

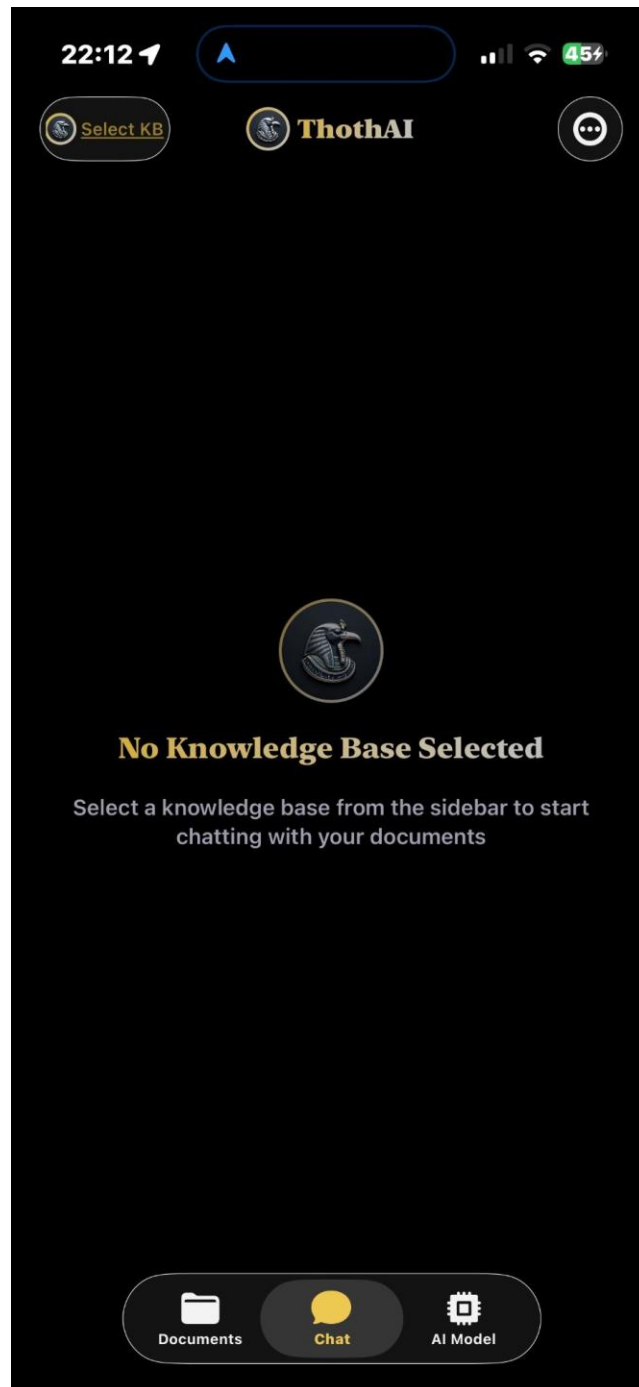
ThothAI

User Manual

Your private, offline AI document assistant.

1. Getting Started

When you first open ThothAI, you'll see the welcome screen.



ThothAI has two main modes:

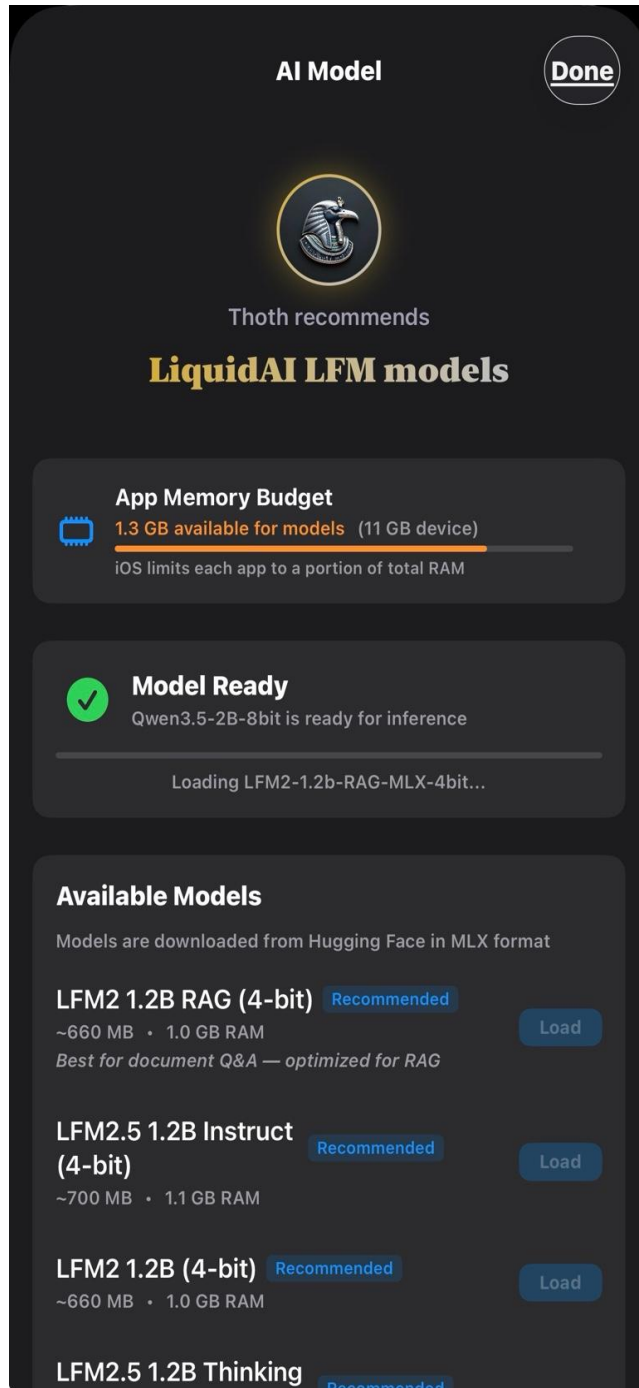
- **RAG Mode** — Ask questions about your PDF documents. The AI searches your files and gives answers based on what's in them.
- **Chatbot Mode** — Have a general conversation with the AI, no documents needed.

Before you can use either mode, you need to download the AI model.

2. Downloading the AI Model

The AI model is what powers ThothAI's intelligence. It's about 843 MB and downloads once. After that, everything works offline.

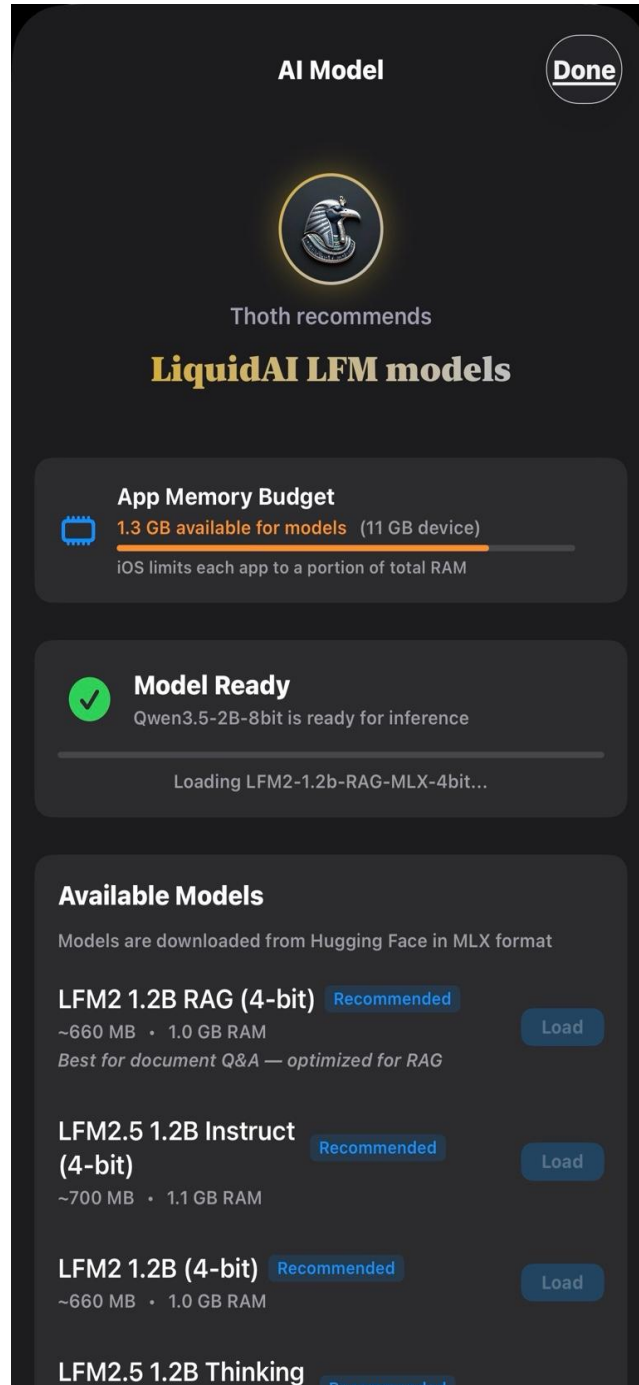
1. Tap Download AI Model on the welcome screen (or go to the AI Model tab)
2. Wait for the download to complete — you'll see a progress bar with download speed
3. Once finished, you'll see a green checkmark and "Model Ready"



