



Hochschule für
Wirtschaft und Recht Berlin
Berlin School of Economics and Law

Praxistransferbericht

Vereinfachung von Entwicklungsprozessen durch Large Language Models

vorgelegt am 13. August 2024

| | |
|-------------------------------|-------------------------------|
| Name: | Justin Becker |
| Matrikelnummer: | 77204368351 |
| Ausbildungsbetrieb: | SAP SE |
| Fachbereich: | FB2: Duales Studium — Technik |
| Studienjahrgang: | 2023 |
| Studiengang: | Informatik |
| Betreuer Unternehmen: | Jonathan Busshoff |
| Betreuerin Hochschule: | Carl Dolling |
| Wortanzahl: | 2166 |

Sperrvermerk

Der vorliegende Bericht enthält vertrauliche Daten des Unternehmens SAP SE. Auf Wunsch des Unternehmens SAP SE ist der vorliegende Bericht für die öffentliche Nutzung zu sperren. Veröffentlichung, Vervielfältigung und Einsichtnahme sind ohne ausdrückliche Genehmigung des Unternehmen SAP SE, in 14469 Potsdam und des Verfassers Justin Becker nicht gestattet. Der Bericht ist nur den Gutachtern und des Mitgliedern des Prüfungsausschlusses zugänglich zu machen.

Ort, Datum

Justin Becker

Von der betrieblichen Betreuung zur Kenntnis genommen:

Ort, Datum

Jonathan Busshoff

Zusammenfassung

Mit einem Aufschwung der KI-Technologien findet sich diese in vielen Anwendungsfällen wieder. Dies gilt auch für die Software-Entwicklung. Dieser Bericht untersucht, wie Large Language Models (LLMs) zur Durchsuchung von Programmcode eingesetzt werden können, um einen Überblick über Projekte zu gewinnen und zu bewahren sowie Entwicklungsprozesse zu optimieren. Dafür wird der Ansatz verfolgt, einen Index aus sogenannten Vektor Embeddings zu erstellen, welcher dann mit Fragen natürlicher Sprache durchsucht werden kann. Für das Erstellen der Vektor Embeddings werden LLMs genutzt. Es wird ein Testprojekt und Testfragen herangezogen, um die im folgenden Bericht erarbeiteten Verfahren dafür zu evaluieren.

Inhaltsverzeichnis

| | |
|---|------------|
| Zusammenfassung | I |
| Inhaltsverzeichnis | II |
| Abbildungsverzeichnis | III |
| Tabellenverzeichnis | IV |
| 1 Einleitung | 1 |
| 2 Theoretischen Grundlagen | 2 |
| 2.1 Large Language Models | 2 |
| 2.2 Vektor Embeddings | 2 |
| 2.3 Vektor Embedding Raum | 3 |
| 3 Indizieren | 4 |
| 3.1 Suche | 4 |
| 3.2 Indizierung von Code | 4 |
| 3.2.1 Problem der Code Semantik | 6 |
| 3.2.2 Seperate Bedeutungserfassen via Completions | 6 |
| 3.2.3 Datei spezifische Prompts | 8 |
| 3.2.4 Embeddings aus erweiterten Erklärungen | 8 |
| 4 Bewertung | 10 |
| 5 Fazit | 11 |
| Literatur | 12 |
| Ehrenwörtliche Erklärung | 14 |

Abbildungsverzeichnis

| | | |
|---|--|---|
| 1 | Darstellung von einem ausgedachten Vektor Embedding Raum über zwei Konzeptachsen (Eigene Darstellung) | 3 |
| 2 | Indexing Prozess aus einem Code Abschnitt(Selbst genzeichnet) . . . | 6 |

Tabellenverzeichnis

| | | |
|---|---|---|
| 1 | Testfragen für die Suchen im Index des Testprojekts | 5 |
| 2 | Ergebnisse der Index- und Suchtests | 5 |

Hinweis

In der vorliegenden Arbeit wird auf die Verwendung gendergerechter Sprachformen verzichtet, um eine bessere Lesbarkeit zu gewährleisten. Es wird in der Regel die maskuline Form verwendet, wobei alle Geschlechter gleichermaßen adressiert werden

