AGENTUR FÜR FORSCHUNG

GPT-4 und Alternativen in der Praxis: Qualität, Implementierung, Kosten und Datenschutz

Paul Simmering

Succeet 2023

Agenda

1. Einleitung (5 Minuten)
2. Arbeit anhand des Fallbeispiels (25 Minuten)
    a. Zielsetzung und Anforderungen
    b. Lösungsfindung
3. Besprechung (15 Minuten)

Fragebogen

- Nützliche Fragen zur Planung eines Projekts
- Fallbeispiel besprechen
- Optional: Parallel Fragebogen zu eigenem Projekt ausfüllen

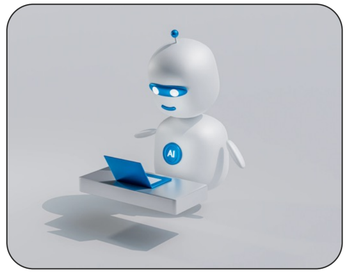
→ Ausgefüllter Fragebogen als Basis für konkrete Empfehlungen

Fragen zum Fragebogen und dem Fallbeispiel gerne zwischendurch stellen.
Auf Ihre eigenen Projekte gehe ich im dritten Teil ein.

Anwendungsbeispiele für Sprachmodelle

Eigener Chatbot

Unternehmens-eigener Chatbot in der eigenen Cloud. Auf eigenen Daten trainiert und autark.

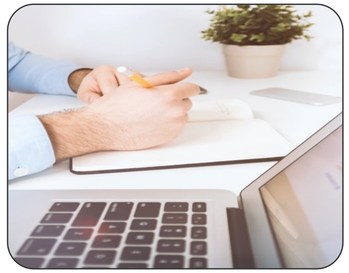


AI Personas

Chatbots bauen, welche Kunden-personas verkörpern.

Inhaltsanalyse

Analyse von Umfrage-antworten, Reviews und anderen Texten nach Inhalt, genannten Entitäten und Sentiments.



Zusammen-fassungen

Texte zusammenfassen und die wichtigsten Aspekte hervorheben.

Protokolle

Transkription, Codierung und Zusammenfassung automatisieren.

Datenbanken per Chat abfragen

Ohne SQL-Kenntnisse Analysen aus Datenbanken generieren.

Fallbeispiel: Analyse von Kundenbewertungen

1. Was soll das Modell tun?

„Die **Grafikkarte** des Gaming-Laptops ist fantastisch, alle meine Spiele laufen darauf problemlos. Große **Festplatte**. Aber die **Lieferung** hat 2 Monate gedauert ☹”

→ Aspekte und Sentiment extrahieren und richtig interpretieren

2. Wie wird die Aufgabe bisher erledigt?

- Suche nach Keywords

Preis, teuer, günstig
Grafikkarte, GPU
Festplatte, Speicher
Lieferung, Versand, Paket

„Die **Grafikkarte** des Gaming-Laptops ist fantastisch, alle meine Spiele laufen darauf problemlos. Große **Festplatte**. Aber die **Lieferung** hat 2 Monate gedauert ☹”

3. Welche Daten soll das Modell zur Bearbeitung der Aufgabe nutzen?

- Texte der Kundenbewertungen

Kauf eines iPhone 14 Pro

★★★★★ 7. Dezember 2022

Als ich mir am 28. September das iPhone 14 Bestellt habe war ich mir unsicher ob es sich für den Preis so sehr lohnt aber als es angekommen ist war ich direkt verliebt es hat mir sehr gut gefallen von der Optik sowohl auch von der Qualität. Wenn ich das Handy in der Hand habe oder ein Handy spiel spiele stört es mich etwas an der Kamera dadurch das sie so groß ist. Mir gefällt sehr die Qualität der Bilder und der Actionmode ist auch ganz Innordnung jedoch ist der Modus für dunkle Orte nicht zu gebrauchen denn er braucht viel Licht. Die Dynamic Island ist sehr gut Designed und gut gehalten, mir gefällt es sehr wenn ich Musik höre und die Dynamic Island mir ein Bild zeigt und sobald ich auf sie lange gedrückt halte kann ich direkt die Musik wechseln und sonstiges. Das Always on Display gefällt mir mega es ist nicht zu dunkel und nicht zu hell es passt sich gut an die Helligkeit der Umgebung an und die Farben des Hintergrundbild sind auch schön gehalten und gehen nicht verloren. Durch immer neue Updates kann man bestimmt schon bald immer mehr mit der Dynamic Island machen. Ich hab das iPhone knappe 3 Monate und im ganzen gefällt und würde eigentlich 5/5 Sternen aber durch den zu hohen Preis kriegt es nur. 4/5. Weniger