📶 You need a Wi-Fi connection for the initial download. After that, ThothAI works completely offline.

Memory Requirements

The AI model needs about 1.5 GB of RAM to run. TothAI shows you how much memory is available on the Model Manager screen.

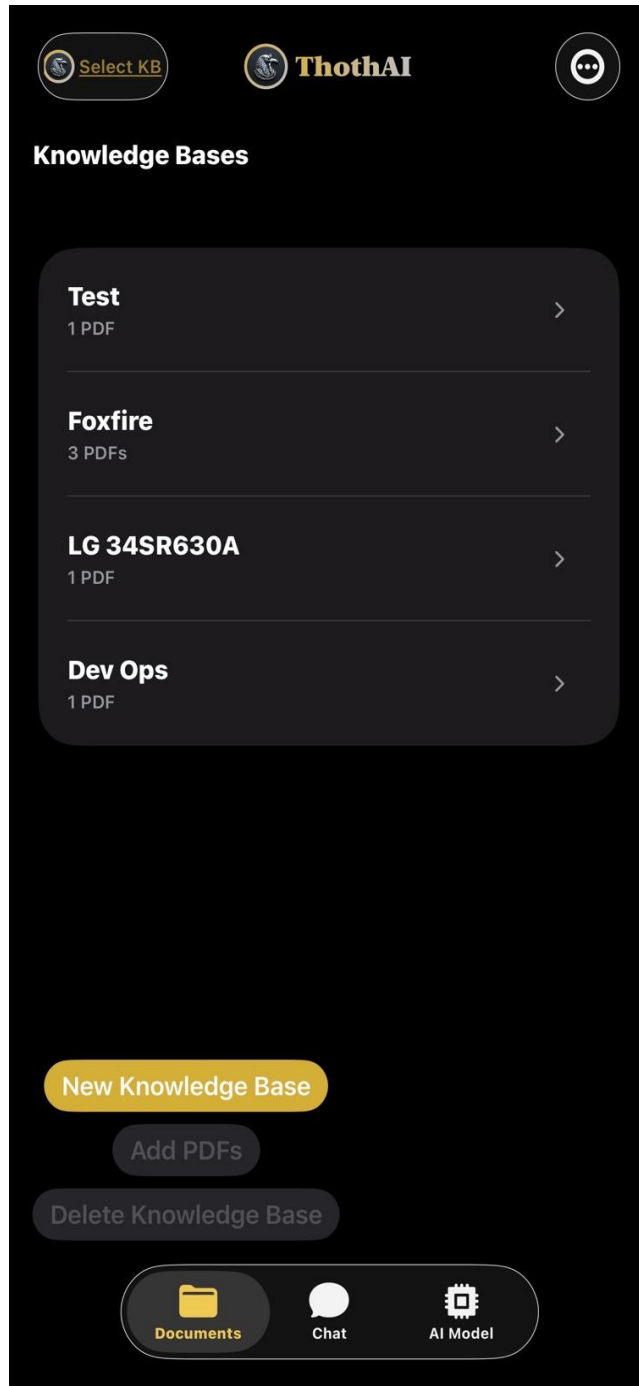


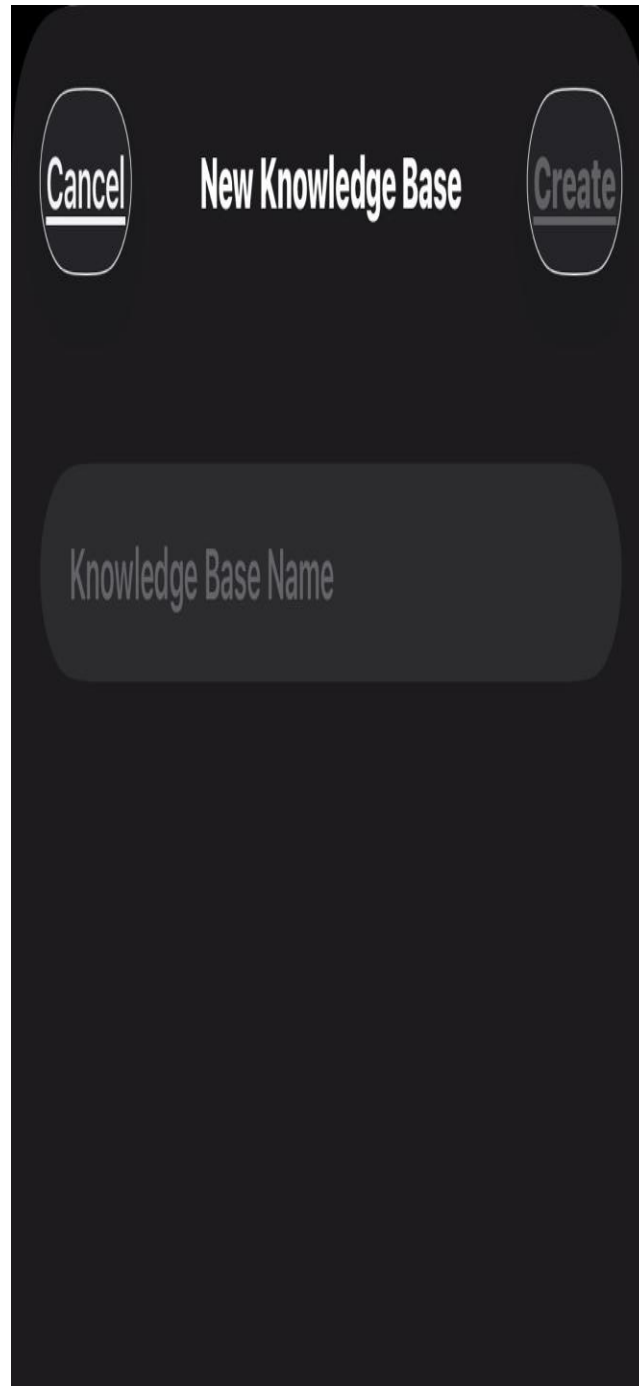
For best results, close other apps before using TothAI, especially if you notice slow performance.

3. Creating a Knowledge Base

A Knowledge Base is a collection of PDFs that ThothAI can search through. You can create multiple Knowledge Bases for different topics.

