Intel® AI Assistant Builder

Product overview and user guide

February 2025

Contents

1.0 **Introduction** **4**

1.1 Prerequisites 4

1.1.1 Hardware requirements 4

1.1.2 Software requirements 4

1.2 Simple steps to get up and running quickly 5

1.3 Download the Intel® AI Assistant Builder 5

2.0 **Getting started** **6**

2.1 Select AI Assistant type and Install the application 6

2.2 First use and initial setup 7

2.2.1 Adding documents to the assistant 8

2.2.2 Customize the assistant - Settings 8

2.2.3 Interact with the AI Assistant 9

2.2.4 New chat and chat history 10

3.0 **Knowledge base / File library** **11**

4.0 **Model selection** **13**

4.1 Upload a model 13

5.0 **Parameter settings ("advanced" users only)** **15**

6.0 **Special Query Types** **18**

6.1 Query Summary 18

6.2 Query Tabular Data 19

6.3 Query Images 19

6.4 Additional Notes 20

7.0 **Troubleshooting** **21**

7.1 Updating or re-installing the application 21

7.2 Initial load time 21

7.3 Not responsive 21

7.4 Grayed Out Interface 21

7.5 Model loading error 22

Revision history

Date	Revision	Description
February 2025	1.1	Major revision for consistency, simplicity, and clarity.
January 2025	1.0	Added sections for Knowledge Base, Chat Sessions, Model Selection, Parameter Settings and Special Query Types, User Model Upload
December 2024	0.8.0	Added NPU support for Phi-3-4k model on Lunar Lake system, Model parameter tuning, Hugging Face model repo support
November 2024	0.7.1	Added instructions for proxy server configuration in section 1.2
November 2024	0.7	Initial release.

§

1.0 Introduction

The Intel® AI Assistant Builder is a software framework which enables AI Assistants customized to your specific needs and use cases. These AI assistants can dramatically streamline day-to-day tasks and provide answers to questions based on libraries of information you provide. Your proprietary data and processes are completely secure as the AI assistant runs locally using the best large language models (LLMs), retrieval-augmented generation (RAG), and other components optimized for performance and accuracy running on Intel® based AI PCs.

1.1 Prerequisites

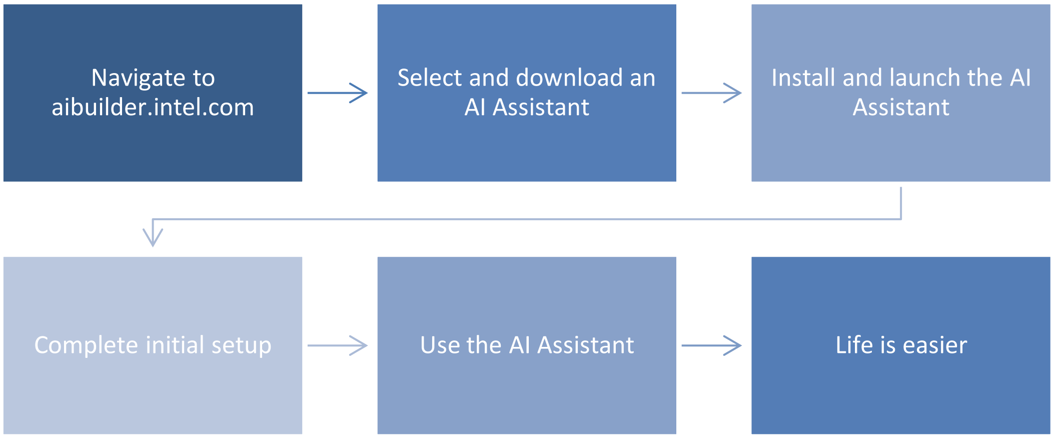
1.1.1 Hardware requirements

COMPONENT	MINIMUM REQUIREMENTS	RECOMMENDED REQUIREMENTS
Processor	Intel® Core™ Ultra processor Series 1 (Meteor Lake)	Intel® Core™ Ultra 200V series (Lunar Lake)
Memory (RAM)	16GB	32GB
Storage	4GB for AI Assistant with 1 LLM	12GB for AI Assistant with 3 LLMs
Graphics	Integrated Intel® Graphics	Integrated Intel® Arc™ Graphics
Network	Broadband connection for LLMs and other components’ download	

1.1.2 Software requirements

Intel® AI Assistant Builder has been validated for use on Microsoft Windows 11 version 23H2 or newer. During the installation process Intel® AI Assistant Builder application may download and install required components.

1.2 Simple steps to get up and running quickly


Navigate to aibuilder.intel.com
Select and download an AI Assistant
Install and launch the AI Assistant
Complete initial setup
Use the AI Assistant
Life is easier

1.3 Download the Intel® AI Assistant Builder

Navigate to https://aibuilder.intel.com in your web browser to learn about and download the application.

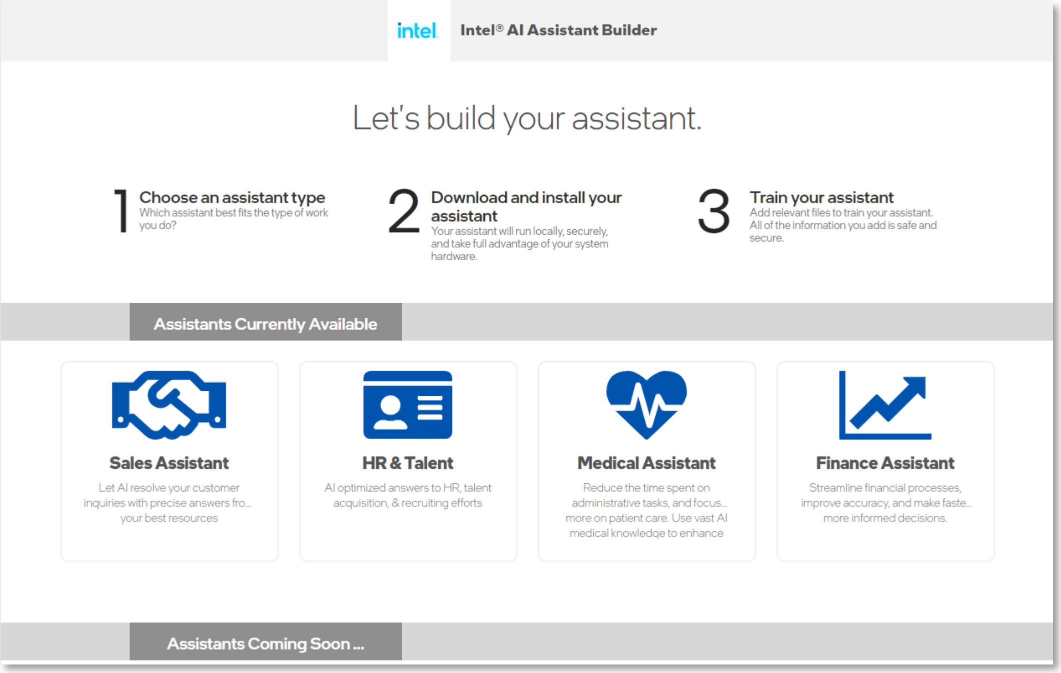
§

2.0 Getting started

2.1 Select AI Assistant type and Install the application

Step 1. Navigate to https://aibuilder.intel.com in your web browser.

