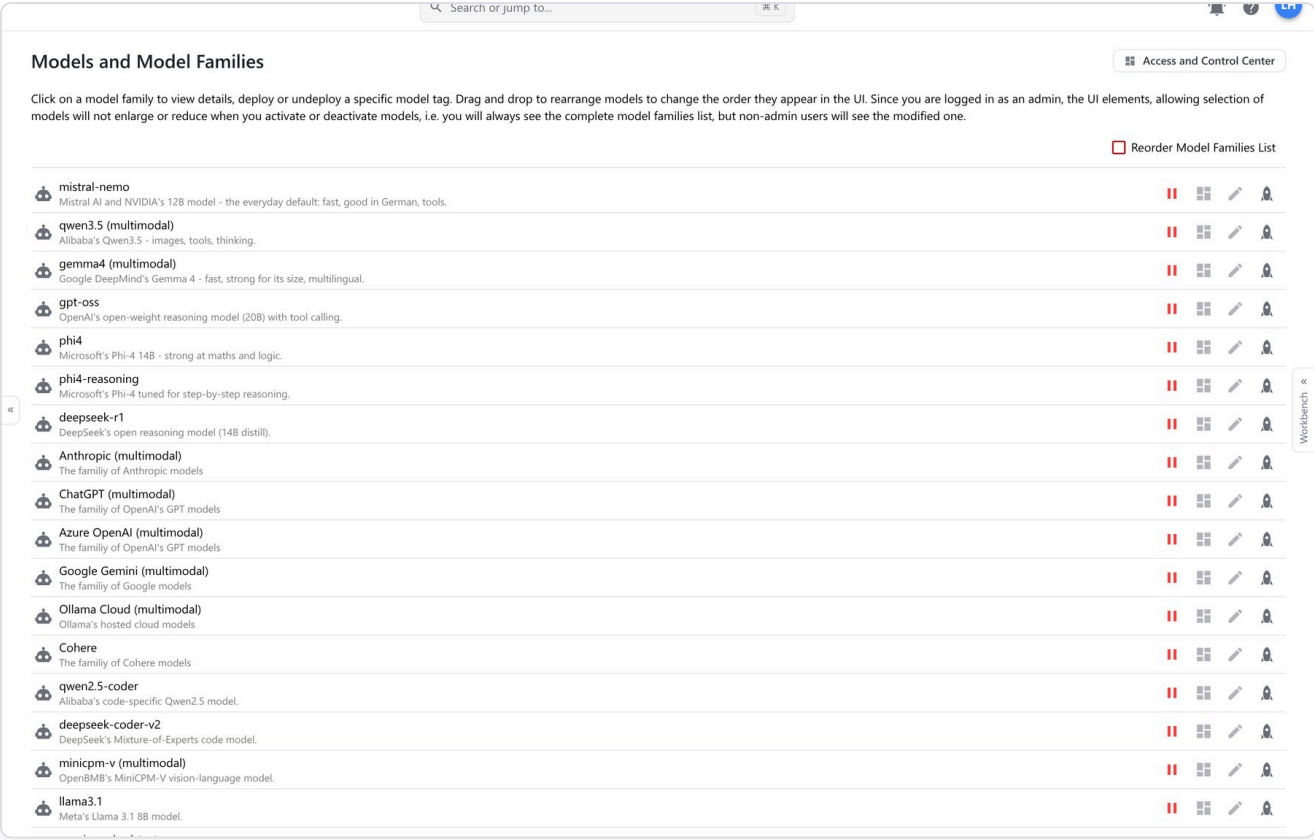
Das Access and Control Center

Kachel	Wer sie sieht
Models and Model Families	Superuser
Loaded models	Rollen, die Modelle bereitstellen dürfen
Assistant instructions	Superuser
User Management	Rollen, die Benutzer verwalten dürfen
Organizational Units, Teams and Tenants	Rollen, die Kontexte verwalten dürfen
Token Management	Rollen, die Tokens verwalten dürfen