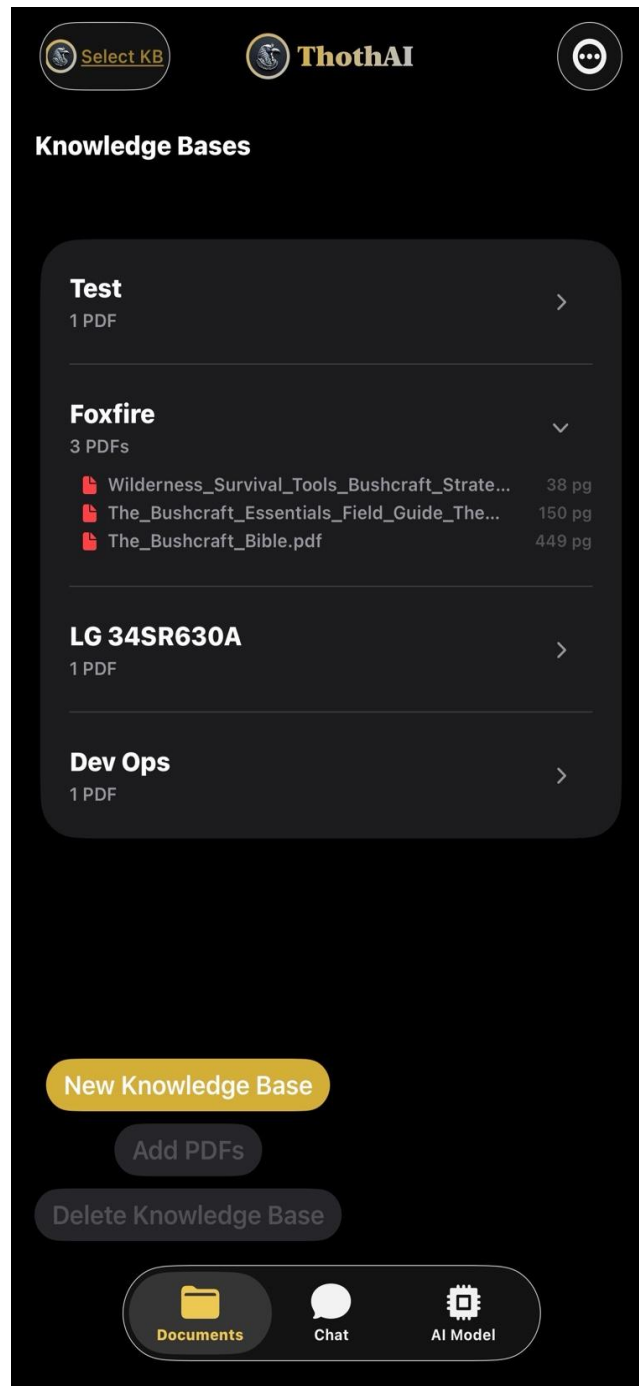
4. Go to the Documents tab
5. Tap New Knowledge Base
6. Enter a name (e.g., "Work Documents", "Recipes", "Study Notes")
7. Tap Create





Managing Knowledge Bases

- **Select** a Knowledge Base by tapping it
- **Expand** it to see the PDFs inside by tapping the arrow
- **Swipe left** on a Knowledge Base to delete it
- **Swipe right** to repair it (rescans and restores any missing files)

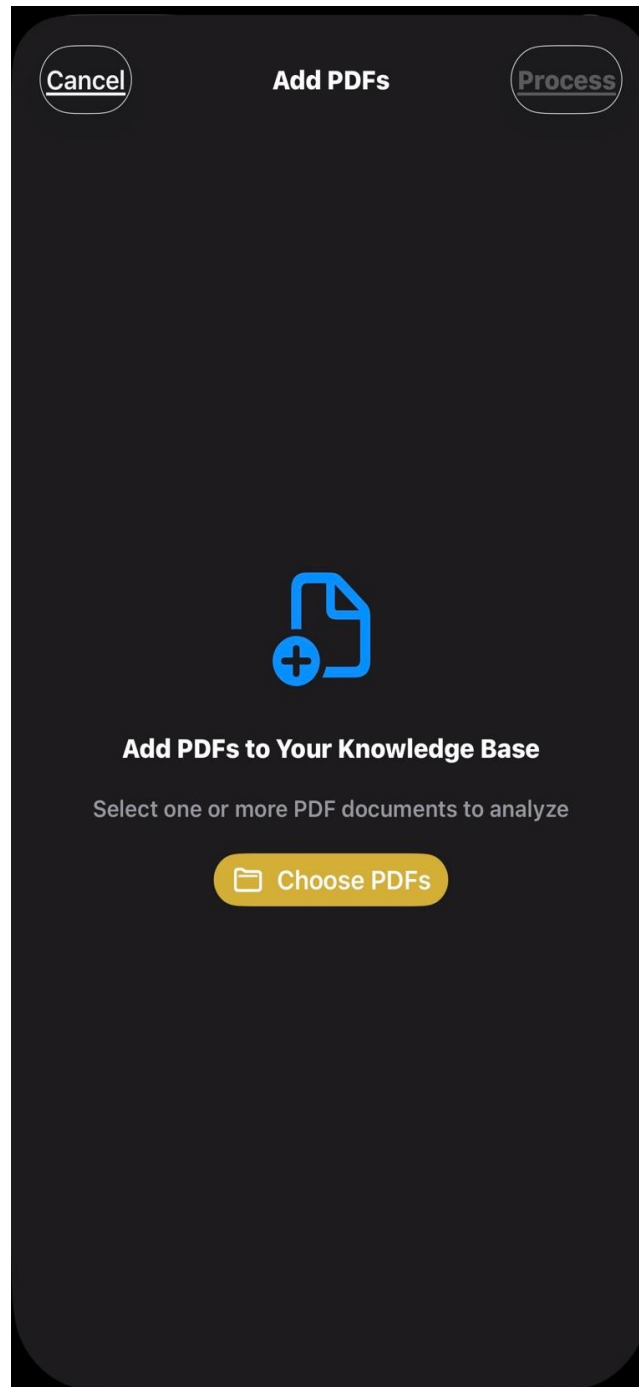


4. Adding PDFs

Once you have a Knowledge Base, you can add PDFs to it.

8. Select a Knowledge Base from the sidebar

9. Tap Add PDFs
10. Choose one or more PDF files from your device, iCloud Drive, or any connected cloud storage (Dropbox, Google Drive, etc.)
11. Wait for processing — ThothAI extracts the text and prepares it for AI search



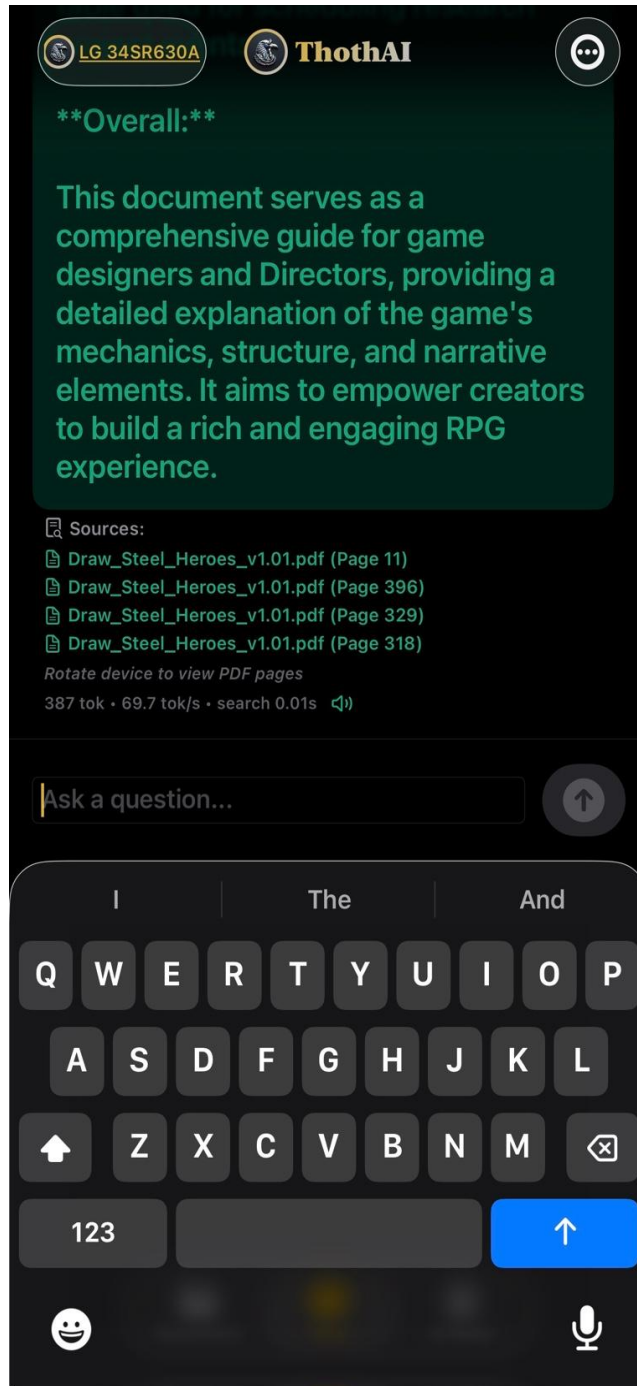
Processing time depends on the size of your PDFs. A typical document takes a few seconds; larger collections (hundreds of pages) may take a minute or two.

