Towards Improving ASR Outputs of Spontaneous Speech with LLMs

Manuel Karner[1], Julian Linke[2], Mark Kröll[1], Barbara Schuppler[2], Bernhard C. Geiger[1,2]

[1]Know-Center GmbH, Sandgasse 34, 8010 Graz, Austria,
[2]Signal Processing and Speech Communication Laboratory,
Graz University of Technology, Inffeldgasse 16c, 8010 Graz, Austria
Correspondence: geiger@ieee.org

Abstract

This paper presents ongoing work towards an initial understanding of how large language models (LLMs) can assist automatic speech recognition (ASR) tasks. More concretely, we investigate if LLMs can improve hypotheses obtained from ASR systems, and if so, which patterns in the hypothesis allow for a correction. Our results show that LLMs can mainly correct syntax errors or errors caused by ASR systems splitting long words. We further find that in the majority of cases the word error rates with respect to the human annotation increase when an LLM is applied, while the semantic similarity with the human annotation improves.

1 Introduction

As artificial intelligence continues permeating our lives, reliable performance of automatic speech recognition (ASR) of conversational and spontaneous speech becomes more and more important as an enabler for natural conversations with social robots and automatic meeting transcripts, among other things. While ASR systems now achieve human-level performance for read or prepared speech (Szymański et al., 2020a), for which multiple benchmark datasets are available (Librispeech (Panayotov et al., 2015), Common Voice (Ardila et al., 2020), Multilingual Librispeech (Pratap et al., 2020)), ASR performance is still unsatisfactory for spontaneous speech. This is particularly true for face-to-face conversations of less-resourced languages, where word error rates (WERs) of 21.0-16.3% for Hungarian (Mihajlik et al., 2023, 2024) and up to 35.71% to 16.09% for Austrian German (Linke et al., 2022) are common.

Modern ASR systems like wav2vec (Baevski et al., 2020) or Whisper (Radford et al., 2023) rely on transformer architectures and often achieve excellent performance on read speech without requiring an explicit, powerful language model (LM). Indeed, common implementations of wav2vec contain only a simple n-gram LM. At the same time, large LMs (LLMs) have shown impressive performance on a variety of natural language tasks. Llama2 was even shown to be capable of ASR, if it is provided with embeddings of the acoustic signal (Fathullah et al., 2024).

In this paper, we investigate if LLMs can be used to correct errors in ASR outputs (Section 3) and which error patterns are easiest to correct (Section 4). Since we find that WERs are insufficient to fully evaluate original and corrected ASR outputs, we also analyze how the semantic similarity to the ground truth changes if an LLM is applied. In the future, we will incorporate LLMs into ASR systems based on the results presented here, aiming at coupling the power of LLMs with the acoustic signal available to the ASR system (cf. discussion in Section 5).

2 GRASS corpus and ASR systems

The experiments of this paper are based on data from the Graz Corpus of Read and Spontaneous Speech (GRASS) (Schuppler et al., 2014, 2017). More concretely, we used the conversational speech component of GRASS, which contains one hour long conversations from 19 pairs of speakers, summing up a total of 220.000 word tokens. Since the speakers knew each other well prior to the recordings and since they chatted with each other without topic instruction and without any experimenter in the recording room throughout the whole conversation, the speaking style of GRASS is highly spontaneous and casual compared to other data sets used in speech technology (Linke et al., 2023). Its challenging characteristics for ASR are the high degree of pronunciation variation and dialectal pronunciation, the highly varying speech rate, and the highly frequent occurrence of broken words, fillers, incomplete and/or grammatically wrong

Proceedings of the 20th Conference on Natural Language Processing (KONVENS 2024), pages 339–348

p1	I will give you a part of an Austrian German sentence. Please correct it for me.
p2	I will give you a part of a german sentence. Please correct it for me but preserve austrian dialect.
p3	I had to write down this text in austrian german I heard, but it could be that excactly one word is wrong.
p4	I had to write down this Austrian German text I heard, but there could be one or two errors in it. I need your help to correct it. I will provide you the text and you approach the problem step by step. First check if the sentence is grammatically correct. Secondly decide which word is probably wrong, in rare cases there could be two wrong words. Thirdly exchange the wrong word with what you think is the right word and would make the sentence grammatically correct.
p5	I have to write down a sequence of Austrian words I listened to, but there are some problems with it. Since my hearing is bad it could be that I split a long word like "holzbungalows" into two smaller words that sound similar together like "halt" and "pomelos". I make other errors too, often they are grammar related. Therefore, I need your help to correct my mistakes.
p6	I have a part of a sentence in austrian german but it is grammatically incorrect. I need your help to improve it but there are three problems with it. The principal part of every word could be wrong, a word could be missing and in rare cases you have to delete a certain word. Do the best you can to form a grammatically correct sequence of words while preserving anything that you think is true.

