FH AACHEN
UNIVERSITY OF APPLIED SCIENCES

Fachhochschule Aachen
Campus Jülich

Fachbereich 9: Medizintechnik und Technomathematik
Studiengang: Angewandte Mathematik und Informatik

Contrasting Llama-2 and LeoLM Models for Speaker Attribution in German Parliamentary Discourse: Investigating Model Efficacy and Language-Specific Prompt Influence

Seminararbeit

von

Eric Thor

Erstprüfer:	Prof. Dr. rer. nat. Stephan Bialonski
Zweitprüfer:	Patrick Gustav Blaneck B. Sc.
Matrikelnummer:	3529472

Aachen, December 27, 2023

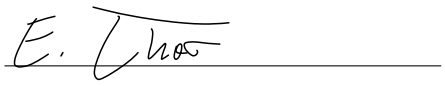
Eidesstattliche Erklärung

Ich versichere hiermit, dass ich die vorliegende Seminararbeit mit dem Thema

> *Contrasting Llama-2 and LeoLM Models for Speaker Attribution in German Parliamentary Discourse: Investigating Model Efficacy and Language-Specific Prompt Influence*

selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe, wobei ich alle wörtlichen und sinngemäßen Zitate als solche gekennzeichnet habe. Die Arbeit wurde bisher keiner anderen Prüfungsbehörde vorgelegt und auch nicht veröffentlicht.

Aachen, den December 27, 2023



Eric Thor

Abstract

The release of ChatGPT has created significant public discourse surrounding Large Language Models (LLMs). This period marked a significant milestone in the history of artificial intelligence, as it witnessed a substantial increase in the discourse of the topic. The amount of models available for fine-tuning rises continuously, predominantly caused by research advancements in the USA. Consequently, these models exhibit a primary proficiency in English, with other languages always receiving secondary consideration in their development. This development raises a difficult question: can monolingual models, those fine-tuned for a specific language, offer a robust and efficient alternative to address the limitations posed by the primary focus on English in multilingual models? This situation poses a considerable challenge due to the scarcity of monolingual models, datasets, and data sources currently available in the market. This is further exaggerated by a lack of research dedicated to evaluating their performance. This work addresses this question through an analysis of the multilingual foundation model Llama-2 and the model LeoLM, which is the first German monolingual derivation of Llama-2, published by LAION from Hamburg. In the analysis, this work specifically focuses on their performance in the task of speaker attribution in German parliamentary debates. The findings shed a light on the future importance of developing monolingual models and offer an insight into their comparative effectiveness, particularly in tasks that involve data exclusively in the language specific to the monolingual model. The study concludes that, in the current landscape, monolingual LLMs do not demonstrate a substantial performance advantage, when compared to multilingual models.

Contents

1. Introduction

Since the introduction of the transformer architecture [1] in 2017, the concept of Self-attention, integrated within most-recent LLMs, allows the computation of individual tokens as well as contextual interdependencies within a sentence. In contrast to Long Short-Term Memory (LSTM) [2] networks, which similarly to transformer-based LLMs can utilize information from adjacent tokens, these LLMs also have the capability for parallel processing. LLMs gained considerable public attention upon the unveiling of ChatGPT in November 2022 [3], signifying a crucial moment in their development and subsequent widespread adoption across diverse sectors, e.g. education [4] and medicine [5]. Concurrently, while the weights and architectures of many LLMs remain unknown to the public, other openly accessible models such as Llama-2 [6] and Falcon [7] have been made available to the public domain. These open-source models offer versatility, enabling users to adapt them according to the specific requirements of individual tasks or applications, e.g. for lexical simplification [8] and detecting predatory sexual behavior in online chats [9].The quantity of available texts has significantly expanded in recent years due to advancements in data collection methodologies [10] and advancing digitalization [11].
According to [12], political texts are usually inherently unstructured. This presents a big challenge in terms of automated analysis, rendering the process notably tedious. One potential methodology designed to tackle this challenge involves the integration of automated speaker attribution, as defined in the context of the GermEval 2023 [13] competition. This approach enables the development of a computational model that can attribute the specific interactions in dialogue, characterized by identifying the speaker, the addressee, and the content of the communication, within a predefined context. The organizers provided a dataset containing German parliamentary debate speeches, serving as the basis for testing and evaluating the aforementioned analysis method-

ology in this work. A team of researchers at Aachen University of Applied Sciences has achieved the second place at this competition [12] using rigorous preprocessing, postprocessing and a QLoRA [14] fine-tuning of Llama-2. In the latter half of 2023, a monolingual model, denoted as LeoLM [15], was published. This model, a fine-tuned iteration derived from the foundational Llama-2 [6] architecture, was specifically trained using an exclusive corpus of German language data [16]. This work is going to compare the effects of fine-tuning multilingual and monolingual models on task of automated speaker attribution and whether it makes a difference if the model prompts are formulated in German or English. The foundation models used are LeoLM and Llama-2.