When complete, you'll see a confirmation and be returned to the main screen.

5. Asking Questions (RAG Mode)

RAG (Retrieval-Augmented Generation) mode is ThothAI's core feature. It searches your documents and generates answers based on what it finds.

12. Make sure **RAG Mode** is selected (check Settings > AI Mode)
13. Select a Knowledge Base
14. Type your question in the text field at the bottom
15. Tap the send button



Source Citations

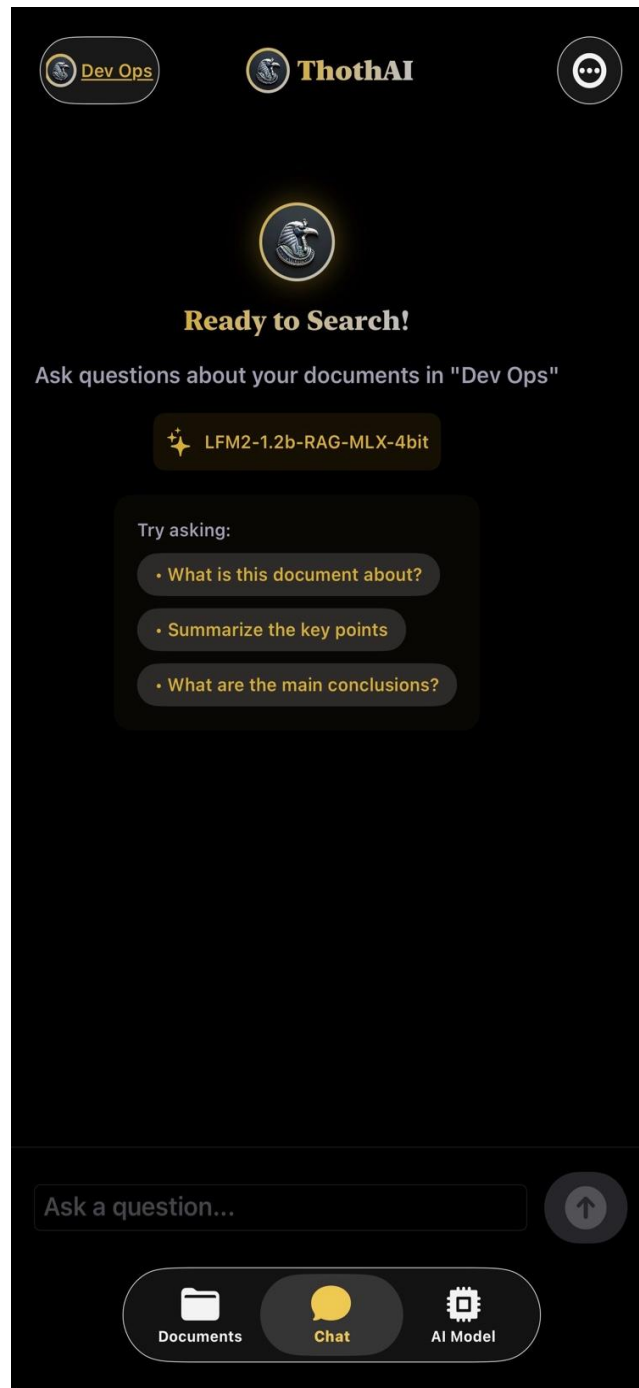
Every RAG answer includes source citations showing which documents and pages the AI used. Tap a source to jump to that page in the PDF viewer.

Sample Questions

When you first open a Knowledge Base, ThothAI suggests sample questions to get you started:

- "What is this document about?"

- "Can you summarize the key points?"
- "What are the main topics covered?"



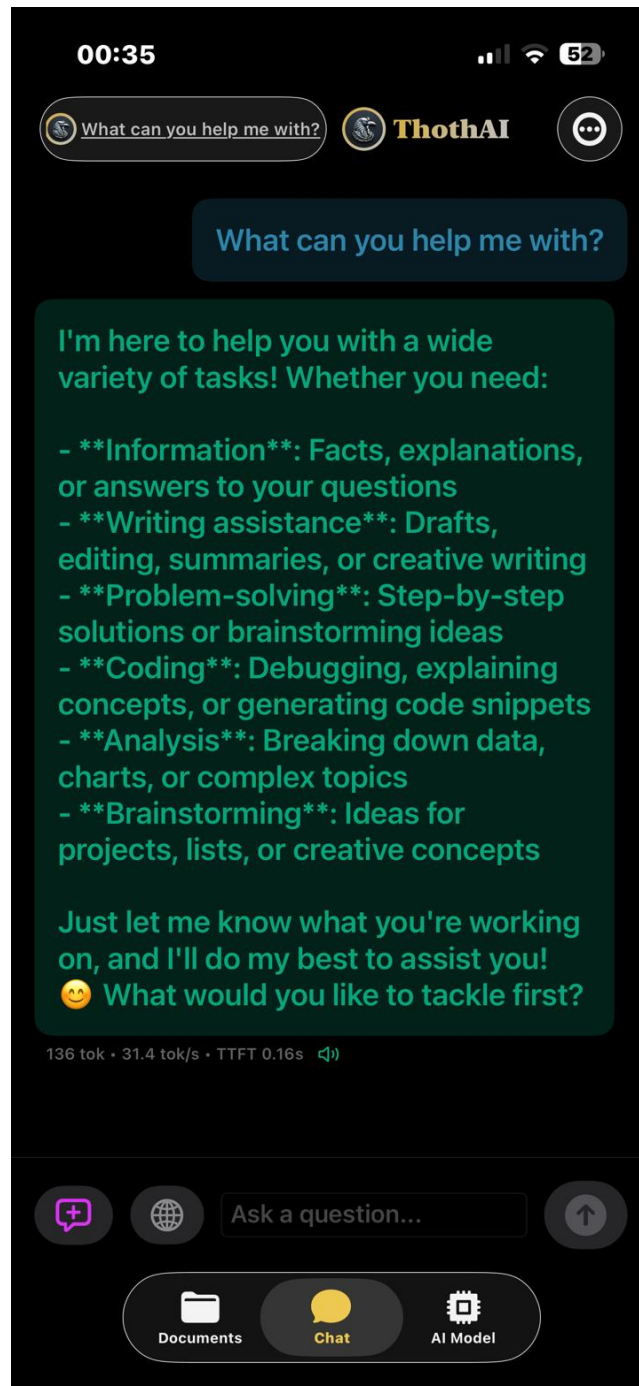
Generation Statistics

Each response shows performance stats: token count, generation speed (tokens/sec), and retrieval time. Tap to see details.

6. Chatbot Mode

Chatbot mode lets you have general AI conversations without needing any documents.

16. Go to **Settings > AI Mode** and select **Chatbot Mode**
17. Start typing in the chat
18. Responses stream in with a typing effect



Chatbot mode maintains conversation history, so the AI remembers what you've talked about within each conversation.

