Letta Local Server Setup – Quick Guide

Generated on August 14, 2025 04:00

This guide shows how to run a Letta server locally for testing and learning, verify it, connect the ADE (web UI), and talk to it from Python.

Option A — Docker (recommended)

```
# 1) Make a place for persistent data (Postgres inside the container)
mkdir -p ~/.letta/.persist/pgdata

# 2) Start the Letta server on localhost:8283 with your model keys
docker run \
  -v ~/.letta/.persist/pgdata:/var/lib/postgresql/data \
  -p 8283:8283 \
  -e OPENAI_API_KEY="sk-..." \
  letta/letta:latest
```

This launches the Letta API at **http://localhost:8283/v1** and persists data under ~/.letta/.persist/pgdata.

Using other providers / local models

```
# Mac/Windows (Docker Desktop)
-e ANTHROPIC_API_KEY="sk-ant-..." \
-e OLLAMA_BASE_URL="http://host.docker.internal:11434"

# Linux (prefer host networking)
--network host \
-e OLLAMA_BASE_URL="http://localhost:11434"
```

Secure mode (optional)

```
docker run ... \
  -e SECURE=true \
  -e LETTA_SERVER_PASSWORD="yourpassword" \
  letta/letta:latest

# When secure mode is on, use a Bearer token with the password for API calls.
```

Option B — Pure Python (pip)

```
pip install -U letta
export OPENAI_API_KEY=sk-...   # or other provider keys
letta server
```

This starts the same server on **http://localhost:8283**. The letta package installs the server/CLI; the Python SDK is a separate package (letta-client).

Verify the server

```
# Add: -H "Authorization: Bearer <password>" if SECURE=true
curl http://localhost:8283/v1/health/
```

A 200 response with a small JSON body indicates the server is up.

Connect the ADE (Letta’s web UI)

Open **app.letta.com**, sign in, and choose the **Self-hosted** tab; it can connect to your local server at http://localhost:8283. If you prefer everything local, Letta Desktop bundles UI + server (beta).

Talk to your server from Python

```
# Install the official SDK
pip install letta-client

from letta_client import Letta

# If you did NOT enable secure mode:
client = Letta(base_url="http://localhost:8283")

# If you DID enable secure mode:
# client = Letta(base_url="http://localhost:8283", token="yourpassword")

# Create a simple agent
agent = client.agents.create(
    model="openai/gpt-4.1",
    embedding="openai/text-embedding-3-small",
    memory_blocks=[
        {"label": "human", "value": "The human is John and likes systems design."},
        {"label": "persona", "value": "You are a helpful test agent."}
    ],
    tools=["web_search"]  # or your own tools later
)

# Send a message
resp = client.agents.messages.create(
    agent_id=agent.id,
    messages=[{"role": "user", "content": "Hello from my local Letta!"}]
)
print(resp.messages[-1].get("content"))
```

Handy extras

- Default API base: http://localhost:8283/v1 (e.g., /v1/agents, /v1/messages).
- ADE with local server: the cloud/web ADE can connect to localhost; Letta Desktop bundles UI + server for fully local workflows.
- OpenAI-compatible proxies are supported but direct provider APIs are recommended.

Tip: Keep your provider API keys in a .env file and pass them via Docker’s --env-file to avoid exposing secrets in shell history.