Example Intel® AI Assistant Builder Web Portal



Intel® AI Assistant Builder

Let's build your assistant.

1 Choose an assistant type
Which assistant best fits the type of work you do?

2 Download and install your assistant
Your assistant will run locally, securely, and take full advantage of your system hardware.

3 Train your assistant
Add relevant files to train your assistant. All of the information you add is safe and secure.

Assistants Currently Available

Sales Assistant
Let AI resolve your customer inquiries with precise answers fro... your best resources

HR & Talent
AI optimized answers to HR, talent acquisition, & recruiting efforts

Medical Assistant
Reduce the time spent on administrative tasks, and focus... more on patient care. Use vast AI medical knowledge to enhance

Finance Assistant
Streamline financial processes, improve accuracy, and make faste... more informed decisions.

Assistants Coming Soon ...

Step 2. Click on one of the four AI Assistants to start the download. If none of the assistants are a good match for your needs we recommend using the "Sales Assistant" for general use. Much of the assistants capability (and appearance) can be customized after installation.

Note: Going forward, this guide assumes you are using the Sales Assistant. The experience with the other AI Assistants is largely similar but there may be minor differences.

Step 3. Once the installer is downloaded, locate it in your Downloads folder.

- The file name of the installer "**Intel(R) AI Assistant Builder_Installer_XX_1.0.0.YYYY.exe**" has the following meaning:

  - **XX** defines the type of assistant: SA for Sales Assistant, HR for HR & Talent, and MA for Medical Assistant
  - **YYYY** is the build number

Step 4. Launch the installation wizard by double clicking the downloaded file. The wizard will guide you through the required steps to successfully complete the installation

Note: Various supporting components are required. Depending on your individual system you may or may not need to install these components as part of the installation process. You must accept the End User License Agreement (EULA) to complete the installation.

Step 5. If the “Launch when ready” checkbox was selected during installation, the Intel® AI Assistant Builder application will launch automatically. Otherwise, manually launch it from the Windows start menu.

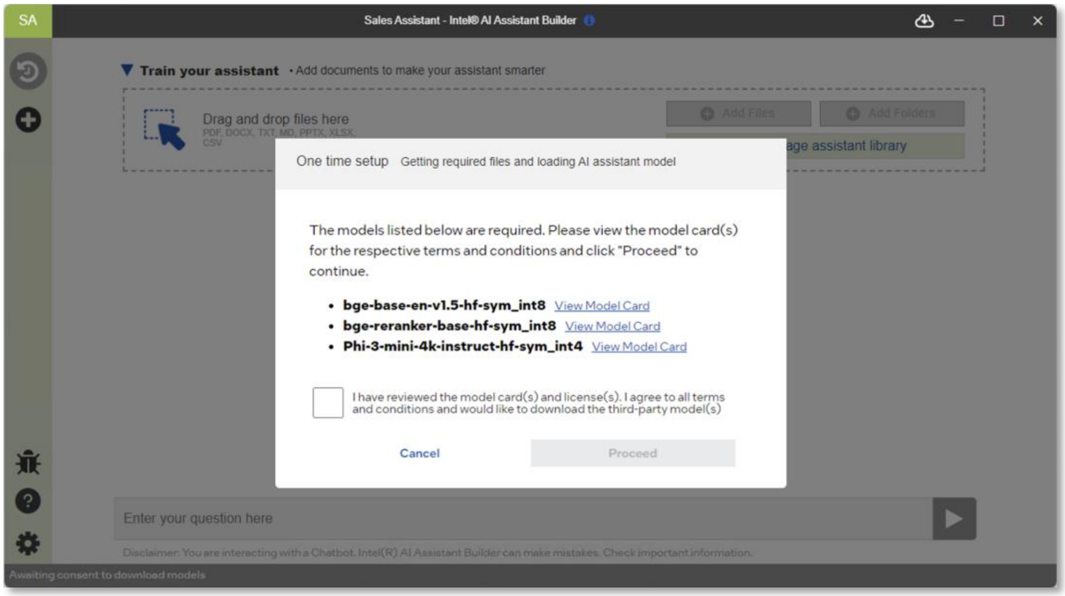
2.2 First use and initial setup

Downloading required files

The first time the application runs certain required files will need to be downloaded. Accept the terms and conditions of use and then click “Proceed” for the download to begin. You may close the download progress window if desired. To re-open the download progress window click the ☁ icon in the title bar at the top of the main application window.

The status bar at the bottom of the window indicates the application status and is removed when the download is complete.

Intel® AI Assistant Builder Application Main Screen at the First Launch


SA
Sales Assistant - Intel® AI Assistant Builder
Train your assistant
• Add documents to make your assistant smarter
Drag and drop files here
PDF, DOCX, TXT, MD, PPTX, XLSX, CSV
Add Files
Add Folders
age assistant library
Cancel
Proceed
Enter your question here

One time setup Getting required files and loading AI assistant model

The models listed below are required. Please view the model card(s) for the respective terms and conditions and click "Proceed" to continue.

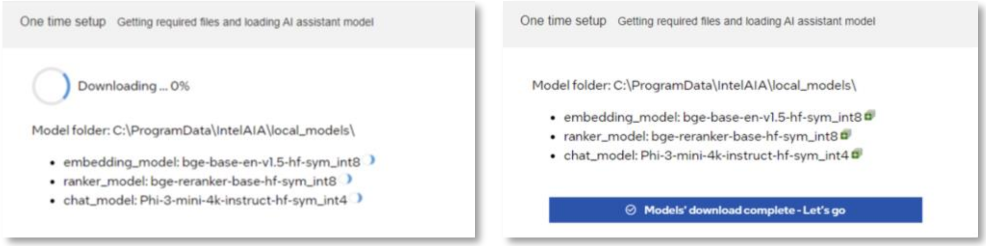
- **bge-base-en-v1.5-hf-sym_int8** View Model Card
- **bge-reranker-base-hf-sym_int8** View Model Card
- **Phi-3-mini-4k-instruct-hf-sym_int4** View Model Card

□ I have reviewed the model card(s) and license(s). I agree to all terms and conditions and would like to download the third-party model(s)

Disclaimer: You are interacting with a Chatbot. Intel(R) AI Assistant Builder can make mistakes. Check important information.