Stopping a Response

If the AI is generating a response and you want to stop it, tap the **red stop button** that replaces the send button during generation.

7. Conversations

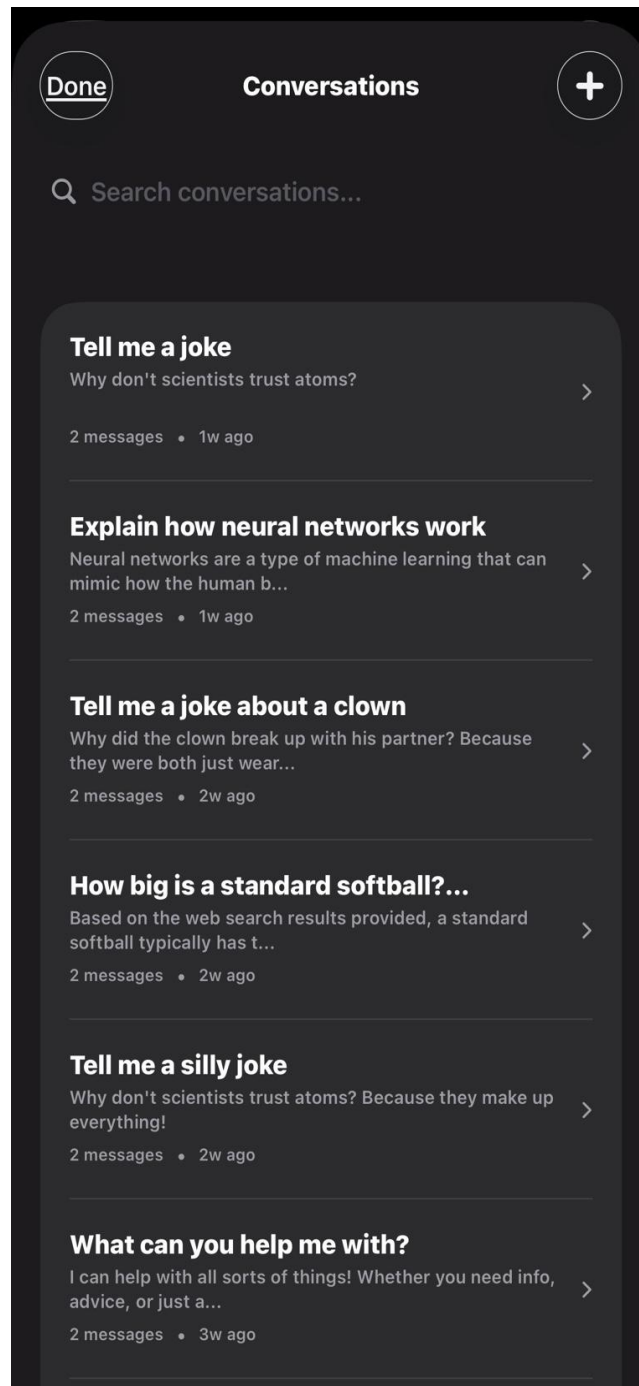
In Chatbot mode, your conversations are saved and organized.

Starting a New Conversation

Tap the **+ button** (bubble with plus) next to the text input to start a fresh conversation.

Switching Conversations

Tap the conversation title in the top navigation bar to open the **Conversation Picker**.



From here you can:

- **Search** conversations by title or content
- **Select** a conversation to continue it
- **Swipe left** to delete a conversation
- **Tap + New** to start a new one

8. Web Search

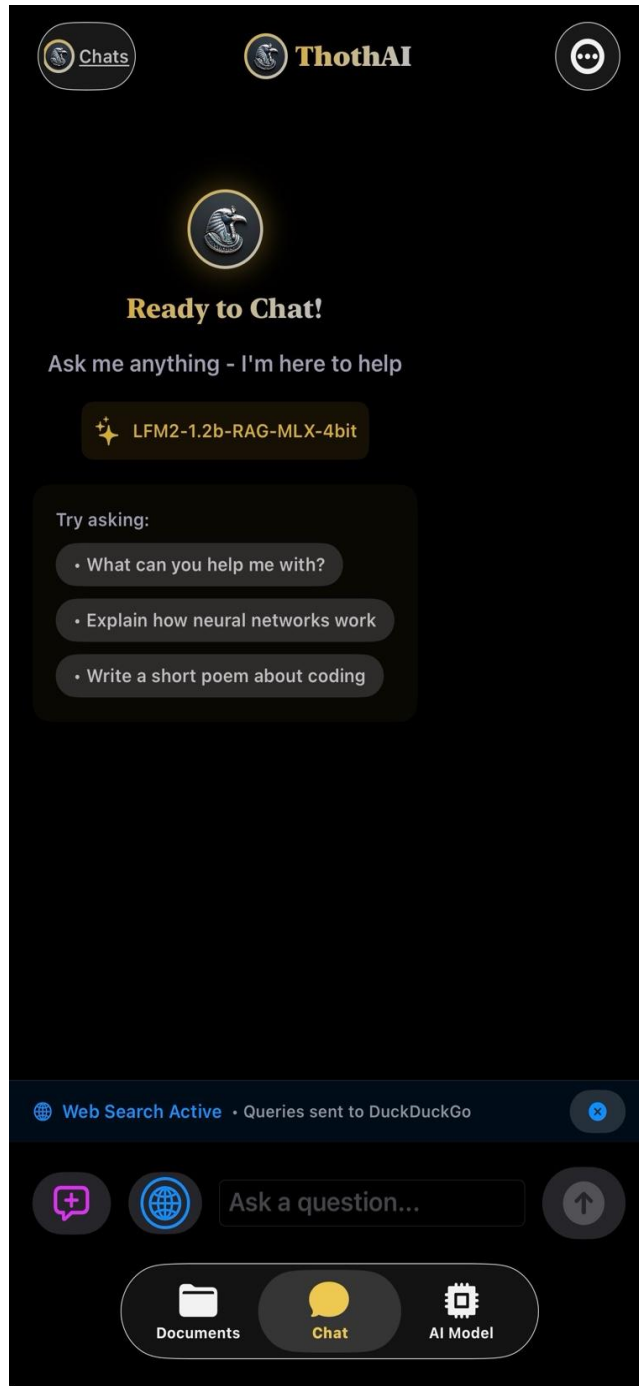
ThothAI can search the web to enhance AI responses in Chatbot mode. Searches go through DuckDuckGo for privacy.

