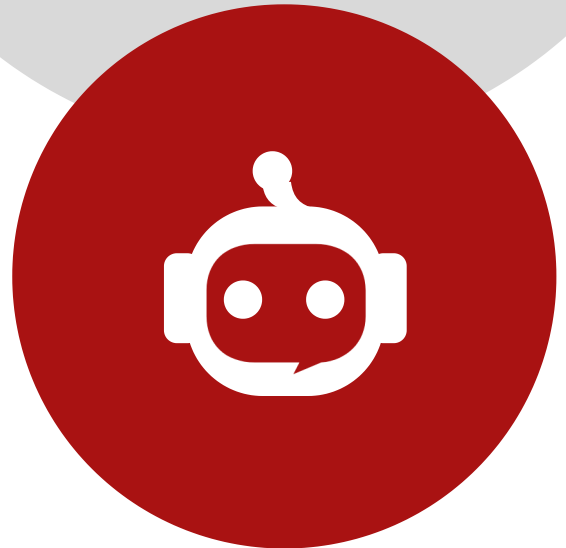


■ brandcompete

AIMan

User Guide — every leading AI model in one workspace, governed from the first prompt

RELEASE YOUR INNER PROMPT ENGINEER.



Contents

1	Welcome to AIman	2
2	Getting started	4
	Signing in	4
	Finding your way around	5
	Your first prompt	6
3	Models	7
	Choosing a model in the picker	8
	Which model for which task	9
4	Chats, projects and tabs	10
	Several models side by side: tabs	10
	Working with answers	11
	Projects: sharing a chat with colleagues	12
5	The composer	13
	Thinking: letting a model reason first	15
6	Web search	16
7	Data sources: answers from your documents	18
	Creating a data source	18
	Knowledge bases: Confluence, Jira, Bitbucket	20
	Asking questions about your documents	21
	Sharing data sources, and current limits	22
8	Workbench, prompt repository and custom instructions	23
9	Chatbots for websites and applications	25
10	Plugins	26
11	Your preferences	27
12	Administration	28
	Models, loaded models and assistant instructions	28
	Users, roles and organisational units	30
	Keys, knowledge bases, branding and sign-in	31
13	Using the API	33
14	Data and privacy: what goes where	35
15	Troubleshooting & FAQ	36
A	Reference: roles, statuses, shortcuts, glossary	38

1 Welcome to AIMan

AIMan puts the leading AI models in one workspace for the whole company: one sign-in, one interface, answers grounded in your documents and the web, and rules that IT sets once.

You sign in with your work account and chat with two kinds of model through the same interface. **Open models** such as Mistral NeMo, Qwen, Gemma, gpt-oss, Phi and DeepSeek run on brandcompete's dedicated GPU servers in Germany, or on your company's own servers if it runs AIMan on-premise. **Vendor models** – Anthropic Claude, OpenAI GPT, Azure OpenAI, Google Gemini, Cohere and Ollama Cloud – run at the vendor, under your company's own key. Nobody needs a private AI account to work.

Answers can draw on your company's knowledge: documents you upload, web pages, and Confluence, Jira or Bitbucket. Each answer cites what it used, document and page. With web search switched on, a model can also answer about today's news, with links to its sources.

The things you work with

Term	What it is
Chat	A conversation of your own. Nobody else sees it.
Project	A shared chat. You invite colleagues, teams, organisational units or the whole company, and work in it together.
Tab	Every chat and project has one or more tabs. Each tab has its own model, data source and conversation, so you can compare models side by side.
Model	The AI that answers. Models are grouped in families (for example <i>Anthropic</i> or <i>qwen3.5</i>); a family has one or more concrete versions, called tags .
Composer	The message box at the bottom of a chat, with the model, web search, attachments and options.
Data source	A collection of documents, web pages or a connected knowledge base that a model can look things up in before it answers.
Workbench	The panel on the right with your saved prompts, your data sources and your custom instructions.
Prompt repository	Saved prompts, with <code>{{placeholders}}</code> for the parts that change, categories, tags and sharing.
Custom instruction	Standing context the model gets with every prompt: who you are, how it should answer.
Chatbot	A chat tab exported, with its model and data source, as a chatbot you embed in a website or application.

What you can do depends on your role

Every account has a **role**, such as Staffmember, Teamlead or Administrator, and the role decides what you may create and manage: chats, data sources, prompts, chatbots, or users and models. Areas your role does not include are not shown. Everything you create is **private until you share it** – with named colleagues, a team, an organisational unit or the whole company, to read, to write, or to administer. The roles are listed in Appendix A.

Where your data goes, in one paragraph

The picker shows open models and vendor models in separate tabs, so you always see where a prompt goes. With an open model, the prompt, the answer and any document passages stay on the GPU servers that run your AlMan. With a vendor model, the prompt and the passages it needs go to that vendor, under your company's contract and key. Reading and indexing your documents always happens on those servers, whichever model answers. Chapter 14 has the details.

NOTE

This guide describes AlMan as of October 2026. The interface is in English; the labels in this guide are quoted exactly as they appear on screen. Models answer in the language you write in. A few areas you may see in the product are not finished yet; they are marked **Not available yet** .

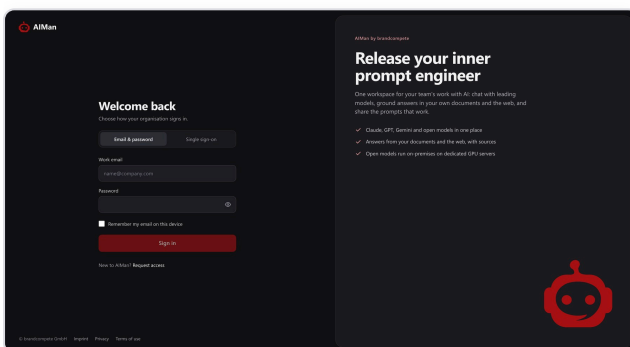
2 Getting started

Sign in, find your way around the window, and send your first prompt.

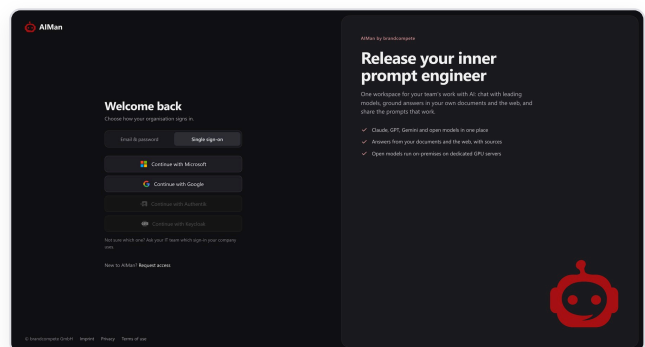
Signing in

Open the AIman address your company gave you. The sign-in page has two tabs:

- **Email & password** – enter your **Work email** and **Password** and click **Sign in**. The eye icon shows the password while you type. Tick **Remember my email on this device** to have the address filled in next time. The email address is not case-sensitive.
- **Single sign-on** – click **Continue with Microsoft** or **Continue with Google** and sign in with the work account you already have. AIman supports Microsoft Entra ID and Google; administrators set them up per installation. Buttons for providers that are not configured are greyed out.



Signing in with email and password



The single sign-on tab with Microsoft and Google configured

Single sign-on only admits addresses from the email domains and directories your company has named. If you see an error after signing in with Microsoft or Google, it tells you why – for example that your domain is not allowed – and the single sign-on tab opens again.

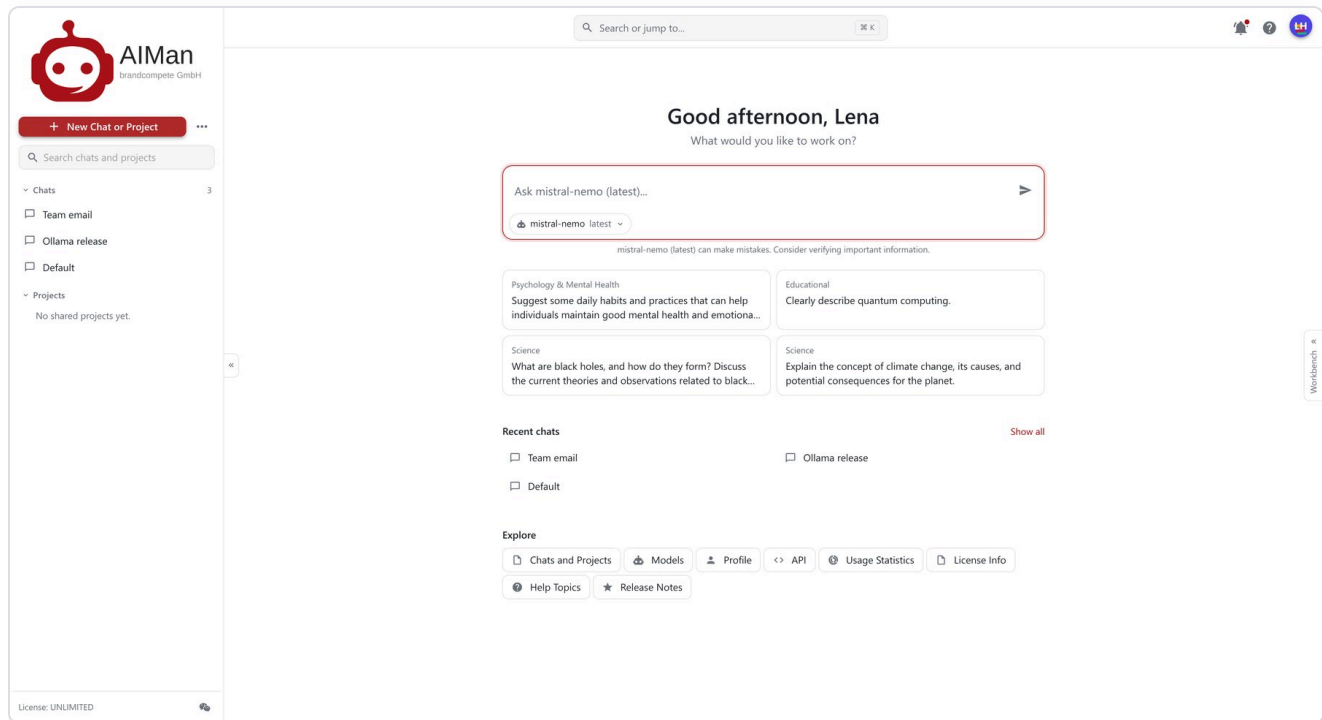
NOTE

The first time you sign in with single sign-on, AIman creates your account with the role **Guest**. A guest can work with AIman in a limited way; ask your administrator for the role your work needs. **New to AIman? Request access** on the sign-in page leads to a contact form.

