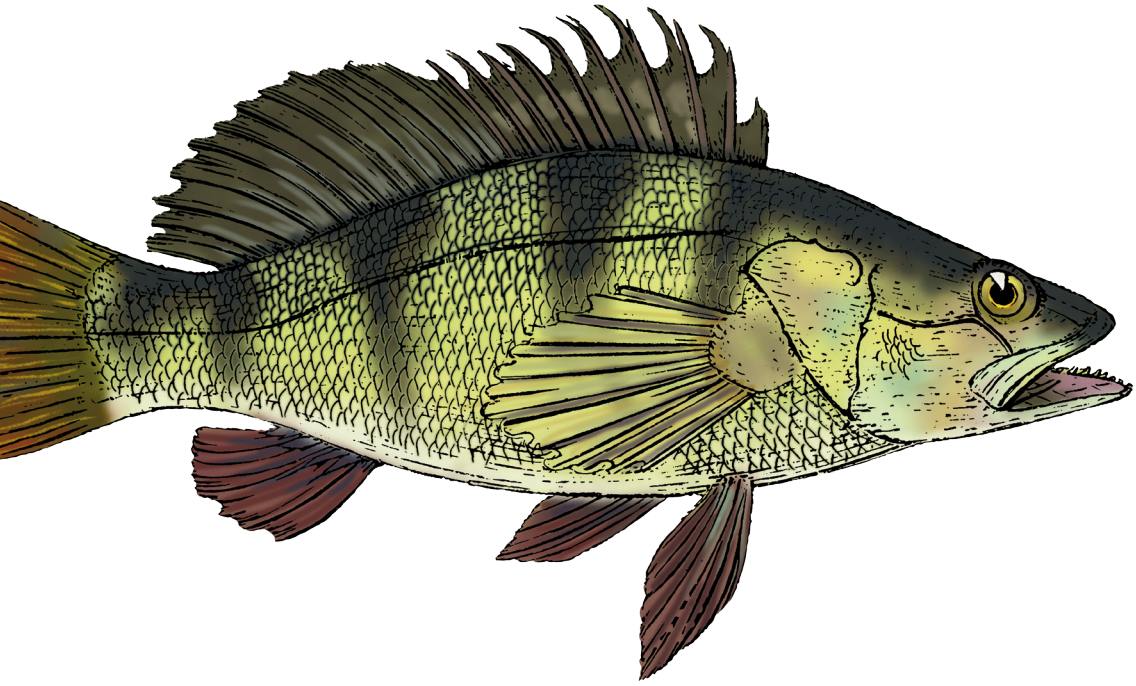
O'REILLY®

Deutsche Ausgabe

Praxiseinstieg Large Language Models

Strategien und Best Practices für den Einsatz von ChatGPT und anderen LLMs



Sinan Ozdemir

Übersetzung von Frank Langenau

Lob für »Praxiseinstieg Large Language Models«

»Indem er das Potenzial sowohl von Open-Source- als auch von Closed-Source-Modellen abwägt, präsentiert sich *Praxiseinstieg Large Language Models* als umfassender Leitfaden für das Verständnis und die Verwendung von LLMs, der die Kluft zwischen theoretischen Konzepten und praktischer Anwendung überbrückt.«

– Giada Pistilli, Principal Ethicist bei Hugging Face

»Eine erfrischende und inspirierende Ressource. Vollgepackt mit praktischen Anleitungen und klaren Erläuterungen, die Sie in diesem spektakulären Gebiet klüger machen.«

– Pete Huang, Autor von The Neuron

»Wenn es darum geht, große Sprachmodelle (*Large Language Models*, LLMs) zu erstellen, erweist es sich mitunter als schwierig, umfassende Ressourcen zu finden, die alle wesentlichen Aspekte abdecken. Meine Suche nach einer solchen Ressource hatte jedoch kürzlich ein Ende, als ich dieses Buch entdeckte.

Sinan zeichnet sich unter anderem durch seine Fähigkeit aus, komplexe Konzepte auf einfache Weise zu präsentieren. Der Autor hat hervorragende Arbeit geleistet, indem er komplizierte Ideen und Algorithmen aufgeschlüsselt hat, sodass Leser sie verstehen können, ohne sich überfordert zu fühlen. Er erklärt jedes Thema sorgfältig und baut dabei auf Beispielen auf, die als Sprungbrett für ein besseres Verständnis dienen. Dieser Ansatz bereichert die Lernerfahrung und macht selbst die kompliziertesten Aspekte der LLM-Entwicklung für Leserinnen und Leser mit unterschiedlichem Wissensstand zugänglich.

Eine weitere Stärke dieses Buchs ist die Fülle an Coderessourcen. Das Einbeziehen von praktischen Beispielen und Codefragmenten ist ein Gamechanger für jeden, der experimentieren und die gelernten Konzepte anwenden will. Diese Coderessourcen vermitteln dem Leser praktische Erfahrungen und ermöglichen ihm, die eigenen Kenntnisse zu testen und aufzubessern. Dies ist von unschätzbarem Wert, da es ein tieferes Verständnis der Materie fördert und es dem Leser erlaubt, sich wirklich mit dem Inhalt auseinanderzusetzen.