1 Einleitung

Bereits als 2017 das Paper „Attention is all you need“ von Mitarbeitern bei Google veröffentlicht wurde, gab es Optimismus gegenüber dem vorgestellten Aufmerksamkeitsmechanismus und der Transformer Architektur, welche es ermöglichten, semantische Zusammenhänge in unter anderem Text zu erfassen. Nach erfolgreichen Tests dieser Prinzipien mit Englisch-Deutsch-Übersetzungen wurde der Schluss gezogen: „We are excited about the future of attention-based models and plan to apply them to other tasks. We plan to extend the Transformer to problems involving input and output modalities other than text[...]“[16]. Doch bereits die Textverarbeitung durch KI-Modelle erlebte einen Boom in der Öffentlichkeit mit dem Erscheinen von ChatGPT im November 2022. Ein Chatbot, der die Technologie zur Textverarbeitung von natürlicher Sprache, sogenannte LLMs, die sich den Aufmerksamkeitsmechanismus und die Transformer Architektur zu nutzen machen, leicht zugänglich präsentiert. [3]

Durch die hohe Publizität wurde schnell nach Anwendung der Möglichkeiten in allen Bereichen des Lebens gesucht. So wurde die Technologie im Bereich der Bildung genutzt, um mit ihr zu lernen und z. B. Ergebnisse von Übungsaufgaben abzugleichen. Auch in der Medizin wurden Experimente dazu ausgeführt, um Patienten mit dem Chatbot einen Ansprechpartner zu geben oder sogar Krankheiten durch diesen zu identifizieren. [7]

So kam es ebenfalls dazu, dass LLMs Einzug in die Entwicklungsprozesse von Software gefunden haben und dort eine Reihe von Aufgaben erledigen. Dazu gehört unter anderem die Generierung von Code, das Beheben von Fehlern im Code oder das Erklären von Code. Empirisch konnte belegt werden, dass die Technologie teilweise besser abschneidet, als es ein Mensch durchschnittlich in diesem Anwendungsbereich machen würde[14]. Vor allem in größeren Softwareprojekten kann es zu einem Problem werden, einen Überblick über die Codebasis zu gewinnen und zu behalten. Deshalb wird dieser Praxistransferbericht untersuchen, wie man LLMs nutzen kann, um Sourcecode mit Fragen in natürlicher Sprache zu durchsuchen.

2 Theoretischen Grundlagen

2.1 Large Language Models

LLMs sind KI-Systeme, die darüber charakterisiert sind, dass sie Informationen über natürliche Sprachen erfassen können. Dazu gehören nebst Sprachmustern und Syntax auch die Erfassung von Bedeutung und Zusammenhängen. [19]

Einen Ausdruck dieser Verständnisfähigkeit bringen LLMs durch sogenannte Completions (dt. Vervollständigungen). Dabei wird ein Textabschnitt vorgegeben und das LLM generiert auf dessen Grundlage sukzessiv Folgetext. [10]

Außerdem LLMs in der Regel so konfiguriert, dass sie nicht deterministisch arbeiten, d. h., dass selbst bei einer gleichen Eingabe unterschiedliche Ausgaben entstehen können. Diese Konfiguration ist bei der Arbeit mit LLMs in diesem Bericht vorgenommen.

Da die technische und mathematische Funktionsweise dieser Modelle keine Relevanz für diese Arbeit hat, wird sie hier nicht weiter erläutert.

2.2 Vektor Embeddings

Embeddings sind hochdimensionale Vektoren, die Syntax, Semantik und Zusammenhänge zwischen semantischen Objekten (im Folgenden meist als Bedeutung bezeichnet) darstellen können. Diese Darstellung kann für unterschiedlichste Datenformen wie Text, Bild oder auch Audio gelten. [9]

Dieser Bericht konzentriert sich auf die Darstellung von Text und vernachlässigt andere Formen in den Beschreibungen, aber die Grundprinzipien sind unabhängig von der Form der Daten.

