Mittendrin in einer weiteren digitalen Revolution

Veröffentlicht am 17. Mai 2023

Ihre Ansprechpersonen: Philipp Krauspe, Martin Spielmann

Tags: Fachbeiträge, Innovation, Künstliche Intelligenz (KI), Schulung und Training

Im November 2022 veröffentlicht die open AI Stiftung ChatGPT und macht eine neue Entwicklungsstufe von IT-Systemen mit künstlicher Intelligenz publik. An dieser Stelle lassen wir die Diskussion um Intelligenz aussen vor. Fakt ist, dass diese Veröffentlichung von ChatGPT schon heute grosse Folgen hat.

Innerhalb der APP führen wir regelmässig interne IT-Schulungen zur Weiterbildung unserer Mitarbeitenden durch, Anfang 2023 zum Thema moderne KI-Systeme. Unser Fokus lag dabei auf ChatGPT und dem Online-Bildgenerator Stable Diffusion. Die Rückmeldungen zu den Fähigkeiten von ChatGPT waren durchmischt, ein paar gingen in die Richtung: «Das kann kein Hermes, das ist kein gutes System».

Im einfachen online Zugang ist ChatGPT ein Tool, das prinzipiell den Turing Test bestehen sollte. Es ist darauf optimiert, wie ein Mensch zu antworten. Es erfindet aber Antworten, falls die Informationen dafür im System nicht vorhanden sind.

Trotz fehlender Quellenangaben waren wir neugierig, wie man die dahinterliegende Technologie GPT-3.5 per Programmierschnittstelle dafür nutzen könnte, um Fragen zum Projektmanagementstandard Hermes zu beantworten. Der naheliegendste Ansatz wäre, die gesamte Dokumentation und eine Frage per Schnittstelle zu schicken. Im Standardzugang ist das aufgrund des Umfangs der HERMES-Dokumentation nicht denkbar.

Unsere Versuche endeten entweder erfolglos oder führten zu Kosten von einigen US-Dollar pro Anfrage.

Der Plan: Wir nutzen GPT und lassen uns Fragen zu Hermes 2022 beantworten

Wir betreiben seit einigen Jahre einen Chatbot, der Fragen zu Hermes beantwortet. Unser Plan war, mit Langchain einen mindestens so leistungsfähigen Chatbot zu bauen, der auf Basis von Hermes 2022 Fragen zum Projektmanagement beantwortet.

Das Ziel war, jede Anfrage kostengünstig und korrekt zu beantworten und die passenden Quellen mitzuliefern, damit Nutzende selbst die Orginaldokumentation nachlesen können. Konkret heisst das:

- Die Kosten betragen weniger als 1 Rp. pro Anfrage
- Das System liefert korrekte Antworten unter Zuhilfenahme der Hermes 2022 Dokumentation zurückliefern
- Das System liefert zusammen mit der Antwort einen Link zur Quelle innerhalb der Dokumentation

Die Entwicklung rund um GPT ist rasant

Wir haben in einer Kleingruppe einen Hackathon durchgeführt, um in einem Tag ein System zu schaffen, was diese Anforderungen erfüllt. Seit November 2022 wurden zahlreiche Beispiele zur Nutzung der GPT-API demonstriert.

Wir wollen hier nicht zu sehr auf technische Details eingehen, nur so viel: Im Prinzip haben wir die Idee von Knowledge GPT übernommen. Knowledge GPT erlaubt es, Dokumente hochzuladen und, unter Nutzung von GPT, Fragen zum Inhalt des Dokuments zu beantworten. Wir wollten in unserem System die passende Quelle selbst zur Beantwortung der Fragen mitschicken.

Unser GPT Hermes Chatbot

Wir vereinfachen zur einfachen Lesbarkeit diese Beschreibung und gehen auf die Technologie nicht näher ein. Falls Sie Interesse an weiteren technischen Details haben, nehmen sie gerne Kontakt mit uns auf.

