AI CONTENT OPTIMIZATION AND GENERATION CASE STUDY

An experiment in Gen AI-powered content optimization and localization

4 AI TOOLS TESTED

PROMPTS WRITTEN AND TESTED IN 2 LANGUAGES

AVERAGE QUALITY SCORE OF GERMAN CONTENT OPTIMIZED WITH ENGLISH PROMPTS: 86%

BEHIND THIS LIONBRIDGE EXPERIMENT

Lionbridge performed this experiment with its internal team, resources, and content. Specifically, we used our German content: existing blogs. Though this case study focused on Lionbridge content (German in particular), the intention is to create protocol and gather learnings that can be applied to our thousands of customers across every vertical and language. We will use the results and observations from this experiment to serve customers with a wide variety of content needs and spoken languages. We'll also use this information in our own internal marketing content creation and optimization processes.

BACKGROUND

These four generative AI tools, Amazon Bedrock, ChatGPT (Lionbridge Azure instance, using ChatGPT 4), Google Bard, and Meta's Llama, all have the ability to create and optimize content.

They analyze search patterns, competitor data, and user behavior to make data-driven suggestions for keywords and content.

The tools help ensure optimized content is engaging and has strong SEO value, aligning as much as possible with user expectations. In some instances, it may even deliver the required copy better than humans do.

For example, GenAI tools are sometimes equally skilled as humans in keeping to the strict character limits of writing and translating content (especially for SEO value).

THE CHALLENGE

Optimizing content, especially in other languages, can be time-consuming and costly. Additionally, translations of existing content don't always perform at a top level for SEO. Sometimes, they are simply direct translations. They are not built with the nuances of language and region-specific keywords and user behavior and/or data. Generative AI tools might help us optimize multilingual content while reducing costs, timelines, and labor.

THE SOLUTION

We executed a three-phase experiment with GenAI tools to see which was most impactful for research and content generation, especially in other languages. Our team focused on learning best practices for developing prompts, especially for multilingual content. We also explored best practices for optimizing pieces of older content in new languages, which boosts a website's SEO.

METHODOLOGY

Lionbridge used four AI tools in this study: Amazon Bedrock, ChatGPT (Lionbridge Azure instance, using ChatGPT 4), Google Bard, and Meta's Llama. We ran our three phases twice: a research, creation, and analysis phase.

Research: In this phase, we asked each AI tool to complete keyword research and select the two best questions for creating content in German.

We fed each of the tools these items:

- A persona prompt, asking it to act like a German content writer and to "provide 3 new German keywords and 3 German questions."
- Source content to be optimized
- Existing keywords

Additionally, we chose two supplementary questions we'd later use to generate small paragraphs of content.

For keyword research, including a human in the loop, commonly referred to as human-in-the-loop, appears essential. This is especially true for Bard. The tool does excellent research, but provides too many options and requires a subject matter expert to select the keywords or questions necessary for content generation.

Creation: In this phase, we developed two sets of prompts in German. These content creation prompts included the keywords identified in the previous step. One set was based on the persona of a content writer. The other set was for the task instructions.

We asked a Lionbridge AI department member to review and help tweak all prompts for peak efficacy. Once we finalized the prompts (listed below in the "Prompts" section), we used Lionbridge's resources to localize the prompts into German. We ran these completed two sets of German prompts through all four AI tools, gathering and marking their suggested content.

We used these inputs to develop the prompts:

- Questions and Keywords from the research phase (German)
- Examples of good content we wanted to emulate (German)
- The text of the source article (German)
- Persona prompts (English and German)
- Task prompts (English and German)

Because we had two sets of prompts, we needed to run the project twice in each AI tool.

Analysis: In this phase, we asked ten German-speaking business and language specialists to review the content. To ensure a "blind review," we did not mark which tool generated content.

We added outputs for evaluation to a spreadsheet using two scoring mechanisms:

- A dropdown menu with three values: "Unusable," "Usable with Edits," and "Usable without Edits"
- A field to score the content from 1–100

We gathered the feedback and reviewed it to determine:

- Which tool was the most effective?
- What are some best practices for building prompts? What about building prompts in other languages?
- How well can these tools do keyword research?
- How well do the tools generate questions for the content?
- How well do the tools generate angles for the articles?
- How well do these tools create answers to the questions they generated?
- How can we ensure optimal AI content creation processes?
- How do you develop effective prompts?

