Olaf Kopp

WIE SCHAFFT MAN ES IN DIE ANTWORTEN VON KI-SYSTEMEN WIE GOOGLES AI OVERVIEWS, CHATGPT ODER PERPLEXITY?

In der sich schnell entwickelnden Welt der generativen KI wird es immer wichtiger, in den Ergebnissen von Plattformen wie AI Overviews, ChatGPT und Perplexity sichtbar zu sein. Diese Technologien nutzen generative KI, um Informationen bereitzustellen und Inhalte zu generieren, die auf den Bedürfnissen und Interessen der Nutzer basieren. Um in diesen Ergebnissen präsent zu sein, müssen Unternehmen und Einzelpersonen strategische Ansätze entwickeln, die Suchmaschinenoptimierung, relevante und qualitativ hochwertige Inhalte sowie eine aktive Präsenz in digitalen Medien einschließen. Der Beitrag von Olaf Kopp beleuchtet die wesentlichen Schritte und Überlegungen, um die Sichtbarkeit in der Zukunft der generativen KI zu sichern.

Foto: kemalbas / gettyimages.de

DER AUTOR

Olaf Kopp ist Online-Marketing-Experte mit mehr als 15 Jahren Erfahrung in Google Ads, SEO und Content-Marketing. Olaf ist Co-Founder, Chief Business Development Officer (CBDO) und Head of SEO & Content bei der Online-Marketing-Agentur Aufgesang GmbH.

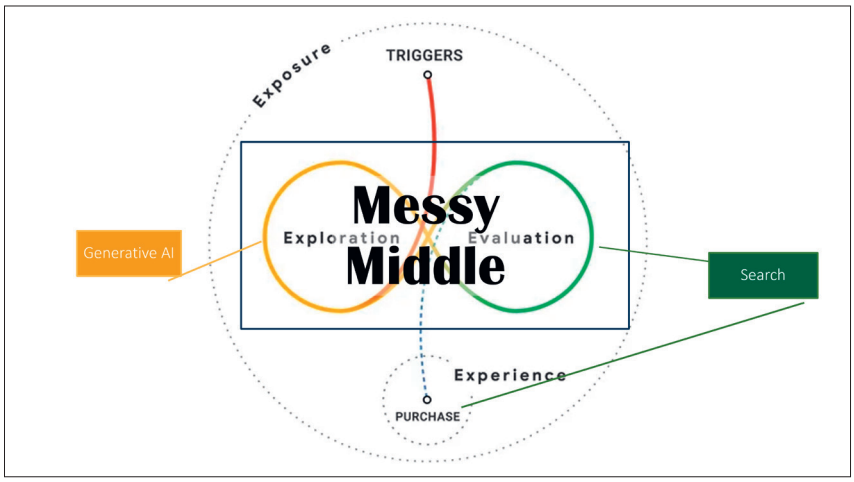


Abb. 1: Nutzung generativer KI im Rechercheprozess, *© Olaf Kopp, Aufgesang GmbH*

Eine neue Sau wird durch das Online-Marketing-Dorf getrieben! Die Sau hat viele Namen: Large Language Model Optimization (LLMO), Generative Engine Optimization (GEO), Generative AI Optimization (GAIO). Es geht dabei um die Optimierung von Ergebnissen aus Anwendungen, die auf generativer KI basieren. Das Ziel dabei ist, mit den eigenen Produkten, Marken oder Website-Inhalten in den AI-Ergebnissen aufzutauchen. Ich werde es in diesem Beitrag LLMO nennen.

Plattformen wie ChatGPT, Google AI Overviews, Microsoft Copilot und Perplexity verändern nicht nur die Art und Weise, wie Nutzer Informationen suchen und konsumieren, sondern auch, wie Unternehmen und Marken in diesen KI-generierten Inhalten sichtbar werden können.

Dieser Beitrag beleuchtet Ansätze und Strategien, die helfen, um von generativen KI-Modellen aufgegriffen zu werden.

Ein Disclaimer vorweg. Es gibt bisher keine Methoden, die sich in der Praxis bewährt haben. Dazu ist dieses Feld zu neu und es erinnert etwas an die frühen Tage der SEO, in denen keine Ranking-Faktoren der Suchmaschinen bekannt waren und man sich das Thema über Testing, Recherche und technologisches Verständnis von Information Retrieval und Suchmaschinen in Pionierarbeit erschließen musste.

Dazu bedarf es eines genaueren Blicks auf die Funktionsweise von Natural Language Processing und Large Language Models (LLMs). Ein technologisches Verständnis dieser Themenfelder ist in dieser frühen Phase unabdingbar, um Potenziale für die Zukunft von SEO, digitalem Markenaufbau und Content-Strategien abzuleiten.

Alle hier vorgestellten Ansätze basieren auf eigener Recherche von wissenschaftlichen Dokumenten, Patenten zu generativer KI und meiner über zehnjährigen Arbeit rund um die semantische Suche.

Wie funktioniert generative KI?

Bevor man sich mit LLMO beschäftigt, empfehle ich jedem, sich ein Grundverständnis zur Technologie hinter LLMs anzueignen. Ähnlich wie bei Suchmaschinen verhindert ein Verständnis der Technologie das Hinterherlaufen hinter vermeintlichen Hacks und falschen Empfehlungen. Lieber ein paar Stunden mehr in dieses Verständnis investieren, als unnötige Maßnahmen umzusetzen, die sich dann als Sackgassen darstellen und Ressourcen unnötig verbrauchen.

Large Language Models (LLMs) wie die GPT-Modelle, Claude oder LLaMA stellen einen bedeutenden Fortschritt in der Suchtechnologie und generativen KI dar und verändern die Art und Weise, wie Suchmaschinen und AI-Assistenten Anfragen und Aufgaben verarbeiten und beantworten, grundlegend.

Laut verschiedenen Forschungen wie zum Beispiel dem Microsoft-Forschungsbericht *„Large Search Model: Redefining Search Stack in the Era of*

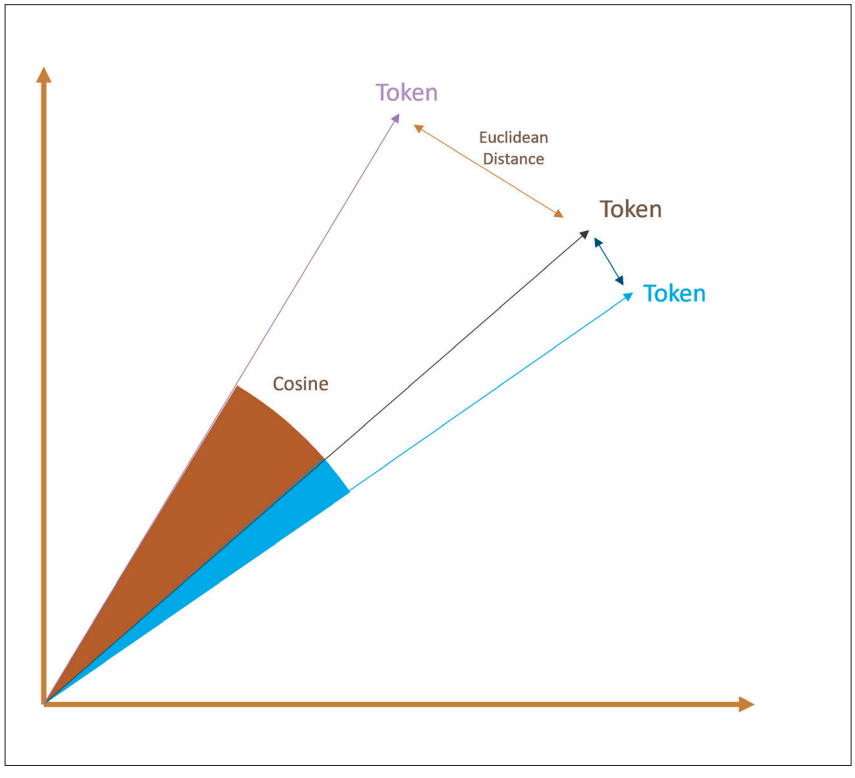


Abb. 2: Embeddings im Vektorraum, *© Olaf Kopp, Aufgesang GmbH*

LLMs“ zeigen diese Modelle bemerkenswerte Fähigkeiten im Bereich des Sprachverständnisses und der Argumentation, die über die einfache Textübereinstimmung hinausgehen und nuancierte und kontextbezogene Antworten liefern.

Die Kernfunktionalität von LLMs bei der Suche besteht darin, Anfragen zu verarbeiten und Zusammenfassungen in natürlicher Sprache zu erstellen. Anstatt nur Informationen aus vorhandenen Dokumenten zu extrahieren, können diese Modelle umfassende Antworten generieren und dabei Genauigkeit und Relevanz beibehalten. Dies wird durch ein einheitliches Framework erreicht, das alle (suchbezogenen) Aufgaben als Probleme der Textgenerierung behandelt.

Was diesen Ansatz besonders leistungsstark macht, ist seine Fähigkeit, Antworten durch Eingabeaufforderungen in natürlicher Sprache anzupassen. Das System generiert zunächst eine erste Reihe von Abfrageergebnissen, die dann vom LLM verfeinert und verbessert werden. Wenn zusätzliche Informationen benötigt werden, kann das LLM ergänzende Abfragen generieren, um umfassendere Daten zu sammeln.

Ohne zu tief ins Detail zu gehen, soll hier kurz der Encoding- und Decoding-Prozess erklärt werden.

Das Encoding ist für das Prozessieren und „Verstehen“ der Trainingsdaten zuständig. Von einem wirklichen Verständnis kann man hier allerdings nicht ausgehen. LLMs besitzen kein Wissen oder Verständnis im menschlichen Sinn!

Die Daten werden extrahiert und via Chunking in Tokens eingeteilt.

Tokens sind der elementare Bestandteil von Sprachmodellen. Dabei können Tokens je nach Anwendungsfall Wörter, N-Gramme, Entitäten, Bilder, Videos oder ganze Dokumente sein.