Awaiting consent to download models

LLM, RAG, and Other Components' Download Dialogue Window:
Download in Progress (left), Download Complete (right)



One time setup Getting required files and loading AI assistant model

Downloading ... 0%

Model folder: C:\ProgramData\IntelAIA\local_models\

- embedding_model: bge-base-en-v1.5-hf-sym_int8
- ranker_model: bge-reranker-base-hf-sym_int8
- chat_model: Phi-3-mini-4k-instruct-hf-sym_int4

One time setup Getting required files and loading AI assistant model

Model folder: C:\ProgramData\IntelAIA\local_models\

- embedding_model: bge-base-en-v1.5-hf-sym_int8
- ranker_model: bge-reranker-base-hf-sym_int8
- chat_model: Phi-3-mini-4k-instruct-hf-sym_int4

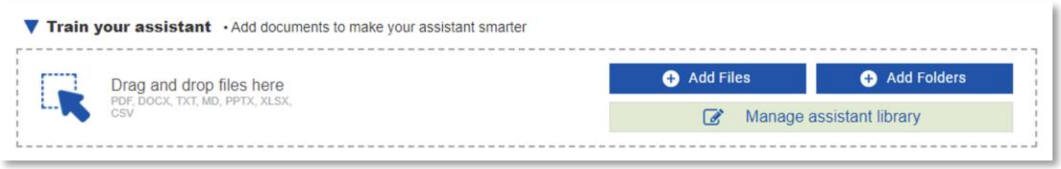
Note: The AI Assistant can't be used until all required files are downloaded and "activated". Many features will be disabled until the download process is complete. This process is only required the first time the assistant runs and typically takes a few minutes at most to complete, however depending on network conditions and specific system hardware the process could take longer.

2.2.1 Adding documents to the assistant

"Train your assistant" by adding the information (documents) it should know in order to best help you. To add documents, do one of the following:

- drag and drop a file to the target area
- click the "Add Files" button to add one file at a time
- click the "Add Folders" button to add several files from the same folder.

Intel® AI Assistant Builder Application Interface for File and Folder Loading



Train your assistant • Add documents to make your assistant smarter

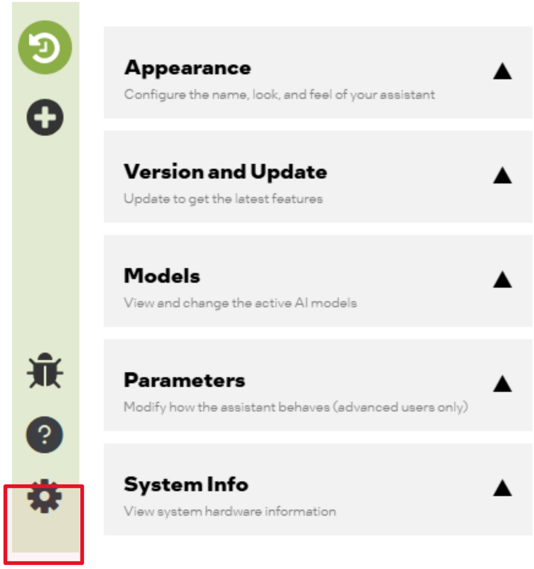
Drag and drop files here
PDF, DOCX, TXT, MD, PPTX, XLSX, CSV

Note: The list of supported file types is shown towards the left side of the "Train your assistant" section.

To manage files added to the assistant, click the "Manage assistant library" button.

2.2.2 Customize the assistant - Settings

You can further customize your AI Assistant by clicking the settings icon on the left sidebar.



Appearance ▲
Configure the name, look, and feel of your assistant

Version and Update ▲
Update to get the latest features

Models ▲
View and change the active AI models

Parameters ▲
Modify how the assistant behaves (advanced users only)

System Info ▲
View system hardware information

Through the settings menu you can:

- **Customize the appearance of the assistant** using a custom name and/or color theme.
- **Change the LLM and RAG models** which power your AI Assistant. You can even upload* your own models as desired.
  *This is an advanced feature and should be used only by those who understand the impact.
- **Adjust LLM and RAG parameters***.
  *This is an advanced feature and should be used only by those who understand the impact.
- **View the hardware** details of your AI PC.

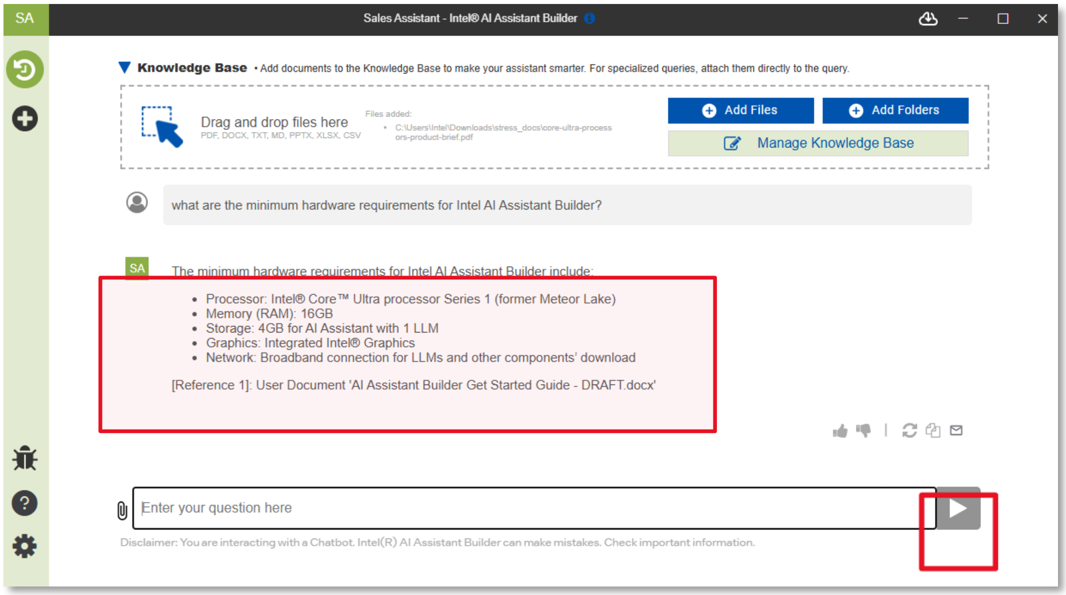
Note: When the model is changed the download for the newly selected model will begin (if needed). While the model is being downloaded and activated use of the assistant is disabled. Download progress can be checked from the download status icon in the title bar.

