CAN LANGUAGE MODELS BEHAVE LIKE WINE SOMMELIERS? – USING MULTIPLE AGENTS TO EVALUATE THE QUALITY OF WINE DESCRIPTORS GENERATED BY LLAMA 2

Siddarth Venkateswaran[1], *Ronald Böck*[1]

[1]*Genie Enterprise Inc., Branch Office Germany*
{venkat, rboeck}@genie-enterprise.com

Abstract: Wines are complex beverages whose taste can be described either numerically or textually, with the former involving the rating of the intensities of different aroma characteristics often with the help of a wine tasting wheel, and the latter with the help of crisp terms often in a poetic fashion. These are often done with the help of wine sommeliers who with one sniff can describe the wine. Usually, each sommelier has a unique style when it comes to textually describing a wine, research has shown that such differences have no negative impact in correctly classifying wines on the basis of their color, grape variety, region etc. Given the recent advancements in the field of Natural Language Processing, especially with the emergence of Large Language Models, we aim to check the capability of Llama 2 in its ability to generate texts pertaining to a specific color of a wine, given a list of aroma intensities as input prompts. In our experiments, we relied on data from Meininger and Falstaff, and on a combination of domain adaptation and pseudo-labeling techniques to create the corpus to train the Llama 2 model on. Also, we relied on a voting scheme of three differently trained classifiers to evaluate the wine-color specific text generation capabilities of Llama 2. Additionally, we employed the services of domain experts to evaluate the quality of a sample set of texts that was generated by Llama 2.

1 Introduction

Wine is a complex beverage full of aromas which can be tasted or smelled. To tap the wines characteristics, experts as well as customers usually rely on (flowery) textual descriptions or (somehow objective) tasting wheel representations [1]. Particularly, wine descriptions can be made in two ways - by rating the intensities of different aroma characteristics with the help of a wine tasting wheel, or in a purely textual format, usually also being based on terms from a tasting wheel. Spider-graphs form a perfect representation with which the aroma intensities are assigned with the help of a wine tasting wheel (cf. Figure 1). The rating of wine aroma characteristics involves assigning numerical values to its different properties like their fruitiness (presence of fruitful flavors like apple, vanilla, etc.), the influence of oak, the presence of minerals, etc., all on a scale of, e.g., 0 (no sensation) to 100 (strong sensation). Textual descriptions, however, involve the use of crisp terms, sometimes in a poetic fashion, describing the influence of various aromas of a wine. The process of collecting such data incorporates the knowledge of wine sommeliers, who already with one sniff can describe the wine - numerically as well as textually.

Considering the “standardized” tasting wheel [1], we assert that in fact at least two tasting wheels are necessary to cover proper assessments. The difference is based on color (red and

DOI: 10.35096/othr/pub-7091, URN: urn:nbn:de:bvb:898-opus4-70911

white) since both wine types show multiple variations in their characteristics in terms of numerical and textual ratings. For instance, a red wine could contain several berries, and a less content of citrus fruits. For white wines, it is rather the other way around.

As part of the BMEL funded project "PINOT"[1], we intend to check the influence of AI, especially in the domain of Natural Language Processing, across different strata in the wine supply chain. The recent advancements in Large Language Models (LLM) have sparked an interest in checking their text generation capabilities across different domains, and in our case, in the field of wines.

On manual inspection of spider-graphs, wines can be easily classified on the basis of their color based on the intensities of specific aromas. Likewise, textual descriptors contain terms specifically used for red or white wines. This research used two corpora (cf. Section 3) - Meininger which contained human-labeled aroma intensities as spider-graphs (cf. Figure 1), and Falstaff for which the intensities were obtained using domain adaptation (cf. Section 4.1.2) and pseudo-labeling techniques (cf. Section 4.1.3). Also, both the corpora contained textual descriptors, with those provided in Falstaff (cf. Figure 2) being more informative and descriptive than those in Meininger (cf. Figure 1). Therefore, given human-labeled (Meininger) or a pseudo-labeled (Falstaff) aroma-intensity as an input, we aimed to generate wine-color specific textual descriptions being at a level to those provided in Falstaff. An evaluation was conducted by three Multi-Layer Perceptron Classifiers (MLPC)[2] that were trained separately to classify wines on the basis of color (cf. Section 4.3), either using of aroma intensities as an input, or textual descriptors, or a combination of both. The experiments conducted yielded a high average weighted F1 score when it came to classifying the wines on the basis of the generated text. The color-specific text generation capabilities also yielded a high score when a sample set of generated texts were evaluated by a panel of domain experts.