Modelle, geladene Modelle und Assistenten-Anweisungen

Models and Model Families

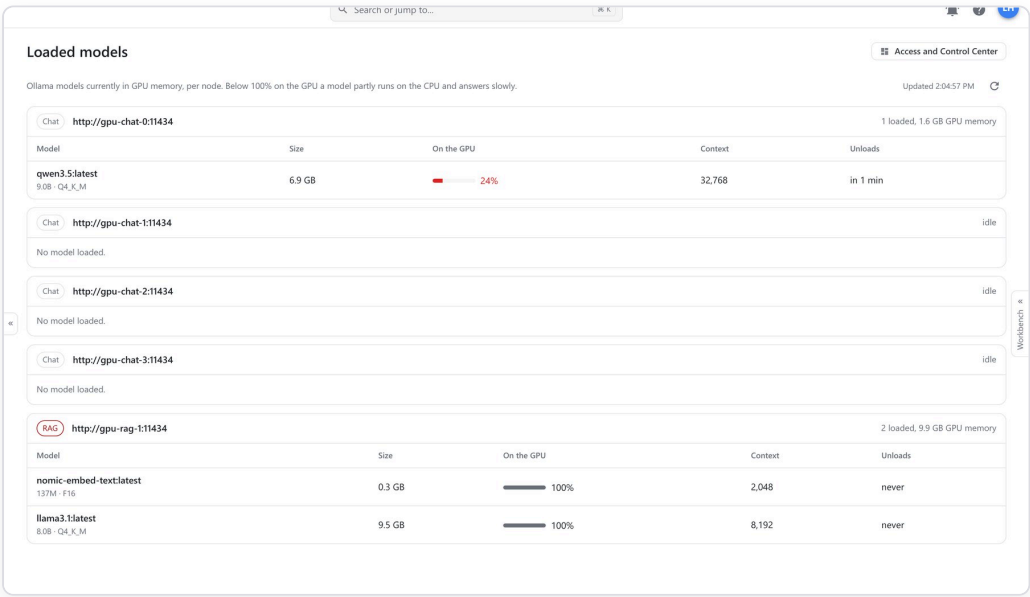
Die Liste aller Modellfamilien, in der Reihenfolge der Modellauswahl. Ein Klick auf eine Familie zeigt ihre Versionen (Tags), die Sie einzeln bereitstellen oder zurückziehen. Mit den Symbolen rechts aktivieren oder deaktivieren Sie eine Familie, bearbeiten ihre Beschreibung (die Zeile, die Nutzer in der Auswahl und in der Modellhilfe sehen) und stellen bereit. Mit **Reorder Model Families List** ziehen Sie Familien in eine neue Reihenfolge. Als Admin sehen Sie immer die vollständige Liste; andere sehen nur aktive Familien.



Models and Model Families

Loaded models

Zeigt je GPU-Server, welche offenen Modelle gerade im GPU-Speicher liegen: Größe, Anteil auf der GPU, Kontext und wann das Modell entladen wird. Unter 100 % auf der GPU (rot dargestellt) läuft ein Modell teilweise auf dem Prozessor und antwortet langsam. Chat-Server und der für die Dokumentverarbeitung reservierte Server (markiert mit **RAG**) stehen getrennt. Die Seite aktualisiert sich alle 15 Sekunden.



Geladene Modelle je GPU-Server

Assistant instructions

Jedes Modell erhält dieselbe kurze Grundanweisung: dass es der KI-Assistent Ihres Unternehmens in brandcompete AI Manager ist, das heutige Datum, seinen Trainingsstand, die Antwortsprache und – bei

Dokumenten oder Webergebnissen – wie es diese nutzt und zitiert. Unter **Addition for this AIman** ergänzen Superuser, was jedes Modell zusätzlich wissen soll: Unternehmenskontext, Tonalität, zu verwendende Begriffe (bis zu 4.000 Zeichen). Mit **Save** gilt der Text ab dem nächsten Prompt, ohne Neustart. Die eigenen Anweisungen der Nutzer folgen danach.

Assistant instructions

Every model gets the same short base instruction: who it is, today's date and that it should say so instead of guessing about recent events. Add text here that every model in this AIman should know as well: company context, tone, terms to use. Users' own custom instructions are added after it.

Base instruction (always sent)

"You are the AI assistant for <company> in brandcompetete AI Manager. Today is Monday, 12 October 2026 (Europe/Berlin)." Plus the model's training cutoff, the answer language and, when documents or web results are attached, how to use and cite them.

Addition for this AIman

e.g. "We are a media agency. Answer in a friendly, concise tone and use British English for English answers."