2. Data & Methods

To answer this research question, we use a dataset of German news articles and parliamentary debates [17]. Cues are components within a sentence that signal the occurrence of a speech event. Roles, in this context, refer to the specific elements of the speech event, including attributes such as the speaker, the addressee, the message, and more. Within the training dataset, roles and cues have previously been extracted. The cues and roles undergo a prompt engineering process wherein they are structured and formatted to create model prompts, ensuring they qualify as viable input data for our model. In contrast, the test dataset necessitates the final model to autonomously extract these cues and roles as part of its evaluative task. For tuning our state-of-the-art Large-Language-Models, we use Quantized Low Rank Adaptation [14] (QLoRA)-based fine-tuning. In this work, separate models are used for the extraction of cues and roles, respectively, and the output of the cues model is utilized for the input of the roles model

2.1. Data

The dataset was provided by the organizers of the GermEval 2023 Shared Task on Speaker Attribution in German News Articles and Parliamentary Debates. It comprises 267 speeches extracted from the German parliament, originating from six political factions (namely: CDU/CSU, SPD, AfD, FDP, DIE LINKE, BÜNDNIS 90/DIE GRÜNEN), and independent speakers [13]. The entire dataset is partitioned into three distinct datasets labeled as Dev, Train, and Eval as seen in table 2.1. Each text underwent an automated segmentation process into sentence-like structures using the SpaCy [18] tool [19]. Each sentence then underwent segmentation, separating its components into words

and punctuation marks [13]. Human annotators subsequently assigned either zero, one, or multiple annotations in adherence to the annotation guidelines [13]. The Trial dataset is entirely within the Train dataset, thus it is not displayed in table 2.1 [12].

Split	Speeches	Sentences	Annotations
Train	177	9093	5399
Dev	18	927	515
Eval	72	3067	1792
Total	267	13087	7706

Table 2.1.: The dataset's speech, sentence, and annotation counts, as given by the organizers [13]. Each speech consists of several structures, that contain roles and cues.

The annotations consist of cues and roles, and are to be interpreted according to a detailed annotation guideline, provided by the organizers of GermEval 2023 [13]. A cue within a sentence refers to the presence of specific lexical items or linguistic constructs that serve as indicators, signaling the representation of speech, writing, or cognitive processes [13]. These cues occur in various linguistic forms, such as phrases, words, or syntactical structures, effectively marking the inception or reiteration of communicated content within the given discourse. The cue is underlined in the following examples: "She didn't say much", "Merkel told the people", "He proposed to change the law".
The roles consist of the source, medium, message, topic, addressee, and evidence. The source pertains to the individual from whom the message originates. For a role to be observed, it requires the presence of one identifiable cue that is linked to it. The medium functions as the container, conveying the message. The message is the fundamental essence, defining "what is articulated" within the communication. The topic contains the information regarding the subject of the statement. The addressee represents the designated recipient within the communication process. The evidence, akin to the medium, serves as a channel of transmission. For instance, a written statement or document can function not only as a means of expression but also as evidence for a claim.

Role	Examples
Source	"She didn't say much"
	"Merkel told the people"
Medium	"It said on Twitter that"
	"On television he said"
Message	"She didn't say much"
	"He suggested postponing the topic"
Topic	"They talked about finances"
	"We discussed the impact of the war"
Addressee	"They talked about finances"
	"We discussed the impact of the war"
Evidence	"The law states that it is impossible"
	"The statistic clearly proves that ..."

Table 2.2.: Examples from the annotation guide [13] showcasing individual annotations. They are Representing distinct roles are underlined within the corresponding parts of the sentence. Two examples per role.

Each individual excerpt does not have to contain every role.

