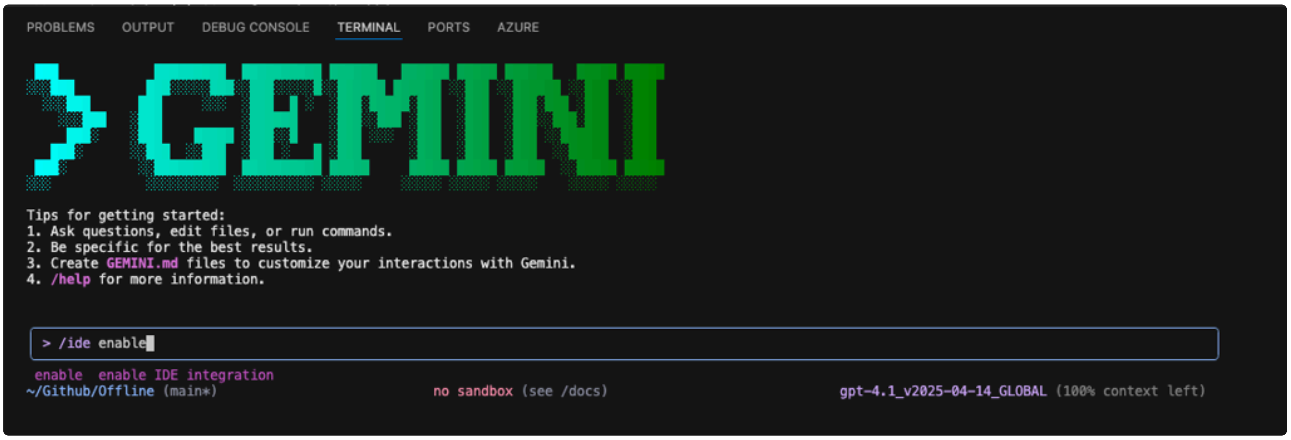
Setup Gemini CLI with Group available LLMs



```
Tips for getting started:
1. Ask questions, edit files, or run commands.
2. Be specific for the best results.
3. Create GEMINI.md files to customize your interactions with Gemini.
4. /help for more information.

> /ide enable

 enable  enable IDE integration
~/Github/Offline (main*)                    no sandbox (see /docs)                    gpt-4.1_v2025-04-14_GLOBAL (100% context left)
```

⚠ Before proceeding, note that Gemini CLI currently can only read plain-text files in a repo and answer questions *(in my setup at least…)* so not very agentic.

I’ve reported the issue here: [Bug] Gemini CLI limited to current-dir text reads despite permissive config · Issue #9529 · google-gemini/gemini-cli.

Until it’s fixed, only Claude Code and Codex seem viable as CLIs (no checkpointing and fewer models).

Setup Codex with GPT-5

Setup Claude Code with Group available LLMs

If you decide to deploy GeminiCLI anyway, please let me know if you encounter the same issues. 👍

Requirements

The below was tested with gemini-cli 5.5 , gemici-cli 6.0 and `.6.1` , LiteLLM 1.77.3

As everything in the Group, CLI connectivty is dependant on having the right proxy and SSL settings.
Sounds easy, it isn’t… To solve proxy issues, you should run Alpaca: 🦙 Setting up Alpaca on MacOS or 🦙 Setting up Alpaca on Windows 🤦 CLI - Why you should not set CNTLM nor Prisma

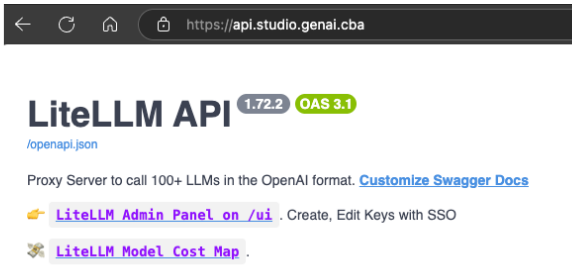
To solve SSL issues, you should build a PEM certificate store and point your CLI to it using various environment variables… 🔐 Group Internal Root CAs Certificate Store

Or if you’re `on MacOS,` you can head there and let my script do it for you: 🦙 Proxy and SSL Certificate issues

Request `AR - GenAI Studio - Prod - User` access in identity.cba

Prior to going forward, let’s state the facts! Our current Corporate LiteLLM version is not (yet - 3 Jul 2025) compatible with GeminiCLI. Expand the below for more information…

Corporate LiteLLM Compatibility issues



https://api.studio.genai.cba

LiteLLM API 1.72.2 OAS 3.1

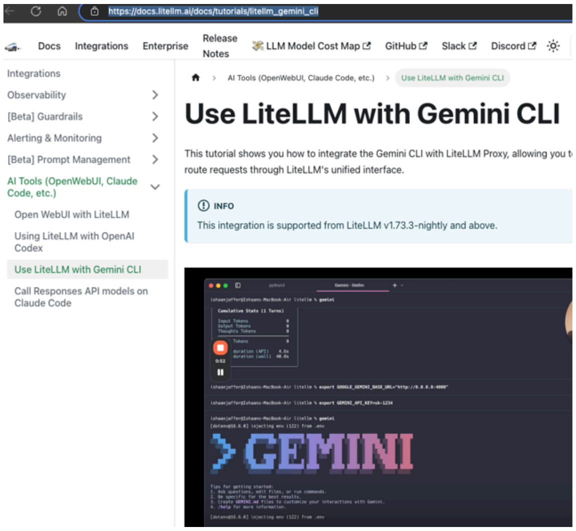
/openapi.json

Proxy Server to call 100+ LLMs in the OpenAI format. Customize Swagger Docs

👉 `LiteLLM Admin Panel on /ui` . Create, Edit Keys with SSO

💸 `LiteLLM Model Cost Map` .