· Rezension bereitgestellt von mediamarkt.de

4. Wie soll der Output aussehen?

„Die Grafikkarte des Gaming-Laptops ist fantastisch, alle meine Spiele laufen darauf problemlos. Große Festplatte. Aber die Lieferung hat 2 Monate gedauert ☹”

Aspekt	Sentiment
Grafikkarte	Positiv
Festplatte	Positiv
Lieferung	Negativ

5. Warum soll die Aufgabe mit einem großen Sprachmodell gelöst werden?

- Höhere Genauigkeit als Keyword-Suche erzielen
- Höhere Effizienz als manuelle Auswertung

6. Was macht die erfolgreiche Bearbeitung der Aufgabe aus?

„Die Grafikkarte des Gaming-Laptops ist fantastisch, alle meine Spiele laufen darauf problemlos. Große Festplatte. Aber die Lieferung hat 2 Monate gedauert ☹”

Aspekt	Sentiment
Grafikkarte	Positiv
Festplatte	Positiv
Lieferung	Negativ

→ Korrekte Extraktion der Aspekte und Einschätzung des Sentiments

gemessen an **Präzision** (nur richtige Aspekte extrahieren) und **Sensitivität** (alle gesuchten Aspekte finden)

7. Gibt es Beispiel Input-Output Paare, anhand derer das Modell evaluiert werden kann?

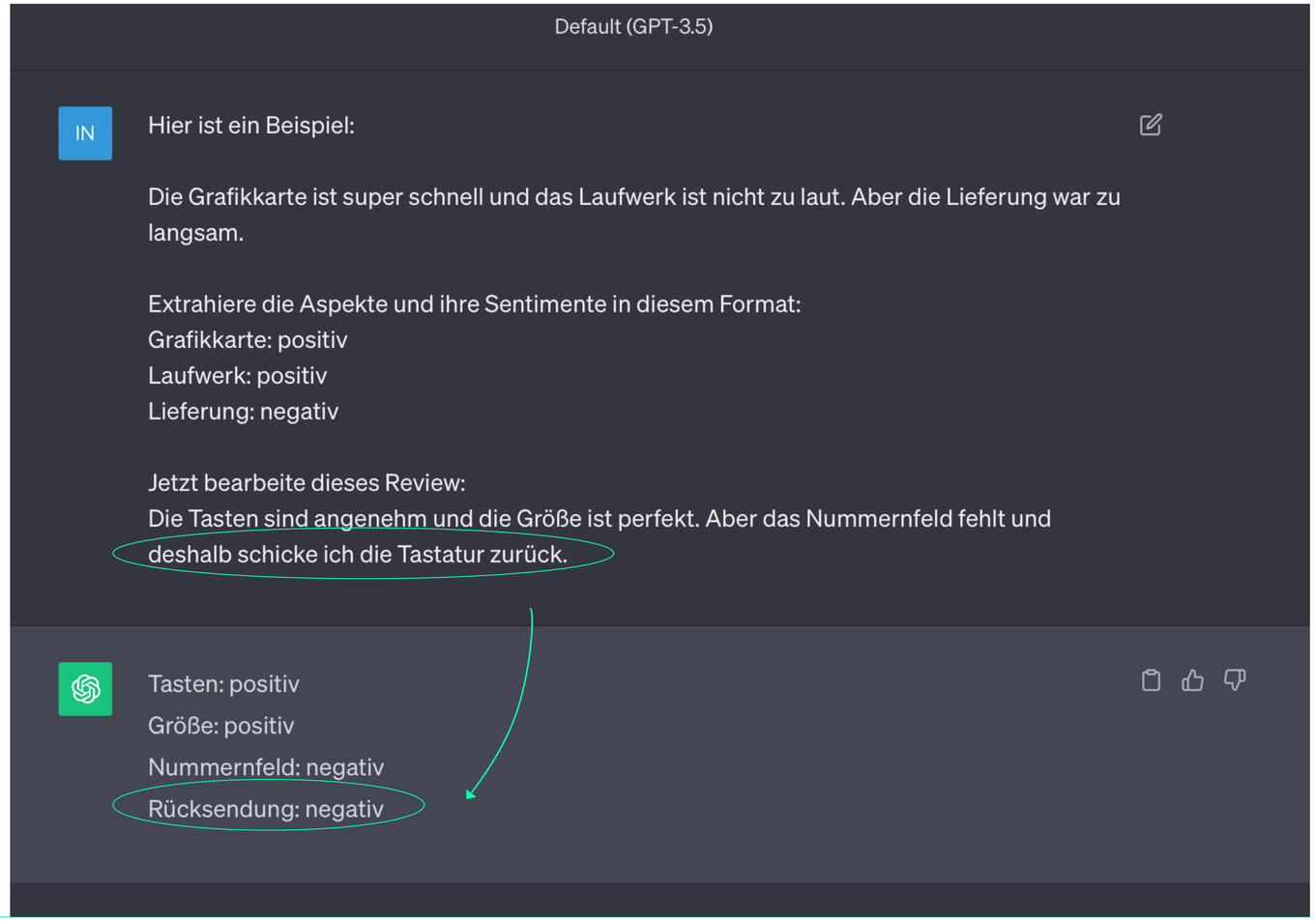
Ja, einen Datensatz von manuell bearbeiteten Beiträgen

