

POSTPRINT

Christian Lang
Roman Schneider
(corresponding author)
Ngoc Duyen Tanja Tu

Leibniz Institute for the German Language, Mannheim, Germany
lang.schneider,tu@ids-mannheim.de

Automatic Question Answering for the Linguistic Domain – An Evaluation of LLM Knowledge Base Extension with RAG

Abstract:

We investigate the extent to which Retrieval Augmented Generation improves the quality of Large Language Models' answers to technical questions in the field of linguistics—a domain known for its broad terminological inventory and theory-dependent use of technical terms. Furthermore, this application is not only about terminological information on language, but also about information on its well-formedness. We present the results of an empirical evaluation of automatically generated answers based on authentic data from a language consulting service, with special emphasis on different question types.

Keywords:

question answering; large language model; retrieval augmented generation; quality evaluation; domain specificity

1 Introduction

Large Language Models (LLMs) have been applied to a broad range of NLP-related tasks, such as automated Question Answering (QA). Among other things, the use of QA systems appears promising for the education sector—for example to create low-threshold natural-language access to specialist content in teaching, learning, and information systems. Our contribution takes the renowned information system on German grammar and orthography, *grammis* [19], as a use case.

When using automated QA systems for educational purposes, it is immanently vital to avoid misinformation and misunderstandings. In their current state, however, pre-trained LLMs suffer from gaps in their underlying resources. Such gaps arise if certain information is not widely accessible for training, be it for legal reasons or because it belongs to a niche topic. When LLMs are asked about such expert knowledge, they often produce answers that seem convincing at first glance, but turn out to be vague, contradictory or even made-up on closer inspection—so-called hallucinations [23]. This is due to the fact that LLMs have no semantic knowledge in a true sense, but merely predict words.

So their ability to answer queries correctly depends on the textual coverage of the subject and its particular terminology in the underlying training data.

Combining LLMs with domain-specific knowledge that goes beyond the original training data seems to be a key to prevent hallucinations and thus misinformation. One approach is *Retrieval Augmented Generation* (RAG) [20], i.e. combining the generative capability of an LLM with a specialized knowledge base. One of the main advantages of RAG is that it allow to determine the materials on which an answer is based, therefore it does not work like a black box.

In our contribution, we evaluate how expanding an LLM with RAG improves QA quality. We take into account that in linguistics, natural language is not only a means of communication, but also the object of investigation. One consequence is that questions about linguistic content may relate not only to terminological definitions (e.g. *What is a definite object?*), but also to concrete language objects, their position in a communication unit, and the well-formedness of language (e.g. *Is the following expression correct? What is the definite object in the following sentence?*).

For widespread practical reasons, we implement the QA system on a local machine: Textual data required for model extension is often legally protected, especially in specialized domains, and may not be made publicly available or uploaded to external servers. This is the case with parts of the data we use for our knowledge base (see Sect. 3.1). By focusing on local deployment, we provide basic research for similar application contexts.

Subject of our study are German-language texts; findings should be transferable to other natural languages.

2 Related Work

Within deep learning, QA is obviously one prominent focus of interest, with a huge ecosystem of datasets and models proposed [1, 6, 18, 25]. The integration into learning activities has also already come into focus [4, 15]. Along with this, quality benchmarks have been proposed in order to study performance [9, 10], to investigate the impact of punctuation [17], and to understand to which extent models learn commonsense knowledge [13].

Pre-trained models have been investigated regarding In-Context Learning (ICL), i.e. regarding their ability to consider given examples and to learn from analogy [3, 16]. Since this can be conducted without direct model access, ICL seems related to (hard) prompt tuning, i.e., to the modification of input questions in hope to improve answering quality [14].

Large-scale models, such as those from the GPT or LLaMA family, are trained on general domain data, so their performance for specialized domains is not necessarily good. Two very different approaches to enhance LLMs with domain-specific competence are (1) Retrieval Augmented Generation (RAG, [12]) and (2) Parameter-Efficient Fine-Tuning (PEFT, [8]). With regard to often limited local resources, and as an alternative to full fine-tuning, PEFT improves

models in a computational efficient manner by only considering a small number of parameters. One PEFT approach is Low Rank Adaption (LoRA). It decomposes the weight matrices through low-rank approximation into two smaller matrices. These can then be fine-tuned to domain-specific data. Both, the adapted weights and the original weights, are merged for the final result.

