Go GenAI Studio

User Guide

Go Reply
Bartholomäusweg 26 - 33334 Gütersloh - Deutschland - tel +49 5241 5009-0 - fax +49 5241 5009-1099
go@reply.de

www.reply.de

STATE OF THE DOCUMENT

Review	Date	Changes	(Approved by)
0.1	16.09.2025	Initial creation	Andrej Maya
0.2	19.09.2025	Glossary added	Andrej Maya

CHANGES SUMMARY

Main changes list compared to the previous version:	

CONTENTS

1 Introduction

1.1 Purpose

This document is the official user guide for "Go GenAI Studio," a generative AI chatbot application from Go Reply GmbH. Its purpose is to provide you, the end-user, with clear instructions on how to use the application's features and functionalities effectively. By following this guide, you will learn how to navigate the interface, perform key tasks, and get the most out of the software.

1.2 About Go Reply GmbH

Go Reply GmbH is a Premier Google Cloud Partner specializing in consulting and IT services in the DACH region. With a focus on generative AI (GenAI) solutions, the company supports its customers in leveraging cutting-edge technologies to enhance their business processes and user experiences.

1.3 Product Overview

Go GenAI Studio is a chatbot application based on the open-source solution LibreChat.io (MIT licensed). It integrates seamlessly with various GenAI models, including Google's Gemini, Anthropic, and Imagine, as well as OpenAI's GPT models. The solution is tailored for deployment in dedicated customer environments on the Google Cloud Platform.

2 Key Features

2.1 Chatbot Functionality

Go GenAI Studio offers a rich set of features designed to empower users in creating and deploying sophisticated, AI-driven chatbot solutions. These functionalities are geared towards providing a seamless and intuitive experience, while also offering deep customization options to meet diverse business needs.

Core Chatbot Functionalities:

- **Multi-Model Support:** Go GenAI Studio provides access to a variety of powerful language models, allowing users to select the best fit for their specific needs and budget. This includes:
  - **Google's Gemini:** Leverage Google's cutting-edge, multimodal model for advanced capabilities in understanding and generating text, code, and images.
  - **Imagine:** Tap into Google's powerful image generation model to create visuals within your chatbot interactions.
  - **GPT Models:** Integrate with OpenAI's GPT models by providing your own API key, offering further flexibility in language model selection.
  - and many other models like Claude Sonnet, Mistral and others
- **User Authentication:** Securely authenticate users through their Google accounts. This simplifies access while leveraging Google's robust security infrastructure.
- **Contextual Awareness:** Maintain conversation history and context, allowing the chatbot to provide personalized and relevant responses throughout the interaction.

Advanced Chatbot Features:

- **Integration with Vertex AI:** Seamlessly connect with your data stores in Vertex AI, including vector databases and cloud storage buckets containing unstructured data (PDFs, Excel, Word, etc.). This allows the chatbot to access and process relevant information from your knowledge base.
- **Google Search Integration:** Provide the most up-to-date information in chatbot responses by integrating with Google Search. This enables the chatbot to retrieve relevant data from the internet, ensuring accuracy and comprehensiveness.
- **Image and Document Upload:** Enhance chatbot interactions by allowing users to upload images and documents. The chosen language model can then analyze these attachments, extract relevant information, and incorporate it into the conversation.
- **API Integration:** Connect with various third-party APIs (e.g., CRM systems, payment gateways) to extend the chatbot's functionality and integrate it with existing business processes.
- **Custom AI Assistants**: AI Agents feature provides a flexible framework for creating custom AI assistants powered by various model providers.

Go GenAI Studio's comprehensive chatbot functionalities empower businesses to create sophisticated, AI-driven conversational experiences that drive engagement, improve customer satisfaction, and automate key tasks.

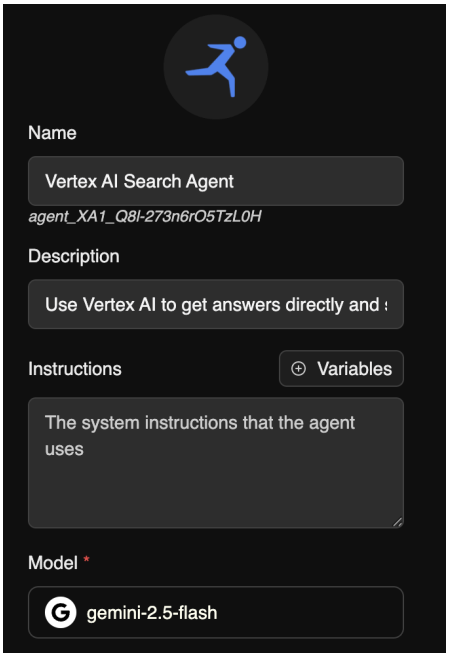
2.2 Agents

2.2.1 Getting Started

To create an agent, open the Agent Builder panel located in the side panel. A creation form will then appear, which you will need to complete.