2.2. Methods

This work employs models from the Llama-2 family, specifically Llama-2 by Meta as a multilingual foundation model [6] and LeoLM [15] as a monolingual foundation model by LAION. These models undergo training utilizing QLoRA [14]. To assess the potential impact of prompt language on model performance, this work also implements prompt translation from English to German, for the cue and role extraction prompts, utilizing the DeepL translation service [20] for this purpose.
Furthermore, the analysis looks at different model sizes, namely the 7B, 13B, and 70B versions, to understand how model parameter amount affects performance outcomes.
In [13], the shared task was structured into two distinct subtasks. Subtask 1, termed "Full Annotation", focused on comprehensive labeling and analysis. Subtask 2, named "Role Detection", concentrated on identifying and classifying various roles within the respective context. In the "Full Annotation" subtask., the task is predicting a set of cues and their corresponding roles for each

data excerpt [13]. The "Role Detection" subtask provided participants with predetermined "gold" cues, with the objective being to accurately predict solely the roles associated with each excerpt. This work only handles Subtask 1. A dual-model approach was implemented where one model was dedicated to detecting roles and the other to identifying cues [12]. The evaluation metric was subsequently derived from the combined results of both models [12].

2.2.1. Preprocessing

The preprocessing used here is identical to [12]. Each excerpt is preprocessed, which means individual annotations are parsed, resulting in the creation of distinct lists containing their respective elements [12]. Subsequently, all these elements are concatenated to form the textual representation of a single excerpt. Given the potential occurrence of roles extending across preceding excerpts, these excerpts are integrated into the text by concatenating them with the corresponding text excerpts. This process is limited to adding a maximum of two additional roles, see Figure 2.1

Now I give you again the sentence only in addition with the two following sentences, because the roles can be partially contained in the following sentences. Text: (text)

Figure 2.1.: Model prompt structure, if roles are contained in previous excerpts

Due to the models behaving in varied and sometimes unpredictable patterns, if the text ends on a colon, a specific handling procedure is implemented: when the text ends on a colon, it is swapped with a period. Distinct prompt structures are designated for cues and roles individually. In addition to the text the model is given, it also needs to be given a prompt as a framework of how to process the text. The prompts designed for the model commence with an introductory segment, explaining the nature of the data it is given. Subsequently, they articulate the intended task that the model is expected to perform, followed by specifications outlining the desired format for the resulting output. Furthermore, these prompts include directives regarding the

expected behavior to be adopted in case of errors present within the data. Finally, an imperative statement finalizes the input, prompting the model to execute the designated task as seen in Figure 2.2

A cue is the lexical items in a sentence that indicate that speech, writing, or thought is being reproduced.

I want you to extract all cues in the text below. If you find multiple words for one cue, you output them separated by commas. If no cue can be found in the given text, you output the string #UKN# as cue.

Now extract all cues from the following sentence.
Use the prefix "Cues: "Sentence: ".

"Now find all roles in the sentence associated with the cue: (respective cue), you found in the beginning sentence."

Figure 2.2.: Model prompt structure example, to also extract roles from previous excerpts, because sometimes roles from previous excerpts are associated with the current cue.

Examples presented in Figure 2.3 and Figure 2.4 are derived from [12], where both the roles and cues prompt structure has been systematically applied.

Input:
User: A cue is the lexical items in a sentence that indicate that speech, writing, or thought is being reproduced.
I want you to extract all cues in the text below.
If you find multiple words for one cue, you output them separated by commas.
If no cue can be found in the given text, you output the string #UKN# as cue.
Now extract all cues from the following sentence.

Use the prefix "Cues: ".
Sentence: denn wir wissen: Neben ihren rassistischen Positionen
Assistant:
Output:
Cues: [wissen], [Positionen]</s>

Figure 2.3.: Example of cue extraction prompt. The end-of-sentence token, "</s>" denotes the conclusion of the output sequence.

Input:
User: Now I give you again the sentence only in addition with the two following sentences, because the roles can be partially contained in the following sentences.

Text: denn wir wissen : Neben ihren rassistischen Positionen ‡ haben die Rechtsradikalen nicht nur Klimawandelleugnung
im Angebot , ‡ sie haben auch die rechtspopulistischen Positionen eines Donald Trump gepachtet . ‡ Als Linke übernehmen
wir Verantwortung .

Now find all roles in the sentence associated with the cue 'wissen' you found in the
beginning sentence.