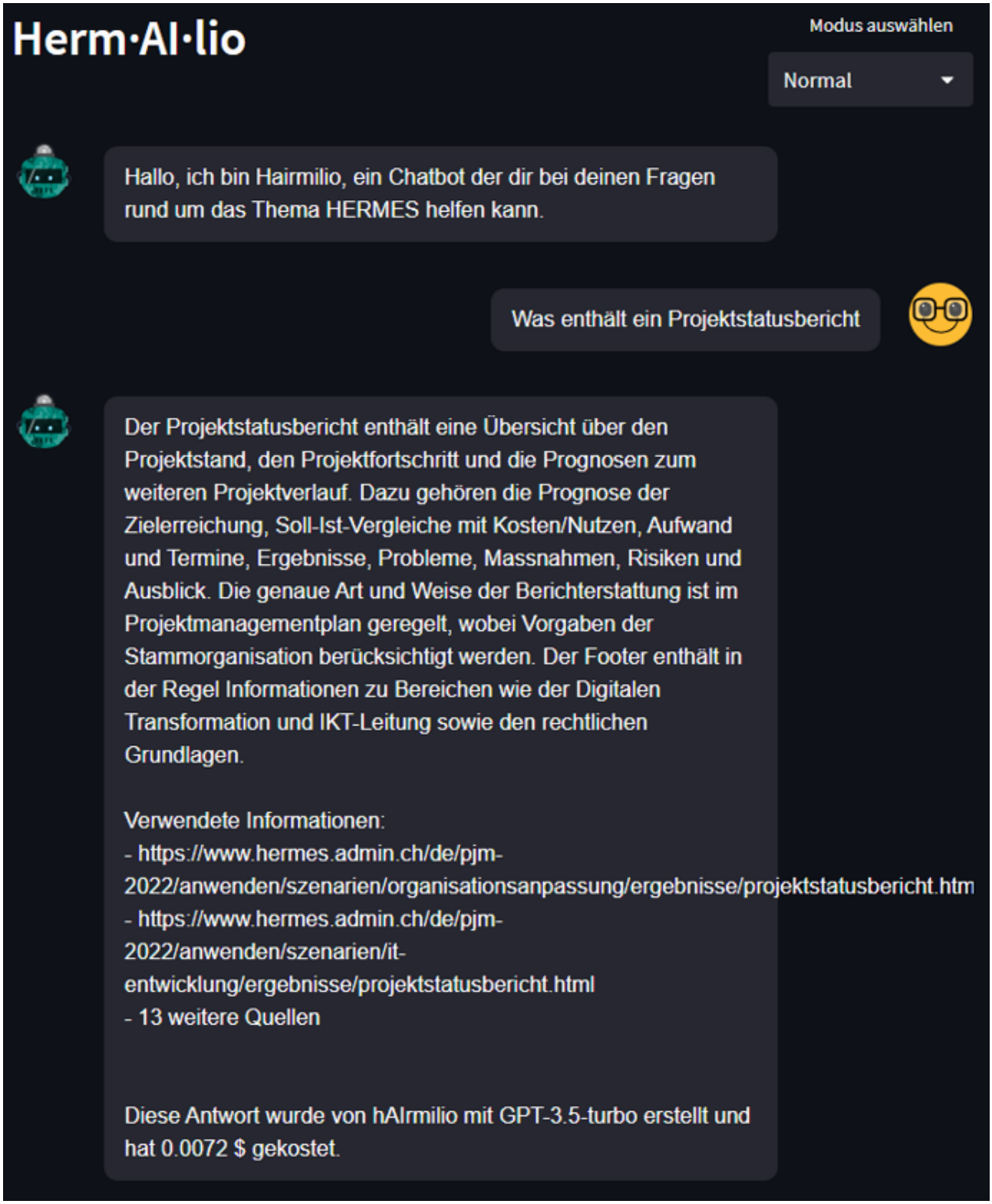
Wir haben die gesamte Hermes 2022 Dokumentation durch GPT indizieren lassen. Mit dem Index (ein sogenannter Vectorstore) konnten wir die Segmente aus der Dokumentation auswählen, die Ähnlichkeiten zur eingegebenen Frage an unser System aufweisen. Dafür sind wir prinzipiell der Anleitung zur Erstellung eines Chroma Vectorstores gefolgt.