The agent creation form is a comprehensive interface designed to empower users in defining and customizing their AI agents. This form encompasses several key fields, each playing a crucial role in shaping the agent's identity, purpose, and operational behavior:

- **Avatar:** This field allows for the upload of a custom avatar, providing a visual representation that personalizes your agent. A unique avatar helps distinguish your agent and adds a touch of individuality.
- **Name:** A distinctive name can be chosen for your agent. This name serves as the primary identifier and should be memorable and relevant to the agent's function.
- **Description:** This optional field offers an opportunity to provide additional details about your agent's purpose. A well-crafted description can clarify the agent's role and expected interactions for other users or for your own reference.
- **Instructions:** This critical field is where system instructions are defined, dictating your agent's behavior. These instructions are fundamental in shaping how the agent processes information, responds to queries, and interacts within its environment. Precise and clear instructions are essential for ensuring the agent performs as intended.
- **Model:** This section allows you to select from a range of available providers and models. The choice of model directly influences the agent's capabilities, performance, and underlying AI technology. Users can select the most suitable model based on their specific requirements and the complexity of the tasks the agent is expected to handle.



Example Agent: Vertex AI Search Agent

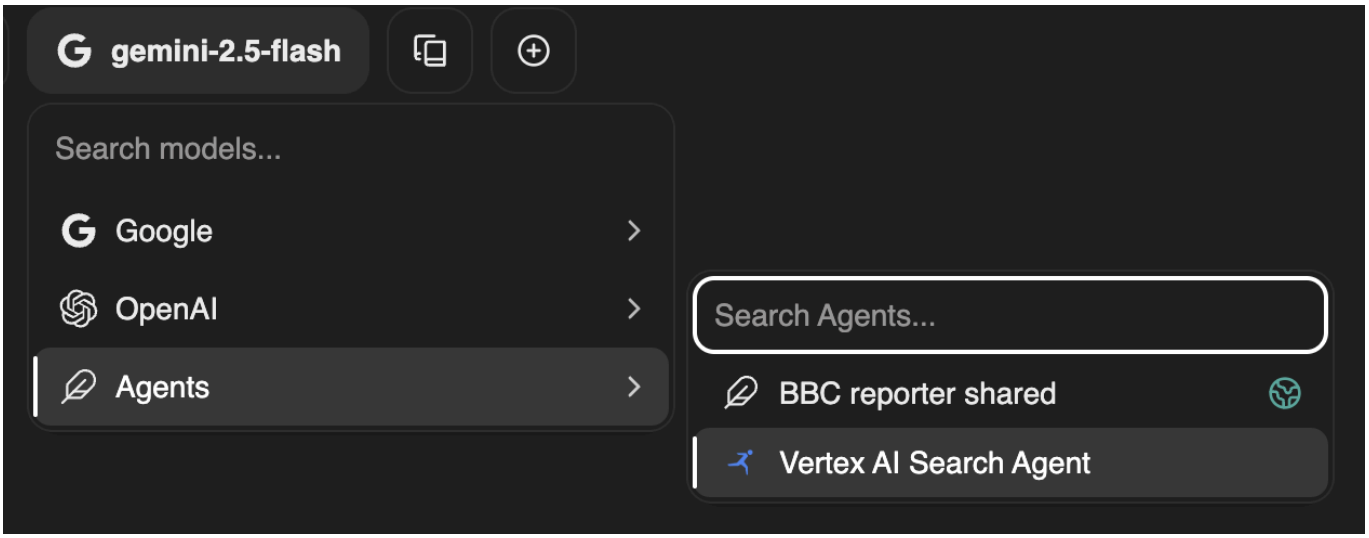
This example agent is specifically crafted for the purposes of this guide, demonstrating a powerful and secure method for data retrieval.

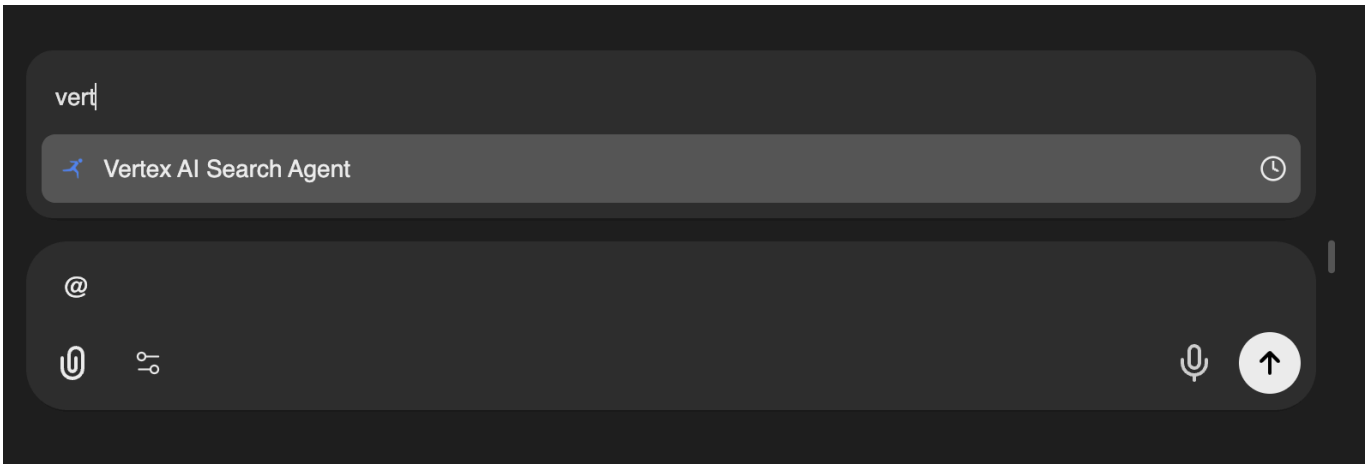
It is engineered to interface directly with the **Google Vertex AI API**, ensuring that all data interactions are handled with the highest level of security and efficiency.