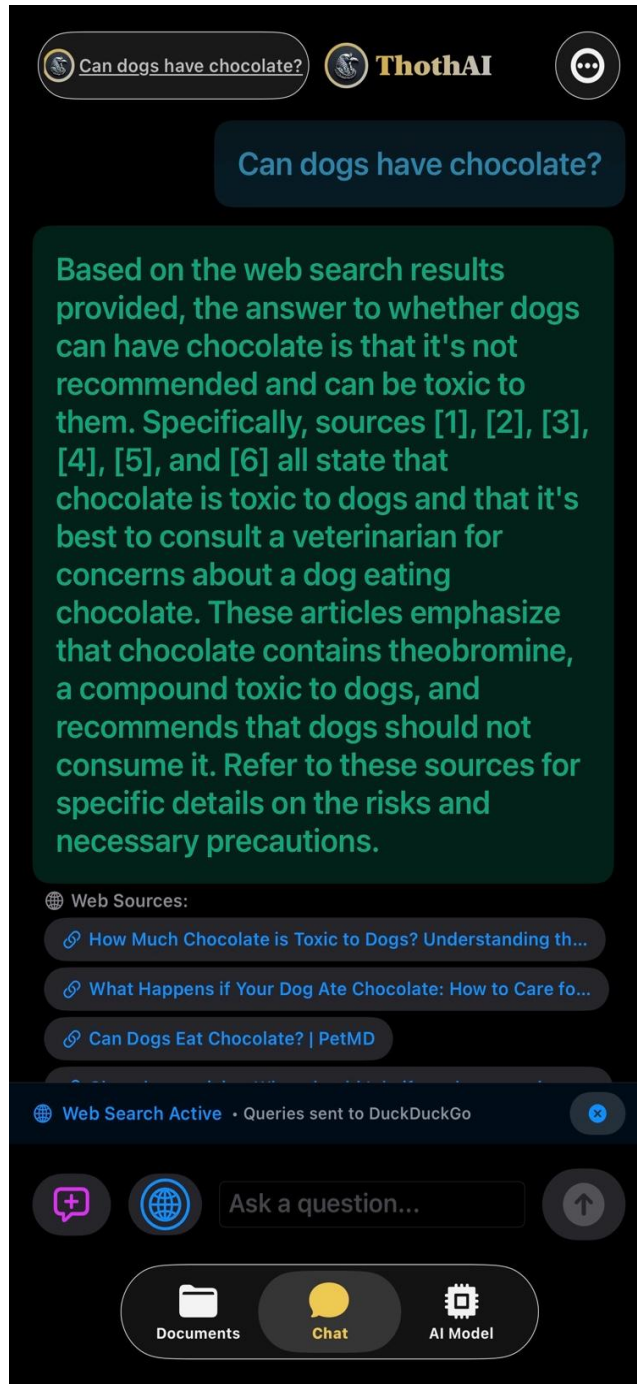
Enabling Web Search

19. Go to **Settings > Web Search**
20. Toggle **Enable Web Search** on


Using Web Search

21. In Chatbot mode, tap the **globe icon** next to the text input
22. The icon turns blue and a banner appears: *"Web Search Active"*
23. Ask your question — ThothAI will search the web first, then generate an answer using the results





Web search results appear as clickable links below the response. Tap a link to open it in your browser.

 Privacy note: When web search is active, your search query is sent to DuckDuckGo. All AI processing still happens entirely on your device.

9. Text-to-Speech

ThothAI can read AI responses aloud.

Enabling TTS

24. Go to **Settings > Audio & Accessibility**
25. Toggle **Enable Text-to-Speech** on
26. Optionally enable **Auto-play Responses** to have every AI response read automatically

Using TTS

Tap the **speaker icon** on any AI message to hear it read aloud. Tap again to stop.

TTS Settings

- **Voice:** Choose from available system voices
- **Speech Rate:** Adjust from 0.5x (slow) to 2.0x (fast), default is 1.0x

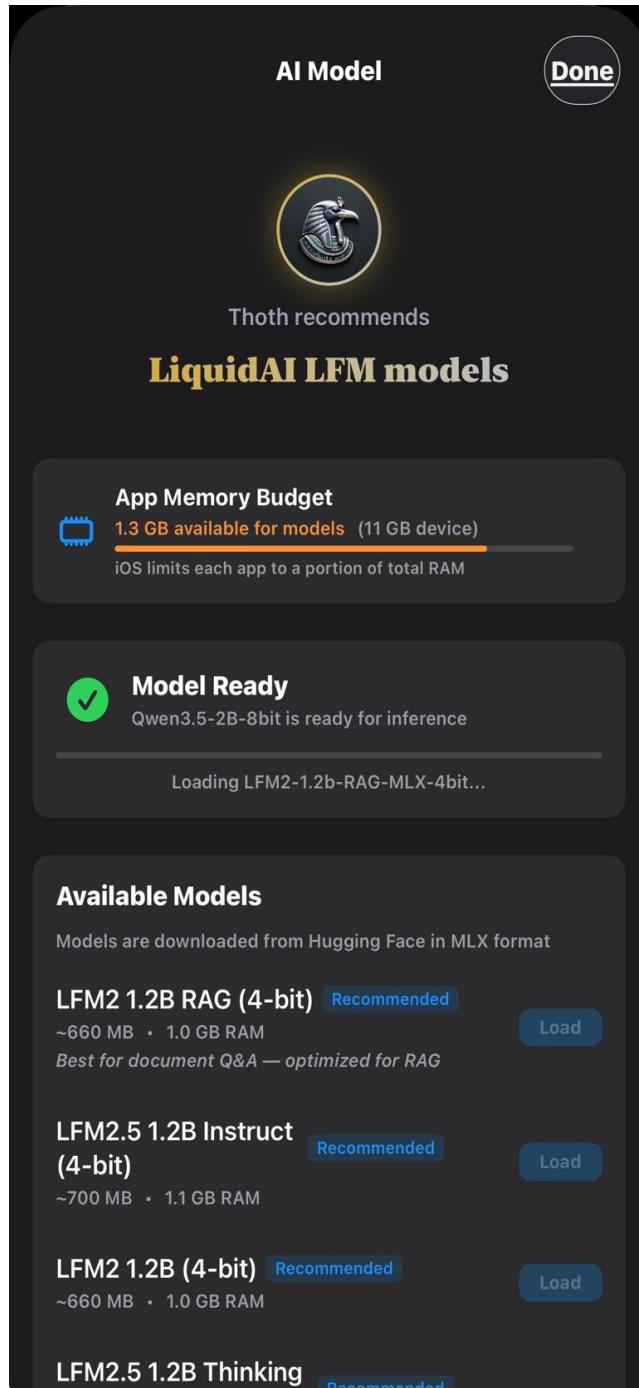
10. Managing Models

ThothAI comes with the recommended **LFM2-1.2B-RAG** model, optimized for document Q&A. You can also load other models.

ThothAI recommends the LFM2 (or LFM2.5), as they are optimized for mobile/edge devices, but any MLX model that fits in memory will work.

Model Manager

Go to the **AI Model** tab or **Settings > AI Model** to access the Model Manager.



Available Models

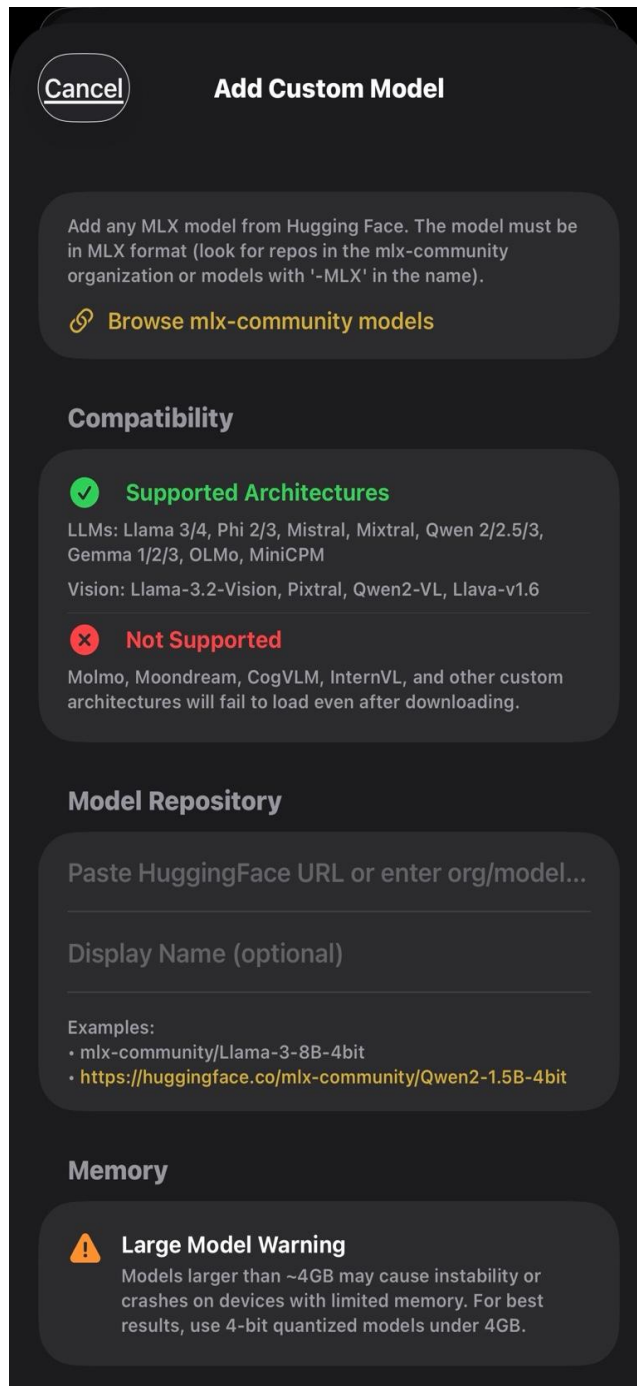
Model	Best For	Size
LFM2-1.2B-RAG	Document Q&A (Recommended)	~843 MB
LFM2-1.2B	General use	~843 MB


