



DIAL: Local setup

JUNE 2025

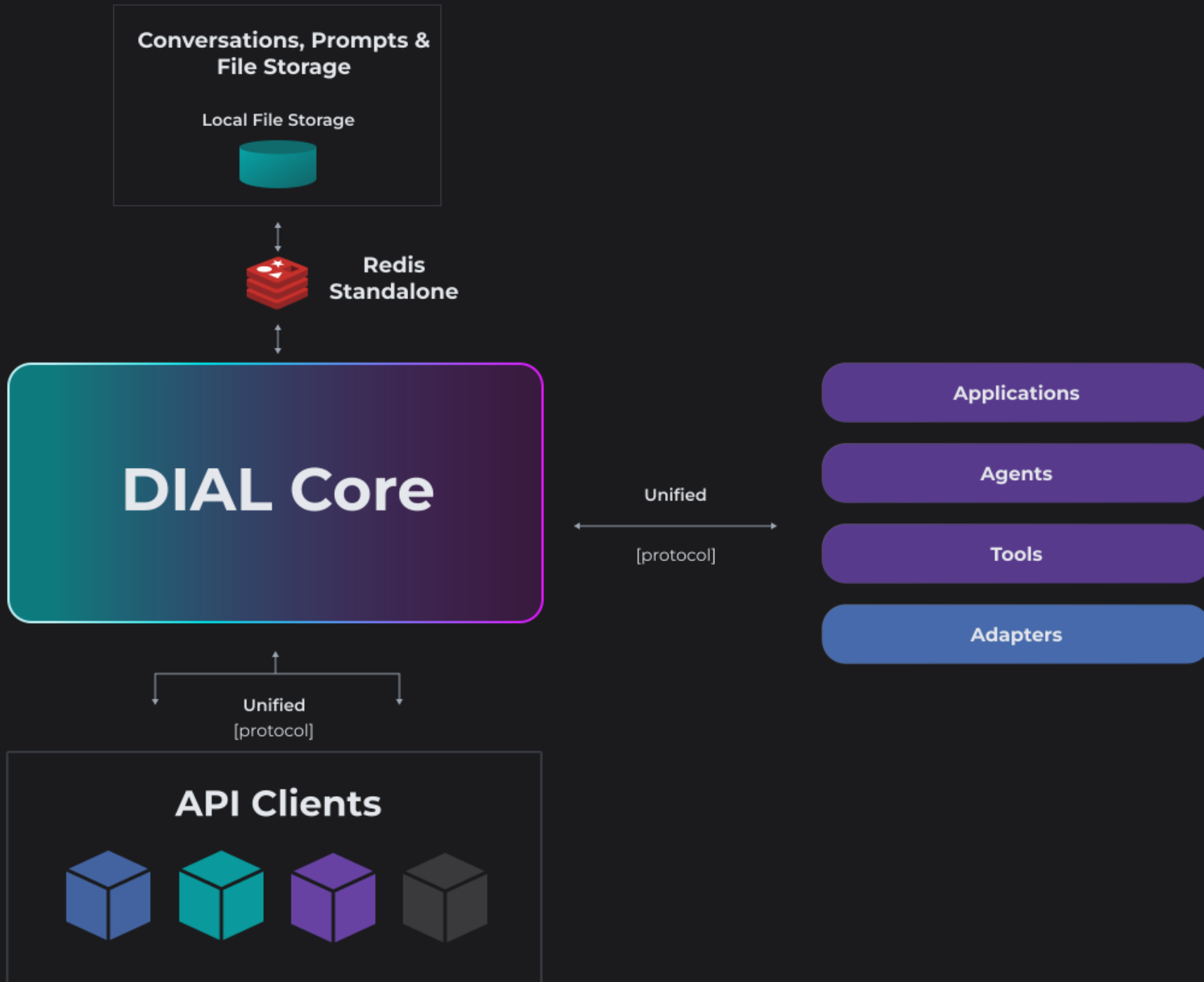
Before we start:

- **Raise your hand and ask questions if you have any**
- **It is better to ask questions when you have**
- **Also, type them in chat**
- **We will need DIAL API key for this session**

Agenda:

- **Presentation:**
 - **About DIAL Infrastructure**
 - **DIAL Core configuration**
- **Workshop:**
 - **Setup locally DIAL Chat with Core, Redis (to store conversations and files) and Themes for Chat**
 - **Configure and add Echo application to DIAL Core**
 - **Configure and add openai, anthropic and gemini models to DIAL Core**
 - **Setup dial-adapter-dial**
 - **Write and configure Essay Assistant**