The primary function of this agent is to extract precise and relevant answers from an organization's proprietary datasets, providing a streamlined and reliable way to access critical information.

For this particular use-case no additional instructions are necessary.

You can select existing agents from the top dropdown menu or by mentioning them using **"@"** in the chat input.

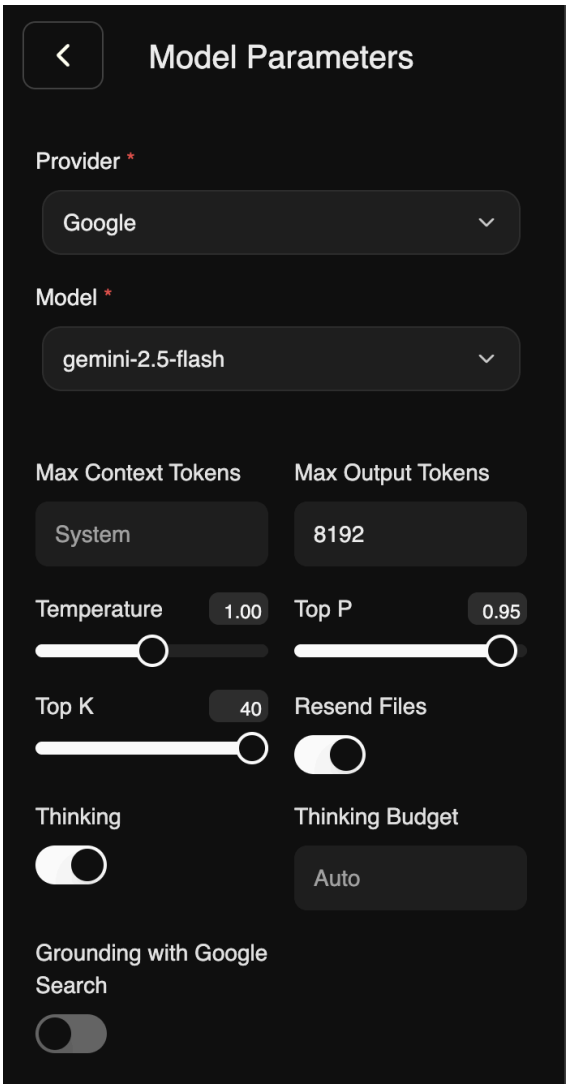




2.2.2 Model Configuration

The model parameters interface offers control over your agent's response generation, including:

- **Temperature:** A 0-1 scale to adjust response creativity.
- **Max context tokens:** Sets the maximum number of tokens for context.
- **Max output tokens:** Defines the maximum number of tokens in the output.
- **Additional provider-specific settings:** Other configurable options based on the provider.



As of version **0.7.9**, the Google provider offers the following adjustable parameters:

Top P: A nucleus sampling parameter (0-1) that controls token generation randomness.

Top K: Limits next token selection to the top K most probable tokens.

Resend Files: Determines if files are resent when persistent sessions are not maintained.

Thinking: Model engages in internal reasoning before responding.

Thinking Budget: Sets max tokens for internal reasoning; higher budget can improve quality for complex problems.

Grounding With Google Search: Model can use Google Search for answers.

2.2.3 Agent Capabilities

File Search:

File Search offers the following capabilities:

- **RAG (Retrieval-Augmented Generation) functionality:** Enhance responses by retrieving relevant information from uploaded documents.
- **Semantic search:** Conduct intelligent searches across your uploaded documents.
- **Context-aware responses:** Generate responses that are informed by the content of your files.
- **File attachment support:** Attach files at both the agent and chat thread levels.

Artifacts:

Artifacts allow agents to create and display interactive content, enabling them to:

- Generate React components, HTML code, and Mermaid diagrams.
- Show content in a dedicated UI window for enhanced clarity and interaction.
- Set up artifact-specific instructions at the agent level.

When artifact use is enabled, specific additional instructions are automatically included. These options are:

- **Enable shadcn/ui instructions:** This adds guidance for utilizing shadcn/ui components, which are reusable components built with Radix UI and Tailwind CSS.
- **Custom Prompt Mode:** Activating this mode bypasses the default artifact system prompt, allowing you to input your own custom instructions.

Configuring artifacts at the agent level is the preferred approach, as it allows for more granular control.