LFM2-1.2B-Extract	Data extraction	~843 MB
LFM2-1.2B-Tool	Function calling	~843 MB

Adding Custom Models

You can add models from Hugging Face in MLX format:

27. Tap **+** **Add** in the Custom Models section
28. Paste the Hugging Face repository name (e.g., *mlx-community/Llama-3.2-1B-Instruct-4bit*)
29. Tap **Add Model**



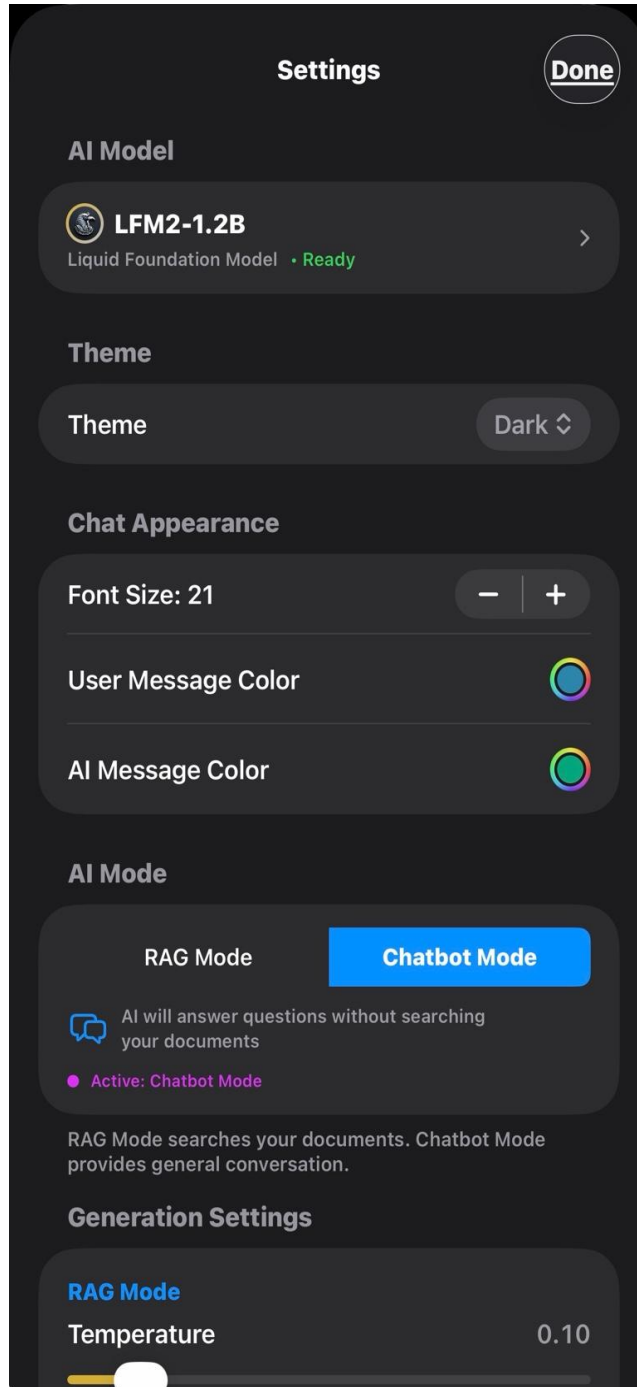
 Custom models must be in MLX format. Models larger than 4 GB may cause instability on devices with limited memory.

Deleting Downloaded Models

To free up storage, swipe left on any downloaded model in the **Downloaded Models** section to delete it.

11. Settings

Access Settings from the **menu button (...)** in the top navigation bar.



Theme

Choose between **Light**, **Dark**, or **System** (follows your device settings).

Chat Appearance

- **Font Size:** Adjust the text size of chat messages (8–32pt)
- **User Message Color:** Customize the color of your messages
- **AI Message Color:** Customize the color of AI responses

AI Mode

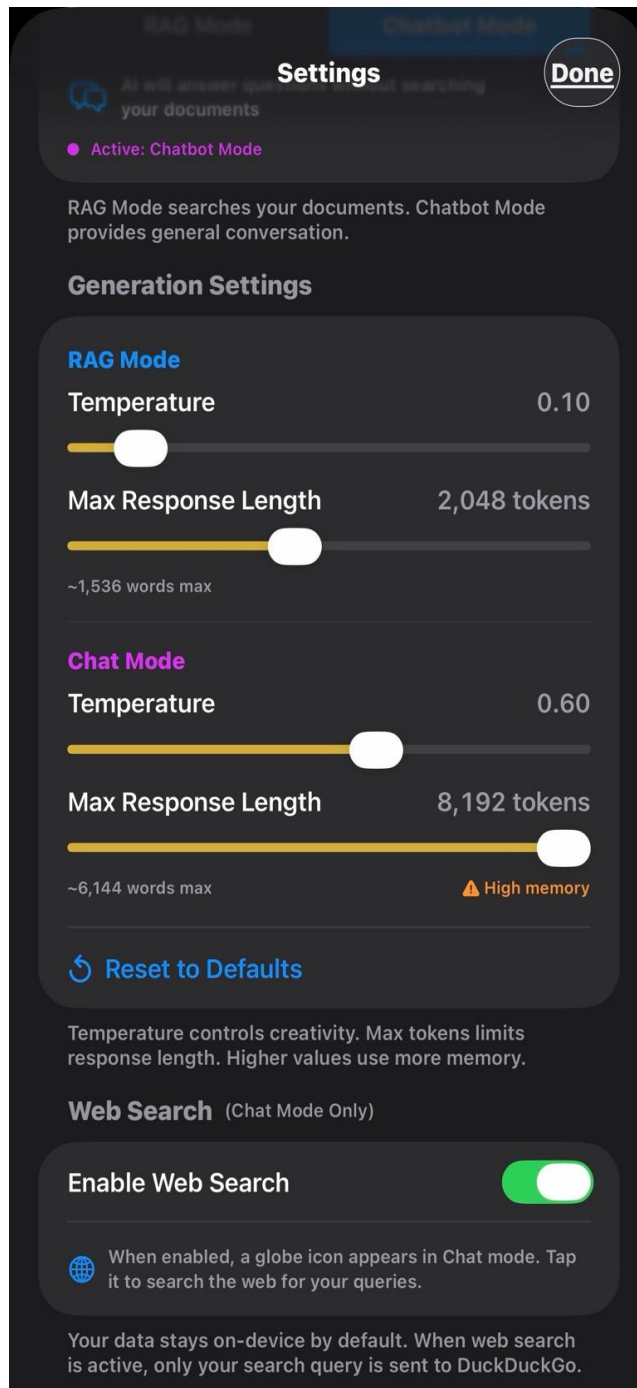
Switch between **RAG Mode** and **Chatbot Mode**. Switching modes clears the current chat.

Generation Settings

Fine-tune how the AI generates responses:

- **Temperature:** Controls creativity. Lower values (0.0–0.3) give more focused, factual answers. Higher values (0.5–1.0) give more creative, varied responses.
- **Max Response Length:** Controls how long responses can be, measured in tokens (roughly 0.75 words per token).