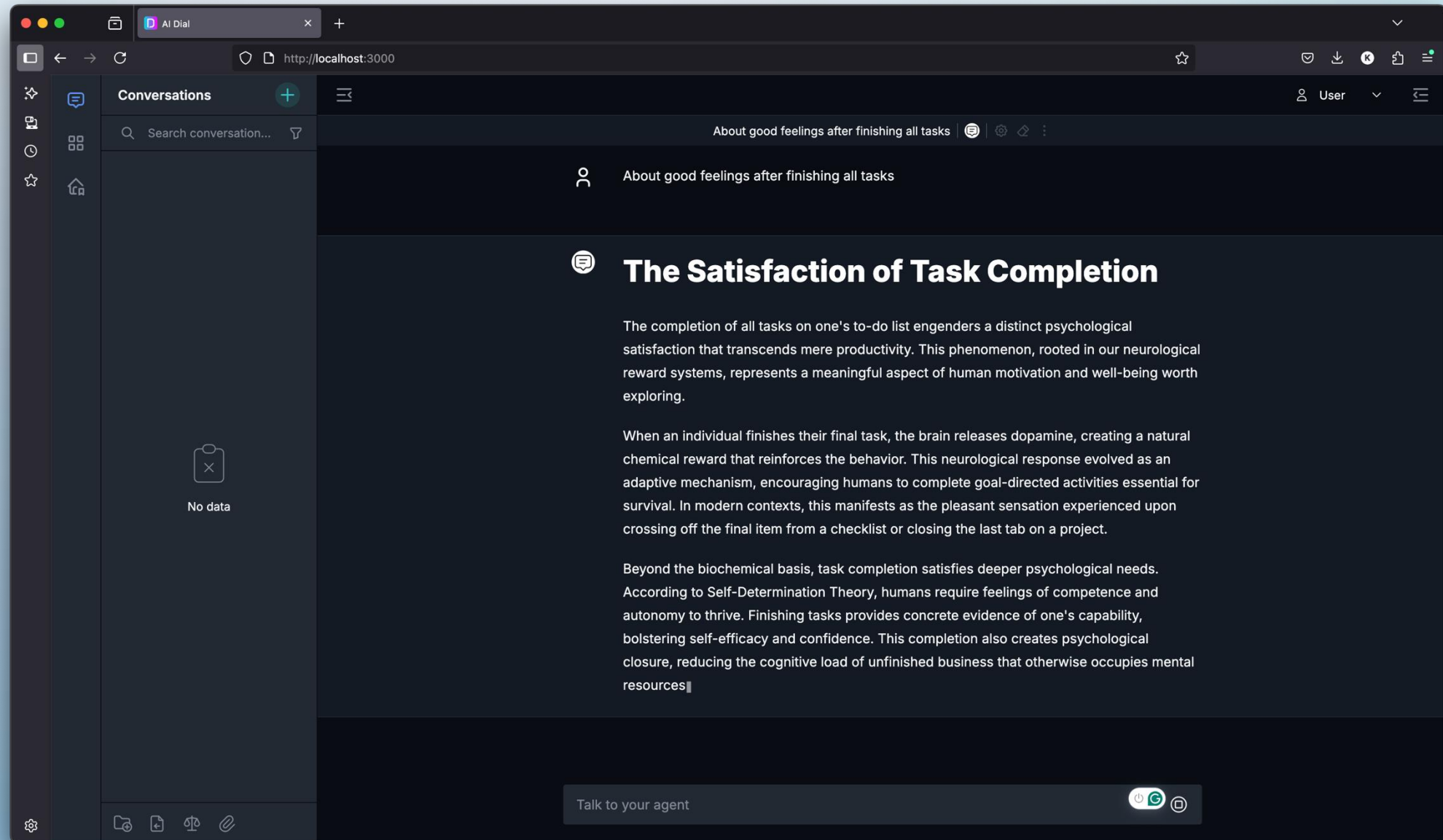
About Infrastructure



DIAL Core is the main integration hub that uses a **Unified Protocol** (OpenAI compatible) to enable governed, unified access to all features for internal/external clients, LLM models, and applications.

