Generating Search-Engine-Optimized Headlines for Sports News

Frank Zalkow[1], Benedikt Schäfer[2], Thomas Moissl[1], Jonas Bücherl[1], Kerstin Markl[1], Sebastian Bothe[2], Francois Duchateau[2], Julia Dollase[2], Patric Kabus[3], Daniel Steinigen[4], Oliver Schmitt[5], Fabian Küch[1]

[1] Fraunhofer IIS, Erlangen, Germany, [2] Westdeutscher Rundfunk, Cologne, Germany,
[3] Hessischer Rundfunk, Frankfurt am Main, Germany,
[4] Fraunhofer IAIS, Sankt Augustin, Germany, [5] Südwestrundfunk, Stuttgart, Germany

Correspondence: frank.zalkow@iis.fraunhofer.de

Abstract

Search engines rank websites based on various features, such as headline wording, keyword usage, links, and site structure. Achieving a high rank is crucial for increasing website traffic, a key goal of search engine optimization (SEO). Automating content adjustments for improved SEO can significantly enhance workflows in media publishing. In this paper, we explore the use of large language models (LLMs) to automatically generate SEO-optimized headlines and toplines for German sports news articles. We compare the results from medium- and large-sized LLMs, both finetuned and non-finetuned, against headlines crafted by journalists. Our evaluation, based on a survey with SEO experts, reveals that finetuning is crucial for effectiveness and that medium-sized LLMs perform well in this task. These findings suggest promising opportunities for optimizing workflows in online media publishing.

1 Introduction

Search engine optimization (SEO) is a critical strategy for enhancing a website's visibility and increasing traffic by improving its ranking in search engine results. Search engines use various features of a website to rank it, given a relevant search query, including headline wording, keyword usage, links, and site structure.

Automatically adjusting a website's content for better SEO suitability presents an excellent opportunity to improve the workflows within media publishing houses for three reasons. First, not all editorial departments employ SEO specialists who can revise the journalistic output for the target publication medium. Second, the increasing demand for SEO expertise can overwhelm existing specialists, especially during peak content production periods. Third, optimizing content for SEO can be regarded as a initial step toward automatic content adaptation, which may lead to adapting content for different target audiences, potentially tailoring content for diverse audiences and further boosting website traffic.

In this paper, we explore the application of large language models (LLMs) to automatically generate SEO-optimized head- and toplines for German sports news articles. We assess several models, including large, closed-source models not specifically finetuned for this task, such as Gemini (Gemini Team Google et al., 2024), GPT-4 (OpenAI et al., 2023), and GPT-4o (OpenAI et al., 2024). We also evaluate a version of GPT-4o finetuned for the specific task, and two different task-specific finetuned versions of the medium-sized, open-source Teuken-7B-Instruct model (Ali et al., 2024). The results generated from these models are compared against human-written head- and toplines created by journalists and SEO experts.

Our application scenario focuses on the German sports website Sportschau[1], where SEO is a key growth factor since a significant portion of their reach comes from users who discover the content through search engines. We conduct a survey with SEO experts from the Sportschau department to assess the models' effectiveness for the given task. Our results indicate that task-specific finetuning is essential and that medium-sized LLMs are well-suited for this task. These findings open up possibilities for improved workflows in online media publishing houses, enabling editors to automatically select SEO-optimized headlines, regardless of their depth of SEO knowledge.

While several studies have adressed the generation of headlines (Ding et al., 2023; Gu et al., 2020) and related tasks like extreme summarization (Narayan et al., 2018), only few works expicitely focused on the SEO-compatibility of generated headlines. Unlike our research, machine learning has been previously applied in the SEO domain for tasks such as web page classification (Shaffi

[1] https://www.sportschau.de

Proceedings of the 21st Conference on Natural Language Processing (KONVENS), Volume 1: Long and Short Papers, pages 59–65, 2025

and Muthulakshmi, 2022; Matošević et al., 2021) or analysis (Roumeliotis and Tselikas, 2023), content optimization through commercial tools like Clearscope[2] or MarketMuse[3], automatic page creation (Jie et al., 2022), and enhancing search engine functionality (Xiong et al., 2024). More related to our work, a few previous studies have explored the generation of SEO-optimized titles for web pages. For instance, Mathur et al. (2018) focused on the large-scale production of headlines in the e-commerce sector, where manually crafting titles is impracticale for quantitative reasons. In contrast to their approach, our goal is to generate high-quality headlines that closely resemble those written by human SEO experts. Anastasiu et al. (2021) pursued a similar objective to ours but employed a multi-step process. They used a BERT-based model to create a one-sentence summary of an article, followed by generating a list of potential keywords, and finally combining these elements. In contrast, our approach is a data-driven, end-to-end method that leverages the capabilities of modern pre-trained LLMs. Some preliminary steps of our experiments have been described in a Master's thesis (Moissl, 2024).

