Interspeech 2024
1-5 September 2024, Kos, Greece

Automated content assessment and feedback for Finnish L2 learners in a picture description speaking task

Nhan Phan[1], *Anna von Zansen*[2], *Maria Kautonen*[3], *Ekaterina Voskoboinik*[1], *Tamás Grósz*[1], *Raili Hildén*[2], *Mikko Kurimo*[1]

[1]Aalto University, Finland [2]University of Helsinki, Finland [3]University of Jyväskylä, Finland

nhan.phan@aalto.fi

Abstract

We propose a framework to address several unsolved challenges in second language (L2) automatic speaking assessment (ASA) and feedback. The challenges include: 1. ASA of visual task completion, 2. automated content grading and explanation of spontaneous L2 speech, 3. corrective feedback generation for L2 learners, and 4. all the above for a language that has minimal speech data of L2 learners. The proposed solution combines visual natural language generation (NLG), automatic speech recognition (ASR) and prompting a large language model (LLM) for low-resource L2 learners. We describe the solution and the outcomes of our case study for a picture description task in Finnish. Our results indicate substantial agreement with human experts in grading, explanation and feedback. This framework has the potential for a significant impact in constructing next-generation computer-assisted language learning systems to provide automatic scoring with feedback for learners of low-resource languages.

Index Terms: low-resource language, L2 speaking, content feedback, Automatic Speech Assessment, LLM

1. Introduction

Content assessment is an important part of evaluating L2 learners' spontaneous speech and can suggest valuable improvement for them [1]. For popular languages, such as English, there are several previous studies on the assessment of task completion with variable corpora [2, 3, 4]. However, much less data are available for L2 Finnish and other low-resource languages, presenting a significant challenge [5]. In addition, the beginner-level L2 speakers have shown less interest and motivation to participate in the data collection, which results in the under-representation of this group [6]. This is particularly problematic as they are the ones who would benefit most from the ASA.

Integrated speaking tasks, which combine the interpretation of supplementary visual materials (e.g. picture and video) with speaking [7], are used in language assessment to increase the authenticity of the speaking test. For example, pictures in speaking tasks can provide a starting point or content for spontaneous speech. These aspects are considered when assessing how the learner performs in task completion. Conventional ASA models typically rely only on audio data, which makes picture-based speaking task assessment challenging [4]. Inspired by research that combines image captioning with language models to solve the visual question answering [8, 9], we use a visual NLG model, capable of converting images to natural language based on a given context [10], to solve the ASA for integrated speaking tasks.

A notable challenge persists in ASA using deep learning: the scoring produced by the model may not be immediately understandable or beneficial to language learners. Efforts have been made to improve the interpretability of ASA model outputs based on the spoken content. However, those studies predominantly focus on the English language [1]. For Finnish and other languages with smaller L2 learner populations, obtaining sufficient high-quality data to provide automatic explanations, i.e. justifications or reasoning behind automatic scoring, is expensive. This discourages research on low-resource languages. Moreover, the development of ASA that can only provide scores is rarely sufficient for self-learning: the L2 learners need detailed, personalized feedback to improve their speaking skills [11]. Yet, the research focus on personalized feedback remains skewed towards the English language [12, 13].

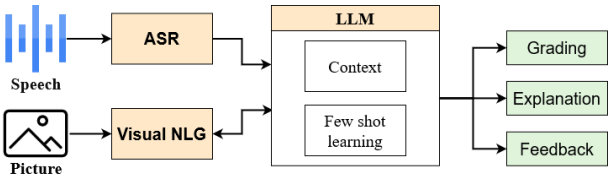

Figure 1: *Overview of our task completion grading framework*

In theory, some LLMs can potentially address those issues for low-resource languages. These models have been trained to evaluate written language tasks in popular languages and for translation purposes [14]. They have demonstrated effectiveness in rating advanced English writing tests [15]. The LLM demonstrated reasoning and explainability capabilities in evaluating complex English dialogues, as evidenced by Zheng et al, [16]. Consequently, the knowledge embedded in these models can be leveraged and applied to ASA tasks in low-resource languages. In ASR, there is recent innovative research in synthesising ASR output as LLM input for decoding speech with outstanding results [17, 18, 19].