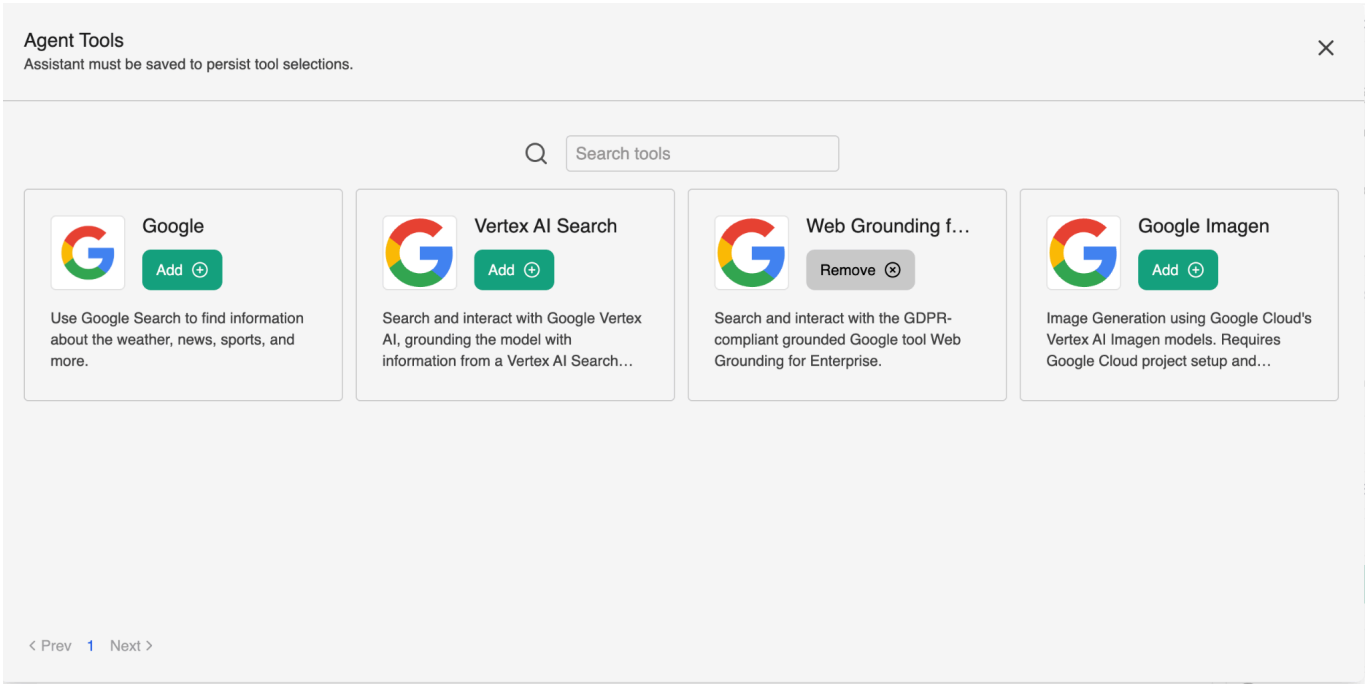
If you enable **Custom Prompt Mode**, you should include at minimum the basic artifact format in your **Instructions**.

```
None
When creating content that should be displayed as an artifact, use the following format:

:::artifact{identifier="unique-identifier" type="mime-type" title="Artifact Title"}
```
Your artifact content here
```
:::

For the type attribute, use one of:
- "text/html" for HTML content
- "application/vnd.mermaid" for Mermaid diagrams
- "application/vnd.react" for React components
- "image/svg+xml" for SVG images
```

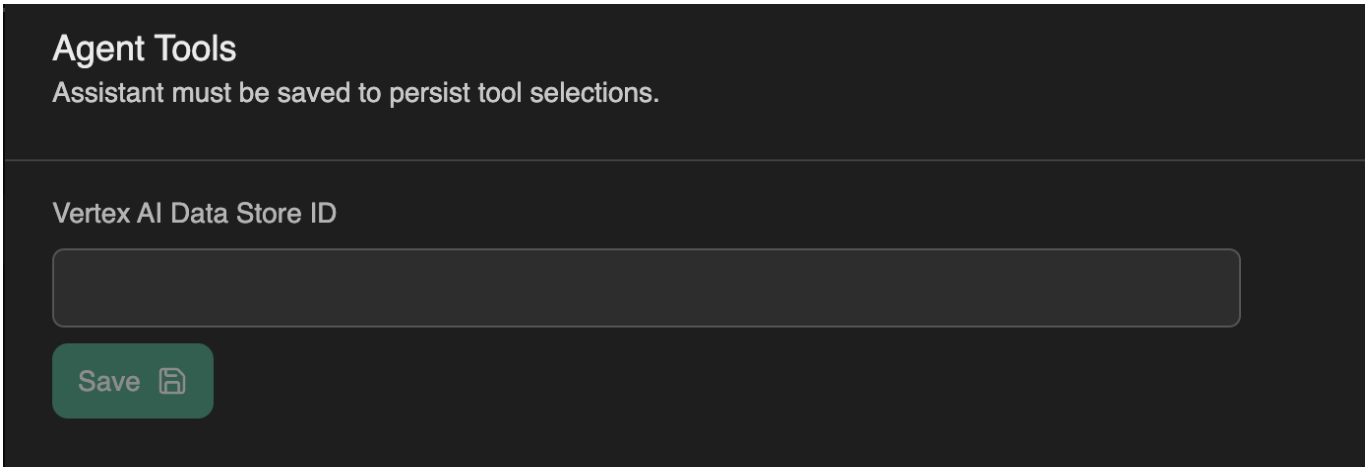
Tools:

Agents can also be enhanced with various built-in tools:

- **Google:** Access to web search functionality
- **Vertex AI Search:** Grounds AI responses by retrieving precise information from your custom enterprise data.
- **Web Grounding:** GDPR-Compliant alternative to Google Search (learn more here)
- **Google Imagen:** Generates images using Imagen 3



NOTE: For the Vertex AI Search tool, you need to provide a valid Data Store Path and click on save.



2.2.4 Advanced Settings

Additional agent settings, beyond core capabilities, are accessible within the advanced view of the agent form.