Für das Erfassen von Bedeutung bildet nebst einem Textteil selbst der Kontext eine entscheidende Rolle [13]. Da das Ziel der Generierung von einem Embedding ist, ist die Qualität der Embeddings auch bestimmt durch die Einarbeitung des Kontextes in das Embedding. [5]

2.3 Vektor Embedding Raum

Die Vektor Embeddings werden selten alleine betrachtet, sondern als Teil eines Vektorraums. In diesem Raum wird durch die Embeddings Richtung und Entfernung Bedeutung zugewiesen. So liegt Text, der ähnliche Semantik hat, nahe aneinander in seiner Embedding Darstellung in diesem Vektorraum. Außerdem stellt das Verhältnis zwischen zwei Vektoren semantische Konzepte dar und es können „Konzeptachsen“ gebildet werden, die verallgemeinernd Konzepte erfassen, wie in Abbildung 1 zu erkennen. [4, 9]

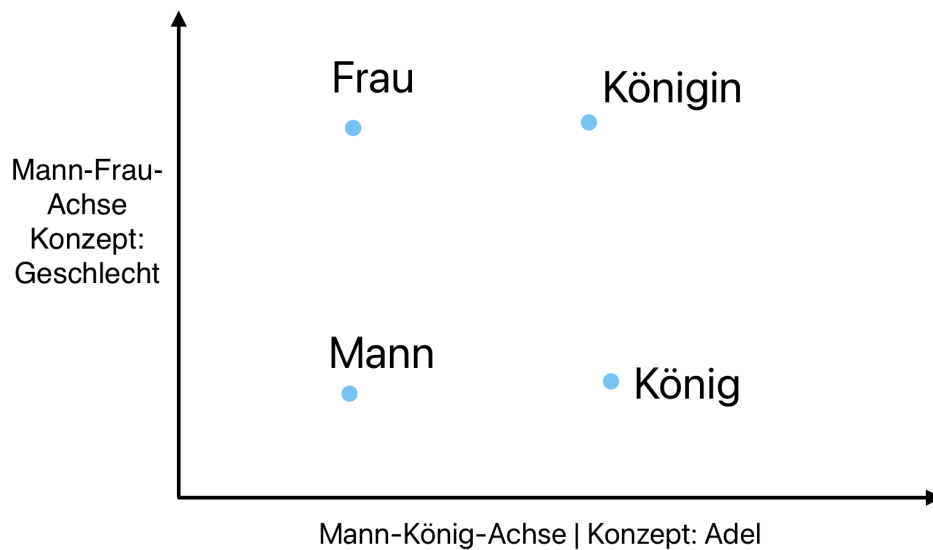


Abbildung 1: Darstellung von einem ausgedachten Vektor Embedding Raum über zwei Konzeptachsen (Eigene Darstellung)

Um einen Vektor Embedding Raum sinnvoll aufzubauen, ist eine einheitliche Methode zur Generierung von Embeddings nötig, in der die Bedeutungszuweisung von Richtungen kohärent zwischen den Embeddings ist. Dafür können LLMs durch ihr Textverständnis genutzt werden, um aus einer Texteingabe ein Ausgabe-Vektor Embedding zu erzeugen. [18]

3 Indizieren

Das Wort „Index“ hat in unterschiedlichsten Anwendungsbereichen unterschiedliche Bedeutung. Gemein haben diese, dass ein Index eine Form von Daten hält, dessen Sinn eine Schlussfolgerung aus eben diesen Daten ist. [2, 8, 17]

Unter dieser Definition stellt auch ein Vektor Embedding Raum, so wie hier vorgestellt, ein Index dar. In dem Vektor Embedding Raum werden Daten als Ausprägung von Information über Text gehalten und es kann aus diesen Daten geschlussfolgert werden, wie bedeutungsnah z. B. zwei Textabschnitte sind.

Der Ablauf der Indizierung mit LLMs wird im Allgemeinen so funktionieren, dass Text vorverarbeitet und in Abschnitte geteilt wird, aus denen für jeden Abschnitt durch ein LLM sein dazugehöriges Vektor Embedding erstellt wird, das dann die Bedeutung dieses Abschnitts repräsentiert. Die resultierenden Vektor Embeddings werden dann in einer Datenbank o. Ä. gespeichert, welche damit den Index beziehungsweise den Vektor Embedding Raum hält. [6]

3.1 Suche

Zur Suche ist eine Frage in natürlicher Sprache gegeben. Aus dem Fragetext wird dann mit dem selben LLM, das für das Indizieren genutzt wurde, ein Vektor Embedding erzeugt. Dieses Embedding wird mit den anderen Embeddings abgeglichen und die zu den ähnlichsten Embeddings gehörenden Textabschnitte werden als Suchergebnis zurückgegeben. Für den Abgleich wird die Kosinus-Ähnlichkeit[12] zwischen allen Vektoren im Index und dem Frage-Vektor berechnet.