Weitere Quellen für Input-Output Paare:

- Öffentliche Datensätze
- Nutzer-Interaktionen (Bewertungen, Likes, Käufe, Klicks…)
- Von einem anderen Modell
- Synthetische Daten

8. Kann ChatGPT die Aufgabe mit Beispiel-Inputs lösen?

- Mit Beispielen ausprobieren
- Kann Aufgabe lösen
- Problem: Falschinterpretation bei der Rücksendung



9. In welchen Sprachen liegen die Inputs vor?

- Deutsch und Englisch

Modell	Merkmal	Sprachen	Urheber	Open Source
GPT-4	Qualität	EN, DE, mehr	OpenAI (Microsoft)	✗
Claude	Kontextlänge	EN, DE, mehr	Anthropic (Amazon)	✗
Luminous	Aus DE	EN, DE, mehr	Aleph Alpha	✗
Mistral	Effizient, offen	EN, DE, mehr	Mistral AI	✓
Llama-2	Offen, Qualität	Primär EN	Meta	✓
LeoLM	Offen, deutsch	DE	LAION, hessian.ai	✓

Führende Modelle (Stand Oktober 2023)

10. Wie lange soll das Modell genutzt werden?

Langfristige Einbindung in eine Data Pipeline

Datenbank Modell BI-Tool

	Modell entwickeln	**Modell anpassen**	**Vorgefertigtes Modell nutzen**
Anschaffung	Extrem teuer	Teuer	Billig
Nutzung	Billig	Billig	Teuer
Genauigkeit	Optimal	Optimal	Kann gut sein

11. Wie häufig wird das Modell im ersten Jahr voraussichtlich aufgerufen?

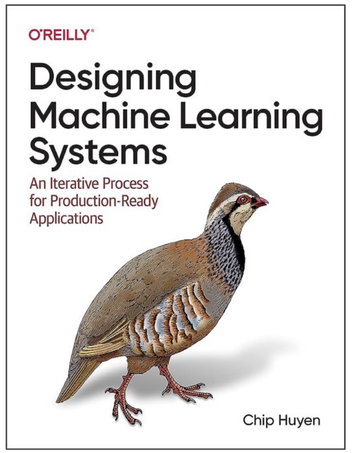
- Bisher sind 150.000 angelaufen
- Monatlich kommen ca. 10.000 Reviews hinzu
- Im ersten Jahr also 270.000 Aufrufe