Table 1: Prompts that were investigated in our experiments. We only report results for base prompts p1, p4, p5.

utterance structures, laughter and non-lexical tokens. Moreover, the lively turn-taking dynamics result in disrupted turns, overlapping speech, one-word-utterances (e.g., *hmh, ja, sicher*) and overall shorter utterances than for instance in spontaneous interviews. We use GRASS as an example for a database that 1) contains speech from a language variety that is low-resourced, and 2) for a speaking style that is highly casual and spontaneous, both posing (different types of) challenges to ASR systems. Reason to use GRASS for this study is not only to improve WER, but also to gain insights with respect to how an LLM in general deals with disfluent and even grammatically wrong structures that are highly frequent in GRASS.

Here, we compare ASR results for GRASS from four ASR systems comprising Whisper (Radford et al., 2023), Kaldi (Povey et al., 2011), and wav2vec2 (Baevski et al., 2020) with and without a lexicon and LM (w2v/w2vLM). For all experiments, we excluded utterances containing laughter, singing, imitations/onomatopoeia, unintelligible word tokens, and artefacts leading to 33734 utterances (14.4h).

Training/fine-tuning these ASR systems as described in Appendix C, we achieved similar conversation-dependent WERs with high-resourced zero-shot Whisper ($41.78\% \pm 8.23\%$) and low-resourced Kaldi ($42.86\% \pm 4.78\%$), while best WERs were achieved with the fine-tuned w2v ($29.81\% \pm 4.80\%$) and w2vLM ($22.79\% \pm 4.02\%$). Interestingly, mean WERs with Whisper were worst for utterances including only two word tokens (approx. 55%) but decreased for utterances with more word tokens (mean WER was approx. 30% for utterances with 15 word tokens).

3 Approach

We are interested in whether and to what extent generative capabilities of LLMs can be utilized to correct hypotheses obtained from ASR systems. As we are analyzing German utterances, we opted for a recent version of SauerkrautLM specifically fine-tuned for the German-speaking region as well as aligned to human preferences by direct preference optimization (Rafailov et al., 2023). The SauerkrautLM-Mixtral-8x7B-Instruct model, optimized to follow instruction-based prompting, is a Mixture of Experts model with the foundational model being Mixtral-8x7B-Instruct[1]. Each of the 8 experts is using the Mistral-7B architecture; resource efficiency was achieved by using a quantized variant of the LLM, i.e., `gptq-4bit-32g-actorder_True`.

Effective prompt engineering remains an open research challenge (Gonen et al., 2022). LLM outputs can vary significantly and unpredictably, for instance, depending on choice (Zhang et al., 2022) as well as on ordering (Lu et al., 2022) of (in-context) examples.

Informed by best practices from the literature, we initially designed six instruction base prompts (BP) from which we selected three for our experiments (see Table 1). Prompt p1 only emphasizes that GRASS contains Austrian German. Prompt p4 was inspired by (Zhang et al., 2023), where the authors recommended to add the phrase "Let´s think step by step" to "facilitate the reasoning chains in LLMs". Prompt p5 emphasizes an error pattern we named "long word splitting error" (cf. Observation 1 in Section 4). A typical example of this would be that the ASR system splits the word "*erzähle*" into the words "*er*" and "*zählt*". The three omitted prompts were either redundant or suffered from performance issues. For example, the omitted prompt p2 used the wording "preserve Austrian dialect" instead of "Austrian German" (hence is redundant) and yielded worse corrections than p1.

[1]Model Card: `https://huggingface.co/VAGOsolutions/SauerkrautLM-Mixtral-8x7B-Instruct` last accessed: 19.7.2024