We have 1.72.2



https://docs.litellm.ai/docs/tutorials/litellm_gemini_cli
Docs
Integrations
Enterprise
Release Notes
LLM Model Cost Map
GitHub
Slack
Discord
Integrations
Observability
[Beta] Guardrails
Alerting & Monitoring
[Beta] Prompt Management
AI Tools (OpenWebUI, Claude Code, etc.)
Open WebUI with LiteLLM
Using LiteLLM with OpenAI Codex
Use LiteLLM with Gemini CLI
Call Responses API models on Claude Code
AI Tools (OpenWebUI, Claude Code, etc.)
Use LiteLLM with Gemini CLI

Use LiteLLM with Gemini CLI

This tutorial shows you how to integrate the Gemini CLI with LiteLLM Proxy, allowing you t
route requests through LiteLLM's unified interface.

INFO

This integration is supported from LiteLLM v1.73.3-nightly and above.

```
ishaanjaffer@Ishaans-MacBook-Air litellm % gemini
  Cumulative Stats (1 Turns)

  Input Tokens          0
  Output Tokens         0
  Thoughts Tokens       0
  ────────────────────────
        Tokens          0

  duration (API)     4.6s
  duration (wall)   40.8s

ishaanjaffer@Ishaans-MacBook-Air litellm % export GOOGLE_GEMINI_BASE_URL="http://0.0.0.0:4000"

ishaanjaffer@Ishaans-MacBook-Air litellm % export GEMINI_API_KEY=sk-1234

ishaanjaffer@Ishaans-MacBook-Air litellm % gemini
[dotenv@16.6.0] injecting env (122) from .env
> GEMINI

Tips for getting started:
1. Ask questions, edit files, or run commands.
2. Be specific for the best results.
3. Create GEMINI.md files to customize your interactions with Gemini.
4. /help for more information.

[dotenv@16.6.0] injecting env (122) from .env
```

We need 1.73.3 min

Using proxy chaining will solve these limitations.
The article below outlines the steps to run GeminiCLI on your SOE by deploying Docker images built by our AIPE teams.

Value Proposition

GeminiCLI supports far more model families (Gemini, Mistral Anthropic, OpenAI, and AWS Bedrock) than Claude Code, making it easier to standardise and compare workflows across providers.
It also includes out-of-the-box, opt-in checkpointing (disabled by default) to persist conversation and tool state, enabling resumption, branching, and reproducible runs, and preventing lost progress in long-running tasks.

Setup Guide

Local GenAI Stack Installation

Because our LiteLLM proxy is currently below the supported version for GeminiCLI (It requires LiteLLM 1.73.3 minimum), we'll need to run our own gateway... To do so:

Follow this guide: https://playbook.genai.cba/develop/quickstart/
All credits goes to @Blair Hudson @Leopoldo Venegas Rubio @Alex Brown @Tamara Gunawan @Sherin Mary Mathew *for this excellent guide!*

Artificatory only has LiteLLM 1.73 and below which is not new enought for Gemini CLI.
So we need to replace line 45 of the `docker-compose.yaml` to run our own LiteLLM container in the latest version
from `image: analyticsinformation-genaihub.docker.internal.cba/litellm:v1.60.2-20250304-prd`
to `image: ghcr.io/berriai/litellm:main-latest`

For this, you can copy the edited `docker-compose.yaml` below... (expand the section)