Zusammenfassend lässt sich sagen, dass dieses Buch ein Glückstreffer für jeden ist, der sich für den Aufbau von LLMs interessiert. Die außergewöhnliche Qualität der Erklärungen, der klare und prägnante Schreibstil, die reichhaltigen Coderessourcen und die umfassende Abdeckung aller wesentlichen Aspekte machen es zu einer unverzichtbaren Ressource. Ob Sie nun Anfänger oder erfahrener Praktiker sind, dieses Buch wird zweifellos Ihr Verständnis und Ihre praktischen Fertigkeiten in der LLM-Entwicklung erweitern. Ich empfehle *Praxiseinstieg Large Language Models* jedem, der sich auf die aufregende Reise begeben will, LLM-Anwendungen zu erstellen.«

– Pedro Marcelino, Machine Learning Engineer,
Mitbegründer und CEO @overfit.study

Praxiseinstieg Large Language Models

Praxiseinstieg Large Language Models

Strategien und Best Practices für den Einsatz von ChatGPT und anderen LLMs

Sinan Ozdemir

Deutsche Übersetzung von Frank Langenau

O'REILLY®

Sinan Ozdemir

Lektorat: Alexandra Follenius
Übersetzung: Frank Langenau
Copy-Editing: Sibylle Feldmann, *www.richtiger-text.de*
Satz: III-satz, *www.drei-satz.de*
Herstellung: Stefanie Weidner
Umschlaggestaltung: Karen Montgomery, Michael Oréal, *www.oreal.de*

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über *http://dnb.d-nb.de* abrufbar.

ISBN:
Print 978-3-96009-240-7
PDF 978-3-96010-853-5
ePub 978-3-96010-854-2

1. Auflage 2024

Wieblinger Weg 17
69123 Heidelberg

Authorized German translation of the English edition of *QUICK START GUIDE TO LARGE LANGUAGE MODELS: Strategies and Best Practices for Using ChatGPT and Other LLMs* 1st Edition by Sinan Ozdemir, published by Pearson Education, Inc, publishing as Addison-Wesley Professional © 2024 Pearson Education, Inc.

Dieses Buch erscheint in Kooperation mit O'Reilly Media, Inc. unter dem Imprint »O'REILLY«.
O'REILLY ist ein Markenzeichen und eine eingetragene Marke von O'Reilly Media, Inc. und wird mit Einwilligung des Eigentümers verwendet.

Schreiben Sie uns:
Falls Sie Anregungen, Wünsche und Kommentare haben, lassen Sie es uns wissen: *kommentar@oreilly.de*.

Inhalt

Vorwort

Obwohl die Verwendung von großen Sprachmodellen – *Large Language Models* (LLMs) – schon in den letzten fünf Jahren stetig zugenommen hat, ist das Interesse daran geradezu explodiert, als OpenAI sein Produkt ChatGPT veröffentlichte. Der KI-Chatbot hat die Leistungsfähigkeit von LLMs demonstriert und eine einfach zu bedienende Schnittstelle eingeführt, die es Menschen aus allen Gesellschaftsschichten ermöglicht, die Vorteile dieses bahnbrechenden Tools zu nutzen. Jetzt, da diese Untergruppe der Verarbeitung natürlicher Sprache – *Natural Language Processing* (NLP) – zu einem der meistdiskutierten Bereiche des maschinellen Lernens geworden ist, wollen viele Menschen sie in ihre eigenen Angebote integrieren. Diese Technologie fühlt sich tatsächlich so an, als könnte es sich um künstliche Intelligenz handeln, auch wenn es lediglich um die Vorhersage von aufeinanderfolgenden Token anhand eines probabilistischen Modells geht.

Praxiseinstieg Large Language Models ist ein exzellenter Überblick über das Konzept der LLMs sowie deren praktische Anwendung, und zwar für Programmiererinnen und Programmierer mit und ohne Vorkenntnisse in Data Science. Die Mischung aus Erklärungen, visuellen Darstellungen und praktischen Codebeispielen macht das Buch zu einer fesselnden und leicht verständlichen Lektüre, die dazu anregt, immer wieder umzublättern. Sinan Ozdemir deckt viele Themen in einer anschaulichen Art und Weise ab und macht dieses Buch damit zu einer der besten Informationsquellen, die zur Verfügung stehen, um etwas über LLMs, ihre Fähigkeiten und den Umgang mit ihnen zu lernen und damit die besten Ergebnisse zu erzielen.

Sinan wechselt geschickt zwischen verschiedenen Aspekten von LLMs und gibt dem Leser alle Informationen, die er braucht, um LLMs effektiv zu nutzen. Beginnend mit der Diskussion, wo LLMs innerhalb von NLP angesiedelt sind, und der Erklärung von Transformern und Encodern, geht er auf Transfer Learning und Feintuning, Attention und Tokenisierung in einer verständlichen Art und Weise ein. Außerdem befasst er sich mit vielen weiteren Aspekten von LLMs, zu denen gehören: die Kompromisse zwischen Open-Source-Modellen und kommerziellen Optionen, wie man Vektordatenbanken nutzt (schon für sich genommen ein sehr beliebtes Thema), das Schreiben eigener APIs mit Fast API, das Erstellen von Embeddings

und das Überführen von LLMs in die Produktion – etwas, das sich für jede Art von Machine-Learning-Projekt als Herausforderung erweisen kann.