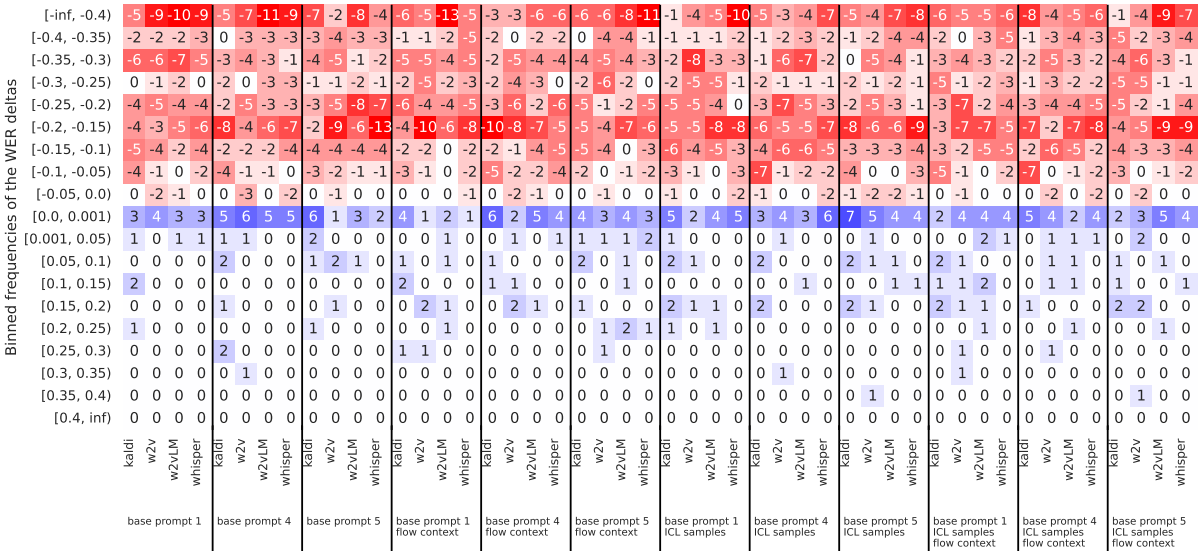

Figure 1: Effect of LLM corrections (instructed by 12 prompt combinations) on outputs of four ASR systems: Binned differences (deltas) are illustrated between the WERs of ASR hypotheses and the WERs of 1776 LLM corrections (4*12*37 = #ASR models * #experiments * baseline dataset size). Negative differences (counts in red) indicate a WER increase - positive differences (counts in blue) indicate a WER decrease.

We provided the LLM with two types of context. Flow Context (FC) represents short-term information dependencies from the conversation flow, i.e. the last 350 characters before the hypothesis to be corrected. In-Context Learning (ICL) leverages the ability of the LLM to learn from task demonstrations without fine-tuning the model. Thus, SauerkrautLM was provided with four hypothesis/reference pairs to better understand the correction task. Those were matched to the source of the input utterance, to incorporate the differences of the four ASR systems, i.e. in how they apply substitution, deletion, and insertion operations. While most of these differences are too subtle and diverse for a proper qualitative analysis, we noticed that Whisper sometimes keeps >70% of the letters in the correct order if it substitutes a word.

Combinations of three BPs with either FC, ICL, or both, led to a total of 12 experiments. Figure 2 in Appendix C shows an example for one full prompt used in the experiments including a system prompt, base prompt p1, FC, and one input utterance.

4 Results

After an initial data analysis and a preliminary experimental phase, one conversation from the GRASS corpus was selected to test how LLMs respond to different kinds of inputs. From this conversation, a baseline dataset (Table 4) was built, consisting of 37 utterances that 1.) contain at least three words, 2.) lead to (mostly) wrong ASR hypotheses, and 3.) appear to be improvable by a human. While the third criterion is quite subjective, we differentiated between utterances that probably contain enough information (within themselves and/or within the FC) to be improvable and those utterance for which a human correction would certainly not be possible with a high confidence. For a detailed explanation and examples on this matter we refer to the Appendix C.4.

Figure 1 shows differences between the WERs of the ASR hypotheses and the LLM corrections. These differences are binned into WER ranges and counted; each column represents resulting WER differences for the 37 utterances in the baseline dataset. Negative differences (counts in red) indicate a WER increase - positive differences (counts in blue) indicate a WER decrease. As a general result, WERs rather increased for the majority of considered utterances, independent of prompts, the amount of provided context, or the ASR system that generated the hypotheses. Notably, some hypotheses were amenable to correction by the LLM's inherent language understanding capabilities. In the remainder, we will analyze these utterances in detail to better understand under which circumstances LLMs can contribute to correcting ASR outputs. The results look qualitatively similar for independent execution of the chosen prompts and even when using a different LLM. To allow for