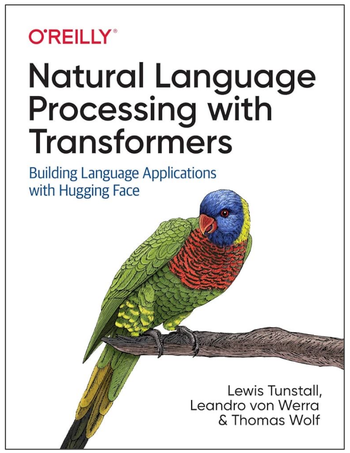
Aufrufe	**1.000**	**10.000**	**100.000**	**270.000**	**10.000.000**
gpt-3.5	1€	10€	100€	270€	10.000€
gpt-3.5 finetuned	8€	80€	800€	2.160€	80.000€
gpt-4	20€	200€	2.000€	5.400€	200.000€

Grobe Rechnung mit 500 Tokens Input und 100 Tokens Output je Review

12. Ist Personal mit Programmierkenntnissen verfügbar?

- Grundzüge Machine Learning sind nötig, aber keine mathematischen Details
- Python ist Standard
- Alternative No-Code Tools:
  - AWS SageMaker Canvas
  - Google Cloud AutoML
  - Azure AutoML
  - Viele viele mehr, sehr unterschiedliche Qualität





13. Hat Ihr Unternehmen eine Cloud?

- Ja, wir benutzen Microsoft Azure

Zahlreiche weitere Dienstleister für Machine Learning Dienstleistungen
Besonders gut: HuggingFace 🤗

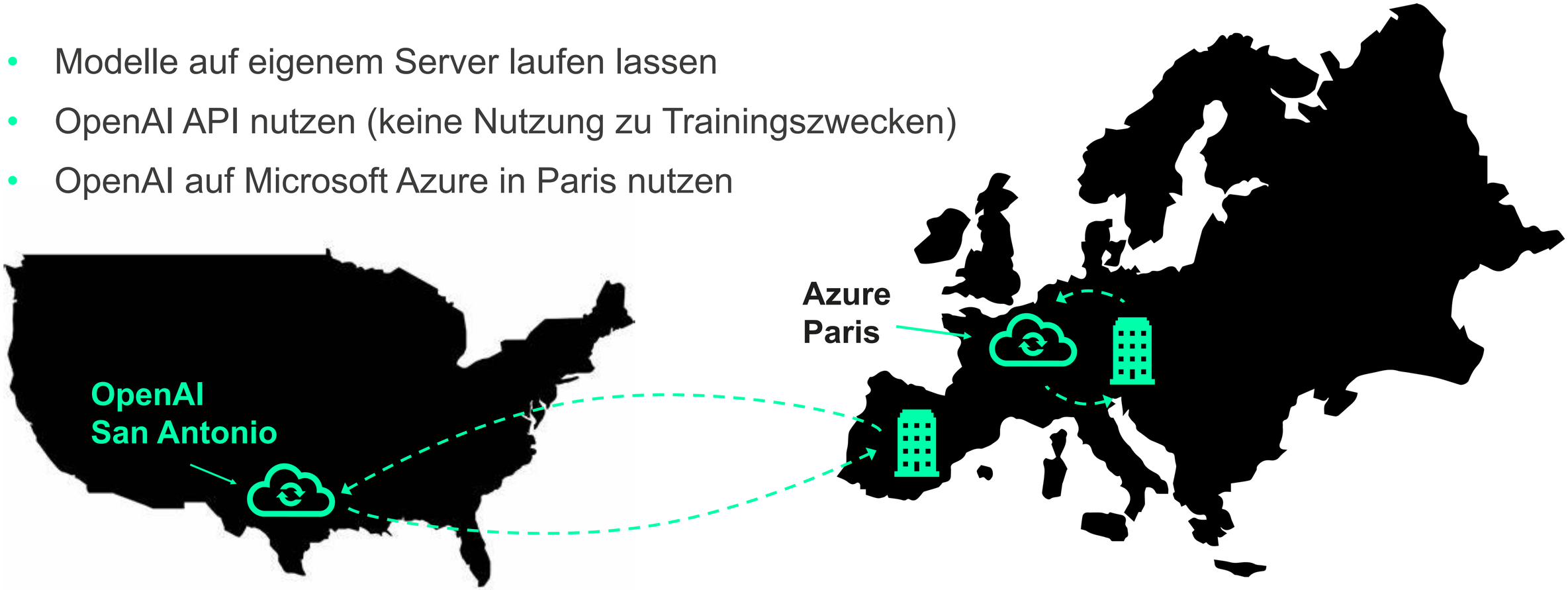
14. Um welche Art von Daten handelt es sich aus Sicht des Datenschutzes?