While (pre-trained and fine-tuned) LLMs only generate answers based on their parametric knowledge (which always will be outdated at some point), RAG adds an external knowledge base that is customizable without further training, and leads to reduced hallucinations [20]. Besides, annotators seem to find RAG-improved answers more factual and specific in comparison to the parametric BART model [12]. What is more, RAG offers transparency with regard to the underlying materials. We build upon these studies, and especially evaluate RAG.

3 Implementation

As initial model, we choose Llama 2 (Large Language Model Meta AI, [22]) as a powerful open source alternative to openAI’s GPT 3.5, the state of the art model at the time of our evaluation. We use the 4bit quantized¹ 13B model version in order to enable local deployment² and acceptable response times.

Using a domain-specific text collection (cf. Sect. 3.1), we implement RAG (cf. Sect. 3.2) and evaluate the answering quality by determining inter-annotator agreement (cf. Sects. 3.3 and 4). We lay out our work in such a way that it can be replicated locally, and make the annotated evaluation dataset publicly accessible under Creative Commons License (CC BY-NC).

3.1 Data Set for Extending an LLM for the Linguistic Domain

To extend Llama 2 with linguistic knowledge, we use a scientifically sound domain-specific dataset which covers a variety of linguistic sub-areas (grammar, spelling, punctuation, word formation, etc.). Dataset compilation is based on our use case and considers two sources:

(1) Linguistic literature and online materials from *grammis* (grammis.ids-mannheim.de). 78 % of these high-quality contents address specialists, 22 % are intended for a wider audience [19].

(2) Anonymized question-answer pairs from a commercial language consulting service. The questions are asked by (supposed) laypersons and answered by professionals; we concatenate questions and answers as joint data points.

All texts (with the obvious exception of the question part in the question-answer pairs) are based on scientific evidence, include specialized terminology,

¹ Quantization reduces the model size by reducing the precision of the data types that represent the model weights (e.g. 16-bit floating point to 4-bit integer). We prefer the quantized version of the 13B model over the non-quantized version of the smaller 7B model, as exploratory tests show better results for the larger model even with quantization.

² Ryzen 3600X, 16GB of RAM and an Nvidia RTX 3090 with 24GB of VRAM.

and have been either directly written or content-checked by experts. Table 1 indicates the number of texts, the total number of tokens, and the average number of tokens per text, broken down by source. The question-answer pairs make up the majority of the dataset; we deal with different average token numbers by segmenting all texts to chunks of 1,000 characters (see Sect. 3.2).³

Table 1. Extension dataset.

| Source | Texts | Tokens | Tokens/Text |
|----------------------------|--------|-----------|-------------|
| <i>Contents grammis</i> | 1,775 | 1,006,021 | 566.77 |
| <i>Language consulting</i> | 45,912 | 6,425,253 | 139.95 |

3.2 Model Extension and Application

We apply LocalGPT⁴, a Python-based toolkit with LangChain as its main component, to implement RAG locally. The pipeline is structured as follows:

(1) Instructor-xl [21] embeds texts in our data set (cf. Sect. 3.1) minus the 24 evaluation questions (cf. Sect. 3.3). Texts are segmented with a chunk size of 1,000 characters and an overlap of 200 characters. This model is not specialized and achieves state-of-the-art results in many downstream embedding tasks. Its biggest advantage is that no fine-tuning is required for different tasks and domains, because it is instruction-based fine-tuned, i.e. every input is embedded along with task instructions.

(2) We store the embedded data from (1) in a local vector database, which functions as knowledge base. Whenever a question is asked, it is also embedded and used to find the closest vectors in the database, according to the cosine similarity.⁵ Thus, we get the most related sequences, which can then be passed as context to Llama, together with the original input.

Figure 1 schematically shows the process of answering a user query with RAG, i.e. utilizing an external knowledge base. It illustrates with an authentic example—that concerns a rather niche topic of German grammar—how extending the initial input with relevant information retrieved from the knowledge base can improve the answer’s accuracy; the answer without using the knowledge base (bottom right in Fig. 1) seems convincing at first glance, but is actually incorrect. On the other hand, the answer that was created with the help of the knowledge base (bottom left in Fig. 1) is correct.

³ While the specialist literature is edited throughout, spelling and punctuation in the authentic questions are not consistently correct.

⁴ <https://github.com/PromptEngineer/localGPT>.

⁵ A suitable way to include lexical similarity would be to use the linguistic terms from the inputs for the comparison. However, terms used by laypeople are often ‘vague’ terms like *sentence* [11] and therefore not useful for a lexical comparison.