Assistant:
Output:
cue: enquotewissen
ptc: #UKN#
evidence: #UKN#
medium: #UKN#
topic: #UKN#
addr: #UKN#
message: Neben, ihren, rassistischen, Positionen, haben, die, Rechtsradikalen, nicht, nur, Klimawandelleugnung, im,
Angebot, sie, haben, auch, die, rechtspopulistischen, Positionen, eines, Donald, Trump, gepachtet
source: wir</s>

Figure 2.4.: Concrete role extraction prompt and output, for the cue: "wissen". The end-of-sentence token, "</s>" denotes the conclusion of the output sequence. The token "‡" marks the separation between two samples [12]

As previously mentioned, the prompts have also been translated into German, as seen in Figure A.2 and Figure A.1

2.2.2. QLoRA

QLoRA [14] is the combination of Quantization [21] and Low-Rank-Adaptation (LoRA) [22]. LoRA works by integrating low-rank matrices (matrices with

dimensions larger than their rank) into each layer of the model [22]. Instead of modifying all weights within the model, the process selectively adjusts solely the weights associated with these added matrices during the backpropagation process [2]. Quantization is a method to optimize computational efficiency [21]. The model's weights are converted from float16 to the int8 datatype which, according to [21], reduces computational requirements while preserving a reasonable level of precision for subsequent operations and calculations. The implementation of this method was instrumental in the execution of the experiment, given the constraints posed by limited computational resources.

2.2.3. Postprocessing

In a postprocessing phase, various output anomalies were addressed as per the methodology outlined by [12]. Initially, outputs in invalid formats are reclassified as "unknown". Furthermore, in cases where multiple overlapping cues are identified, these are compressed into a singular cue. Should the model generate words not present in the provided example, these additional words are removed. In instances where a segment is simultaneously identified as both a cue and a role, the attributions of the adjacent annotations are counted. Subsequently, the segment is assigned to the category, either cue or role, where this count is higher. The model's tendency to overlook punctuation marks in its output was noted. To address this, punctuation was incorporated into the output retrospectively.

Problem	Example	Fixed Example
Output Format	cues:“;`##he's saying`”	cues:`#UKN#`
Overlapping Cues	cues:“she said that he said”, “he said”	cues:“she said that he said”
Model Hallucination	”cues: he's saying ” (not contained in excerpt)	cues:`#UKN#`
Ambiguities	cues: ”he says” message: ”he says”	cues: “he says” message: `#UKN#`
Surrounding Punctuation	message: ”that they are traitors” (originally ending on punctuation)”	message: ”that they are traitors.”

Table 2.3.: Examples highlighting issues within the model's output, accompanied by the corresponding post-processing measures, deployed to solve them.

3. Results

3.1. Evaluation Metric

The selected metric for assessing the performance of our models is the F1 score (F1), as used in [12]. This evaluation employs a classification framework where a correct identification of an annotation as an annotation is categorized as a True Positive (TP), a misclassification of a non-annotation as an annotation is termed a False Positive (FP), a non-annotation wrongfully identified as an annotation is labeled a False Negative (FN), and an accurate recognition of a non-annotation as a non-annotation stands as a True Negative (TN). The formal definition of this metric is as follows:

$$F_1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{3.1}$$

The Precision metric is formally defined as:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{3.2}$$

The Recall metric is formally defined as:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{3.3}$$

A high precision indicates that out of all the instances the model predicted as positive, a larger proportion of them were correct. In contrast, if precision is low, the model predicted a lot more instances as positive, than there actually were. A high recall shows, that the model is effectively identifying a large proportion of the actual positive cases. A low recall indicates, that the model is missing a lot of the positive cases. F1 scores are computed separately for roles and cues, resulting in two distinct metrics. The overall F1 score is calculated

by taking the mean of these two metrics.