IMPORTANT

Forgotten your password? Ask your administrator to set a new one. Once signed in, you can change your password yourself in your profile (chapter 11).

Finding your way around



The start page: sidebar on the left, the composer in the middle, the workbench handle on the right edge

- **Sidebar** — **New Chat or Project** at the top, then one search field (**Search chats and projects**) and the sections **Pinned**, **Chats** and **Projects**. Click a section heading to fold it. Every row has a ... menu to pin the chat to the top, see who it is shared with, or delete it. A pulsing dot shows that an answer is still being written. The « handle on the sidebar's edge, or **Ctrl+Shift+S**, shrinks the sidebar to a slim icon rail.
- **Header** — in the middle the search **Search or jump to...** (**Ctrl+K**, on a Mac **⌘K**): it finds any chat, project, saved prompt, model or page and runs actions such as “new chat” or “focus mode”. On the right: notifications (the bell), the ? help menu (**Help Topics**, **Keyboard Shortcuts**, **Start Feature Tour**, **YouTube Tutorial Videos**, **Release Notes**, **About**) and your account menu (**Profile**, **Bookmarks**, **Plugins**, the **AIMan Assistant**, **Access and Control Center** for admins, the light/dark switch and **Sign Out**).
- **Workbench** — the handle labelled **Workbench** on the right edge opens the panel with your prompts, data sources and custom instructions (chapter 8).

The start page

After signing in you land on the start page. It greets you by name and asks **What would you like to work on?** Below the composer you find four **prompt ideas** to try (a new set every visit), your **Recent chats** with **Show all**, and an **Explore** row with shortcuts to **Chats and Projects**, **Models**, **Profile**, **API**, **Help Topics**, **Release Notes** and more. You can choose a different landing page in your startup options (chapter 11).

NOT AVAILABLE YET

The **Usage Statistics** tile opens a page that is still being built: AIMan does not show usage or cost figures yet. The in-app **API** help page is still empty; chapter 13 of this guide describes the API.

Your first prompt

- 1 Type your question in the box on the start page. The placeholder tells you which model will answer, for example **Ask mistral-nemo (latest)...**
- 2 To use a different model, click the model chip below the text (**mistral-nemo latest**) and pick one (chapter 3). mistral-nemo, an open model on the GPU servers, is preselected.
- 3 Press **Enter** or click the arrow. AIman creates a new chat, names it after your prompt, and the answer appears word by word.
- 4 Ask follow-up questions in the composer at the bottom of the chat. The model sees the conversation so far.

Below every composer AIman reminds you that the model *can make mistakes*. Check anything important – the citations under an answer (chapters 6 and 7) make that easy.

TIP

Every model knows today's date and is told to say so when a question is beyond what it can know, instead of guessing. For current events, switch on web search (chapter 6).

3 Models

Two kinds of model, in one picker: open models on dedicated GPU servers, and the vendors' hosted models under your company's key.

Local models and vendor models

Local models are open models that run on the GPU servers of your AIMan installation — brandcompete's dedicated servers in Germany, or your company's own. Prompts to them do not leave those servers, and there are no per-token fees. Every local model handles a context of 32,000 tokens, enough for long conversations and several pages of document passages.

Vendor models are the hosted models of Anthropic, OpenAI, Azure OpenAI, Google, Cohere and Ollama Cloud. They are often the strongest for difficult tasks. Your prompt, and the passages it needs, are sent to the vendor under your company's key and contract.

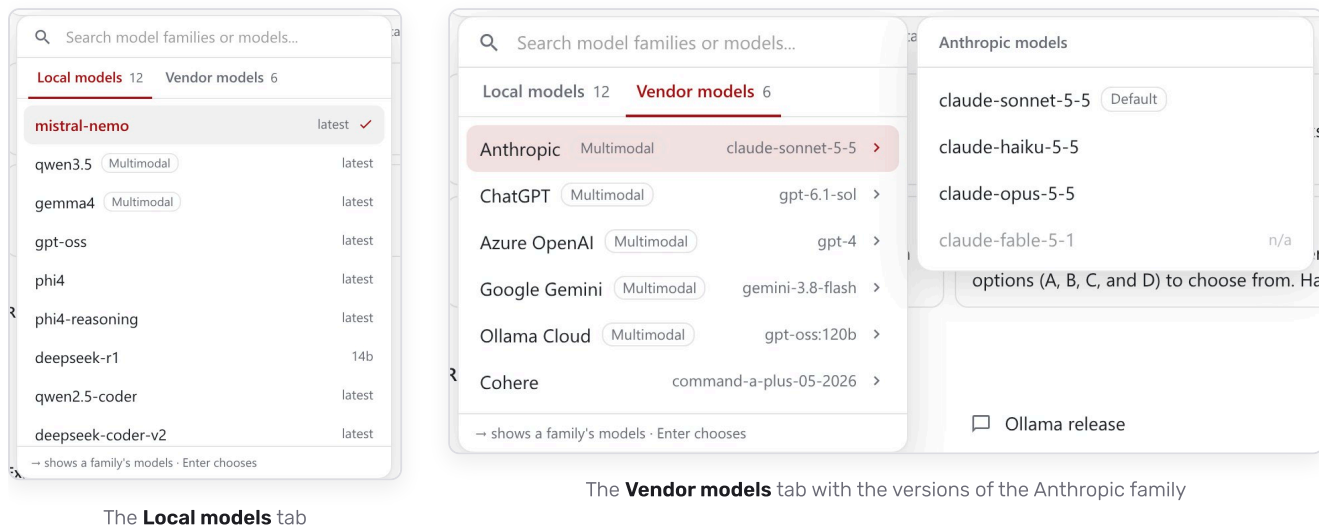
Which families and versions you see is decided by your administrators. The table shows the catalogue as of October 2026.

Local model	Good at
mistral-nemo	The everyday default. Mistral AI and NVIDIA's 12B model: fast, good in German, can use tools (for example web search).
qwen3.5	Alibaba's Qwen 3.5: understands images, uses tools, can think step by step.
gemma4	Google DeepMind's Gemma 4: fast, strong for its size, multilingual, understands images.
gpt-oss	OpenAI's open-weight reasoning model (20B), with tool calling.
phi4	Microsoft's Phi-4 (14B): strong at maths and logic.
phi4-reasoning	Phi-4 tuned for step-by-step reasoning.
deepseek-r1	DeepSeek's open reasoning model (14B distill).
qwen2.5-coder	Code: writing, explaining and reviewing it.
deepseek-coder-v2	Code, DeepSeek's mixture-of-experts code model.
minicpm-v	Vision: describing and reading images.
llama3.1	Meta's Llama 3.1 (8B), a compact general model.

Vendor family	Default version	Other versions
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

Choosing a model in the picker

Click the model chip in the composer – or the **Model** field in the **New Chat or Project** dialog – to open the picker.



- Two tabs, **Local models** and **Vendor models**, each with a count. The picker opens on the tab of the model you are using, so a chat on Claude opens the vendor tab.
- Type into **Search model families or models...** to search both tabs at once.
- A family with several versions has a chevron. Click it, or press **→**, to see its versions on the right; the family's standard version is marked **Default**. A version marked **n/a** is not deployed and cannot be chosen. A family with only one version (most local models) is chosen with a single click.
- The **Multimodal** badge marks families that understand images.
- Keyboard: arrows to move, **Enter** to choose, **←** or **Esc** to go back.

The model you choose applies to the current tab only. To compare models, give each tab its own (chapter 4). The page **Available Models** in the help (the **Models** tile on the start page) lists every model with a one-line description and whether it is deployed.

Which model for which task

Task	Good choices
Everyday writing, emails, summaries, German texts	mistral-nemo (default); gemma4 for other languages
Confidential material	Any local model: the prompt stays on the GPU servers
Difficult analysis, long reasoning, best quality	Claude (claude-opus-5-5, claude-sonnet-5-5), GPT, Gemini; locally qwen3.5 or gpt-oss with Thinking switched on
Maths and logic puzzles	phi4, phi4-reasoning, deepseek-r1
Programming	qwen2.5-coder, deepseek-coder-v2; Claude and GPT for larger tasks
Images: describe, read a chart or a photo	qwen3.5, gemma4, minicpm-v; the vendor families marked Multimodal
Current events (with web search)	Claude, Gemini and GPT use their own search; locally qwen3.5 or mistral-nemo
Answers from your documents	mistral-nemo or qwen3.5 locally; any vendor model if your company allows passages to go to it

Images

Models marked **Multimodal** can look at images. Attach a picture with the paperclip in the composer (JPEG, PNG, GIF, WebP or SVG) and ask about it: describe it, read the text in a screenshot, compare two diagrams. AIMan does not generate images.

4 Chats, projects and tabs

A chat is yours alone; a project is a chat you share. Both can hold several tabs, each with its own model.

Creating a chat or a project

The fastest way is to type on the start page: the first prompt creates a chat. For more control, click **New Chat or Project** in the sidebar (**Ctrl+Shift+P**):

- 1 Pick the **Model** for the first tab.
- 2 Enter a **Chat / Project Name** (required).
- 3 For a shared chat, tick **This is a Project (Shared Chat)** and choose whom to share it with under **Share With...** — colleagues, teams, organisational units or the company.
- 4 Optionally open **Advanced Settings**: a **Description**, a **Data Source** for the chat, and your **Custom Instructions**.
- 5 Click **Create**.

The **New Chat or Project** dialog

Above an open chat its name is shown as **Chat: ...** or **Project: ...**. The icons at the top right edit it (name, description, sharing), open **Prompt All Tabs** (below) and delete it. Deleting cannot be undone. The **Chats and Projects** page (start page, **Explore**) shows all your chats and projects as cards.