2.2.3 Interact with the AI Assistant

After the initial setup and any optional customizations are complete, you are ready to chat with your AI Assistant. Type your question into the text box toward the bottom of the window and hit “Enter” or click ▶ . You will see the response in the middle portion of the window.

Your assistant is much smarter and more helpful when you add documents to the “knowledge base” in order to give it context and references for your topic of interest . The process for this is covered in the next section of this document.

Example Interaction with the AI Assistant



• Add documents to the Knowledge Base to make your assistant smarter. For specialized queries, attach them directly to the query.

- C:\Users\Intel\Downloads\stress_docs\core-ultra-processors-product-brief.pdf

what are the minimum hardware requirements for Intel AI Assistant Builder?

The minimum hardware requirements for Intel AI Assistant Builder include:

- Processor: Intel® Core™ Ultra processor Series 1 (former Meteor Lake)
- Memory (RAM): 16GB
- Storage: 4GB for AI Assistant with 1 LLM
- Graphics: Integrated Intel® Graphics
- Network: Broadband connection for LLMs and other components' download

[Reference 1]: User Document 'AI Assistant Builder Get Started Guide - DRAFT.docx'

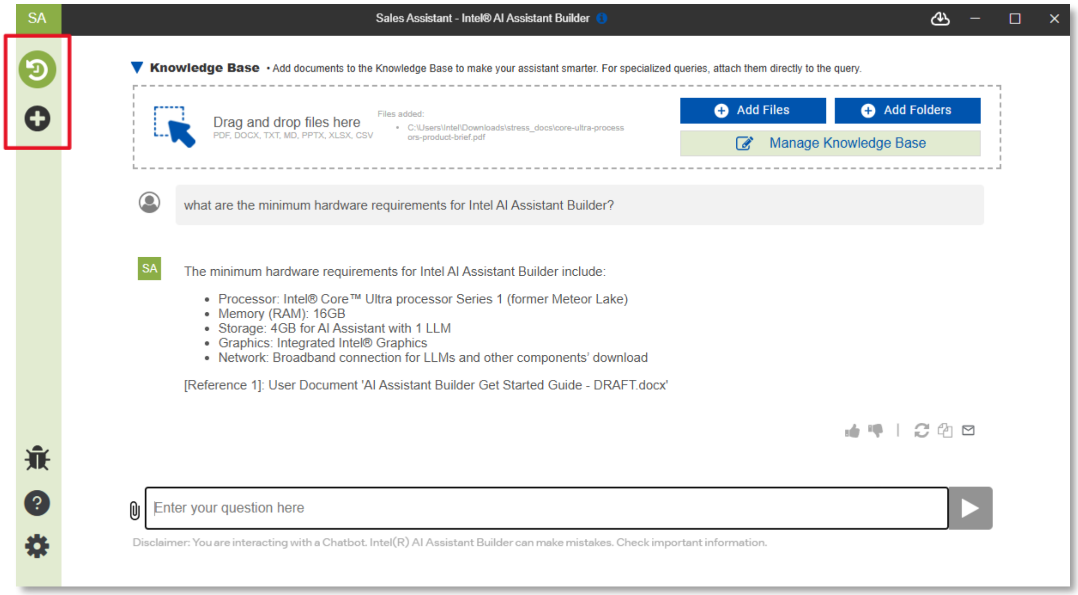
Enter your question here

Disclaimer: You are interacting with a Chatbot. Intel(R) AI Assistant Builder can make mistakes. Check important information.

2.2.4 New chat and chat history

You may start a new chat by clicking the”New chat” ⊕ icon toward the top of the left side bar.

You may recall/resume prior chats by clicking the “History” ⟲ icon just above the “New chat” icon.



• Add documents to the Knowledge Base to make your assistant smarter. For specialized queries, attach them directly to the query.

- C:\Users\Intel\Downloads\stress_docs\core-ultra-processors-product-brief.pdf

what are the minimum hardware requirements for Intel AI Assistant Builder?

The minimum hardware requirements for Intel AI Assistant Builder include:

- Processor: Intel® Core™ Ultra processor Series 1 (former Meteor Lake)
- Memory (RAM): 16GB
- Storage: 4GB for AI Assistant with 1 LLM
- Graphics: Integrated Intel® Graphics
- Network: Broadband connection for LLMs and other components' download

[Reference 1]: User Document 'AI Assistant Builder Get Started Guide - DRAFT.docx'

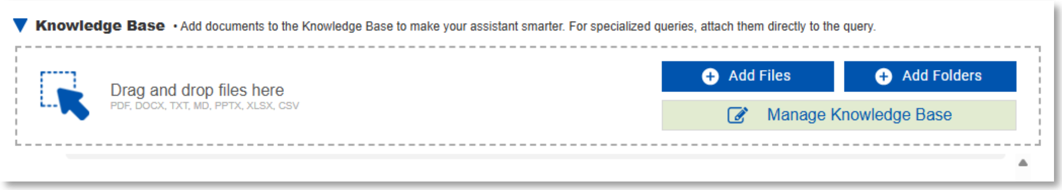
Enter your question here

Disclaimer: You are interacting with a Chatbot. Intel(R) AI Assistant Builder can make mistakes. Check important information.

3.0 Knowledge base / File library

You may train your assistant by adding files and documents to the common “Knowledge Base”. The assistant will use this knowledge base as a reference to find relevant information to answer your questions. The assistant will also note which document it used to answer any questions you ask. The knowledge base is shared across chat sessions and remains in place unless you choose to remove or replace the document.