Ein großer Teil dieses Buchs beschäftigt sich sowohl mit visuellen Schnittstellen – wie zum Beispiel ChatGPT – als mit auch Schnittstellen für die Programmierung. Sinan stellt hilfreichen Python-Code zur Verfügung, der leicht verständlich ist und klar veranschaulicht, was im Einzelnen passiert. Im Rahmen des Prompt Engineering führt er vor, wie sich drastisch bessere Ergebnisse von LLMs erzielen lassen, und – was noch besser ist – er demonstriert, wie man diese Prompts sowohl in der visuellen GUI als auch über die Python-Bibliothek von OpenAI bereitstellen kann.

Dieses Buch hat mich so inspiriert, dass ich versucht war, dieses Vorwort mit ChatGPT zu schreiben, um all das zu demonstrieren, was ich gelernt habe. Dies zeigt, wie gut geschrieben, ansprechend und informativ das Buch ist. Auch wenn ich dazu in der Lage gewesen wäre, habe ich dieses Vorwort doch selbst geschrieben, um meine Gedanken und Erfahrungen über LLMs auf die authentischste und persönlichste Art und Weise zu formulieren, die ich kenne. Mit Ausnahme des letzten Teils des letzten Satzes, der von ChatGPT stammt, einfach weil ich es konnte.

Für jemanden, der mehr über die vielen Aspekte von LLMs lernen möchte, ist dies das richtige Buch. Es wird Ihnen helfen, die Modelle zu verstehen und sie in Ihrem täglichen Leben effektiv zu nutzen. Und was vielleicht am wichtigsten ist: Sie werden diese Reise genießen.

– Jared Lander, Editor der Reihe bei Addison-Wesley

Einleitung

Hallo! Mein Name ist Sinan Ozdemir. In bin ein ehemaliger theoretischer Mathematiker, der zum Universitätsdozenten wurde, dann zum KI-Enthusiasten, zum erfolgreichen Start-up-Gründer, zum KI-Lehrbuchautor und zum Berater für Risikokapitalgeber. Heute bin ich auch Ihr Reiseleiter durch das riesige Museum des Wissens, das die Entwicklung von *Large Language Models* (LLMs), also großen Sprachmodellen, und deren Anwendungen darstellt. Mit diesem Buch verfolge ich zwei Ziele: das Gebiet der LLMs zu entmystifizieren und Sie mit praktischem Wissen auszustatten, damit Sie in der Lage sind, mit LLMs zu experimentieren, zu programmieren und zu bauen.

Aber dies ist kein Schulungsraum, und ich bin kein typischer Professor. Ich bin nicht hier, um Sie mit komplizierter Terminologie zu überschütten. Vielmehr möchte ich komplexe Konzepte leicht verdaulich, nachvollziehbar und – was noch wichtiger ist – anwendbar machen.

Aber jetzt genug von mir. Dieses Buch ist nicht für mich – es ist für Sie. Ich möchte Ihnen einige Tipps dazu geben, wie Sie dieses Buch lesen können, wie Sie dieses Buch noch einmal lesen können (wenn ich meine Arbeit richtig gemacht habe) und wie Sie sicherstellen können, dass Sie alles, was Sie brauchen, aus diesem Text herausholen.

Leserkreis und Voraussetzungen

Für wen ist dieses Buch gedacht, werden Sie fragen. Nun, meine Antwort ist einfach: für jeden, der neugierig auf LLMs ist, den ehrgeizigen Programmierer, die unermüdlich Lernende. Ganz gleich, ob Sie sich bereits mit maschinellem Lernen (Machine Learning) beschäftigt haben oder erst am Rand stehen und Ihre Zehenspitzen in diesen riesigen Ozean tauchen, dieses Buch ist Ihr Leitfaden, Ihre Landkarte, um in den Gewässern der LLMs zu navigieren.

Aber ich will ehrlich zu Ihnen sein: Um das meiste aus dieser Reise herauszuholen, ist eine gewisse Erfahrung mit Machine Learning und Python von unschätzbarem Vorteil. Das heißt nicht, dass Sie ohne diese Kenntnisse nicht überleben werden, aber ohne diese Werkzeuge könnten die Gewässer ein wenig unruhig erscheinen.

Wenn Sie unterwegs lernen, ist das aber auch prima! Einige der Konzepte, die wir erforschen werden, erfordern nicht unbedingt eine umfangreiche Programmierung, die meisten jedoch schon.

Ich habe auch versucht, in diesem Buch ein Gleichgewicht zwischen tiefem theoretischem Verständnis und praktischen Fertigkeiten herzustellen. Jedes Kapitel ist mit Analogien gefüllt, um das Komplexe einfach zu machen, gefolgt von Codeauszügen, die die Konzepte zum Leben erwecken. Im Wesentlichen habe ich dieses Buch als Ihr LLM-Dozent und Tutor geschrieben, um dieses faszinierende Gebiet zu entwirren und zu vereinfachen, anstatt Sie mit akademischem Fachjargon zu überhäufen. Ich möchte, dass Sie aus jedem Kapitel mit einem klareren Verständnis des Themas und dem Wissen, wie es in der Praxis anzuwenden ist, herausgehen.