Agent Chain:

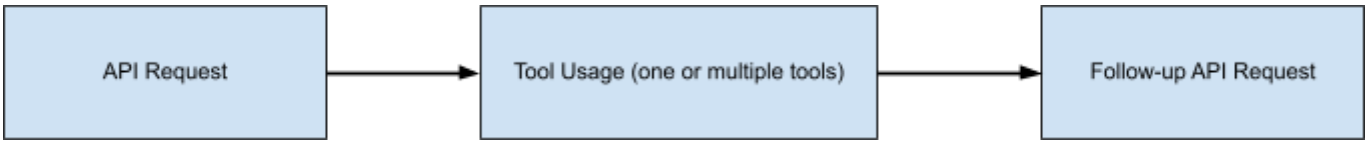
The Agent Chain feature facilitates a Mixture-of-Agents (MoA) methodology, enabling the sequential collaboration of agents. This allows for:

- **Chaining specialized agents** to tackle intricate tasks.
- **Access for each agent** in the chain to outputs from preceding agents.
- **Configuration of a maximum number of steps** for the agent chain.
- A current limit of **10 agents per chain**, with potential for future configurability.
- **Note:** This feature is currently in beta and may undergo changes.

Max Agent Steps:

This setting controls the maximum number of steps an agent can execute within a "run" (or agent loop) before generating a final response. By default, this limit is set to 25 steps, but it can be customized as needed. Each "step" represents either an AI API request or a round of tool usage, which may involve one or multiple tool calls from a single Large Language Model (LLM) request.

A single step is defined as a non-tool response. A round of tool usage typically involves three steps:

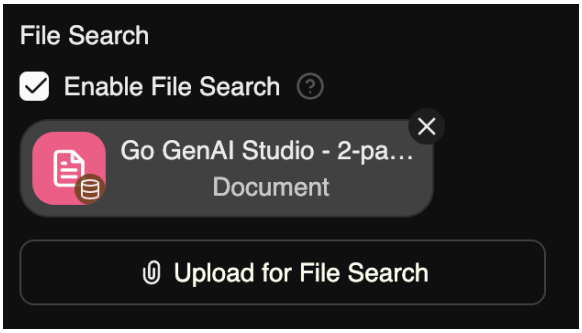


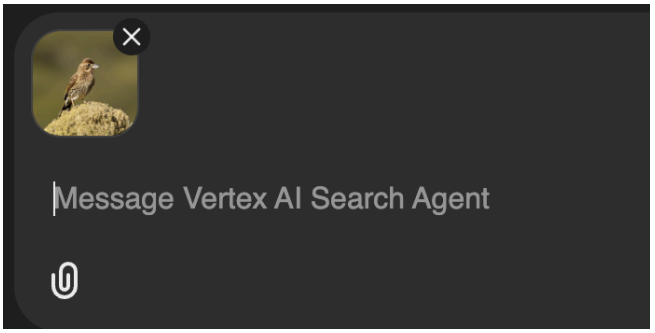
2.2.5 File Management

Agents accept file uploads in two main categories:

- **Image Upload:** For processing visual content.
- **File Search Upload:** For documents used in Retrieval Augmented Generation (RAG).

Files can be attached either directly to the agent's configuration or within individual chat threads.





2.2.6 Sharing and Permissions

Administrator Controls:

Administrators can access global permission settings within the agent builder UI, allowing them to:

- Enable or disable agent sharing for all users.
- Control agent usage permissions.
- Manage agent creation rights.
- Configure platform-wide settings.

Note: Per default all users are non-admin users. If you need an administrator account, please request via go.de.genai.studio@reply.de.

User-Level Sharing:

Individual users have the ability to:

- Share their agents with all other users, provided this feature is enabled.
- Manage editing permissions for any agents they have shared.
- Control access to the agents they have created.

Notes:

- Instructions, model parameters, attached files, and tools are accessible only to users with editing permissions.
- Agents may inadvertently share sensitive information, including instructions or files, during conversations. Therefore, ensure your instructions are designed to prevent such leaks.
- Only the original authors and administrators have the authority to delete shared agents.
- Agents are private by default and remain so unless explicitly shared by their authors.