Index	Reference	ASR Hypothesis	LLM Correction	WER		Similarity	
				ASR	LLM	ASR	LLM
1.1	sich was einprägen und auswendig lernen	sich was einprägen und auswenn die lernen	sich etwas einprägen und auswendig l	0.29	0.33	0.90	0.93
1.2	sich was einprägen und auswendig lernen	sich was einprägen und aus wenn die lernen	sich etwas einprägen und es auswendig lern	0.38	0.43	0.89	0.98
2	und es verschickt natürlich automatisch	und es war schickt natürlich automatisch	und es wäre natürlich automatisch	0.33	0.2	0.79	0.70
3	und danach ist er scheinbar irgendwie	und dann nach ist er schon immer irgendwie	und dann ist er danach irgendwie gewesen	0.5	0.43	0.79	0.89
4	wenn_du jetzt einen fehler machst beim eingeben	wenn du jetzt einen fehler machst beim einsteigen	wenn du nun bei der eingabe einen fehler	0.38	1.0	0.76	0.90
5	die haben im hafen irgendein anderes boot	die haben im hafen irgendein anderes bus	die haben im hafen ein anderes schiff	0.14	0.29	0.68	0.89
6	sogar mitten in der nacht am hafen unten	sogar mitten in der nacht am haufen unten	sogar mitten in der nacht am hafen unten	0.12	0.0	0.90	1.0
7	zu ihrer verteidigung nur der kapitän hat eine schusswaffe falls piraten	zu ihrer verteidigung hat der kapitän eine schusswaffe falls es möglich ist	zu ihrer verteidigung hatte der kapitän eine schusswaffe falls nötig	0.38	0.27	0.90	0.89
8	ja ich frage mich auch immer	ja ich frag mich auch immer	ja ich frage mich auch immer	0.17	0.0	0.99	1.0
9	so wie die dort hausen	so wie die dort hause	so wie die dort hausen	0.2	0.0	0.94	1.0
10	das schiff zu entern	der schiff zu entscheiden	das schiff zu ents	0.5	0.25	0.77	0.82

Table 2: Experiments showing the effect of LLMs on ASR outputs. We refer to Observations 1 - 4 for more details on used prompts as well as respective ASR systems.

another perspective, we introduce the semantic similarity between reference and ASR hypothesis as well as reference and LLM correction as additional metric. The similarity values are calculated by first creating embedding vectors using `German_Semantic_STS_V2 model`[2], followed by calculating the cosine similarity. In Table 2, we illustrate selected utterances, the WERs as well as the semantic similarities before and after correction with LLMs. While the WERs in many cases increase, so does the semantic similarity (see Figure 3 in Appendix C for a heatmap similar to Figure 1, but with a focus on semantic similarity).

Observation 1: ASR Systems Split Long Words (Index 1-3). All ASR systems except Whisper tend to split long words. Since Kaldi and w2v (idx 1.1) often introduce syntactic errors into the split words, these errors are easier to correct, in comparison to w2vLM (idx 1.2) which only produces correct syntax. For w2vLM, this "long word splitting error" increases the number of words in the utterance, which makes correction even harder. This can lead to cases where the WER improves but the semantic overlap decreases (idx 2), or to different wordings with correct semantics (idx 3)

Observation 2: Relevance of FC (Index 4-7). Providing conversational context (FC) can lead to situations where the LLM output is semantically closer to the reference. While in some cases this also leads to fewer word errors (idx 6, idx 7), sometimes the WER increases for the sake of correcting the semantics of the hypotheses (idx 4, idx 5). Referring to Figure 1, FC had this positive effect in only approx. half of the used prompts.

Observation 3: Syntax Errors Are Easy to Correct (Index 8-10). As expected, syntax errors in the hypotheses produced by Kaldi and w2v are easily corrected by the LLM (idx 8, 9). The same holds for wrong articles (idx 10). These types of errors are corrected quite reliably (in our small set of experiments), which suggests a direction for future prompt engineering efforts.

Observation 4: Whisper is Rarely Corrected. SauerkrautLM almost never improved hypotheses resulting from Whisper. The main reason behind this is that our dataset consists mainly of (comparably) long utterances, and we can observe that for Whisper the WERs decrease with utterance length.

5 Discussion and Outlook

Our attempts at correcting ASR hypotheses with LLMs led us to rethink what it means to "correct ASR output". The main goal of an ASR system might depend on the application scenario: (i) it could be to transcribe a conversation as accurately as possible (e.g., in case of court protocols), or (ii) it could be to summarize the content of a conversation in a comprehensive, inclusive way (e.g., in case of meeting minutes). This appears to be highly relevant for setting up the ASR framework with respect to LLM selection as well as prompt engineering. To give an example, the German verb "*frag*" (idx 8 in Table 2) may be adequate for one, but inadequate for another scenario. This directly relates to the used metric to evaluate ASR outputs. WER as a metric may be inappropriate in certain scenarios

[2]Model card: https://huggingface.co/aari1995/German_Semantic_STS_V2 last accessed: 19.7.2024

(and indeed, WER has often been criticized in the literature (Aksënova et al., 2021; Szymański et al., 2020b; Wang et al., 2003)). For these other scenarios, utilizing semantic similarity as metric might be better suited as it generally measures whether an output shares more (idx 4) or less (idx 2) meaning with the reference.