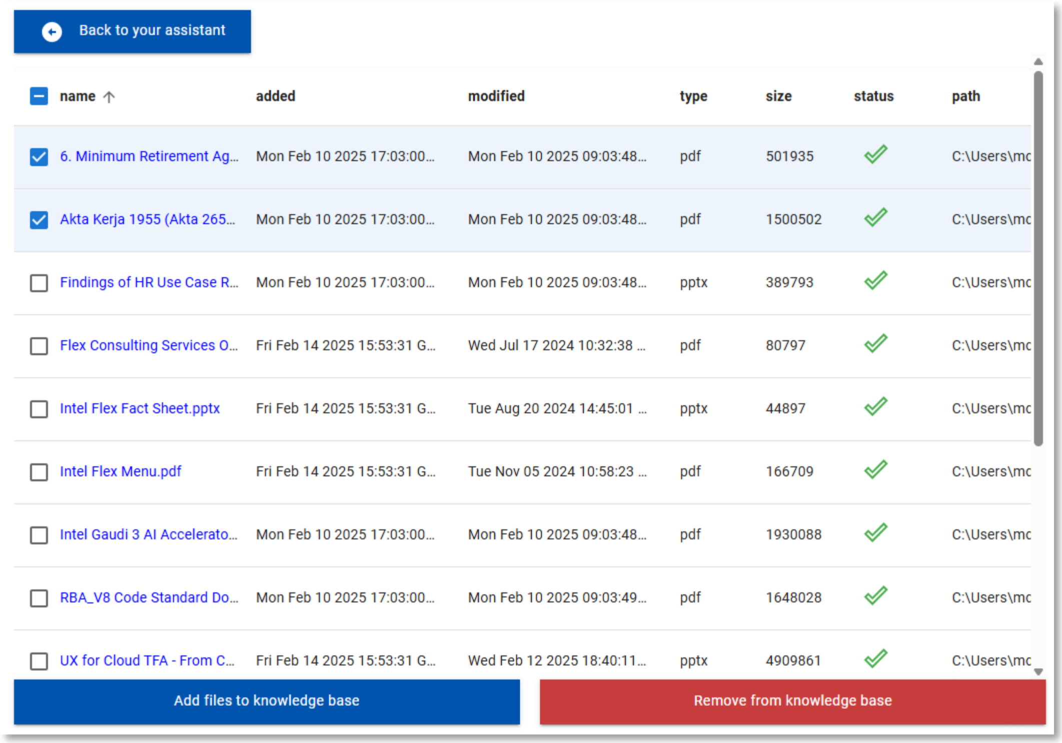
Supported file formats include PDF, DOCX, TXT, MD, PPTX, XLSX, CSV



▼ **Knowledge Base** • Add documents to the Knowledge Base to make your assistant smarter. For specialized queries, attach them directly to the query.

Drag and drop files here
PDF, DOCX, TXT, MD, PPTX, XLSX, CSV

The Knowledge Base is managed by clicking the “Manage Knowledge Base” button where you may add or remove files as needed.



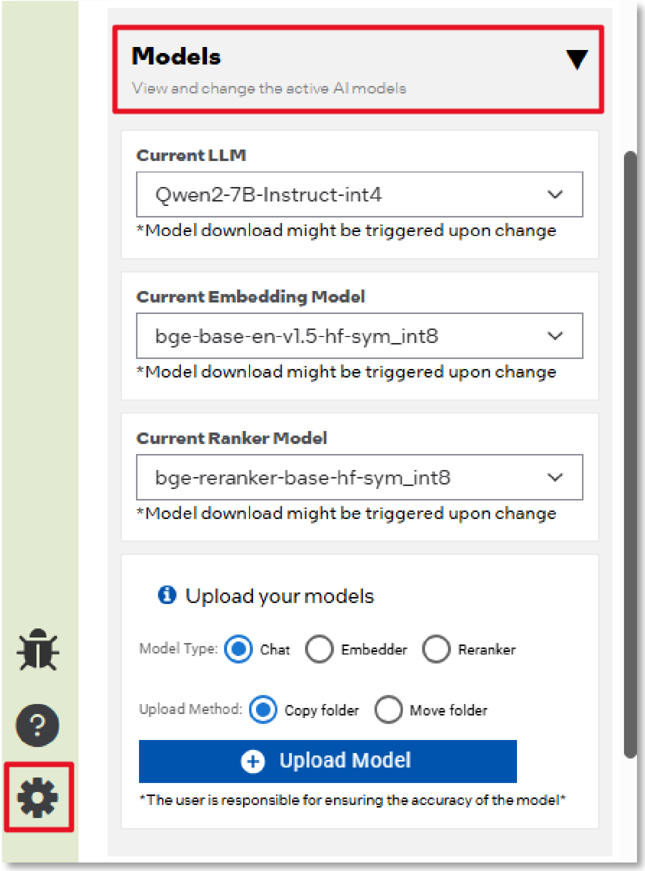
☐ name ↑	added	modified	type	size	status	path
☑ 6. Minimum Retirement Ag...	Mon Feb 10 2025 17:03:00...	Mon Feb 10 2025 09:03:48...	pdf	501935	✓	C:\Users\mc
☑ Akta Kerja 1955 (Akta 265...	Mon Feb 10 2025 17:03:00...	Mon Feb 10 2025 09:03:48...	pdf	1500502	✓	C:\Users\mc
☐ Findings of HR Use Case R...	Mon Feb 10 2025 17:03:00...	Mon Feb 10 2025 09:03:48...	pptx	389793	✓	C:\Users\mc
☐ Flex Consulting Services O...	Fri Feb 14 2025 15:53:31 G...	Wed Jul 17 2024 10:32:38 ...	pdf	80797	✓	C:\Users\mc
☐ Intel Flex Fact Sheet.pptx	Fri Feb 14 2025 15:53:31 G...	Tue Aug 20 2024 14:45:01 ...	pptx	44897	✓	C:\Users\mc
☐ Intel Flex Menu.pdf	Fri Feb 14 2025 15:53:31 G...	Tue Nov 05 2024 10:58:23 ...	pdf	166709	✓	C:\Users\mc
☐ Intel Gaudi 3 AI Accelerato...	Mon Feb 10 2025 17:03:00...	Mon Feb 10 2025 09:03:48...	pdf	1930088	✓	C:\Users\mc
☐ RBA_V8 Code Standard Do...	Mon Feb 10 2025 17:03:00...	Mon Feb 10 2025 09:03:49...	pdf	1648028	✓	C:\Users\mc
☐ UX for Cloud TFA - From C...	Fri Feb 14 2025 15:53:31 G...	Wed Feb 12 2025 18:40:11...	pptx	4909861	✓	C:\Users\mc

PRO Tip: When documents are added or removed from the knowledge base the source documents are not moved and remain unchanged. Only the embeddings of the files are stored in the knowledge base.

4.0 Model selection

You may choose to use a different model for your assistant by clicking the “Settings” icon and opening the “Models” section. Use the drop down menus to select a supported model. As needed when a new model is selected it will be downloaded and activated for use. You may also upload a model* for use by the assistant.

