



FALLSTUDIE: OPTIMIERUNG UND ERSTELLUNG VON KI-CONTENT

Ein Experiment zur KI-basierten Optimierung und Lokalisierung von Inhalten

VIER GETESTETE KI-TOOLS

PROMPTS IN ZWEI SPRACHEN
VERFASST UND GEPRÜFT

DURCHSCHNITTLICHER QUALITÄTSWERT DEUTSCHEN CONTENTS, DER MIT ENGLISCHEN PROMPTS OPTIMIERT WURDE: 86 %

HINTERGRUND DES LIONBRIDGE-EXPERIMENTS

Lionbridge hat dieses Experiment mit seinem internen Team sowie eigenen Ressourcen und eigenem Content durchgeführt. Konkret haben wir deutschen Content aus existierenden Blogs verwendet. Obwohl bei dieser Fallstudie unser Content (vor allem deutscher Content) im Mittelpunkt stand, wollten wir ein Protokoll erstellen und Erkenntnisse sammeln, die für unsere Tausenden von Kunden in allen Branchen und Sprachen relevant sind. Wir werden die Ergebnisse und Beobachtungen aus diesem Experiment nutzen, um unsere Kunden in ihren vielfältigen Bereichen und mit ihren verschiedenen Sprachen zu unterstützen. Außerdem lassen wir diese Informationen in unsere internen Prozesse zur Erstellung und Optimierung von Marketingcontent einfließen.

DIE HERAUSFORDERUNG

Das Optimieren von Content, insbesondere in anderen Sprachen, kann zeitaufwendig und kostspielig sein. Hinzu kommt, dass die Übersetzungen vorhandenen Contents nicht immer im Hinblick auf die SEO-Werte optimiert wurden. In vielen Fällen sind die Übersetzungen zu literal. Sie wurden nicht unter Berücksichtigung der Feinheiten sprach- und regionalspezifischer Schlüsselwörter, des Benutzerverhaltens und/oder der richtigen Daten angefertigt. KI-Tools können uns helfen, mehrsprachigen Content zu optimieren, dabei Kosten und Arbeitsaufwand zu reduzieren und vorgegebene Fristen einzuhalten.

HINTERGRUND

Die vier KI-Tools Amazon Bedrock, ChatGPT (Lionbridge Azure-Instanz mit ChatGPT 4), Google Bard und Metas Llama können Content erstellen und optimieren.

Sie analysieren Suchmuster, Daten von Mitbewerbern und das allgemeine Nutzerverhalten, um datenbasierte Vorschläge für Schlüsselwörter und Content zu machen.

Die Tools gestalten optimierten Content, der möglichst weitgehend den Erwartungen der Benutzer entspricht, ansprechend ist und einen hohen SEO-Wert hat. Gelegentlich liefern sie besseren Text als Menschen.

KI-Tools sind oft ebenso gut wie Menschen in der Lage, strikte Längenbeschränkungen (Zeichenzahl) beim Schreiben und Übersetzen von Content einzuhalten (vor allem im Hinblick auf den SEO-Wert).

DIE LÖSUNG

Wir führten ein dreistufiges Experiment mit KI-Tools durch, um herauszufinden, welches für die Recherche und die Erstellung von Content – insbesondere in anderen Sprachen – am besten geeignet ist. Unser Team konzentrierte sich auf das Erlernen von Best Practices für die Entwicklung von Prompts, vor allem für mehrsprachigen Content. Wir testeten zudem Best Practices zur Optimierung älteren Contents in neuen Sprachen, um das SEO-Ranking einer Website zu verbessern.

METHODIK

Lionbridge setzte in dieser Studie vier KI-Tools ein: Amazon Bedrock, ChatGPT (Lionbridge Azure-Instanz, mit ChatGPT 4), Google Bard und Metas Llama. Wir haben drei Phasen – Recherche, Erstellung und Analyse – zweimal durchgeführt.

Recherche: In dieser Phase haben wir jedes KI-Tool aufgefordert, eine Keywordrecherche durchzuführen und die beiden besten Fragen für die Erstellung von Content auf Deutsch auszuwählen.

Wir haben bei jedem Tool diese Elemente eingegeben:

- Persona-Prompt mit der Anweisung, wie ein Verfasser deutschen Contents zu agieren: „provide 3 new German keywords and 3 German questions“.
- Zu optimierender Quellcontent
- Vorhandene Keywords

Außerdem haben wir zwei Zusatzfragen ausgewählt, die wir später für die Erstellung kleiner Absätze mit Content verwenden wollten.

Für die Keywordrecherche ist die Einbeziehung eines Menschen in den Prozess (als Human-in-the-Loop bezeichnet) unerlässlich. Dies gilt insbesondere für Bard. Das Tool liefert hervorragende Ergebnisse bei der Recherche, stellt jedoch zu viele Optionen bereit, sodass ein Experte entscheiden muss, welche Keywords oder Fragen für die Contenterstellung verwendet werden sollen.