0 / 4000

Save

Assistant instructions

Benutzer, Rollen und Organisationseinheiten

User Management

Create, view, delete, modify or change user roles. **Add User** legt ein Konto an; ein Klick auf eine Zeile bearbeitet es. Jedes Konto hat genau eine Rolle (Anhang A), und pro Konto legen Sie fest, ob es die Oberfläche, die API oder beides nutzen darf – ein Dienstkonto für eine Anwendung braucht kein Chatfenster. Wer sich zum ersten Mal per Single Sign-on anmeldet, erscheint hier als **Guest**; vergeben Sie die passende Rolle. Ein Rollenwechsel wirkt bei der nächsten Aktion der Person, ohne neue Anmeldung. Die Rolle **Nobody** behält ein Konto, erlaubt aber nichts – praktisch, wenn jemand ausscheidet.

User Management

Create, view, delete, modify or change user roles.

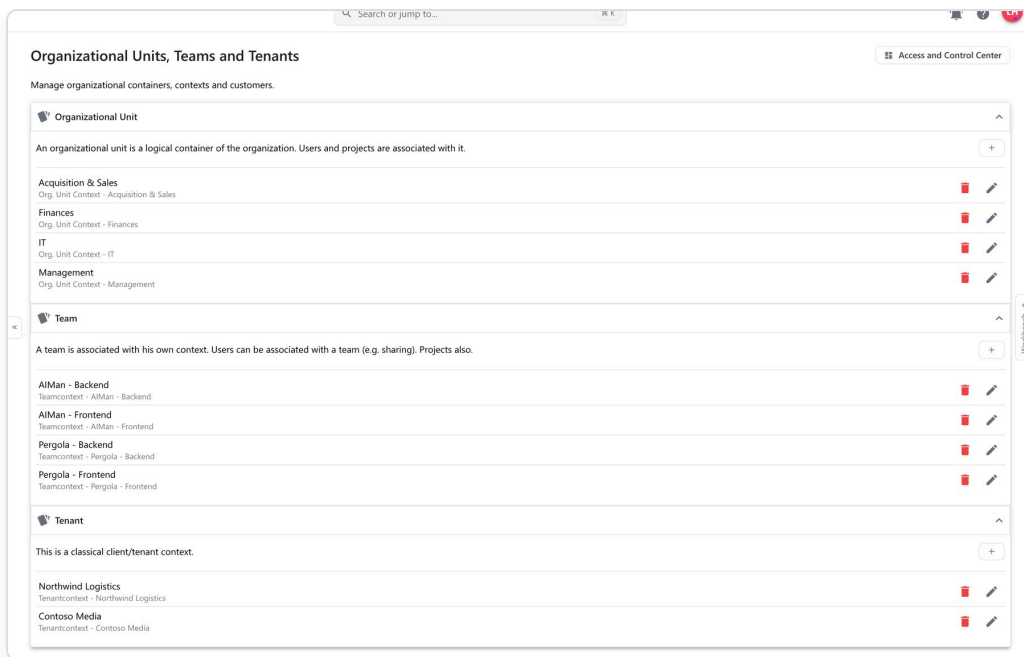
Filter by Name...

NAME	ID	UUID	CREATED	MODIFIED	ACTIONS
[Avatar]		6f77c7ca-c7fd-4d8b-92f9-14f6b675bb7a	13/02/24	13/02/24	[Edit] [Delete]
[Avatar]		67610307-55c3-45c3-8411-d03c9372d8e3	13/02/24	13/02/24	[Edit]
[Avatar]		6349c424-23f9-44fa-aa79-80fb2bde101c	05/02/25	05/02/25	[Edit]
[Avatar]		ecc2933f-d5cd-4cb5-af13-396c06ab3e8e	05/02/25	05/02/25	[Edit] [Delete]
[Avatar]		42dfd117-53ae-4d09-8468-483b4b5dae84	05/02/25	05/02/25	[Edit] [Delete]
[Avatar]		85f56cf3-42f6-4d30-a7ce-a8a03970eae7	05/02/25	05/02/25	[Edit] [Delete]
[Avatar]		5ec6d701-7e2e-4036-be19-23e0b2d4da9f	05/02/25	11/10/26	[Edit] [Delete]
[Avatar]		c481a7ad-d467-4019-b4e5-dcd0ca482775	05/02/25	11/10/26	[Edit] [Delete]