3.2. Result Discussion

Figure 3.1 shows the final results of the training, as determined by the metric provided by [12]. In the evaluation, the leading LeoLM model, which utilized a 7B parameter configuration and was prompted in English, attained an F1 score of 82.47%. Correspondingly, the best Llama-2 model, mirroring the same parameter configuration and language prompt settings, achieved an F1 score of 83.58%. The observed correlation between enhanced performance and increased parameter size was anticipated, but likely didn't exist due to too short amounts of training time and too small training datasets, as with increasing parameters, models also require more training data and training time. However, in the specific context of the Speaker attribution task utilizing German data, LeoLM, a German monolingual model, did not demonstrate a substantial performance advantage over Llama, a multilingual model. Given the benchmarks established in [15], it is important to interpret the results with caution. This recommendation stems from the assumption that a monolingual model would typically surpass a multilingual model in tasks conducted in the monolingual model's language. The analysis reveals that, in scenarios where prompts are provided in German, LeoLM models consistently score a substantially higher recall. In contrast, Llama-2 models are characterized by a significantly greater precision. This indicates that the LeoLM models tended to identify a greater number of cues than were actually present, leading to an increased amount of false positives. Contrary to the LeoLM models, the Llama-2 models tended to recognize fewer of the actual cues, resulting in a lower detection rate of true positives. The analysis of [12] underscores the criticality of precise cue prediction, emphasizing that inaccuracies in cue identification affects the performance of the roles model due to the transfer of errors from the cues model to the roles model.

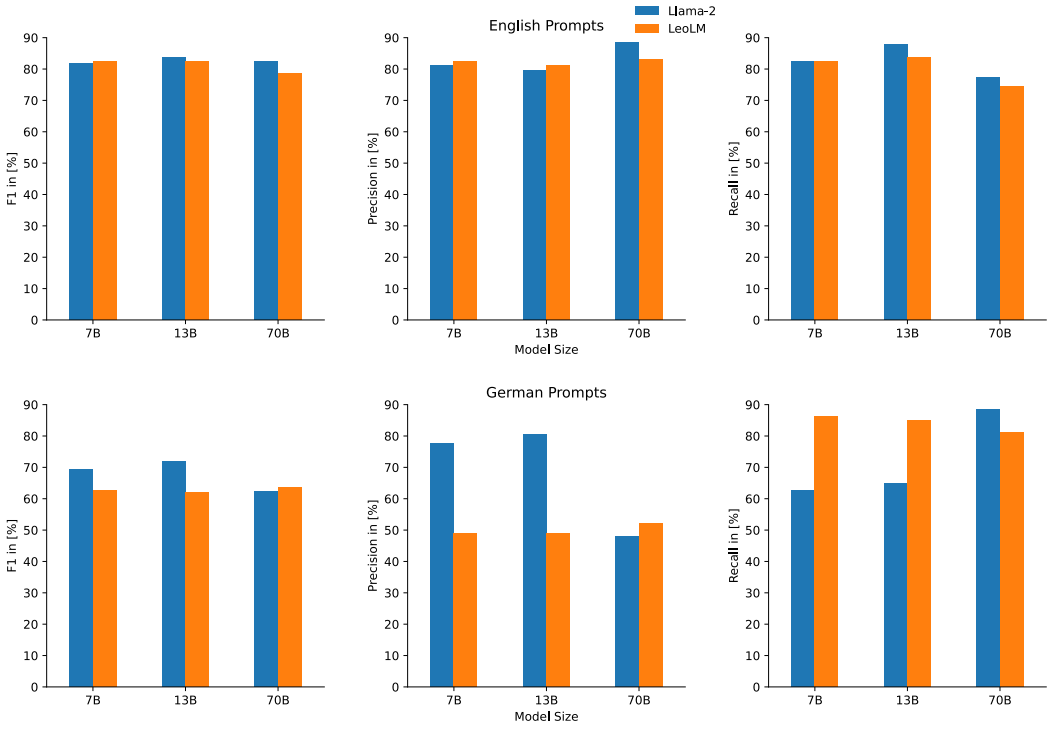


Figure 3.1.: Comparative performance (F1-score, precision, recall) of trained models in different prompt languages and sizes. Comparative analyses indicate that larger models do not demonstrate a statistically significant improvement in overall performance. Notably, there is a marked discrepancy between precision and recall metrics in LeoLM and Llama-2 models, when prompts are German. Furthermore, on the main performance metric (F1-Score), LeoLM and Llama-2 do not exhibit significant differences when evaluated under comparable conditions.

(A)	Llama-2-13B-Ger	Llama-2-13B-Eng
Recall	83.58	87.88
Precision	71.78	79.69
F1	71.78	83.58
(B)	Llama-2-13B-Eng	LeoLM-7B-Eng
Recall	87.88	82.35
Precision	79.69	82.58
F1	83.58	82.47

Table 3.1.: Comparison of the highest-performing language models for generating responses to prompts in German vs. English (A) and for evaluating the performance of Llama-2 vs. LeoLM models (B).