Wie man an dieses Buch herangeht

Wie eben erwähnt, werden Sie einen leichteren Zugang zu diesem Buch haben, wenn Sie bereits Erfahrung in Machine Learning mitbringen, als wenn Sie komplett bei null anfangen. Dennoch steht der Weg offen für jeden, der in Python programmieren kann und bereit ist zu lernen. Dieses Buch ermöglicht verschiedene Stufen der Beteiligung, je nach Ihrem Hintergrund, Ihren Zielen und Ihrer verfügbaren Zeit. So können Sie tief in die praktischen Abschnitte eintauchen, mit dem Code experimentieren und die Modelle optimieren, oder Sie beschäftigen sich mit den theoretischen Teilen und eignen sich ein solides Verständnis von der Funktionsweise der LLMs an, ohne eine einzige Zeile Code zu schreiben. Sie haben die Wahl.

Wenn Sie das Buch durcharbeiten, sollten Sie daran denken, dass jedes Kapitel in der Regel auf vorherigen Arbeiten aufbaut. Die Kenntnisse und Fertigkeiten, die Sie in einem Abschnitt erwerben, werden in den nachfolgenden Kapiteln zu wertvollen Werkzeugen. Die Herausforderungen, denen Sie sich stellen müssen, sind Teil des Lernprozesses. Es kann sein, dass Sie manchmal etwas durcheinanderkommen, frustriert sind und vielleicht auch gar nicht weiterkommen. Als ich das visuelle Frage-Antwort-System (*Visual Question-Answering*, VQA) für dieses Buch entwickelte, hatte ich wiederholt mit Fehlschlägen zu kämpfen. Das Modell hat nur Unsinn ausgespuckt, immer wieder die gleichen Phrasen. Aber dann, nach unzähligen Wiederholungen, begann es, sinnvolle Ergebnisse zu erzeugen. Dieser Moment des Triumphs, das Hochgefühl, einen Durchbruch erzielt zu haben, war jeden Fehlversuch wert. Dieses Buch bietet Ihnen ähnliche Herausforderungen und folglich auch die Chance auf ähnliche Triumphe.

Aufbau dieses Buchs

Das Buch umfasst vier Teile.

Teil I: Einführung in Large Language Models

Die Kapitel in Teil I bieten eine Einführung in LLMs (Large Language Models) oder mit großen Datenmengen trainierte Sprachmodelle.

- **Kapitel 1: Überblick über Large Language Models**

  Dieses Kapitel bietet einen breiten Überblick über die Welt von LLMs. Es behandelt die Grundlagen: Was sind sie, wie funktionieren sie, und warum sind sie wichtig? Am Ende dieses Kapitel besitzen Sie solide Grundkenntnisse, um den Rest des Buchs zu verstehen.

- **Kapitel 2: Semantische Suche mit LLMs**

  Aufbauend auf den in Kapitel 1 gelegten Grundlagen, untersucht Kapitel 2, wie sich LLMs für eine der einflussreichsten Anwendungen der Sprachmodelle einsetzen lassen – die semantische Suche. Wir erstellen ein Suchsystem, das die Bedeutung Ihrer Abfrage versteht und nicht nur Schlüsselwörter vergleicht.

- **Kapitel 3: Erstes Prompt Engineering und ein Chatbot mit ChatGPT**

  Die Kunst und Wissenschaft, effektive Prompts zu erstellen, ist entscheidend, um die Vorzüge von LLMs nutzen zu können. Kapitel 3 bietet eine praktische Einführung in das Prompt Engineering mit Richtlinien und Techniken, um das Beste aus Ihren LLMs herauszuholen. Zum Schluss erstellen wir einen Chatbot, der auf ChatGPT aufsetzt und die API nutzt, die wir in Kapitel 2 aufgebaut haben.

Teil II: Das Beste aus LLMs herausholen

In Teil II erklimmen Sie die nächste Ebene.

- **Kapitel 4: LLMs mit individuellem Feintuning optimieren**

  In der Welt der LLMs gibt es keine Einheitslösung. Kapitel 4 erläutert, wie Sie LLMs mit Ihren eigenen Datensets feintunen können. Anhand von praktischen Beispielen und Übungen lernen Sie, wie Sie Ihre Modelle im Handumdrehen anpassen.

- **Kapitel 5: Fortgeschrittenes Prompt Engineering**

  Jetzt tauchen wir tiefer in die Welt des Prompt Engineering ein. Kapitel 5 befasst sich mit fortgeschrittenen Strategien und Techniken, die Ihnen helfen, noch mehr aus Ihren LLMs herauszuholen – zum Beispiel Validierung der Ausgabe und semantisches Few-Shot-Learning.