User Management (Namen unkenntlich gemacht)

Organizational Units, Teams and Tenants

Die Gruppen, mit denen geteilt wird. Eine **Organisationseinheit** ist ein Teil der Organisation, etwa eine Abteilung; ein **Team** ist eine Gruppe von Personen; ein **Tenant** ist ein Mandantenkontext. Mit **+** legen Sie eine Gruppe an und ordnen ihr Benutzer und Projekte zu. Was mit einer Einheit oder einem Team geteilt ist, erreicht alle Mitglieder – auch spätere –, und wer ein Objekt auf mehreren Wegen erreicht, behält den höchsten dieser Zugriffe.

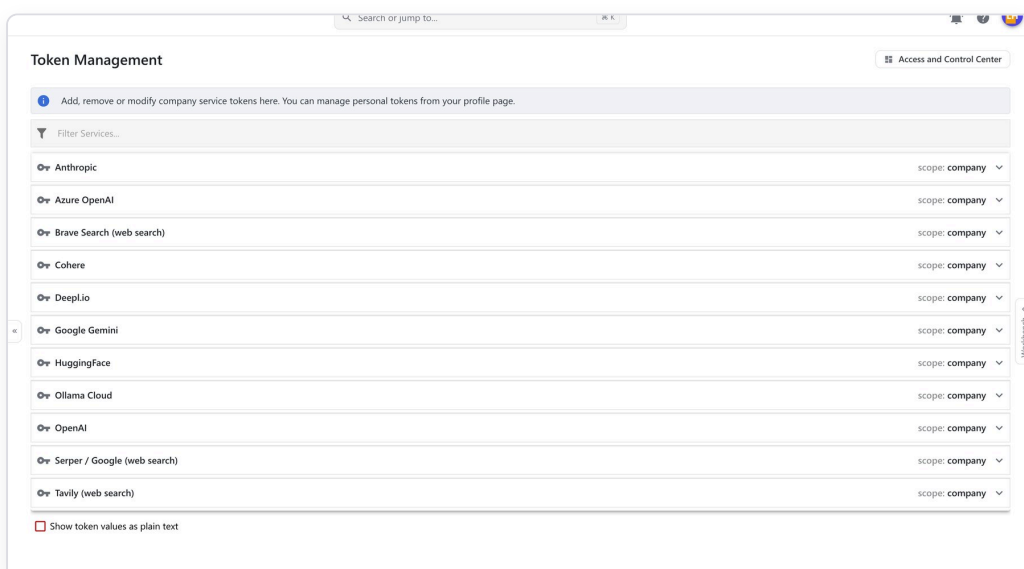


Organizational Units, Teams and Tenants

Schlüssel, Wissensdatenbanken, Branding und Anmeldung

Token Management

Die Anbieterschlüssel, mit denen jeder Aufruf läuft: Anthropic, Azure OpenAI, Cohere, DeepL (**Deepl.io**), Google Gemini, Ollama Cloud, OpenAI sowie die Suchdienste Brave Search, Serper / Google und Tavily. Klappen Sie einen Dienst auf und tragen Sie seinen Schlüssel ein; er gilt ohne Neustart. Schlüssel mit **scope: company** gelten für alle, niemand braucht einen eigenen; persönliche Schlüssel verwaltet jede Person in ihrem Profil. Die Werte bleiben verborgen, außer Sie haben **Show token values as plain text** an. Ein Anbieter ohne Schlüssel kann nicht antworten.



Token Management

TIPP

Ein Schlüssel für eine kostenpflichtige Such-API (Tavily, Brave oder Serper) verbessert die Antworten lokaler Modelle zu aktuellen Themen spürbar gegenüber den freien öffentlichen Suchmaschinen.

Wissensdatenbanken

Nur Rollen, die Wissensdatenbanken verwalten dürfen, sehen **Connect Knowledgebase** im Datenquellen-Dialog (Kapitel 7). Legen Sie die Verbindung an und teilen Sie sie mit dem Kontext **Company**, damit alle sie nutzen können: **This Knowledgebase is only visible to you until you share it below...** Verwenden Sie das Zugriffstoken eines Kontos, das genau das lesen darf, was alle sehen dürfen.

