
# Large Language Model Optimierung (LLMO) für Websites

03 | 2025
netzleuchten GmbH | Kiel

# Warum Sie ohne KI-Optimierung **Besucher verlieren werden**

Eine aktuelle Studie von Gartner[1] prognostiziert einen dramatischen Rückgang klassischer Suchanfragen um 25% bis zum Jahr 2026. Dieser Trend wird hauptsächlich durch die zunehmende Nutzung von KI-gestützten Chatbots und virtuellen Assistenten getrieben.

Die jahrzehntelange Dominanz von Google im Bereich der Internetsuche gerät durch KI-gestützte Alternativen zunehmend unter Druck. Laut einer aktuellen Analyse von similarweb[2] hat besonders Microsofts Bing durch die Integration von ChatGPT-Technologie aufgeholt. Am Tag nach Einführung des Features stieg der Zugriff auf die Suchmaschine um 15%.

Die Art und Weise, wie wir im Internet nach Informationen suchen, befinden sich in einem fundamentalen Wandel. Anders als traditionelle Suchmaschinen, die lediglich eine Liste von Links präsentieren, bieten KI-Chatbots wie ChatGPT eine deutlich intuitivere und effizientere Nutzererfahrung. Sie verstehen den Kontext einer Anfrage und können komplexe Zusammenhänge erfassen. Statt durch verschiedene Webseiten zu navigieren, erhalten Nutzer sofort eine maßgeschneiderte, zusammenfassende Antwort.

## Inhalt



# **Veränderungen** im Suchverhalten

Jüngere Nutzer*Innen (18-26 Jahre) bevorzugen laut einer Umfrage von ExpressVPN[3] bereits zu 33% KI-gestützte Suchmaschinen gegenüber Google. Sie schätzen vor allem die Möglichkeit, Fragen in natürlicher Sprache zu stellen und sofort verwertbare Antworten zu erhalten, ohne sich durch verschiedene Webseiten klicken zu müssen.

Die Länge der Suchanfragen hat sich daher gewandelt: Eine umfangreiche Analyse von SearchEngineLand[4], die 41 Millionen Suchanfragen untersuchte, offenbart bedeutende Veränderungen im Suchverhalten seit der Einführung von ChatGPT. Vor dem KI-Boom im Jahr 2022 bestanden noch 60% aller Suchanfragen aus lediglich 1-4 Wörtern. Die neuesten Daten aus 2024 zeigen einen bemerkenswerten Trend: die Suchanfragen mit 7-8 Wörtern haben sich seit der Einführung von Chat GPT verdoppelt, auf Kosten der Suchanfrage mit 5-6 Worten.

52% der Deutschen nutzen schon ChatGPT mindestens einmal täglich, 9% sogar mehrmals am Tag[5]. Als häufigste Nutzung von ChatGPT in Unternehmen wurde von 29,4% der befragten Mitarbeiter die Recherche genannt.[6]. Dies zeigt, dass die Informationsbeschaffung durch ChatGPT bereits eine zentrale Rolle in Unternehmen spielt.

# **LLMO** oder **SEO?**

Die LLM-Optimierung und traditionelles SEO ergänzen sich gegenseitig. Während SEO sich darauf konzentriert, Websites für Suchmaschinen wie Google zu optimieren, zielt die LLM-Optimierung darauf ab, Inhalte für KI-gestützte Sprachmodelle verständlicher zu machen. Beide Ansätze arbeiten Hand in Hand: SEO sorgt für die technische Grundlage und die Auffindbarkeit einer Website, während die LLM-Optimierung die inhaltliche Tiefe und kontextuelle Relevanz verstärkt.

Die klare und präzise Informationsarchitektur einer Website bildet das Fundament für die erfolgreiche Verarbeitung durch LLMs. Eine gut strukturierte Website sollte wie eine Pyramide aufgebaut sein, bei der die wichtigsten Informationen an der Spitze stehen und sich nach unten hin weiter verzweigen. Dies ermöglicht es LLMs, die Beziehungen zwischen verschiedenen Inhaltselementen besser zu verstehen und relevante Informationen schneller zu identifizieren.