While constrained by the quantity and quality of the Finnish L2 datasets for ASA, we developed a framework to utilise LLMs, specifically GPT-4 [14], to address the aforementioned issues in task completion assessment (Fig. 1). This framework aims to provide low-resource language learners with transparent grading and corrective feedback on their spontaneous speech. Our research focuses on the integration of the L2 ASR model, Visual NLG, LLM, and applying in-context learning [20] and chain-of-thought prompting [21].

2. Dataset

The data in this study consists of spoken responses from L2 Finnish learners engaged in picture description speaking tasks

10.21437/Interspeech.2024-1166

Figure 2: *Picture for the task of describing the missing hoodie*

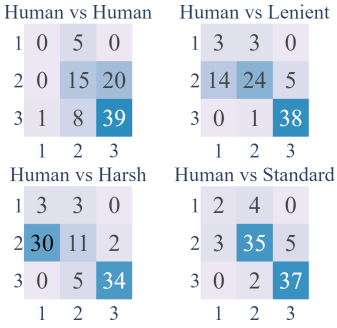

Figure 3: *Confusion matrices to compare original grades provided by the raters and the models (88 samples in the test set). In Human vs Models, the vertical axis denotes the human grade; when humans disagreed, the lower one of their grades was used.*

from the DigiTala dataset[1] [6]. Data from these integrated tasks share the common challenges - data imbalance and limited quantities - with each task having about 100 samples predominantly for grade 3 (excellent) and grade 2 (good) and less than 10 samples for grade 1 (partially). Due to resource constraints, we selected one of the tasks for this experiment: task 8c, as it has the largest amount of data with 114 samples. In task 8c, participants were presented with the picture shown in Fig. 2 and given the task description in Finnish. The English translation of the task is as follows:

Situation: You planned a visit of a friendship school [ystävyyskoulu - a school your school is collaborating with] at a café yesterday evening. At home, you notice your hoodie is missing. You call the café (max 30sec): Introduce yourself. Politely state your matter. Describe the hoodie, it looks like this: [Fig. 2]. Note: Do not disclose your real name or personal matters. Instead of your own name, you can use the name Maija / Matti Meikäläinen.

The dataset includes 114 audio recordings of L2 Finnish learners, transcripts made by humans, and task completion grades by human raters. The audio has an average duration of 24.5 seconds, with the shortest and longest recordings being 9.8 seconds and 30 seconds, respectively. 92 samples were graded by two raters, while the remaining 22 were graded by one rater, with 26 raters in total. The criteria for assessing task completion, as given to the original raters, are in Appendix B of Al-Ghezi et al. [6]. The grading distribution among the samples is as follows: 56 samples received a grade of 3, 51 were average graded as 2 or 2.5, and only seven samples were graded as 1, noting that no two raters concurred on a grade 1 assessment.

We used eight samples for training and development, the remaining 106 samples, formed the test set. The inter-rater agreement, measured by the Quadratic Weighted Cohen's Kappa (κ) [22] was 0.34 for the 88 samples in the test set that were rated by two raters. The baseline for this task (8c) is $\kappa = 0.30$ [6], noting that the baseline never predicted grade 1 due to the lack of training samples for this grade.

A significant challenge in developing an ASA system for this task is that the raters occasionally awarded the highest grade (3), even for the wrong colour or an entirely different hoodie. Some participants might not have seen or looked at the picture, leading them to improvise their descriptions. Consequently, certain raters might not have fully considered the accuracy of hoodie descriptions in their grading, which may be one reason for the low inter-rater agreement. For example, a description of the hoodie as "pure white" was inaccurately awarded grade 3.

The confusion matrix depicting rater discrepancies is illustrated in the top left matrix in Fig. 3. We observed that most disagreements occur near the boundary between grades (1-2 or 2-3). Conversely, fluent speakers whose proficiency exceeds the task requirements, generally receive unanimous grade 3 assessments. Nevertheless, a few fluent descriptions of an incorrect hoodie may have received grade 3.

[1] https://www.kielipankki.fi/corpora/digitala/

3. Framework and Experiment

First, the student's speech is transcribed by ASR and the output is given to the LLM (see Fig. 1). The ASR model is Wav2Vec 2.0 [23] pre-trained on the unlabeled 42.5K hours Uralic languages subset of the Voxpopuli corpus [24], then finetuned with 100 hours of native Finnish speech [25] and continually finetuned with the DigiTala dataset [6]. The average character error rate (CER) for the test samples is 6.35%. ASR does not use language model, because it may correct student's mistakes. As we cannot modify the GPT-4, we rely on its inherent robustness for interpreting imperfect spoken language [26]. To test this, we gave the human transcripts instead of ASR to the GPT-4 and observed no significant difference in results (see Table 3).