Our preliminary analyses indicate that LLMs may indeed be capable of improving certain error patterns in ASR outputs (such as syntax errors or errors due to long words being split). While the results of these analyses still must be reproduced using a larger variety of prompts and confirmed with statistical tests, we take the liberty to reflect on promising directions for future work. On the one hand, targeting only specific error patterns could lead to more stable corrections, by using Chain of Thought (Wei et al., 2022) or even Tree of Thought (Yao et al., 2023) based prompting. On the other hand, in our current implementation, LLMs attempt to correct ASR hypotheses without taking into account the speech signal, i.e. decoupling acoustics from text. Ignoring this important piece of information may be one of the reasons behind the sub-par performance exhibited in Figure 1. Having shown that LLMs can nevertheless improve ASR outputs in some cases suggests that including LLMs in ASR systems, thus coupling acoustic and language models, is a promising approach for automatic recognition of conversational speech. Conducting respective experiments, especially with longer hypotheses for which LLMs should be most useful, is within the scope of future work.

Acknowledgments

The work by M. Karner and B. C. Geiger was funded by grant P-32700 from the Austrian Science Fund.

References

Alëna Aksënova, Daan van Esch, James Flynn, and Pavel Golik. 2021. How might we create better benchmarks for speech recognition? In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pages 22–34. Association for Computational Linguistics.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proc. of the 12th Language Resources and Evaluation Conference*, pages 4218–4222.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2021. Unsupervised Cross-Lingual Representation Learning for Speech Recognition. In *Proc. Interspeech 2021*, pages 2426–2430.

Facebook Research. 2022. Fairseq Model (XLSR). `https://github.com/facebookresearch/fairseq/tree/main/examples/wav2vec`. Last Accessed: 2024-03-05.

Yassir Fathullah, Chunyang Wu, Egor Lakomkin, Junteng Jia, Yuan Shangguan, Ke Li, Jinxi Guo, Wenhan Xiong, Jay Mahadeokar, Ozlem Kalinli, Christian Fuegen, and Mike Seltzer. 2024. Prompting large language models with speech recognition abilities. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 13351–13355.

Hila Gonen, Srini Iyer, Terra Blevins, Noah A Smith, and Luke Zettlemoyer. 2022. Demystifying prompts in language models via perplexity estimation. *arXiv preprint arXiv:2212.04037*.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, page 369–376.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197.

Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. 2015. Audio augmentation for speech recognition. In *Proc. Interspeech 2015*, pages 3586–3589.

Julian Linke, Philip N. Garner, Gernot Kubin, and Barbara Schuppler. 2022. Conversational speech recognition needs data? Experiments with Austrian German. In *Proc. of the 13th Language Resources and Evaluation Conference*, pages 4684–4691.

Julian Linke, Saskia Wepner, Gernot Kubin, and Barbara Schuppler. 2023. Using Kaldi for Automatic Speech Recognition of Conversational Austrian German. *arXiv preprint arXiv:2301.06475*.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 8086–8098.

Péter Mihajlik, Máté Soma Kádár, Gergely Dobsinszki, Yan Meng, Meng Kedalai, Julian Linke, Tibor Fegyó, and Katalin Mády. 2023. What kind of multi- or cross-lingual pre-training is the most effective for a spontaneous, less-resourced asr task? *2nd Annual Meeting of the ELRA/ISCA SIG on Under-resourced Languages (SIGUL 2023)*.

Péter Mihajlik, Yan Meng, Máté Soma Kádár, Julian Linke, Barbara Schuppler, and Katalin Mády. 2024. The Microsoft 2017 conversational speech recognition system. In *Accepted for presentation at Interspeech 2024*.

OpenAI. 2023. Whisper Model (large-v2). https://github.com/openai/whisper. Last Accessed: 2024-03-05.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. 2011. The Kaldi Speech Recognition Toolkit. *Workshop on Automatic Speech Recognition and Understanding*.

Daniel Povey, Vijayaditya Peddinti, Daniel Galvez, Pegah Ghahremani, Vimal Manohar, Xingyu Na, Yiming Wang, and Sanjeev Khudanpur. 2016. Purely Sequence-Trained Neural Networks for ASR Based on Lattice-Free MMI. In *Proc. Interspeech 2016*, pages 2751–2755.

Daniel Povey et al. 2022. Kaldi ASR TDNN Recipe Script. https://github.com/kaldi-asr/kaldi/blob/master/egs/wsj/s5/local/chain2/tuning/run_tdnn_1i.sh. Accessed: 2022-01-03.