Im nächsten Schritt werden die Tokens in Vektoren transformiert. Dieser Vorgang war die Grundlage für die von Google entwickelte Transformer-Technologie und die darauf basierenden Sprachmodelle. Es war der Gamechanger in der KI und ist grundlegend dafür verantwortlich, dass KI-Modelle heute die Welt erobern.

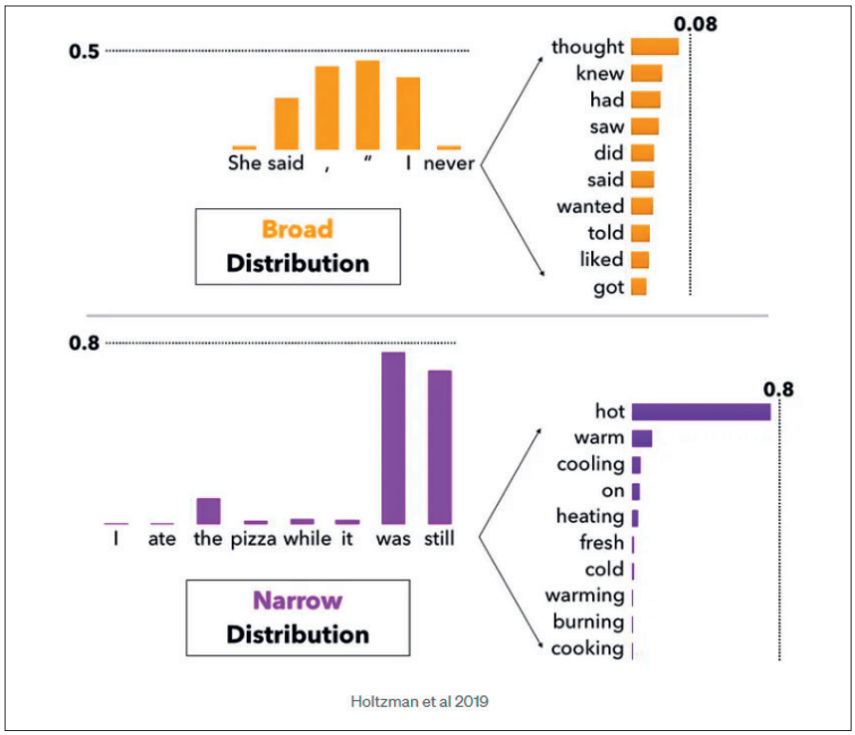


Abb. 3: Decoding-Beispiel über K-Sampling, *Quelle: https://arxiv.org/abs/1904.09751*

Vektoren sind in Nummern transformierte Tokens. Die Nummern stehen für bestimmte Attribute, die die Eigenschaften des Tokens beschreiben. Über diese Eigenschaften lassen sich die Vektoren in semantische Räume einordnen und Beziehungen zu anderen Vektoren herstellen. Hier spricht man auch von Embeddings.

Über verschiedene Verfahren wie zum Beispiel das Cosinus-Maß oder die Euclidean Distance lässt sich die semantische Ähnlichkeit, aber auch Beziehung zwischen Vektoren feststellen.

Beim Decoding geht es darum, die Wahrscheinlichkeiten zu interpretieren, die das Modell für jedes mögliche nächste Token (Wort oder Symbol) berechnet. Das Ziel ist es, die sinnvollste oder natürlichste Sequenz zu erzeugen. Es gibt verschiedene Methoden, die beim Decoding angewendet werden können, wie zum Beispiel das Top-K-Sampling oder Top-P-Sampling.

Hierbei werden potenziell nachfolgende Wörter mit einem Wahrscheinlichkeits-Score bewertet. Je nachdem wie hoch der „Kreativitätsspielraum“ des Modells ist, werden die Top-K-Wörter als mögliches nächstes Wort in Betracht gezogen. Modelle mit einer breiteren Auslegung können neben der Top-One-Wahrscheinlichkeit auch die nachfolgenden Wörter berücksichtigen und damit kreativer in der Ausgabe sein.

Das erklärt auch mögliche unterschiedliche Ergebnisse zu demselben Prompt. Bei Modellen, die sehr „streng“ ausgelegt sind, wird man dementsprechend immer ähnliche Ergebnisse bekommen.

Sowohl für das Encoding als auch für das Decoding wird die Methode des Natural Language Processing genutzt. Über Natural Language Processing kann das Kontextfenster größer ausgelegt werden, um auch die grammatikalische Satzstruktur im Natural-Language-Understanding-Prozess zu berücksichtigen. Dadurch lassen sich Haupt und Neben-Entitäten identifizieren.

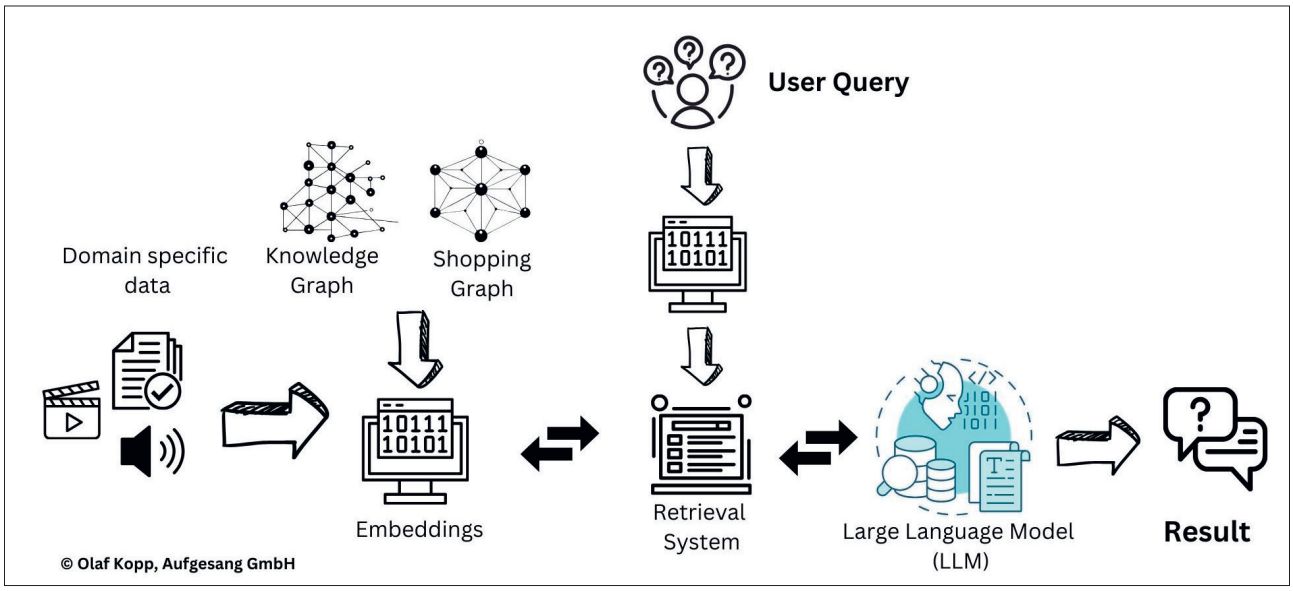


Abb. 4: Retrieval-Augmented-Generation-Prozess

Generative KI ist multimedial und kann neben Text viele andere Formate umfassen. Diese Ausführungen fokussieren sich auf textbasierte generative KI, da diese mit Blick auf LLMO die wichtigste Methode ist. Das beinhaltet auch Audio und teilweise visuelle Formate, die von der KI im Encoding-Prozess zur Weiterverarbeitung auch in Text-Tokens transformiert werden.

Die größte Herausforderung ist die Aktualität der Informationen, die Vermeidung des sogenannten Halluzinierens und die Ausgabe von tiefgehenden Informationen zu bestimmten Themenfeldern. Die grundlegenden LLMs sind meistens mit recht oberflächlichen „Basisinformationen" angelernt, was sie oft bei spezifischen Anfragen sehr generisch und teilweise auch falsch antworten lässt.

Für diese Herausforderung hat sich mit Retrieval Augmented Generation (RAG) eine Methode durchgesetzt, bei der den LLMs zusätzliche themenfeldspezifische Informationen zugeführt werden.

Durch RAG können LLMs die beschriebenen Herausforderungen besser bewältigen.

Diese Ergänzung von themenspezifischen Informationen kann neben Dokumenten auch über Knowledge Graphen beziehungsweise in Vektoren transformierte Entitäten-Knoten erfolgen. Dadurch besteht der Vorteil, dass ontologische Informationen zu den Beziehungen der Entitäten zueinander mitgegeben werden können und man einem wirklichen semantischen Verständnis näher kommt.

Durch RAG ergeben sich mögliche Ansatzpunkte für LLMO. Während bei den initialen Trainingsdaten die Grundlagen für die Ermittlung der Quellen nicht klar zu ermitteln sind beziehungsweise es nicht so einfach möglich ist, diese zu beeinflussen, ist der Ansatzpunkt bei LLMO, die bevorzugten themenspezifischen Quellen zu beeinflussen.

Die große Frage hierbei ist, wie die unterschiedlichen Plattformen diese Quellen auswählen.

Zur Beantwortung dieser Frage hängt es davon ab, ob die Anwendungen Zugriff auf ein Retrieval-System haben, über das die Quellen hinsichtlich Relevanz und Qualität bewertet und ausgewählt werden können.

Retrieval-Modelle fungieren als Informations-Gatekeeper in der RAG-Architektur. Ihre Hauptfunktion besteht darin, einen großen Datenbestand zu durchsuchen, um relevante Informationen zu finden, die für die Textgenerierung verwendet werden können. Diese Modelle kann man sich als spezialisierte Bibliothekare vorstellen, die genau wissen, welche „Bücher" sie zu einem Thema aus den „Regalen" holen müssen.