4. Summary and Outlook

This work showed that in the context of speaker attribution within German parliamentary debates, the performance of the LeoLM model, which was fine-tuned for German, does not significantly surpass that of the multilingual Llama-2 model. Additionally, the findings of this study indicate that the language used in prompts has a considerable influence on the performance of both monolingual and multilingual models. Moreover, the study highlights a substantial divergence in recall and precision metrics, where each model demonstrated significant improvement in one of the two, while concurrently showing notable deterioration in another. Looking ahead, these results suggest that relying predominantly on multilingual models may be adequate for effectively addressing most monolingual tasks. Consequently, this reduces the necessity to allocate computational resources for the training of monolingual models.

Bibliography

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. "Attention is All you Need". In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html (visited on 12/13/2023).

[2] S. Hochreiter and J. Schmidhuber. "Long Short-Term Memory". en. In: *Neural Computation* 9 (1997), pp. 1735–1780. ISSN: 0899-7667, 1530-888X. DOI: 10.1162/neco.1997.9.8.1735.

[3] OpenAI. *Introducing ChatGPT*. en-US. URL: /web/20231222104632/https://openai.com/blog/chatgpt (visited on 12/14/2023).

[4] C. K. Lo. "What Is the Impact of ChatGPT on Education? A Rapid Review of the Literature". en. In: *Education Sciences* 13 (2023). Number: 4 Publisher: Multidisciplinary Digital Publishing Institute, p. 410. ISSN: 2227-7102. DOI: 10.3390/educsci13040410.

[5] B. Puladi, C. Gsaxner, J. Kleesiek, F. Hölzle, R. Röhrig, and J. Egger. "The impact and opportunities of large language models like ChatGPT in oral and maxillofacial surgery: a narrative review". In: *International Journal of Oral and Maxillofacial Surgery* (2023). ISSN: 0901-5027. DOI: 10.1016/j.ijom.2023.09.005.

[6] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A.

Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, and T. Scialom. *Llama 2: Open Foundation and Fine-Tuned Chat Models*. arXiv:2307.09288 [cs]. 2023. DOI: 10.48550/arXiv.2307.09288.

[7] E. Almazrouei, H. Alobeidli, A. Alshamsi, A. Cappelli, R. Cojocaru, M. Debbah, É. Goffinet, D. Hesslow, J. Launay, Q. Malartic, D. Mazzotta, B. Noune, B. Pannier, and G. Penedo. *The Falcon Series of Open Language Models*. arXiv:2311.16867 [cs]. 2023. DOI: 10.48550/arXiv.2311.16867.

[8] A. Baez and H. Saggion. "LSLlama: Fine-Tuned LLaMA for Lexical Simplification". In: *Proceedings of the Second Workshop on Text Simplification, Accessibility and Readability*. Varna, Bulgaria: INCOMA Ltd., Shoumen, Bulgaria, 2023, pp. 102–108. URL: https://aclanthology.org/2023.tsar-1.10 (visited on 12/14/2023).

[9] T. T. Nguyen, C. Wilson, and J. Dalins. *Fine-Tuning Llama 2 Large Language Models for Detecting Online Sexual Predatory Chats and Abusive Texts*. arXiv:2308.14683 [cs]. 2023. DOI: 10.48550/arXiv.2308.14683.

[10] W. Lan, S. Qiu, H. He, and W. Xu. "A Continuously Growing Dataset of Sentential Paraphrases". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, 2017, pp. 1224–1234. DOI: 10.18653/v1/D17-1126.

[11] M. Kuusisto. "Organizational effects of digitalization: A literature review". In: *International Journal of Organization Theory and Behavior* 20 (2017). Publisher: Emerald Publishing Limited, pp. 341–362. ISSN: 1093-4537. DOI: 10.1108/IJOTB-20-03-2017-B003.

[12] P. G. Blaneck, T. Bornheim, N. Grieger, and S. Bialonski. "Automatic Readability Assessment of German Sentences with Transformer

Ensembles". In: *Proceedings of the GermEval 2022 Workshop on Text Complexity Assessment of German Text*. Potsdam, Germany: Association for Computational Linguistics, 2022, pp. 57–62. URL: https://aclanthology.org/2022.germeval-1.10 (visited on 12/13/2023).