- **Kapitel 6: Embeddings und Modellarchitekturen anpassen**

  In Kapitel 6 erkunden wir die eher technische Seite von LLMs. Wir zeigen, wie man Modellarchitekturen und Embeddings modifiziert, um sie besser auf die eigenen spezifischen Anwendungsfälle und Anforderungen abzustimmen. Außer-

dem passen wir LLM-Architekturen an unsere Bedürfnisse an und führen ein Feintuning an einer Empfehlungsengine durch, die die Modelle von OpenAI übertrifft.

Teil III: Fortgeschrittene LLM-Nutzung

- **Kapitel 7: Jenseits der Basismodelle: LLMs kombinieren**

  Kapitel 7 untersucht einige der Modelle und Architekturen der nächsten Generation, die die Grenzen dessen verschieben, was mit LLMs möglich ist. Wir kombinieren mehrere LLMs und richten ein Framework ein, damit Sie Ihre eigenen LLM-Architekturen mit PyTorch aufbauen können. Außerdem stellt dieses *Reinforcement Learning* (bestärkendes Lernen) aus Rückkopplungen vor, um LLMs auf Ihre Bedürfnisse auszurichten.

- **Kapitel 8: Feintuning fortgeschrittener Open-Source-LLMs**

  In Fortsetzung von Kapitel 7 bietet Kapitel 8 praktische Richtlinien und Beispiele für das Feintuning fortgeschrittener Open-Source-LLMs, wobei der Schwerpunkt auf der praktischen Umsetzung liegt. Wir werden LLMs nicht nur mithilfe von generischer Sprachmodellierung feintunen, sondern auch mit fortgeschrittenen Methoden wie Reinforcement Learning aus Rückkopplungen, um unsere eigenes auf Anweisungen ausgerichtetes LLM namens SAWYER zu kreieren.

- **Kapitel 9: LLMs in die Produktion überführen**

  Dieses letzte Kapitel fasst alles zusammen, indem es die praktischen Überlegungen zur Bereitstellung von LLMs in Produktionsumgebungen untersucht. Unter anderem geht es darum, wie man Modelle skaliert, Echtzeitanfragen verarbeitet und sicherstellt, dass unsere Modelle robust und zuverlässig sind.

Teil IV: Anhänge

Die drei Anhänge enthalten eine Liste mit häufig gestellten Fragen (FAQs), ein Glossar mit Fachbegriffen und eine Referenz auf Archetypen von LLM-Anwendungen.

- **Anhang A: LLM-FAQs**

  Als Berater, Ingenieur und Dozent erhalte ich täglich eine Menge von Fragen zu LLMs. Einige der wichtigsten Fragen habe ich hier zusammengestellt.

- **Anhang B: LLM-Glossar**

  Das Glossar bietet einen Überblick über einige der wichtigsten Begriffe, die in diesem Buch verwendet werden.

- **Anhang C: Archetypen von LLM-Anwendungen**

  In diesem Buch erstellen wir viele Anwendungen mit LLMs, sodass Anhang C als Ausgangspunkt für jeden gedacht ist, der eine eigene Anwendung bauen möchte. Für einige häufige Anwendungen von LLMs schlägt dieser Anhang vor, auf welche LLMs Sie sich konzentrieren sollten und welche Daten Sie möglicherweise benötigen. Und Sie erfahren ebenfalls, auf welche häufig vorkommenden Fallstricke Sie eventuell stoßen und wie Sie mit ihnen umgehen können.

Was unterscheidet dieses Buch von anderen?

Zunächst einmal habe ich eine Vielzahl von Erfahrungen in dieses Werk einfließen lassen: von meinem Hintergrund in theoretischer Mathematik über meinen Einstieg in die Welt der Start-ups und meine Erfahrungen als ehemaliger Hochschullehrer bis hin zu meinen derzeitigen Rollen als Unternehmer, Machine Learning Engineer und Risikokapitalberater. Jede dieser Erfahrungen hat mein Verständnis von LLMs geprägt, und ich habe all mein Wissen in dieses Buch einfließen lassen.

Eine der Besonderheiten, die Sie in diesem Buch finden, ist die praktische Anwendung von Konzepten. Und ich meine es ernst, wenn ich »praktisch« sage. Dieses Buch ist voll von praktischen Erfahrungen, die Ihnen helfen werden, die Realität der Arbeit mit LLMs zu verstehen.

Darüber hinaus geht es in diesem Buch nicht nur darum, das Gebiet zu verstehen, wie es sich heute darstellt. Wie bereits häufig gesagt: Die Welt der LLMs ändert sich stündlich. Dennoch bleiben einige Grundlagen konstant, und ich lege großen Wert darauf, diese im gesamten Buch hervorzuheben. Auf diese Weise sind Sie nicht nur für das Hier und Jetzt, sondern auch für die Zukunft gerüstet.

Im Wesentlichen spiegelt dieses Buch nicht nur mein Wissen wider, sondern auch meine Leidenschaft für die Entwicklung von KI und LLMs. Es ist eine Destillation (Wortspiel beabsichtigt – siehe Kapitel 8) meiner Erfahrungen, meiner Einsichten und meiner Begeisterung für die Möglichkeiten, die LLMs uns eröffnen. Es ist eine Einladung an Sie, gemeinsam mit mir dieses faszinierende, sich schnell entwickelnde Gebiet zu erforschen.