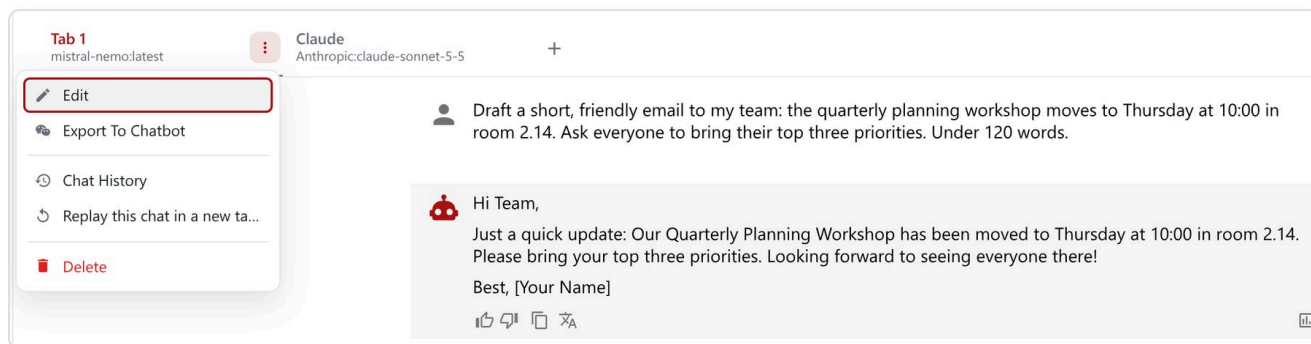
Several models side by side: tabs

A chat with two tabs: “Tab 1” on mistral-nemo and “Claude” on claude-sonnet-5-5, answering the same prompt

Every tab shows its name and, underneath, its model (for example **Anthropic:claude-sonnet-5-5**). Each tab has its own conversation, model, data source and settings. Add a tab with **+** next to the tabs (**Add New Tab**, **Ctrl+Shift+,**) and choose its model.

Prompt All Tabs sends one prompt to every tab at once. Click its icon at the top right of the chat, type the prompt in the field that opens, and send it: each tab answers with its own model. This is the quickest way to compare how Claude, GPT and a local model handle the same task.

The tab menu



The  menu of the selected tab

Entry	What it does
Edit	Rename the tab and change its model, data source and custom instructions. Clicking the model chip in the composer opens the same dialog.
Export To Chatbot	Turns the tab into an embeddable chatbot (chapter 9).
Share Chat	Creates a link to a read-only copy of the conversation. Shown only where link sharing is enabled for your installation (see “Share links, export and print” below).
Chat History	Lists the prompts you sent in this tab, to reuse one: put it into the composer, or Send directly to chat .
Replay this chat in a new tab with another model	Sends the whole conversation's prompts again, in a new tab, to a model you choose – the easiest way to see how another model would have answered.
Delete	Deletes the tab and its conversation.

Working with answers

Answers appear as they are written, with formatting, tables and code highlighting; code blocks have their own copy button. Under each answer:

Control	What it does
Thumbs up / down	Good Response / Bad Response : rates the answer. Ratings are stored with the answer.
Copy	Copies the answer to the clipboard (Copy Response).
Translate	Translates the answer into another language with DeepL.
Statistics (right)	Stats For Nerds : tokens per second and response time of this answer.
Sources	With web search or a data source: the numbered sources the answer used (chapters 6 and 7).

- **Answers keep running.** You can switch to another chat, add a tab or start a new chat while a model is writing; the answer is finished and saved in the background. A progress bar on the tab and a dot in the

sidebar show that it is still running. Reloading the page loses only the live view, not the answer.

- **Stop** – while a model is writing, the send button becomes a red square. Click it to stop the answer.
- **Long conversations** – when a conversation becomes too long for the model, AIman replaces the oldest turns with a short summary instead of silently dropping them, so the model still knows what was agreed at the start.

Projects: sharing a chat with colleagues

A project is shared with the people, teams, organisational units or the company you name under **Share With...** when you create or edit it. Each of them gets a level of access:

- **Read** – can open the project and follow the conversation;
- **Write** – can also send prompts and work in its tabs;
- **Admin** – can also change the project and whom it is shared with.

Projects shared with you appear under **Projects** in the sidebar, marked with an icon: **This project has been shared with you by another person. Click here to view details.** The details show who else has access and with which rights. If you no longer need a project, **Leave Project** removes it from your list; the others keep it. Sharing with a team reaches every member of the team, including people who join later.

Share links, export and print

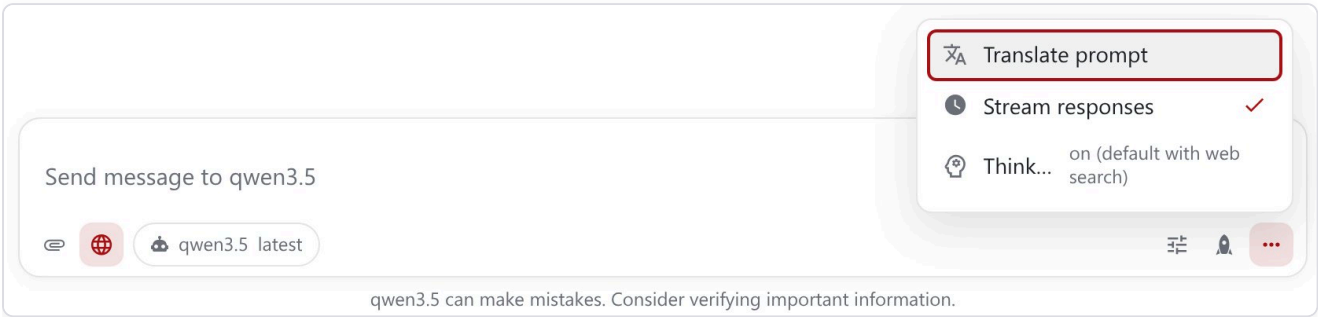
Where your installation enables link sharing, the tab menu offers **Share Chat** and the project header **Share Project**. **Copy URL** copies a link to a read-only page of the conversation. That page offers a download of the conversation as text, HTML, an image (PNG) or JSON, and **Print Chat**. Without link sharing, use **Copy Response** on single answers, or share the project itself.

Pinning, bookmarks and search

Pin chats you use every day: ... on the row → pin. Pinned chats stay at the top of the sidebar and in the slim icon rail. **Bookmarks** (account menu) keeps links to any page in AIman: **Bookmark This Page**. And **Ctrl+K** finds every chat and project by name.

5 The composer

Everything that shapes the next answer sits in the message box at the bottom of the chat.



The composer with web search on and the ... menu open

Control	What it does
Paperclip	Attaches files to the prompt: images for multimodal models, or a document the model should read for this one question.
Globe	Web search on or off for this tab (chapter 6). Highlighted in red when on.
Model chip	Shows the tab's model, for example qwen3.5 latest . Click it to change the model or the data source.
Sliders icon	Prompt Options: temperature and answer length (below).
Rocket icon	Opens and closes the workbench (chapter 8).
...	More options: Translate prompt , Stream responses , Thinking , and with a data source Amount of Chunks and Only from the documents . A small dot on the button shows that something differs from the default.
Arrow / red square	Sends the prompt (Enter); while an answer is being written, stops it.

The placeholder in the box names the model that will answer: **Send message to qwen3.5**. **Shift+Esc** puts the cursor into the composer from anywhere in the chat.

Attaching files and images

Click the paperclip, choose files, and send your prompt. An image is passed to the model itself, so choose a model marked **Multimodal**. A document attached this way is meant for the question at hand. For documents you want to ask about again and again, or share with colleagues, create a data source instead (chapter 7): it is indexed once and searched with every prompt.

Prompt options

The sliders icon opens **Prompt Options** for the current chat (**Fine-tune your prompt options to get the best responses.**):

- **temperature** — lower values make answers more focused and repeatable, higher values more varied and creative;
- **num predict** — the maximum length of the answer, in tokens.

Leave a field empty to use the model's own default; this is recommended for most chats, because every model is tuned differently. Only values you fill in are sent. **Defaults** empties the fields, **Apply** saves them. Some current vendor models no longer accept a temperature; AIman leaves it out for them.

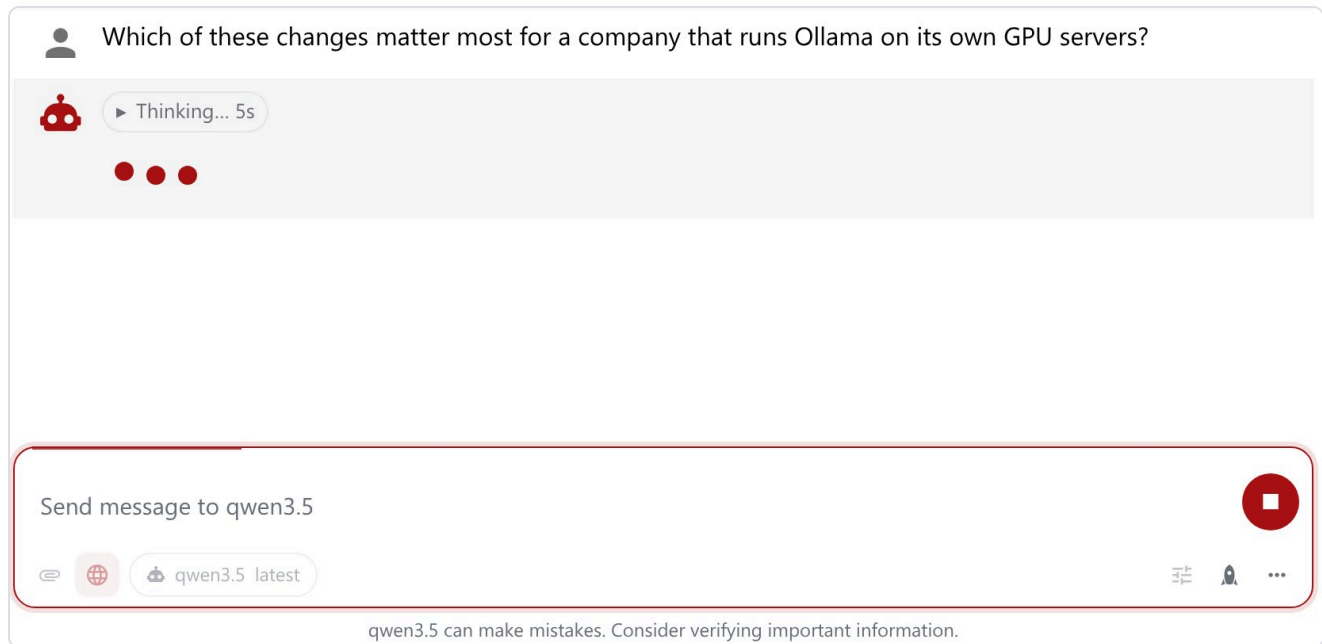
The **Prompt Options** dialog

The “...” menu

Entry	What it does
Translate prompt	Translates your prompt with DeepL before it is sent — useful because many models answer best in English. You choose the target language; the choice applies to the current session unless you save it.
Stream responses	On (ticked): the answer appears word by word. Off: it appears complete when it is finished.
Thinking	Only for models that can reason. The hint shows the current state, for example off - faster (default) or on (default with web search) . Click to switch.
Amount of Chunks	With a data source: how many passages are looked up per question, 1 to 10, default 5 (chapter 7).
Only from the documents	With a data source: ticked (default) answers only from the documents; unticked lets the model add general knowledge, marked as such (chapter 7).