Diese Modelle verwenden Algorithmen, um die relevantesten Daten zu bewerten und auszuwählen, und bieten so die Möglichkeit, externes Wissen in den Textgenerierungsprozess einzubringen. Auf diese Weise schaffen Abrufmodelle die Voraussetzungen für eine fundiertere, kontextreichere Sprachgenerierung und erweitern die Möglichkeiten herkömmlicher Sprachmodelle.

Diese Retrieval-Systeme können über verschiedene Mechanismen implementiert werden. Eine der gängigsten Techniken ist die Verwendung von Vektor-Embeddings und Vektor-Suche, aber auch Dokumentindex-Datenbanken, die Technologien wie BM25 (Best Match 25) und TF-IDF (Term Frequency – Inverse Document Frequency) nutzen, sind weitverbreitet.

Nicht jedes System hat Zugriff auf solche Retrieval-Systeme, was RAG nur schwer möglich macht. Das könnte auch der Grund sein, warum **Meta** jetzt auch eine eigene Suchmaschine einführen will. Das würde Meta die Möglichkeit geben, RAG für die eigenen LLaMA-Modelle mittels eines eigenen Retrieval-Systems zu nutzen.

Perplexity nutzt laut eigener Aussage einen eigenen Index und Ranking-Systeme. Hier gibt es den Vorwurf, dass Perplexity die Suchergebnisse anderer Suchmaschinen wie Google scrapt beziehungsweise kopiert.

Claude: Ob Claude RAG über den von den Nutzern bereitgestellten Informationen aus einem eigenen Index nutzt, ist unklar.

Gemini, Copilot, ChatGPT: Sowohl Microsoft als auch Google besitzen die Möglichkeit, die eigene Suche als Quelle für RAG beziehungsweise das domänenspezifische Training zu nutzen. ChatGPT hat lange die Bing-Suche genutzt. Mit der Einführung von SearchGPT ist nicht ganz klar, ob ein eigenes Retrieval-System betrieben wird. Laut Aussagen seitens OpenAI wird bei SearchGPT auf eine Mischung aus Suchmaschinentechnologien zurückgegriffen, die auch Microsoft Bing umfasst.

„The search model is a fine-tuned version of GPT-4o, post-trained using novel synthetic data generation techniques, including distilling outputs from OpenAI o1-preview. ChatGPT search leverages third-party search providers, as well as content provided directly by our partners, to provide the information users are looking for." (openai.com/index/introducing-chatgpt-search/)

Der Verge gegenüber wird Bing als einer der Kooperationspartner genannt (*einfach.st/verge53*).

Fragt man ChatGPT nach den besten Laufschuhen, gibt es zwar eine Überlappung zwischen den in den Bing-Suchergebnissen top rankenden Seiten und den für die Antwort genutzten Quellen, aber der Overlap ist deutlich unter 100 %.

Prompt versus Suchanfrage

Ein Prompt ist komplexer und entspricht mehr der natürlichen Sprache als viele Suchanfragen, die meistens als Aneinanderreihung von Schlüsselbegriffen eingegeben werden. Ein Prompt wird häufig durch eine explizite Frage und/oder zusammenhängende Sätze formuliert. Dadurch wird mehr Kontext mitgeliefert und die Antwort kann noch zielgenauer ausgegeben werden.

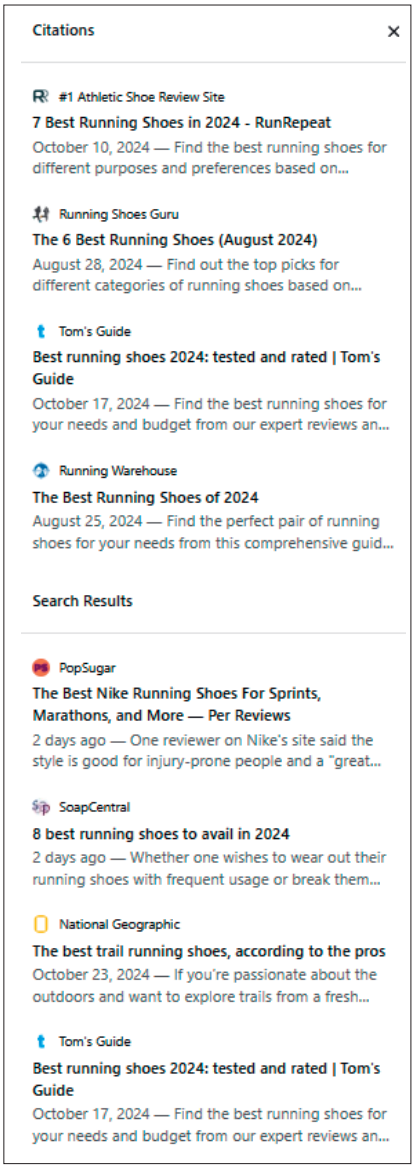


Abb. 5: Screenshot von Quellenangaben in SearchGPT

Wichtig ist, zu verstehen, dass es einen Unterschied macht, ob man für AI Overviews optimieren will oder direkt für die Ergebnisse in KI-Assistenten. AI Overviews sind als SERP-Feature Teil der Google-Suche und werden in der Regel durch Suchanfragen angesprochen, während bei der Eingabe in KI-Assistenten komplexere Prompts in natürlicher Sprache genutzt werden. Das bedarf für den RAG-Prozess, dass der Prompt im Hintergrund in eine Suchanfrage umgeschrieben werden muss, ohne dass wichtiger Kontext verloren geht, damit geeignete Quellen identifiziert werden können.

Was sind Ziele von LLMO?

In den vielen Beiträgen zu LLMO wird oft nicht klar, um welche Ziele es denn eigentlich geht. Die einen sprechen von der Nennung der eigenen Inhalte in den referenzierten Quellen-Links. Bei den anderen geht es um die Erwähnung des eigenen Namens, der Marke oder der Produkte in den Ausgaben von generativer KI.

Beides sind mögliche Ziele, aber für beide Ziele braucht es unterschiedliche Ansätze. Während es beim Ziel, in den Link-Referenzen genannt zu werden, eher darum geht, mit seinem Content referenziert zu werden, geht es bei den Mentions in der KI-Ausgabe um die Erhöhung der Wahrscheinlichkeit, mit den eigenen Personen-, Organisations- oder Produkt-Entitäten in bestimmten Kontexten genannt zu werden.

Ziel sollte es im ersten Schritt auf jeden Fall sein, in den favorisierten Quellen stattzufinden und/oder selbst zu den ausgewählten Quellen zu gehören. Das ist für beide Zielsetzungen Voraussetzung.

Müssen wir alle LLMs fokussieren?

Die unterschiedlichen Ergebnisse der KI-Anwendungen zeigen, dass jede Plattform ihre eigenen Prozesse und Kriterien anwendet, um Empfehlungen für benannte Entitäten auszugeben und Quellen auszuwählen.

Demnach werden wir uns zukünftig wohl mit mehreren Large Language Models beziehungsweise AI-Assistenten und deren Funktionsweise beschäftigen müssen. Für Google-verwöhnte SEO-Experten wird das eine Umstellung sein. In den nächsten Jahren ist daher zu beobachten, in welchen Märkten

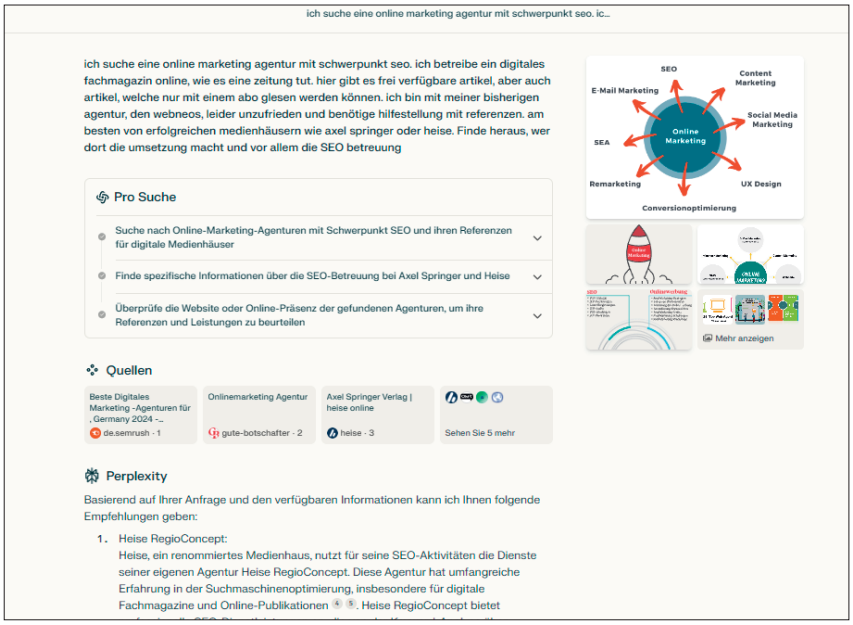


Abb. 6: Screenshot aus Perplexity

und/oder Branchen sich welche Anwendungen durchsetzen werden.

Hierbei ist es entscheidend, herauszufinden, wie die unterschiedlichen Anwendungen die Quellenauswahl durchführen.

Warum werden bestimmte Personen, Marken oder Produkte von der generativen KI genannt?

Man kann ziemlich sicher sein, dass in den nächsten Jahren immer mehr Menschen auch für die Suche nach Produkten und Dienstleistungen in der Recherche KI-Anwendungen nutzen werden.

Hier ein Beispiel aus der Praxis. Ein potenzieller Kunde hat vor ein paar Monaten bei Aufgesang angefragt. Er hatte zuvor den folgenden Prompt bei Perplexity eingegeben:

„ich suche eine online marketing agentur mit schwerpunkt seo. Ich betreibe ein digitales fachmagazin online, wie es eine zeitung tut. hier gibt es frei verfügbare artikel, aber auch artikel, welche nur mit einem abo gelesen werden können. Ich bin mit meiner bisherigen agentur, den .....(die genannte Agentur wurde entfernt), leider unzufrieden und benötige hilfestelllung mit referenzen, am besten von erfolgreichen medienhäusern wie axel springer oder heise. Finde heraus, wer dort die umsetzung macht und vor allem die SEO Betreuung.“

Als Ergebnis hat Perplexity unter anderem Aufgesang empfohlen.