Codebeispiele

Zusätzliches Material (Codebeispiele in Jupyter Notebooks, Daten und Abbildungen) finden Sie zum Herunterladen unter *https://github.com/sinanuozdemir/quick start-guide-to-llms*.

Dieses Buch soll Ihnen bei Ihrer Arbeit helfen. Ganz allgemein gilt: Wenn in diesem Buch Beispielcode angeboten wird, können Sie ihn in Ihren Programmen und Dokumentationen verwenden. Sie müssen sich dafür nicht unsere Erlaubnis einholen, es sei denn, Sie reproduzieren einen großen Teil des Codes. Schreiben Sie zum Beispiel ein Programm, das mehrere Teile des Codes aus diesem Buch benutzt, brauchen Sie keine Erlaubnis. Verkaufen oder vertreiben Sie Beispiele aus O'Reilly-Büchern, brauchen Sie eine Erlaubnis. Beantworten Sie eine Frage, indem Sie dieses Buch und Beispielcode daraus zitieren, brauchen Sie keine Erlaubnis. Binden Sie einen großen Anteil des Beispielcodes aus diesem Buch in die Dokumentation Ihres Produkts ein, brauchen Sie eine Erlaubnis.

Wir freuen uns über eine Erwähnung, verlangen sie aber nicht. Eine Erwähnung enthält üblicherweise Titel, Autor, Verlag und ISBN, zum Beispiel: »*Praxiseinstieg Large Language Models* von Sinan Ozdemir, O'Reilly 2024, ISBN 978-3-96009-240-7.«

Falls Sie befürchten, zu viele Codebeispiele zu verwenden oder die oben genannten Befugnisse zu überschreiten, kontaktieren Sie uns unter *kommentar@oreilly.de*.

In diesem Buch verwendete Konventionen

Die folgenden typografischen Konventionen kommen in diesem Buch zum Einsatz:

Kursiv
: Steht für neue Begriffe, URLs, E-Mail-Adressen, Dateinamen und Dateierweiterungen.

`Nichtproportionalschrift`
: Wird für Programmlistings verwendet, aber auch innerhalb von Absätzen, um sich auf Programmelemente wie Variablen oder Funktionsnamen, Datenbanken, Datentypen, Umgebungsvariablen, Anweisungen und Schlüsselwörter zu beziehen.

`Fette Nichtproportionalschrift`
: Steht für Befehle oder anderen Text, der genau so einzugeben ist.

`Kursive Nichtproportionalschrift`
: Steht für Text, der von den Benutzerinnen und Benutzern durch Werte ersetzt werden soll, die sich eventuell aus dem Kontext ergeben.

Dieses Element enthält einen allgemeinen Hinweis.

Zusammenfassung

Damit sind wir nun am Ende des Vorworts oder am Beginn unserer gemeinsamen Reise angekommen, je nachdem, wie Sie es betrachten. Sie haben einen Eindruck davon bekommen, wer ich bin, warum es dieses Buch gibt, was Sie erwarten können und wie Sie das Beste aus ihm herausholen können.

Jetzt liegt der Rest bei Ihnen. Ich lade Sie dazu ein, in die Welt der LLMs einzutauchen. Ob Sie nun ein erfahrener Data Scientist oder eine neugierige Enthusiastin sind – es ist mit Sicherheit etwas für Sie dabei. Ich möchte Sie ermutigen, sich aktiv mit dem Buch zu beschäftigen – den Code auszuführen, ihn zu optimieren, ihn zu zerstören und wieder zusammenzusetzen. Erkunden Sie, experimentieren Sie, machen Sie Fehler, lernen Sie.

Lassen Sie uns eintauchen!

Danksagung

Familie: An meine unmittelbaren Familienmitglieder: Danke, Mom, dass du immer wieder die Kraft und den Einfluss des Lehrens verkörpert hast. Es war deine Leidenschaft für Bildung, die mich den tiefen Wert der Weitergabe von Wissen erkennen ließ, was ich nun in meiner Arbeit umzusetzen versuche. Dad, dein lebhaftes Interesse an neuen Technologien und ihrem Potenzial hat mich immer dazu inspiriert, die Grenzen auf meinem eigenen Gebiet zu erweitern. Meine Schwester, deine ständigen Ermahnungen, die menschlichen Auswirkungen meiner Arbeit zu berücksichtigen, haben mich auf dem Boden der Tatsachen gehalten. Deine Einsichten haben mir bewusster gemacht, auf welche Weise meine Arbeit das Leben der Menschen berührt.

Zuhause: An meine Lebensgefährtin Elizabeth: Deine Geduld und dein Verständnis waren von unschätzbarem Wert, als ich mich in unzähligen Nächten ins Schreiben und Programmieren vertieft habe. Danke, dass du mein Geschwafel ertragen und mir geholfen hast, komplexen Ideen einen Sinn zu verleihen. Du warst eine Stütze, ein Resonanzboden und ein Leuchtturm, wenn der Weg unklar erschien. Deine Standhaftigkeit während dieser Reise hat mich inspiriert, und ohne dich wäre dieses Werk nicht das, was es ist.