Web search, Thinking and **Only from the documents** are remembered per tab, also after a reload.

Thinking: letting a model reason first



A model thinking before it answers; the red square stops it

Reasoning models can think a problem through step by step before they answer. That takes longer, but helps with calculations, logic, planning and questions with several conditions. With **Thinking** on, the model's reasoning appears live in a folded block with a timer (**Thinking... 5s**); the answer follows below it. Click the block to read the reasoning. The reasoning is shown while you are on the page but is not stored with the answer.

- Thinking is **off by default**, because answers come faster.
- With web search on, it is **on by default** for local models: weighing several web pages benefits from reasoning.
- The **Thinking** entry appears only for models that can reason, for example qwen3.5 and gemma4; mistral-nemo, phi4 and llama3.1 have none.
- If a model thinks but produces no answer, AIman automatically asks it again.

6 Web search

Models only know what they learned up to their training cutoff. Web search grounds an answer in today's pages, with the sources listed.

Switch on web search with the **globe** in the composer (**Web search: use current web results for the answers**). It stays on for this tab until you switch it off. Use it for news, prices, releases, opening hours – anything that changes.



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](#)



An answer from the web: claims numbered [5], the source listed underneath with its site and an **excerpt** link

What happens when you ask

AIMan searches the web for your question – follow-up questions such as “and in Germany?” take the earlier messages into account – reads the best pages and gives the passages that answer the question to the model. While it does so, the chat shows **Searching the web and reading the results...**; this adds a few seconds before the answer starts. How the search is done depends on the model:

Model	How it searches
Claude, Gemini, GPT (OpenAI)	The vendor's own web search, run by the vendor.
Local models that can use tools (for example mistral-nemo, qwen3.5, gemma4)	The model decides itself whether to search, what to search for and which pages to read. A question like “what is 12 × 12?” is answered without searching.
All other models (for example phi4, Azure OpenAI, Cohere, Ollama Cloud)	AIMan's own search: it writes one or two search queries, reads the top pages and ranks their passages before the model answers.

Sources

The answer cites web pages like documents, with numbers such as [3]. Under **Sources** each number links to its page, with the site's name; **excerpt** shows the passage that was used. The sources are stored with the answer. Web search also works together with a data source: the answer can then use both, and documents and pages are numbered without clashes.

Good to know

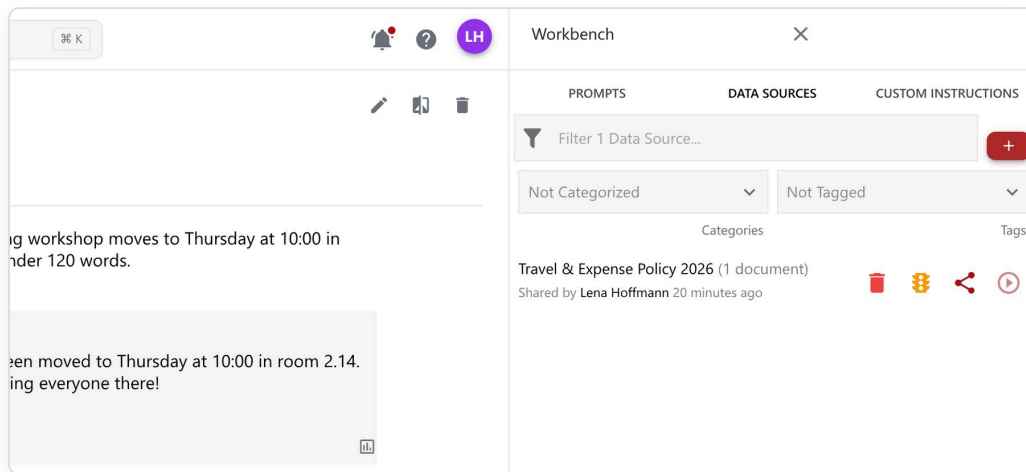
- **Privacy** – your question, or a short search query derived from it, is sent to a search service: with Claude, Gemini and GPT to the vendor's search, otherwise through AIman's own search service to public search engines or a paid search API your company has set up. Do not switch web search on for questions that contain confidential details.
- **Pages can be wrong or out of date.** Check the linked sources for anything important.
- **Web content is information, never instructions.** A page cannot tell the model what to do; AIman only reads public addresses.
- If nothing useful is found, the model says so and answers without current results.
- Admins can switch web search off for the whole installation; the globe then disappears. Chatbots embedded in websites do not offer web search.

7 Data sources: answers from your documents

Attach your own documents, web pages or a company knowledge base to a chat, and every answer is built from the most relevant passages – with document and page cited.

A model normally only knows what it was trained on. With a **data source** attached, AIman first searches the data source for the passages that best match your question, hands them to the model together with the question, and tells it to cite them. You do nothing extra: pick a data source and chat as usual.

Data sources live in the **workbench**, tab **DATA SOURCES**. The list shows each data source with its number of documents and who shared it, and four icons: delete, status, sharing info, and **Use Data Source**.

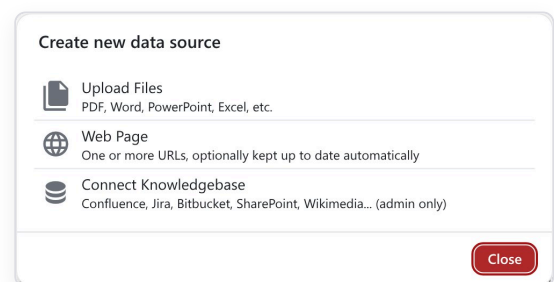


A data source in the workbench: delete, status (yellow: indexing), sharing info and **Use Data Source**. On the left, the chat header icons: edit, **Prompt All Tabs**, delete

Creating a data source

Click **+** in the workbench's data source tab. **Create new data source** offers three kinds:

- **Upload Files** – PDF, Word, PowerPoint, Excel, etc.
- **Web Page** – one or more URLs, optionally kept up to date automatically.
- **Connect Knowledgebase** – Confluence, Jira or Bitbucket; for admins only.



Choosing the kind of data source

Uploading files

- 1 Enter a **Name** (required) and a short **Summary** of what the documents contain.
- 2 Drop files onto the red area, or click **Browse**. Each file appears as a card; the **x** removes it again.
- 3 Optionally choose whom to share it with (**Share Data Source With...**) and add **TAGS** and **CATEGORIES** to find it again.
- 4 Click **Create**. AIman reads and indexes the documents; the status changes from **Pending** and **Indexing** to **Ready**, and a notification tells you when it is done.

A data source with one PDF, ready to be created

Format	File types
PDF	.pdf (with a text layer)
Word	.doc, .docx
PowerPoint	.ppt, .pptx
Excel and CSV	.xls, .xlsx, .csv
Text and Markdown	.txt, .md
Web page	.html

Tables stay intact: a table, from a spreadsheet or inside another document, is kept together as one passage, so its header row is never separated from its data.

Advanced settings

Advanced Settings holds **Chunk Size** and **Chunk Overlap** (how long the passages are and how much they overlap) and **Use Contextual Retrieval Augmented Generation**. With contextual indexing on, a model first writes a short note for every passage that places it in the whole document ("this is from the pricing section of the Q3 report..."). Questions then find the right passages more reliably in long, dense or ambiguous documents such as contracts, reports and manuals. Indexing takes longer. Leave it off for short, self-contained documents.

Web pages

Connect Web Page takes a **Name** and one or more URLs: type each into **Add New URI Data Source...** and click **+**. AIman fetches each page, turns it into readable text and indexes it like a document; citations link to the page.

Switch on **Keep this up to date automatically** and set **Recrawl Every (Days)** (at least one day) to have the pages read again on a schedule. The data source stays usable while it is being refreshed.

The **Connect Web Page** dialog

Knowledge bases: Confluence, Jira, Bitbucket

A knowledge base connects AIman directly to a system your company already runs, instead of uploading exports. Administrators set it up once with an access token; it is then shared like any other data source – usually with the whole company. Until it is shared it is visible only to the admin who created it.

System	Editions	What is indexed
Confluence	Cloud, and Server / Data Center 7.9 or later	The pages of one or more spaces (Space Key). Attachments and images are not read.
Jira	Server / Data Center 8.14 or later	The issues of one or more projects (Project Key(s)).
Bitbucket	Server / Data Center 5.5 or later	One or more repositories (Project/Repository Key(s) , for example PROJ/my - repo).

The admin enters the **Base URL**, chooses the **Authentication (Personal Access Token**, or for Confluence Cloud an **Account Email** with an API token) and clicks **Connect**; the first crawl starts at once. With **Keep this up to date automatically** the content is read again every few days, as set. SharePoint and Wikimedia are shown in the dialog but marked **Coming soon**.

NOTE

Answers from a knowledge base reflect what was there at the last crawl. The stored token is never shown again; leave the field empty when editing to keep it. Jira Cloud and Bitbucket Cloud are not supported.

Status

Status	Icon	Meaning
Pending	grey	Waiting to be processed.
Indexing	yellow	The documents are being read, split into passages and indexed.
Ready	green	Can be used in chats. Use Data Source is only active now.
Failed	red	Processing failed, for example for an unreadable or password-protected file. Point at the icon to see why; you also get a notification.

Asking questions about your documents

There are three ways to attach a data source:

- **In the workbench** – **Use Data Source** on a ready data source. Choose **Set as default Data Source for the current project or chat**, or **Set to individual Tab(s) from the current project or chat** and tick the tabs. A project default is offered to the tabs; to use it in a tab, apply it to that tab.
- **In the composer** – click the model chip and choose the **Data Source**.
- **When creating** a chat or project – **Advanced Settings** → **Data Source**.

Then ask as usual. The answer marks what it took from the documents with numbers such as [1], [2], and lists the sources underneath: document and, for PDFs and other paged formats, the page. Several passages from the same page share one number.

Screenshot to follow: answer from a data source with citations

An answer drawn from an uploaded document, with numbered references in the text and the document and page listed under
Sources

Only from the documents – or general knowledge too

By default answers come **only from the attached documents**. If you want the model to add what it knows in general – for example to explain a term the documents use without defining it – untick **Only from the documents** in the composer's ... menu. The model then uses the documents first and marks everything else as general knowledge (*From general knowledge: ...*). The setting is remembered per tab.

