Human Latency Conversational Turns for Spoken Avatar Systems

Derek Jacoby
University of Victoria
Victoria, BC, Canada

Tianyi Zhang
University of Victoria
Victoria, BC, Canada

Aanchan Mohan
Northeastern University
Vancouver, BC, Canada

Yvonne Coady
Northeastern University
Vancouver, BC, Canada

ABSTRACT

A problem with many current Large Language Model (LLM) driven spoken dialogues is the response time. Some efforts such as Groq address this issue by lightning fast processing of the LLM, but we know from the cognitive psychology literature that in human-to-human dialogue often responses occur prior to the speaker completing their utterance. No amount of delay for LLM processing is acceptable if we wish to maintain human dialogue latencies.

In this paper, we discuss methods for understanding an utterance in close to real time and generating a response so that the system can comply with human-level conversational turn delays. This means that the information content of the final part of the speaker's utterance is lost to the LLM. Using the Google NaturalQuestions (NQ) database, our results show GPT-4 can effectively fill in missing context from a dropped word at the end of a question over 60% of the time. These results indicate that a simple classifier could be used to determine whether a question is semantically complete, or requires a filler phrase to allow a response to be generated within human dialogue time constraints.

1 INTRODUCTION

In this paper, we would like to explore some methods of evaluating spoken language avatar dialogue systems with respect to human-to-human dialogues. We will discuss response quality, and some mechanisms of evaluating that quality, but our ultimate focus will be on latencies in human dialogues and corresponding latencies in automated systems. In a 2009 study on latencies in conversational turn taking [23] it was found that on average in English there is a 239 msec delay between the original interlocutor in a dialogue and the start of the answering utterance. The variance is quite high, though, with one standard deviation in response time being 519 msec. So human dialogue expectations are that responses occur between -280 and +758 msecs from the end of an utterance, on average. In other languages, these expectations of response time are even more challenging for automated systems to meet, for instance in Japanese the average response is only 7 msec after the original speaker finishes speaking.

The current standard for architectures for automated dialogue systems is one where the speech recognition pass begins when the utterance ends, and then the recognized speech is passed to a Large Language Model (LLM) which composes the response to be sent to a synthetic speech generation module. This architecture is completely unable to conform to the latency expectations in human dialogues. This serial processing of responses is also counter to the means by which humans process speech. Humans form an understanding in real time and engage in a number of complex turn-taking behaviours to negotiate when to take control of the conversation [16]. In this paper we will not delve too deeply into those cues, such as eye contact and non-verbal utterances, which might signal a desire to begin to speak, but we will concentrate on the timings when it is turn for the system to speak. Violating the conversational expectations of human to human dialogue detracts from the perception of naturalness and engagement in human-machine spoken dialogues, although users do have some tolerance for longer delays than would be acceptable in a human conversational partner[19].

There is a deep literature on real-time speech recognition systems which will be briefly reviewed in the related work section. In general, these systems lose some accuracy in comparison with speech recognition systems that can search from both ends of an utterance in generating its recognition candidates. We will discuss semantic redundancy in dialogue utterances, though, and provide experimental evidence that in many cases the loss of recognition accuracy is not a detriment to the quality of spoken avatar responses. We will also discuss heuristics for the use of filler phrases when the construction of an utterance leads the avatar system to suspect that critical information is being withheld until the end of an utterance. Finally, we will speculate on the impact of these conversational behaviours in the construction of spoken dialogues in the construction of an avatar that we are in the midst of building to support engaging the general public in a museum experience.

To summarize, the realization that this work is being driven from is that in American English a conversational turn generally begins from 239 msec before the first speaker is done speaking to about 758 msec after. There have been attempts to on each dialogue turn use a filler phrase of some sort to start the automated dialog response while the system runs the speech recognition and generates the response, but this gets unnatural quickly. Our current attempt is to take a question answering dataset, chop the final 1, 2 or 3 words off the initial utterance and generate a response, use an automated framework to judge the quality of the response, and then model how much accuracy we lost on the responses by lopping the initial question off early. Next, to use the instances that lost accuracy to form a binary classifier so that when we are part way through the question we can take the classifier output to determine whether we should respond with a filler phrase and then process the entire question (late informative questions) or skip the filler phrase and process an answer based on the truncated question (late uninformative questions). The hypothesis being that by classifying and making a decision on whether to complete processing that we can recover answer quality with a minimum of use of filler phrases by taking

advantage of a statistical understanding of sentence structure to determine the likelihood of a twist at the end of the sentence.