Prozess der Buchveröffentlichung: Ein herzliches Dankeschön an Debra Williams Cauley, die mir die Möglichkeit gegeben hat, einen Beitrag zur KI- und LLM-Community zu leisten. Das Wachstum, das ich als Pädagoge und Autor während dieses Prozesses erfahren habe, ist unermesslich. Ich entschuldige mich zutiefst für die wenigen (oder doch mehr) Abgabetermine, die ich verpasst habe, weil ich mich in den Feinheiten der LLMs und des Feintunings verloren hatte. Ich schulde auch Jon Krohn Dank dafür, dass er mich für diese Reise empfohlen hat, und für seine kontinuierliche Unterstützung.

TEIL I

Einführung in Large Language Models

KAPITEL 1

Überblick über Large Language Models

Im Jahr 2017 stellte ein Team von Google Brain ein fortschrittliches Deep-Learning-Modell für *künstliche Intelligenz* (KI) namens Transformer vor. Seitdem ist der Transformer zum Standard geworden, um verschiedenste Aufgaben bei der Verarbeitung natürlicher Sprache (*Natural Language Processing*, NLP) in Wissenschaft und Industrie zu bewältigen. Höchstwahrscheinlich haben Sie in den letzten Jahren bereits mit dem Transformer-Modell interagiert, ohne sich dessen bewusst zu sein, denn Google verwendet BERT, um seine Suchmaschine zu verbessern, indem es die Suchanfragen der Nutzer besser versteht. Die Modelle der GPT-Familie von OpenAI haben ebenfalls Aufmerksamkeit erregt, da sie in der Lage sind, wie von Menschen geschaffene Texte und Bilder zu erzeugen.

Diese Transformer treiben nun Anwendungen voran wie etwa Copilot von GitHub (eine Entwicklung von OpenAI in Zusammenarbeit mit Microsoft), der es ermöglicht, Kommentare und Codefragmente in voll funktionsfähigen Quellcode umzuwandeln, der sogar andere große Sprachmodelle (*Large Language Models*, LLMs) aufrufen kann, um NLP-Aufgaben zu erfüllen (siehe Beispiel 1-1).

Beispiel 1-1: Mithilfe des Copilot-LLM eine Ausgabe vom BART-LLM von Facebook erhalten

```
from transformers import pipeline
def classify_text(email):
    """
    Use Facebook's BART model to classify an email into "spam" or "not spam"
    Args:
        email (str): The email to classify
    Returns:
        str: The classification of the email
    """
 # COPILOT START. EVERYTHING BEFORE THIS COMMENT WAS INPUT TO COPILOT
    classifier = pipeline(
        'zero-shot-classification', model='facebook/bart-large-mnli')
    labels = ['spam', 'not spam']
    hypothesis_template = 'This email is {}.'
    results = classifier(
        email, labels, hypothesis_template=hypothesis_template)
    return results['labels'][0]
    # COPILOT END
```

In Beispiel 1-1 habe ich Copilot verwendet, um nur eine Python-Funktionsdefinition und einige von mir verfasste Kommentare zu übernehmen. Und ich habe den ganzen Code geschrieben, damit die Funktion das tut, was ich geschrieben habe. Hier gibt es kein Rosinenpicken, sondern nur eine voll funktionsfähige Python-Funktion, die ich wie folgt aufrufen kann:

```
classify_text('hi I am spam')  # spam
```

Es scheint, dass wir von LLMs umgeben sind, aber was machen sie hinter den Kulissen? Finden wir es heraus!

Was sind Large Language Models?

Large Language Models (*LLMs*, große Sprachmodelle) sind KI-Modelle, die in der Regel (aber nicht unbedingt) von der Transformer-Architektur abgeleitet sind und dazu dienen, menschliche Sprache, Code und vieles mehr zu verstehen und zu erzeugen. Diese Modelle werden anhand großer Mengen von Textdaten trainiert, sodass sie die Komplexität und die Nuancen menschlicher Sprache erfassen können. LLMs können ein breites Spektrum sprachbezogener Aufgaben erfüllen – von der einfachen Textklassifizierung bis hin zur Texterzeugung –, und das mit hoher Genauigkeit, Geläufigkeit und mit Stil.

Im Gesundheitswesen nutzt man LLMs, um elektronische Krankenakten (*Electronic Medical Record*, EMR) zu verarbeiten, klinische Studien abzugleichen und Medikamente zu entwickeln. Im Finanzwesen setzt man sie bei der Betrugserkennung, zur Stimmungsanalyse von Finanznachrichten und sogar für Handelsstrategien ein. Außerdem werden LLMs zur Automatisierung des Kundendiensts durch Chatbots und virtuelle Assistenten herangezogen. Aufgrund ihrer Vielseitigkeit und hohen Leistungsfähigkeit werden auf Transformer basierende LLMs in einer Vielzahl von Branchen und Anwendungen immer wertvoller.