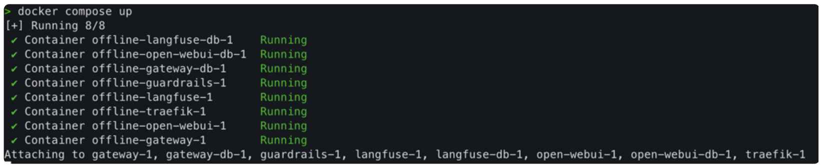
docker-compose.yaml

```yaml
# Run the GenAI Platform services locally using Docker Compose

services:
  traefik:
    image: hub.docker.internal.cba/traefik:v3.3
    command:
      - "--api.dashboard=true"
      - "--api.insecure=true"
      - "--providers.docker=true"
      - "--providers.docker.exposedbydefault=false"
      - "--entrypoints.web.address=:80"
    labels:
      - "traefik.enable=true"
      - "traefik.http.routers.traefik.rule=Host(`localhost`)"
      - "traefik.http.routers.traefik.service=api@internal"
      - "traefik.http.services.traefik.loadbalancer.server.port=8080"
    ports:
      - "80:80"
    volumes:
      - /var/run/docker.sock:/var/run/docker.sock:ro
    networks:
      - genai-platform

  guardrails:
    image: analyticsinformation-guardrails.docker.internal.cba/genai-hub-guardrails-cosumer-main:1.2.0
    networks:
      - genai-platform
    labels:
      - "traefik.enable=true"
      - "traefik.http.routers.guardrails.rule=Host(`guardrail.local1.dev.ai.dhp.cba.localhost`)"
      - "traefik.http.services.guardrails.loadbalancer.server.port=8000"
    ports:
      - "9202:8000"
    environment:
      - GAAS__ENVIRONMENT=Development
      - GAAS__AUTH_DISABLED=true
      - DISABLE_ADAPTOR=false
      - GAAS__AUTH_JWT_SIGNING_KEY=test
      - GAAS__AUTH_REQUIRED_SCOPES=test-scope
      - GAAS__AUTH_ALLOWED_ISSUERS=test-issuer1 test-issuer2
      - GPT_ENDPOINT=http://gateway:4000/openai/deployments/gpt-4o_v2024-05-13_NOFILTER_GaaS/chat/completions?api-version=2024-05-01-preview
      - GPT_API_KEY=sk-123changeme

  gateway:
    image: ghcr.io/berriai/litellm:main-latest
    networks:
      - genai-platform
    labels:
      - "traefik.enable=true"
      - "traefik.http.routers.gateway.rule=Host(`local1.dev.ai.dhp.cba.localhost`)"
      - "traefik.http.services.gateway.loadbalancer.server.port=4000"
    ports:
      - "9201:4000"
    depends_on:
      gateway-db:
```

```
    condition: service_healthy
  volumes:
   - ./proxy_server_config.yaml:/tmp/base_config.yaml:ro
  entrypoint: |
   bash -c '
   set -eou pipefail
   echo "Updating LiteLLM model configuration..."
   python3 -c "
   import os, sys, requests, json, logging, yaml

   logging.basicConfig(level=logging.INFO)
   logger = logging.getLogger(__name__)

   CONFIG_PATH = \"/tmp/base_config.yaml\"
   WORKING_CONFIG = \"/tmp/proxy_server_config.yaml\"

   try:
       # Copy base config to working location
       with open(CONFIG_PATH, \"r\") as f:
           config = yaml.safe_load(f)

       # Add dynamic models from API
       headers = {
           \"accept\": \"application/json\",
           \"Authorization\": \"Bearer \" + os.environ.get(\"OPENAI_API_KEY\", \"\")
       }
       api_base = os.environ.get(\"OPENAI_API_BASE\", \"\")

       # Try to fetch models from API
       try:
           logger.info(f\"Fetching models from {api_base}/models\")
           response = requests.get(f\"{api_base}/models\", headers=headers, verify=False)
           response.raise_for_status()
           models = response.json()[\"data\"]

           # Add API models to existing model list
           for model in models:
               model_id = model[\"id\"]
               config[\"model_list\"].append({
                   \"model_name\": model_id,
                   \"litellm_params\": {
                       \"model\": f\"openai/{model_id}\",
                       \"api_key\": \"os.environ/OPENAI_API_KEY\",
                       \"api_base\": \"os.environ/OPENAI_API_BASE\"
                   }
               })
           logger.info(f\"Added {len(models)} API models\")
       except Exception as e:
           logger.warning(f\"Failed to fetch API models: {str(e)}\")

       # Try to add Ollama models
       try:
           ollama_response = requests.get(\"http://host.docker.internal:11434/v1/models\")
           ollama_response.raise_for_status()
           ollama_models = ollama_response.json()[\"data\"]
           for model in ollama_models:
                  model_name = model[\"id\"]
                  config[\"model_list\"].append({
                      \"model_name\": model_name,
                      \"litellm_params\": {
                          \"model\": \"ollama_chat/\" + model_name,
                          \"api_base\": \"http://host.docker.internal:11434\"
                      }
                  })
              logger.info(f\"Added {len(ollama_models)} Ollama models\")
          except Exception as e:
              logger.warning(f\"Failed to fetch Ollama models: {str(e)}\")

          with open(WORKING_CONFIG, \"w\") as f:
              yaml.dump(config, f, default_flow_style=False)

          logger.info(\"Configuration updated successfully\")

      except Exception as e:
          logger.error(f\"Failed to update configuration: {str(e)}\")
          sys.exit(1)
      "

      echo "Starting LiteLLM server..."
      exec litellm --config /tmp/proxy_server_config.yaml --port 4000'
    environment:
      - OPENAI_API_KEY=${OPENAI_API_KEY}
      - OPENAI_API_BASE=${OPENAI_BASE_URL:-https://api.studio.genai.cba}
      - DATABASE_URL=postgres://postgres:postgres@gateway-db:5432/genai_gateway
      - LITELLM_MASTER_KEY=sk-123changeme
      - STORE_MODEL_IN_DB=False
      - NUM_WORKERS=1
      - LITELLM_PORT=4000
      - LANGFUSE_PUBLIC_KEY=pk-lf-d8825033-3edb-4767-9786-2fce2495bcf5
      - LANGFUSE_SECRET_KEY=sk-lf-000bf592-eec9-4fc7-938e-59c60d61e7b1
      - LANGFUSE_HOST=http://langfuse:3000
    healthcheck:
      disable: true

  gateway-db:
    image: hub.docker.internal.cba/postgres:16-alpine
    networks:
      - genai-platform
    volumes:
      - gateway-db-data:/var/lib/postgresql/data
    healthcheck:
      test: ["CMD-SHELL", "pg_isready -U postgres"]
      interval: 5s
      timeout: 5s
      retries: 5
    environment:
      - POSTGRES_DB=genai_gateway
      - POSTGRES_USER=postgres
      - POSTGRES_PASSWORD=postgres

  langfuse:
    image: analyticsinformation-genaihub.docker.internal.cba/langfuse:dhp-v2.85
    networks:
      - genai-platform
    labels:
      - "traefik.enable=true"
      - "traefik.http.routers.langfuse.rule=Host(`langfuse.local1.dev.ai.dhp.cba.localhost`)"
      - "traefik.http.services.langfuse.loadbalancer.server.port=3000"
    ports:
      - "9203:3000"
    depends_on:
      langfuse-db:
        condition: service_healthy
    healthcheck:
      disable: true
    environment:
      - DATABASE_URL=postgres://postgres:postgres@langfuse-db:5432/langfuse
      - SALT=mysalt
      - TELEMETRY_ENABLED=false
      - NEXTAUTH_URL=http://langfuse.local1.dev.ai.dhp.cba.localhost/api/auth
      - NEXTAUTH_SECRET=mysecret
      - LANGFUSE_ENABLE_EXPERIMENTAL_FEATURES=true
      - LANGFUSE_INIT_ORG_ID=local
      - LANGFUSE_INIT_ORG_NAME=local
      - LANGFUSE_INIT_PROJECT_ID=local
      - LANGFUSE_INIT_PROJECT_NAME=local
      - LANGFUSE_INIT_USER_EMAIL=local@genai.cba
      - LANGFUSE_INIT_USER_PASSWORD=123changeme
      - LANGFUSE_INIT_PROJECT_PUBLIC_KEY=pk-lf-d8825033-3edb-4767-9786-2fce2495bcf5
      - LANGFUSE_INIT_PROJECT_SECRET_KEY=sk-lf-000bf592-eec9-4fc7-938e-59c60d61e7b1

  langfuse-db:
    image: hub.docker.internal.cba/postgres:16-alpine
    networks:
      - genai-platform
    volumes:
      - langfuse-db-data:/var/lib/postgresql/data
    healthcheck:
      test: ["CMD-SHELL", "pg_isready -U postgres"]
      interval: 5s
      timeout: 5s
      retries: 5
    environment:
      - POSTGRES_DB=langfuse
      - POSTGRES_USER=postgres
      - POSTGRES_PASSWORD=postgres

  open-webui:
    image: analyticsinformation-genaihub.docker.internal.cba/open-webui:v1.1.0
    networks:
      - genai-platform
    labels:
      - "traefik.enable=true"
      - "traefik.http.routers.webui.rule=Host(`studio.local1.dev.ai.dhp.cba.localhost`)"
      - "traefik.http.services.webui.loadbalancer.server.port=8080"
    ports:
      - "9204:8080"
    depends_on:
      open-webui-db:
        condition: service_healthy
    healthcheck:
      disable: true
    environment:
      - OPENAI_API_BASE_URL=http://gateway:4000
      - OPENAI_API_KEY=sk-123changeme
      - WEBUI_AUTH=False
      - DATABASE_URL=postgres://postgres:postgres@open-webui-db:5432/open_webui
      - ENABLE_OLLAMA_API=False
      - WEBUI_SECRET_KEY=123changeme
      - OFFLINE_MODE=True
      - ENV=dev

  open-webui-db:
    image: hub.docker.internal.cba/postgres:16-alpine
    networks:
      - genai-platform
    volumes:
      - open-webui-db-data:/var/lib/postgresql/data
    healthcheck:
      test: ["CMD-SHELL", "pg_isready -U postgres"]
      interval: 5s
      timeout: 5s
      retries: 5
    environment:
      - POSTGRES_DB=open_webui
      - POSTGRES_USER=postgres
      - POSTGRES_PASSWORD=postgres

volumes:
  gateway-db-data:
  langfuse-db-data:
  open-webui-db-data:

networks:
  genai-platform:
    driver: bridge
```