Given a spider-graph of aroma intensities and a textual descriptor for a wine, we try to analyze the underlying meaning of these descriptors, and link these words to aroma intensities. As can be seen in Figure 1, the presence of aromas like "Kaffee schokoladig" at medium intensity in the spider-graph could have led to the presence of terms like "geröstete Kaffeebohnen" and "Kakaopulver". Also, the presence of aromas like "Beerenfrüchte", "Kirsche", "Marmeladig" and "Süße" could have led to the presence of "Brombeere" in the textual descriptor. However, there is no limit to the number of aroma profiles that can be referred to while rating a wine [1]. Also, every single aroma profile for a given wine cannot be expressed in words. A separate Multi-Layer Perceptron Regressor (MLPR)[2] was trained to find a mapping between these textual descriptors and the aroma intensities (cf. Sections 4.1.2 and 4.1.3). Average cosine similarities were computed between the input aroma intensities, and the aroma intensities generated from the MLPR as a result of the text generated from the LLM. Our experiments also yielded high average cosine similarity scores, highlighting the ability of LLMs to generate meaningful texts, when prompted with aroma intensities as an input.

Extensive steps were taken in this research - using different open-sourced models - to curate a corpus, in order to fine-tune the LLM on. This research also reviews the pros and cons of the chosen techniques, and what additional steps could be taken to carefully curate a corpus.

2 Related Work

The recent advances with prompt-based text generation models has sparked interest in generating domain-specific texts [2]. A new paradigm of fine-tuning called Instruction Tuning (IT) has made it easier for LLMs that were pre-trained on a large corpus of general domain texts, to

[1] https://pinot-ai.com/ (last accessed 10th of January 2024).

[2] Use of sklearn.neural_network from the scikit-learn toolkit.

easily adapt to domain specific texts [3]. IT involves curating datasets in a JSON structure of the form {Instruction, Output} in which the LLMs are fine-tuned in such a way that for a given instruction as an input prompt, the losses between the model generated output and the actual output in the IT curated dataset are minimized, which eventually helps in generating texts that are preferred by humans [3]. Some of the interesting but challenging domains in which IT has been implemented include those of the text generation in the medical domain, solving arithmetic problems, and code generation [3]. Likewise, we intend to make use of model-generated IT technique [3] to curate a corpus in the domain of wines, using which we intend to fine-tune Llama 2 [2] to generate wine sommelier like texts when prompted with aroma intensities.

3 Data Set

For this research, we rely on tastings data collected by two online accessible wine magazines - Meininger (11,000 textual descriptions in German accompanied with aroma intensities as spider-graphs) and Falstaff (122,000 textual descriptions only, all in German). Additional metadata like the color, grape variety, etc. were also included. For our experiments, we considered only those wines whose color was either red or white. As input for the text generation, aroma intensities that contained a dictionary of individual aroma names and their corresponding intensity ratings were used (cf. Figures 1 and 2).

4 Methods and Experimental Setup

4.1 Corpus Creation Steps

The following subsections highlight the methods implemented to curate a corpus, on which Llama 2 was fine-tuned:

4.1.1 Aroma Intensities Extraction from Meininger Corpus

The aroma intensities in the Meininger data were provided in the form of spider-graph images, with the lowest value per intensity being 0, and the highest value being 10 (cf. Figure 1). The following steps were taken to extract information from them:

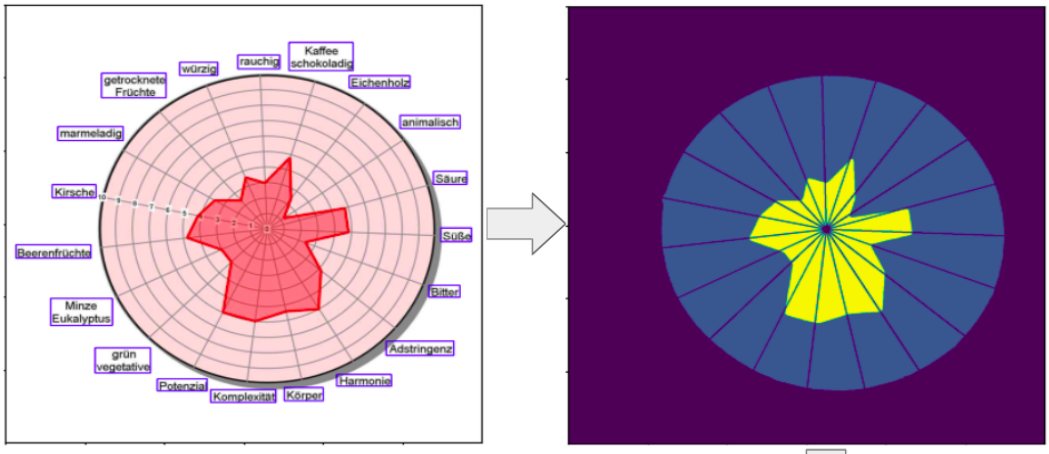
Step 1 - Getting Aroma Names: Using Label Studio, a sample set of 200 spider-graph images were manually annotated with bounding boxes drawn across each name of the aroma characteristic (cf. Figure 1). Detectron 2 [4], an object detection model from Meta AI was fine-tuned on these images, and predictions were made on the remaining unlabeled images. The bounding box regions from each image were then cropped and fed separately to EasyOCR[3] – an open-sourced Optical Character Recognition model – to fetch the aroma names.

Step 2 - Getting Aroma Intensities: The polygon and the circular regions of each spider-graph was segmented using Segment Anything Model [5], an open-source model from Meta AI for semantic segmentation. Radial lines were then drawn from the origin, to the circumference of the circle, in the direction of the aroma bounding-boxes (cf. Figure 1). The hitting point between the radial lines per category and the polygon represented the aroma intensities.

Step 3 - Normalizing Aroma Intensities: Based on the the data extracted from steps 1 and 2, it was observed that not all wines had a similar set of aroma names. Therefore, the normalization process involved identifying the unique aroma names that were present across all data points, which led to 38 aroma names being identified. In total, this results in 11,000 data points from

[3]https://github.com/JaidedAI/EasyOCR (last accessed for cross-checking 10th of January 2024).

Figure 1 – Steps involved in extracting the actual values of aroma intensities from the spider-graphs provided in a sample data from the Meininger corpus. *Top-left*: Original image with bounding boxes drawn per aroma name. *Top-right*: Image transformation after segmenting the circular and the polygon regions, and after drawing the radial lines from the origin to the circumference of the circle. *Bottom*: JSON structure of the extracted aroma intensity values, and the corresponding textual descriptor of the sample wine.



{'**Aroma Intensities**' : ['Adstringenz': 4.3, 'Animalisch': 1.5, 'Beerenfrüchte': 4.8, 'Bitter': 2.7, 'Eichenholz': 2.7, 'Getrocknete früchte': 2.5, 'Grün vegetative': 3.1, 'Harmonie': 6.1, 'Kaffee schokoladig': 4.8, 'Kirsche': 4.2, 'Komplexität': 6.1, 'Körper': 5.6, 'Marmeladig': 3.6, 'Minze eukalyptus': 3.2, 'Potenzial': 5.9, 'Rauchig': 3.1, 'Säure': 4.5, 'Süße': 4.9, 'Würzig': 3.5],
'**Text**' : 'geröstete Kaffeebohnen, Kakaopulver, Brombeere, Wacholder; recht straff mit Holzbiss; ein schöner Essensbegleiter'}

Meininger containing textual descriptions, with each description being accompanied by the numerical intensities of 38 unique aromas.

4.1.2 Domain Adaptation

Given the textual descriptions present across both the datasets, a German-based DistilBERT [6] model was fine-tuned on them for domain adaptation. With the help of the textual embeddings extracted from the fine-tuned DistilBERT model, a separate MLPR was trained on the Meininger corpus to map the textual descriptions to their corresponding numerical aroma intensities (cf. Section 4.1.1), only on those wines that were tagged either as red or white. Also, for the same set of wines, using a combination of the textual embeddings and the numerical aroma intensities as input, a separate MLPC was trained classifying the wine's color.

4.1.3 Pseudo-labeling Falstaff Corpus

Based on our experience of pseudo-labeling in a different domain [7], a similar approach was implemented on the Falstaff corpus to generate aroma intensities as pseudo-labels for each wine descriptor. Using the knowledge gained from the Meininger corpus in the domain adaptation phase (cf. Section 4.1.2), the aroma intensities for the Falstaff corpus were generated iteratively through the process of pseudo-labeling.