The rest of the paper is structured as follows. In Section 2, we detail the task addressed in our paper. Next, we explain how we used and finetuned the models in Section 3. Then, we describe how we set up the survey in Section 4. Next, in Section 5, we discuss our results. Finally, we close with a short summary in Section 6.

2 Task

The task approached in this paper is to generate SEO-optimized headlines and toplines for German sports news articles. The headline serves as the primary entry point for an article, designed to convey the article's key message and integrate relevant keywords to maximize visibility for both readers and search engines. The topline, positioned above the headline, provides additional context and plays a crucial role in helping both users and search engines in understanding the article's relevance, often by including the competition name, event, or sport category. Together, the topline and headline form the technical headline used by search engines for indexing.

The key challenge of the task is to craft concise, informative, truthful, and engaging headlines that achieve high search engine rankings, without resorting to clickbait or sensationalism. At the same time, they should capture user interest and encourage readership. Specifically, the objective of SEO is to ensure that a user can immediately discern whether their search intent is met by a given search result. Unlike print headlines, which can rely on accompanying images and layout to attract attention, digital headlines must function independently in search results, mobile news feeds, and app notifications.

An essential consideration is the target tone of the publication medium. In our case study, we aim to meet the tone of the German sports news website Sportschau, known for its reputable style with high linguistic and journalistic standards, as opposed to more tabloid-style sports websites, which presents a challenge: On the one hand, the content needs to remain discoverable and competitive in search rankings. On the other hand, the aim is to maintain quality-driven, fact-based, and serious journalism, rather than being led by short-term SEO trends or sensational language.

Character count limitations are also a critical aspect for SEO headlines. While we acknowledge this factor, we do not address it beyond prompting (see Appendix A) and the typical lengths in our training corpus. Although we conducted internal experiments to shorten the generated LLM outputs with an additional postprocessing step utilizing another LLM call, this aspect is not the focus of our paper.

3 Models

In this section, we present the models employed in our experiments, with an overview provided in Table 1.

It is well-known that modern state-of-the-art LLMs exhibit robust performance on unseen tasks and are well-suited for zero-shot inference (Bubeck et al., 2023). Therefore, we utilize such large commercial models in a zero-shot setting, specifically Gemini, GPT-4, and GPT-4o. The simple prompts used for these models are given in Appendix A.

To tailor the models for our specific task, we compiled a dataset comprising 3576 web articles from the German sports news website Sportschau[4], covering diverse sports genres such as soccer, cricket, and tennis. During dataset curation, we

[2]https://www.clearscope.io
[3]https://www.marketmuse.com
[4]https://www.sportschau.de

Short Name	Model	Version	Open Source	Finetuned
`Gemini_Orig`	Gemini	gemini-1.5-pro-001	✗	✗
`GPT-4_Orig`	GPT-4	0613	✗	✗
`GPT-4o_Orig`	GPT-4o	2024-05-13	✗	✗
`GPT-4o_Tuned`	GPT-4o	2024-08-06	✗	✓ (API)
`Teuken_LoRA`	Teuken-7B-Instruct	commercial-v0.4	✓	✓ (LoRA)
`Teuken_Full`	Teuken-7B-Instruct	commercial-v0.4	✓	✓ (Full)

Table 1: Models used in the experiments of this paper.

ensured only articles with headlines reviewed or revised by SEO experts were included (having internal advice from the Sportschau department).

The collected articles serve as training data for finetuning. First, we employ OpenAI's commercial finetuning API to develop a finetuned version of GPT-4o. While some specifications of the finetuning procedure remain undisclosed, the documentation and job task description indicate that the training is conducted over three epochs using a LoRA-based approach (Hu et al., 2022) with a batch size of three.