We are taking a statistical pattern recognition approach to this problem, there have also been linguistic processing approaches related to our efforts, but they require a deeper semantic modeling of the input phrases which seem to require more hand-processed rules to execute [25]. This rule-based versus statistical approach in some ways mirrors ongoing theoretical positions within the linguistics community which we will elaborate upon in related work.

2 BACKGROUND AND RELATED WORK

In this review we are going to examine different methods for assessing machine dialogues including some discussion of data sets. Next we will review some latencies and factors affecting latency in human-to-human dialogues, this will also include a discussion of underlying neural correlates of speech in humans. Finally we will identify some of the limiting factors in the latency of machine dialogues, and detail some previous attempts to overcome these barriers.

2.1 Assessing machine dialogues

We later propose some methods for speeding up machine dialogues, but it is not possible to do this without first carefully considering the impact on the quality of those dialogues. There are a variety of ways of assessing dialogue quality, and this has an interplay with the type of dataset used for that assessment. In our case, we will be mostly using the Google NaturalQuestions database [14] which has large number of factual questions (drawn from Wikipedia) with short, long, and yes/no answers supplied by human annotators. This gold standard approach is effective in that a human annotator supplies the ground truth, but this annotation is expensive. This particular data set has, in some ways, been supplanted by more challenging tasks [22] but we have elected to set a baseline on our approach before going to a more variable dataset.

Some approaches have tried to automate the comparison of human responses to the automatically generated responses[5]. The other alternatives are reference-free approaches, for instance LLM-eval [15] is a common reference-free approach that uses a multi-dimensional set of automated metrics to evaluate dialogue quality. We have chosen to use a method that relies on a human annotated gold standard, but uses an LLM to calculate a semantic similarity score, semScore, to that gold standard [24]. This method, LLM-as-judge, allows us to compare against the human annotated standard, and then from there against our questions manipulations in each case giving a 0 to 1 score for how semantically similar the two answers are based on a cosine distance measurement.

2.2 Latencies in Human-to-Human Dialogue

The analysis of human-to-human dialogue latencies is complex both in terms of which language is under discussion, and what the intent of the dialogues are [23]. A more in-depth analysis of dialogue intents, and the means of grounding dialogues, is found in a book chapter on Cognitive Mathematics, which also does a good job of describing the types of dialogues [18]. In our case, we are focusing on information seeking dialogues.

Even the measurement of turn taking has some complexities, and are referred as overlaps (where the reply begins before the prior utterance is finished) and gaps (where there is some silence between turns) [16]. Some Scandinavian studies have re-enforced our ideas of the importance of smooth turn taking in establishing pleasing machine dialogues in both a paper [13] and a doctoral thesis [11].

2.3 Spoken Dialogue Theories

As far back as the mid-1990's researchers began studying the nature of human-machine spoken dialogues [12] Given the technology of the day, this involved Wizard of Oz studies, with a human with a script simulating the machine side of the dialogue. A fundamental realization of that, and most subsequent studies, is that human-machine dialogues are simply different that human-to-human dialogues. The expectations are different and the interlocutory acts are different. We maintain that this is an artifact of the state of the technology and that as this technology improves that human-machine spoken dialogue will more closely resemble the dialogue between people. Of course, some linguistic theorists would ardently disagree, and in fact would maintain that today's statistically-generated dialogue behaviours are fundamentally different to human language mechanisms, and thus a natural dialogue is largely impossible [8]. In fact, the very existence of Large Language Models in some ways threatens the Chomsky view that language depends on innate structures, and is not simply an emergent property of a probabilistic system [20].