PLANNING

Choosing 3 GenAI tools to use

Choosing 3 articles to optimize for the experiment

Finding 10 experts for the review panel

Researching for keywords and questions

Developing prompts and localizing them into German

EXAMPLES OF PROMPTS USED

Persona Prompts

- You are a content writer for Lionbridge, you create content related to language services.
- You are German-speaking and based in Germany.
- You create German language content, targeting German language speakers in Germany, Austria, and Switzerland...

Audience Prompts

- People who work in Marketing or communications
- People who have a senior position, Senior Manager, Director, Vice President, VP, President, CMO, CEO
- People who are German-speaking. They are based in Germany, Austria, or Switzerland...

Creation Prompts

- Please create 2 additional paragraphs of text, to append to this article...
- Each paragraph should be an answer to these questions...
- The primary keyword for the URL is...
- The secondary keywords are...

MULTIDISCIPLINARY TEAM

- 10 German-speaking Prompt Engineers
- 1 Lionbridge SEO Expert
- 1 Lionbridge Digital Marketing Expert

THE RESULTS

AI should become a critical element of any optimization project.
Our experiment confirmed that AI tools can help optimize and update existing content — even non-English content — effectively and efficiently.

Encouragingly, the quality of outputs improved over the course of the experiment. Including a human-in-the-loop of any content optimization or creation process is vital. They must be there to select keywords from the AI output for optimization. A human reviewer is also required to validate final content outputs. (This may change in the future, as the quality of outputs is noticeably improving.)

These are some best practices we identified:

- Currently, a human-in-the-loop must validate final AI outputs.
- The quality of our inputs advanced throughout the experiment as we learned new best practices.
- AI is more likely to yield high-quality content generation if requests are submitted as:
  - Shorter pieces
  - Content that can easily be broken down (such as top ten lists, etc.)
- Communicating with AI technology is currently most effective in English. (This isn't surprising, as the tools are being trained in English.) Giving German prompts didn't help AI tools generate better German content — notably, the tools produced better German content with English prompts. Note: it still helps to provide source content, keywords, questions, and examples, in the (non-English) target language.
- Content creation processes should utilize a thorough description of the content creator and target audience. Furthermore, it's ideal to build prompt libraries. These allow users to reuse prompts and develop consistent content for clients.

Bard seemed to provide the best outputs for research. Its results were similar to the caliber of research generated by a tool like MarketMuse, which is developed expressly for this purpose. Notably, Bard's research outputs require human intervention to select nuggets for the content creation part of the process. Bard, using German instructions, was static over the two phases. However, results forged from the English instructions went from 82/100 to 66/100. This was a small dataset, so averaging the scores for Bard (in English) results in 74/100.

ChatGPT's content garnered slightly more negative comments. However, reviewers also sometimes praised its content. Meta (Llama) didn't perform as well, and probably isn't ready yet to optimize or generate non-English content reliably. Output quality seemed to improve over the course of this short, month-long project. Bard improved its research outputs, and ChatGPT improved its German content creation. ChatGPT outputs originated by German instructions initially scored 47/100.

Within just a month, their outputs scored 76/100. For outputs produced by English instructions, outputs scored 71/100. Within a month, the score increased to 86/100. (The average score for English is 78.5/100.)

Both Bard and ChatGPT were skilled at producing short content (e.g., just one or two sentences). There was very minimal negative feedback on their short pieces.

The scores shown are for long-form content in the two phases of the experiment.

PHASE 1 SCORES

Source	Unusable	Usable With Edits	Usable Without Edits	Average Score
ChatGPT (German Input)	9	9	0	47.61
Bard (German Input)	2	14	2	68.33
ChatGPT (English Input)	3	9	6	71.67
Bard (English Input)	0	9	9	82.22
Meta (Llama) (English Input)	7	8	3	46.50
TOTAL	21	49	20	

PHASE 2 SCORES

Source	Unusable	Usable With Edits	Usable Without Edits	Average Score
ChatGPT (German Input)	1	7	0	76.25
Bard (German Input)	2	4	2	66.25
ChatGPT (English Input)	0	4	4	86.875
Bard (English Input)	3	4	1	59.375
TOTAL	5	19	7	

Note: Meta (Llama) was removed in Phase 2 because it's not ready yet to optimize or generate non-English content.