DIAL provides a single **unified** OpenAI-compatible API for accessing all language models, embeddings, and applications, creating a unification layer that makes models and applications interchangeable for cohesive conversational experiences and future-proof GenAI development.
https://dialx.ai/dial_api

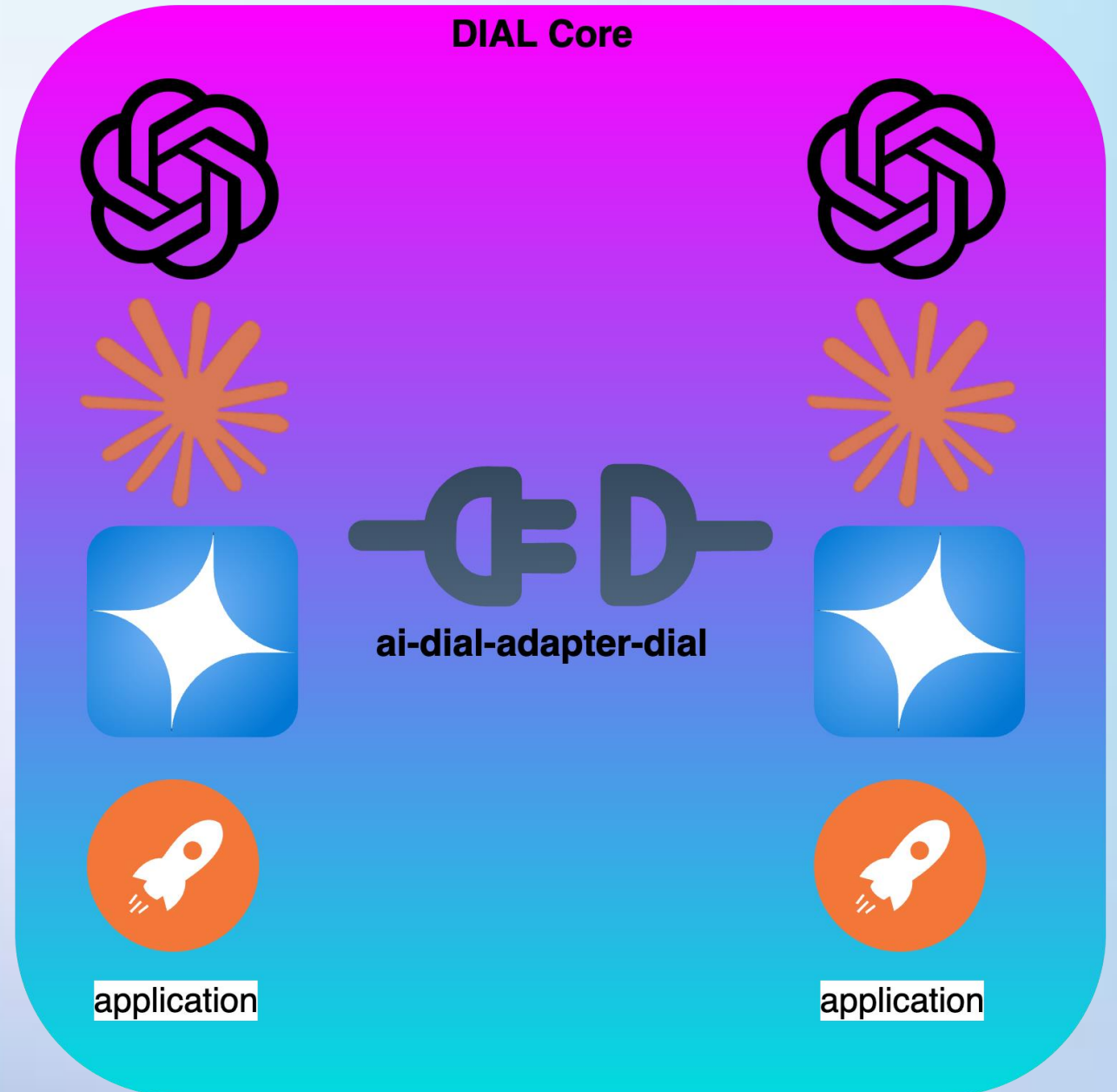
DIAL Chat is a customizable enterprise chat application with access control and extensible GenAI functionality. Overlay enables seamless embedding into existing web applications.



DIAL Adapter:

The DIAL adapter simplifies DIAL application development and local deployment by enabling communication between two DIAL Core instances, eliminating manual LLM model setup and providing access to any remote DIAL Core model within your API key scope.