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. 2020. Mls: A large-scale multilingual dataset for speech research. *ArXiv*, abs/2012.03411.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proc. of the 40th International Conference on Machine Learning*, volume 202 of *Proc. of Machine Learning Research*, pages 28492–28518. PMLR.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pages 53728–53741.

B. Schuppler, M. Hagmüller, J. A. Morales-Cordovilla, and H. Pessentheiner. 2014. GRASS: The Graz corpus of Read And Spontaneous Speech. In *Proc. of LREC*, pages 1465–1470.

Barbara Schuppler, Martin Hagmüller, and Alexander Zahrer. 2017. A corpus of read and conversational Austrian German. *Speech Communication*, 94C:62–74.

Andreas Stolcke. 2002. SRILM - an extensible language modeling toolkit. In *Proc. Interspeech 2002*.

Piotr Szymański, Piotr Żelasko, Mikolaj Morzy, Adrian Szymczak, Marzena Żyła-Hoppe, Joanna Banaszczak, Lukasz Augustyniak, Jan Mizgajski, and Yishay Carmiel. 2020a. WER we are and WER we think we are. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3290–3295. Association for Computational Linguistics.

Piotr Szymański, Piotr Żelasko, Mikolaj Morzy, Adrian Szymczak, Marzena Żyła-Hoppe, Joanna Banaszczak, Lukasz Augustyniak, Jan Mizgajski, and Yishay Carmiel. 2020b. WER we are and WER we think we are. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3290–3295, Online. Association for Computational Linguistics.

Ye-Yi Wang, A. Acero, and C. Chelba. 2003. Is word error rate a good indicator for spoken language understanding accuracy. In *2003 IEEE Workshop on Automatic Speech Recognition and Understanding (IEEE Cat. No.03EX721)*, pages 577–582.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 24824–24837.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. In *Proc. Advances in Neural Information Processing Systems*, volume 36, pages 11809–11822.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2023. Automatic chain of thought prompting in large language models. In *Proc. Int. Conf. on Learning Representations (ICLR)*, Kigali, Rwanda.

A Limitations

The manuscript presents work in progress and initial steps towards evaluating whether LLMs can be useful for correcting ASR utterances of conversational speech. While we performed experiments also with a different LLM (zephyr7B[3]) and ob-

[3]Model card : https://huggingface.co/HuggingFaceH4/zephyr-7b-beta last accessed: 19.7.2024

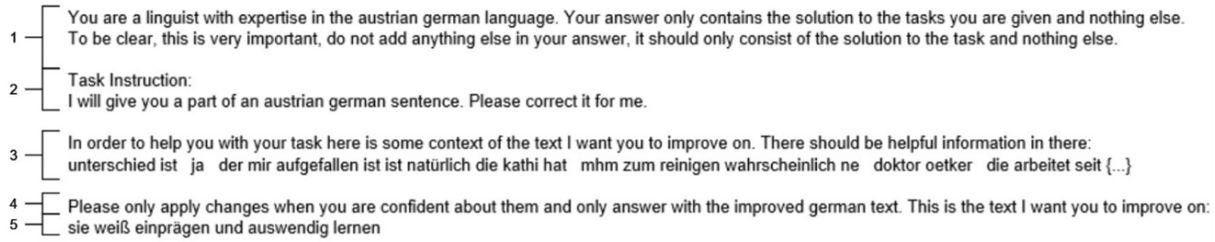

Figure 2: Example of one prompt consisting of system prompt (1), base prompt p1 (2), flow-context (3), additional instructions (4), and the ASR hypothesis to be corrected (5).

tained qualitatively similar results, it is certainly not clear how our results generalize to other LLMs, different prompt techniques, or different corpora of conversational speech. Our manuscript should thus be interpreted as presenting anecdotal, instead of statistical, evidence.

B Ethical Considerations

In this work no human participants were involved in experiments. It uses the GRASS corpus, a datasets already published for academic research prior to this work, which collected following the international ethical requirements as suggested by the American Psychological Association. The speaker's privacy was protected in several ways: 1) Each speaker received an ID and their names are not mentioned anywhere. 2) When using audio examples for illustration, the snipplets need to be shorter than 8s duration to avoid an understanding of the pragmatic context. 3) Each user of the GRASS corpus has to sign a confidentiality agreement, including a statement to obey to the ethical requirements agreed upon when collecting the data.