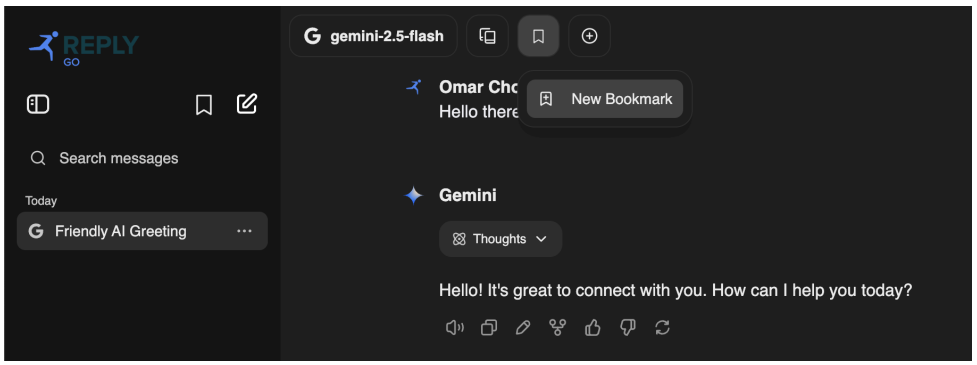
2.3 Bookmarks

2.3.1 Introduction

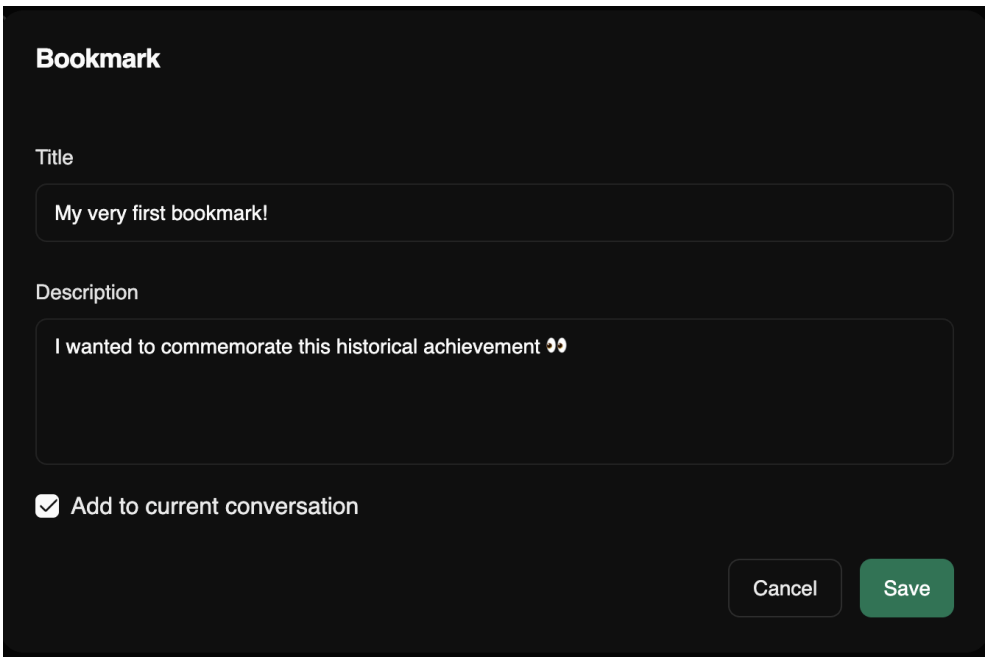
Enhance your workflow and streamline your communication management with the exciting new ability to filter conversations using bookmarks. This powerful new feature allows you to effortlessly organize and prioritize your interactions, ensuring that you can quickly locate and focus on the most important discussions. By leveraging bookmarks, you gain greater control over your conversation view, leading to improved efficiency and a more productive experience.

2.3.2 How-to

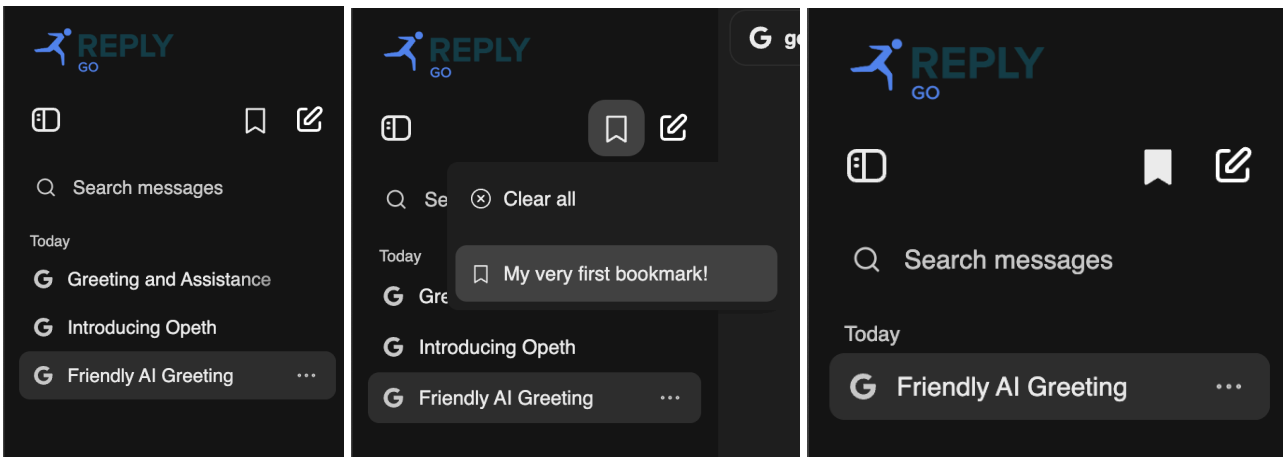
To create a new bookmark, go to the top panel, click the bookmark icon, and then select "New Bookmark."