⚠ *Optional but important → Remember to change the* `sk-123changeme` *password…*

Now, after saving this file somewhere, run `docker compose up` to start building the components!

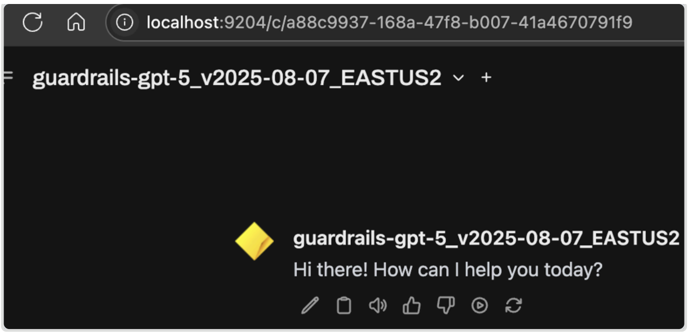


```
> docker compose up
[+] Running 8/8
 ✔ Container offline-langfuse-db-1    Running
 ✔ Container offline-open-webui-db-1  Running
 ✔ Container offline-gateway-db-1     Running
 ✔ Container offline-guardrails-1     Running
 ✔ Container offline-langfuse-1       Running
 ✔ Container offline-traefik-1        Running
 ✔ Container offline-open-webui-1     Running
 ✔ Container offline-gateway-1        Running
Attaching to gateway-1, gateway-db-1, guardrails-1, langfuse-1, langfuse-db-1, open-webui-1, open-webui-db-1, traefik-1
```

If all goes well, it'll look like this…

Local GenAI Stack Testing

Before proceeding, ensure you can access the various components.



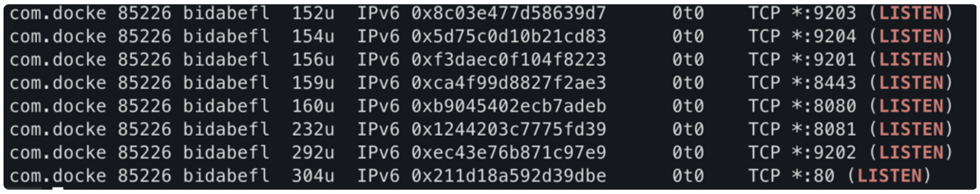
localhost:9204/c/a88c9937-168a-47f8-b007-41a4670791f9
guardrails-gpt-5_v2025-08-07_EASTUS2

guardrails-gpt-5_v2025-08-07_EASTUS2

Hi there! How can I help you today?

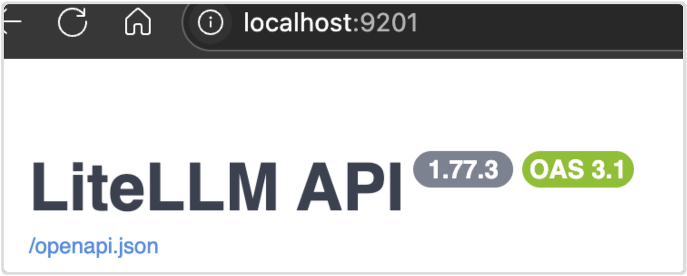
Check if you can query models from your Local OpenWeb-UI, this will be a conclusive test that LiteLLM and Traefik are working as expected

Some more testing...



```
com.docke 85226 bidabefl  152u  IPv6 0x8c03e477d58639d7      0t0     TCP *:9203 (LISTEN)
com.docke 85226 bidabefl  154u  IPv6 0x5d75c0d10b21cd83      0t0     TCP *:9204 (LISTEN)
com.docke 85226 bidabefl  156u  IPv6 0xf3daec0f104f8223      0t0     TCP *:9201 (LISTEN)
com.docke 85226 bidabefl  159u  IPv6 0xca4f99d8827f2ae3      0t0     TCP *:8443 (LISTEN)
com.docke 85226 bidabefl  160u  IPv6 0xb9045402ecb7adeb      0t0     TCP *:8080 (LISTEN)
com.docke 85226 bidabefl  232u  IPv6 0x1244203c7775fd39      0t0     TCP *:8081 (LISTEN)
com.docke 85226 bidabefl  292u  IPv6 0xec43e76b871c97e9      0t0     TCP *:9202 (LISTEN)
com.docke 85226 bidabefl  304u  IPv6 0x211d18a592d39dbe      0t0     TCP *:80 (LISTEN)
```

Docker should be listening on a bunch of ports



LiteLLM API 1.77.3 OAS 3.1

/openapi.json

Note how were now have LiteLLM version that is compatible with GeminiCLI

Now that we are confident local GenAI is running smoothly, it's time to set up Gemini CLI to use it.