Branding

Jede Installation trägt Farbe und Logos des Unternehmens für die helle und die dunkle Darstellung, gewählt anhand der aufgerufenen Adresse; auch die Anmeldeseite zeigt sie. Ein Unternehmen kann AIMan unter seiner eigenen Domain betreiben. Das Branding richtet brandcompete beim Aufsetzen der Installation ein; sprechen Sie uns für Änderungen an.

Single Sign-on und Websuche

Single Sign-on mit Microsoft Entra ID und Google wird pro Installation konfiguriert: die zugelassenen E-Mail-Domains (eine leere Liste lässt niemanden herein), die Entra-Mandanten-IDs und die Google-Domains. Neue Nutzer werden Gäste. Die Websuche lässt sich für die ganze Installation abschalten. Beides stellt ein, wer die Installation betreibt – im Managed Service brandcompete.

13 Die API nutzen

Anwendungen nutzen dieselben Modelle, Datenquellen und Schlüssel wie Menschen – über eine REST-API mit Streaming-Antworten.

Die REST-API von AIMan liegt unter `https://<Ihre AIMan-API-Adresse>/api/v1`. Sie spricht JSON, authentifiziert mit Bearer-Tokens und ruft jedes bereitgestellte Modell über seinen Namen auf – eine Anwendung kann also das Modell wechseln, ohne den Code zu ändern. Die OpenAPI-Beschreibung wird mit der API ausgeliefert (`/api/v1/spec.html`).

WICHTIG

Die API von AIMan ist eine eigene REST-API. Einen OpenAI-kompatiblen Endpunkt (`/v1/chat/completions`) bietet sie **nicht**; Werkzeuge, die einen solchen erwarten, lassen sich nicht direkt auf AIMan richten.

Ein Konto mit API-Zugriff

Die Administration legt ein Konto an und gibt ihm die API frei (User Management, Kapitel 12); für ein Dienstkonto nur die API. Seine Rolle bestimmt, was es darf – genau wie bei Menschen.

1. Anmelden

```
POST /api/v1/auth/authenticate
Content-Type: application/json

{ "userName": "service@ihrefirma.de", "userPassword": "..." }
```

Die Antwort enthält `messageContent.data.access_token` und ein `refresh_token`. Senden Sie das Access-Token mit jeder Anfrage als `Authorization: Bearer <token>`. Es ist begrenzt gültig; `POST /api/v1/auth/refresh` mit dem Refresh-Token im Header `Authorization` liefert ein neues Paar, `/auth/logout` widerruft beide.

2. Das Modell finden

`GET /api/v1/models` listet die Modelle, die das Konto nutzen darf. Jedes hat einen `fullName` wie `mistral-nemo:latest` oder `Anthropic:claude-sonnet-5-5` – der Name, mit dem Sie prompten. `GET /api/v1/models/{modelId}/tags` listet die Versionen einer Familie.

3. Ein Modell über seinen Namen aufrufen

```
POST /api/v1/prompts/models/mistral-nemo:latest
Authorization: Bearer <token>
Content-Type: application/json

{ "prompt": "Fasse unsere Reiserichtlinie in drei Stichpunkten zusammen.",
  "datasourceId": 42 }
```

Der Name lautet `<Familie>:<Tag>`, ohne Rücksicht auf Groß- und Kleinschreibung – oder nur die Familie für ihre Standardversion. Ein unbekannter Name ergibt 404, ein mehrdeutiger 400. Der Antworttext steht in `messageContent.data.responseText`.

Feld	Bedeutung
<code>prompt</code>	Der zu sendende Text (Pflicht).
<code>datasourceId</code> , <code>amountOfChunks</code>	Eine Datenquelle, aus der geantwortet wird, und wie viele Passagen (1–10, Standard 5).
<code>ragAssisted</code>	<code>true</code> erlaubt Allgemeinwissen neben den Dokumenten (Standard <code>false</code> : nur die Dokumente).
<code>webSearch</code>	<code>true</code> stützt die Antwort auf aktuelle Webergebnisse.
<code>think</code>	<code>true</code> / <code>false</code> für Modelle, die nachdenken können; weglassen für den Standard von AIMan.
<code>attachments</code>	Dateien, etwa Bilder für multimodale Modelle, Base64-kodiert.
<code>custInstWhoId</code> , <code>custInstRespondId</code>	Anzuwendende eigene Anweisungen.
<code>options</code>	Prompt-Optionen wie <code>temperature</code> und <code>num_predict</code> ; gesendet wird nur, was Sie setzen.