There is a class of sentence that is studied which is known as a sluiced sentence. This means that there is a portion of the sentence that is ommitted and filled in from context. This type of sentence is very common where the context exists in a proximal sentence and requires additional processing by the person listening to fill in the missing context [9]. One can consider the questions where we are truncating the last words as a form of sluiced sentence. If the context is too hard to recover, however, processing of the question will fail.

2.4 Latencies in Machine Dialogues

Most spoken dialogue avatar systems attempt to simply process as quickly as possible in order to generate a response within an acceptable amount of time for the user. The elements of that processing generally include the speech recognition pass (which for the best recognition quality cannot start until the utterance is complete), followed by a language understanding pass (generally an LLM currently) to produce the reply, followed by a Test-to-Speech (TTS) system to generate the output audio. Each of these steps take time.

Let us first consider the recognition speed. The OpenAI whisper speech recognition engine has made a large impact on the community, but generally provides only offline recognition - meaning that the entire utterance is processed at once. Some efforts have been made to perform online, or ongoing, recognition using whisper [17]. In many situations it would be ideal to use a cloud service for the speech recognition, but there are large differences in the latency of the offerings by different providers. Microsoft tends to be the faster in providing both initial hypotheses and final recognitions [3]. In

general on Azure, the first hypotheses come back within 150ms and those hypotheses then to be 95% stable within 500 msec. So we could depend on final recognitions within approximately 650ms from the end of the utterance. Google and IBM on their cloud services would take on the order of two seconds to produce their final recognitions.

The next item to look at is the response generation. LLMs can very widely in their latencies, but one of the fastest is the new Groq API. There are some restrictions, but on the open source models that they have optimized 240 tokens per second is not unreasonable with only a few msec of initial latency in the response whereas the speed of generation on OpenAI is closer to 94 ms per token [1, 2]. Given that the responses in this study range from about 20 tokens to about 60 tokens in length, the response generation time in the worst cases will be approximately 250 msec on Groq or 650 msec on OpenAI.

On the speech production side, numbers are very situation dependent but we have found that between 80 and 100 msec response times are a reasonable estimate, and this aligns with other experiences in the literature [4].

So an overall response time, independent of network delays, is likely in the 1000 to 1500 msec range for most questions.

3 METHODOLOGY AND EXPERIMENTAL FRAMEWORK

This section gives an overview of the experimental setup for this work. Our goal is to simulate two cases in human dialogue when it comes to question answering, namely the late informative question scenario and late uniformative question scenario. We would then like to understand the impact this has on responses generated by an LLM. This section first briefly describes the Google NaturalQuestions (NQ) dataset, followed by our setup to mimic the two stated question answering scenarios. Furthermore, we describe the process of scoring similarities between responses generated by the LLM for the different scenarios compared to the ground-truth and the baseline replies generated by the LLM.

3.1 Google NaturalQuestions Dataset

The Google NaturalQuestions(NQ) dataset [14] as mentioned before consists of answers to a large number of factual questions. The dataset consists of 307,383 training examples, 7830 development examples which contain human annotations and 7842 test examples. For our experimental study we chose the first 1,000 examples in the development set as our subset for the purposes of this experimental study. Each example consists of a question along with a long answer that is generated by human annotators which we treat as the gold-standard answer (which will be referred to as 'ref'). We dropped those questions which were negative controls (meaning that they were designed not to be answerable only from the question text). We also intentionally did not use the context text from the dataset, instead preferring to use only the background training data from ChatGPT to provide context for the question. This was to not favorably bias the correct answering of questions from the context provided.

3.2 Responses from Large Language Models

In our experimental study we used ChatGPT [7] as our LLM to generate responses for each of the 1,000 example questions taken from the NQ dataset. For each question in our subset, the response that that is generated our LLM is referred to as 'res-0' or response-0. This refers to responses without any sentence truncation in the original question. This response simulates our late informative response scenario in natural dialogue, where a speaker giving the answer listens to the entire phrase before responding.