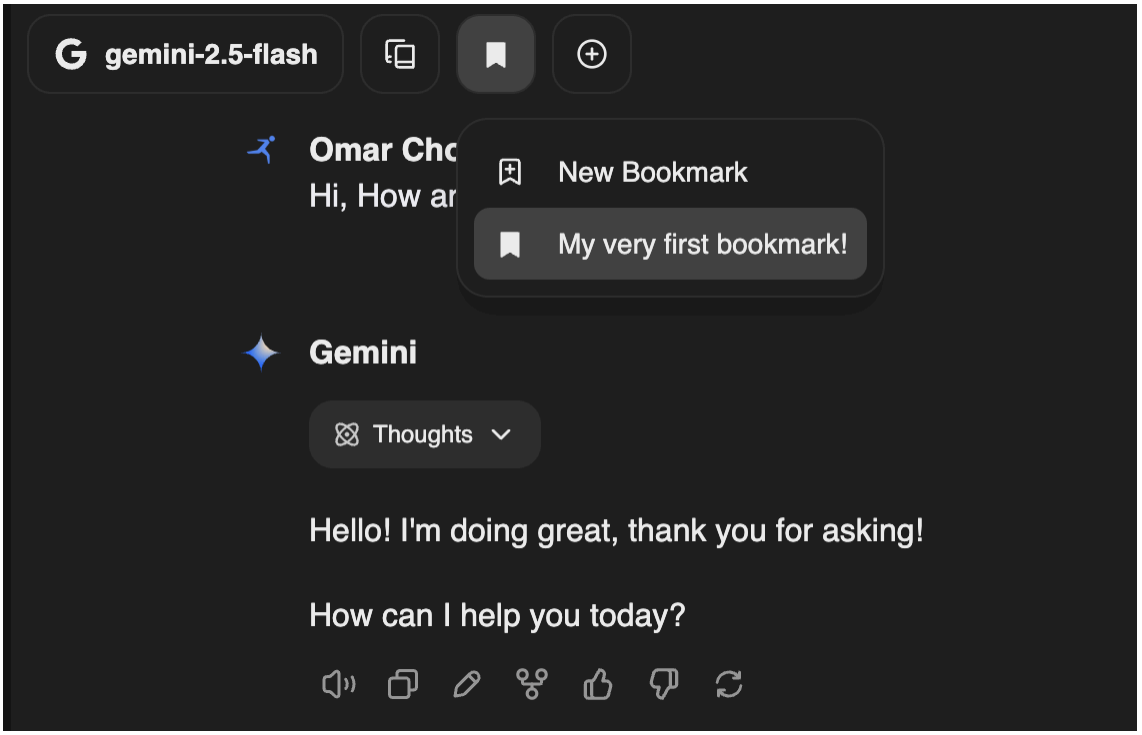
A form will appear for you to complete and save.



Conversations can now be filtered using bookmarks.



To bookmark a conversation, click on it, then select the bookmark icon in the top panel, and add it to your desired bookmark collection.



2.4 Prompting

2.4.1 Introduction

In the world of artificial intelligence, the quality of your input (a.k.a. the "prompt") directly dictates the quality of the output. A vague prompt leads to a generic answer. A well-crafted prompt, however, can unlock precise, creative, and highly useful responses.

One of the most powerful techniques for achieving this is the **"You are..., detailed description of the problem"** framework. This method involves two key steps: assigning a role or persona to the AI, and then providing a clear and comprehensive task.

2.4.2 The Core Principle: Role + Task

The fundamental structure of this prompt style is simple:

1. **`You are...` (The Persona):** You assign the AI a specific role, complete with expertise, a point of view, and sometimes even a personality. This tells the AI *how* to think and from what perspective to answer.
2. **`Your task is...` (The Detailed Problem):** You provide a clear, specific, and detailed description of what you want the AI to do.

By combining these two elements, you move from simply asking a question to directing a specialist.

2.4.3 Part 1: Assigning the Persona

Assigning a persona is the single best way to narrow the AI's focus and improve the tone, style, and relevance of its response. Instead of a generic AI, you get an expert tailored to your needs.

Why is this so effective?

- **Sets the Tone:** A `senior physician` will use clear and empathetic language, while a `senior research scientist` will be formal, precise, and data-driven.
- **Focuses Knowledge:** The AI will draw upon the specific knowledge, vocabulary, and conventions associated with that role.
- **Defines the Audience:** The persona often implies a **target audience**, helping the AI craft a response that resonates with the right people.

How to Formulate the Persona:

Be specific. The more detail you provide, the better.

Weak Persona (Too Broad)	Strong Persona (Specific & Effective)
You are a doctor.	You are a board-certified attending oncologist specializing in gastrointestinal cancers, consulting on a complex case for a multidisciplinary tumor board.
You are a writer.	You are a professional travel blogger with a witty, humorous, and engaging writing style.

You are a researcher.	You are a senior clinical research scientist drafting the 'Methods' section for a manuscript on a neurology clinical trial for submission to a peer-reviewed journal.

Checklist for a Strong Persona:

- **Role/Profession:** What is their job title? (e.g., *Attending Physician, Clinical Nurse Specialist, Hospital Pharmacist*)
- **Expertise/Specialization:** What is their specific field of knowledge? (e.g., *in pediatric cardiology, in surgical oncology, in critical care nursing*)
- **Key Traits/Style:** What is their tone? (e.g., *concise and professional, enthusiastic and encouraging, skeptical and analytical*)

2.4.4 Part 2: Describing the Task

This is where you provide the AI with everything it needs to successfully complete the task. Clarity and context are your primary goals. Don't assume the AI knows what you're thinking.

Key Components of a Detailed Task Description:

1. **The Goal (The "What"):** State the primary objective clearly. What is the single most important thing you want to achieve?
    - *Example:* `Your goal is to write an email that persuades invited guests to RSVP to an event.`
2. **Context (The "Why"):** Provide background information. Why is this task important? Who is the audience? What is the situation?
    - *Example:* `This is for a corporate charity gala. The audience is high-level executives who are very busy. The tone should be respectful, elegant, and brief.`
3. **Constraints & Rules (The "How"):** Set boundaries and give specific instructions.
    - **Format:** `Structure the output as a bulleted list. Provide the answer in a table.`
    - **Length:** `The entire response must be under 200 words. Generate three distinct options.`
    - **Inclusions/Exclusions:** `Do not mention our competitor “Competitor Corp”. Be sure to include a call to action at the end.`
4. **Step-by-Step Instructions (For Complex Tasks):** If the task has multiple parts, break it down.
    - *Example:* `First, analyze the attached customer feedback for common themes. Second, categorize these themes into three groups: Positive, Negative, and Suggestions. Finally, write a summary of your findings.`

By consistently applying the **"You are..., detailed description of the problem"** method, you can transform the AI from a simple tool into a powerful, specialized assistant for any task.