4. Streaming

`POST /api/v1/prompts/models/{modelName}/stream` nimmt denselben Body und antwortet mit Server-Sent Events (`text/event-stream`). Ereignisse namens `Response` liefern die Antwort Stück für Stück und enden mit `[STREAM-DONE]` ; `Thinking` transportiert die Überlegungen eines Reasoning-Modells, `Status` Fortschritte wie eine laufende Websuche, `Citations` die Quellen und `Error` einen Fehler. Einen Stream stoppen Sie mit `POST /api/v1/prompts/streams/{streamId}/cancel` , wenn Sie mit dem Prompt eine `streamId` geschickt haben.

HINWEIS

Aufrufe über die Modell-ID statt des Namens (`/api/v1/prompts/{modelTagId}`), die Übersetzung (`/api/v1/tools/translate/...`) und die übrigen Ressourcen beschreibt die OpenAPI-Spezifikation.

14 Daten und Datenschutz: was wohin geht

AIman macht die Wahl sichtbar: Lokale und Anbietermodelle sind getrennte Tabs in der Auswahl. Dieses Kapitel sagt genau, was die Server Ihrer AIman-Installation verlässt, und wann.

Was Sie tun	Wo die Daten verarbeitet werden
Chat mit einem lokalen Modell	Prompt, Antwort und alle Dokumentpassagen bleiben auf den dedizierten GPU-Servern Ihrer Installation – den Servern von brandcompete in Deutschland oder Ihren eigenen on-premise.
Chat mit einem Anbietermodell	Der Prompt, die Unterhaltung und die für die Antwort nötigen Passagen gehen an diesen Anbieter (Anthropic, OpenAI, Microsoft Azure, Google, Cohere oder Ollama Cloud), unter dem Vertrag und mit dem Schlüssel Ihres Unternehmens.
Eine Datenquelle anlegen	Lesen, Zerlegen, Indexieren, Embeddings und Reranking laufen immer auf den Servern Ihrer Installation, egal welches Modell später antwortet. Wissensdatenbanken werden aus dem eigenen Confluence, Jira oder Bitbucket Ihres Unternehmens gelesen.
Übersetzen	Der zu übersetzende Text geht an DeepL.
Websuche	Eine aus Ihrer Frage abgeleitete Suchanfrage geht an einen Suchdienst: bei Claude, Gemini und GPT an die Suche des Anbieters, sonst an den Suchdienst von AIman, der öffentliche Suchmaschinen oder eine von Ihrem Unternehmen eingerichtete kostenpflichtige Such-API abfragt.

- **Eine Installation pro Kunde.** Jeder Kunde hat seine eigene Datenbank, seinen eigenen Suchindex, seine eigene Warteschlange und sein eigenes Netz. Im Managed Service werden nur die GPUs geteilt.
- **Privat, bis geteilt.** Chats, Datenquellen, Prompts und Anweisungen sieht nur, wer sie angelegt hat, bis er sie teilt.
- **Schlüssel zentral verwahrt.** Anbieterschlüssel werden einmal pro Unternehmen gespeichert und auf dem Server verwendet; niemand braucht ein privates Konto.
- **Bereinigte Antworten.** Modellausgaben werden vor der Anzeige bereinigt, damit eine präparierte Antwort keinen Code in Ihrem Browser ausführen kann.

WICHTIG

Faustregeln: Für vertrauliches Material ein lokales Modell nutzen; die Websuche nicht bei Fragen mit vertraulichen Details einschalten; und die Richtlinie Ihres Unternehmens befolgen, welche Daten an welchen Anbieter gehen dürfen.

15 Fehlerbehebung & FAQ

Die häufigsten Fragen, und was zu tun ist.