3.2 Indizierung von Code

Der erste Schritt zum Indexing ist das Einteilen der Codebasis in Abschnitte. Dies wird hier mit der Tree-Sitter Bibliothek durchgeführt, die das Aufteilen von Code bereitstellt [15]. Als Nächstes wird aus diesen Codeabschnitten Vektor Embeddings erzeugt und im Index gespeichert.

Zum Erfassen der Funktionstüchtigkeit wird eine Metrik eingeführt. Es gibt ein im Anhang enthaltendes Testprojekt und folgende zehn Testfragen. Jede dieser Fragen, die eine „Where“-Formulierungen aufweisen, zielt auf einen Codeabschnitt des Projektes ab.

Tabelle 1: Testfragen für die Suchen im Index des Testprojekts

| | |
|-----|---|
| 1. | „Where is the logger implemented?“ |
| 2. | „Where are the setting changes for the logger handled?“ |
| 3. | „Where is the linting configured?“ |
| 4. | „Where is typescript configured?“ |
| 5. | „Where are the styles for my app?“ |
| 6. | „Where is my app created?“ |
| 7. | „Where do I create the IDs for my web view context?“ |
| 8. | „Where do I manage the state of my web view panels?“ |
| 9. | „Where do I configure what commands will be available in my vs code extension?“ |
| 10. | „Where do I retrieve information about vs code settings?“ |

Für einen ersten Test 1 wird für das Projekt nach vorgestelltem Verfahren ein Index erstellt, welcher dann mit den Fragen durchsucht wird. Enthalten die drei besten Ergebnisse, also jene, dessen Embedding am meisten Kosinus-Ähnlichkeit mit dem Embedding der Frage habe, den Textabschnitt, der durch die Frage gemeint ist, wird das Ergebnis als richtig gewertet. Für richtige Ergebnisse wird die Kosinus-Ähnlichkeit notiert, sonst eine 0.

Alle Ergebnisse wurden mithilfe der gpt-35-turbo [1] und text-embedding-ada-002 [11] Modelle erzeugt. Für Completions wurde eine Temperature von 0.2 und ein Top-P-Wert von 0.3 gewählt. Alle anderen Einstellungen blieben unverändert.

Tabelle 2: Ergebnisse der Index- und Suchtests

| Frage | Test 1 | Test 2 | Test 3 | Test 4 | Test 5 | Test 6 | Test 7 |
|-------|--------|--------|--------|--------|--------|--------|--------|
| 1. | 0,7993 | 0 | 0,7993 | 0,7937 | 0,8008 | 0,7925 | 0,8062 |
| 2. | 0,8172 | 0,793 | 0,8172 | 0,7956 | 0,8307 | 0,8049 | 0,8229 |
| 3. | 0 | 0 | 0 | 0 | 0,8436 | 0 | 0 |
| 4. | 0,8036 | 0,8039 | 0,8606 | 0,8606 | 0,8556 | 0,852 | 0,8428 |
| 5. | 0 | 0 | 0 | 0 | 0,8138 | 0,7576 | 0,7912 |
| 6. | 0,7816 | 0,7896 | 0,777 | 0,777 | 0 | 0,7895 | 0,7912 |
| 7. | 0 | 0 | 0,8087 | 0 | 0,8087 | 0,7844 | 0 |
| 8. | 0 | 0 | 0 | 0 | 0,8596 | 0 | 0,7533 |
| 9. | 0,8004 | 0 | 0,854 | 0,8928 | 0,8516 | 0 | 0 |
| 10. | 0,7558 | 0,7589 | 0,8434 | 0,8336 | 0,8434 | 0,8337 | 0,8227 |

3.2.1 Problem der Code Semantik

In Test 1 konnten sechs von zehn Fragen richtig beantwortet werden. Zur weiteren Untersuchungen wurde Test 2 durchgeführt. In diesem Test war der Prozess des Indexing und der Suche gleich, doch wurde das Testprojekt insofern angepasst, dass sämtlich Dokumentation gelöscht wurde und Namen von Variablen nicht deskriptiv, z. B. „logger“ zu „l“, waren. Die vorgestellte Indexing und Suchmethode konnte dieser Änderung nicht standhalten und nur noch vier von zehn Fragen richtig beantworten. Außerdem ist anzumerken, dass die Fragen, die richtig beantwortet wurden, jene sind, bei denen der dazugehörige Codeabschnitt weiterhin viele deskriptive Namen enthält, da dort externe Bibliotheken genutzt werden, deren Namen nicht angepasst werden können.