Warum ist das so?

Das Stichwort hier heißt Kookkurrenzen oder Co-Nennungen und Kontext. Je häufiger bestimmte Tokens miteinander genannt werden, desto wahrscheinlicher ist es, dass sie zusammengehören beziehungsweise kontextuell eng miteinander in Verbindung stehen. Anders ausgedrückt: Der Wahrscheinlichkeits-Score beim Decoding steigt. Da Aufgesang eine SEO-Agentur ist und zur heise Gruppe gehört, gibt es in den Trainingsdaten sowohl Kookkurrenzen mit heise als auch SEO.

Ein anderes Beispiel für einen möglichen Prompt:

„I am 47, weigh 95 kilogram with a height of 180 cm. I go running three times a week 6 to 8 kilometers. What are the best jogging shoes for me?“

In diesem Prompt sind wichtige Kontextinformationen vorhanden:

Alter, Gewicht, Größe, Distanz als Attribute und Jogging-Schuhe als Haupt-Entität.

Produkte, die in diesen oder ähnlichen Kontexten und Themenfeldern häufig genannt werden, werden eine größere Wahrscheinlichkeit haben, von der generativen KI genannt zu werden.

Ein Test mit Gemini, Copilot, ChatGPT und Perplexity zeigt auf, welche dieser Kontexte berücksichtigt werden.

Anhand der Überschriften der zitierten Quellen lassen sich folgende Schlussfolgerungen ziehen. Alle vier Systeme konnten aus den Attributen ableiten, dass ich zu schwer bin bezie-

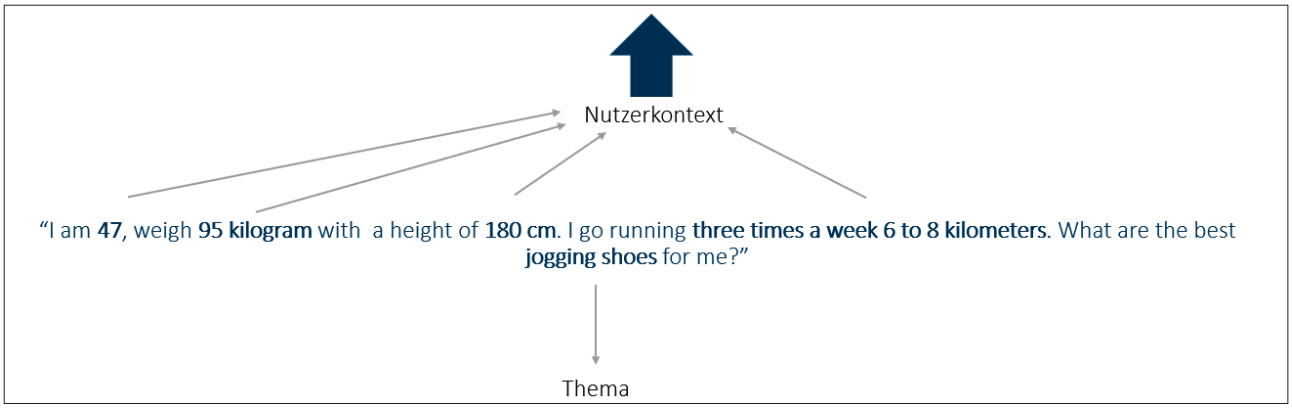


Abb. 7: Prompts und Kontext, © Olaf Kopp, Aufgesang GmbH

hungsweise Übergewicht habe, und haben dementsprechend die Informationen aus Beiträgen mit zum Beispiel folgenden Überschriften generiert:

- Best Running Shoes for Heavy Runners (August 2024)
- 7 Best Running Shoes For Heavy Men in 2024
- Best Running Shoes for Heavy Men in 2024
- Best running shoes for heavy female runners
- 7 Best Long Distance Running Shoes in 2024

Copilot

Copilot berücksichtigt die Attribute Alter und Gewicht und es wird laut den referenzierten Quellen aus den Angaben der Kontext Übergewicht abgeleitet. Alle Quellen sind informationsorientierte Inhalte wie Tests, Reviews und Listicles. Keine der Quellen ist eine typische Shopseite wie Kategorie oder PDP.

ChatGPT

ChatGPT berücksichtigt die Attribute Distanz und Gewicht und es wird laut den referenzierten Quellen aus den Angaben der Kontext Übergewicht und lange Distanz abgeleitet. Alle Quellen sind informationsorientierte Inhalte wie Tests, Reviews und Listicles. Keine der Quellen ist eine typische Shopseite wie Kategorie oder PDP.

Perplexity

Perplexity berücksichtigt das Attribut Gewicht und es wird laut den referenzierten Quellen aus den Angaben der Kontext Übergewicht abgeleitet. Quellen sind informationsorientierte Inhalte wie Tests, Reviews und Listicles, aber auch typische Shopseiten.

Gemini

Nur Gemini gibt keine Quellen bei dem Prompt an. Aber über „Weitere Informationen suchen“ sieht man,