PRO Tip: *Uploading a model is an advanced feature and should be done by those who understand the impact of using a “custom” model. It is your responsibility to ensure compatibility and accuracy.



Models ▼

View and change the active AI models

Current LLM: Qwen2-7B-Instruct-int4

*Model download might be triggered upon change

Current Embedding Model: bge-base-en-v1.5-hf-sym_int8

*Model download might be triggered upon change

Current Ranker Model: bge-reranker-base-hf-sym_int8

*Model download might be triggered upon change

ℹ Upload your models

Model Type: ● Chat ○ Embedder ○ Reranker

Upload Method: ● Copy folder ○ Move folder

The user is responsible for ensuring the accuracy of the model

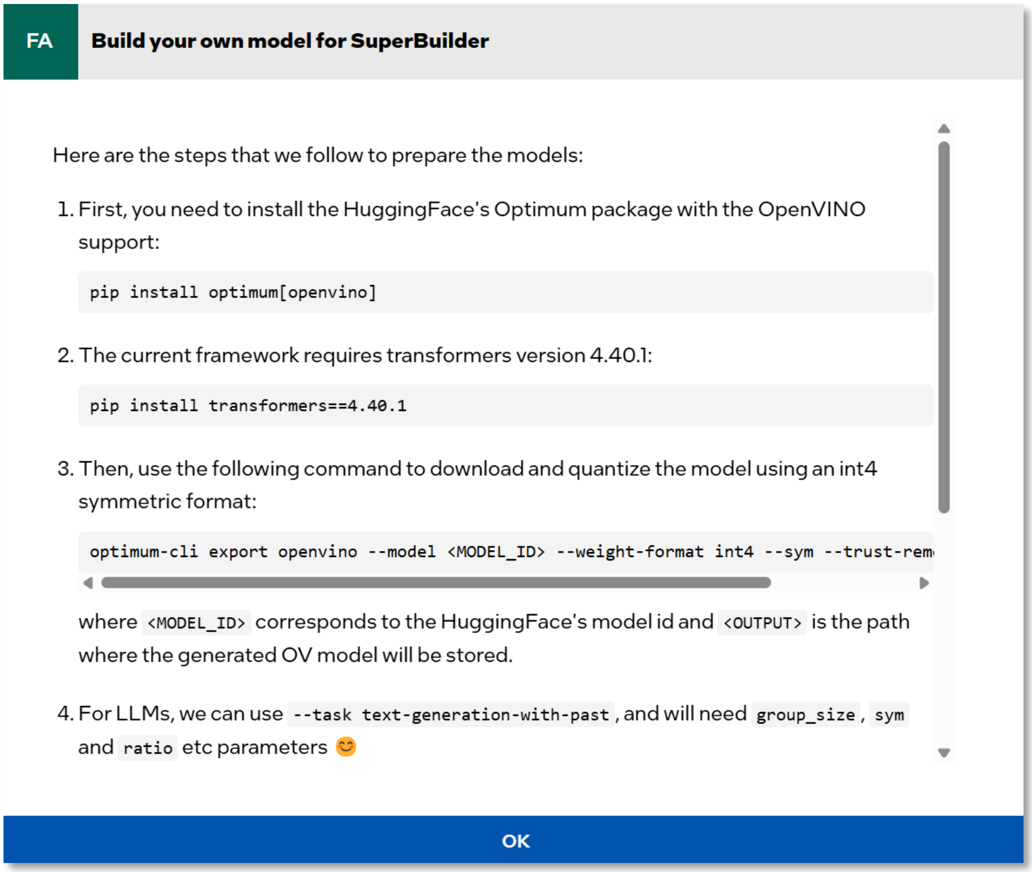
4.1 Upload a model

If you prefer to use a model which isn’t in the list of available models you may upload a model of your choice. The ℹ icon next to “Upload your models” label will provide some guidance in doing this. Please note the model needs to be properly converted to run with the Intel® AI Assistant

Builder software. The provided procedure may not work with all models. Users are responsible for verifying the feasibility and correctness of the procedure with their specific model(s). If you would like to receive personalized recommendations, please contact us.

Illustration of the model conversion instructions.

This is an illustration only. The full set of instructions is available in the application.



Build your own model for SuperBuilder

Here are the steps that we follow to prepare the models:

1. First, you need to install the HuggingFace's Optimum package with the OpenVINO support:

```
pip install optimum[openvino]
```

2. The current framework requires transformers version 4.40.1:

```
pip install transformers==4.40.1
```

3. Then, use the following command to download and quantize the model using an int4 symmetric format:

```
optimum-cli export openvino --model <MODEL_ID> --weight-format int4 --sym --trust-rem
```

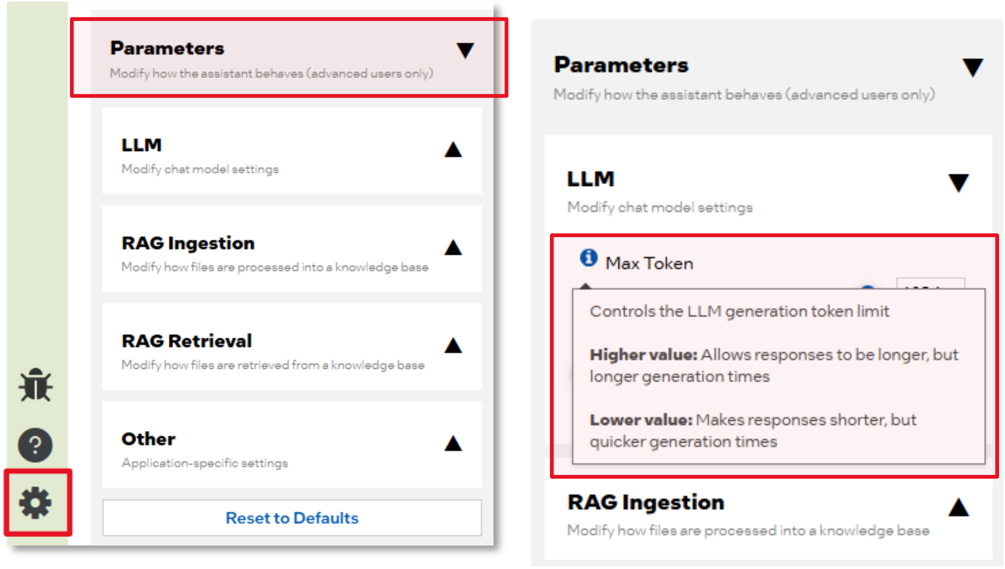
where `<MODEL_ID>` corresponds to the HuggingFace's model id and `<OUTPUT>` is the path where the generated OV model will be stored.