Zum Beispiel helfen HTML-Elemente wie header, article, section und eine gut durchdachte Überschriftenstruktur h1 bis h6 dabei, den Aufbau und die Wichtigkeit von Inhalten zu vermitteln. Bei einer Produktseite müssen die technischen Spezifikationen in strukturierten Datentabellen table oder Listen (ul, ol) organisiert sein, damit LLMs wie bei der Suche nach einem Produkt präzise Produktempfehlungen geben können.

Die Implementierung von strukturierten Daten, zum Beispiel mittels Schema.org-Markup, macht es KI-Modellen leichter zu verstehen, wie die Inhalte zusammenhängen und einzelne Seiten zu bewerten und einzuordnen. Zusätzlich müssen grundlegende SEO-Aspekte wie Ladegeschwindigkeit, mobile Optimierung und HTTPS-Sicherheit gewährleistet sein, da diese Faktoren auch die Zugänglichkeit für LLMs beeinflussen.

Diese technischen Grundlagen, kombiniert mit qualitativ hochwertigem Content, bilden das Fundament für eine erfolgreiche LLM-Optimierung von Websites. Darauf aufbauend sorgt die LLM-Optimierung durch umgangssprachliche Formulierungen und umfassende Themen Kontexte für eine bessere Erfassung durch KI-Systeme.

## Wie eine **KI-gestützte Suche** im Web abläuft

Die jahrzehntelange Dominanz von Google im Bereich der Internetsuche gerät durch KI-gestützte Alternativen zunehmend unter Druck.

# Semantische **Relevanz**

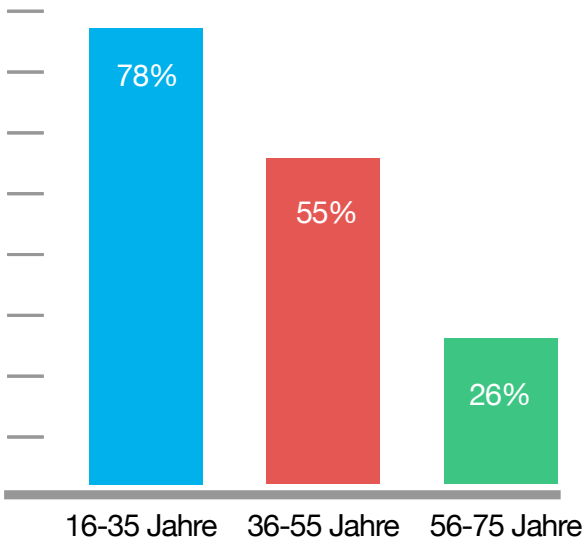
Die Bedeutung der Wortwahl in Webtexten hat sich durch den Einfluss von Large Language Models fundamental verändert. Während traditionelles SEO sich auf Keywords und deren Dichte konzentrierte, geht es nun verstärkt um natürliche Sprache und kontextuelle Zusammenhänge. Besonders wichtig bei der Verwendung von Fachterminologie ist deren verständliche Erklärung. LLMs wie ChatGPT erkennen Inhalte nicht an Keywords, sondern am gesamten Sprachkontext. Ein Beispiel aus der Medizin: Statt einer klassischen Keywordphrase “Kopfschmerzen behandeln” wird ein Text, der präzise von “Migräne-Prophylaxe durch nicht-medikamentöse Interventionen” spricht und diese detailliert erklärt, von LLMs als höherwertige Quelle eingestuft.