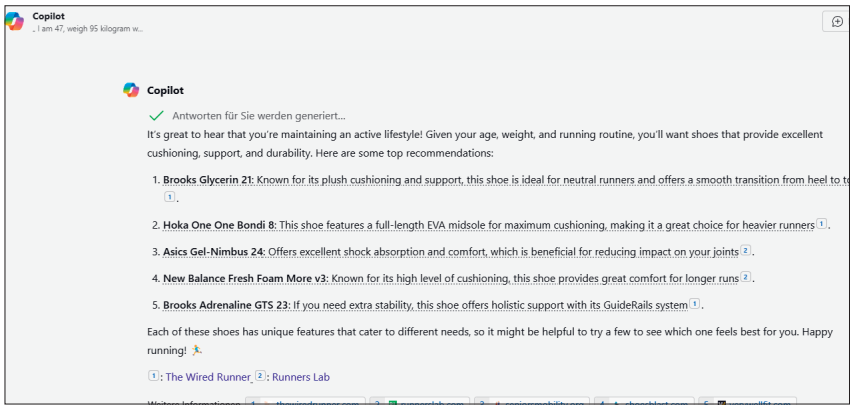


Abb. 8: Screenshot aus Microsoft Copilot

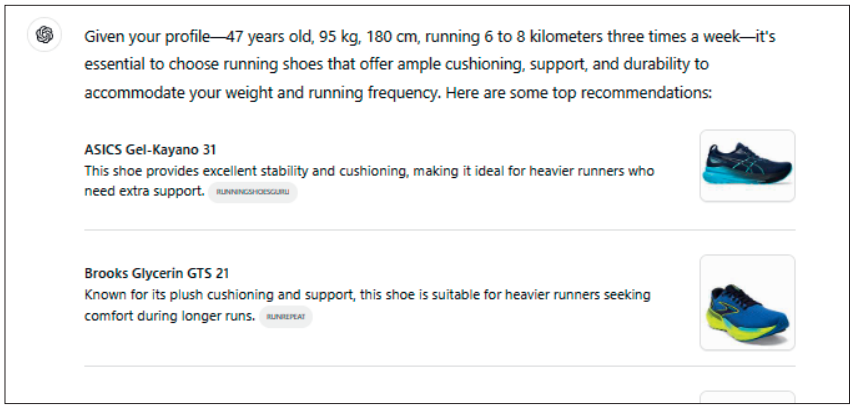


Abb. 9: Screenshot aus SearchGPT

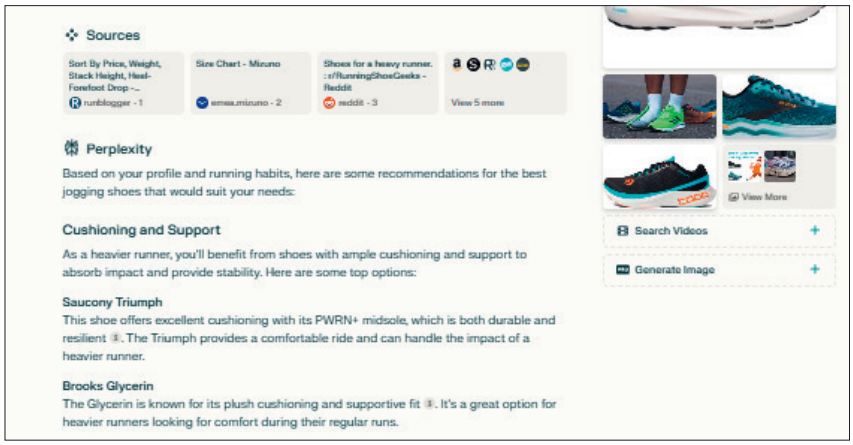


Abb. 10: Screenshot aus Perplexity

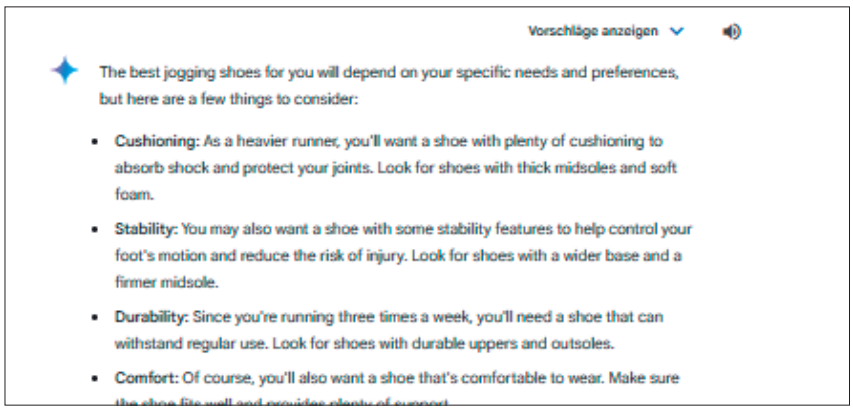


Abb. 11: Screenshot aus Gemini

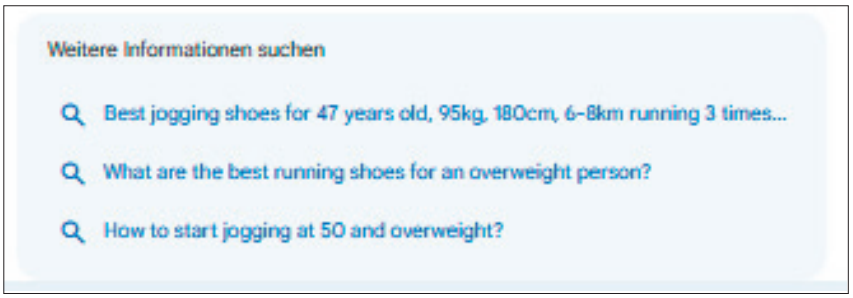


Abb. 12: Screenshot aus Gemini

ChatGPT	Copilot	Gemini	Perplexity
1. Session	1. Session	1. Session	1. Session
▪ ASICS Gel-Kayano 31	▪ Brooks Glycerin 21	▪ Brooks Glycerin GTS 21	▪ Saucony Triumph
▪ Brooks Glycerin GTS 21	▪ Hoka One Bondi 8	▪ Hoka One Bondi 8	▪ Brooks Glycerin
▪ Saucony Echelon 9	▪ Asics Gel Nimbus 24	▪ ASICS Gel-Kayano 31	▪ Asics Nimbus 25
▪ HOKA Skyflow	▪ New Balance Fresh Foam More v3	▪ New Balance Fresh Foam 1080v12	▪ Saucony Ride 15 or 16
	▪ Brooks Adrenaline GTS 23		▪ New Balance 1080
2.Session	2.Versuch:	2.Versuch:	2.Versuch:
▪ ASICS Gel-Nimbus 26	▪ Asics Gel-Kayano 31	▪ Brooks Glycerin 20	▪ Brooks Glycerin 20
▪ Brooks Glycerin GTS 21	▪ Hoka One Gaviota 4	▪ Hoka Bondi 8	▪ Saucony Triumph 20
▪ Hoka Skyflow	▪ New Balance Fresh Foam More v3	▪ ASICS Gel-Kayano 31	▪ Asics Gel-Nimbus 25
▪ Nike Vomero 17	▪ Brooks Adrenaline GTS 22	▪ New Balance Fresh Foam 1080v12	▪ Nike Invincible 3
▪ ASICS Gel-Kayano 31	▪ Saucony Triumph 20	▪ Saucony Triumph 2	

Abb. 13: Vorgeschlagene Laufschuhe bei den unterschiedlichen KI-Anwendungen in mehreren Versuchen

I want to be transparent: I don't have access to current product catalogs or real-time information sources. These recommendations come from my training data, which only goes up to April 2024, and I can't be certain that all these specific models are still current or available.

Abb. 14: Screenshot aus Claude AI

dass auch Gemini die Kontexte Alter, Gewicht beziehungsweise Übergewicht verarbeitet hat.

Es fällt auf, dass jedes der großen LLMs unterschiedliche Produkte listet:

Nur ein Schuh wird von allen vier getesteten KI-Systemen einheitlich vorgeschlagen.

Alle KI-Systeme besitzen einen gewissen „Kreativitätsspielraum“ und schlagen in verschiedenen Sessions unterschiedliche Produkte vor. Auffällig ist, dass sich Copilot, Perplexity und ChatGPT in erster Linie auf nicht kommerzielle Quellen wie Shop-Kategorieseiten oder PDPs beziehen, was auf den Zweck des Prompts hinweist.

Claude ist das einzige der bekannten Sprachmodelle, das ich nicht weiter getestet habe. Es schlägt zwar auch Schuhmodelle vor, hat aber nur Zugriff auf initiale Trainingsdaten und keine Realtime-Daten. Claude hat bisher keine Möglichkeit, auf ein eigenes Retrieval-System zuzugreifen.

Wie man an den unterschiedlichen Ergebnissen sieht, wird jedes LLM seinen eigenen Prozess haben, Quellen und Inhalte auszuwählen, was die Herausforderung bei der LLMO noch größer macht.

Wie können Quellen und Informationen für den RAG-Prozess ausgewählt werden?

Bei LLMO geht es vor allem um Positionierung. Positionierung der eigenen Produkte, Marken und Inhalte in den Trainingsdaten der Large Language Models. Deswegen ist es wichtig, zu verstehen, wie der Trainingsprozess bei LLMs funktioniert, um mögliche Ansatzpunkte zu identifizieren.

Die nachfolgenden Gedanken entstammen verschiedenen Untersuchungen, Patenten und wissenschaftlichen Dokumenten, meinen Recherchen zu E-E-A-T und eigenen Gedanken.

Die zentralen Fragen sind, wie groß der Einfluss der Retrieval-Systeme im RAG-Prozess ist, wie entscheidend die initialen Trainingsdaten sind und welche Faktoren darüber hinaus eine Rolle spielen können.

Hierzu gab es in den vergangenen ein bis zwei Jahren verschiedene Untersuchungen vor allem zu den ausgewählten Quellen bei den AI Overviews beziehungsweise SGE bei Google, Perplexity und Copilot.

Aktuell scheint es bei den Google AI Overviews einen Overlap von circa 50 % (*einfach.st/rich52* und *einfach.st/rich53*) zu geben.

Die Schwankungsbreite ist sehr hoch. So lag der Overlap bei Untersuchungen von Anfang 2024 noch bei circa 15 %. Es gab aber auch Untersuchungen, die 99 % Overlap festgestellt haben.

Man kann davon ausgehen, dass sich der Einfluss des Retrieval-Systems eher im Bereich um 50 % bewegt. Das zeigt, dass hier noch viel experimentiert wird, um die Ausgaben besser werden zu lassen. Nach der berechtigten Kritik an der Qualität der Ergebnisse der AI Overviews ist das verständlich.

Die Auswahl der referenzierten Quellen in den KI-Antworten gibt einen Aufschluss darüber, an welchen Orten es sinnvoll ist, die eigenen Marken oder Produkte kontextuell passend zu positionieren. Dabei muss man zwischen den Quellen unterscheiden, die als Teil der initialen Trainingsdaten für das Anlernen der Modelle genutzt werden, und den Quellen, die im Verlauf des RAG-Prozesses themenspezifisch ergänzt werden.

Hierzu muss man sich mit dem Anlernen der Modelle genauer beschäftigen. Nachfolgend wird das Training von Gemini exemplarisch im Detail erklärt. Andere Modelle mit Anschluss an ein Retrieval-System funktionieren ähnlich.

Google trainiert sein Gemini-Modell, ein multimodales großes Sprachmodell, das verschiedene Datentypen wie Text, Bilder, Audio, Video und Code verarbeiten kann, auf umfangreichen und vielfältigen Datensätzen. Diese umfassen Webdokumente, Bücher, Code sowie Bild-, Audio- und Videodaten. Durch die Kombination dieser unterschiedlichen Datenquellen kann Gemini komplexe Aufgaben effizienter erlernen und ausführen.

Aus den verschiedenen Untersuchungen zu den AI Overviews und den am häufigsten referenzierten Quellen unabhängig vom Thema bekommt man einen Einblick, welche Quellen Google zu den eigenen Indizes und dem Knowledge Graph noch für das initiale Pre-Training nutzen kann.

Domänenspezifische Quellen werden dann im RAG-Prozess ergänzt.

Eine bemerkenswerte Charakteristik von Gemini ist die Verwendung einer Mixture-of-Experts(MoE)-Architektur. Im Gegensatz zu traditionellen Transformern, die als ein großes neuronales Netzwerk fungieren, ist ein MoE-Modell in kleinere „Experten"-Netzwerke unterteilt. Je nach Art der Eingabe aktiviert das Modell selektiv die relevantesten Expertenpfade, was die Effizienz und Leistungsfähigkeit des Modells erheblich steigert. Hier wird wahrscheinlich auch der RAG-Prozess verortet sein.

Gemini wird von Google durch eine Kombination aus mehreren Trainingsphasen entwickelt, die auf öffentlich zugänglichen Daten und einer speziellen Technik basieren, um die Relevanz und Präzision der generierten Inhalte zu maximieren:

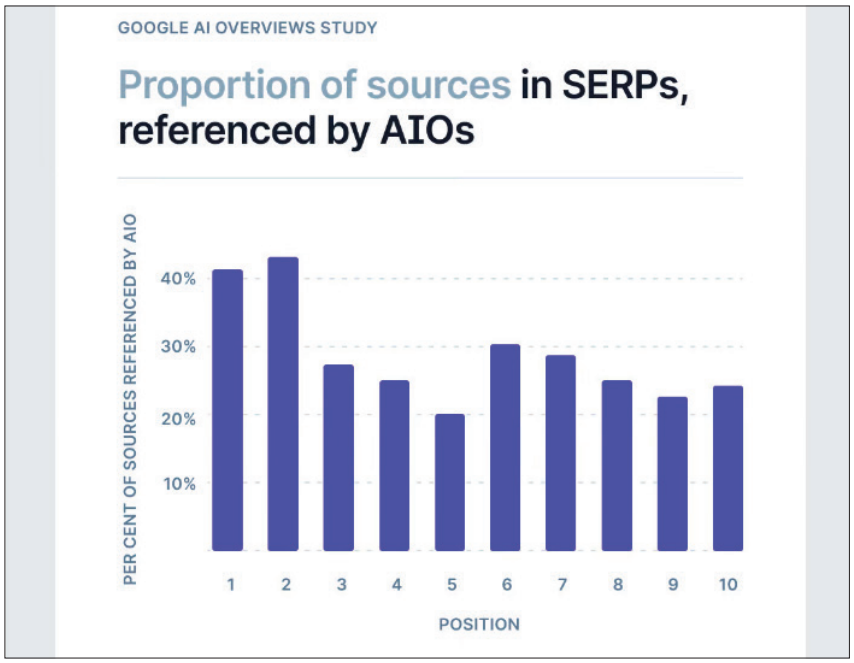
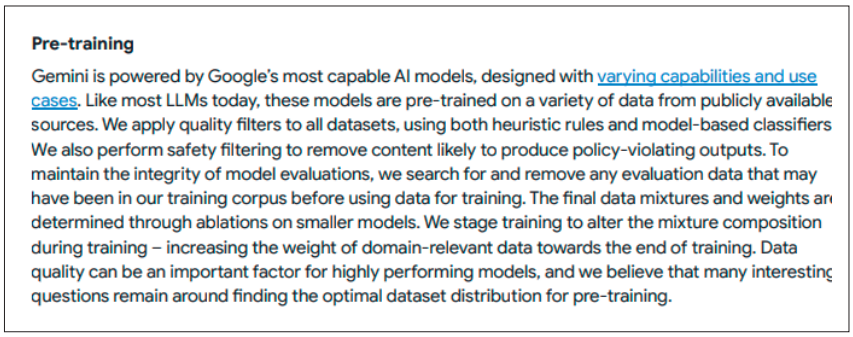


