

# LLM Grounding: Vertex AI

## Grounding

Grounding stands out by creating a dynamic link between models and external data, ensuring real-time adaptability, and reducing inaccuracies. This strategy is emphasized in the project documentation to enhance the precision and reliability of language models, especially through the integration of Vertex AI Search and Conversation.

The grounding strategy focuses on text-bison and chat-bison models, anchoring generated content in designated data stores within Vertex AI Search. This approach mitigates model hallucinations, secures anchored responses, and increases the credibility of the generated content.

Grounding transforms AI models, making them more focused, accurate, and efficient by reducing hallucinations, anchoring responses, and enhancing trustworthiness. Grounded models differ from non-grounded ones in data dependence, response accuracy, and adaptability.

Implementing grounding involves activating Vertex AI Search, creating a data store, and grounding the model using the API. Regular monitoring and iteration to enhance grounding, and ensure compliance with data use terms and security standards are vital considerations.

In conclusion, integrating grounding in Vertex AI highlights the commitment to delivering accurate, relevant, and trustworthy content. The documentation serves as a comprehensive guide for implementing, monitoring, and optimizing grounding features, aligning with the broader goal of providing high-quality, reliable, and innovative solutions in the dynamic world of generative AI projects.

## Steps for Crafting Grounding:

1. Define Data Source: Initiate this process by defining a data source in Vertex AI Search.
2. Data Source ID: Obtain the unique data source ID, an essential element for the grounding process.
3. Enable API: Ensure that the necessary API is enabled, facilitating grounding functionality.

## Utilizing Vertex AI Search App to Ground Model Responses:

Vertex AI Search and Conversation lets developers tap into the power of Google's foundation models, search expertise, and conversational AI technologies to create enterprise-grade generative AI applications.

Three types of applications can be created from the service

- 1. Search** – an LLM-powered service that enables it to give a Google-level search experience from random data
- 2. Chat** – this service is empowered with chatbots and apps (search apps) to have a dialog flow with the end user.
- 3. Recommendations** – this system recommends media (related videos, images, news) to have personalized content and generic recommendations to recommend non-media content.

The Vertex AI Search enables 4 different data sources to crawl and retrieve the data:

- Website URLs
- BigQuery table
- Cloud Storage (Cloud Buckets)
- API – calling an API to import the data manually.

# Grounding ‘text-bison’ to website URLs

## Creating Search application

1. Navigate to Search & Conversation through the console search bar.
2. Agree to the conditions and activate the service.
3. Create an app for the service to be accessed.
4. Select the search type (here the aim is to search from the websites).

The screenshot shows the Google Cloud interface with the 'Search & Conversation' service selected. A modal window titled 'Create App' is open, specifically the 'Type' step. It displays three options: 'Search', 'Chat', and 'Recommendations'. Each option has a brief description and a 'SELECT' button. The 'Search' option is highlighted.

5. Fill in the required credentials and create the app.
6. Create a new data store.

The screenshot shows the 'Create a new data store' wizard. Step 1, 'Source', is selected. The 'Native sources' section contains three items: 'Website URLs' (selected), 'BigQuery', and 'Cloud Storage'. Each item has a 'SELECT' button. Below this is an 'API' section with an 'API' description and a 'SEE DOCUMENTATION' link.

7. Select the Website URL data store, and fill in the website URLs to be searched and the URLs to be excluded.

**Note:** Adding ‘/\*’ at the end of the main page URL allows the service to navigate to all the

web pages available in it.

The screenshot shows the 'Data' step of a three-step wizard. The left sidebar has tabs for 'Source' (checked), 'Data' (selected), and 'Configuration'. The main area is titled 'Specify the websites for your data store' with the sub-instruction 'Specify the list of websites you wish to index for your data store'. It includes a section 'Want advanced website indexing?' with a toggle switch (disabled). Below is a 'DATA MANAGEMENT INFORMATION' section with links to 'features' and 'pricing'. The 'Specify the URLs to index' section contains two text areas: 'Sites to include \*' (empty) and 'Sites to exclude' (empty). Both have character limits of 0/500. Below these is a note about operations like 'Entire site: www.mysite.com/\*'. At the bottom are 'CONTINUE' and 'CANCEL' buttons.

8. Give a name and create the data store.

The screenshot shows the 'Configuration' step of the wizard. The left sidebar has tabs for 'Source' (checked), 'Data' (checked), and 'Configuration' (selected). The main area is titled 'Configure your data store' with the sub-instruction 'Configure additional settings for your data store'. It includes a 'Location of your data store' section set to 'Multi-region global (Global)' and a 'Your data store name' section where 'Data store name \*' is filled with 'Enter a data store name'. A note says 'A data store ID will be generated based on the data store name. It cannot be changed later.' At the bottom are 'CREATE' and 'CANCEL' buttons.

9. Now select the created data source and complete the app creation.

## Pointing data store to the model

1. Navigate to the model garden in Vertex AI and select the 'PaLM 2 Text Bison' model.
2. Enable the API and select 'OPEN PROMPT DESIGN'.
3. Select any of the text bison models with a grounding option in the rightmost column of the page.

4. Drop down the Advanced menu and enable grounding.

The screenshot shows the Vertex AI interface with the 'Advanced' configuration panel open. The 'Enable Grounding' checkbox is checked, and the 'SOURCE' dropdown is set to 'Vertex AI search'. Other options like 'CUSTOMIZE' and 'Streaming responses' are also visible.

5. Customize the grounding option to set the path to the data store.

Fill the path in the format:

projects/{project\_id}/locations/global/collections/default\_collection/  
dataStores/{data\_store\_id}

Get the data store ID from the created search application.

### Customize Grounding

Select your grounding source from the drop-down below

Grounding source — **Vertex AI Search**

Vertex AI datastore path \*

**SAVE**   **CANCEL**

Now the model is ready to be queried related to the websites linked.