GeminiCLI Installation

Follow the official documentation for the most up-to-date installation instructions: GitHub - google-gemini/gemini-cli: An open-source AI agent that brings the power of Gemini directly into your terminal.

Current installation command: `npm install -g @google/gemini-cli`

GeminiCLI Configuration

Follow the official documentation for the most up-to-date authentication instructions: Gemini CLI | liteLLM

All you need is to declare some environment variables.

on MacOS

1. Add the following line to your shell configuration file (`~/.profile` , `~/.zshrc` , `~/.bashrc` , etc.):
   `export GEMINI_API_KEY=sk-123changeme` */!\ Update this with your password as per the Docker compose*
   `export GOOGLE_GEMINI_BASE_URL="http://localhost:9201"`
2. Reload `source ~/.zshrc` (or equivalent for your shell)

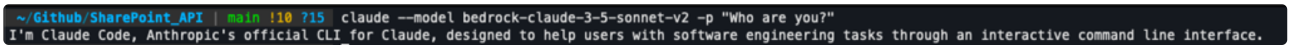
> For a safer way to locally store credentials, consider using Keychain Access to store the secret as per Setup Claude Code with Group available LLMs | Authentication
>
> *Alternatively, you can use 3rd-party password manager like MacPass*

on Windows

Same as above but similarly to Group Internal Root CAs Certificate Store | How to Set Up on Windows

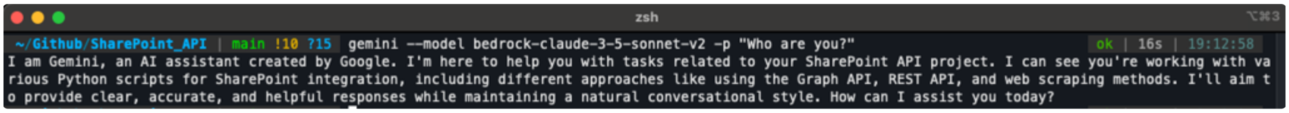
GeminiCLI Testing

Run Gemini with your desired model...



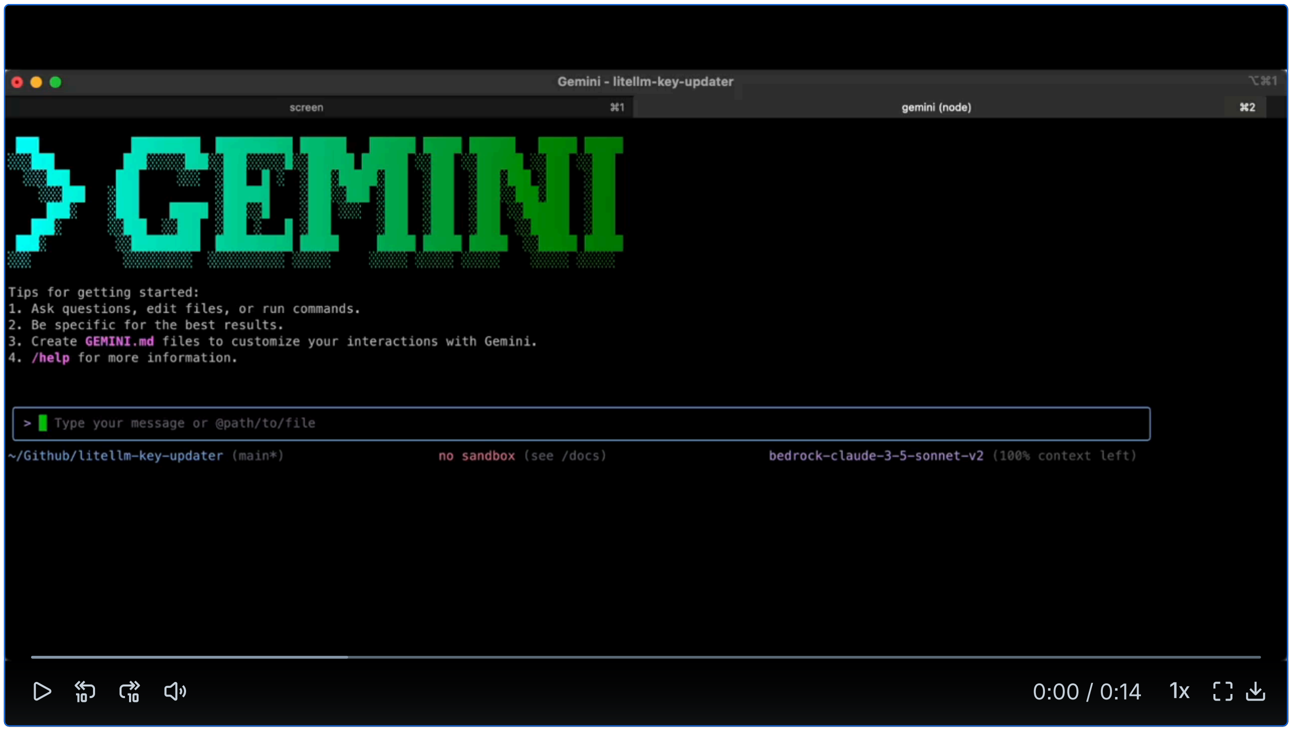
```
~/Github/SharePoint_API | main !10 ?15  claude --model bedrock-claude-3-5-sonnet-v2 -p "Who are you?"
I'm Claude Code, Anthropic's official CLI for Claude, designed to help users with software engineering tasks through an interactive command line interface.
```

Since Claude Code works with `bedrock-claude-3-5-sonnet-v2` using `localhost:9201` ...



```
zsh
~/Github/SharePoint_API | main !10 ?15  gemini --model bedrock-claude-3-5-sonnet-v2 -p "Who are you?"          ok | 16s | 19:12:58
I am Gemini, an AI assistant created by Google. I'm here to help you with tasks related to your SharePoint API project. I can see you're working with va
rious Python scripts for SharePoint integration, including different approaches like using the Graph API, REST API, and web scraping methods. I'll aim t
o provide clear, accurate, and helpful responses while maintaining a natural conversational style. How can I assist you today?
```

GeminiCLI should as well, and... that’s looking good 🙂



```
> GEMINI

Tips for getting started:
1. Ask questions, edit files, or run commands.
2. Be specific for the best results.
3. Create GEMINI.md files to customize your interactions with Gemini.
4. /help for more information.

> Type your message or @path/to/file
~/Github/litellm-key-updater (main*)          no sandbox (see /docs)          bedrock-claude-3-5-sonnet-v2 (100% context left)
```

Success!