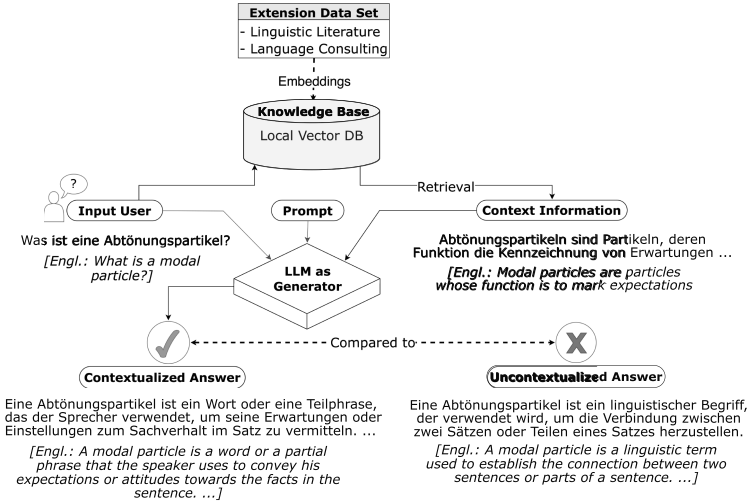


Fig. 1. Schematic representation of RAG

In addition to RAG, we also experimented with QLoRa fine-tuning [5] by applying the Text Generation Web UI⁶ and using the aforementioned dataset. An initial evaluation of the results shows that this approach produces significantly poorer results compared to RAG.⁷ For this reason, we do not provide a detailed description of this flanking approach.

3.3 Evaluation Dataset

To evaluate whether and how a RAG extension increases answer accuracy, we use 24 questions based on the language consulting data described in Sect. 3.1. We remove these 24 questions from the knowledge base to avoid a direct overlap with the evaluation questions.

In order to stay close to the domain characteristics (see Sect. 1), we distinguish between two question types, and evaluate 12 questions per type:

- **Usage questions** concern concrete language objects and their formal correctness, status or function, e.g.: *Schreibt man "Online Shop" mit oder ohne Bindestrich?* [Engl.: *Do you write "online store" with or without a hyphen?*]

⁶ <https://github.com/oobabooga/text-generation-webui>.

⁷ However, we observe that the language model imitates linguistic mannerisms of the human experts in the data set as a result of QLoRA fine-tuning. The model uses greeting formulas that are commonly used in emails and refers to books and dictionaries of the publishing house, which provided the language consultant service included in the data set (these products are partly fictional and partly appear in the data set).

- **Definition questions** concern the clarification of linguistic concepts and their function, e.g.: *Was sind starke und schwache Verben?* [Engl.: *What are strong and weak verbs?*]

Within usage questions, language objects play the central role (e.g. “online store” above), as the question refers to their correctness, spelling, function, etc. Sometimes, usage questions also contain technical terms (e.g. “hyphen” above). In definition questions, on the other hand, terminology plays the central role (such as “strong verb”/“weak verb” above); in some cases, they also contain linguistic objects for further illustration.

The two question types can be roughly associated with the two target groups that a linguistic QA system is intended to cover: Definition questions seem more typical for domain specialists (this is suggested by taking a closer look on search behavior in *grammis*), while usage questions are more typical for the general public (this is suggested by the language consulting data). Regarding syntax and semantics, definition questions are open-ended questions, typically starting with “wh” words and asking for an explanation, assessment, or example. Usage questions are often binary, close-ended questions that can be basically answered with yes or no, or multiple choices between two or more variants.

4 Quality Evaluation

We compare the answering performance of Llama 2 13B in the 4-bit quantized variant in two scenarios: with the extension by an external knowledge base with RAG (henceforth *4bit RAG*) and without such an extension (henceforth *4bit*).⁸ The answering performance of *4bit* serves as a baseline for *4bit RAG*. As additional baseline, we also include a non quantized version of Llama 2 13B (henceforth *no quant*).

Quality evaluations of natural language have content and language aspects to consider, and are prone to subjective judgments. Empirical studies therefore ideally work with several human raters. The assessment of our automatically generated answers is conducted by six unpaid native speakers with linguistic education. They receive the question list, the answers generated by each of the models, and adequate annotation guidelines; the list of questions does not contain any information that makes it possible to assign individual answers to a specific model. In order to provide some orientation, annotators are also provided with an exemplary correct answer for each question, originally given by human experts. Each of the 24 answers is evaluated based on five attributes. Attribute values range from 1 to 3 and express increasing quality (cf. Table 2).