When nothing relevant is found

If none of the passages really answers your question, AIMan says so instead of letting the model guess: **No relevant content was found in the attached data source for this question**. Rephrase the question, attach another data source, untick **Only from the documents**, or remove the data source if you meant a general question.

How the right passages are found

- **Hybrid search** – every question is matched both by keywords (exact terms, names, codes) and by meaning (paraphrases, your own words), and the two are combined.
- **Reranking** – a second model reads your question together with each candidate passage and keeps only the best ones. Like all document processing, this runs on the servers of your AIMan installation.
- **Follow-up questions** – “And who is responsible for it?” means little on its own, so AIMan rewrites it with the help of the last messages into a complete question before it searches. The model still answers your own wording.
- **Amount of Chunks** (composer, ...) sets how many passages are passed on: 1 to 10, default 5. More passages help with broad questions (“summarise...”), fewer keep answers focused.

Sharing data sources, and current limits

Share a data source when you create or edit it, under **Share Data Source With...**: with colleagues, teams, organisational units or the company, to read, write or administer. With read access, colleagues can use it in their chats. With write access they can also add or remove documents and start re-indexing, but not change its name, summary or sharing (**You have read and write access to this data source...**). Only admins of a data source can delete it; deleting removes all its documents and shares.

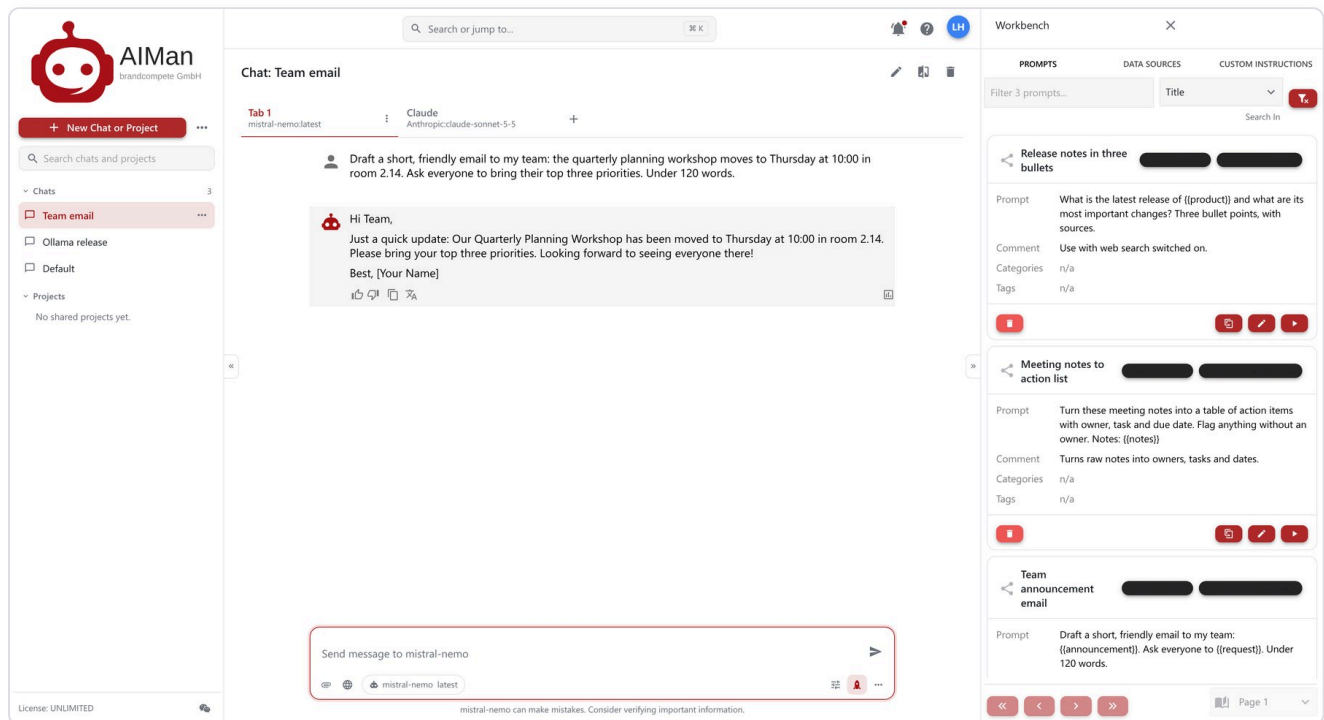
NOT AVAILABLE YET

Scanned documents are not read yet. AIMan reads the text layer of a document; there is no text recognition (OCR) yet. A scanned PDF without a text layer, or a photo of a document, cannot be searched; the data source reports that OCR is not supported. Images inside documents and Confluence attachments are not read either. Text recognition is on the roadmap.

8 Workbench, prompt repository and custom instructions

The workbench keeps what you reuse next to every chat: saved prompts, data sources and standing instructions.

Open the workbench with the handle on the right edge of the window, the rocket icon in the composer, or **Ctrl+K** → “Show/hide prompt repository”. On wide screens it sits next to the chat; on smaller screens it slides over it. It has three tabs: **PROMPTS**, **DATA SOURCES** (chapter 7) and **CUSTOM INSTRUCTIONS**.



The prompt repository in the workbench, with placeholders in the prompts

The prompt repository

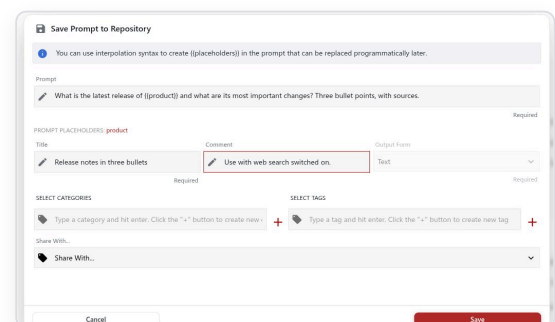
A good prompt is worth keeping. Each saved prompt has a title, the prompt text, a comment, categories and tags, and its sharing. Filter the list by text (**Filter ... prompts...**) and choose what to search in (title, prompt, comment...). The buttons on each card delete, copy and edit the prompt, and the ► button runs it in the current chat.

Saving a prompt

Save a prompt you have sent, or create a new one in the repository. In **Save Prompt to Repository**, give it a **Title** (required), an optional **Comment** – for example “Use with web search switched on.” – categories, tags, and whom to share it with (**Share With...**).

Placeholders

Write the parts that change as placeholders in double curly braces: `What is the latest release of {{product}}...`. AIman lists them under **PROMPT PLACEHOLDERS**. When you



Saving a prompt with the placeholder `{{product}}`

run the prompt, you fill in each placeholder before it is sent.
One prompt can serve every product, customer or document.

TIP

Share your best prompts with your team. Colleagues then start from what already works instead of rediscovering it – and everybody's prompts improve together.

Custom instructions

Custom instructions are standing context that the model receives with every prompt, so you do not have to repeat it. There are two types (**Instruction Type**):

- **What would you like AIMan to know about you?** – your role, your department, your customers, terms you use;
- **How would you like AIMan to respond?** – tone, length, format, language, for example “answer in British English, in short paragraphs, with a summary first”.

Create one in the workbench (**CUSTOM INSTRUCTIONS** → +, **Create New Custom Instruction**) with an **Instruction Name**, a description and the **Instruction Content**. Then use it for a project or a tab: in the **New Chat or Project** dialog under **Advanced Settings**, or by editing a tab. Like prompts and data sources, custom instructions can be shared with colleagues, teams or the company.

Your custom instructions are added after the company's **assistant instructions**, which your administrators set for every model (chapter 12).

9 Chatbots for websites and applications

Any chat tab – its model, data source and instructions – can become a chatbot on an intranet page or in an application.

Exporting a tab as a chatbot

- 1 Prepare a tab with the model, the data source and the custom instructions the chatbot should use, and test it with a few questions.
- 2 Open the tab menu (⋮) and choose **Export To Chatbot**. Confirm with **Export**.
- 3 Set it up in the editor (table below) and click **Get Code**.
- 4 **Copy Embed Code** and paste it into your page, at the bottom, right before the closing `</body>` tag.

Setting	What it does
Title, Description	The text in the chatbot's title bar; the description is only for you, in Chatbots Management.
Color, Background Color, Color Presets, Dark Theme	Colours to match your website.
Chatbot Width, Messages Area Height, Position, Full Screen	Size in pixels and left or right on the screen – or the whole page, for pages that show only the chatbot.
Start Minimized, Minimize in Icon	Starts folded, as a button or an icon, until a visitor opens it.
Realtime Streaming, Sample Prompts	Answers appear word by word; example questions visitors can click before they start.
Clearable, Downloadable, Play Sound	Visitors may clear the chat, download it as an HTML page; a sound plays on a new message.
Language of the UI	The language of the chatbot's own buttons and labels.

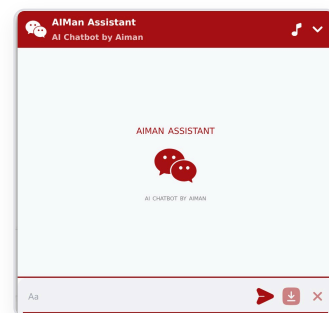
Chatbots Management

Chatbots Management (from your profile, under your shortcuts and tools) lists the chatbots you have created: **View, update or delete the custom chatbots you have created**. Updating a chatbot's settings does not change its look on websites where it is already embedded; deleting it stops it working everywhere.

The AIMan Assistant

AIMan uses its own chatbot technology: the **AIMan Assistant** in your account menu opens a chatbot window inside AIMan, at the bottom right. It looks the way an exported chatbot looks on your own website: a title bar, the conversation, and a message field (**Aa**) with send, download and close.

Chatbots do not offer web search. A chatbot answers with the model, data source and instructions of the exported tab. Chatbot conversations are archived automatically after a period your administrators set.