GeminiCLI Supported Models

Tested with `gemini --model <model> -p "Hello"` where model comes from `http://localhost:9201/v1/models`

Total Models Tested: 55

Working Models: 24 (43.6% success rate)

Failed Models: 31 (56.4% failure rate)

Add your prefered model to your Shell config G Setup Gemini CLI with Group available LLMs | GeminiCLI Configuration

For instance: `export GEMINI_MODEL="gpt-4.1_v2025-04-14_GLOBAL"`

⌄ Models thar are supported...

Working Model Ordered by Most to Least powerful	Response
bedrock-claude-4-sonnet	Hello! I can see you're working in your `/Users/bidabefl/Github/Offline` directory on macOS. It looks like you have an interesting project set up here with: - Docker

	configuration (`docker-compose.yml`) - Gemini model testing scripts and results - A Qwen 2.5 Coder model file (7B parameters, quantized) - Python virtual environment - Some configuration files for proxy server and Claude settings This appears to be some kind of offline AI/LLM setup or testing environment. Is there anything specific you'd like to work on or discuss about this project?
guardrails-bedrock-claude-4-sonnet	Hello! I can see you're working in your `/Users/bidabefl/Github/Offline` directory on macOS. It looks like you have an interesting project set up here with: - Some Gemini model testing scripts and results - A Qwen 2.5 Coder model file (GGUF format) - Docker configuration - A Python virtual environment - Cache directories for Hugging Face downloads Is there something specific you'd like to work on or discuss regarding this offline AI/ML setup? I'm here to help!
gpt-4.1_v2025-04-14_GLOBAL	Hello! How can I assist you today?
guardrails-gpt-4.1_v2025-04-14_GLOBAL	Hello! How can I assist you today?
bedrock-claude-3-7-sonnet	Hello! I'm Gemini, your AI assistant. I see you're working in your "Offline" directory which appears to contain some AI model testing files and configurations. Is there something specific you'd like help with today? I notice you have test scripts for Gemini models, a Docker configuration, and what looks like a Qwen 2.5 model file. I'd be happy to assist with any questions about these files or help with other tasks.
guardrails-bedrock-claude-3-7-sonnet	Hello! I'm Gemini, your AI assistant. I see you're working in your "Offline" directory on macOS. It looks like you have some files related to AI models and testing, including Gemini model test results and a Qwen 2.5 model file. How can I help you today? I'd be happy to assist with questions about your files, help with coding tasks, or discuss anything else you're interested in.
bedrock-claude-3-sonnet	Hello! I'm ready to assist you with any tasks or questions you may have related to the current working directory and project setup. Please feel free to ask, and I'll do my best to help.

guardrails-bedrock-claude-3-sonnet	Hello! I'm ready to assist you with any tasks or questions you may have related to the current working directory and project setup. Please feel free to ask, and I'll do my best to help.
gpt-4o_v2024-11-20	Hello! How can I assist you today? 😊
gpt-4o_v2024-11-20_USEAST	Hello! How can I assist you today? 😊
guardrails-gpt-4o_v2024-11-20	Hello! How can I assist you today? 😊
guardrails-gpt-4o_v2024-11-20_USEAST	Hello! How can I assist you today? 😊
gpt-4o_v2024-05-13_USEAST	Hello! How can I assist you today?
guardrails-gpt-4o_v2024-05-13_USEAST	Hello! How can I assist you today?
o3-mini_v2025-01-31_EASTUS2	Hello! 👋 How can I help you today?
guardrails-o3-mini_v2025-01-31_EASTUS2	Hello there! How can I help you today?
gpt-4o-mini_v2024-07-18	Hello! How can I assist you today?
guardrails-gpt-4o-mini_v2024-07-18	Hello! How can I assist you today?
gemini-2.0-flash-001	Hello! How can I help you today?
guardrails-gemini-2.0-flash-001	Hello! How can I help you today?

aipe-gemini-2.0-flash-001	Hello! How can I help you today?
bedrock-mistral-large-2402-v1	Hello! How can I assist you today? Just a reminder, here's the current context: * Today's date is Tuesday, September 23, 2025 ∗ My operating system is: darwin * I'm currently working in the directory: /Users/bidabefl/Github/Offline * Folder structure of the current working directories is as shown earlier. Please let me know if there's anything specific you'd like help with.
guardrails-bedrock-mistral-large-2402-v1	Hello! How can I assist you today? Just a reminder, here's the current context: * Today's date is Tuesday, September 23, 2025 ∗ My operating system is: darwin * I'm currently working in the directory: /Users/bidabefl/Github/Offline * Folder structure of the current working directories is as shown earlier. Please let me know if there's anything specific you'd like help with.
bedrock-claude-3-haiku	Hello! I'm ready to assist you. How can I help you today?
guardrails-bedrock-claude-3-haiku	Hello! I'm ready to assist you. How can I help you today?
guardrails-bedrock-mistral-7b-instruct-v0	Hello! How can I help you today? If you have any questions or tasks related to the Gemini CLI or the directories and files you've listed, feel free to ask! Here's a brief overview of the current working directory: - The directory is located at /Users/bidabefl/Github/Offline - There are several files and folders, including a .dockercompose.yml file, a .claude settings file, and a .venv virtual environment. - There are also several scripts for testing Gemini models and a PDF file for a Quickstart guide. - The .cache and .claude folders contain configuration files and settings for various tools and applications. Let me know if you have any specific questions or tasks!