4. For LLMs, we can use `--task text-generation-with-past` , and will need `group_size` , `sym` and `ratio` etc parameters 😊

5.0 Parameter settings (“advanced” users only)

You can change the default settings for LLM, RAG, and other app processing from the “Parameters” section of the application settings.

Pro Tip: The “i” icon provides important detail for each parameter setting. It’s important to understand the impact of modifying these settings. As such we recommend only “advanced users” change these settings.

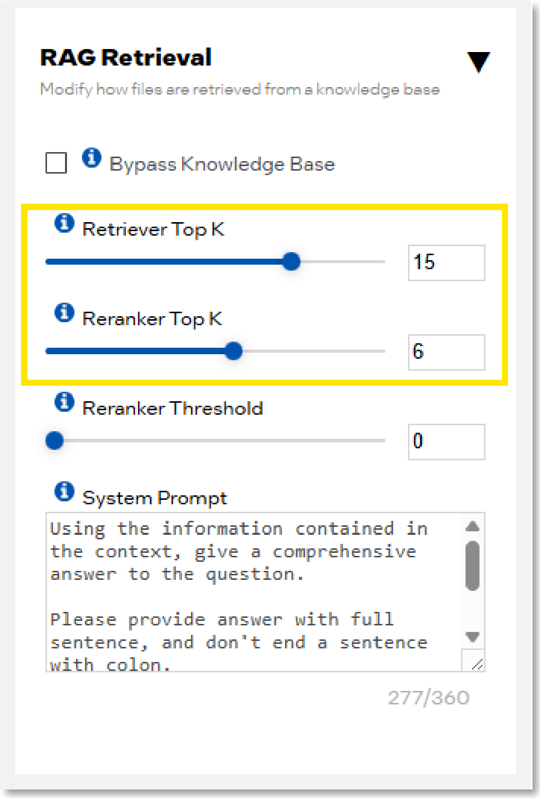

Parameters
Modify how the assistant behaves (advanced users only)
LLM
Modify chat model settings
RAG Ingestion
Modify how files are processed into a knowledge base
RAG Retrieval
Modify how files are retrieved from a knowledge base
Other
Application-specific settings
Reset to Defaults
Parameters
Modify how the assistant behaves (advanced users only)
LLM
Modify chat model settings
Max Token
RAG Ingestion
Modify how files are processed into a knowledge base

Controls the LLM generation token limit

Higher value: Allows responses to be longer, but longer generation times

Lower value: Makes responses shorter, but quicker generation times

Example: To retrieve more context, increase the Retriever TopK and Reranker TopK.



RAG Retrieval

Modify how files are retrieved from a knowledge base

□ Bypass Knowledge Base

Retriever Top K: 15

Reranker Top K: 6

Reranker Threshold: 0

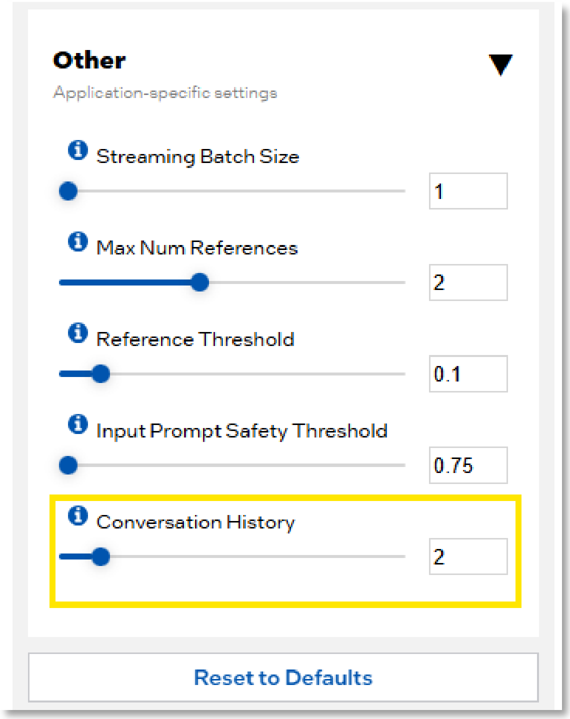
System Prompt

```
Using the information contained in
the context, give a comprehensive
answer to the question.

Please provide answer with full
sentence, and don't end a sentence
with colon.
```

277/360

Example: To include conversation history in the Chat context, update the Conversation History setting.



Other

Application-specific settings

Streaming Batch Size: 1

Max Num References: 2

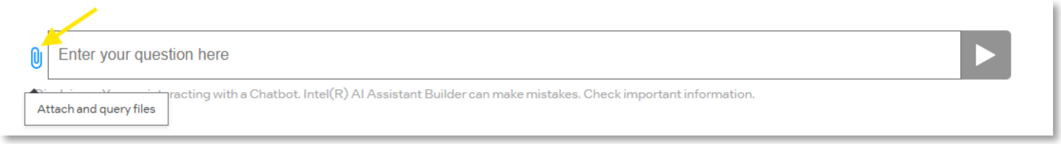
Reference Threshold: 0.1

Input Prompt Safety Threshold: 0.75

Conversation History: 2

6.0 Special Query Types

Special queries are used to “focus” the assistant on specific documents. This can yield better, more accurate results and help the assistant answer more sophisticated questions based on the context of the specified document(s). To use the special queries click the attach file icon (the paperclip icon next to the question entry field).

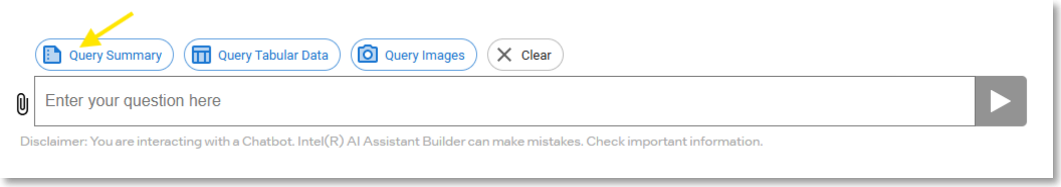


racting with a Chatbot. Intel(R) AI Assistant Builder can make mistakes. Check important information.

PRO TIP: Using small files which only contain information relevant to the topic at hand will yield the best results.

