OpenAI API-Nutzung im Projekt – Leitblatt für Studierende

✅ Erlaubte Modelle:

Ihr dürft in diesem Projekt nur folgende Modelle verwenden:

- **Chat:** `gpt-4o-mini`, `gpt-3.5-turbo`
- **Embeddings:** `text-embedding-3-small`

➡️ Andere Modelle (z. B. GPT-4, DALL·E, Whisper, TTS etc.) sind **gesperrt**.
Diese Auswahl ist bewusst getroffen: Die Modelle sind **günstig**, **schnell** und **technisch ausreichend** für typische RAG- oder Agentensysteme.

🧱 So könnt Ihr Euren Tokenverbrauch schätzen:

Tokens = GPTs interne „Spracheinheit“. Sie sind **nicht identisch mit Wörtern**.

Faustregeln:

- 1 Wort ≈ 1,3 Tokens
- 1 Satz ≈ 20 Tokens
- 1 GPT-Antwort ≈ 300–500 Tokens (je nach Länge)

Beispielcode:

```
from openai import OpenAI
client = OpenAI()

response = client.chat.completions.create(
    model="gpt-4o-mini",
    messages=[{"role": "user", "content": "Was ist ein Agentensystem in der KI?"}],
)

print(response.usage.total_tokens)  # zeigt den Verbrauch
```

➡️ Nutzt diesen Mechanismus, um den Verbrauch eurer Abfragen einzuschätzen!

Wichtige Metriken im Zusammenhang mit Kosten: TPM und RPM

OpenAI limitiert eure Nutzung mit zwei technischen Größen:

TPM
(Token pro Minute): Wie viele Tokens (Input + Output!) Ihr pro Minute verbrauchen dürft
RPM
(Requests pro Minute): Wie viele API-Anfragen Ihr pro Minute senden dürft

Beispiel:

Wenn der Account `TPM = 50.000` und `RPM = 100` erlaubt, dann dürft Ihr als Gruppe:

- **max. 100 Anfragen pro Minute** senden
- dabei insgesamt **max. 50.000 Tokens pro Minute** verbrauchen
  (z. B. 50 Anfragen à 1.000 Tokens oder 100 Anfragen à 500 Tokens)

⚠️ Diese Limits sind **nicht direkt sichtbar**, aber wenn Ihr sie überschreitet, bekommt Ihr z. B. diesen Fehler:
`openai.RateLimitError: You exceeded your current quota`
oder
`HTTP 429: Too Many Requests`

💡 Tipps für sparsamen Umgang:

- Nutzt `gpt-4o-mini`, nicht `gpt-3.5-turbo`, wenn es nicht nötig ist
- Setzt `max_tokens=200`, um lange Antworten zu vermeiden
- Gebt GPT nur den nötigsten Kontext
- Für Embeddings: große Texte **in Chunks aufteilen** (TextSplitter etc.)
- Keine unkontrollierten Schleifen mit API-Aufrufen

⚠️ Achtung: Missbrauch vermeiden

Bitte **vermeidet alles**, was unnötige Last oder Kosten erzeugt:

- ❌ Endlosschleifen mit GPT-Abfragen
- ❌ Prompt-Spamming oder absichtliches Tokenfüllen
- ❌ Automatisierte oder sinnlose Tool-Use-Tests

📊 Hintergrund

Das OpenAI-Konto läuft über den Dozenten.
Alle Gruppen teilen sich **ein gemeinsames Budget & Limit** (Tokens und Requests).

➡️ Ziel: **Kollaborativ & verantwortungsvoll** arbeiten
➡️ Kein Stress durch hohe Kosten oder Sperren