

AethronChat v2

Private Intelligence. Anywhere.

INSTRUCTION MANUAL | VERSION 2.0.0

A private Android assistant with an embedded 1.5-billion-parameter model. All conversation processing runs locally on the phone CPU. No account, internet connection or model download is needed.

Start in four steps

Step	Action
1. Install	Open the supplied AethronChat-v2.0.0-arm64.apk on your phone. If an older version is installed, choose **Update**.
2. Prepare	Open AethronChat v2. On first launch, let it verify and prepare the model in its private app directory.
3. Chat	Wait for the CPU-ready indicator. Type a message and tap the send arrow.
4. Exit	From the main chat with the keyboard closed, press the phone Back button twice within two seconds.

Updating? Do not uninstall first. An in-place update keeps conversations, memories and saved settings. A verified existing model file is reused.

Inside this guide

2 Chat and response display
3 Documents, exports and history
4 Personalities and persistent memory
5 Context, CPU settings and model storage
6 Troubleshooting, privacy and reference

Developed by S. van Ingen
Copyright © 2026 S. van Ingen
Manual issued 22 November 2026

Chat and response display

The main screen

Control	What it does
Plus (+)	Starts a new conversation. The next message uses your current personality settings.
History	Opens saved conversations and search.
Settings	Opens appearance, personality, model, CPU performance, memory and About options.
Paperclip	Imports a text document for the assistant to read.
Send / Stop	Sends your message, or stops a response that is being generated.
Clear input (x)	Clears the draft in the message box.
Export	Saves the current conversation as plain text or Markdown.

Read, copy and regenerate

Responses appear progressively. Streaming text is kept plain and Unicode-safe while generation is running; completed responses receive Markdown and formula formatting. This avoids repeatedly reformatting half-written equations or code blocks. Allow the response to finish before judging its final layout.

Tap the copy icon in a message to copy its original text. **Regenerate** replaces the latest assistant response. Swipe through the conversation to read older text; use the down-arrow control to follow the latest response again.

Equations, code and drawings

Completed replies support Markdown and LaTeX equations, including fractions, powers, subscripts, integrals and matrices. Common math delimiters and fenced math/LaTeX blocks are recognized. Unrenderable formulas remain available as source rather than disappearing.

PostScript/EPS code remains readable and may show a basic vector preview. The preview supports a limited set of single-page drawing commands; unsupported commands show a preview-unavailable message. It is not a full PostScript interpreter.

Back, keyboard and backgrounding

Back first dismisses the keyboard or an open dialog. From a settings/history page, Back returns to chat. At the main chat, the first Back shows **Press Back again to exit**; the second within two seconds closes the app. If a response is running, it is stopped and its partial text is saved before exit. Switching to another app also requests cancellation. Rotation keeps the current conversation.

Documents, exports and history

Ask questions about a text document

1. Tap the paperclip while the assistant is idle.
2. Select a local text file using Android's file picker.
3. Review the filename, preview and prompt-token estimate.
4. Enter what you want done, then tap **Send document**.

Example task	Suggested instruction
Answer questions	Answer every numbered question. Keep the original numbering.
Summarise	Summarise this document in five bullet points, then list open questions.
Rewrite	Rewrite this document in clear, professional English. Preserve all figures and dates.
Review	Identify inconsistencies and explain each proposed correction.

Supported inputs are UTF-8 text and BOM-marked UTF-16 text, up to **256 KB**. TXT, Markdown, CSV, JSON and XML are text-based examples. Word documents, PDFs, images and binary files are not supported by this importer.

The preview may be shortened, but the full accepted document is sent. If it cannot fit the context, sending is disabled: choose a smaller document, increase the context in Model settings, or reduce maximum output tokens. Imported content becomes a saved user message. The original file is never modified.

Export a conversation

At the bottom of chat, tap **Export**, choose **Text (.txt)** or **Markdown (.md)**, and select a destination. Export is available when generation has finished. The saved file contains the title and ordered user/assistant messages. Export preserves source text, not rendered images of formulas or drawings.

Manage history

Open History to search titles and message contents. Tap a conversation to resume it. The pencil renames it; the bin deletes it after confirmation. **Delete all conversations** also requires confirmation and cannot be undone. Persistent memory is separate from conversation history.

To keep document handling entirely local, select an on-device folder in Android's picker. The picker may also list storage-provider apps installed on your phone.

Personalities and memory

Choose how the assistant responds

Open **Settings** → **Personality**. Set an assistant name, then choose Default Assistant, Technical Expert, Conversational, Creative or Custom. For presets, you can choose response length, response style, technical detail and additional behavioural instructions. Tap **Save settings**.

Personality changes apply to **new conversations**. Existing conversations retain the system prompt they were created with. Use the plus button to begin a chat with your new personality.

Use a complete custom prompt

The Custom preset gives you the complete editable application-level system prompt. For example:

Example custom prompt

You are Aethron, a practical electronics tutor. Explain calculations step by step, state units, and ask for missing circuit values. Use concise paragraphs and include equations where helpful.