During each iteration of the pseudo-labeling process, textual embeddings for the unlabeled Falstaff corpus were extracted by making use of the fine-tuned DistilBERT model. The aroma intensities for these embeddings were predicted by feeding them to the pre-trained MLPR model. Then, using a combination of these embeddings and their corresponding aroma intensities, their class was predicted either as red or white wines using the pre-trained MLPC model. Only those data points that were correctly predicted by the MLPC were considered as confidently pseudo-labeled data.

After each iteration, MLPR and MLPC models were fine-tuned with the newly pseudo-

Figure 2 – Example of a sample descriptor from the Falstaff corpus converted into the IT format.

{**'instruction'**: 'Given these aroma intensities, generate a wine descriptor',
'input': "{"Adstringenz': 0.6, 'Apfel': 0.2, 'Beerenfrüchte': 1.8, 'Bitter': 1.8, 'Brioche Hefe': 0.0, 'Eichenholz': 3.0, 'Harmonie': 6.9, 'Kaffee schokoladig': 2.5, 'Kirsche': 2.3, 'Komplexität': 7.5, 'Körper': 6.6, 'Potenzial': 5.4, Säure': 4.9, 'gelbe Früchte': 1.0, 'getrocknete Früchte': 7.5, 'marmeladig': 1.4, 'rauchig': 3.9, 'würzig': 6.1}",
'output': 'Tiefdunkles Rubingranat, opaker Kern, violette Reflexe, zarte Randaufhellung. Attraktives Bukett nach Edelholz, reife schwarze Beeren, nach Brombeeren und Cassis, zart nach Nougat und Tabak. Komplex, saftig, reife Kirschen, feine Tannine, frisch und mineralisch, sehr gut balanciert, ein toller Begleiter bei Tisch.'}

labeled Falstaff corpus. This led to 109,000 data points from Falstaff being confidently pseudo-labeled with their corresponding aroma intensities, being verified by their color as a confidence parameter (herein referred to as pseudo-labeled corpus).

4.1.4 Final Corpora

The pseudo-labeled corpus created in Section 4.1.3 was then converted into the IT format (herein referred to as IT-curated corpus), a sample of which is shown in Figure 2. While the IT-curated corpus was used to fine-tune the Llama 2 model, the pseudo-labeled corpus was used to train the three classifier models that were eventually used to validate the color-specific text generated by the fine-tuned Llama 2 model. Also across each of Falstaff (pseudo-labeled and IT-curated) and Meininger corpora, a sample set of 1000 data - 500 each for red and white wines - were set aside as test data. This was done to compare the model performances on pseudo-labeled (Falstaff) and human-labeled (Meininger) data for all the objectives covered within this research.

4.2 Fine-Tuning Llama 2

For our experiments, we used the 7 Billion parameters version of Llama 2, a public LLM by Meta AI. The fine-tuning was done with the help of a model quantization technique called Low-Rank Adaptation of Large Language Models [8], on NVIDIA Tesla V100 with 16GB GPU RAM. On manual inspection, it was observed that fine-tuning Llama 2 for 1 epoch showed the most suitable performance.

4.3 Training Multi-Layer Perceptron Classifiers

In order to evaluate the wine-color specific text generated by the fine-tuned Llama 2, three separate MLPC models - named as MLPC-1, MLPC-2, and MLPC-3 - were trained on the pseudo-labeled corpus. The input data data fed to these classifiers can be referred to in Table 1, where the DistilBERT embeddings refer to the embeddings of the textual descriptors, while the aroma intensities refer to the numeric values of those intensities fetched during the pseudo-labeling process (cf. Section 4.1.3). These models were trained with the objective of predicting the class of wine-color based on the provided inputs. The evaluation techniques with which these models were used are mentioned in detail in Section 4.4.

4.4 Evaluating Generated Text

To evaluate the wine descriptor generation capabilities of Llama 2, when prompted with aroma intensities as an input (input intensities), the following steps were taken: Textual embeddings for the generated text, and their corresponding aroma intensities (output intensities) were fetched from the DistilBERT and the MLPR models that were fine-tuned during the domain adaptation

Table 1 – Data used to train the different MLPCs, and their Weighted F1 scores on unseen pseudo-labeled corpus of Falstaff, and on a sample set of 1,000 original data from Meininger.

