

What changed in Gnosis v2

Audit of v1 → fixes

v1 problem	v2 fix
Chat failed with <i>"input length exceeds the context length"</i>	<code>num_ctx</code> sent explicitly, token-budgeted prompt (excerpts, history, giant-paste trimming)
Replies ignored the conversation	Multi-turn history is sent, newest-first within budget
Memory/vault weren't really used; "not learning"	Hybrid retrieval (BM25 + embeddings + RRF) with contextual headers; past chats indexed; source chips show what was used; profile + advisory style in every prompt
Every save re-embedded everything; save failed if Ollama was down	Chunks are searchable instantly; embeddings fill in the background, incrementally; Ollama outages never block saving
"Remember that..." unreliable	Deterministic capture grammar + de-duplication + confirmation badge
Self-learning was implicit	Suggestions inbox (accept/dismiss) — transparent, never silent
Plain-text API key	OS keychain / encrypted file; never returned by the API; excluded from backups
Any website could call the local API (CSRF / DNS rebinding)	Per-run token, Host and Origin checks, CSP
<code>innerHTML</code> with chat text (XSS)	DOM-built UI, escaped Markdown allow-list, strict CSP
Backup import could write outside the data folder	Allow-list extraction (flat names, regular files only)
Data folder inside Documents/OneDrive (SQLite corruption, pip errors on Windows)	OS app-data folder; Windows launcher keeps its venv in <code>%LOCALAPPDATA%</code>
Stale cloud model name; cloud sent whatever was typed	Current model default; mandatory redacted preview → confirm; audit log
Basic UI, no history, alignment issues	Claude-style workspace: collapsible chat history, pinned/grouped chats, left-aligned messages, light/dark, responsive
Errors were raw 500s	Typed errors with one-click fixes (install model, retry), global JSON error handling
No tests	57 automated tests incl. real-browser end-to-end

Behaviour you will notice

- **First run is guided:** banners offer **Install** for the chat and memory models.
- **Source chips** under answers; click to open the memory.
- **Memory page** handles every file type, shows "Reading.../Needs attention" per file, and lets you edit extracted text.

- **Pinned memories** ("Always keep in mind") are in every chat.
- **Cloud** mode shows a review window with exactly what leaves the machine.

Upgrading from v1

v2 keeps its data in a new location and a new schema, so v1 data is not touched.

1. Keep your v1 folder as it is (it is your backup).
2. Start v2 (`start.command` / `start.bat`).
3. Bring your files across: v1's uploaded documents and notes can be re-imported in one go:
`python run.py --import-folder "/path/to/old/gnosis/memory"` (or upload them in *Memory*).
4. Re-enter your profile in *Settings* → *Profile* and, if you use it, your cloud API key in *Settings* → *Cloud*.
5. Old v1 `.gnosis` backups are **not** readable by v2 (different format); import the underlying files as above.