To simulate a fully automatic system, the picture is sent to a Visual NLG model for an in-context English description (note the two-way arrow in Fig. 1). For this experiment, we manually reduced the picture description length and focused on the grading, explanation and feedback performance. See Table 3 for the result of using automated output by GPT-4V.

The LLM is prompted with detailed instructions, including the task description, grading criteria, the picture's natural language description, and the ASR transcription of the spoken response. We instruct the LLM to provide reasoning for their grading decision. For feedback, we also requested that LLM return a corrected version of the student's speech. To enhance the LLM's reasoning capacity, we implement the chain-of-thought approach by breaking down the task into smaller steps. We also apply in-context learning by providing two examples for grade 3, two for grade 2, and one for grade 1. In our experiment, we found a significant performance increase from 1-shot to 2-shot learning for grade 2 and 3.

One notable open-source LLM is Llama 2 [27]. Our preliminary results, however, indicated that Llama 2 was unable to follow the complex instructions required for grading the task. Moreover, it was not sufficiently competent at understanding the complexities of the Finnish language, particularly when the context involved spoken Finnish, which has distinct differences from written Finnish.

For this experiment, only five samples were selected for

few-shot learning. Three more were used to change the model instructions and in discussion with the language assessment experts, who subsequently re-grade the data. Details of our experiments and prompt setups can be found in our repository [2].

3.1. Grading mode of ASA models

In our preliminary experiments, we found that the more lenient the grading criteria, the higher the Quadratic κ between the model and the original data, suggesting that the original raters may prioritize linguistic fluency over task-specific accuracy. To thoroughly investigate the reason and for an objective evaluation, we prepared three distinct ASA models: Lenient, Harsh, and Standard. They had almost similar grading instructions, with the only difference being in the criteria for grade 3.

The Lenient model was instructed to give grade 3 even with an incorrect colour description. In one of the training examples we gave to the LLM for grade 3, the speaker merely describes the hoodie as colourful ("värikäs") and does not introduce themself. In contrast, our Harsh model was told not to accept answers with incorrect colours for grade 3. It was given two training examples where speakers correctly describe the colours of the hoodie as well as its striped or lined patterns. Both the Lenient and Harsh models were also instructed to provide personalized feedback along with the grading and explanation.

For testing the grading performance, we used the Standard model which requires the speaker to describe the correct hoodie with the correct colours, but does not require the description of the pattern or texture of the hoodie (as these expressions exceed the targeted proficiency level [6]). One of the training examples given for grade 3 presents a comprehensive answer, with speakers describing the correct colours and patterns. The other example demonstrates the minimum answer to achieve grade 3, created by modifying a sample that only meets the grade 2 requirement (the speaker did not introduce themself). This augmented text response includes introducing the speaker, explaining the context, and politely asking to find the colourful hoodie.

3.2. Re-grading the data for this study

Since the human ratings were rather inconsistent with Quadratic κ only 0.35, we asked three experts, who trained the original raters and possess higher expertise in language assessment, to re-grade the data for this study. Deviating from the original rating process, the raters of this study graded the human transcripts without listening to the speech samples. With limited resources, we could not have all the data re-graded, so we selected only those samples that were most likely unreliable. The samples where both original raters agreed on grade 1 or 3 would presumably be reliable. However, as explained in Section 2, it was possible for both raters to award a grade 3 to samples describing the incorrect hoodie. Thus, we decided to leverage our experimental framework to select the samples for re-grading. Based on the Lenient and Harsh models' outputs, we selected all five samples graded as 3 by both human raters, but where the Lenient or Harsh model recommended a different grade. Among the other 34 samples where the raters mutually agreed with both models on grade 3, we still randomly selected one sample for re-grading. We also chose all six samples marked by at least one rater as grade 1. And finally, we randomly selected 34 samples from the rest of the data. As a result, we re-graded 45 samples, by asking each of the three raters to assess 30 overlapping samples. The remaining 61 samples, including 33 that were collectively graded 3 by raters and the Lenient and Harsh models, were not re-graded.