Erstellung: In dieser Phase haben wir zwei Promptsets in deutscher Sprache entwickelt. Diese Prompts zur Erstellung von Content enthielten die im vorherigen Schritt ermittelten Keywords. Ein Set basierte dabei auf der Persona eines Content Writers. Das andere Set enthielt die Anweisungen.

Wir haben ein Teammitglied der KI-Abteilung von Lionbridge gebeten, alle Prompts zu überprüfen und zu optimieren, damit sie möglichst effizient sind. Nach dem Fertigstellen der Prompts (siehe unten im Abschnitt „Prompts“) nutzten wir die Ressourcen von Lionbridge, um diese ins Deutsche zu übersetzen. Wir ließen die beiden fertigen Sets deutscher Prompts durch alle vier KI-Tools laufen und sammelten und kennzeichneten den vorgeschlagenen Content.

Für die Entwicklung der Prompts haben wir folgenden Input verwendet:

- Fragen und Keywords aus der Recherchephase (Deutsch)
- Beispiele für guten Content, der als Vorbild dienen sollte (Deutsch)
- Text des Ausgangsartikels (Deutsch)
- Personaprompts (Englisch und Deutsch)
- Aufgabenprompts (Englisch und Deutsch)

Da wir zwei Promptsets hatten, mussten wir das Projekt in jedem KI-Tool zweimal ausführen.

Analyse: In dieser Phase haben wir zehn deutschsprachige Wirtschafts- und Sprachspezialisten gebeten, den Content zu überprüfen. Die Analyse wurde als Blindstudie durchgeführt: Die Spezialisten wussten nicht, welches Tool den jeweiligen Content erstellt hatte.

Wir fügten die KI-Ausgabe zur Bewertung in eine Tabelle ein, die zwei Bewertungsmechanismen aufwies:

- Drop-down-Menü mit den drei Werten „Unbrauchbar“, „Nach Bearbeitung brauchbar“ und „Ohne Bearbeitung brauchbar“
- Feld zur Bewertung des Contents von mit Werten von 1 – 100

Das Feedback wurde dann unter folgenden Aspekten ausgewertet:

- Welches Tool war am effektivsten?
- Welche Best Practices bieten sich zum Erstellen von Prompts an? Was gilt für die Erstellung von Prompts in anderen Sprachen?
- Wie gut eignen sich diese Tools für die Keywordrecherche?
- Wie gut generieren die Tools Fragen für den Content?
- Wie gut generieren die Tools Blickwinkel für die Artikel?
- Wie gut können die Tools die von ihnen generierten Fragen beantworten?
- Wie können wir sicherstellen, dass die KI-Prozesse zur Contenterstellung optimal funktionieren?
- Wie können effektive Prompts entwickelt werden?

PLANUNG



Auswahl von drei KI-Tools



Auswahl von drei zu optimierenden Artikeln für das Experiment



Auswahl der zehn Experten für die Analysephase



Recherchieren der Keywords und Fragen



Entwickeln der Prompts und Lokalisierung ins Deutsche

BEISPIELE FÜR VERWENDETE PROMPTS

Persona-Prompts

- Angenommen, Sie sind ein Content Writer für Lionbridge und Sie erstellen Inhalte im Zusammenhang mit Sprachdienstleistungen.
- Sie sind deutschsprachig und leben in Deutschland.
- Sie erstellen deutschsprachige Inhalte, die sich an Deutsch sprechende Menschen in Deutschland, Österreich und der Schweiz richten.

Zielgruppen-Prompts

- Personen, die im Bereich Marketing oder Kommunikation arbeiten
- Personen, die eine leitende Position haben, Senior Manager, Direktor, Vice-President, VP, Leiter/Leiterin, CMO, CEO, Geschäftsführer/Geschäftsführerin
- Sie sind in Deutschland, Österreich oder der Schweiz ansässig

Prompts für die Erstellung

- Bitte erstellen Sie 2 zusätzliche Textabsätze, die diesem Artikel angehängt werden sollen
- Jeder Absatz sollte eine Antwort auf folgende Fragen sein
- Das primäre Keyword für die URL lautet ...
- Die sekundären Keywords lauten ...

MULTIDISZIPLINÄRES TEAM

- 10 deutschsprachige Promptentwickler
- 1 Lionbridge-Experte für SEO
- 1 Lionbridge-Experte für digitales Marketing

DIE ERGEBNISSE

Künstliche Intelligenz wird elementarer Bestandteil jedes Optimierungsprojekts werden.

Unser Experiment hat bestätigt, dass KI-Tools bei der Optimierung und Aktualisierung von Content – auch in anderen Sprachen als Englisch – effektiv und effizient helfen können.