In order to simulate the long uninformative response scenario, where a speaker giving the answer starts before the question is finished being said. The typical speaking rate for English is 4 syllables per second [10]. It is well known that some of the most common words in the English language are between 1-7 syllables in the length with a large number of words being not more than 3-4 syllables long. Considering this, we consider two subsets of our 1,000 example subset. In the first subset, we take each of the examples and remove the last word in the each question and record the responses from our LLM which we refer to as 'res-1'. This simulates removal of approximately between 500ms-1sec. of speech audio. Similarly, in our second subset, we truncate the last two words in each question and record the responses from our LLM which we refer to as 'res-2'. This simulates the removal of approximately 1sec or more of speech audio. In our third subset we truncate the last 3 words and we refer to this as 'res-3'.

3.3 Scoring similarity between responses

In order to evaluate the quality of responses we use SEMSCORE [6] a measure of semantic textual similarity. In the calculation of the SEMSCORE, the ground truth response to a question, the target response, and the response from the LLM, the so-called model response are both converted to sentence embeddings using a sentence transformer model 'all-mpnet-base-v2' [21]. The SEMSCORE then consists of calculating the cosine similarity between the ground truth target response embedding and the embedding corresponding to the LLM response. The value of the SEMSCORE lies between [-1,1]. If the scores are close to 1, then the two sentences are considered semantically similar. Negative values imply semantically opposite sentences.

4 RESULTS

This section describes our experiments. In order to understand how closely the responses from the LLM match to the ground truth responses, the SEMSCORE is first calculated between the responses received from the LLM ('res-0') and the human annotated ground truth responses ('ref'). Figure 1 shows a histogram plot of the SEMSCORES from our set 'res-0' scored against the ground truth responses. The LLM used to generate responses in this figure is ChatGPT with the GPT-4 model. The histogram is seen to provide a central tendency of scores closer to a mean of 0.68 to indicate that the returned responses and the ground truth responses bear semantically similarity. The standard deviation is 0.16 with the minimum SEMSCORE being 0.03, and the maximum SEMSCORE being 0.97. The 75th percentile of these scores is 0.81.

In order to understand the impact of word truncation, our next experiment looks at the distribution of scores for examples whose

complete responses from the LLM very closely matched the human annotated ground-truth responses. For this we selected those 'res-0' examples whose 'res-0' vs 'reference' scores were above the 75th percentile score of 0.81. We then plotted the SEMSCOREs for these examples comparing :

- The original response from the LLM ('res-0') and its similarity compared to the response obtained when the last word was removed from the original question ('res-1').
- The original response from the LLM ('res-0') and its similarity compared to the response obtained when the last two words were removed from the original question ('res-2').
- The original response from the LLM ('res-0') and its similarity compared to the response obtained when the last three words were removed from the original question ('res-3').

The distribution of these scores for these specific examples are captured in the box and whisker plots in Figure 2. The counts of scores rated above 75th percentile for each truncation condition are shown in Figure 3.

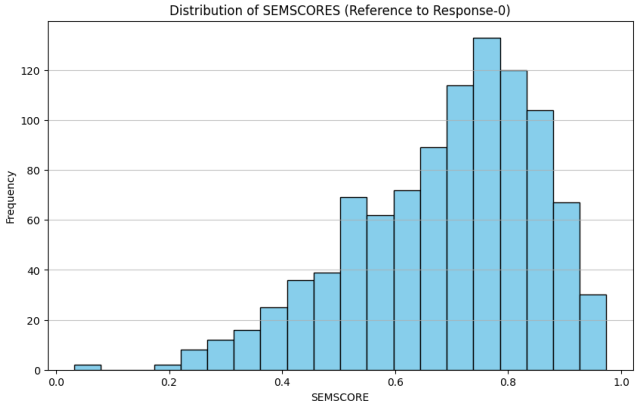

Figure 1: Histogram of SEMSCORES for LLM returned responses against the ground truth reference responses