Abb. 15: Zusammenhang zwischen Top-Ten-Rankings und Referenzierung in den AI Overviews, *Quelle: einfach.st/rich53*

Pos.	Domain	Anzahl Keywords
1	youtube.com	329.924
2	wikipedia.org	329.276
3	nih.gov	191.096
4	healthline.com	159.688
5	clevelandclinic.org	147.102
6	webmd.com	119.280
7	medicalnewstoday.com	115.342
8	study.com	101.975
9	mayoclinic.org	101.970
10	britannica.com	86.084
...	...	...
196	texas.gov	3.248
197	spotify.com	2.570
198	netflix.com	2.503
199	ssa.gov	1.945
200	paypal.com	1.791

Abb. 16: Bevorzugte Quellen in den AI Overviews, *Quelle: einfach.st/sistrix72*

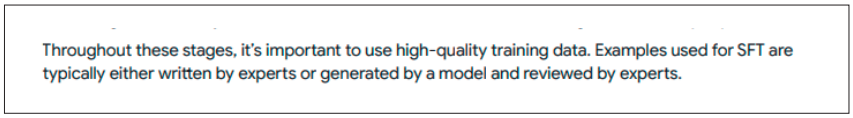
1. Pre-Training: Ähnlich wie bei anderen großen Sprachmodellen (LLMs) wird Gemini zunächst auf einer Vielzahl öffentlicher Datenquellen vortrainiert. Dabei wendet Google verschiedene Filter an, um die Datenqualität sicherzustellen und problematische Inhalte zu vermeiden. Das Training berücksichtigt eine flexible Auswahl wahrscheinlicher Wörter, was kreativere und kontextuell passende Antworten ermöglicht.



Pre-training

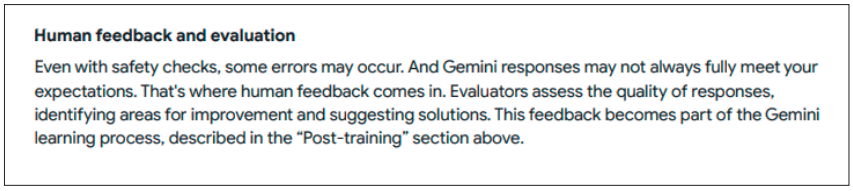
Gemini is powered by Google's most capable AI models, designed with varying capabilities and use cases. Like most LLMs today, these models are pre-trained on a variety of data from publicly available sources. We apply quality filters to all datasets, using both heuristic rules and model-based classifiers We also perform safety filtering to remove content likely to produce policy-violating outputs. To maintain the integrity of model evaluations, we search for and remove any evaluation data that may have been in our training corpus before using data for training. The final data mixtures and weights ar determined through ablations on smaller models. We stage training to alter the mixture composition during training – increasing the weight of domain-relevant data towards the end of training. Data quality can be an important factor for highly performing models, and we believe that many interesting questions remain around finding the optimal dataset distribution for pre-training.

Abb. 17: Google über das Trainieren von Gemini im Pre-Training, *Quelle: An overview of the Gemini app*



Throughout these stages, it's important to use high-quality training data. Examples used for SFT are typically either written by experts or generated by a model and reviewed by experts.

Abb. 18: Google über das Trainieren von Gemini im Post-Training, *Quelle: An overview of the Gemini app*



Human feedback and evaluation

Even with safety checks, some errors may occur. And Gemini responses may not always fully meet your expectations. That's where human feedback comes in. Evaluators assess the quality of responses, identifying areas for improvement and suggesting solutions. This feedback becomes part of the Gemini learning process, described in the "Post-training" section above.

Abb. 19: Google über das Trainieren von Gemini im Post-Training, *Quelle: An overview of the Gemini app*

2. Feinabstimmung (Supervised Fine-Tuning – SFT): Nach dem Vortraining wird das Modell mithilfe von hochwertigen Beispielen optimiert, die entweder von Experten erstellt oder von Modellen generiert und dann von Experten überprüft wurden. Dieser Prozess ist vergleichbar mit dem Lernen von guten Textstrukturen und Inhalten, indem man Beispiele von gut geschriebenen Texten sieht.
3. Reinforcement Learning from Human Feedback (RLHF): Hierbei wird das Modell anhand menschlicher Bewertungen weiterentwickelt. Ein Reward-Modell, das auf Präferenzen der Nutzer basiert, hilft Gemini, bevorzugte Antwortstile und -inhalte zu erkennen und zu lernen.
4. Erweiterungen und Abruf-Augmentation (RAG): Gemini kann externe Datenquellen wie Google Search, Maps, YouTube oder spezifische Erweiterungen durchsuchen, um kontextbezogene Informationen zur Antwort zu liefern. Beispielsweise könnte Gemini bei einer Anfrage nach den aktuellen Wetterbedingungen oder Nachrichten direkt auf Google Search zugreifen, um aktuelle, verlässliche Daten zu finden und diese in die Antwort einfließen zu lassen.

» Um die relevantesten Informationen für die Antwort auszuwählen, führt Gemini eine Filterung der Suchergebnisse durch. Dabei berücksichtigt das Modell die Kontextualität der Anfrage und filtert die Daten so, dass sie möglichst genau zur Fragestellung passen. Ein Beispiel hierfür wäre eine komplexe technische Frage, bei der das Modell Ergebnisse auswählt, die von wissenschaftlicher oder technischer Natur sind, anstatt allgemeine Webinhalte zu verwenden.

» Gemini kombiniert die abgerufenen Informationen aus externen Quellen mit dem Modell-Output. Dieser Prozess beinhaltet das Erstellen eines optimierten Antwortentwurfs, der sowohl auf das Vorwissen des Modells als auch auf die Informationen der abgerufenen Datenquellen zurückgreift. Das Modell strukturiert die Antwort dabei so, dass die Informationen logisch zusammengeführt und lesbar präsentiert werden. Jede Antwort wird zusätzlich geprüft, um sicherzustellen, dass sie den Qualitätsstandards von Google entspricht und keine problematischen oder unangemessenen Inhalte enthält. Diese Sicherheitsüberprüfung wird durch ein Ranking ergänzt, bei dem die qualitativ besten Versionen der Antwort bevorzugt werden. Das Modell präsentiert dem Benutzer dann die höchstrangige Antwort.

5. Nutzerfeedback und ständige Optimierung: Google integriert fortlaufend Rückmeldungen von Nutzern und Experten, um das Modell anzupassen und eventuelle Schwachstellen zu beheben.

Die erste Möglichkeit ist, dass die KI-Anwendungen auf bestehende Retrieval-Systeme zugreifen und deren Suchergebnisse nutzen. Allerdings steigt laut verschiedenen Untersuchungen die Chance, dass man bei einem guten Ranking in der jeweiligen Suchmaschine auch als Quelle in den angeschlossenen KI-Anwendungen genannt wird, aber die Overlaps zeigen wie bereits erläutert bisher keine klare Korrelation zwischen den Top-Rankings und den referenzierten Quellen.

Hier muss noch ein anderes Kriterium genutzt werden, um die Quellen auszuwählen.

Hier fällt der Verweis seitens Google auf, bei der Auswahl der Quellen sowohl im Rahmen des Pre-Trainings als auch bei RAG auf die eigenen Qualitätsstandards zu achten. Zudem werden Klassifikatoren genannt.

Bei der Nennung von Klassifikatoren kann die Brücke zu E-E-A-T

geschlagen werden, wo auch Qualitäts-Klassifikatoren genutzt werden. Auch zum Post-Training findet man in den Informationen von Google Hinweise auf einen Einsatz von E-E-A-T bei der Klassifizierung der Quellen nach Qualität.

Der Verweis auf die Evaluatoren lässt die Brücke zu den Quality Ratern schlagen, die sich mit der Bewertung von E-E-A-T beschäftigen.

Die Rankings in den meisten Suchmaschinen werden nach Relevanz und Qualität auf den Ebenen des Dokuments, der Domain und des Urhebers oder der Source Entity ermittelt.

Es könnte sein, dass Quellen weniger über die Relevanzkriterien ausgewählt werden als über die Qualitätskriterien auf Domain- und Source-Entity-Ebene. Das wäre auch sinnvoll, da komplexere Prompts im Hintergrund umgeschrieben werden müssen, damit entsprechende Suchanfragen für die Abfrage der Rankings entstehen. Relevanz ist abhängig von der Suchanfrage, Qualität nicht.

Das würde erklären, warum man nur eine kleine Korrelation zwischen Rankings und durch die generative KI referenzierten Quellen feststellen kann und auch Quellen referenziert werden, die nicht in den Top-Positionen zu finden sind.

Zur Bewertung der Qualität nutzen Suchmaschinen wie Google oder auch Bing Qualitäts-Classifier wie zum Beispiel bei Google E-E-A-T.

Google hat immer wieder betont, dass sich E-E-A-T auf Themenfelder bezieht. Deswegen muss auch bei LLMO-Strategien eine themenspezifische Strategie berücksichtigt werden.