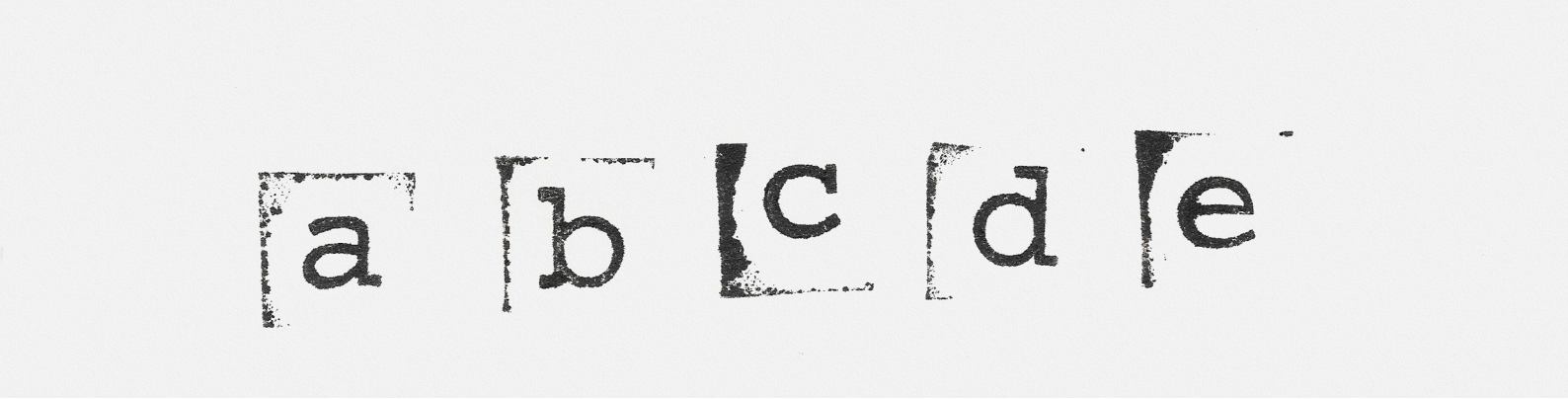
Entscheidend ist auch die Präzision der Formulierungen: LLMs bewerten Texte höher, die komplexe Sachverhalte klar und unmissverständlich darstellen. Präzise formulierte Texte werden häufiger als Quellen in KI-generierten Antworten zitiert werden als Texte mit vagen oder mehrdeutigen Formulierungen. Die Wortwahl beeinflusst so also die Wahrscheinlichkeit, in Konversation basierten Suchanfragen gefunden zu werden [7]. Texte, die natürliche Frage-Antwort-Muster enthalten, werden häufiger als Quellen herangezogen.

Dabei spielt auch die Verwendung von Synonymen und verwandten Begriffen eine wichtige Rolle. Ein Text, der beispielsweise das Thema “Nachhaltigkeit” behandelt, wird besser gefunden, wenn er auch verwandte Konzepte wie “Kreislaufwirtschaft”, “CO2-Neutralität” und “regenerative Energien” natürlich einbindet. Dabei zeigt sich, dass präzise formulierte FAQs besonders effektiv von den KI-Systemen verarbeitet werden können. Die natürliche Frage-Antwort-Struktur von FAQs entspricht dabei dem Dialogformat, das ChatGPT für seine Interaktionen nutzt. Dies führt zu genaueren und relevanteren Antworten für Endnutzer.

Webseitenbetreiber sollten ihre FAQ-Bereiche daher nicht nur für menschliche Besucher optimieren, sondern auch für ChatGPT und Co. Dies bedeutet konkret, dass Fragen und Antworten in natürlicher Sprache formuliert, regelmäßig aktualisiert und mit relevanten Kontextinformationen angereichert werden sollten. Besonders wichtig ist dabei, wie schon erwähnt, die Verwendung von präzisen Formulierungen und die Integration von Fachbegriffen, die von LLMs erkannt und korrekt interpretiert werden können.

**Tägliche Nutzung von KI**
nach Altersgruppen [8]






# Die Rolle des **Vektorraums**

Die Umwandlung von Daten in Vektoren erfolgt durch komplexe neuronale Netzwerke und spezialisierte Embedding-Modelle. Bei der Verarbeitung von Textdaten analysiert das System zunächst die linguistischen Eigenschaften wie Wortbedeutungen, grammatikalische Strukturen und kontextuelle Zusammenhänge, wobei jedes Wort oder jede Phrase in einen hochdimensionalen Zahlenraum übertragen wird, in dem ähnliche Konzepte nahe beieinander liegen.

Ein praktisches Beispiel hierfür ist die Verarbeitung des Wortes “Bank”, bei der das System je nach Kontext unterschiedliche Vektoren erzeugt - einmal für das Finanzinstitut und einmal für die Sitzgelegenheit. Diese Vektoren bestehen typischerweise aus mehreren hundert bis tausend Dimensionen, wobei jede Dimension einen bestimmten Aspekt der Bedeutung oder Eigenschaft repräsentiert. Moderne Embedding-Modelle verwenden dabei fortgeschrittene Transformer-Architekturen, die durch das Training an Millionen von Textdokumenten gelernt haben, diese semantischen Beziehungen präzise in Vektorform abzubilden.