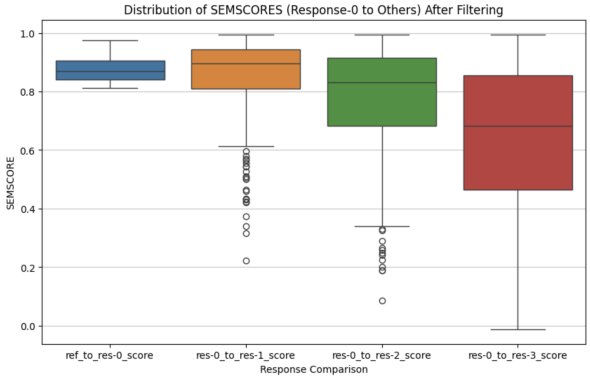

Figure 2: Box and whisker plots of SEMSCORES for the similarity of 'res-0' vs truncated utterances, specifically those utterances whose 'res-0' scores when compared to the ground truth were above the 75th percentile score of 0.81

.

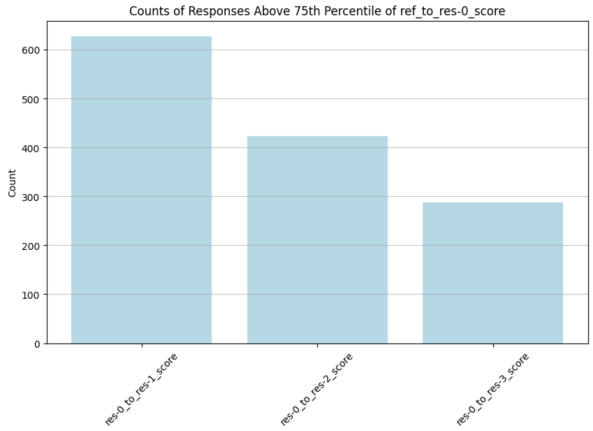

Figure 3: Number of questions (out of 1000) where the response was rated within 75th percentile of untruncated responses

.

5 DISCUSSION

We have presented data showing that the use of state of the art LLM's such as GPT-4 can quite effectively fill in missing context from dropped words in a question. In the case of a single dropped word, this results in little impact on the quality of the response over 60% of the time. In pursuit of human latencies in spoken avatar dialogues, we are in the midst of taking advantage of this in the preparation of filler phrases which will allow the avatar to respond within latency expectations if it is still processing. For instance, in the question "How do I get from New York to Chicago?" plainly the question cannot be answered until after the entire question is uttered. This is an example of a late informative question, and a human conversational partner might answer "Well, the way I would go would be to take interstate..." with the first part of that response ("Well, the way I would go") essentially playing the role of a filler phrase while the rest of the response is composed. If the question were asked as "I want to drive from New York to Chicago, can you give me directions?" then no filler phrase would be needed.

In some sense, the capabilities of the language model to fill in missing context (deal with sluiced sentences) determines how often filler phrases will be needed to maintain appropriate conversational norms. That GPT-4 is so superior to previous language model variants allows us to rely less on context directly in the questions and instead depend on the LLM to fill in the gaps. What this gap filling cannot do, though, is allow the avatar to engage in other conversation mediating activities. A human conversational partner will make eye contact, nod, make non-verbal utterances of agreement in an attempt to re-assure the speaker that their message is being received. It is only if the speech recognizer is running on the fly and the avatar paying attention to the semantics of the question that these types of conversation reinforcing behaviors can be produced.

The production of filler phrases and adherence to human conversational norms is only one of the small benefits of processing, as humans do, while the conversational turn is ongoing.

The use of heuristics, like filler phrases and rules for semantic non-interrupting responses, can be thought of as something of a bridge between purely statistical language learning in the LLM, and higher level language rules that are so prominent in Chomsky's approach to language [8].

6 CONCLUSIONS AND FUTURE WORK

This work is ongoing as we develop an avatar for responding to questions about the environment that will be on display in a museum setting. Before we release the avatar we will need to move our current text-based investigations into spoken language inputs and responses. There are several datasets of spoken language questions, but we intend to develop our own that are in-domain with the environmental questions that we expect our users to ask.