Model	MLPC-1	MLPC-2	MLPC-3
Input Data	DistilBERT Embeddings	Aroma Intensities	DistilBERT Embeddings + Aroma Intensities
Pseudo-labeled Falstaff	0.99	0.86	0.99
Original Meininger	0.93	1.00	1.00
Average Scores	0.96	0.93	1.00

Table 2 – Weighted F1 scores of wine color specific texts generated by the fine-tuned Llama 2, evaluated by separately trained MLPCs - individually as well as using a majority voting scheme. The average cosine similarities were computed between the input aroma intensities, and the aroma intensities of the generated text extracted using the pre-trained DistilBERT and MLPR models.

Corpus	**Weighted F1 Scores**				**Average Cosine Similarities**
	MLPC-1	**MLPC-2**	**MLPC-3**	**Majority Vote**	
Falstaff	0.90	0.84	0.90	0.90	0.86
Meininger	0.96	0.88	0.97	0.97	0.83
Average Scores	0.93	0.86	0.94	0.94	0.85

phase (cf. Section 4.1.2). These embeddings and output intensities were in-turn fed to the three MLPCs that were used during the model training process (cf. Section 4.3), to predict the color class of the test data. On the basis of the outputs from each of the MLPCs, a majority vote was taken to identify the class of each test data. These majority voted output classes were compared with their ground-truth (color statement), and the weighted F1 score was used as a metric to check if Llama 2 could generate wine descriptions, varying on the basis of color.

To check the informativeness of the Llama 2 generated texts, an average cosine similarity score was computed between the input and the output intensities. Additionally, the individual performance of MLPC-2 was compared on its ability to classify wines, when fed with human-labeled or pseudo-labeled aroma intensities, as compared to the output intensities extracted from Llama 2 generated text.

Additionally, a sample set of ten Llama 2 generated descriptors (five each for red and white wines) were evaluated during a preliminary study with the help of three domain experts - proficient in their knowledge of wines as well as in the German language - on the following aspects: the quality of the text generated on a scale of 0 (worst) to 10 (best), and the wine color that the descriptor pertains to. The text quality evaluation was done to check two aspects: the grammatical correctness of the descriptors as well as the amount of information provided in it.

Table 3 – Overall evaluation of sample texts generated by Llama 2 by a panel of domain experts.

Color	**Average Descriptor Rating**	**Color Accuracy %**
Red	6.07	0.80
White	7.00	1.00
Average Scores	6.54	0.90

5 Results

Given the results obtained by the MLPCs in Table 1, the following observations can be made: Based on our observations made while comparing a sample set of textual descriptors across both corpora (cf. Figures 1 and 2), it can be assumed that MLPC-1 has some benefits from the extra information provided in those of Falstaff, as compared to those of Meininger (cf. Table 1). Additionally, it can be inferred that MLPC-2 benefited from the human-labeled aroma intensities of Meininger, as compared to those pseudo-labeled for Falstaff (cf. Table 1). MLPC-3 however attained an average weighted F1 score of 1.0 across both the corpora (cf. Table 1). Here, it can be deduced that aroma intensities - human-labeled or pseudo-labeled - when used in combination with the textual descriptors have a greater influence in classifying wines on the basis of color. On the basis of these results, we can hypothesise that the steps followed in the pseudo-labeling of Falstaff corpus (cf. Section 4.1.3) forms a reasonable seed to be able to fine-tune Llama 2 on.

Given that Llama 2 was fine-tuned on the pseudo-labeled Falstaff Corpus, it attained an average weighted F1 score of 0.94 across both corpora (cf. Table 2), when evaluated using a majority voting scheme by the MLPCs. Also, considering the overall feedback given by the panel of domain experts, 90% of the sample set evaluated by them had descriptions that defined a specific color of a wine (cf. Table 3). This shows that Llama 2 has the ability to generate wine descriptions that pertain to a specific color when prompted with originally labeled (Meininger data) or pseudo-labeled (Falstaff data) aroma intensities. Also, given that higher average weighted F1 scores were attained by the MLPCs on the Llama 2 outcomes of the Meininger corpus (cf. Table 2) - whether individually or on using a majority voted scheme - it can be assumed that better descriptors were generated from better quality aroma intensities.