Die resultierenden Vektoren ermöglichen es dann, Ähnlichkeitsberechnungen durchzuführen, indem mathematische Operationen wie die Kosinus-Ähnlichkeit angewendet werden. Hierdurch kann die inhaltliche Nähe verschiedener Bedeutungen oder die Relevanz von Suchanfragen bestimmt werden. Die Vektorisierung ermöglicht es somit LLMs, auch “ähnliche” Inhalte zu finden, selbst wenn diese nicht exakt die gleichen Begriffe verwenden. Dies ist ein entscheidender Vorteil gegenüber einer klassischen Keyword-basierter Suche.

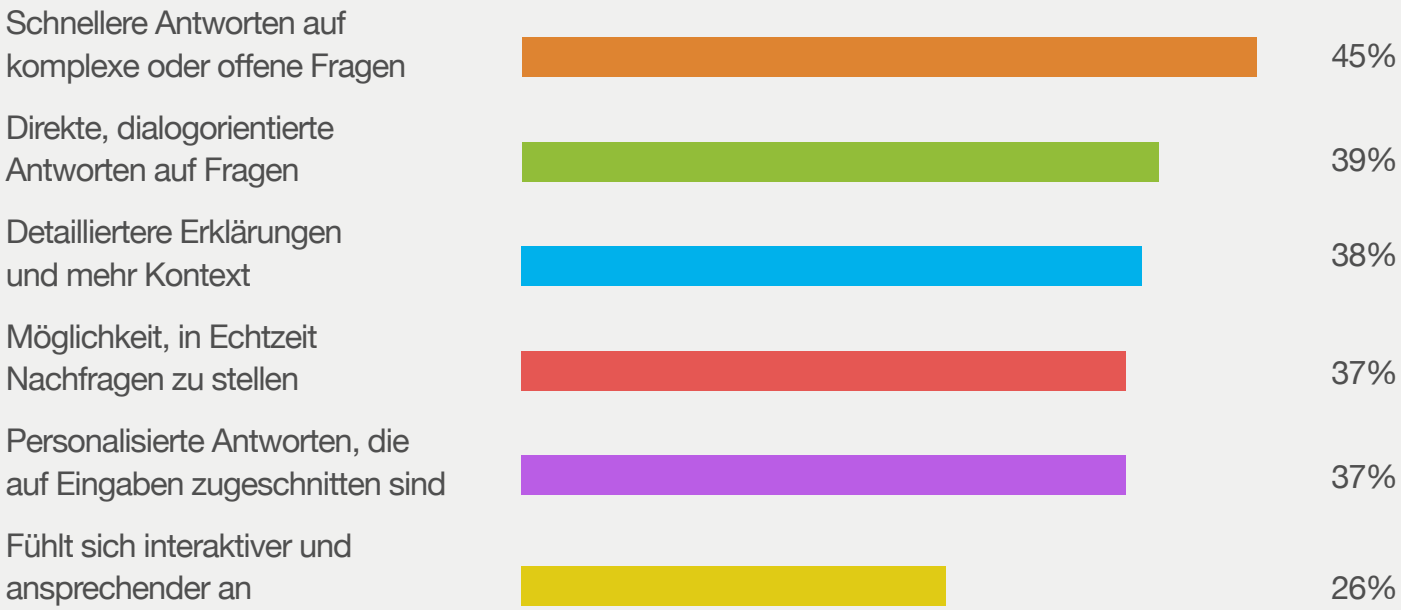
Fachspezifische und Kontext reiche Formulierungen erzeugen eindeutigere Vektormuster als allgemeine oder vage Ausdrücke. Besonders wichtig ist dabei die “semantische Dichte” - also wie viele bedeutungsvolle Informationen pro Textabschnitt in Vektoren umgewandelt werden können.

Die Vektorisierung von Webinhalten wird fundamental auch durch die strukturelle Organisation einer Website beeinflusst. Bei der Umwandlung von Webinhalten in Vektoren spielt die hierarchische Struktur eine entscheidende Rolle für die Qualität der resultierenden Vektordatenbank. Eine klare Hierarchie von Hauptthemen, Unterthemen und detaillierten Inhalten ermöglicht es den Vektorisierungsalgorithmen, semantische Beziehungen präziser zu erfassen und in den mehrdimensionalen Vektorraum zu übertragen.