Models that are not supported

Model	Error Message
gpt-5_v2025-08-07_EASTUS2	Unsupported parameters error

aipe-gpt-5_v2025-08-07_EASTUS2	Unsupported parameters error
guardrails-gpt-5_v2025-08-07_EASTUS2	Unsupported parameters error
aipe-bedrock-claude-4-sonnet	AIPE-prefixed models are forbidden
bedrock-claude-3-5-sonnet-v2	Model returned response but was marked as failed (exit code 0)
guardrails-bedrock-claude-3-5-sonnet-v2	Model returned response but was marked as failed (exit code 0)
aipe-bedrock-claude-3-7-sonnet	AIPE-prefixed models are forbidden
aipe-gpt-4.1_v2025-04-14	AIPE-prefixed models are forbidden
gpt-4o_v2024-05-13	Model not found error
gpt-4o_v2024-05-13_NOFILTER_GaaS	Model not found error
guardrails-gpt-4o_v2024-05-13	Model returned response but was marked as failed (exit code 0)
aipe-gpt-4o_v2024-11-20	AIPE-prefixed models are forbidden
o3-mini_v2025-01-31_EASTUS2	Model returned response but was marked as failed (exit code 0)
guardrails-o3-mini_v2025-01-31_EASTUS2	Model returned response but was marked as failed (exit code 0)
bedrock-mistral-large-2402-v1	AWS Bedrock authentication error - signature mismatch
bedrock-mistral-small-2402-v1	AWS Bedrock authentication error - signature mismatch
guardrails-bedrock-mistral-small-2402-v1	Bad request error
bedrock-mistral-7b-instruct-v0	AWS Bedrock authentication error - signature mismatch
bedrock-amazon-titan-text-express-v1	AWS Bedrock authentication error - signature mismatch

guardrails-bedrock-amazon-titan-text-express-v1	Model returned response but was marked as failed (exit code 0)
bedrock-amazon-titan-text-lite-v1	AWS Bedrock authentication error - signature mismatch
guardrails-bedrock-amazon-titan-text-lite-v1	Model returned response but was marked as failed (exit code 0)
text-embedding-3-large_v1	Model not compatible with chat completion (embedding model)
text-embedding-3-small_v1	Model not compatible with chat completion (embedding model)
text-embedding-ada-002_v2	Model not compatible with chat completion (embedding model)
bedrock-cohere-embed-eng-v3	AWS Bedrock authentication error - signature mismatch
bedrock-cohere-embed-mul-v3	AWS Bedrock authentication error - signature mismatch
bedrock-titan-embed-text-v2	Gemini API communication error
bedrock-amazon-titan-embed-text-v2	AWS Bedrock authentication error - signature mismatch
bedrock-amazon-titan-embed-image-v1	AWS Bedrock authentication error - signature mismatch
GenAI Assistant	Restricted to GenAI Playbook information only

GeminiCLI configuration

To each their preference, but here's mine. I built it manually based on: gemini-cli/docs/cli/configuration.md at main · google-gemini/gemini-cli

Expand it to see it.

settings.json (click to expand)

```
{
```

```json
    "general":
    {
        "checkpointing": { "enabled": true },
        "disableAutoUpdate": true,
        "vimMode": true,
        "preferredEditor": "sublime"
    },

    "ui":
    {
      "theme": "ANSI",
      "showMemoryUsage": true,
      "disableLoadingPhrases": true
    },

    "ide":
    {
        "hasSeenNudge": true,
        "enabled": false
    },

    "privacy":{ "usageStatisticsEnabled": false },

    "tools":
    {
      "sandbox": false,
      "core":
      [
        "list_directory",
        "read_many_files",
        "read_file",
        "write_file",
        "glob",
        "replace",
        "search_file_content",
        "run_shell_command",
        "web_fetch",
        "web_search",
        "ShellTool"
      ],
      "exclude":
      [
        "ShellTool(rm -rf)",
        "ShellTool(dd)",
        "ShellTool(mkfs)",
        "ShellTool(fdisk)",
        "ShellTool(gpt)",
        "ShellTool(halt)",
        "ShellTool(reboot)",
        "ShellTool(shutdown)"
      ]
    },

    "security":
    {
        "auth": { "selectedType": "gemini-api-key" },
        "folderTrust": { "enabled": true }
    }
```

```
}
```

Save the below in `$HOME/.gemini/settings.json` for MacOS or `%userprofile%\.gemini\settings.json` for Windows

> ⚠ Please consider reviewing the below settings… you are responsible for your setup whether over-permissive or not and its consequences…
>
> Sandbox can be enabled and YOLO mode should be prevented where possible.

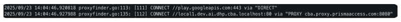
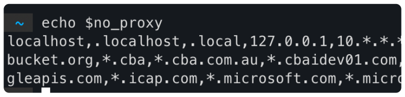
Gemini CLI Companion

Head there: Gemini CLI Companion (VSCode Extension)

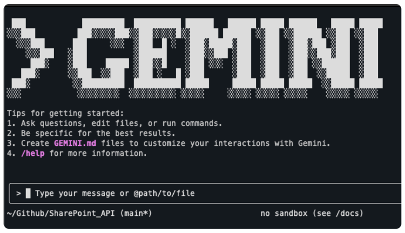
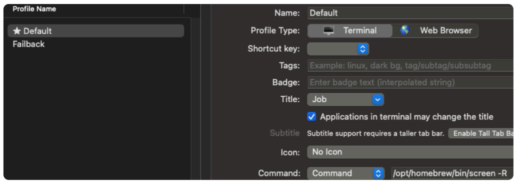
Troubleshooting

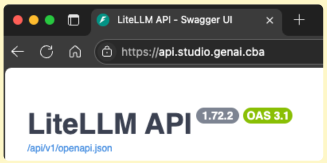
Expand to see the various possible errors and resolution