Eine Untersuchung von Brightedge (www.brightedge.com/perplexity) hat ergeben, dass sich neben der generellen Wichtigkeit von Wikipedia, Foren wie Reddit und Amazon bei Perplexity die referenzierten Domain-Quellen nach Branche/Thema unterscheiden.

Bei der Strategie der Positionierung müssen demnach branchen- beziehungsweise themenspezifische Begebenheiten berücksichtigt werden.

Taktische und strategische Ansätze für LLMO?

Wie eingangs erwähnt, gibt es noch keine nachweisbaren Erfolgsstorys, was die Beeinflussung der Ergebnisse von generativer KI angeht.

Zudem scheinen sich die Plattformbetreiber selbst noch nicht sicher zu sein, wie sie die Quellen qualifizieren, die sie dann im Rahmen des RAG-Prozesses auswählen.

Die bisherigen Erläuterungen sollen zeigen, dass es wichtig ist, mehr darüber zu erfahren, wo man optimieren sollte. Es ist demnach fraglich, welche Quellen so vertrauenswürdig und relevant sind, dass man dort ansetzen sollte. Die zweite Frage ist, wie man selbst zu so einer Quelle wird.

Eine erste wissenschaftliche Abhandlung, ob sich Ausgaben generativer KI beeinflussen lassen und welche Faktoren dafür verantwortlich sein können, gibt es in dem Research-Paper mit dem Titel „GEO: Generative Engine Optimization“. Diese Untersuchung hat auch den Namen „GEO“ geprägt.

Laut dem Research-Paper können die Sichtbarkeit und Effektivität der Generative Engine Optimization (GEO) durch folgende Faktoren erhöht werden:

» Autorität in der Schreibweise: Sie verbessert die Leistung insbesondere bei debattierenden Fragen und Anfragen im historischen Kontext, da eine überzeugendere Schreibweise in debattenähnlichen Kontexten wahrscheinlich mehr wert ist.
» Zitierungen (Cite Sources): besonders vorteilhaft für faktische Fragen, da Zitierungen eine Quelle der Überprüfung für die präsentierten Fakten bieten, wodurch die Glaubwürdigkeit der Antwort erhöht wird
» Einfügung von Statistiken (Statistics Addition): besonders nützlich in Bereichen wie Recht und Regierung und bei Meinungsfragen, da die Einbeziehung von relevanten Statistiken in den Website-Inhalt die Sichtbarkeit einer Website in bestimmten Kontexten, insbesondere diesen, erhöhen kann
» Zitat hinzufügen (Quotation Addition): am effektivsten in den Bereichen Menschen und Gesellschaft, Erklärungen und Geschichte. Dies könnte daran liegen, dass diese Bereiche oft persönliche Erzählungen oder historische Ereignisse umfassen, bei denen direkte Zitate Authentizität und Tiefe zum Inhalt hinzufügen können.

Diese Faktoren variieren in ihrer Effektivität je nach Bereich, was darauf hindeutet, dass die Einbeziehung von domänenspezifischen, zielgerichteten Anpassungen auf Websites für eine höhere Sichtbarkeit wesentlich ist.

Aus dem Paper lassen sich folgende taktische Dos für GEO beziehungsweise LLMO ableiten:

» Zitierbare Quellen nutzen: Integriere zitierbare Quellen in deinen Inhalt, um die Glaubwürdigkeit und Authentizität zu erhöhen, besonders bei faktischen Fragen.
» Statistiken einfügen: Füge relevante Statistiken hinzu, um deine Argumente zu stärken, insbesondere in Bereichen wie Recht und Regierung und bei Meinungsfragen.
» Zitate hinzufügen: Nutze Zitate, um den Inhalt in Bereichen wie Menschen und Gesellschaft, Erklärungen und Geschichte zu bereichern, da sie Authentizität und Tiefe hinzufügen.
» Domänenspezifische Optimierung: Berücksichtige die Besonderheiten deiner Domäne bei der Optimierung, da die Effektivität von GEO-Methoden je nach Bereich variiert.
» Fokus auf Inhaltsqualität: Konzentriere dich auf die Erstellung hochwertiger, relevanter und informativer

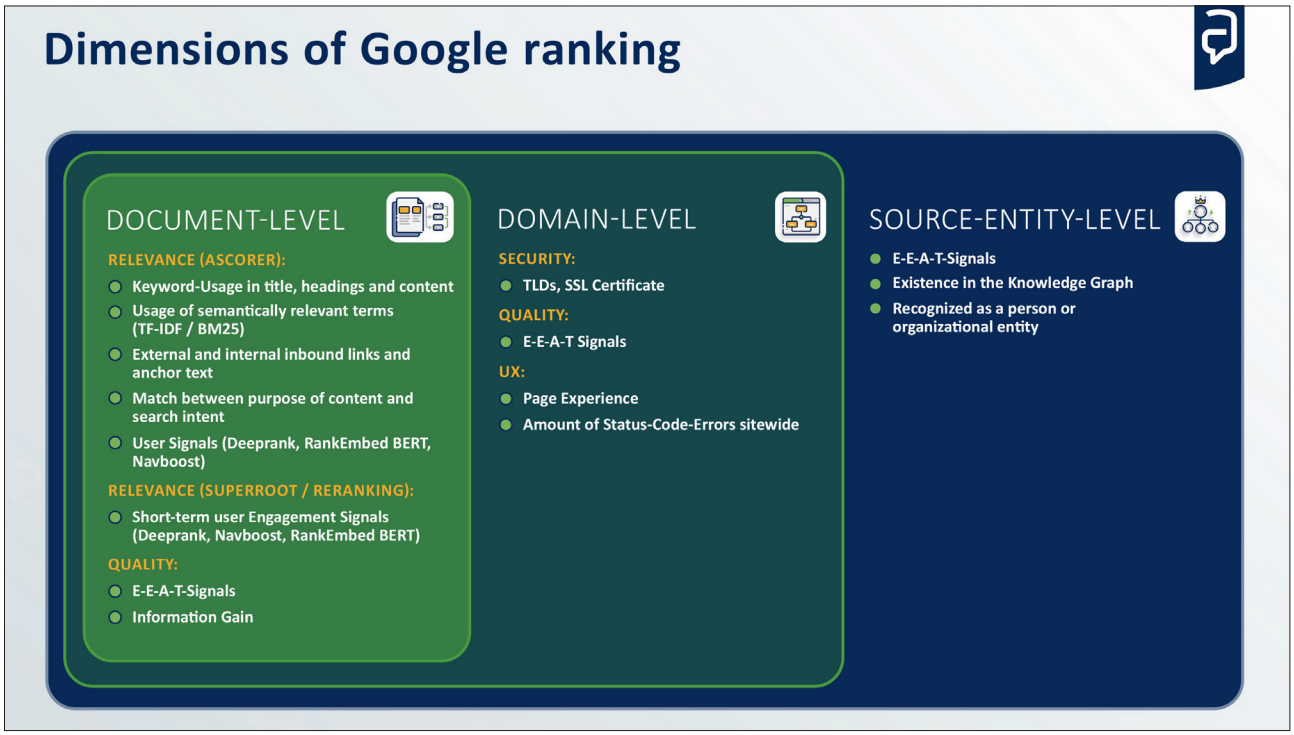


Abb. 20: Ranking-Dimensionen bei Google, © Olaf Kopp, Aufgesang GmbH

Inhalte, die den Nutzern einen Mehrwert bieten.

Auch taktische Don'ts lassen sich ableiten:

» Vermeide Keyword-Stuffing: Traditionelles Keyword-Stuffing zeigt wenig bis keine Verbesserung bei den Antworten generativer Suchmaschinen und sollte vermieden werden.
» Ignoriere nicht den Kontext: Vermeide die Generierung von Inhalten, die keinen Bezug zum Thema haben oder keinen Mehrwert für den Nutzer bieten.
» Übersehe nicht die Nutzerintention: Vernachlässige nicht die Absicht hinter Suchanfragen. Stelle sicher, dass dein Inhalt die Fragen der Nutzer tatsächlich beantwortet.

Brightedge hat aus der bereits erwähnten Untersuchung folgende strategische Überlegungen abgeleitet:

» **Unterschiedliche Auswirkungen von Backlinks und Co-Citations:** SGE (AI Overviews) und Perplexity haben jeweils ihre eigenen Domain-Sets, die sie je nach Branche bevorzugen. Im Gesundheits- und Bildungswesen schätzen beide Plattformen vertrauenswürdige Quellen wie mayoclinic.org und coursera.com. Die Ausrichtung der SEO-Strategien auf diese Domains oder ähnliche Domains kann effektiv sein. Im Gegensatz dazu zeigt Perplexity für Sektoren wie E-Commerce und Finanzen eine Präferenz für Domains wie reddit.com, yahoo.com und marketwatch.com. SEO-Bemühungen sollten an diese Präferenzen angepasst werden, indem Backlinks oder Co-Citations genutzt werden, um die SEO-Leistung zu steigern.
» **Maßgeschneiderte Strategien für die KI-gestützte Suche:** Der Ansatz für die KI-gestützte Suche muss für jede Branche individuell angepasst werden. Die Vorliebe von Perplexity für reddit.com unterstreicht den Wert von Community-Erkenntnissen im E-Commerce, während SGE etablierte Bewertungs- und Q&A-Websites wie consumerreports.org und quora.com priorisiert. Marketingfachleute und SEO-Experten sollten Content-Strategien entwickeln, die auf diese spezifischen Tendenzen eingehen, wie zum Beispiel die Erstellung detaillierter Produktbewertungen oder die Förderung von Q&A-Foren für E-Commerce-Marken.
» **Änderungen der Zitierlandschaft vorhersehen:** SEO-Experten müssen die bevorzugten Domains von Perplexity aufmerksam verfolgen, insbesondere die Nutzung von reddit.com für Inhalte aus der Community. Durch die Partnerschaft von Google mit Reddit könnte Perplexity auch seinen Algorithmus ändern, um die Inhalte von Reddit stärker zu berücksichtigen. Dies könnte ein Hinweis darauf sein, dass Inhalte, die durch die Interaktion der Nutzer entstehen, stärker in den Vordergrund rücken. SEO-Experten müssen proaktiv und anpassungsfähig bleiben, um sicherzustellen, dass ihre Content-Strategie angesichts der sich ändernden Zitiermuster von Perplexity relevant und effektiv bleibt.