The work to build a classifier to allow us to determine on an ongoing basis as the user speaks whether the question will be late informative is work that is upcoming. Specifically, we intend to use the questions in our dataset that were labeled late informative through the use of the semScore measure differing from complete questions to train a classifier. This will allow us to develop a score that tends towards 1 for identifying when a question is semantically complete, and as we pass a cutoff that we define we can then prepare the response. In those cases where the question ends without a response being ready we will have used the intermediate recognitions (and semantic understanding) to generate an appropriate filler phrase with as much specificity as the question allows.

Eventually, we also expect to use these intermediate semantics to introduce dialogue reinforcing behaviors. This is particularly the case as our graphical avatar is produced to go along with our responses.

The mix of systems latency questions and user expectations make this an exciting time to continue to craft machine conversational systems that meet user expectations of human conversational partners.

REFERENCES

[1] 2023. GPT-3.5 and GPT-4 response times. https://www.taivo.ai/__gpt-3-5-and-gpt-4-response-times/

[2] 2024. ArtificialAnalysis.ai LLM Benchmark Doubles Axis - Groq. https://wow.groq.com/artificialanalysis-ai-llm-benchmark-doubles-axis-to-fit-new-groq-lpu-inference-engine-performance-results/ Section: Blog.

[3] Angus Addlesee, Yanchao Yu, and Arash Eshghi. 2020. A Comprehensive Evaluation of Incremental Speech Recognition and Diarization for Conversational AI. In *Proceedings of the 28th International Conference on Computational Linguistics*, Donia Scott, Nuria Bel, and Chengqing Zong (Eds.). International Committee on Computational Linguistics, Barcelona, Spain (Online), 3492–3503. https://doi.org/10.18653/v1/2020.coling-main.312

[4] Syed Rameel Ahmad. 2024. Enhancing Multilingual Information Retrieval in Mixed Human Resources Environments: A RAG Model Implementation for Multicultural Enterprise. https://doi.org/10.48550/arXiv.2401.01511 arXiv:2401.01511 [cs].

[5] Chris Alberti, Kenton Lee, and Michael Collins. 2019. A BERT Baseline for the Natural Questions. http://arxiv.org/abs/1901.08634 arXiv:1901.08634 [cs].

[6] Ansar Aynetdinov and Alan Akbik. 2024. SemScore: Automated Evaluation of Instruction-Tuned LLMs based on Semantic Textual Similarity. http://arxiv.org/abs/2401.17072 arXiv:2401.17072 [cs].

[7] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. https://doi.org/10.48550/arXiv.2005.14165 arXiv:2005.14165 [cs].

[8] Noam Chomsky, Ian Roberts, and Jeffrey Watumull. 2023. Noam chomsky: The false promise of chatgpt. *The New York Times* 8 (2023). https://edisciplinas.usp.br/pluginfile.php/7614933/mod_resource/content/1/Opinion%20_%20Noam%20Chomsky_%20The%20False%20Promise%20of%20ChatGPT%20-%20The%20New%20York%20Times.pdf

[9] Lyn Frazier Charles Clifton. 1998. Comprehension of Sluiced Sentences. *Language and Cognitive Processes* 13, 4 (Aug. 1998), 499–520. https://doi.org/10.1080/016909698386474

[10] Alan Cruttenden. 2014. *Gimson's pronunciation of English.* Routledge. https://www.taylorfrancis.com/books/mono/10.4324/9780203784969/gimson-pronunciation-english-alan-cruttenden

[11] Anna Hjalmarsson. 2010. Human interaction as a model for spoken dialogue system behaviour. (2010). https://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-24258 Publisher: KTH.