Die Navigation einer Website fungiert dabei als semantischer Wegweiser. Gut strukturierte Navigationsebenen helfen bei der Erstellung von Embedding-Clustern, wobei thematisch verwandte Inhalte im Vektorraum näher beieinander positioniert werden. Dies verbessert die Effizienz von Ähnlichkeitssuchen in der Vektordatenbank erheblich. Beispielsweise werden Produktkategorien in E-Commerce-Systemen als zusammenhängende Cluster im Vektorraum abgebildet, was präzisere Empfehlungen und Suchergebnisse ermöglicht.

Die technische Strukturierung durch HTML-Semantik (wie <header>, <nav>, <main>, <article>) und eine saubere Hierarchie unterstützt die Vektorisierungs-Engines bei der Gewichtung von Inhalten. Hauptüberschriften (H1) und deren hierarchische Unterebenen (H2-H6) beeinflussen direkt die Dimensionalität und Gewichtung im Vektorraum. Eine logische URL-Struktur und Breadcrumb-Navigation tragen zusätzlich zur kontextuellen Einordnung bei, was sich in der Qualität der Vektorrepräsentationen widerspiegelt.

Was empfinden Sie als **besonders vorteilhaft bei der Nutzung von KI-Tools** wie ChatGPT oder Bard im Vergleich zu Google für Suchanfragen? [9]




# Redundante **Informationen**

Dopplungen, redundante Informationen und Datenmüll stellen eine große Herausforderung für die effektive Vektorisierung von Webinhalten in Vektordatenbanken dar. Wenn identische oder stark ähnliche Textpassagen mehrfach vektorisiert werden, entstehen überflüssige Vektoren, die den Vektorraum aufblähen und die Effizienz beeinträchtigen. Statt kompakter und präziser Vektor-Cluster bilden sich diffuse Wolken von Vektoren, die semantisch gleiche Informationen repräsentieren. Dies führt zu einer Verwässerung der Suchergebnisse und einer Verschlechterung der Antwortqualität von LLMs.

Ein konkretes Beispiel verdeutlicht das Problem: Wenn ein Onlineshop dieselbe Produktbeschreibung auf mehreren Unterseiten oder in verschiedenen Kategorien dupliziert, werden bei der Vektorisierung redundante Vektoren erzeugt. Eine Suchanfrage nach diesem Produkt liefert dann eine Vielzahl von Treffern mit identischen Informationen, statt eine präzise und diverse Ergebnismenge zu präsentieren. Dies beeinträchtigt nicht nur die Nutzererfahrung, sondern erschwert auch die Identifikation zusätzlicher relevanter Inhalte.

Datenmüll, wie veraltete oder irrelevante Informationen, verstärkt dieses Problem zusätzlich. Wenn obsolete Produktbeschreibungen, fehlerhafte Metadaten oder unzusammenhängende Textfragmente in die Vektorisierung einfließen, verschlechtern sie die Qualität des Vektorraums insgesamt. Die resultierenden Vektoren repräsentieren dann keine kohärenten semantischen Einheiten mehr, sondern ein Rauschen irrelevanter Informationen. Dies beeinträchtigt die Fähigkeit von LLMs, präzise Antworten zu generieren und relevante Inhalte zu identifizieren.

Um diese Probleme zu minimieren, ist eine sorgfältige Datenaufbereitung und Qualitätskontrolle vor der Vektorisierung unerlässlich. Duplikate sollten identifiziert und bereinigt, redundante Informationen zusammengefasst und Datenmüll entfernt werden. Nur durch eine konsistente und aussagekräftige Datenbasis kann eine effektive Vektorisierung erreicht werden, die präzise und hochwertige LLM-Antworten ermöglicht.

# **Barrierefrei**
# im Vorteil

Die Bewertung von ALT-Texten durch Large Language Models hat die Bedeutung von Barrierefreiheit im Web weiter erhöht. LLMs analysieren nicht nur die bloße Existenz von ALT-Texten, sondern bewerten auch deren qualitative Aussagekraft und kontextuelle Einbettung. Anders als traditionelle SEO-Tools, die hauptsächlich die Präsenz von ALT-Attributen prüfen, verstehen LLMs den semantischen Zusammenhang zwischen Bild und Beschreibung. Ein gut formulierter ALT-Text sollte dabei nicht nur das Bild beschreiben, sondern auch dessen Funktion und Bedeutung.