[2] https://doi.org/10.5281/zenodo.11385109

Table 1: *The Quadratic κ between raters and models. The top rows represent the original, and the bottom rows represent the re-graded grades. *Quadratic κ between human raters are from 88 samples (original) and 96 samples (re-graded).*

Human*	Lenient	Harsh	Standard
0.34	0.51 [0.42-0.61]	0.41 [0.33-0.51]	**0.56** [0.48-0.65]
0.69	0.64 [0.57-0.72]	0.65 [0.59-0.71]	**0.73** [0.68-0.79]

3.3. Evaluation of the explanation and feedback

In addition to re-grading the 45 samples, we requested the language assessment experts to evaluate the explainability, i.e. the reasoning behind the automatic score, and the feedback provided by our model. 50% of the explanatory and feedback outputs were taken from the Lenient model and the remaining 50% from the Harsh models. We asked the experts to mark whether they agreed (Yes/No) with the generated explanation. This agreement was solely on the clarity of the rationale behind the grading (Explainability), even if the raters did not agree with the grading. Additionally, raters were tasked to evaluate the "Grammatical correctness and fulfilling the criteria for grade 3" (Accuracy) and "Does the feedback improve the student's answer specific to their response and not just provide the correct answer?" (Usefulness) in the generated feedback.

The raters were aware of the inconsistencies in the original ratings, but did not know the details of the assessment models. They agreed on the grading criteria and the few-shot training examples of the Lenient model. They were also informed that the outputs come from different models, but not the allocation between models or the method used for sample selection.

4. Results

The Quadratic κ between our language experts in 45 re-graded samples is 0.70. We replaced the original grades with the new grades. The grading performance of our models is presented in Table 1. Our Standard model is in substantial agreement with human experts with $\kappa = 0.73$ [22]. The Quadratic κ is determined by randomly choosing one rater's grade as the true label and comparing it with the model output. The value is averaged over 1,000 runs. We calculated the 95% Confidence Interval (CI) by selecting the 2.5th and 97.5th percentiles of these runs.

We noticed that the framework is quite robust for grade 1. E.g., even if we do not give any grade 1 training examples to the LLM, the results remain almost as good, as shown in Table 3.

4.1. Explanation and Feedback

The results of the transparency and feedback quality outputs for 45 samples are detailed in Table 2. Given that this data consists solely of Yes/No responses, we chose to report the percentage of agreement, calculated from the count of raters' consensus [22].

The ability of our models to explain their decisions was well received by the language assessors. The explanations were clear and straightforward and the assessors noticed the harsh ratings

Table 2: *The average scores given by the raters. The scale is from 0 to 1, where 1 represents satisfaction with the model output and 0 dissatisfaction.*

	Explainability	**Accuracy**	**Usefulness**
Avg. score	0.93	0.74	0.86
% agreement	86.7%	53.3%	75.6%

proposed by the Harsh model. It is worth noting that even when they disagreed with the rating proposed by the model, the explanations still aided them in their own rating process.

As an example, Lenient model gave one sample grade 2 because student *"described the hoodie as 'harmoan näköinen' (grey-looking [3]) with a 'Gant' brand, which does not match the given description"*. It also provided the corrected feedback: "*Hei, olin eilen illalla teidän kahvilassa ystävyyskoulun ohjelmassa ja huomasin kotona, että hupparini jäi sinne. Huppari on erittäin värikäs, siinä on punaista, sinistä ja vihreää raitaa. Se on koko M. Voisitteko tarkistaa, onko se löytynyt?*".

While the feedback generated by the models was generally considered accurate and useful for supporting learning, its performance was hindered by a few limitations. The first stemmed from mistranslations between English and Finnish. For instance, the Finnish context "suunnittelitte ystävyyskoulun vierailua kahvilassa" (you planned a visit of a friendship school in a café) was mistranslated by LLM in the feedback as "kävin teidän kahvilassanne ystävyyskoulun kanssa" (I visited your café with a friendship school).

Secondly, the feedback failed to retain the "spoken" characteristic of students' speech. There are notable differences between spoken Finnish and written Finnish. For example, a colloquial version in spoken Finnish for "your café" is "teiän kahvilassa", while the written form is "teidän kahvilassanne". Generally, they can be used interchangeably, and speakers can use both spoken and written Finnish in their answers ("teidän kahvilassa" is also acceptable). However, our models show a tendency towards modifying the student's answer to the written form, failing our Usefulness criteria.