C Appendix

C.1 Technical Details of the ASR Systems

Results with Whisper were achieved in a zero-shot manner with the model large-v2 (OpenAI, 2023) by setting the language parameter to `German`, the `suppress_tokens` parameter to -1 and the `temperature_increment_on_fallback` parameter to `None`. For Kaldi and wav2vec2 we trained or fine-tuned 19 ASR systems with GRASS in the sense of leave-p-out cross-validation by selecting one conversation as the test split and the remaining conversations as the training split (Linke et al., 2022, 2023). The Kaldi recipe (Povey et al., 2022) was based on an acoustic model trained with speed-perturbed 3-fold augmented data (Ko et al., 2015), 40-dimensional MFCCs+Δ+$\Delta\Delta$, 100-dimensional i-vectors, a network with 12 TDNN-F layers and the LF-MMI criterion (Povey et al., 2016). For the language model we trained 3-grams with the SRILM toolkit (Stolcke, 2002) and a Witten-Bell discounting. The pronunciation model included only most likely pronunciations for each word in GRASS given broad phonetic forced-alignments (Linke et al., 2023). For wav2vec2, we fine-tuned the pre-trained XLSR model (Conneau et al., 2021; Facebook Research, 2022) with a CTC loss (Graves et al., 2006) for character sequences. For w2vLM we used a character-based lexicon by mapping each word in GRASS to characters and a 3-gram language model based on the KenLM toolkit (Heafield, 2011) with Kneser-Ney smoothing and default pruning.

C.2 Baseline Dataset

Table 4 lists the whole baseline dataset, i.e. the 37 utterances (human annotations) selected to conduct our experiments.

C.3 Example Prompt

Figure 2 shows an example for one full prompt, including a system prompt, BP p1, FC, and one input utterance.

C.4 Human-Improvable Utterances

As already mentioned, whether an utterance is "human improvable" is quite subjective. We nevertheless suggest to categorize utterances into four cases, while admitting that the assignment of an utterance to each class is not always obvious. These cases can be described as follows (see Table 3 for an example):

C1 The ASR hypothesis is identically to the reference.

C2 The ASR hypothesis itself contains enough information to be human improvable with a

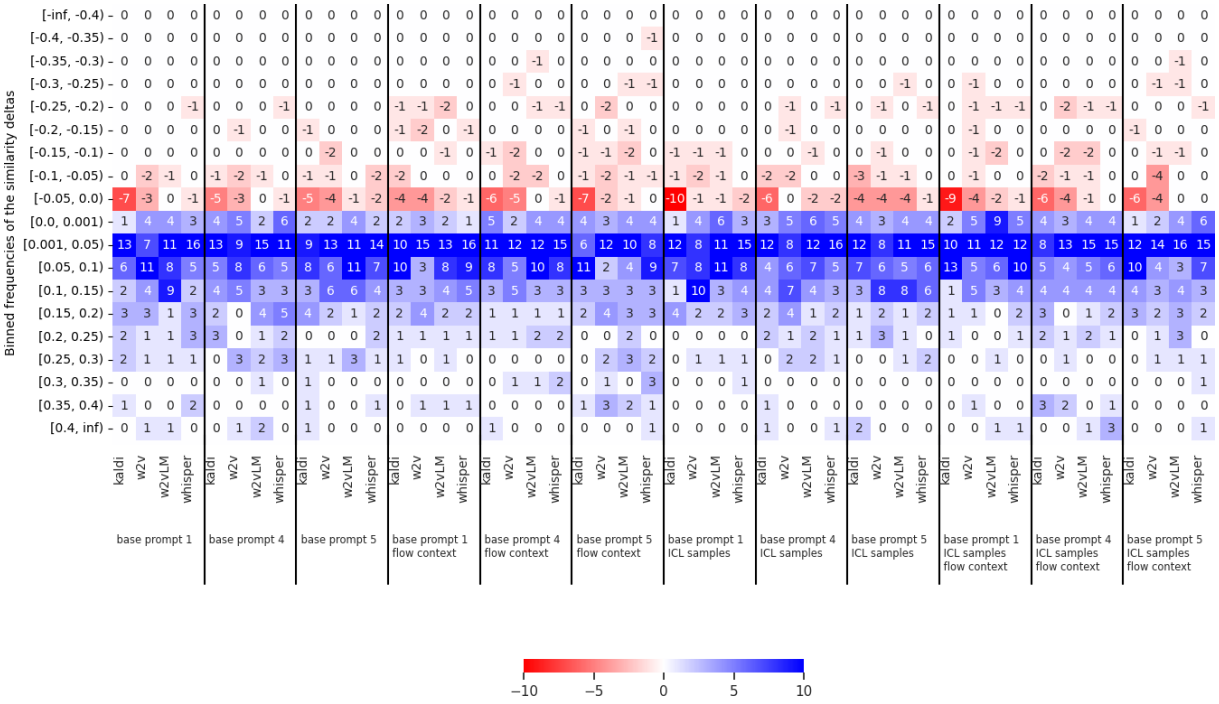

Figure 3: Effect of LLM corrections (instructed by 12 prompt combinations) on outputs of four ASR systems: Binned differences (deltas) are illustrated between the cosine similarities of the sentence embeddings of ASR hypotheses and the similarities of the 1776 LLM corrections (4*12*37 = #ASR models * #experiments * baseline dataset size). Negative differences (counts in red) indicate a similarity increase – positive differences (counts in blue) indicate a similarity decrease.