LLMs berücksichtigen dabei auch die Konsistenz der ALT-Texte mit dem umgebenden Content der Website. Ein ALT-Text, der thematisch zum Seiteninhalt passt und relevante Keywords natürlich einbindet, wird höher bewertet als isolierte oder generische Beschreibungen. Dies fördert einen ganzheitlichen Ansatz zur Barrierefreiheit, bei dem ALT-Texte nicht nur als Notwendigkeit, sondern als Vorteil bei LLM-Optimierung verstanden werden. Sie führen nicht nur zu einer verbesserten Zugänglichkeit für alle Nutzer, sondern unterstützen gleichzeitig die semantische Erschließung von Bildinhalten durch KI-Systeme.

Beispielsweise wird ein ALT-Text wie “Bild eines Autos in Mitternachtsblau vor einer Ladesäule” auf einer Ökostrom Seite von LLMs als weniger wertvoll eingestuft als “CO2 neutrales SUV Model X in Mitternachtsblau mit geöffneten Flügeltüren vor einer unserer Multi-Norm Ladestationen”. Die detaillierte Beschreibung ermöglicht es nicht nur sehbehinderten Nutzer*Innen den Inhalt besser zu verstehen, sondern hilft auch LLMs, den Kontext und die Relevanz des Bildes präziser zu erfassen. Diese verbesserte Kontexterfassung wirkt sich positiv auf die Vektorisierung aller Inhalte aus.

Und es gibt schon den ersten LLMO Hack: Im Podcast „Digital Kompakt“ [10] waren Hamid Hosseini und Paul Krauss von ECODYNAMICS eingeladen, die gerade eine aufwändige Studie zum Thema KI durchgeführt haben. Ein Ergebnis ihrer Untersuchungen: Websites, die einen Chatbot über Open AI API nutzen, werden in den „Ergebnissen“ von Chat GPT deutlich bevorzugt. Wir können außerdem nach unseren umfangreichen Tests bestätigen, dass bei der Vektorisierung von Websites fast immer nur die ersten Navigationsebenen berücksichtigt werden.

# Die netzleuchten
# **LLMO-Analyse**

Wir haben ein Verfahren entwickelt, das die wichtigsten Erkenntnisse und Studien aufgreift und umfangreiche eigene Test durchgeführt. Dies gibt uns die Möglichkeit das Optimierungspotenzial qualifiziert zu analysieren und zu bewerten. Das Verfahren beschränkt sich dabei auf eine Analyse der Vektoren, die nach der „Semantic Search“ an das Large Language Modell für die Generierung der Antworten übermittelt werden.

LLMs und Suchmaschinen mit KI-Funktionen nutzen unterschiedliche Verfahren. Daher sind die ermittelten Ergebnisse nur eine Annäherung und ein erster Einstieg in die Optimierung einer Website. Grundsätzlich lässt sich aber feststellen, dass von der Vektorisierung in vielen Fällen nur die ersten Ebenen der Website erfasst werden, anders als bei der klassischen Suche. Mögliche „Treffer“ sind auch dann nicht im Vektorraum vorhanden, wenn die Keywords im Inhalt der Website vorliegen.

## Das Verfahren

1. Vektorisierung der Website
2. Ermittlung der drei häufigsten Suchanfragen zur Website
3. Umwandlung der Suchbegriffe in das umgangssprachliche Frageformat
4. Prompting
5. Ermittlung und Analyse der gefundenen Vektoren
6. Vergleich der Ergebnisse
7. Bewertung

## Die Bewertung

1. Strukturierung und Klassifizierung
2. Semantische Klarheit
3. Hierarchische Übersichtlichkeit
4. Bildtexte/ALT-Texte
5. Füllworte und Dopplungen
6. Verständlichkeit für Laien
7. Menge der fehlenden Informationen

jeweils 1-5 Punkte

# **Es ist Zeit,** zu handeln

Die Auseinandersetzung mit KI-gestützten Suchsystemen und LLMs ist für Webseitenbetreiber nicht länger optional, sondern zwingend notwendig. Es ist aber wichtig zu betonen, dass die Optimierung für LLMs kein Ersatz für klassisches SEO ist, sondern vielmehr eine Ergänzung darstellt. Webseitenbetreiber sollten beide Ansätze kombinieren, um sowohl in traditionellen Suchmaschinen als auch in KI-gestützten Systemen gut aufgestellt zu sein.