Considering the informativeness of the texts generated by Llama 2, from our perspective, an average cosine similarity of 0.85 (cf. Table 2) already shows confidence in the methods adopted during the domain adaptation phase (cf. Section 4.1.2), to learn a mapping between the textual descriptors, and their corresponding aroma intensities. However, looking at the individual performance of MLPC-2, there was a drop in the F1 scores when tested on the aroma intensity outcomes of Llama 2 generated texts (cf. Table 2), as compared to those that were human-labeled or pseudo-labeled (cf. Table 1). This loss in information could be attributed to model hallucination, in which LLMs tend to learn the textual patterns, but at the same time also tend to deviate from factual reality [9]. This can also be attributed by an average rating of 6.54 on the sample set of descriptors by domain experts, in which some common feedback included occasional grammatical errors, and the presence of random information.

Given the above results, we can argue as follows: a combination of different models in an iterative framework helped curate an IT corpus. This in turn helped in fine-tuning Llama 2 to generate color-specific wine descriptors. However, the loss of information shows that instead of a model-generated approach, taking the services of a domain expertise, also called as a human-model-mixed approach [3] could help curate a better IT corpus, thereby alleviating the issue of model hallucination.

6 Conclusion

Given the high F1 score attained using a majority vote amongst three differently trained MLPCs, (cf. Table 2), and the high color-specific accuracy when a sample set of data was evaluated during a preliminary study with the help of three domain experts (cf. Table 3), the hypothesis holds true that LLMs have the ability to generate wine descriptions, especially with respect to a specific color. Also, the steps taken to curate a corpus to fine-tune an LLM on, can be attributed by the individual performances of three differently trained MLPCs - whether on human-labeled

Meininger, or pseudo-labeled Falstaff corpus (cf. Table 1). A high average cosine similarity score between the input and the model generated aroma intensities also highlights the steps taken during the domain adaptation phase (cf. Table 2). Additionally, employing the services of a domain expertise during the corpus curation phase could help alleviate issues related to loss of information, and grammatical errors (cf. Table 3).

Additional research can be focused to find the correlations between numerical values of different aroma characteristics and their generated text, and the changes that occur in the textual description based on the changes made in numerical values.

7 Acknowledgements

We acknowledge support by the PINOT project funded by the German Federal Ministry of Food and Agriculture (BMEL) under grant number 28DK107C20. We also thank Divya Yalavarthi for helping with the bounding-box annotations and Abdulla Al Foysal and Nazeer Basha Shaik for helping with the data collection steps.

References

[1] PARADIS, C.: *Conceptual spaces at work in sensory cognition: Domains, dimensions and distances. Applications of conceptual spaces: The case for geometric knowledge representation*, pp. 33–55, 2015.

[2] TOUVRON, H., L. MARTIN, K. STONE, P. ALBERT, A. ALMAHAIRI, Y. BABAEI, N. BASHLYKOV, S. BATRA, P. BHARGAVA, S. BHOSALE ET AL.: *Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288*, 2023.

[3] ZHANG, S., L. DONG, X. LI, S. ZHANG, X. SUN, S. WANG, J. LI, R. HU, T. ZHANG, F. WU ET AL.: *Instruction tuning for large language models: A survey. arXiv preprint arXiv:2308.10792*, 2023.

[4] WU, Y., A. KIRILLOV, F. MASSA, W.-Y. LO, and R. GIRSHICK: *Detectron2*. `https://github.com/facebookresearch/detectron2`, 2019.

[5] KIRILLOV, A., E. MINTUN, N. RAVI, H. MAO, C. ROLLAND, L. GUSTAFSON, T. XIAO, S. WHITEHEAD, A. C. BERG, W.-Y. LO ET AL.: *Segment anything. arXiv preprint arXiv:2304.02643*, 2023.

[6] SANH, V., L. DEBUT, J. CHAUMOND, and T. WOLF: *Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108*, 2019.

[7] VENKATESWARAN, S., R. BÖCK, T. KESSLER, and O. KRINI: *Pseudo-labelling and transfer learning based speech emotion recognition*. In *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2021*, pp. 19–26. TUDpress, Dresden, 2021.

[8] HU, E. J., Y. SHEN, P. WALLIS, Z. ALLEN-ZHU, Y. LI, S. WANG, L. WANG, and W. CHEN: *Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685*, 2021.

[9] RAWTE, V., A. SHETH, and A. DAS: *A survey of hallucination in large foundation models. arXiv preprint arXiv:2309.05922*, 2023.