There are 3 types of special queries:

6.1 Query Summary

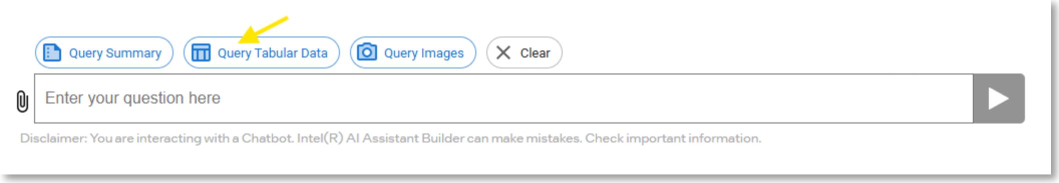


Disclaimer: You are interacting with a Chatbot. Intel(R) AI Assistant Builder can make mistakes. Check important information.

- Attach up to 3 files (PDF, DOCX, TXT)
- Generates a summary of each file, allowing users to ask follow-up questions based on the summary.

PRO Tip: This query works best when used with smaller files like resumes or whitepapers. If a file is too large, the system optimizes the experience by focusing on content from the beginning, middle, and end.

6.2 Query Tabular Data



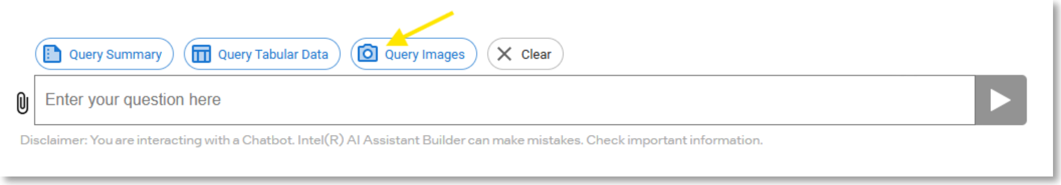
Disclaimer: You are interacting with a Chatbot. Intel(R) AI Assistant Builder can make mistakes. Check important information.

- Works with 1 file (XLSX, CSV)
- Requires data to be in tabular format, with headers in the first row and data in subsequent rows.
- Generates a description of the tabular data in each worksheet.
- Internally, converts the data into SQL tables, enabling users to ask natural language questions that are translated into SQL queries.

PRO Tip:

- Be as specific as possible when asking questions. E.g. Instead of "How many units did Alexander sell?", specify column names such as "How many units did the **salesman** Alexander sell?", where salesman is a column name.
- Also try to be specific with the column values. E.g. Instead of "Which salesman sold the most Cell Phones?", specify exact column values such as "which salesman sold the most units of **Cell Phone**? Where "salesman" and "Cell Phone" are column names"
- Often you can "get away" with asking natural language questions and still get an accurate answer, however accuracy and reliability will be improved if questions are specific and use words the assistant can reference within the document.

6.3 Query Images



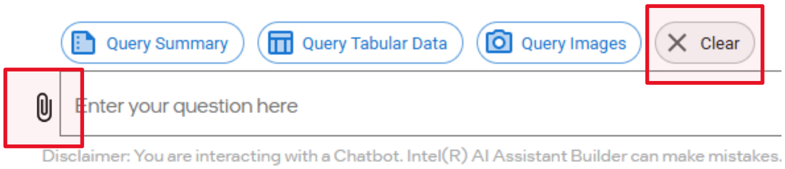
Disclaimer: You are interacting with a Chatbot. Intel(R) AI Assistant Builder can make mistakes. Check important information.

- Can attach up to **3 files (PNG, JPEG, JPG)**
- Generates a description of each image.
- This requires a vision model such as Phi-3.5-vision-instruct-int4-ov. Ensure it is selected in the App Settings → Models section.

PRO Tip: The model may occasionally misinterpret text extraction, especially with numbers or illegible text. Also, once the vision model is selected, it will become the default chat model for all app features, until the user switches to a different model.

6.4 Additional Notes

- Switching to a new query type overrides the previous one.
- If the attached file list changes for a selected query type, the system regenerates the task.
    - Example: If summaries were generated for files A and B, and the user later selects A, C, and D, all files will be summarized again.
- To close “Special query mode” click the “Clear” icon from the list of special queries. Starting a new chat session will also exit special query mode.



Enter your question here

Disclaimer: You are interacting with a Chatbot. Intel(R) AI Assistant Builder can make mistakes.

7.0 Troubleshooting

7.1 Updating or re-installing the application

Updating the application using the “update” function in the application settings does not currently work. This is a future enhancement we’re working on. To update to a new version or to re-install the application:

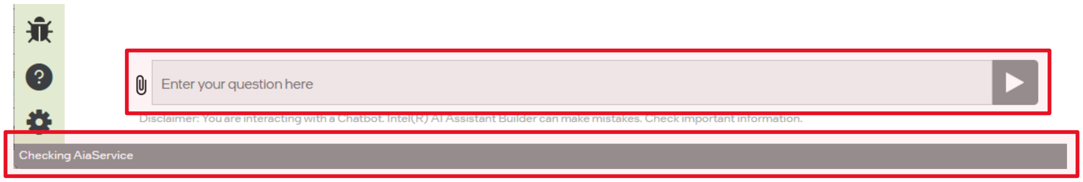
- Uninstall the application using the “Add or remove programs” function from the Microsoft® Windows control panel.
- Download the Intel® AI Assistant Builder
- Install the downloaded application

Note: In rare cases you may need to ensure a complete uninstall by verifying the following folders are removed. If these folders remain after uninstalling, you may delete them manually:

- C:\Program Files\Intel Corporation\Intel(R) AI Assistant Builder
- C:\ProgramData\IntelAIA (this folder can be hidden).

7.2 Initial load time

When the assistant is started the assistant service and models must be initialized before the assistant can be used. During this time the chat text entry field will be disabled and a status message at the bottom of the window is displayed indicating what is happening. When the assistant is ready the status bar at the bottom of the window is removed and the chat text entry is enabled. There are numerous factors that can affect the initialization time but typically it is from 20 seconds to 1 minute.



Disclaimer: You are interacting with a Chatbot. Intel(R) AI Assistant Builder can make mistakes. Check important information.

Checking AiaService

7.3 Not responsive

This rarely happens but if the assistant becomes unresponsive close and re-open the assistant.

7.4 Grayed Out Interface

On rare occasion, the assistant interface will disable and become “greyed out”. Typically right clicking anywhere in the interface and choosing “Refresh” will recover the assistant.

7.5 Model loading error

If a “model loading error” occurs, please make sure to update the GPU and NPU drivers to the latest version. The NPU model requires NPU Driver 32.0.100.3714 at a minimum.