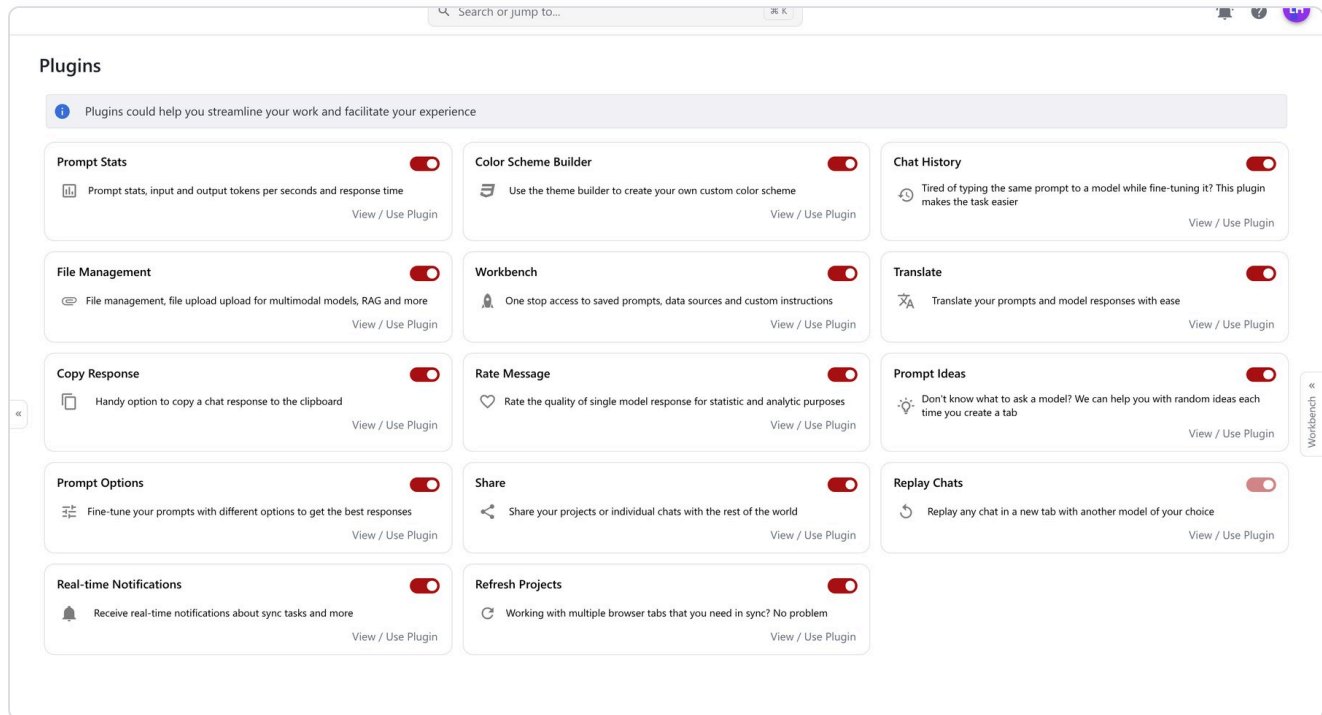


The AIMan Assistant, an embedded chatbot

10 Plugins

Many features are plugins. Switch on what you use, and the interface stays as simple as your work.

Open **Plugins** from your account menu. Each card has a switch and **View / Use Plugin**. Switching a plugin off removes its buttons and menu entries for you; it does not affect your colleagues.



The **Plugins** page

Plugin	What it adds
Prompt Stats	Input and output tokens per second and response time under each answer (Stats For Nerds).
Color Scheme Builder	Your own colour scheme for AIman (chapter 11).
Chat History	Reuse prompts you have sent in a tab.
File Management	File upload for multimodal models, data sources and more.
Workbench	The panel with saved prompts, data sources and custom instructions.
Translate	Translation of prompts and answers with DeepL.
Copy Response	Copies an answer to the clipboard.
Rate Message	Thumbs up and down for each answer.
Prompt Ideas	Random prompt suggestions on the start page and in new tabs.
Prompt Options	Temperature and answer length per chat.
Share	Share links for chats and projects, where enabled.
Replay Chats	Replay a chat in a new tab with another model.
Real-time Notifications	Notifications as they happen, for example when a data source is ready or has failed.
Refresh Projects	Keeps several browser tabs with AIman in sync.

11 Your preferences

Your profile, your start page, your colours and the shortcuts that save time.

Profile

Account menu → **Profile**. Under **Account Details** you change your name and photo (**Upload Photo**) and your password (**Change Password**). **Your AIman Shortcuts and Tools** leads to your startup options, bookmarks, chatbots and personal service tokens.

Startup options

Startup Options – Choose how to start AIman.

Option	Where you land after signing in
Start Page	Start a new chat right away, see your recent chats and jump to models, profile, help and more (the default).
Your Chats and Projects	The overview of your chats and projects.
New Project	A new chat with a model you choose, created automatically if no default exists.
Your Profile	Straight to your profile.
Custom URI	Any page in AIman, for example one particular project. If you delete that project later, remove it here too.

Themes and colours

AIman has a light and a dark theme; switch them in the account menu (**Set dark or light theme**). Your company's colour and logo are set by your administrators. With the **Color Scheme Builder** plugin you can create a colour scheme of your own (**Theme Builder**, with a preview for dark and light); it overrides the company colours for you only. **Set Default Theme** returns to the company's.

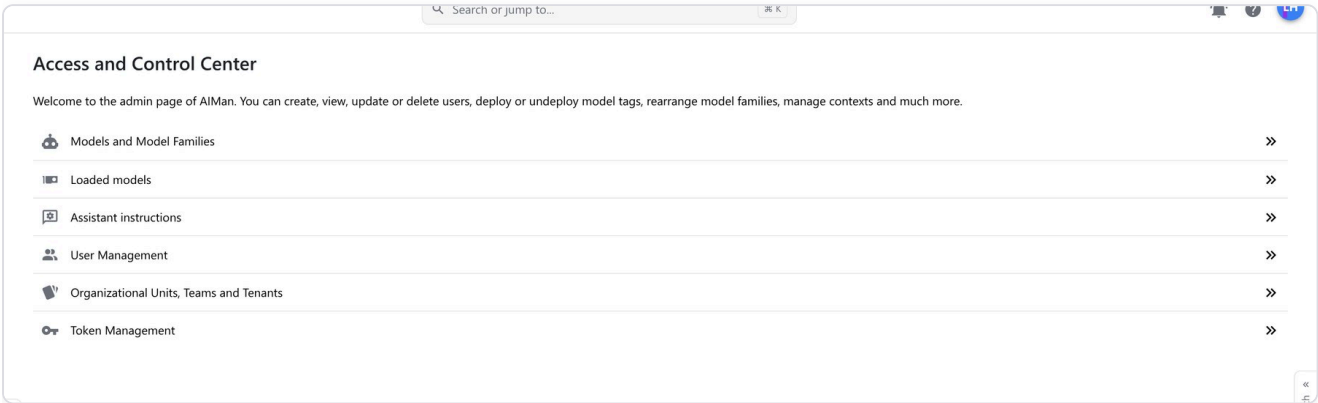
Working faster

- **Command palette** – **Ctrl+K** (**⌘K**) searches chats, projects, saved prompts, models and pages, and runs actions. The fastest way to anything.
- **Focus mode** – **Ctrl+Shift+.** hides the header, sidebar and workbench for long chats where you only need the conversation. Leave it with **Esc**, the same shortcut, or **Exit focus mode** at the top right. Everything comes back as it was.
- **Icon rail** – **Ctrl+Shift+S** folds the sidebar to a slim rail that keeps “new chat”, search and pinned chats in reach.
- **Keyboard shortcuts** – **Ctrl+V** shows them all; the full list is in Appendix A.
- **Feature tour** – **Start Feature Tour** in the help menu walks you through the window.
- **Notifications** – the bell in the header collects messages, for example when a data source is ready.

12 Administration

The Access and Control Center is where administrators decide which models people see, who may do what, and which keys every call runs under.

Open it from the account menu: **Access and Control Center**. You see only the tiles your role allows:



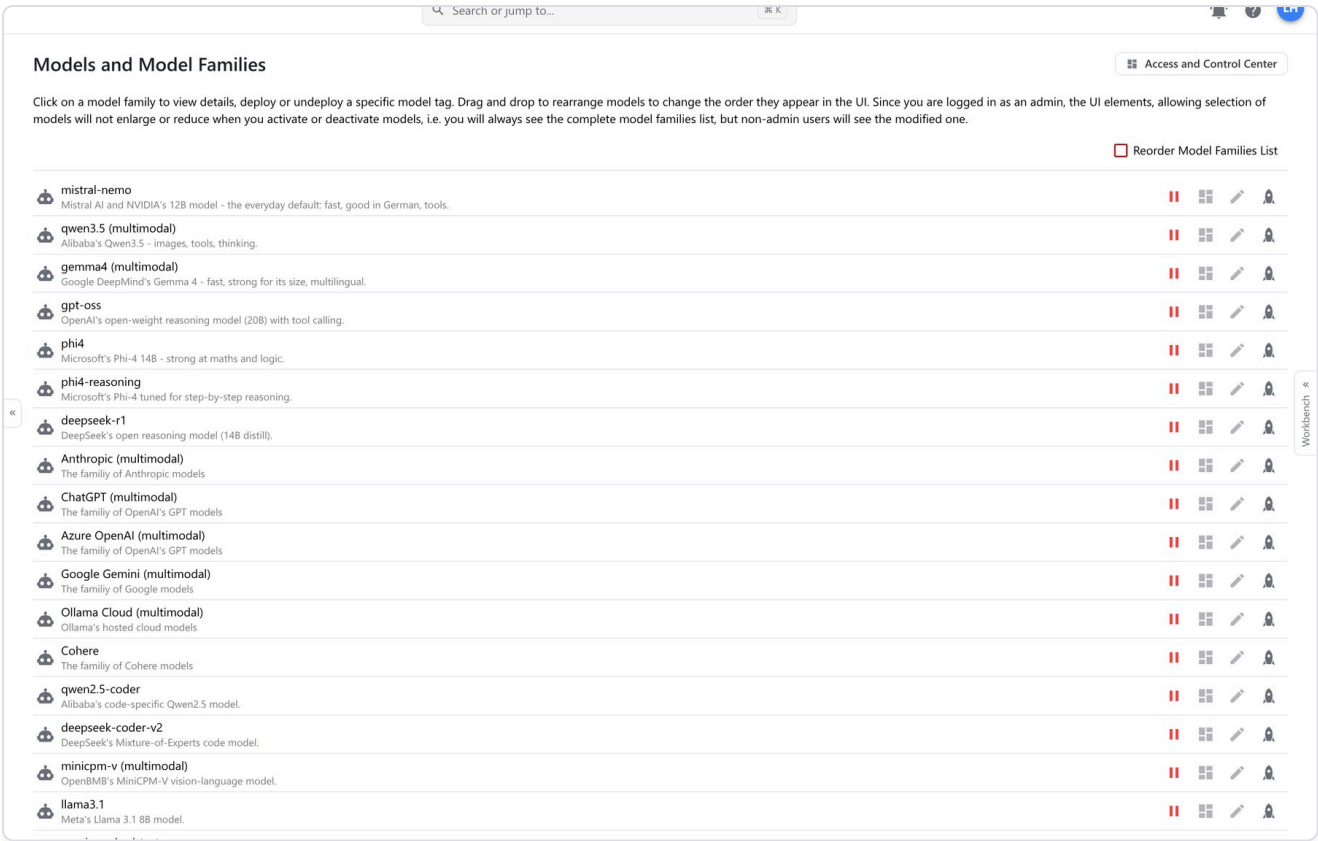
The Access and Control Center

Tile	Who sees it
Models and Model Families	Superusers
Loaded models	Roles allowed to deploy models
Assistant instructions	Superusers
User Management	Roles allowed to manage users
Organizational Units, Teams and Tenants	Roles allowed to manage contexts
Token Management	Roles allowed to manage tokens