[12] Anne Johnstone, Umesh Berry, Tina Nguyen, and Alan Asper. 1995. There was a long pause: influencing turn-taking behaviour in human-human and human-computer spoken dialogues. *International Journal of Human-Computer Studies* 42, 4 (April 1995), 383–411. https://doi.org/10.1006/ijhc.1995.1018

[13] Gudny Ragna Jonsdottir, Kristinn R. Thorisson, and Eric Nivel. 2008. Learning Smooth, Human-Like Turntaking in Realtime Dialogue. In *Intelligent Virtual Agents*, Helmut Prendinger, James Lester, and Mitsuru Ishizuka (Eds.). Springer, Berlin, Heidelberg, 162–175. https://doi.org/10.1007/978-3-540-85483-8_17

[14] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural Questions: A Benchmark for Question Answering Research. *Transactions of the Association for Computational Linguistics* 7 (Nov. 2019), 453–466. https://doi.org/10.1162/tacl_a_00276

[15] Yen-Ting Lin and Yun-Nung Chen. 2023. LLM-Eval: Unified Multi-Dimensional Automatic Evaluation for Open-Domain Conversations with Large Language Models. In *Proceedings of the 5th Workshop on NLP for Conversational AI (NLP4ConvAI 2023)*, Yun-Nung Chen and Abhinav Rastogi (Eds.). Association for Computational Linguistics, Toronto, Canada, 47–58. https://doi.org/10.18653/v1/2023.nlp4convai-1.5

[16] Rebecca Lunsford, Peter A. Heeman, and Emma Rennie. 2016. Measuring Turn-Taking Offsets in Human-Human Dialogues. In *Interspeech 2016*. ISCA, 2895–2899. https://doi.org/10.21437/Interspeech.2016-1350

[17] Ke-Ming Lyu, Ren-yuan Lyu, and Hsien-Tsung Chang. 2024. Real-time multilingual speech recognition and speaker diarization system based on Whisper segmentation. *PeerJ Computer Science* 10 (March 2024), e1973. https://doi.org/10.7717/peerj-cs.1973 Publisher: PeerJ Inc..

[18] Bernardo Magnini and Samuel Louvan. 2022. Understanding Dialogue for Human Communication. In *Handbook of Cognitive Mathematics*, Marcel Danesi (Ed.). Springer International Publishing, Cham, 1159–1201. https://doi.org/10.1007/978-3-031-03945-4_20

[19] Zhenhui Peng, Kaixiang Mo, Xiaogang Zhu, Junlin Chen, Zhijun Chen, Qian Xu, and Xiaojuan Ma. 2020. Understanding User Perceptions of Robot's Delay, Voice Quality-Speed Trade-off and GUI during Conversation. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems (CHI EA '20)*. Association for Computing Machinery, New York, NY, USA, 1–8. https://doi.org/10.1145/3334480.3382792

[20] Steven Piantadosi. 2023. Modern language models refute Chomsky's approach to language. *Lingbuzz Preprint, lingbuzz* 7180 (2023). https://lingbuzz.net/lingbuzz/007180/current.pdf

[21] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. http://arxiv.org/abs/1908.10084 arXiv:1908.10084 [cs].

[22] Corby Rosset, Ho-Lam Chung, Guanghui Qin, Ethan C. Chau, Zhuo Feng, Ahmed Awadallah, Jennifer Neville, and Nikhil Rao. 2024. Researchy Questions: A Dataset of Multi-Perspective, Decompositional Questions for LLM Web Agents. https://doi.org/10.48550/arXiv.2402.17896 arXiv:2402.17896 [cs].

[23] Tanya Stivers, N. J. Enfield, Penelope Brown, Christina Englert, Makoto Hayashi, Trine Heinemann, Gertie Hoymann, Federico Rossano, Jan Peter de Ruiter, Kyung-Eun Yoon, and Stephen C. Levinson. 2009. Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences of the United States of America* 106, 26 (June 2009), 10587–10592. https://doi.org/10.1073/pnas.0903616106

[24] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. https://doi.org/10.48550/arXiv.2306.05685 arXiv:2306.05685 [cs].

[25] Jiawei Zhou, Jason Eisner, Michael Newman, Emmanouil Antonios Platanios, and Sam Thomson. 2022. Online Semantic Parsing for Latency Reduction in Task-Oriented Dialogue. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, Dublin, Ireland, 1554–1576. https://doi.org/10.18653/v1/2022.acl-long.110