[13] R. Ines, P.-F. Fynn, B. Annelen, R. Josef, B. Chris, and P. Simone paolo. "Overview of the GermEval 2023 Shared Task on Speaker Attribution in Newswire and Parliamentary Debates". Deutsch. In.

[14] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer. *QLoRA: Efficient Finetuning of Quantized LLMs*. arXiv:2305.14314 [cs]. 2023. DOI: 10.48550/arXiv.2305.14314.

[15] LAION. *LeoLM: Ein Impuls für Deutschsprachige LLM-Forschung | LAION*. en. URL: /web/20231222163114/https://laion.ai/blog-de/leo-lm/ (visited on 12/14/2023).

[16] A. Barbaresi. "A corpus of German political speeches from the 21st century". In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA), 2018. URL: https://aclanthology.org/L18-1127 (visited on 12/14/2023).

[17] G. Abrami, M. Bagci, L. Hammerla, and A. Mehler. "German Parliamentary Corpus (GerParCor)". In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, 2022, pp. 1900–1906. URL: https://aclanthology.org/2022.lrec-1.202 (visited on 12/14/2023).

[18] *spaCy · Industrial-strength Natural Language Processing in Python*. en. URL: https://spacy.io/ (visited on 12/18/2023).

[19] C. Rauh and J. Schwalbach. *The ParlSpeech V2 data set: Full-text corpora of 6.3 million parliamentary speeches in the key legislative chambers of nine representative democracies*. en. 2020. DOI: 10.7910/DVN/L4OAKN.

[20] *DeepL Übersetzer: Der präziseste Übersetzer der Welt*. de. URL: https://web.archive.org/web/20231217234742/https://www.deepl.com/de/translator (visited on 12/18/2023).

[21] B. Jacob, S. Kligys, B. Chen, M. Zhu, M. Tang, A. Howard, H. Adam, and D. Kalenichenko. "Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference". In: 2018, pp. 2704–2713. URL: https://openaccess.thecvf.com/content_cvpr_2018/html/Jacob_Quantization_and_Training_CVPR_2018_paper.html (visited on 12/17/2023).

[22] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. "LoRA: Low-Rank Adaptation of Large Language Models". en. In: 2021. URL: https://openreview.net/forum?id=nZeVKeeFYf9 (visited on 12/17/2023).

A. Appendix

Nun gebe ich Ihnen den Satz nur zusätzlich mit den beiden folgenden Sätzen wieder, denn die roles können teilweise in den folgenden Sätzen enthalten sein.
Text:
Nun finden Sie alle roles in dem Satz, die mit dem cue verbunden sind

Figure A.1.: Translation of role prompt framework using DeepL [20]

Ein cue ist ein lexikalisches Element in einem Satz, das anzeigt, dass gesprochen,
geschrieben oder gedacht wird.

Ich möchte, dass Sie alle cues aus dem folgenden Text extrahieren.
Wenn Sie mehrere Wörter für ein cue finden, geben Sie diese durch Kommas getrennt aus.
Wenn kein cue in dem gegebenen Text gefunden werden kann, geben Sie die Zeichenfolge #UNK#
als cue aus.

Nun extrahieren Sie alle cue aus dem folgenden Satz.
Verwenden Sie das Präfix “cue”:
Satz:

Figure A.2.: Cue prompt framework translated from English to German using DeepL [20]

	Llama-2	LeoLM
7B	69.27	63.57
13B	71.78	64.11
70B	62.28	63.57

Table A.1.: F1 Scores of final results on German prompts.

	Llama-2	LeoLM
7B	81.66	82.47
13B	83.58	82.33
70B	82.54	78.49

Table A.2.: F1 Scores of final results on English prompts.

	Llama-2	LeoLM
7B	77.65	52.23
13B	80.36	50.75
70B	48.10	52.23

Table A.3.: Precision Scores of final results on German prompts.

	Llama-2	LeoLM
7B	81.05	82.58
13B	79.69	81.05
70B	88.54	83.18

Table A.4.: Precision Scores of final results on English prompts.

	Llama-2	LeoLM
7B	62.26	81.19
13B	64.86	87.70
70B	88.33	81.19

Table A.5.: Recall Scores of final results on German prompts.

	Llama-2	LeoLM
7B	82.28	82.36
13B	87.88	83.65
70B	77.30	74.29

Table A.6.: Recall Scores of final results on English prompts.