Thirdly, the feedback provided by the machine failed to serve learners who already performed well in terms of task completion; it did not encourage them or add any new aspects.

Despite these drawbacks, the language assessment experts generally considered the feedback to be useful for learners, i.e. the experts thought, that the feedback provided by the models would likely improve the learner's performance. Only in a few cases, the original response was considered to be better, usually because the automated feedback failed to use spoken language. Most of the improvised context made by speakers was correctly kept or improved (e.g. checking camera footage, the time of the visit, etc.). However, in its current form, the learners might need some guidance on interpreting the feedback, e.g., the differences between spoken and written Finnish. Our experts pointed out that this kind of feedback might have undesirable washback effects on practising oral skills. In the future, we aim to enhance feedback quality by including Finnish translations for complex phrases and separating feedback generation from grading and explanation.

[3]User said 'harmoan', but correct Finnish word is 'harmaan'

Table 3: *The Quadratic κ between raters and models using the re-graded data. Transcript: using human transcripts instead of ASR transcripts. 0-shot grade 1: do not use any training examples for grade 1. GPT-4V: use GPT-4V output directly.*

Transcript	**0-shot grade 1**	**GPT-4V**
0.73	0.73	0.68
[0.67 - 0.79]	[0.67 - 0.79]	[0.62 - 0.75]

4.2. Implications for language education

Our study has significant implications for L2 education. It introduces an innovative framework to use ASA and to provide transparent feedback automatically to L2 learners, even for languages and tasks with minimal training data. In our best model, we only used five training samples. The integration of Visual NLG underscores the framework's versatility. Potentially, the picture of this task can be switched to have a different item to be described, with the only modification in the grading instruction being the few-shot training examples.

Our work can bring new perspectives to language assessment, by supporting the work of the human raters and providing them with reasoning behind a certain score. In addition, using automated solutions requires technical and pedagogical expertise and will change the way how we design tasks and develop assessment criteria.

4.3. Limitations and Reproducibility

While this paper introduces a novel framework to apply ASR and LLMs in L2 education, it is subject to several limitations, primarily due to the low-resource nature of our study. We only used one proprietary model (GPT-4) as LLM. Finetuning the LLM was not feasible, due to its proprietary nature and the absence of adequate datasets. The scale of our experiment is small, with just one task. We also lack the resources to re-grade all data, and our test set may still have a few unreliable samples.

At the time of the writing, it is possible to reproduce the result using identical settings and the same infrastructure in the experiments. We also ran similar experiments using Azure OpenAI while keeping the same grading criteria (see Table 3).

5. Conclusion

In this work, we proposed a framework that combines LLM, Visual NLG, and ASR to automatically score task completion in a picture description speaking task as well as to provide explanations and corrective feedback to L2 learners. We also utilised this framework to identify problematic samples, investigate raters' discrepancies in the original data, and reduce the evaluation workload. Though our experiments were constrained by limited resources, we demonstrated promising results in the context of L2 learning and assessment, especially when handling low-resource and imbalanced datasets.

This paper also highlights the importance of LLMs for low-resource languages, suggesting their potential in designing innovative and resource-effective solutions. Considering this, we hope to attract more multidisciplinary research on ASA and computer-assisted language learning systems for low-resource languages. While the resources might not always be available, it is possible to leverage LLMs to develop a reliable system that would ultimately benefit low-resource L2 learners.

6. Acknowledgements

We would like to thank the following projects and funding agencies: NordForsk through the funding to "Technology-enhanced foreign and second-language learning of Nordic languages" (project number 103893); Research Council of Finland through the funding to "Digital support for training and assessing second language speaking" (grant no 322619, 322625, 322965), and "Automatic assessment of spoken interaction in second language" (grant no 355586, 355587, 355588).

7. References

[1] X. Wang, K. Zechner, and C. Hamill, "Targeted content feedback in spoken language learning and assessment." in *INTERSPEECH*, 2020, pp. 3850–3854.

[2] Y. Qian, R. Ubale, M. Mulholland, K. Evanini, and X. Wang, "A prompt-aware neural network approach to content-based scoring of non-native spontaneous speech," in *2018 IEEE spoken language technology workshop (SLT)*. IEEE, 2018, pp. 979–986.