Erfreulicherweise verbesserte sich die Qualität der Ergebnisse im Laufe des Experiments. Die Einbeziehung eines Menschen in den Optimierungs- oder Erstellungsprozess von Inhalten ist dabei unerlässlich, weil nur Menschen die richtigen Schlüsselwörter für die Optimierung aus dem KI-Output auswählen können und den endgültigen Content validieren müssen. (Dies könnte sich in Zukunft ändern, da sich die Qualität der Ergebnisse zusehends verbessert.)

Wir haben die folgenden Best Practices ermittelt:

- Derzeit muss ein Mensch den endgültigen KI-Output validieren.
- Die Qualität unseres Inputs verbesserte sich im Laufe des Experiments, während wir Best Practices entwickelten.
- KI generiert mit zunehmender Wahrscheinlichkeit Content höherer Qualität, wenn Prompts folgenden Kriterien entsprechen:
 - Kurz
 - Content, der sich leicht aufschlüsseln lässt (z. B. Top-Ten-Listen usw.)
- Die Kommunikation mit der KI funktioniert derzeit in Englisch am besten. (Das kann nicht überraschen, weil die Tools mit englischen Daten trainiert werden.) Deutsche Prompts haben nicht dazu geführt, dass die KI-Tools besseren deutschen Content generierten. Tatsächlich generierten die Tools mit englischen Prompts besseren deutschen Content. Hinweis: In jedem Fall ist es aber hilfreich, Quellcontent, Keywords, Fragen und Beispiele in der jeweiligen Zielsprache bereitzustellen.
- Im Prozess der Contenterstellung sollte eine sorgfältige Beschreibung des Verfassers und der Zielgruppe enthalten sein. Sinnvoll ist zudem das Erstellen von Promptbibliotheken. Diese ermöglichen Benutzern die Wiederverwendung von Prompts und die Entwicklung konsistenten Contents für Kunden.

Bard lieferte wohl die besten Rechercheergebnisse.

Die Ergebnisse waren ähnlich gut wie die von MarketMuse, ein Tool, das speziell für diesen Zweck entwickelt wurde. Die Ergebnisse von Bard mussten jedoch von Menschen geprüft werden, die dann die „Perlen“ für die Contenterstellung auswählten. Mit deutschen Prompts arbeitete Bard über beide Phasen hinweg konsistent. Die mit englischen Prompts erzielten Ergebnisse fielen dagegen von 82/100 auf 66/100. Da es sich um einen kleinen Datensatz handelt, ergibt die Durchschnittsbewertung für Bard (in Englisch) 74/100.

Der Content von ChatGPT schnitt nicht ganz so gut ab, erhielt aber von den Korrektorlesern gelegentlich gute Bewertungen. Meta (Llama) lieferte unzureichende Ergebnisse und ist wohl noch nicht in der Lage, nicht-englischen Content zuverlässig zu optimieren oder zu generieren. Die Outputqualität schien sich im Lauf dieses kurzen, einmonatigen Experiments zu verbessern. Bard verbesserte seine Rechercheergebnisse und bei der Erstellung deutschen Contents konnte ChatGPT Fortschritte verzeichnen. Der ChatGPT-Output bei Verwendung deutscher Prompts lag zunächst bei 47/100.

Innerhalb von nur einem Monat lag das Ergebnis bereits bei 76/100. Für Output, der mit englischen Prompts erstellt wurde, lag der Wert zunächst bei 71/100 und stieg innerhalb eines Monats auf 86/100. (Der Durchschnittswert für Englisch beträgt 78,5/100.)

Sowohl Bard als auch ChatGPT konnten recht gut kurzen Content (z. B. nur ein oder zwei Sätze) erstellen. Hier gab es nur sehr wenige negative Rückmeldungen.

Die gezeigten Bewertungen beziehen sich auf längeren Content in zwei Phasen des Experiments.

ERGEBNISSE PHASE 1

Quelle	Unbrauchbar	Nach Bearbeitung brauchbar	Ohne Bearbeitung brauchbar	Durchschnittliche Bewertung
ChatGPT (Deutscher Input)	9	9	0	47.61
Bard (Deutscher Input)	2	14	2	68.33
ChatGPT (Englischer Input)	3	9	6	71.67
Bard (Englischer Input)	0	9	9	82.22
Meta (Llama) (Englischer Input)	7	8	3	46.50
GESAMT	21	49	20	

ERGEBNISSE PHASE 2

Quelle	Unbrauchbar	Nach Bearbeitung brauchbar	Ohne Bearbeitung brauchbar	Durchschnittliche Bewertung
ChatGPT (Deutscher Input)	1	7	0	76.25
Bard (Deutscher Input)	2	4	2	66.25
ChatGPT (Englischer Input)	0	4	4	86.875
Bard (Englischer Input)	3	4	1	59.375
GESAMT	5	19	7	

Hinweis: Meta (Llama) wurde in Phase 2 nicht getestet, weil das Tool Content in anderen Sprachen als Englisch nicht optimieren oder generieren kann.