In addition to large commercial models, we use a medium-sized open-source model, which is convenient for local experimentation, on-premise deployment, and guarantueed data security. We decided on the recently published Teuken model (Ali et al., 2024) that has a multilingual focus on Europe's linguistic diversity and particularly strong performance in the German language. We use two different finetuning procedures: a full finetuning of all model parameters and a parameter-efficient LoRA approach (Hu et al., 2022). Both finetunings are conducted in an instruction-tuning setting, where the task description and the provided article (i.e., the instruction) serve as context, and only the expected outputs from the models (i.e., the headline) are used for computing the loss (Zhang et al., 2023). In both cases, we train for 100 epochs, using a linear learning rate schedule with warmup, an initial learning rate of 10^{-5}, an effective batch size of 512, and the AdamW optimizer (Loshchilov and Hutter, 2019). For our LoRA finetuning, we use a rank of 32 (corresponding to roughly 0.17 % of the parameters, equating approximately to 12.8 M), an α value of 32, and a LoRA dropout rate of 0.05.

4 Survey

We conducted a survey, where we asked members from the Sportschau team to rate several top- and headline pairs for a given sports news article, which was excluded from the training set used for finetuning. In particular, we asked the participants to rate each pair on a discrete scale from 1 (very poor) to 5 (very good), where the primary focus should be on the SEO quality. Additionally, they were instructed to also consider other factors such as the journalistic alignment with the article and the seriousness and orientation of their publication medium. In cases where these criteria conflicted, the participants were encouraged to rely on their subjective judgment to assign an appropriate score. We are aware that this rating task is multi-dimensional and challenging. However, since the participants were experts who actually had a real use case for the system, we believe they could provide meaningful scores that reflect real-life usefulness overall. To give a better impression of the survey content, we show one entry in Appendix B.

Nine Sportschau members participated in the survey. Among them, five self-described as having deep SEO knowledge, three as having basic SEO knowledge, and one as having no SEO knowledge. Each participant reviewed ten articles of representative sports genres, where the distribution of sports genres was determined in collaboration with the Sportschau department. Specifically, the survey included articles on basketball, cycle sport, European handball (two articles), Formula One, Olympic Games, and soccer (four articles). For each article, the participants rated seven top- and headline pairs. Six of them were generated by the models (see Section 3) and one served as a reference, which was human-written and reviewed or revised by an expert on SEO. The participants were not informed about which pair was the reference or which pairs were generated by the models.

5 Results

Figure 1 presents boxplots along with corresponding raw histograms for the scores obtained from our survey. The reference obtained an average score of 3.533, positioned between "fair" and "good." In contrast, the commercial non-finetuned mod-

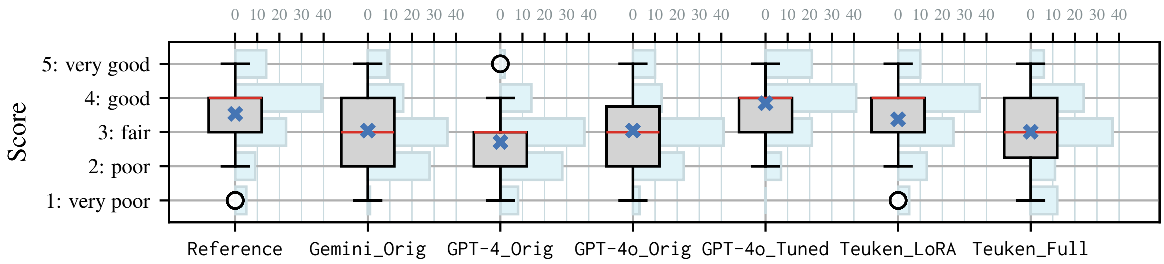

Figure 1: Visualization of the score distributions obtained in our survey: boxplots (gray boxes) with median values (red lines), mean values (dark blue crosses), and underlying raw scores (vertical histograms in light blue).

els Gemini (`Gemini_Orig`), GPT-4 (`GPT-4_Orig`), and GPT-4o (`GPT-4o_Tuned`) received lower average scores of 3.044, 2.711, and 3.044, respectively. The finetuned version of GPT-4o (`GPT-4o_Tuned`) obtained an excellent average score of 3.844, which is even better than that of the reference. Among the finetuned versions of the Teuken model, the LoRA-finetuned variant (`Teuken_LoRA`) yielded a good score of 3.378, and the full finetuning (`Teuken_Full`) scored 3.011. These results suggest that our approach to the task is promising since several models received scores close to the reference's scores, where the output of the finetuned version of GPT-4o was rated even better than the human-written headlines. It is also noteworthy that one finetuned version of the medium-sized Teuken model outperformed all non-finetuned commercial models. A potential reason for the superior performance of the LoRA finetuning compared to the full finetuning of the Teuken model is that our dataset's size was insufficient for a full finetuning.