Aus diesen beiden Punkten lässt sich vermuten, dass bei der Erzeugung der Embeddings vor allem aus der Namensgebung Information gezogen wird und nicht aus der Logik des Codes.

3.2.2 Seperate Bedeutungserfassen via Completions

Um dieses Problem der Erfassung der Semantik von Code anzugehen, wird der Indexing-Prozess für folgende Tests angepasst.

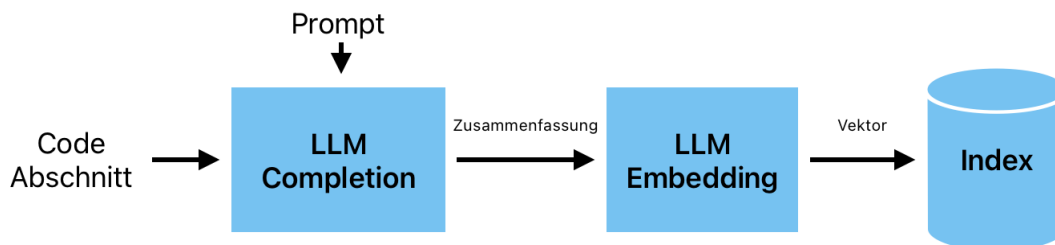


Abbildung 2: Indexing Prozess aus einem Code Abschnitt(Selbst gezeichnet)

Dafür wird, wie in Abbildung 2 illustriert, vor der Embedding Generierung ein weiteres LLM aufgerufen, welches eine Completion durchführt, die als Starttext einen Prompt und den Codeabschnitt erhält. Das Prompt fordert das LLM auf, den erhaltenen Code zusammenzufassen und diese Zusammenfassung wird dann für das Embedding genutzt. Einerseits kann dem ersten LLM durch dieses Prompt mehr Kontext gegeben werden und andererseits ist das Ergebnis dieser ersten Analyse durch das LLM ein Text in natürlicher Sprache, der besser von dem Embedding LLM erfasst werden kann. Als Prompt wird folgender Text genutzt:

You are a senior software developer. I provide you with a fenced code snippet that is part of a larger codebase.

Do the following things:

1. Summarize what the code does in non technical terms, in 2 sentences. (summary)
2. Briefly describe how the code works. (analysis)
3. Imagine a developer needs to find this code based on its functionality in the codebase. List 3-5 "Where" questions a developer might ask to locate this code, based on its functionality/intent. (questions)

Take your time to do a thorough review and identify meaningful questions.

Return nothing but JSON,

Here is an example:

```
{
  "summary": "This code snippet sorts an array of numbers. It supports
    sorting in ascending or descending order.",
  "analysis": "The code implements the quick sort algorithm to recursively
    esort an array in ascending or descending order.",
  "questions": [
    "Where is the code used for sorting data?",
    "Where can I find the array sorting logic?",
    "Where is the code that implements quicksort?"
  ]
}
```

If the code snippet is empty, just return the text ["NOTHING"].

Do not start with "Here is the JSON object" or similar, just output plain JSON no markdown!

<Code Abschnitt wird hier eingefuegt>

Im Test 3 wird dieses neue Verfahren genutzt und sieben von zehn Fragen können richtig beantwortet werden. Beim Test 4 wird mit dem selben Verfahren das Testprojekt im Zustand aus Test 2 untersucht und durch die Verbesserungen werden noch sechs von zehn Fragen richtig beantwortet. Dabei ist anzumerken, dass die Frage, die beim Vergleich von Test 3 und Test 4 wegfällt, selbst von einem Menschen nur durch die Dokumentation zu beantworten wäre, da der Codeabschnitt auch für andere Zwecke eingesetzt werden könnte als in der Dokumentation dargelegt und durch die Frage gemeint.

3.2.3 Datei spezifische Prompts

Um den Kontext für die LLMs weiter zu erhöhen, wird dem Prompt in Test 5 der relative Dateipfad zu den Codeabschnitt mitgegeben, indem der Prompt durch den folgenden Satz ergänzt wird:

The file path for the snippet is "<Pfad zur Datei des Codeabschnittes>".