3. Glossary

Agent	An Agent is a specialized version of the chatbot that you can create for a specific purpose. Think of it as a custom assistant you've trained with its own unique instructions, knowledge, and tools to perform a particular task, like searching through company documents or creating images.
API (Application Programming Interface)	An API is like a messenger that allows different software applications to talk to each other and share information. For example, an API allows the chatbot to connect to Google Search or other third-party systems to get information.
API Key	An API Key is a unique code that acts like a password. It grants the chatbot permission to access a specific software service, like OpenAI's GPT models. You provide the key to prove you have access rights to that service.
Artifacts	Artifacts are interactive or visual elements that an Agent can create and display in a separate window. Instead of just plain text, an Agent can generate content like clickable web components, diagrams, or HTML code that you can see and interact with.
Chatbot	A chatbot is a computer program designed to simulate human conversation through text or voice commands. Go GenAI Studio is an advanced chatbot that uses Generative AI to understand questions and provide detailed, relevant answers.
Cloud Storage Bucket	A cloud storage bucket is an online container for storing files, like PDFs, Word documents, or Excel sheets. Think of it as a folder on the internet where you can keep your data for the chatbot to access.
Data Store	A Data Store is a repository where your organization's specific data is kept. When you connect an Agent to a Data Store (like on Vertex AI), you allow it to search for answers using your private, proprietary information.
Generative AI (GenAI)	Generative AI is a type of artificial intelligence that can create new content, such as text, images, or code, instead of just analyzing existing data. It's the technology that powers the chatbot, allowing it to write emails, summarize documents, and answer complex questions in a human-like way.
Grounding	Grounding is the process of connecting the AI's responses to a reliable source of facts to ensure the information is accurate and up-to-date. For example, an Agent can be "grounded" with Google Search or your company's internal documents to base its answers on real-world information rather than just its own training data.
Knowledge Base	A knowledge base refers to the collection of information that the chatbot can draw upon to answer questions. This can include uploaded documents (PDFs, Word files, etc.) stored in a cloud database that the AI can search through.
Large Language Model (LLM) / AI Model	An AI Model or LLM is the "brain" behind the chatbot. It's a massive digital library trained on vast amounts of text and data that has learned patterns, grammar, and information. Different models (like Gemini or GPT) have different strengths, much like different experts have different specialties.

Multimodal Model	A multimodal model is a type of AI that can understand and work with more than one type of information at once. For example, Google's Gemini is multimodal because it can process not only text but also images and code.
Open-source	Open-source refers to software whose original source code is made freely available for anyone to view, modify, and distribute. Go GenAI Studio is based on an open-source solution, which allows for greater transparency and customization.
Prompt	A prompt is the instruction or question you give to the AI. A well-crafted, specific prompt that assigns the AI a role (e.g., "You are an expert physician") and a detailed task will produce much better results than a vague one.
RAG (Retrieval-Augmented Generation)	RAG is a technique that allows the chatbot to improve its answers by first "retrieving" (or looking up) relevant information from your private, uploaded documents and then using that information to "generate" a more accurate and context-aware response.
Semantic Search	Semantic search is an advanced type of search that understands the meaning and context of your words, not just the keywords themselves. This allows it to find more relevant information across your documents, even if they don't contain the exact words you used in your query.
Temperature	Temperature is a setting that controls the creativity or randomness of the AI's responses, typically on a scale from 0 to 1. A low temperature (e.g., 0.2) makes the AI more focused and predictable, which is good for factual answers. A high temperature (e.g., 0.9) encourages more creative and diverse responses.
Tokens	Tokens are the small pieces of text—like words, parts of words, or punctuation—that an AI model processes. Both your prompt and the AI's response are measured in tokens. Setting limits on tokens helps control the length of the conversation and the cost of usage.
Top K	Top K is a setting that controls the AI's randomness by limiting its word choices. It makes the AI consider only the "top K" (a number you set) most likely words when generating a response. For example, if K is set to 40, the AI will only choose its next word from the 40 most probable options, preventing it from selecting a truly random or nonsensical word.
Top P	Top P, also known as nucleus sampling, is another parameter that controls the AI's randomness, but in a more dynamic way. Instead of picking from a fixed number of words (like Top K), it picks from the smallest possible set of words whose combined probability is at least the value of P (e.g., 0.95). This allows the list of potential words to be long or short, depending on the context, making it a flexible way to balance creativity and coherence.
Unstructured Data	Unstructured data is information that is not organized in a predefined way, like in a neat database. This includes documents like PDFs, Word files, Excel spreadsheets, and images. The chatbot is capable of searching through this type of data to find information.
User Authentication	User Authentication is the process of verifying a user's identity to ensure they are who they say they are before granting them access to the system. In this case, it's done

	securely through your Google account.
Vertex AI	Vertex AI is Google's cloud platform for building and managing artificial intelligence applications. Go GenAI Studio uses Vertex AI to connect to your organization's data, allowing the chatbot to access your private knowledge base securely.