We conducted a significance test to further analyze the relevance of the differences among the scores. Since the scores are non-normally distributed, we conducted pairwise comparisons using the Wilcoxon signed-rank test with Bonferroni correction. Specifically, we applied the correction by adjusting the p-values ($p_{\text{Bonferroni}} = k \cdot p_{\text{orig}}$, where $k = 21$ is the number of comparisons). The test evaluates the null hypothesis that two pairs of ratings come from the same distribution. Table 2 presents the adjusted p-values, where low values indicate pairs of significantly different score distributions. The most critical entries of the table are in the first row, comparing the scores produced by the models with those by the reference. Using a standard significance threshold of $p \leq 0.05$, the scores for Gemini, GPT-4, GPT-4o, and the fully finetuned version of Teuken are considered significantly different from the scores of the reference. Conversely, the scores of the finetuned version of GPT-4o and the LoRA-finetuned version of Teuken are both considered to be not significantly different from the scores of the reference, suggesting that these models produce headlines perceived as comparable to human-written ones by our expert evaluators.

6 Summary

In this paper, we explored the non-standard task of generating SEO-optimized head- and toplines for journalistic web articles. We conducted a case study where we utilized various large language models, including some that were finetuned with use-case-specific data, to generate such headlines. To assess the models' performance, we conducted a survey in which expert participants rated both the model-generated and human-written headlines. Our results show that finetuning is essential for achieving optimal performance. Furthermore, both large commercial models and medium-sized open-

	Reference	Gemini_Orig	GPT-4_Orig	GPT-4o_Orig	GPT-4o_Tuned	Teuken_LoRA	Teuken_Full
Reference	1.000	0.026	0.000	0.033	0.305	1.000	0.011
Gemini_Orig		1.000	0.058	1.000	0.000	0.578	1.000
GPT-4_Orig			1.000	0.124	0.000	0.000	0.841
GPT-4o_Orig				1.000	0.000	0.402	1.000
GPT-4o_Tuned					1.000	0.018	0.000
Teuken_LoRA						1.000	1.000
Teuken_Full							1.000

Table 2: p-values obtained by our significance test. Cells with $p \leq 0.05$. are highlighted.

source models can achieve a strong performance level producing headlines comparable to those written by human experts.

Acknowledgments

This work was funded by the German Federal Ministry for Economic Affairs and Climate Action (BMWK) through the project OpenGPT-X (project no. 68GX21007D). The authors gratefully acknowledge the scientific support and HPC resources provided by the Erlangen National High Performance Computing Center (NHR@FAU) of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) and the NHR Center of TU Dresden (NHR@TUD). The hardware by NHR@FAU is funded by the German Research Foundation (DFG) and the NHR@TUD is supported by the Federal Ministry of Education and Research and the state governments participating in the NHR. Furthermore, we thank all participants of our survey.

References

M. Ali, M. Fromm, K. Thellmann, J. Ebert, A. A. Weber, R. Rutmann, C. Jain, M. Lübbering, D. Steinigen, J. Leveling, K. Klug, J. S. Buschhoff, L. Jurkschat, H. Abdelwahab, B. J. Stein, et al. 2024. Teuken-7B-Base & Teuken-7B-Instruct: Towards European LLMs. *arXiv CoRR*, abs/2410.03730.

C. Anastasiu, H. Behnke, S. Lück, V. Malesevic, A. Najmi, and J. Poveda-Panter. 2021. DeepTitle – leveraging BERT to generate search engine optimized headlines. *arXiv CoRR*, abs/2107.10935.

S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. M. Lundberg, H. Nori, H. Palangi, M. T. Ribeiro, and Y. Zhang. 2023. Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv CoRR*, abs/2303.12712.