**Hans Jörg Bordin**
Ihr Ansprechpartner für KI und LLMO

0431 - 785 895 05
bordin@netzleuchten.com

# Glossar

**Embedding Modell:** KI-System, das Wörter, Sätze oder ganze Texte in Zahlenreihen (Vektoren) umwandelt

**Vektor:** Ein Vektor ist eine Größe in der Mathematik und Physik, die sowohl eine Richtung als auch eine Länge hat und oft verwendet wird, um Bewegungen, Kräfte oder Positionen im Raum darzustellen.

**Transformer:** Der Mechanismus eines Transformers basiert auf der sogenannten „Self-Attention“, bei der das Modell analysiert, wie stark verschiedene Teile eines Satzes oder einer Datenstruktur miteinander verbunden sind, um relevante Informationen in einem Kontext zu priorisieren und besser zu verstehen.

**Kosinus-Ähnlichkeit:** Die Kosinus-Ähnlichkeit ist ein mathematisches Maß, das zeigt, wie ähnlich zwei Dinge (z. B. Texte oder Datenpunkte) sind, indem der Winkel zwischen ihren Darstellungen in einem Raum verglichen wird – je kleiner der Winkel, desto ähnlicher sind sie.

**Breadcrumb-Navigation:** Die Breadcrumb-Navigation ist eine Art „Spur aus Krümeln“ auf Webseiten, die zeigt, wo man sich gerade befindet, und es einfach macht, zu vorherigen Seiten oder Kategorien zurückzukehren.

**ALT-Attribute:** Das ALT-Attribut ist ein Text, der Bildern auf Webseiten hinzugefügt wird, um deren Inhalt zu beschreiben, was besonders nützlich für Menschen mit Sehbehinderung ist.

**Open AI API:** Schnittstelle, die es Entwicklern ermöglicht, leistungsstarke KI-Modelle wie ChatGPT oder DALL·E in ihre Anwendungen zu integrieren, um Aufgaben wie Textgenerierung, Übersetzungen, Bildgenerierung oder Datenanalyse effizient zu automatisieren.

**Semantic Search:** Suchtechnologie, die versucht, die Bedeutung von Suchanfragen zu verstehen, anstatt nur nach exakten Schlüsselwörtern zu suchen, und so relevantere Ergebnisse liefert, indem sie den Kontext und die Beziehungen zwischen den Wörtern berücksichtigt.

**Schema.org-Markup:** Standardisiertes Datenformat, das Webseiten-Betreiber verwenden können, um strukturierte Daten auf ihrer Website bereitzustellen.

# Quellen


(1) www.gartner.com/en/newsroom/press-releases/2024-02-19-gartner-predicts-search-engine-volume-will-drop...

(2) www.similarweb.com/blog/insights/ai-news/bing-chatgpt-ai-chat

(3) www.expressvpn.com/de/blog/social-media-search-tool/

(4) searchengineland.com/keyword-query-length-insights-445376

(5) de.statista.com/statistik/daten/studie/1401298/umfrage/durchschnittliche-nutzung-chat-gpt

(6) de.statista.com/statistik/daten/studie/1401309/umfrage/chat gtp-nutzung-in-unternehmen

(7) rankmagic.net/blog/die-zukunft-des-seo-wie-du-mit-voice-search-und-ki-deine-google-rankings-optimierst/

(8) Umfrage des TÜV-Verbands November 2024

(9) Umfrage 9.1.2025 Express VPN

(10) digitalkompakt.de/podcast

**netzleuchten GmbH**

www.netzleuchten.com
hallo@netzleuchten.com
0431 785 895 00

Mitglied von KI.SH

DiWiSH MITGLIED IM VEREIN DIGITALE WIRTSCHAFT SCHLESWIG-HOLSTEIN

Partner

**Schleswig-Holstein**
Der echte Norden