Dies ist der erste Schritt, der GPT einsetzt (hier zur Ermittlung sogenannter Embeddings). Zum Beispiel gibt der Input «Was enthält ein Projektstatusbericht» Textbausteine zurück, die grosse Ähnlichkeit mit der Frage aufweisen.

Zusätzlich wird die Ähnlichkeit mit einem Zahlenwert gewertet. Für diese Ermittlung der Ähnlichkeit werden die Embeddings genutzt. Zudem kann durch den Vectorstore auch die jeweilige Quelle innerhalb der Hermes Dokumentation angegeben werden. Im finalen Schritt geben wir nun die passenden Textstellen aus der Dokumentation, die Frage der Userin / des Users ('User Input') sowie einen konkreten Auftrag an die API. Wir setzen diesen aus den folgenden Textbausteinen zusammen:

```
   system_message='Du bist ein Frage-Antwort-
Assistent, der faktengetreu Fragen beantwortet. Du beantwortest alle
Fragen nur mit den Informationen, die dir in der Frage als Kontext
mitgeschickt werden.
Du sagst, wenn du nicht genügend Fakten kennst, um die Fragen richtig
zu beantworten.'
  prompt_kontext='Nutze nur die folgenden Inputs als Ressource zum
Beantworten der Frage: '
  context=funktion_kontext_aus_vectorstore(Eingabe='User Input')
  question_intro='Jetzt beantworte folgende Frage: '
  question='User Input'
```

Auf die Frage, was ein Projektstatusbericht enthält, bekommen wir mit diesen wenigen einfachen Schritten folgende Antwort:



Das Besondere in der Arbeit mit GPT ist die natürliche Sprache. Wir waren überrascht, wieviel besser die Ergebnisse durch die Optimierung der Eingabe in natürlicher Sprache werden (weitere Informationen gibt es dazu unter prompt engineering).

Unsere Lessons Learned im Umgang mit der GPT API

Die Geschwindigkeit, mit der neue Möglichkeiten hinzukommen und mit der Start-Ups neue Lösungen präsentieren, ist rasant. Auch für Anwendungen in der Beratung, der Verwaltung sowie in weiteren kreativen und komplizierten Arbeiten kann die Nutzung von GPT von grossem Vorteil sein. Es gibt Möglichkeiten, diese Systeme zum Guten effizient einzusetzen. Wir sind beeindruckt, wie einfach diese Entwicklungen mit GPT realisierbar sind und auch die kleinen Schritte, die wir gegangen sind, werden in Zukunft ohne jegliche Programmierkenntnisse möglich werden. Erste Demonstrationen der nächsten Version GPT-4 überholen alles, was uns Anwenderinnen und Anwender die letzten 6 Monate beeindrucken konnte.

Trotzdem liefert uns diese Selbsterfahrung zwei dominante Erkenntnisse:

- Die digitale Revolution hat eine neue Geschwindigkeit und wird auch unsere Tätigkeit in naher Zukunft berühren. Interne Schulungen sind unsere Methode, um innovationsfähig zu bleiben.
- Es ist gut möglich durch eigene Erfahrungen die Berührungsängste zu reduzieren und den Einsatz dieser neuen Generation von Onlinetools mit Mehrwert einzusetzen, den Einfluss realistisch einzuschätzen und keine schützenswerten Daten (1) dafür freizugeben.

(1) Hier gehen wir auf den Informationsschutz bei der Eingabe nicht tiefer ein. Wir empfehlen sehr strikt mit den Informationen, die in diese Systeme eingegeben werden umzugehen

Dieser Text wurde komplett selbst verfasst, da das Prompt-Engineering am Ende für einen solchen Text unserer Meinung nach länger gedauert hätte, als den Text selbst zu schreiben.

Möchten Sie mehr über dieses spannende Thema erfahren und weitere Details zum h-AI-rmilio erhalten? Oder haben Sie Interesse eine Schulung, einen minimalen Proof-of-Concept im Einsatz von GPT oder einem verwandten «AI-Tool» durchzuführen und so zu erfahren, welche Chancen und Risiken diese digitale Revolution für ihr Unternehmen bietet? Wir freuen uns auf Ihre Kontaktaufnahme.