The attribute *Content Correctness* is of particular importance in terms of avoiding misinformation. *Explanation* reflects whether information necessary for understanding the answer is provided (for example, in the case of a spelling

⁸ We use the GGML format model files (*llama-2-13b-chat.ggmlv3.q4_0.bin*) from <https://huggingface.co/TheBloke/Llama-2-13B-chat-GGML>.

Table 2. Annotation scheme for evaluation, values increasing from 1 to 3 express better quality.

| Attribute | Description | Possible value |
|-------------------------------|---|----------------------------------|
| <i>Content Correctness</i> | Is the answer correct in content? | 1: no 2: partly 3: yes |
| <i>Explanation</i> | Does the answer contain an explanation? | 1: no 2: short 3: detailed |
| <i>Irrelevant Information</i> | Does the answer contain irrelevant information that has nothing to do with the actual question? | 1: many 2: some 3: none |
| <i>Language Correctness</i> | Does the answer contain grammatical errors or nonsense words? | 1: many 2: some 3: none |
| <i>English Elements</i> | Does the answer contain incongruous/unmotivated English words or phrases? | 1: many 2: some 3: none |

question, motivating factors for the correct spelling variant). *Irrelevant Information*, on the other hand, indicates whether the answer contains contexts that are irrelevant to the question. Linguistic quality is assessed using the attributes *Language Correctness* and *English Elements*. Llama 2 was trained primarily on English data [22], as a consequence we observe occasional switches to English, even though both knowledge base and input are in German. In such cases, a response may be correct in terms of English grammar, but not regarding the desired output language, which is why we evaluate these two criteria separately.

4.1 Results

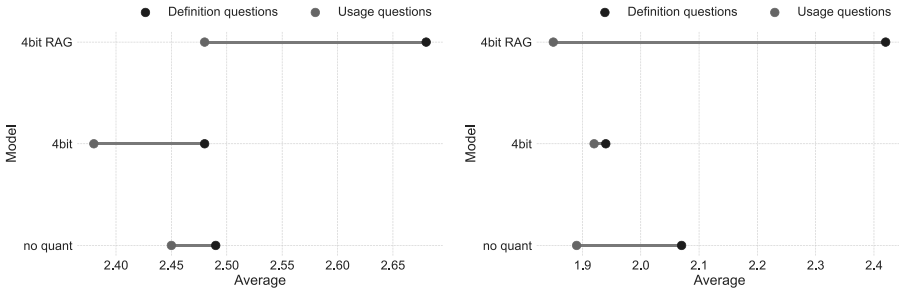
First, we compute inter-annotator agreement using Fleiss’s kappa [7], since we treat all disagreements as equally severe. In order to detect possible influences of the different question types [2], we additionally distinguish between definition questions and usage questions. The separated scores as shown in Table 3 indicate substantial agreement according to the interpretation criteria presented in [24].

Second, we assess the models’ performance by calculating the average of the annotations for the generated answers by each model. Here again, we distinguish between definition questions and usage questions. The results are shown in Fig. 2.

The left side of Fig. 2 shows the average score for each model over all attributes for definition questions and usage questions. We see that *4bit RAG* performs substantially better than its baselines. This is true for both definition and usage questions. However, we discover a difference between the question

Table 3. Inter-Annotator-Agreement scores across all attributes and answers

| Questions data set | Fleiss’s kappa |
|---|----------------|
| <i>Overall agreement (Definition + Usage Questions)</i> | 0.69 |
| <i>Definition Questions</i> | 0.66 |
| <i>Usage Questions</i> | 0.72 |

**Fig. 2.** Left: Average score for each model over all attributes for definition questions and usage questions. Right: Average score for each model over attribute “Content correctness” for definition questions and usage questions

types: Answers to definition questions are consistently rated higher than answers to usage questions. This is particularly pronounced in the RAG-extended model.

Differences between question types become even clearer with a differentiated view of the attributes. The right side of Fig. 2 shows the average score for each model for the attribute *Content Correctness* only. Again, we identify a particularly pronounced difference between definition and usage questions for *4bit RAG*. While definition questions are answered significantly better by the RAG-extended model, this does not apply to usage questions. Rather, *4bit RAG* shows slightly worse ratings with regard to content correctness.