Problem oder Frage	Was zu tun ist
Ich kann mich nicht mit Microsoft oder Google anmelden.	Lesen Sie die Meldung auf der Anmeldeseite. Meist steht Ihre E-Mail-Domain oder Ihr Verzeichnis nicht auf der Freigabeliste des Unternehmens, oder Ihr Konto wurde deaktiviert. Wenden Sie sich an Ihre Administration.
Ich habe nur die Rolle Guest.	Wer sich zum ersten Mal per Single Sign-on anmeldet, beginnt als Gast. Bitten Sie Ihre Administration um die Rolle, die Sie brauchen.
Ein erwartetes Modell fehlt oder ist ausgegraut.	Welche Familien und Versionen angeboten werden, entscheidet Ihre Administration; n/a heißt „nicht bereitgestellt“. Ein Anbietermodell braucht zudem einen Unternehmensschlüssel.
Das Modell sagt, es könne etwas Aktuelles nicht wissen.	So ist es gewollt: Es kennt das heutige Datum, aber nicht, was nach seinem Training geschah. Schalten Sie die Websuche ein.
Eine Antwort braucht lange, bis sie beginnt.	Die Websuche kostet einige Sekunden, Thinking mehr. Ein lokales Modell, das länger nicht genutzt wurde, wird erst in den GPU-Speicher geladen. Anbietermodelle hängen vom Anbieter ab.
Die Antwort bricht mittendrin ab.	Vielleicht ist die Antwortlänge begrenzt (Prompt Options , num predict), oder es wurde gestoppt. Schreiben Sie „bitte weiter“.
„No relevant content was found...“	Die Dokumente decken die Frage nicht ab. Formulieren Sie um, erhöhen Sie Amount of Chunks , hängen Sie eine andere Datenquelle an oder entfernen Sie den Haken bei Only from the documents .
Meine Datenquelle bleibt bei Indexing oder zeigt Failed.	Große Dateien und kontextuelle Indexierung brauchen Zeit. Bei Failed zeigen Sie auf das Statussymbol, um den Grund zu sehen – etwa eine passwortgeschützte Datei. Gescannte PDFs ohne Text sind nicht lesbar (noch keine OCR).
Use Data Source ist ausgegraut.	Nur fertige Datenquellen lassen sich anhängen. Warten Sie auf den grünen Status.
Die Antwort kommt in der falschen Sprache.	Modelle antworten in der Sprache Ihrer Frage. Schreiben Sie in der gewünschten Sprache, sagen Sie es ausdrücklich („antworte auf Deutsch“) oder legen Sie es in einer eigenen Anweisung fest.
Ich habe den Browser geschlossen, während eine Antwort lief.	Die Antwort wird im Hintergrund fertig geschrieben und gespeichert; öffnen Sie den Chat einfach wieder.
Der Eintrag Thinking fehlt.	Das aktuelle Modell kann nicht nachdenken. Wählen Sie etwa qwen3.5 oder gemma4.
Der Globus für die Websuche fehlt.	Die Websuche ist für Ihre Installation abgeschaltet, oder Sie sind in einem eingebetteten Chatbot.
Ich möchte Modelle vergleichen.	Fügen Sie Tabs mit verschiedenen Modellen hinzu und nutzen Sie Prompt All Tabs – oder wiederholen Sie einen Chat mit einem anderen Modell.

Problem oder Frage	Was zu tun ist
Kann AIman Bilder erzeugen oder sprechen? Gibt es eine deutsche Oberfläche?	Keine Bilderzeugung und keine Spracheingabe oder -ausgabe: AIman liest Bilder mit multimodalen Modellen. Die Oberfläche ist vorerst englisch; die Modelle antworten in Ihrer Sprache.
Wo sehe ich unsere Nutzung und Kosten?	Nutzungsstatistiken gibt es noch nicht. Die Nutzung der Anbietermodelle erscheint auf den Rechnungen der Anbieter an Ihr Unternehmen.
An wen wende ich mich?	Zuerst an Ihre Administration; zum Produkt an brandcompete unter sales@brandcompete.com .

A Referenz

Rollen, Status, Tastenkürzel und die Begriffe dieses Handbuchs.

Rollen

Jedes Konto hat eine Rolle. Die Administration vergibt Rollen, keine einzelnen Rechte. Was eine Rolle erlaubt, wird bei jeder Aktion geprüft.