In diesem Text werde ich den Begriff »Verstehen« recht häufig verwenden. In diesem Zusammenhang beziehe ich mich in der Regel auf das Verstehen natürlicher Sprache (*Natural Language Understanding*, NLU) – einen Forschungszweig des NLP, der sich mit der Entwicklung von Algorithmen und Modellen befasst, die menschliche Sprache genau interpretieren können. Wie wir sehen werden, glänzen NLU-Modelle bei Aufgaben wie Klassifizierung, Stimmungsanalyse und Erkennen benannter Entitäten. Allerdings ist es wichtig, zu beachten, dass diese Modelle zwar komplexe Sprachaufgaben erfüllen können, nicht aber über ein wirkliches Verständnis verfügen, wie Menschen es haben.

Der Erfolg der LLMs und Transformer ist auf die Kombination mehrerer Ideen zurückzuführen. Die meisten dieser Ideen gab es schon vor Jahren, und sie wurden auch zur etwa gleichen Zeit aktiv erforscht. Mechanismen wie Attention (Aufmerksamkeit), Transfer Learning und das Heraufskalieren neuronaler Netze, die das Gerüst für Transformer bilden, erlebten etwa zur gleichen Zeit einen Durchbruch. Abbildung 1-1 skizziert einige der größten Fortschritte im NLP der letzten Jahrzehnte, die alle zur Erfindung des Transformers führten.

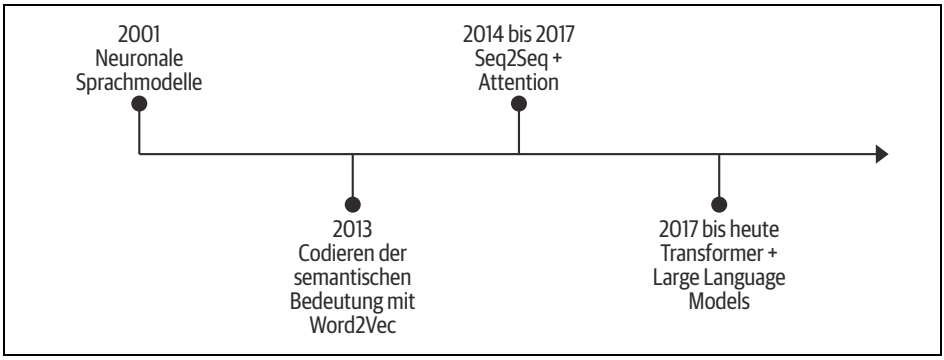


Abbildung 1-1: Ein kurzer geschichtlicher Abriss des modernen NLP verdeutlicht die Verwendung von Deep Learning für die Sprachmodellierung, Fortschritte bei groß angelegten semantischen Token-Embeddings (Word2vec), Sequenz-zu-Sequenz-Modelle mit Attention (worauf wir später in diesem Kapitel ausführlich zurückkommen) und schließlich den Transformer im Jahr 2017.

Die Transformer-Architektur selbst ist ziemlich beeindruckend. Sie lässt sich hochgradig parallelisieren und in einer Weise skalieren, wie es vorhergehenden NLP-Modellen nach dem jeweiligen Stand der Technik nicht möglich war. Somit können auch wesentlich größere Datensets verarbeitet und längere Trainings absolviert werden, als es mit älteren NLP-Modellen realisierbar war. Der Transformer verwendet eine spezielle Art der Attention-Berechnung, die sogenannte *Self-Attention* (Selbstaufmerksamkeit), die es jedem Wort in einer Sequenz erlaubt, alle anderen Wörter in der Sequenz »zu beachten« (nach dem Kontext zu suchen), sodass man weitreichende Abhängigkeiten und kontextuelle Beziehungen zwischen Wörtern erfassen kann. Natürlich ist keine Architektur perfekt. Transformer sind immer noch auf ein Eingabefenster beschränkt, das die maximale Länge des Texts darstellt, den sie zu einem bestimmten Zeitpunkt verarbeiten können.

Seit die Transformer-Architektur im Jahr 2017 eingeführt wurde, ist das Ökosystem rund um die Nutzung von Transformern regelrecht explodiert. Die treffend benannte »Transformers«-Bibliothek und ihre unterstützenden Pakete haben es Praktikern ermöglicht, Modelle zu verwenden, zu trainieren und zu teilen, was die Akzeptanz dieses Modells erheblich beschleunigt hat, sodass es jetzt von Tausenden Organisationen (Tendenz steigend) eingesetzt wird. Beliebte LLM-Repositorys wie Hugging Face sind auf der Bildfläche erschienen und bieten Zugang zu leistungsstarken Open-Source-Modellen für eine breite Nutzerschaft. Kurz gesagt, die Verwendung und das Erzeugen eines Transformers war noch nie so einfach.

Und genau hier kommt dieses Buch ins Spiel.

Ich möchte Ihnen zeigen, wie man alle Arten von LLMs für praktische Anwendungen einsetzt, trainiert und optimiert, wobei Sie genügend Einblicke in die inneren Abläufe des Modells erhalten, damit Sie optimale Entscheidungen in Bezug auf Modellauswahl, Datenformat, Parameter zum Feintuning und vieles mehr treffen können.