Error Message	Explanation	Resolution

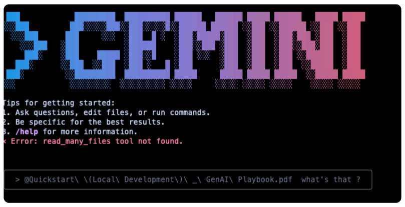
`× [API Error: exception TypeError: fetch failed sending request]`	You've set `GOOGLE_GEMINI_BASE_URL="http://local1.dev.ai.dhp.cba.localhost"` except that Gemini appear to ignore `no_proxy` var and Alpaca route `.localhost` traffic via Prisma while this should be local/direct…
	
	`2025/09/23 14:04:46.920818 proxyfinder.go:113: [111] CONNECT //play.googleapis.com:443 via "DIRECT"`
	`2025/09/23 14:04:46.927988 proxyfinder.go:135: [112] CONNECT //local1.dev.ai.dhp.cba.localhost:80 via "PROXY cba.proxy.prismaaccess.com:8080"`
	
	`~ echo $no_proxy`
	`localhost,.localhost,.local,127.0.0.1,10.*.*.*`
	`bucket.org,*.cba,*.cba.com.au,*.cbaidev01.com,`
	`gleapis.com,*.icap.com,*.microsoft.com,*.micro`
	Gemini seems to simply ignore this env var…
	Just like Claude code used to…
		Since we cannot enforce the .localhost traffic to go direct, do not use `dnsmasq` and instead `GOOGLE_GEMINI_BASE_URL="http://localhost:9201"`

Gemini UI looks borring (black and white)…	I noticed this is the case when using GNU Screen	Use a different Shell profile without `screen`

	
	`Tips for getting started:`
	`1. Ask questions, edit files, or run commands.`
	`2. Be specific for the best results.`
	`3. Create GEMINI.md files to customize your interactions with Gemini.`
	`4. /help for more information.`
	`> Type your message or @path/to/file`
	`~/Github/SharePoint_API (main*)     no sandbox (see /docs)`
		
		Profile Name
		★ Default
		Failback
		Name: Default
		Profile Type: Terminal / Web Browser
		Shortcut key:
		Tags: Example: linux, dark bg, tag/subtag/subsubtag
		Badge: Enter badge text (interpolated string)
		Title: Job
		☑ Applications in terminal may change the title
		Subtitle Subtitle support requires a taller tab bar. Enable Tall Tab Ba
		Icon: No Icon
		Command: Command /opt/homebrew/bin/screen -R

`[API Error: {"detail":"Not Found"}]`
`An unexpected critical error occurred:`
`ApiError: {"detail":"Not Found"}`
`    at throwErrorIfNotOK (file:///opt/homebrew/lib/node_modules/@google/gemini-cli/node_modules/@google/genai/dist/node/index.mjs:14072:30)`
`    at process.processTicksAndRejections (node:internal/process/task_queues:105:5)`
`    at async file:///opt/homebrew/lib/node_modules/@google/gemini-cli/node_modules/@google/genai/dist/node/index.mjs:13848:13`
`    at async GeminiClient.tryCompressChat (file:///opt/homebrew/lib/node_modules/@google/gemini-cli/node_modules/@google/gemini-cli-core/dist/src/core/client.js:560:53)`
`    at async GeminiClient.sendMessageStream (file:///opt/homebrew/lib/node_modules/@google/gemini-cli/node_modules/@google/gemini-cli-`
	That’s what happens if you’re LiteLLM is below 1.73.3 Nightly… It does not support GeminiCLI `GOOGLE_GEMINI_BASE_URL` might be set to `"https://api.studio.genai.cba"`
	⚠ At the time of writing (`23 Sept 2025`), our LiteLLM version is at `1.72.2` ([test](#) is `1.73`)
	
	LiteLLM API - Swagger UI
	https://api.studio.genai.cba
	LiteLLM API 1.72.2 OAS 3.1
	/api/v1/openapi.json
	Once updated, you will no longer need to run your own containerise LiteLLM at `http://localhost:9201`.
		You need `export GOOGLE_GEMINI_BASE_URL="http://localhost:9201"`

`core/dist/src/core/client.js:344:28)`
`    at async file:///opt/homebrew/lib/node_modules/@google/gemini-cli/dist/src/nonInteractiveCli.js:51:34`
`    at async main (file:///opt/homebrew/lib/node_modules/@google/gemini-cli/dist/src/gemini.js:323:5)`		

`× Error: read_many_files tool not found.`	Gemini refuses to read/write files and execute commands…
	This appear to be related to the `settings.json` > After spending far too long troubleshooting npm, node and gemini for permission or settings issues, this appear to be related to… `tools` and the sub entries for `core`, `exclude` and `allowed`
	
	`Tips for getting started:`
	`1. Ask questions, edit files, or run commands.`
	`2. Be specific for the best results.`
	`3. /help for more information.`
	`× Error: read_many_files tool not found.`
	`> @Quickstart\ \(Local\ Development\)\ _\ GenAI\ Playbook.pdf  what's that ?`
		Amend this section and ensure it is aligned to what Gemini CLI expects: As painful as it is… you will need to spend time understanding what is wrong based on: https://github.com/google-gemini/gemini-cli/blob/main/docs/cli/configuration.md https://github.com/google-gemini/gemini-cli/blob/main/docs/tools/index.md and https://github.com/google-gemini/gemini-cli/blob/main/docs/tools/file-system.md
		⚠ Also lots of online resources points to the deprecated format: https://github.com/google-gemini/gemini-cli/blob/main/docs/cli/configuration-v1.md
		Skip this syntax and values to prevent potential anomalies… e.g. avoid having a LLM creating your

		`settings.json`, it’s guaranteed to be bogus! 😞 as per my settings.json file above, you need to declare `tools.core.read_many_files` or `tools.allowed.read_many_files` so Gemini is allowed to use it

Helpful? Drop me a thanks on Achievers ! And if you've got knowledge to share, don't hold back - we all grow when we learn from each other 💡