

Guia de instalação: Hindsight

Memória de longo prazo para **agentes de IA**: o agente guarda o que aprende e lembra na próxima conversa. Gratuito e de código aberto (MIT).

01 Antes de instalar

- Um PC ou servidor com **Docker** (o jeito recomendado), ou **Python** com pip.
- **Uma chave de LLM** (OpenAI, Anthropic, Gemini, Groq e outros) ou um modelo local (Ollama, LM Studio). O Hindsight usa o LLM pra extrair e organizar as memórias.
- Roda em **Windows, Mac e Linux**. Mac com processador Intel, na instalação por Python: use o pacote `hindsight-all-slim`.
- Não quer manter servidor? Existe o **Hindsight Cloud** (pago por uso, com créditos grátis pra começar).

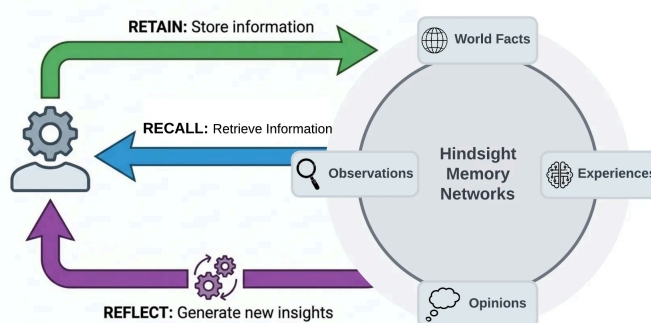


imagem: github.com/vectorize-io/hindsight

02 Instalar

DOCKER (WINDOWS, MAC OU LINUX)

No terminal, troque `sk-xxx` pela sua chave. Outro provedor?
Troque `openai` por `anthropic`, `gemini`, `groq` ou `ollama`.

```
$ docker run -it --pull always --name hindsight --restart unless
-stopped -p 8888:8888 -p 9999:9999 -e HINDSIGHT_API_LLM_PROVIDER
=openai -e HINDSIGHT_API_LLM_API_KEY=sk-xxx -v hindsight-data:/h
ome/hindsight/.pg0 ghcr.io/vectorize-io/hindsight:latest
```

Pronto: API em `localhost:8888` e painel em `localhost:9999`.

SEM DOCKER (PYTHON)

```
$ pip install hindsight-api
```

Defina `HINDSIGHT_API_LLM_PROVIDER` e
`HINDSIGHT_API_LLM_API_KEY` e rode `hindsight-api`.

03 Primeiro uso

- **No seu agente:** instale o `hindsight-client` e use `retain` (guardar) e `recall` (buscar), com um **banco** por cliente.
- **Atalho de 2 linhas:** o `hindsight-litellm` embrulha o seu cliente OpenAI/Anthropic e guarda e busca memória sozinho.
- **Agente de código** (Claude Code, Cursor, Codex):

```
$ npx @vectorize-io/hindsight-coding-agents install
claude-code
```

Também tem integração com n8n, Zapier, LangChain e outros.

! Cuidados

- **Dado pessoal (LGPD):** memória de cliente é dado pessoal. Ligue o **Memory Defense** (opcional, por banco): ele procura dados pessoais e segredos em cada memória e mascara ou bloqueia antes de salvar.
- Use **um banco de memória por cliente**: os bancos são isolados entre si.
- **Custo:** guardar e analisar memórias usa um LLM, então gasta tokens do seu provedor.
- Para automação simples, a própria documentação avisa que pode ser exagero. E é versão 0.x: muda rápido.

04 Links oficiais

- github.com/vectorize-io/hindsight
- Documentação: hindsight.vectorize.io
- Estudo: arxiv.org/abs/2512.12818

Quer colocar IA pra trabalhar na sua empresa?

Diagnóstico gratuito no link da bio de @stratiainelligence.