1. API-Endpunkte

1.1 Modelle auflisten

- **Endpoint:** `GET /api/listModels`
- **Beschreibung:** Gibt eine Liste aller verfügbaren Modelle zurück.

Curl-Befehl:

```
curl -X GET http://localhost:9191/api/listModels
```

-

Beispielantwort (JSON):

```
{
  "models": [
    "tinyllama",
    "llama3.1:8b",
    "llama3.2:3b",
    "llama2-uncensored:7b",
    "moondream"
  ]
}
```

-

1.2 Laufende Modelle auflisten

- **Endpoint:** `GET /api/listRunningModels`
- **Beschreibung:** Gibt eine Liste der aktuell laufenden Modelle zurück.

Curl-Befehl:

```
curl -X GET http://localhost:9191/api/listRunningModels
```

-

Beispielantwort (JSON):

```
{
  "runningModels": [
    {
      "modelName": "llama3.2:3b"
    }
  ]
}
```

-

1.3 Modell laden

- **Endpoint:** `POST /api/loadModel`
- **Beschreibung:** Lädt ein Modell in den Speicher.

Curl-Befehl:

```
curl -X POST http://localhost:9191/api/loadModel \
     -H "Content-Type: application/json" \
     -d '{"modelName": "llama3.2:3b"}'
```

-

Beispielantwort (JSON):

```
{
  "modelName": "llama3.2:3b",
  "loaded": true
}
```

-

1.4 Nicht-streaming Antwort generieren

- **Endpoint:** `POST /api/generateResponseNonStreaming`
- **Beschreibung:** Generiert eine Antwort basierend auf einem Prompt (nicht gestreamt).

Curl-Befehl:

```
curl -X POST http://localhost:9191/api/generateResponseNonStreaming \
     -H "Content-Type: application/json" \
     -d '{"prompt": "Warum ist der Himmel blau?"}'
```

-

Beispielantwort (JSON):

```
{
  "response": "Der Himmel erscheint blau, weil die Atmosphäre
Sonnenlicht streut, insbesondere kürzere Wellenlängen wie Blau."
}
```

-

1.5 Streaming Antwort generieren

- **Endpoint:** `POST /api/generateResponseStreaming`
- **Beschreibung:** Generiert eine Antwort basierend auf einem Prompt und streamt sie zurück.

Curl-Befehl:

```
curl -X POST http://localhost:9191/api/generateResponseStreaming \
     -H "Content-Type: application/json" \
     -d '{"prompt": "Erkläre die Relativitätstheorie"}'
```

-

Beispielantwort (JSON-Streaming):

```
{"response": "Die Relativitätstheorie von Einstein..."}
{"response": "...besagt, dass Zeit und Raum relativ sind..."}
```

-

1.6 Server-URL setzen

- **Endpoint:** `POST /api/setServerUrl`

- **Beschreibung:** Setzt die Server-URL für die Kommunikation mit dem LLM.

Curl-Befehl:

```
curl -X POST http://localhost:9191/api/setServerUrl \
     -H "Content-Type: application/json" \
     -d '{"serverUrl": "http://neuer.server.url"}'
```

-

Beispielantwort (JSON):

```
{
  "serverUrl": "http://neuer.server.url",
  "status": "updated"
}
```

-

---

2. Allgemeine Hinweise

- Der Server läuft standardmäßig unter `http://localhost:9191`.
- Jeder Endpunkt gibt eine JSON-Antwort zurück.