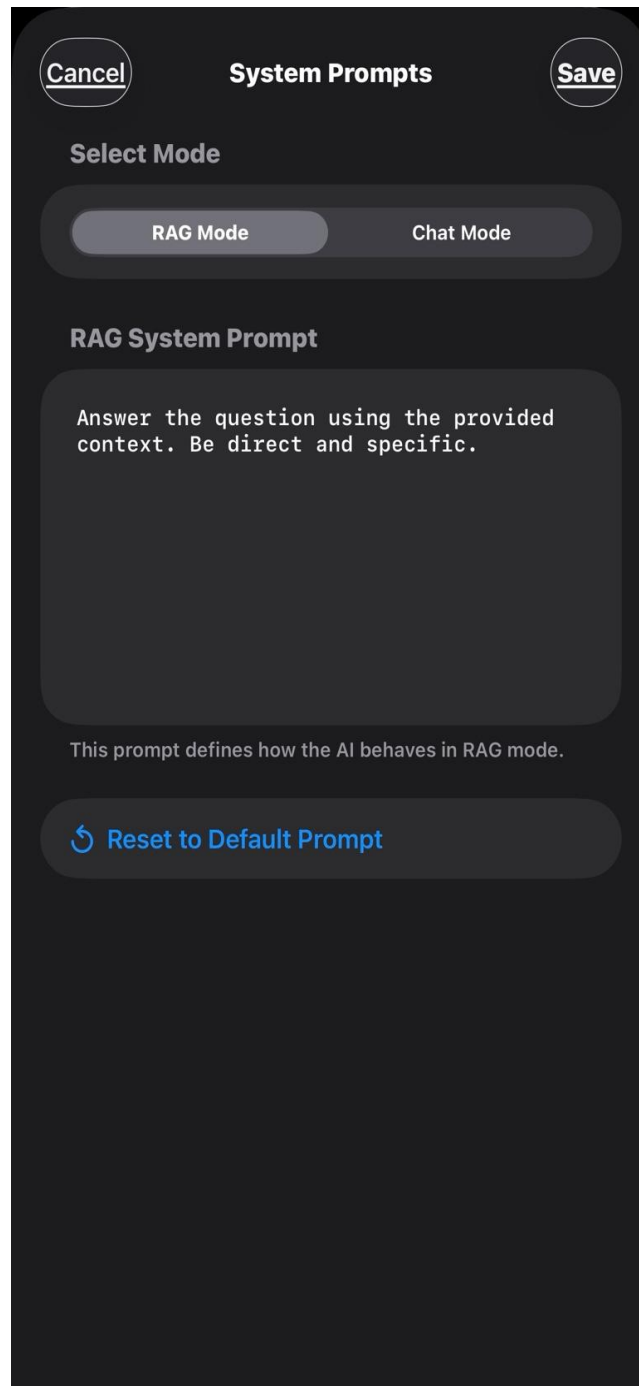
RAG mode and Chatbot mode have independent settings. Tap Reset to Defaults to restore factory settings.

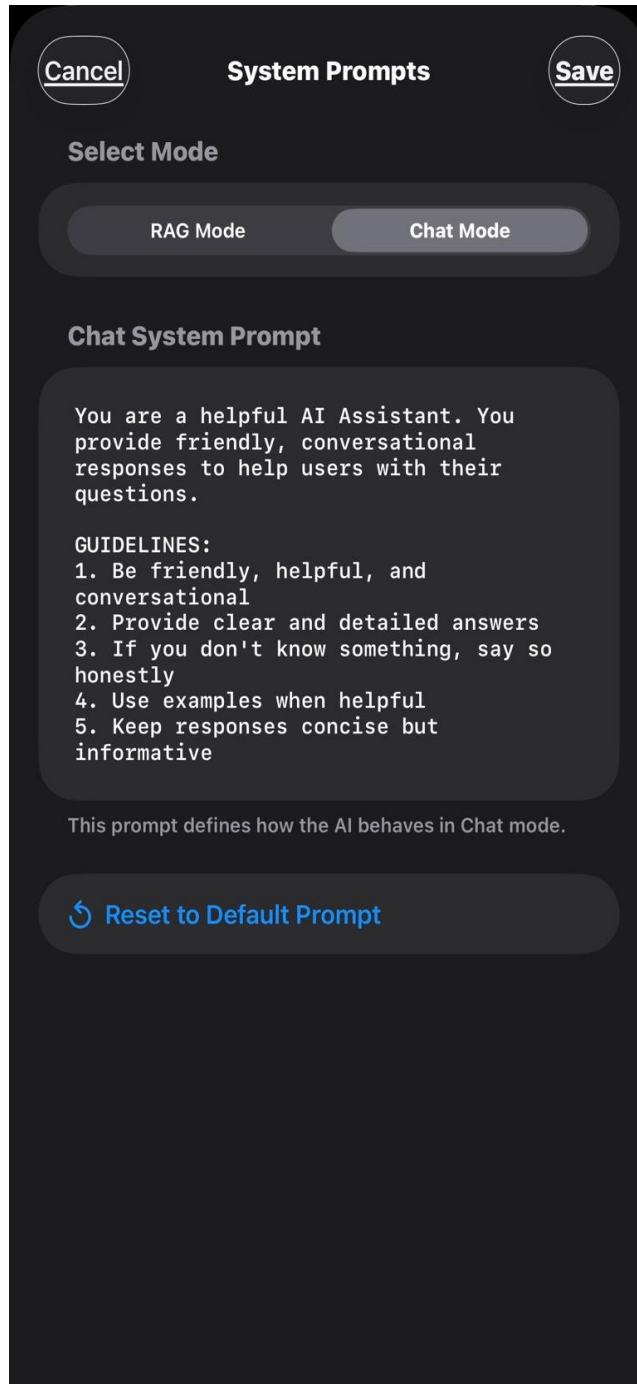


Prompt Settings

Customize the system prompts that guide the AI's behavior. Access from the **menu (...)** > **Prompt Settings**.

- **RAG Mode prompt:** Controls how the AI answers document questions
- **Chat Mode prompt:** Controls the AI's personality in conversations





12. Tips & Troubleshooting

Getting Better Answers

- **Be specific:** "What does section 3.2 say about deadlines?" works better than "Tell me about deadlines"
- **Try rephrasing:** If you don't get a good answer, ask the same question differently

- **Use smaller Knowledge Bases:** Putting related documents together gives more focused results
- **Adjust temperature:** Use lower temperature (0.0–0.2) for factual questions, higher (0.5–0.7) for brainstorming

Performance Tips

- **Close other apps** before using ThothAI to free up memory
- **Avoid very long conversations** in Chatbot mode — start a new conversation periodically
- **Smaller PDFs process faster** — split very large documents if possible

App Crashes or Closes Unexpectedly

If ThothAI closes suddenly during use, it's likely a memory issue. The AI model requires significant RAM, and iOS may terminate the app if memory runs low.

What to do:

30. Force-quit all other apps
31. Reopen ThothAI
32. If the issue persists, try a smaller model from the Model Manager

If the app detects a previous crash during model loading, it will show a warning and suggest choosing a different model.

PDFs Not Working

- **Scanned PDFs** (images of pages) are not currently supported — ThothAI needs PDFs with selectable text
- **Password-protected PDFs** must be unlocked before import
- If a PDF shows 0 chunks after import, the text extraction failed — try a different version of the document

Web Search Not Working

- Make sure **Enable Web Search** is turned on in Settings
- Web search requires an internet connection
- The globe icon must be **blue (active)** for each conversation where you want web search

Reporting Issues

If you encounter a bug, please provide feedback through TestFlight or contact us with:

- What you were doing when the issue occurred
- Your device model
- Whether the issue is reproducible

Privacy

ThothAI is designed with privacy as a core principle:

- **All AI processing happens on your device** — your documents and conversations never leave your iPhone
- **No accounts required** — no sign-up, no login, no tracking
- **No analytics or telemetry** — we don't collect any usage data
- **Web search (when enabled)** sends only the search query to DuckDuckGo — no personal data, no identifiers
- **iCloud sync (if enabled)** keeps your documents within your personal iCloud account

Your data is yours. Always.

ThothAI v1.0 — Powered by Liquid AI's LFM2 and Apple MLX