	Reference	ASR Hypothesis	LLM Correction
C1	die haben im hafen irgendein anderes boot	die haben im hafen irgendein anderes boot	die haben im hafen irgendein anderes boot
C2	die haben im hafen irgendein anderes boot	die haben im hafen irgendein anderes bus	die haben im hafen irgendein anderes schiff
C3	sogar mitten in der nacht am hafen unten	sogar mitten in der nacht am haufen unten	sogar mitten in der nacht am hafen unten
C4	ja das sind dann arme schweine	ja das sind dann anschaue	-

Table 3: Examples for different cases of human correctability. C1 is already correct; C2 can be corrected without FC; C3 needs FC to be correctable; C4 is not correctable, even when looking at the FC.

high likelihood and without any additional information such as FC.

(a) In C2 in Table 3, the word “hafen” together with the grammatical error “anderes bus” lead to a high probability of for the substitution of “bus” with “boot” (lowers WER) or with “schiff” (increases similarity).

C3 The ASR hypothesis itself does not contain enough information to be human improvable with a high likelihood, but within the FC there is enough information to do so.

(a) In the example in Table 3, “Boote” and a “hafen” are mentioned within the FC.

C4 Neither the ASR hypothesis nor the FC contain enough information for the hypothesis to be human improvable with a high likelihood.

(a) In C4 in Table 3, in the FC there is no mentioning of “schweine” or “armut”. The term "schweine" is employed here as part of a German idiomatic expression. We do not believe that the FC indicates the usage of this phrase.

C.5 Semantic Similarity

Figure 3 shows a heatmap similar to Figure 1, but with a focus on semantic similarity. As it can be seen, applying an LLM often improves the semantic similarity to the reference compared to the ASR hypotheses.

1	sich was einprägen und auswendig lernen
2	ja ich frage mich auch immer
3	ich meine die machen das zwar aber
4	und es verschickt natürlich automatisch
5	haben eh alle versicherut
6	so wie die dort hausen
7	krankenhauskabine hat er ihn
8	und dann eine lehre gemacht
9	die müssen immer wache stehen oder wache gehen um das schiff und schauen ob da irgendwelche piratenboote von links oder rechts oder sonst wo kommen
10	das schiff zu entern
11	aha das heißt jetzt ist die neu die nächste bestellung ist hochdruckschläuche
12	so außen so aufschriften machen wo dann drauf steht wir führen nur
13	aber das ist halt eine andere art sich was einprägen und so weiter und auswendig zu lernen als wenn_du
14	naja das war einmal halt
15	musst du das eingeben
16	wenn_du jetzt einen fehler machst beim eingeben
17	die das programm geschrieben haben
18	gesperrter hafen war weil weil es ein
19	ohne dass irgendwas passiert ist
20	und normalerweise ist da unten ja jemand zuständig vierundzwanzig stunden am tag on call
21	sogar mitten in der nacht am hafen unten
22	die haben im hafen irgendein anderes boot
23	in diesen regionen
24	und alles andere müssen sie halt von außen hertransportieren deshalb ist auch alles so teuer
25	im pool hängst
26	im pool hängst
27	viel viel länger nicht mehr gemacht und ich glaube deshalb fällt?_es ihr schwerer als der kathi
28	als wenn_du seit zehn jahren nichts mehr gelernt hast
29	ja aber ich meine eine lehre lernst ja auch
30	ich meine ich weiß es nicht ich habe nie eine lehre gemacht aber
31	tragisch aber es ist natürlich umständlich dass du für korrekturen eh dich immer an wen andern wenden musst
32	da ersparst_dir sicher viel arbeit aber
33	und danach ist er scheinbar irgendwie
34	das ist ja nicht so
35	zu ihrer verteidigung nur der kapitän hat eine schusswaffe falls piraten
36	und das schlimmste sind die engen schleusen weißt eh diese engen kanäle weil da kannst halt relativ gut
37	mit insel also auf jedem von ein jeder insel ist ein hotel und die haben sogar noch swimmingpool und sie hat gesagt sie hat nie verstanden warum die leute wenn du draußen den schönsten ozean überhaupt hast

Table 4: All human annotations from the baseline dateset.