4.2 Discussion

Our results show that for definition questions the overall quality—and in particular content correctness—benefit greatly from a RAG extension. However, the answering of usage questions seems not to benefit in any particular way. On the contrary, the results suggest that when applied to usage questions, the RAG extension leads to poorer answers.

This seems in line with a conceptual difference between question types and the way in which relevant information is retrieved from the knowledge base: Definition questions contain a salient term that verbalizes the concept whose meaning/function etc. is to be explained (e.g. *Was ist eine **Abtönungspartikel**?* [Engl.: *What is a **modal particle**?*

questions, on the other hand, often do not contain a corresponding salient element. And even if they do, the correct interpretation of the question requires consideration of the language objects provided—for instance, the assessment of whether it is correct or not (e.g. *Schreibt man “Online Shop” mit oder ohne Bindestrich?* [Engl.: *Do you write “online store” with or without a hyphen?*]). Our knowledge base implementation (as a collection of text chunks) and the retrieval process based on text similarities reasonably cannot achieve this—unless the exact language object is part of the knowledge base. As a consequence, the retrieved “relevant information” is often not relevant at all for the question, which explains why the ratings for “Content Correctness” are worse for *4bit RAG* compared to the baselines.

5 Conclusion and Outlook

Our results show that RAG is capable to improve answers to definition questions, without sacrificing substantial quality criteria. Nonetheless, further research should cover different basis LLM and key figures for the retrieval of relevant information. RAG thus shows potential for making specialized knowledge available to QA systems, without resource-consuming LLM rebuilds, and as a viable solution for protected content that must not leave a local environment.

Given that no improvement in the quality of answers can be observed for usage questions, our results also suggest that the question type has an influence on the added value. Nevertheless, we believe that even for usage questions, RAG can enhance quality: Based on the remarks in Sect. 4.2, this requires additional knowledge resources and a different way of retrieving relevant information. In particular, we should make productive use of language objects when querying for relevant information. This would require several steps for both pre-processing and retrieval: (1) Identification of language objects contained in the questions.⁹ (2) Classification of the identified language objects (as there is a potentially infinite amount of language objects, this needs to be done on an abstract level, e.g. morphosyntactic structure).¹⁰ (3) Querying of relevant information depending on the classified language objects from suitable knowledge bases (this includes, e.g., specialized corpora for spelling issues or case variants).

Different types of usage questions and different language objects require different knowledge bases. Therefore, a systematic solution would greatly benefit from further investigation of question types and, in particular, the nature of the language objects they contain.

⁹ Based on the nature of the authentic questions dataset described in Sect. 3.1, we cannot assume that language objects are always reliably marked as such (e.g., by the consistent use of quotation marks). Thus, rule-based identification approaches should be complemented by other methods, and we see promising results from the application of machine learning methods, such as deep active learning.

¹⁰ We are currently building a comprehensive collection of language objects with further meta-information based on the dataset described in Sect. 3.1.

We believe that pursuing such research is a worthwhile endeavor, even if it seems plausible that with the further development of LLMs the risk of misinformation on niche knowledge and usage questions will also decrease. Nevertheless, RAG—in contrast to LLMs—promotes transparency and traceability (with regard to the material on which the answers are based), and enables a flexible and comparatively straightforward adaptation of knowledge bases.