Aufgrund dieser Änderung wird nun auch der Dateiname übergeben, der mehr Kontext über den Zweck der Datei bereitstellen kann. Deutlich wird dies am Beispiel von Frage 3, die nach der ESLint Konfiguration fragt. In den vorherigen Tests konnte diese Frage nicht beantwortet werden. Die Antwort des Completion LLM war das vorgegebene „NOTHING“, welches nur zurückgegeben werden soll, wenn kein Code übergeben wurde. Mit dieser Anpassung war es dem LLM allerdings möglich, Konfigurationen richtig einzuordnen und es wurden neun von zehn Fragen richtig beantwortet.

Dabei ist auffällig, dass Frage 6 nicht beantwortet werden konnte, obwohl dies bei allen vorherigen Tests möglich war. Erklärbar ist dieses Verhalten dadurch, dass der Codeabschnitt selber im Wortlaut ähnlich zu der Frage ist und deshalb in den meisten Tests erkannt wird. Doch bei Test 5 wird dieser Abschnitt so abstrakt beschrieben, dass die Embeddings zu fern für eine erfolgreiche Suche sind.

3.2.4 Embeddings aus erweiterten Erklärungen

Ein weiterer Ansatz, die Ergebnisse zu verbessern, war es, neben einer Zusammenfassung auch Beispielfragen zu generieren, die dann zusammen in einen Embedding Vektor gewandelt werden. Dafür wurde mit den Completions stand Test 3 und Test 4 gearbeitet. Die Ergebnisse von diesem Test 6 zeigen allerdings keine Verbesserung im Vergleich zu dem Test 3. Gründe dafür könnten sein, dass die erzeugten Fragen teilweise nur einzelnen Detailspekte des Codes erfassen(1.) oder sehr allgemein formuliert sind(2.). Beispiele für solche Fragen sind:

-
1. "Where is the code that defines the configuration properties for logging level and source location tracking?"
 2. "What is the purpose of the JSON object?"
-

In einem abschließenden Versuch wird nur das übergeben, was im Prompt Analyse bezeichnet ist und meist etwas ausführlicher als die Zusammenfassung war. In diesem Test 7 kann allerdings auch keine Verbesserung festgestellt werden.

4 Bewertung

Bei der Bewertung der Ergebnisse muss zuerst die genutzte Metrik bewertet werden. Sowohl das ausgewählte Testprojekt als auch die zehn Testfragen wurden nur beispielhaft gewählt und haben geringen repräsentativen Aussagewert, da die Ergebnisse für andere Fragen an einem anderen Projekt komplett anders ausfallen könnten. Z. B. hätte Test 7 bei detailreicheren Fragen vielleicht besser abschneiden können. Die Metrik ist dementsprechend nur bedingt aussagekräftig.

Außerdem wurden alle Tests genau einmal durchgeführt. Da die LLMs allerdings nicht deterministisch Ergebnisse erzeugen, könnten auch bei einem zweiten Testlauf andere Ergebnisse entstehen. Zudem wurde vernachlässigt, dass durch z. B. Netzwerkprobleme oder eine Limitierung der Aufruftrate von den LLMs abgefälschte Ergebnisse entstanden sein können. Die Ergebnisse wurden auch nur mit den zwei vorgestellten Modellen ermittelt und es lässt sich nicht verallgemeinernd auf die Funktionsweise bei anderen Modellen schließen.

Insgesamt sind die Ergebnisse deshalb kritisch zu betrachten. Trotzdem kann die Hypothese aufgestellt werden, dass das weitere Analysieren und Zusammenfassen von Code mit einem weiteren Completion-LLM, welches mehr Kontext verarbeiten kann, vorteilhaft ist. Dabei sind die Nachteile anzumerken, dass der weitere LLM-Aufruf viel Zeit verbraucht und auch zusätzliche Kosten verursacht.

5 Fazit

Abschließend lässt sich festhalten, dass empirisch das Potenzial der erarbeiteten Indizierungs-Techniken dargelegt werden konnte. In weiterer Arbeit besteht dabei die absolute Notwendigkeit, die Tests weiter zu vertiefen. Also das Ausweiten auf mehrere Testprojekte und mehr Testfragen zu erstellen, welche nicht nur von einer einzelnen Person, sondern von einer Vielzahl von Entwicklern aufgestellt wurde.

Außerdem können noch weitere Techniken für das Indizieren erkundet werden. Dazu gehört zum Beispiel, dass für unterschiedliche Dateitypen noch mehr Datei spezifischer Kontext gegeben wird, damit die LLMs den Code noch besser einordnen können oder die Verwendung von anderen Suchmethoden.