With [ai-dial-adapter-dial](#) applications and models within DIAL Core can connect other applications and models within DIAL Core



DIAL Adapter:

ai-dial-adapter-openai allow applications and models from DIAL Core to connect openai models that are not within DIAL Core

The same story with vertex and bedrock adapters for Gemini and Anthropic

DIAL Core

ai-dial-adapter-openai



ai-dial-adapter-bedrock



ai-dial-adapter-vertexai



DIAL Core Configuration

Core config:

To add new custom application, you need to add configuration to “applications”.

“endpoint” is URL within DIAL Core for application

Pay attention that key (echo) should be the same as in endpoint, otherwise you will get 404 on access this app or model

```
{
  "routes": {},
  "applications": {
    "echo": {
      "displayName": "My Echo App",
      "description": "Simple application that repeats user's message",
      "endpoint": "http://host.docker.internal:5022/openai/deployments/echo/chat/completions"
    }
  },
  "models": {},
  "keys": {
    "dial_api_key": {
      "project": "TEST-PROJECT",
      "role": "default"
    }
  },
  "roles": {
    "default": {
      "limits": {}
    }
  }
}
```

DIAL Adapter:

To add a new model, you need to add configuration to “models”.

“endpoint” is URL within DIAL Core for model

In “upstreams” you configure endpoint to model.
Don’t forget to add your API Key

```
{
  "routes": {},
  "applications": {},
  "models": {
    "gpt-4o": {
      "displayName": "GPT 4o",
      "endpoint": "http://adapter-dial:5000/openai/deployments/gpt-4o/chat/completions",
      "iconUrl": "http://localhost:3001/gpt4.svg",
      "type": "chat",
      "upstreams": [
        {
          "endpoint": "https://ai-proxy.lab.epam.com/openai/deployments/gpt-4o/chat/completions",
          "key": "{REPLACE_WITH_YOUR_API_KEY}"
        }
      ]
    },
  },
  "keys": {
    "dial_api_key": {
      "project": "TEST-PROJECT",
      "role": "default"
    }
  },
  "roles": {
    "default": {
      "limits": {}
    }
  }
}
```

DIAL Adapter:

In “keys” you can set up api keys for DIAL Core and their roles.

By default we are using dial_api_key as a key, if you change it for local usage – don’t forget to change the env variable for local DIAL Chat

```
{
  "routes": {},
  "applications": {},
  "models": {
    "gpt-4o": {
      "displayName": "GPT 4o",
      "endpoint": "http://adapter-dial:5000/openai/deployments/gpt-4o/chat/completions",
      "iconUrl": "http://localhost:3001/gpt4.svg",
      "type": "chat",
      "upstreams": [
        {
          "endpoint": "https://ai-proxy.lab.epam.com/openai/deployments/gpt-4o/chat/completions",
          "key": "{REPLACE_WITH_YOUR_API_KEY}"
        }
      ]
    }
  },
  "keys": {
    "dial_api_key": {
      "project": "TEST-PROJECT",
      "role": "default"
    }
  },
  "roles": {
    "default": {
      "limits": {}
    }
  }
}
```

DIAL Adapter:

Additionally you can set up limits for model usage. You can set up restrictions of how many tokens one user can use for some duration time.

Pay attention that it works for models, for applications it will be ignored.

```
{
  "models": {
    "gpt-4o": {
      "displayName": "GPT 4o",
      "endpoint": "http://adapter-dial:5000/openai/deployments/gpt-4o/chat/completions",
      "iconUrl": "http://localhost:3001/gpt4.svg",
      "type": "chat",
      "upstreams": [
        {
          "endpoint": "https://ai-proxy.lab.epam.com/openai/deployments/gpt-4o/chat/completions",
          "key": "{REPLACE_WITH_YOUR_API_KEY}"
        }
      ]
    }
  },
  "keys": {
    "dial_api_key": {
      "project": "TEST-PROJECT",
      "role": "default"
    }
  },
  "roles": {
    "default": {
      "limits": {
        "gpt-4o": {
          "minute": 100,
          "month": 1000000,
        }
      }
    }
  }
}
```




Join us:



Subscribe to WeAreCommuntiy

<https://wearecommunity.io/communities/dial>

Keep in touch with our latest updates.
Here you find webinars, workshops and
articles about DIALX features and products.

Subscribe to YouTube

<https://www.youtube.com/@TeamDIALX>

Here we publish videos about our newest
products and features.

Join our Discord community

<https://discord.gg/jvTCQv4E4q>

🌟 AI DIALX Community 🌟 is the place where
you can find help with your questions about
DIALX, direct communication with DIALX team
and contributors.



DIALX
COMMUNITY
POWERED BY <epam>

Thank you!