Des Weiteren können folgende branchenspezifische taktische und strategische Maßnahmen für LLMO abgeleitet werden. Dabei sollten die genannten Domains für den deutschsprachigen Markt geprüft und gegebenenfalls abgeleitet werden.

B2B-Tech

» Präsenz auf autoritativen Tech-Domains aufbauen, insbesondere auf techtarget.com, ibm.com, microsoft.com und cloudflare.com, die von beiden Plattformen als vertrauenswürdige Quellen anerkannt werden
» Content-Syndizierung auf diesen etablierten Plattformen nutzen, um schneller als vertrauenswürdige Quelle zitiert zu werden
» Langfristig die eigene Domain-Autorität durch hochwertige Inhalte aufbauen, da der Wettbewerb um Syndizierungsplätze zunehmen wird
» Partnerships mit führenden Tech-Plattformen eingehen und dort aktiv Content beisteuern
» Expertise durch Credentials, Zertifizierungen und Expertenmeinungen nachweisen, um Vertrauenswürdigkeit zu signalisieren
» E-Commerce
» Eine starke Präsenz auf Amazon aufbauen, da die Plattform von Perplexity häufig als Quelle verwendet wird
» Aktiv Produktbewertungen und User Generated Content auf Amazon und anderen relevanten Plattformen fördern
» Produktinformationen über etablierte Händlerplattformen und Vergleichsseiten verbreiten
» Inhalte syndizieren und Partnerschaften mit vertrauenswürdigen Domains eingehen
» Detaillierte und aktuelle Produktbeschreibungen auf allen Verkaufsplattformen pflegen
» Sich auf relevanten Fachportalen und Communityplattformen wie Reddit engagieren
» Eine ausgewogene Marketingstrategie verfolgen, die sowohl auf externe Plattformen als auch auf die eigene Domain-Autorität setzt

Weiterbildung

» Vertrauenswürdige Quellen aufbauen und mit autoritären Domains wie coursera.org, usnews.com und bestcolleges.com zusammenarbeiten, da diese von beiden Systemen als relevant eingestuft werden
» Aktuellen und qualitativ hochwertigen Content erstellen, der von den KI-Systemen als vertrauenswürdig eingestuft wird. Die Inhalte sollten klar strukturiert und mit Expertenwissen untermauert sein.
» Eine aktive Präsenz auf relevanten Plattformen wie Reddit aufbauen, da communitygetriebene Inhalte zunehmend an Bedeutung gewinnen
» Die eigenen Inhalte für KI-Systeme durch eine klare Strukturierung, eindeutige Überschriften und prägnante Antworten auf häufige Nutzerfragen optimieren
» Qualitätsmerkmale wie Zertifizierungen und Akkreditierungen deutlich hervorheben, da diese die Glaubwürdigkeit erhöhen

Gesundheit

» Inhalte mit vertrauenswürdigen Quellen wie mayoclinic.org, nih.gov und medlineplus.gov verlinken und referenzieren
» Aktuelle medizinische Forschung und Trends in die Inhalte einbinden
» Umfassende und gut recherchierte medizinische Informationen bereitstellen, die von offiziellen Institutionen gestützt werden
» Auf Glaubwürdigkeit und Expertise durch Zertifizierungen und Qualifikationen setzen
» Regelmäßige Updates der Inhalte mit neuen medizinischen Erkenntnissen durchführen
» Eine ausgewogene Content-Strategie verfolgen, die sowohl eigene Domain-Autorität aufbaut als auch etablierte Gesundheitsplattformen nutzt

Finanzen

» Präsenz auf vertrauenswürdigen Finanzportalen wie yahoo.com und marketwatch.com aufbauen, da diese von den KI-Systemen bevorzugt als Quellen genutzt werden
» Aktuelle und präzise Unternehmensinformationen auf führenden Plattformen wie Yahoo Finance pflegen
» Hochwertige, faktisch korrekte Inhalte erstellen und diese durch Referenzen zu anerkannten Quellen untermauern
» Eine aktive Präsenz in relevanten Reddit-Communitys aufbauen, da Reddit zunehmend als Quelle für KI-Systeme an Bedeutung gewinnt
» Partnerschaften mit etablierten Finanzmedien eingehen, um die eigene Sichtbarkeit und Glaubwürdigkeit zu erhöhen

Expertise durch Fachwissen, Zertifizierungen und Expertenmeinungen demonstrieren

Versicherung

» Vertrauenswürdige Quellen nutzen: Inhalte auf anerkannten Domains wie forbes.com und offiziellen Regierungsseiten (.gov) platzieren, da diese von KI-Suchmaschinen als besonders glaubwürdig eingestuft werden
» Aktuelle und präzise Informationen bereitstellen: Die Versicherungsinformationen müssen stets aktuell und faktisch korrekt sein. Dies gilt besonders für Produkt- und Leistungsbeschreibungen.
» Content-Syndikation: Inhalte auf autoritativen Plattformen wie Forbes oder anerkannten Fachportalen veröffentlichen, um schneller als vertrauenswürdige Quelle zitiert zu werden
» Lokale Relevanz betonen: Die Inhalte sollten an regionale Märkte angepasst werden und lokale Versicherungsbestimmungen berücksichtigen.

Restaurants

» Eine starke Präsenz auf wichtigen Bewertungsplattformen wie Yelp, Tripadvisor, OpenTable und GrubHub aufbauen und pflegen
» Positive Bewertungen und Rezensionen von Gästen aktiv fördern und sammeln
» Vollständige und aktuelle Informationen auf diesen Plattformen bereitstellen (Menüs, Öffnungszeiten, Fotos etc.)
» Mit Food-Communitys und spezialisierten Gastronomieplattformen wie Eater.com interagieren
» Lokale SEO durchführen, da KI-Suchen stark auf lokale Relevanz achten
» Umfassende und gut gepflegte Wikipedia-Einträge erstellen und aktualisieren
» Einen nahtlosen Online-Reservierungsprozess über die relevanten Plattformen anbieten
» Hochwertigen Content über das Restaurant auf verschiedenen Kanälen bereitstellen

Tourismus/Reise

» Präsenz auf wichtigen Reiseplattformen wie Tripadvisor, Expedia, Kayak, Hotels.com und Booking.com optimieren, da diese von KI-Suchmaschinen als vertrauenswürdige Quellen angesehen werden
» Umfassende Inhalte mit Reiseführern, Tipps und authentischen Bewertungen erstellen
» Den Buchungsprozess optimieren und benutzerfreundlich gestalten
» Lokale SEO durchführen, da KI-Suchen oft standortbasiert sind
» Auf relevanten Plattformen aktiv sein und Bewertungen fördern
» Qualitativ hochwertige Inhalte mit Mehrwert für den Nutzer bereitstellen
» Mit vertrauenswürdigen Domains und Partnern zusammenarbeiten

Fazit: Wie wichtig ist LLMO für Unternehmen?

Ob LLMO oder GEO wirklich eine legitime Strategie wird, um LLMs im Hinblick auf die eigenen Ziele zu beeinflussen, bleibt abzuwarten.

Wenn dies der Fall ist, müssen folgende Ziele erreicht werden:

» Etablierung von Owned Media als Quelle für LLM-Trainingsdaten über E-E-A-T
» Generierung von Erwähnungen der Marke und der Produkte in qualifizierten Medien
» Generierung von gemeinsamen Nennungen der eigenen Marke mit anderen relevanten Entitäten und Attributen in qualifizierten Medien
» Erstellung von relevanten Inhalten, die gut ranken und dadurch im RAG-Prozess berücksichtigt werden
» Stattfinden in etablierten Graph-Datenbanken wie Knowledge Graph oder Shopping Graph

Die Erfolgschancen der LLM-Optimierung stehen in direktem Zusammenhang mit der Größe des Markts: Je nischenhafter ein Markt ist, desto einfacher ist es, sich als Marke im jeweiligen thematischen Kontext zu positionieren.

Das bedeutet, dass weniger Kookkurrenzen in den qualifizierten Medien erforderlich sind, um mit den relevanten Attributen und Entitäten in den LLMs in Verbindung gebracht zu werden. Je größer der Markt, desto schwieriger ist dies, da viele Marktteilnehmer über große Ressourcen in den Bereichen PR und Marketing sowie über eine lange Geschichte verfügen.

GEO oder LLMO erfordert deutlich mehr Ressourcen als die klassische Suchmaschinenoptimierung, um die öffentliche Wahrnehmung zu beeinflussen.

An dieser Stelle möchte ich auf mein Konzept des Digital-Authority-Managements verweisen, das Unternehmen strukturell und personell so aufstellt, dass sie für die KI-Zukunft gerüstet sind.

Zukünftig werden große Marken aufgrund ihrer PR- und Marketingressourcen in Zukunft erhebliche Vorteile bei der Suchmaschinenpositionierung und den Ergebnissen der generativen KI haben.

Eine andere Perspektive ist, dass die Suchmaschinenoptimierung wie bisher fortgesetzt werden kann, da gleichzeitig gut gerankte Inhalte zum Training von LLMs verwendet werden. Das hängt vom Einfluss der Retrieval Systeme auf die Auswahl der Quellen ab. Haben Relevanz-Faktoren mehr Gewicht oder Qualitäts-Faktoren? Wie groß ist der Einfluss der originären Ranking-Systeme im Verhältnis zu weiteren Auswahl-Kriterien?

Man sollte in jedem Fall auf das gemeinsame Auftreten von Marken/Produkten und Attributen oder anderen Entitäten achten und für diese optimieren. ¶