[3] S.-Y. Yoon and C. Lee, "Content modeling for automated oral proficiency scoring system," in *Proceedings of the fourteenth workshop on innovative use of NLP for building educational applications*, 2019, pp. 394–401.

[4] P. Bamdev, M. S. Grover, Y. K. Singla, P. Vafaee, M. Hama, and R. R. Shah, "Automated speech scoring system under the lens: Evaluating and interpreting the linguistic cues for language proficiency," *International Journal of Artificial Intelligence in Education*, vol. 33, no. 1, pp. 119–154, 2023.

[5] A. v. Zansen and A. Huhta, "Developing automated feedback on spoken performance: Exploring the functioning of five analytic rating scales using Many-facet Rasch measurement," in *Digital Research Data and Human Sciences*. University of Jyväskylä, 2022.

[6] R. Al-Ghezi, K. Voskoboinik, Y. Getman, A. Von Zansen, H. Kallio, M. Kurimo, A. Huhta, and R. Hildén, "Automatic speaking assessment of spontaneous L2 Finnish and Swedish," *Language Assessment Quarterly*, vol. 20, no. 4-5, pp. 421–444, 2023.

[7] S. Luoma, *Assessing speaking*. Cambridge University Press, 2004.

[8] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "VQA: Visual question answering," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2425–2433.

[9] A. Salaberria, G. Azkune, O. L. de Lacalle, A. Soroa, and E. Agirre, "Image captioning for effective use of language models in knowledge-based visual question answering," *Expert Systems with Applications*, vol. 212, p. 118669, 2023.

[10] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, "Survey of hallucination in natural language generation," *ACM Computing Surveys*, vol. 55, no. 12, pp. 1–38, 2023.

[11] L. Gu and L. Davis, "Providing SpeechRater feature performance as feedback on spoken responses," in *Automated Speaking Assessment: Using language technologies to score spontaneous speech*, K. Zechner and K. Evanini, Eds. Routledge, 2020, pp. 159–175.

[12] X. Xi, "Automated scoring and feedback systems: Where are we and where are we heading?" pp. 291–300, 2010.

[13] S.-Y. Yoon, C.-N. Hsieh, K. Zechner, M. Mulholland, Y. Wang, and N. Madnani, "Toward automated content feedback generation for non-native spontaneous speech," in *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 2019, pp. 306–315.

[14] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya *et al.*, "GPT-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[15] B. Naismith, P. Mulcaire, and J. Burstein, "Automated evaluation of written discourse coherence using GPT-4," in *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, 2023, pp. 394–403.

[16] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu *et al.*, "Judging LLM-as-a-judge with MT-bench and chatbot arena," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[17] G.-T. Lin, C.-H. Chiang, and H.-y. Lee, "Advancing large language models to capture varied speaking styles and respond properly in spoken conversations," *arXiv preprint arXiv:2402.12786*, 2024.

[18] K. Everson, Y. Gu, H. Yang, P. G. Shivakumar, G.-T. Lin *et al.*, "Towards ASR robust spoken language understanding through in-context learning with word confusion networks," *arXiv preprint arXiv:2401.02921*, 2024.

[19] Z. Ma, G. Yang, Y. Yang, Z. Gao, J. Wang *et al.*, "An embarrassingly simple approach for LLM with strong ASR capacity," *arXiv preprint arXiv:2402.08846*, 2024.

[20] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[21] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 24 824–24 837, 2022.

[22] A. J. Viera, J. M. Garrett *et al.*, "Understanding interobserver agreement: the kappa statistic," *Fam med*, vol. 37, no. 5, pp. 360–363, 2005.

[23] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[24] C. Wang, M. Riviere, A. Lee, A. Wu, C. Talnikar, D. Haziza, M. Williamson, J. Pino, and E. Dupoux, "Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation," *arXiv preprint arXiv:2101.00390*, 2021.

[25] A. Moisio, D. Porjazovski, A. Rouhe, Y. Getman, A. Virkkunen, R. AlGhezi, M. Lennes, T. Grósz, K. Lindén, and M. Kurimo, "Lahjoita puhetta: a large-scale corpus of spoken Finnish with some benchmarks," *Language Resources and Evaluation*, vol. 57, no. 3, pp. 1295–1327, 2023.

[26] M. He and P. N. Garner, "Can ChatGPT detect intent? Evaluating large language models for spoken language understanding."

[27] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.