Custom replaces the preset instructions. If persistent memory is enabled, relevant saved memories can still be appended as reference information. Disable memory when you want a conversation without those references.

Optional persistent memory

Open **Settings** → **Memory**, enable **Use persistent memory**, and save. **Manage memories** lets you add a category and memory text, edit records, switch them on/off, delete individual records or clear them all after confirmation.

Before answering, the app selects enabled memories by keyword relevance and recency within a token budget. Unrelated memories are not automatically inserted. Memory is stored privately on the phone and survives normal app restarts and in-place updates.

Suggested memories

With **Propose memories after responses** enabled, the local model can suggest durable facts from a conversation. Review each suggestion and tap **Save** only if you want to keep it, or tap **Dismiss**. This optional extra inference can add time after a response.

Appearance and defaults

General settings provide System, Dark and Light appearance, a 12-24 sp font-size setting, resume-last-conversation and optional performance statistics. **Reset defaults** prepares default settings; tap Save settings to apply them. It does not delete conversation or memory records.

Context, CPU and model storage

Context window

The context window is the token budget shared by the system prompt, history, imported document, relevant memories and response. New/default settings use **16,384 tokens**. Existing saved choices remain unchanged after an update. Use **Settings → Model → Use recommended 16,384-token context → Save settings** to apply it.

Choose 4,096 or 8,192 for shorter chats and lower memory allocation. Up to 32,768 is available. Changing context reloads the model. Longer input takes more time to process. If the budget fills, older complete turns are omitted from inference while the full history remains saved; an oversized newest message is rejected rather than silently shortened.

Model setting	Default / purpose
Temperature	0.7. Higher is more varied; 0 selects greedy output.
Top-p / Top-k	0.9 / 40. Limit the token candidates considered.
Repeat penalty	1.1. Discourages repeated tokens.
Maximum output tokens	1,024. Must leave at least 128 context tokens for the prompt.
Seed	-1 chooses a random seed. A fixed value helps repeat a sampling run.

CPU performance

In Performance settings, CPU threads set to **0** use automatic selection, up to six threads. You can choose a manual count. Severe thermal pressure stops generation and limits later requests to two threads; allow a hot phone to cool.

Optional statistics show live token progress, completed-response tokens/second, prompt speed, time to first token, context usage, model load time, approximate RAM, app CPU utilization and readable CPU clocks. Restricted clocks are marked unavailable. Use **Benchmark active backend** for a local CPU test.

Where the model lives

The APK contains the model. On **first launch**, the app verifies it and prepares one file in its own Android-private working directory: **noBackupFilesDir/model/**. Nothing is extracted to Downloads, shared storage or another app's folder.

The same path is reused across updates. Verified files are retained; invalid files are replaced and stale partial files are cleaned up. The APK asset plus the one prepared file use about **1.97 GB** for the model. Allow at least 3 GB free for installation/preparation. Android runs model preparation at first launch, not during package installation.

Troubleshooting and reference

If you see this	What to do
Preparation needs space	Free storage and tap Retry. The model preparation area is private to the app. An interrupted copy is discarded and rebuilt.
Model or context allocation fails	Tap Reset context and retry at startup. Recovery uses a conservative 4,096-token context. Close other heavy apps if necessary.
Document is too large	Use a smaller document, increase context within the model limit, or reduce output tokens. The importer does not silently trim content.
Slow response	Shorten the prompt/history, try automatic CPU threads, and check whether the phone is hot. Longer documents require more prompt-processing time.
Formula is shown as source	The equation may be incomplete or unsupported. Let generation finish, or ask for a simpler LaTeX expression. The original text remains copyable/exportable.
Response says stopped/interrupted	The app was stopped, backgrounded or restarted during generation. Review the saved partial response and use Regenerate if needed.
Update cannot be installed	Use the supplied ARM64 APK and ensure sufficient space. An in-place update requires the same signing identity. Do not uninstall merely to troubleshoot without exporting important chats first.

Privacy and local data

There is no INTERNET permission, cloud inference, analytics, registration or model download. Conversations, memory, configuration and prepared model storage are app-private. Android backup/device transfer is disabled for this data. Exported files are stored wherever you explicitly select.

Uninstalling the app or clearing its storage removes private conversations, memories, settings and the prepared model. Export important conversations before doing so. Exported conversations are readable documents; the app does not currently import them as reconstructed chat history.

About AethronChat v2

Version 2.0.0, build 3. Embedded model: Qwen2.5 1.5B Uncensored Q4_K_M GGUF. Processing: ARM64 CPU with llama.cpp. Model Information displays metadata from the actual GGUF. About contains developer details and third-party licences.

Developed by S. van Ingen. Copyright © 2026 S. van Ingen.
AI-generated answers can contain mistakes. Review answers and calculations before relying on them.