Models, loaded models and assistant instructions

Models and Model Families

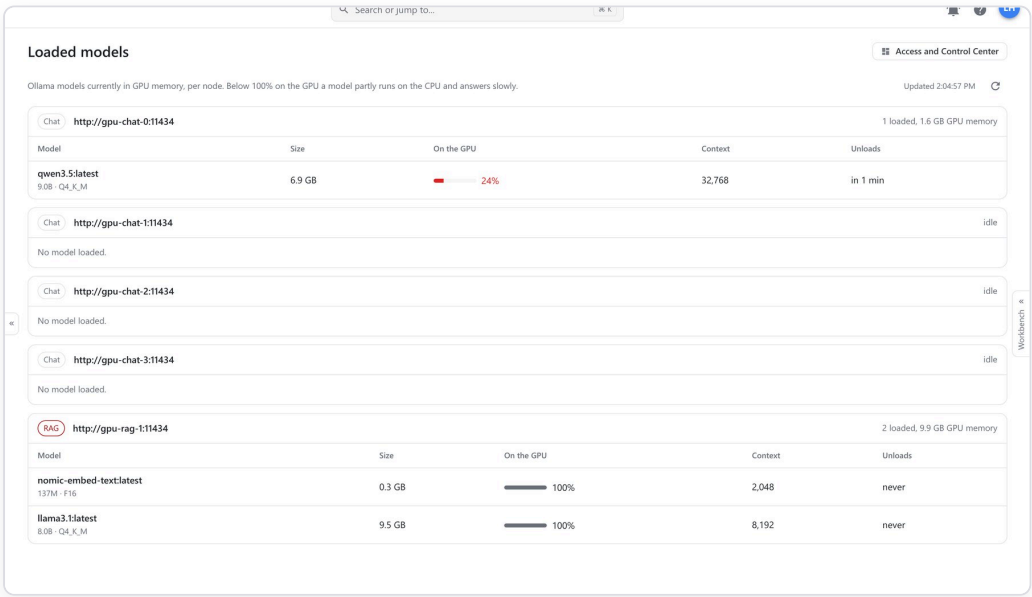
The list of every model family, in the order the picker shows them. Click a family to see its versions (tags) and deploy or undeploy each one; the icons on the right activate or deactivate a family, edit its description (the line users see in the picker and the model help), and deploy. Tick **Reorder Model Families List** to drag families into a new order. As an admin you always see the complete list; other users see only active families.



Models and Model Families

Loaded models

Shows, per GPU server, which open models are currently in GPU memory: size, the share on the GPU, the context and when the model will be unloaded. Below 100% on the GPU (shown in red) a model partly runs on the processor and answers slowly. Chat servers and the server reserved for document processing (marked **RAG**) are listed separately. The page refreshes every 15 seconds.



Loaded models per GPU server

Assistant instructions

Every model receives the same short base instruction: that it is the AI assistant of your company in brandcompete AI Manager, today's date, its training cutoff, the answer language, and – with documents or web

results — how to use and cite them. Under **Addition for this AIMan** superusers add what every model should know as well: company context, tone, terms to use (up to 4,000 characters). Click **Save**; it applies from the next prompt, no restart needed. Users' own custom instructions come after it.

The screenshot shows the 'Assistant instructions' form. At the top, there's a search bar and a title 'Assistant instructions'. Below the title, a paragraph explains that every model gets a base instruction and that users can add custom instructions. The form is divided into two sections: 'Base instruction (always sent)' and 'Addition for this AIMan'. The 'Base instruction' section contains a text box with a pre-filled instruction: "You are the AI assistant for <company> in brandcompet 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. The 'Addition for this AIMan' section has a text box with an example: "e.g. 'We are a media agency. Answer in a friendly, concise tone and use British English for English answers.'" At the bottom right, there is a red 'Save' button. A 'Workbench' sidebar is visible on the right edge.

Assistant instructions

Users, roles and organisational units

User Management

Create, view, delete, modify or change user roles. Add User creates an account; click a row to edit it. Every account has exactly one role (Appendix A), and per account you decide whether it may use the user interface, the API, or both — a service account for an application needs no chat window. People who sign in with single sign-on for the first time appear here as **Guests**; give them the role they need. A role change applies with the person's next action, no new sign-in required. The role **Nobody** keeps an account but allows nothing — useful when someone leaves.

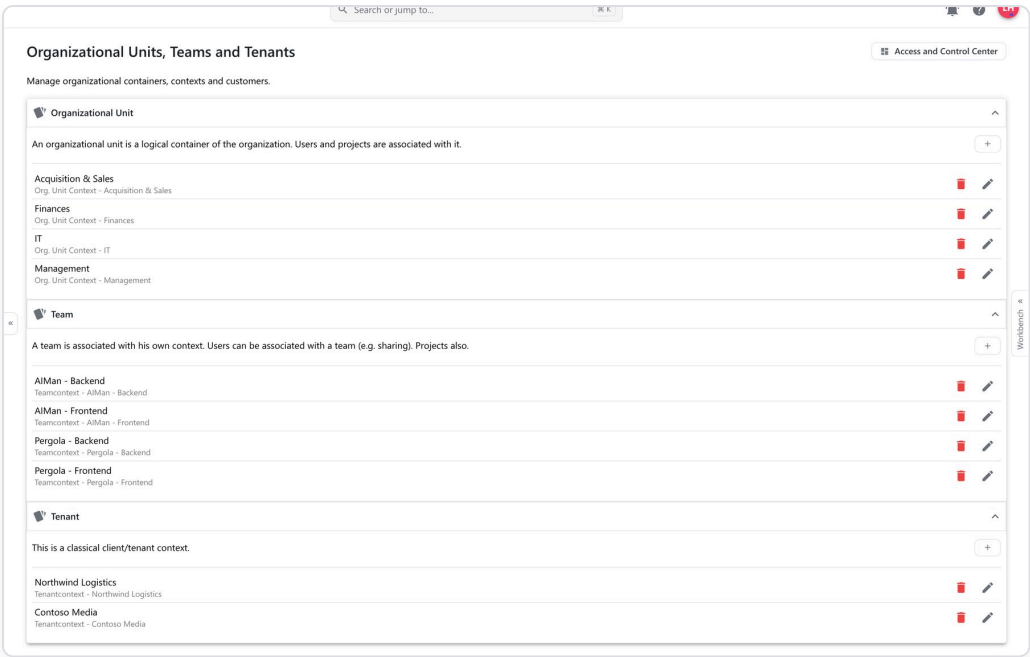
The screenshot shows the 'User Management' table. At the top, there's a search bar and a title 'User Management'. Below the title, a paragraph explains that users can create, view, delete, modify or change user roles. The table has columns: NAME, ID, UUID, CREATED, MODIFIED, and ACTIONS. The table contains 8 rows of user data. The first two rows have a green and orange circle icon respectively. The last two rows have a red and blue circle icon respectively. A 'Workbench' sidebar is visible on the right edge.

NAME	ID	UUID	CREATED	MODIFIED	ACTIONS
[blurred]	[blurred]	6f77c7ca-c7d-d48b-92f9-14f6b675bb7a	13/02/24	13/02/24	[trash] [edit]
[blurred]	[blurred]	67610307-55c3-45c3-8411-d03c9372d8e3	13/02/24	13/02/24	[edit]
[blurred]	[blurred]	6349c424-23f9-44fa-aa79-80fb2bde101c	05/02/25	05/02/25	[edit]
[blurred]	[blurred]	ecc2933f-d5cd-4cb5-af13-396c06ab3e8e	05/02/25	05/02/25	[trash] [edit]
[blurred]	[blurred]	42dfd117-53ae-4d09-8468-483b4b5dae84	05/02/25	05/02/25	[trash] [edit]
[blurred]	[blurred]	85f56cf3-42f6-4d30-a7ce-a8a03970eae7	05/02/25	05/02/25	[trash] [edit]
[blurred]	[blurred]	5ec6d701-7e2e-4036-be19-23e0b2d4da9f	05/02/25	11/10/26	[trash] [edit]
[blurred]	[blurred]	c481a7ad-d467-4019-b4e5-dcd0ca482775	05/02/25	11/10/26	[trash] [edit]

User Management (names blurred)

Organizational Units, Teams and Tenants

The groups that things are shared with. An **organisational unit** is a part of the organisation, such as a department; a **team** is a group of people; a **tenant** is a client context. Create one with **+**, then associate users and projects with it. Whatever is shared with a unit or a team reaches all its members — including people who join later — and someone who reaches an object in several ways keeps the highest access of any.



Organizational Units, Teams and Tenants

Keys, knowledge bases, branding and sign-in

Token Management

The vendor keys every call runs under: Anthropic, Azure OpenAI, Cohere, DeepL (**Deepl.io**), Google Gemini, Ollama Cloud, OpenAI, and the web search services Brave Search, Serper / Google and Tavily. Expand a service and enter its key; it takes effect without a restart. Keys at **scope: company** serve everyone, so nobody needs a key of their own; personal keys are managed from each person's profile. Values stay hidden unless you tick **Show token values as plain text**. A vendor without a key cannot answer.



Token Management

TIP

A paid search API key (Tavily, Brave or Serper) noticeably improves current-events answers from local models compared with the free public search engines.

Knowledge bases

Only roles allowed to manage knowledge bases see **Connect Knowledgebase** in the data source dialog (chapter 7). Create the connection, then share it with the **Company** context so that everyone can use it: **This Knowledgebase is only visible to you until you share it below...** Use an access token of an account that may read exactly what everybody may see.