Rolle	Typischerweise für
Superuser	Alles, einschließlich Modellen, Assistenten-Anweisungen und aller Admin-Bereiche.
Administrator	Den Betrieb von AIMan: Benutzer, Einheiten und Teams, Schlüssel, Modelle.
Teamlead	Alles, was ein Staffmember kann, dazu Teams und andere Gruppen zum Teilen anlegen.
Staffmember	Der Alltag: Chats, Projekte, Datenquellen, Prompts, Anweisungen, Chatbots.
Guest	Ein eingeschränkter Einstieg, etwa nach der ersten Anmeldung per Single Sign-on.
Nobody	Das Konto existiert, darf aber nichts – für ausgeschiedene Personen.
Prompt Repo User	Für die Prompt-Bibliothek vorgesehen; vorher mit der Administration klären.
Modelcrawler	Technisches Konto für den Modellkatalog; nicht für Personen.

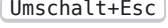
Zusätzlich zur Rolle gewähren Freigaben Zugriff auf einzelne Objekte: **Lesen**, **Schreiben** oder **Verwalten**.

Status von Datenquellen

Status	Bedeutung
Pending (grau)	Wartet auf die Verarbeitung.
Indexing (gelb)	Wird gelesen und indexiert.
Ready (grün)	Kann an Chats angehängt werden.
Failed (rot)	Verarbeitung fehlgeschlagen; das Symbol zeigt den Grund.

Tastenkürzel

Auf dem Mac  statt . In der Oberfläche heißt die Taste .

Kürzel	Was es tut
	Chats, Projekte, Prompts, Modelle und Befehle durchsuchen
	Neuen Chat oder neues Projekt anlegen
	Seitenleiste ein- oder ausklappen
	Fokusmodus: Kopfzeile und Seitenleisten ausblenden ( zum Verlassen)
	Neuen Tab im aktuellen Chat oder Projekt hinzufügen
	Cursor ins Eingabefeld des aktuellen Tabs setzen
	Verfügbare Tastenkürzel anzeigen

Glossar

Begriff	Bedeutung
Access and Control Center	Der Administrationsbereich.
Anbietermodell	Ein gehostetes Modell von Anthropic, OpenAI, Azure OpenAI, Google, Cohere oder Ollama Cloud.
Assistenten-Anweisungen	Der unternehmensweite Text, den jedes Modell erhält, festgelegt von Superusern (Assistant instructions).
Befehlspalette	Die Suche in der Kopfzeile (Strg+K).
Chunk	Eine Passage eines Dokuments, die Einheit, in der eine Datenquelle durchsucht wird.
Eigene Anweisung	Ihr fester Kontext für das Modell: wer Sie sind, wie es antworten soll.
Familie / Tag	Eine Modellfamilie (etwa <i>Anthropic</i>) und eine ihrer konkreten Versionen (etwa <i>claude-sonnet-5-5</i>).
Kontext	Eine Gruppe, mit der geteilt wird: eine Person, ein Team, eine Organisationseinheit, ein Tenant oder das Unternehmen.
Kontextuelle Indexierung	Eine kurze Notiz pro Passage, die sie in ihrem Dokument verortet, für bessere Suchergebnisse.
Lokales Modell	Ein offenes Modell auf den GPU-Servern Ihrer Installation.
Multimodal	Ein Modell, das neben Text auch Bilder versteht.
Projekt	Ein geteilter Chat.
Prompt	Was Sie an ein Modell senden.
Prompt All Tabs	Ein Prompt, der gleichzeitig an alle Tabs eines Chats geht.
Quellenangabe	Eine Nummer wie [2] in einer Antwort, die unter Sources auf eine Dokument- oder Webseite verweist.
RAG	Retrieval-Augmented Generation: Antworten aus Passagen, die in einer Datenquelle gefunden wurden.
Reranking	Ein zweiter Durchgang, der nur die Passagen behält, die die Frage wirklich beantworten.
Streaming	Die Antwort erscheint Wort für Wort, während sie geschrieben wird.
Thinking	Ein Reasoning-Modell durchdenkt das Problem, bevor es antwortet.
Token	(1) Ein Wortstück, die Einheit, in der Modelle Text zählen. (2) Ein Schlüssel oder Zugriffstoken für einen Anbieter oder die API.
Wissensdatenbank	Eine Datenquelle, die mit Confluence, Jira oder Bitbucket verbunden ist (Knowledgebase).
Workbench	Die Leiste rechts mit Prompts, Datenquellen und eigenen Anweisungen.

Fragen oder Rückmeldungen: sales@brandcompete.com • [brandcompete GmbH](#) • www.aiman.rocks • www.brandcompete.com