Z. Ding, A. Smith-Renner, W. Zhang, J. R. Tetreault, and A. Jaimes. 2023. Harnessing the power of LLMs: Evaluating human-AI text co-creation through the lens of news headline generation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3321–3339, Singapore.

Gemini Team Google, P. Georgiev, V. I. Lei, R. Burnell, L. Bai, A. Gulati, G. Tanzer, D. Vincent, Z. Pan, S. Wang, S. Mariooryad, Y. Ding, X. Geng, F. Alcober, R. Frostig, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv CoRR*, abs/2403.05530.

X. Gu, Y. Mao, J. Han, J. Liu, Y. Wu, C. Yu, D. Finnie, H. Yu, J. Zhai, and N. Zukoski. 2020. Generating representative headlines for news stories. In *Proceedings of the Web Conference (WWW)*, pages 1773–1784, Taipei, Taiwan.

E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. 2022. LoRA: Low-rank adaptation of large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, virtual.

C. Jie, D. Xu, Z. Wang, and W. Shen. 2022. Deep learning based page creation for improving e-commerce organic search traffic. *arXiv CoRR*, abs/2209.10792.

I. Loshchilov and F. Hutter. 2019. Decoupled weight decay regularization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, New Orleans, LA, USA.

P. Mathur, N. Ueffing, and G. Leusch. 2018. Multilingual neural title generation for e-commerce browse pages. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 162–169, New Orleans, LA, USA.

G. Matošević, J. Dobša, and D. Mladenić. 2021. Using machine learning for web page classification in search engine optimization. *Future Internet*, 13(1).

T. Moissl. 2024. Adapting large language models for the application in media use cases. Master's thesis, Technical University of Darmstadt.

S. Narayan, S. B. Cohen, and M. Lapata. 2018. Don't give me the details, just the summary! Topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1797–1807, Brussels, Belgium.

OpenAI, J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, R. Avila, I. Babuschkin, S. Balaji, V. Balcom, et al. 2023. GPT-4 technical report. *arXiv CoRR*, abs/2303.08774.

OpenAI, A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow, A. Welihinda, A. Hayes, A. Radford, A. Mądry, A. Baker-Whitcomb, A. Beutel, A. Borzunov, et al. 2024. GPT-4o system card. *arXiv CoRR*, abs/2410.21276.

K. I. Roumeliotis and N. D. Tselikas. 2023. A machine learning Python-based search engine optimization audit software. *Informatics*, 10(3).

S. S. Shaffi and I. Muthulakshmi. 2022. Search engine optimization by using machine learning for web page classification. In *Proceedings of the International Conference on Augmented Intelligence and Sustainable Systems (ICAISS)*, pages 342–349, Trichy, India.

H. Xiong, J. Bian, Y. Li, X. Li, M. Du, S. Wang, D. Yin, and S. Helal. 2024. When search engine services meet large language models: Visions and challenges. *IEEE Transactions on Services Computing*, 17(6):4558–4577.

S. Zhang, L. Dong, X. Li, S. Zhang, X. Sun, S. Wang, J. Li, R. Hu, T. Zhang, F. Wu, and G. Wang. 2023. Instruction tuning for large language models: A survey. *arXiv CoRR*, abs/2308.10792.

A Prompts

Here, we briefly list the prompts used in our experiments. Note that the string "`{input_txt}`" is to be replaced by the article for which the head- and topline is going to be generated.

The following prompt was used for `Gemini_Orig`, `GPT-4_Orig`, `GPT-4o_Orig`, and `GPT-4o_Tuned`.

```
Verfassen Sie eine suchmaschinenoptimierte Dachzeile und Überschrift für den folgenden Artikel. Nutzen Sie dabei maximal 25 Zeichen für die Dachzeile und 45 Zeichen für die Überschrift. Verwenden Sie das Format
Dachzeile: <Dachzeile hier>
SEO-Überschrift: <Überschrift hier>

Artikel:

{input_txt}
```

The following prompt was used for `Teuken_LoRA` and `Teuken_Full`.

```
System: Ein Gespräch zwischen einem Menschen und einem Assistenten mit künstlicher Intelligenz. Der Assistent gibt hilfreiche und höfliche Antworten auf die Fragen des Menschen.
User: Schreibe eine SEO-optimierte Dachzeile und Überschrift für folgenden Artikel.

{input_txt}
Assistant:<s>
```

B Survey Example

In this appendix section, we provide an example entry from the survey. Each survey page included an article, a task description (identical across all pages), and the top- and headline pairs to rate.