Branding

Each installation carries the company's colour and its logos for the light and the dark theme, chosen by the address people open; the sign-in page shows them too. A company can serve AIMan on its own domain. Branding is set up by brandcompete when the installation is created; ask us for changes.

Single sign-on and web search

Single sign-on with Microsoft Entra ID and Google is configured per installation: the allowed email domains (an empty list lets nobody in), the Entra tenant IDs and the Google domains. First-time users become Guests. Web search can be switched off for the whole installation. Both are set by whoever operates the installation – brandcompete in the managed service.

13 Using the API

Applications use the same models, data sources and keys as people, through a REST API with streaming answers.

AIMan's REST API lives under `https://<your AIMan API address>/api/v1`. It speaks JSON, authenticates with bearer tokens, and lets you call any deployed model by name, so an application can switch models without code changes. The OpenAPI description ships with the API (`/api/v1/spec.html`).

IMPORTANT

AIMan's API is its own REST API. It does **not** offer an OpenAI-compatible endpoint (`/v1/chat/completions`); tools that expect one cannot be pointed at AIMan directly.

An account with API access

An administrator creates an account and allows it the API (User Management, chapter 12); for a service account, allow the API only. Its role decides what it may do, exactly as for people.

1. Sign in

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

```
{ "userName": "service@yourcompany.com", "userPassword": "..." }
```

The answer contains `messageContent.data.access_token` and a `refresh_token`. Send the access token with every request as `Authorization: Bearer <token>`. It is valid for a limited time; `POST /api/v1/auth/refresh` with the refresh token in the `Authorization` header returns a new pair, and `/auth/logout` revokes them.

2. Find the model

`GET /api/v1/models` lists the models the account may use. Each has a `fullName` such as `mistral-nemo:latest` or `Anthropic:claude-sonnet-5-5` — the name you prompt with. `GET /api/v1/models/{modelId}/tags` lists a family's versions.

3. Prompt a model by name

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

```
{ "prompt": "Summarise our travel policy in three bullet points.",
  "datasourceId": 42 }
```

The name is `<family>:<tag>`, not case-sensitive, or just the family for its default version. An unknown name returns 404, an ambiguous one 400. The answer text is in `messageContent.data.responseText`.

Field	Meaning
<code>prompt</code>	The text to send (required).

Field	Meaning
<code>datasourceId</code> , <code>amountOfChunks</code>	A data source to answer from, and how many passages (1–10, default 5).
<code>ragAssisted</code>	<code>true</code> allows general knowledge beside the documents (default <code>false</code> : only the documents).
<code>webSearch</code>	<code>true</code> grounds the answer in current web results.
<code>think</code>	<code>true</code> / <code>false</code> for models that can reason; omit it for AIman's default.
<code>attachments</code>	Files, for example images for multimodal models, base64-encoded.
<code>custInstWhoId</code> , <code>custInstRespondId</code>	Custom instructions to apply.
<code>options</code>	Prompt options such as <code>temperature</code> and <code>num_predict</code> ; only what you set is sent.

4. Streaming

`POST /api/v1/prompts/models/{modelName}/stream` takes the same body and answers as server-sent events (`text/event-stream`). Events named `Response` carry the answer piece by piece and end with `[STREAM-DONE]` ; `Thinking` carries a reasoning model's thoughts, `Status` progress such as a running web search, `Citations` the sources, and `Error` a failure. A stream can be stopped with `POST /api/v1/prompts/streams/{streamId}/cancel` when you sent a `streamId` with the prompt.

NOTE

Calls by model id instead of name (`/api/v1/prompts/{modelTagId}`), translation (`/api/v1/tools/translate/...`) and the other resources are described in the OpenAPI specification.

14 Data and privacy: what goes where

AIMan makes the choice visible: local and vendor models are separate tabs in the picker. This chapter says exactly what leaves the servers of your AIMan installation, and when.

What you do	Where the data is processed
Chat with a local model	Prompt, answer and any document passages stay on the dedicated GPU servers of your installation – brandcompete's servers in Germany, or your own on-premise.
Chat with a vendor model	The prompt, the conversation and the passages the answer needs are sent to that vendor (Anthropic, OpenAI, Microsoft Azure, Google, Cohere or Ollama Cloud), under your company's contract and key.
Create a data source	Reading, splitting, indexing, embeddings and reranking always run on the servers of your installation, whichever model later answers. Knowledge bases are read from your company's own Confluence, Jira or Bitbucket.
Translate	The text to translate goes to DeepL.
Web search	A search query derived from your question goes to a search service: the vendor's own search for Claude, Gemini and GPT; otherwise AIMan's search service, which asks public search engines or a paid search API your company has set up.

- **One installation per customer.** Each customer has its own database, search index, queue and network. In the managed service only the GPUs are shared.
- **Private until shared.** Chats, data sources, prompts and instructions are visible only to you until you share them.
- **Keys held centrally.** Vendor keys are stored once per company and used on the server; nobody needs a private account.
- **Sanitised answers.** Model output is cleaned before it is shown, so a crafted answer cannot run code in your browser.

IMPORTANT

Rules of thumb: use a local model for confidential material; do not switch web search on for questions with confidential details; and follow your company's policy on which data may go to which vendor.

15 Troubleshooting & FAQ

The questions people ask most, and what to do.

Problem or question	What to do
I cannot sign in with Microsoft or Google.	Read the message on the sign-in page. Usually your email domain or directory is not on the company's allow-list, or your account was switched off. Ask your administrator.
I only have the role Guest.	First-time single sign-on users start as Guests. Ask your administrator for the role you need.
A model I expect is missing or greyed out.	Your administrators decide which families and versions are offered; n/a means not deployed. A vendor model also needs a company key.
The model says it cannot know something recent.	That is intended: it knows today's date but not what happened after its training. Switch on web search.
An answer takes long to start.	Web search adds a few seconds, Thinking more. A local model that has not been used for a while is loaded into GPU memory first. Vendor models depend on the vendor.
The answer stopped in the middle.	The answer length may be limited (Prompt Options , num predict), or Stop was pressed. Ask "please continue".
"No relevant content was found..."	The documents do not cover the question. Rephrase it, increase Amount of Chunks , attach another data source, or untick Only from the documents .
My data source stays at Indexing, or shows Failed.	Large files and contextual indexing take a while. On Failed, point at the status icon for the reason – for example a password-protected file. Scanned PDFs without text cannot be read (no OCR yet).
Use Data Source is greyed out.	Only ready data sources can be attached. Wait for the green status.
The answer is in the wrong language.	Models answer in the language of your question. Write in the language you want, or say so ("answer in German"), or set it in a custom instruction.
I closed the browser while an answer was running.	The answer is finished and saved in the background; open the chat again.
The Thinking entry is missing.	The current model cannot reason. Choose for example qwen3.5 or gemma4.
The globe for web search is missing.	Web search is switched off for your installation, or you are in an embedded chatbot.
I want to compare models.	Add tabs with different models and use Prompt All Tabs , or replay a chat with another model.
Can AIman create images or speak? Is there a German interface?	No image generation and no voice input or output: AIman reads images with multimodal models. The interface is in English for now; models answer in your language.
Where do I see our usage and costs?	Usage statistics are not available yet. Vendor usage appears on your company's bills from the vendors.

Problem or question	What to do
Who do I contact?	Your administrator first; for the product, brandcompete at sales@brandcompete.com.

A Reference

Roles, statuses, keyboard shortcuts and the terms used in this guide.

Roles

Every account has one role. Administrators assign roles, not individual rights. What a role allows is checked on every action.

Role	Typically for
Superuser	Everything, including models, assistant instructions and every admin area.
Administrator	Running AIMan for the company: users, organisational units and teams, keys, models.
Teamlead	Everything a Staffmember can do, plus creating teams and other groups to share with.
Staffmember	The everyday role: chats, projects, data sources, prompts, custom instructions and chatbots.
Guest	A limited start, for example after the first single sign-on.
Nobody	The account exists but may do nothing – for people who have left.
Prompt Repo User	Reserved for prompt repository access; ask your administrator before using it.
Modelcrawler	A technical account that keeps the model catalogue up to date; not for people.

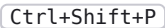
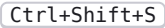
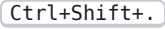
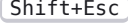
On top of the role, sharing gives access to single objects: **Read**, **Write** or **Admin**.

Data source statuses

Status	Meaning
Pending (grey)	Waiting to be processed.
Indexing (yellow)	Being read and indexed.
Ready (green)	Can be attached to chats.
Failed (red)	Processing failed; the icon shows why.

Keyboard shortcuts

On a Mac use  instead of .

Shortcut	What it does
	Search chats, projects, prompts, models and commands
	Create a new chat or project
	Collapse or expand the sidebar
	Focus mode: hide header and side panels ( to leave)
	Add a new tab to the current chat or project
	Focus the chat input of the current tab
	Show the available shortcuts
	Send the prompt

Glossary

Term	Meaning
Access and Control Center	The administration area.
Assistant instructions	The company-wide text every model receives, set by superusers.
Chunk	A passage of a document, the unit a data source is searched in.
Citation	A number such as [2] in an answer, pointing to a document page or web page under Sources .
Command palette	The search in the header (Ctrl+K).
Context	A group things are shared with: a person, team, organisational unit, tenant or the company.
Contextual indexing	A short note written per passage that places it in its document, for better search results.
Custom instruction	Your standing context for the model: who you are, how it should answer.
Family / tag	A model family (for example <i>Anthropic</i>) and one of its concrete versions (for example <i>claude-sonnet-5-5</i>).
Knowledge base	A data source connected to Confluence, Jira or Bitbucket.
Local model	An open model on the GPU servers of your installation.
Multimodal	A model that understands images as well as text.
Project	A shared chat.
Prompt	What you send to a model.
Prompt All Tabs	One prompt sent to every tab of a chat at once.
RAG	Retrieval-augmented generation: answering from passages found in a data source.
Reranking	A second pass that keeps only the passages that really answer the question.
Streaming	The answer appears word by word as it is written.
Thinking	A reasoning model working through the problem before it answers.
Token	(1) A piece of a word, the unit models count text in. (2) A key or access token for a vendor or the API.
Vendor model	A hosted model of Anthropic, OpenAI, Azure OpenAI, Google, Cohere or Ollama Cloud.
Workbench	The right-hand panel with prompts, data sources and custom instructions.

Questions or feedback: sales@brandcompete.com • [brandcompete GmbH](#) • www.aiman.rocks • www.brandcompete.com