References

1. Abdel-Nabi, H., Awajan, A., Ali, M.: Deep learning-based question answering: a survey. *Knowl. Inf. Syst.* **65**(4), 1–87 (2023)
2. Artstein, R.: Inter-annotator agreement. In: Pustejovsky, J., Ide, N. (eds.) *Handbook of Linguistic Annotation*, vol. 1, pp. 297–313. Springer, Dordrecht (2017). https://doi.org/10.1007/978-94-024-0881-2_11
3. Dai, D., Sun, Y., Dong, L., Hao, Y., Sui, Z., Wei, F.: Why can GPT learn in-context? Language models secretly perform gradient descent as meta optimizers. In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 4005–4019 (2023)
4. Denny, P., Sarsa, S., Hellas, A., Leinonen, J.: *Robosourcing Educational Resources – Leveraging Large Language Models for Learnersourcing* (2022)
5. Dettmers, T., Pagnoni, A., Holtzman, A., Zettlemoyer, L.: QLoRA: efficient fine-tuning of quantized LLMs (2023)
6. Etezadi, R., Shamsfard, M.: The state of the art in open domain complex question answering: a survey. *Appl. Intell.* **53**(4), 4124–4144 (2022)
7. Fleiss, J.L.: Measuring nominal scale agreement among many raters. *Psychol. Bull.* **76**(5), 378–382 (1971)
8. Houlisby, N., et al.: Parameter-efficient transfer learning for NLP (2019). arXiv Version Number: 2
9. Kamaloo, E., Dziri, N., Clarke, C., Rafei, D.: Evaluating open-domain question answering in the era of large language models. In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, vol. 1: Long Papers, pp. 5591–5606. Association for Computational Linguistics, Toronto (2023)
10. Kwiatkowski, T., et al.: Natural questions: a benchmark for question answering research. *Trans. Assoc. Comput. Linguist.* **7**, 453–466 (2019)
11. Lang, C., Tu, N.D.T., Zeidler, L.: Making non-normalized content retrievable – a tagging pipeline for a corpus of expert–Layperson texts. In: *Proceedings of the 4th Conference on Language, Data and Knowledge*, pp. 239–244. NOVA CLUNL, Portugal (2023)
12. Lewis, P., et al.: Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks (2020). arXiv Version Number: 4
13. Li, X.L., Kuncoro, A., Hoffmann, J., de Masson d’Autume, C., Blunsom, P., Nematzadeh, A.: A systematic investigation of commonsense knowledge in large language models. In: Goldberg, Y., Kozareva, Z., Zhang, Y. (eds.) *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11838–11855. Association for Computational Linguistics, Abu Dhabi (2022)
14. Liu, X., et al.: P-tuning: prompt tuning can be comparable to fine-tuning across scales and tasks. In: Muresan, S., Nakov, P., Villavicencio, A. (eds.) *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, vol. 2: Short Papers, pp. 61–68. Association for Computational Linguistics, Dublin (2022)

15. MacNeil, S., et al.: Automatically generating CS learning materials with large language models. In: SIGCSE 2023: Proceedings of the 54th ACM Technical Symposium on Computer Science Education, vol. 2. p. 1176. Association for Computing Machinery, New York (2023)
16. Mosbach, M., Pimentel, T., Ravfogel, S., Klakow, D., Elazar, Y.: Few-shot fine-tuning vs. in-context learning: a fair comparison and evaluation. In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) Findings of the Association for Computational Linguistics: ACL 2023, pp. 12284–12314. Association for Computational Linguistics, Toronto (2023)
17. Rocca, R., de la Vega, A.: Evaluating the role of non-lexical markers in GPT-2’s language modeling behavior. In: Deutsch, D., Udomcharoenchaikit, C., Opitz, J., Gao, Y., Fomicheva, M., Eger, S. (eds.) Proceedings of the 3rd Workshop on Evaluation and Comparison of NLP Systems, pp. 96–102. Association for Computational Linguistics (2022)
18. Rogers, A., Gardner, M., Augenstein, I.: QA Dataset explosion: a taxonomy of NLP resources for question answering and reading comprehension. *ACM Comput. Surv.* **55**(10), 1–45 (2023)
19. Schneider, R., Lang, C.: Das grammatische Informationssystem grammis – Inhalte, Anwendungen und Perspektiven. *Zeitschrift für germanistische Linguistik* **50**(2), 407–427 (2022)
20. Shuster, K., Poff, S., Chen, M., Kiela, D., Weston, J.: Retrieval augmentation reduces Hallucination in conversation. In: Moens, M.F., Huang, X., Specia, L., Wen-tau Yih, S. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2021, pp. 3784–3803. Association for Computational Linguistics, Punta Cana (2021)
21. Su, H., et al.: One embedder, any task: instruction-finetuned text embeddings. In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) Findings of the Association for Computational Linguistics: ACL 2023, pp. 1102–1121. Association for Computational Linguistics, Toronto (2023)
22. Touvron, H., et al.: Llama 2: open foundation and fine-tuned chat models (2023)
23. Tu, N.D.T.: “Hallo ChatGPT, ist das Komma in folgendem Satz richtig?” — Können leistungsstarke Chatbots traditionelle Sprachberatung ersetzen? Staats- und Universitätsbibliothek Göttingen, DHD-Blog – Digital Humanities im deutschsprachigen Raum (2023)
24. Viera, A.J., Garrett, J.M.: Understanding interobserver agreement: the Kappa statistic. *Fam. Med.* **37**(5), 360–363 (2005)
25. Zhao, W.X., et al.: A Survey of Large Language Models (2023)