- Große Sprachmodelle sind rechtlich nicht per se anders als andere Arten der Datenverarbeitung
- Hilfreiche Fragen zur Zusammenarbeit mit Datenschutzbeauftragten:
  - Personenbezogen auf EU-Bürger?
  - Sind es öffentliche Daten?
  - Anonymisierung möglich?

Dies ist keine Rechtsberatung.

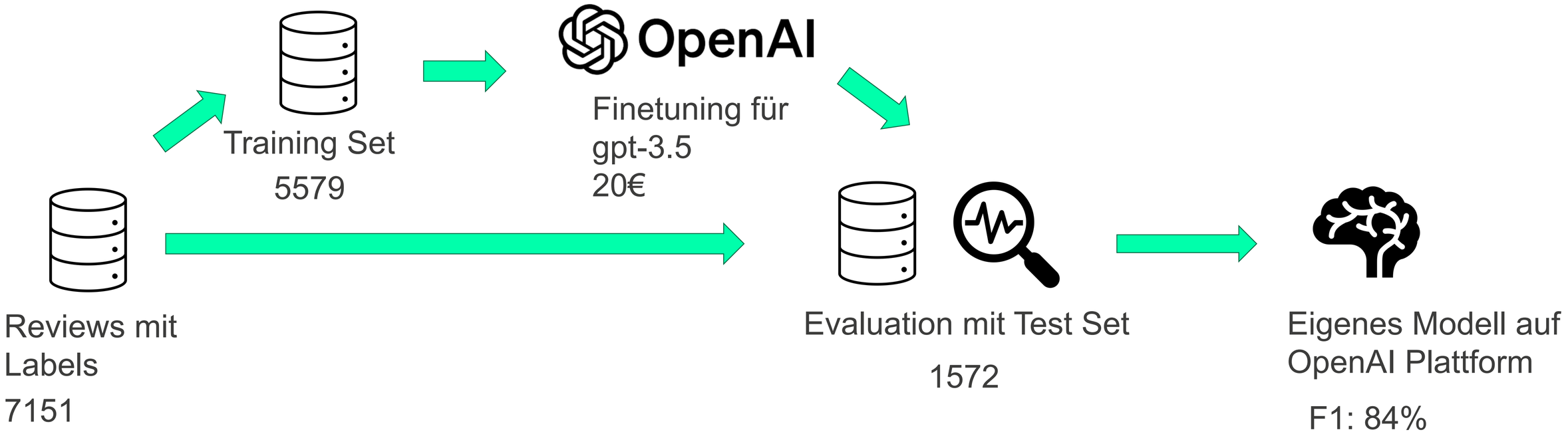
15. Wo sollen diese Daten verarbeitet werden?

- Modelle auf eigenem Server laufen lassen
- OpenAI API nutzen (keine Nutzung zu Trainingszwecken)
- OpenAI auf Microsoft Azure in Paris nutzen



Dies ist keine Rechtsberatung.

Implementierung für das Fallbeispiel



Besprechung Ihrer Projekte

Feedback & Kontakt

Bitte den QR-Code scannen, um Feedback zum Workshop zu geben

Paul Simmering

paul.simmering@teamq.de

www.linkedin.com/in/paulsimmering/

Ich freue mich über neue Kontakte, Fragen und Gespräche zu KI und Data Science.

Q Agentur für Forschung ist an Stand 308.

Q | Agentur für Forschung GmbH
Turley-Strasse 6
D-68167 Mannheim

fon [+49] 06 21 97 65 26-40
fax [+49] 06 21 97 65 26-44

mail info[at]teamq.de
web teamq.de

https://www.facebook.com/Q.Agentur/

https://twitter.com/q_insightagency

https://www.instagram.com/qinsightagency