Below is an example of an article:

```
[Dachzeile]
[Überschrift]

Der 1. FC Heidenheim hat seine Aufstiegsambitionen im letzten Hinrundenspiel der 2. Fußball-Bundesliga mit einem glücklichen Sieg gegen Jahn Regensburg nach einer wilden Schlussphase untermauert.

Das Team von Trainer Frank Schmidt gewann in einem höchst unterhaltsamen Spiel gegen Regensburg mit 5:4 (3:2) und geht als Tabellendritter in die lange WM- und Winterpause.

Spektakel in der Nachspielzeit

Stefan Schimmer (90.+4) setzte in der Nachspielzeit den Schlusspunkt, zwei Minuten zuvor hatte Regensburgs Aygün Yildirim (90.+2) das 4:4 erzielt. Zuvor hatten Tim Kleindienst (21./39.), Adrian Beck (36.) und Denis Thomalla (76.) für Heidenheim sowie Prince Osei Owusu (14.), Charalambos Makridis (45.+1) und Nicklas Shipnoski (55.) für den Jahn getroffen.

In einem Duell mit vielen Chancen auf beiden Seiten waren die Gastgeber die etwas effektivere Mannschaft und hatten das glückliche Ende auf ihrer Seite. Die Regensburger von Trainer Mersad Selimbegovic verpassten es, ihre Negativbilanz gegen Heidenheim aufzubessern. Regensburg hatte vor 9.695 Zuschauern einige große Chancen: Doch Blendi Idrizi (60.) traf nur den Pfosten. Makridis schoss aus sechs Metern am leeren Tor vorbei.
```

Model	**Top- and Headline**	**Average Score**
`Gemini_Orig`	Heidenheim siegt im Spektakel 5:4! Heidenheim dreht nach Rückstand auf und feiert Sieg im Krimi	3.000
`GPT-4_Orig`	Sieg für Heidenheim 1. FC Heidenheim untermauert Aufstiegsambitionen mit Sieg	2.889
`GPT-4o_Orig`	2. Bundesliga Heidenheim siegt dramatisch 5:4 gegen Regensburg	3.889
`GPT-4o_Tuned`	Neun Tore Heidenheim gewinnt turbulentes Spiel gegen Regensburg	4.333
`Teuken_LoRA`	2. Bundesliga Heidenheim besiegt Regensburg in wilder Schlussphase	3.778
`Teuken_Full`	Sieg gegen Jahn Heidenheim mit Happy End gegen Regensburg	3.222
`Reference`	Siegtor in der Nachspielzeit Heidenheim entscheidet Spektakel gegen Regensburg für sich	3.667

Table 3: Example top- and headlines from our survey.

```

Heidenheim nach der Pause gegen Rostock

Am 18. Spieltag geht es für Heidenheim zu Hause gegen
↪  Rostock um wichtige Punkte im Aufstiegskampf (Samstag,
↪  28.01.2023 um 13.00 Uhr). Regensburg ist zum Auftakt
↪  der Rückrunde in Darmstadt gefordert.
```

Following the article, participants received the task description:

```
Bitte bewerten Sie die Qualität folgender Dachzeilen und
↪  Überschriften auf einer Skala von 1 bis 5. Dabei steht
↪  1 für "sehr mangelhaft" und 5 für "sehr gut". Im
↪  Zentrum Ihrer Bewertung sollte die
↪  Suchmaschinenoptimierung stehen, allerdings ohne
↪  Beachtung der typischen Längenbeschränkungen für die
↪  Textelemente. Beziehen Sie auch weitere Kriterien, wie
↪  die journalistische Passgenauigkeit zum Artikel sowie
↪  die Seriösität und Ausrichtung Ihres
↪  Publikationsmediums ein. Sollten sich diese Kriterien
↪  widersprechen, verlassen Sie sich auf Ihr subjektives
↪  Empfinden, um eine angemessene Punktzahl zu vergeben.
```

Finally, the pairs of top- and headline pairs for rating was shown to the participants. The order of these pairs was randomized, and the models used to generate them was not disclosed to the participants. In Table 3, we present these pairs along with the average scores these particular entries received in our survey.