Literatur

- [1] *Azure OpenAI Service-Modelle*. URL: <https://learn.microsoft.com/de-de/azure/ai-services/openai/concepts/models#gpt-35> (besucht am 08.08.2024).
- [2] A. Chatterjee. „Index and Indexing“. In: (2017), S. 125–134. DOI: [10.1016/B978-0-08-102025-8.00009-0](https://doi.org/10.1016/B978-0-08-102025-8.00009-0).
- [3] Will Douglas Heavenarchive. *The inside story of how ChatGPT was built from the people who made it*. MIT Technolog Review. 2023. URL: <https://www.technologyreview.com/2023/03/03/1069311/inside-story-oral-history-how-chatgpt-built-openai/> (besucht am 01.08.2024).
- [4] Florian Heimerl und Michael Gleicher. „Interactive analysis of word vector embeddings“. In: *Computer Graphics Forum*. Bd. 37. 3. Wiley Online Library. 2018, S. 253–265.
- [5] Eric Huang u. a. „Improving Word Representations via Global Context and Multiple Word Prototypes“. In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Hrsg. von Haizhou Li u. a. Jeju Island, Korea: Association for Computational Linguistics, Juli 2012, S. 873–882. URL: <https://aclanthology.org/P12-1092>.
- [6] Xiang Ji u. a. „Speeding Up Question Answering Task of Language Models via Inverted Index“. In: *arXiv preprint arXiv:2210.13578* (2022).
- [7] Yiheng Liu u. a. „Summary of chatgpt-related research and perspective towards the future of large language models“. In: *Meta-Radiology* (2023), S. 100017.
- [8] A. Lo. „What Is an Index?“ In: *The Journal of Portfolio Management* 42 (2016), S. 21–36. DOI: [10.3905/jpm.2016.42.2.021](https://doi.org/10.3905/jpm.2016.42.2.021).
- [9] Tomas Mikolov u. a. „Efficient estimation of word representations in vector space“. In: *arXiv preprint arXiv:1301.3781* (2013).
- [10] Humza Naveed u. a. „A comprehensive overview of large language models“. In: *arXiv preprint arXiv:2307.06435* (2023).
- [11] *New and improved embedding model*. URL: <https://openai.com/index/new-and-improved-embedding-model/> (besucht am 08.08.2024).
- [12] Faisal Rahutomo, Teruaki Kitasuka, Masayoshi Aritsugi u. a. „Semantic cosine similarity“. In: *The 7th international student conference on advanced science and technology ICAST*. Bd. 4. 1. University of Seoul South Korea. 2012, S. 1.

- [13] Herbert Rubenstein und John B Goodenough. „Contextual correlates of synonymy“. In: *Communications of the ACM* 8.10 (1965), S. 627–633.
- [14] Haoye Tian u. a. „Is ChatGPT the ultimate programming assistant—how far is it?“ In: *arXiv preprint arXiv:2304.11938* (2023).
- [15] *Using Parsers*. URL: <https://tree-sitter.github.io/tree-sitter/using-parsers> (besucht am 08.08.2024).
- [16] Ashish Vaswani u. a. „Attention is all you need“. In: *Advances in neural information processing systems* 30 (2017).
- [17] B. Vickery. „THE STRUCTURE OF A CONNECTIVE INDEX“. In: *Journal of Documentation* 6 (1950), S. 140–151. DOI: [10.1108/EB026158](https://doi.org/10.1108/EB026158).
- [18] Xin Zhang u. a. „Language Models are Universal Embedders“. In: *arXiv preprint arXiv:2310.08232* (2023).
- [19] Yizhen Zheng u. a. „Large language models for scientific synthesis, inference and explanation“. In: *arXiv preprint arXiv:2310.07984* (2023).

Ehrenwörtliche Erklärung

Ich erkläre ehrenwörtlich:

1. dass ich meinen Praxistransferbericht selbstständig verfasst habe,
2. dass ich die Übernahme wörtlicher Zitate aus der Literatur sowie die Verwendung der Gedanken anderer Autoren an den entsprechenden Stellen innerhalb der Arbeit gekennzeichnet habe,
3. dass ich meinen Praxistransferbericht bei keiner anderen Prüfung vorgelegt habe.

Ich bin mir bewusst, dass eine falsche Erklärung rechtliche Folgen haben wird.

Ort, Datum

Justin Becker