

Speechworthy Instruction-tuned Language Models

Anonymous ACL submission

Abstract

Current instruction and human preference data do not represent speech-based human preferences because they are almost exclusively collected with a text interface. Hence, instruction-tuned language models (ITLM) trained with this data are suboptimal for generating responses suitable for transforming with text-to-speech systems, yet voice assistants such as Google Assistant or Amazon Alexa are increasingly relying on these responses as direct responses to user queries. We first empirically demonstrate that current ITLMs are suboptimal due to the shift in modality from text to speech. Then, we address their suboptimality by finetuning with preference annotations that are collected by having annotators *listen* to response pairs and thus accurately reflect their speech-based preferences. With only 20K speech-based preference annotations, we develop a speech-centric instruction-tuned language model by finetuning Falcon-Instruct 7B with reinforcement learning with human feedback. Our finetuned model, SpeechFalcon, generates responses that are on average 4.8% more frequently preferred than those from the base model and its prompted counterpart and 19.7% more frequently preferred than the human-written responses from the original instruction dataset when controlled for factuality.

1 Introduction

Instruction-tuned language models (ITLM) are finetuned large language models (LLMs) that follow instructions and they have gained prominence due to their broad applicability in accomplishing tasks without explicit supervision (Brown et al., 2020; Rae et al., 2021; Lieber et al., 2021; Hoffmann et al., 2022). However, speech-based interactions have been largely overlooked in this context, despite being a core factor that predates text and humanizes intelligent systems (Yang, 2004). While prior works have examined the ability of ITLMs

User's query: How can I make cream of mushroom soup?	
Preferred response for text	Preferred response for audio
<p>Certainly! Making cream of mushroom soup is a delicious and comforting process. Here's a simple recipe for you:</p> <p><i>Ingredients:</i></p> <ul style="list-style-type: none">• 1/4 cup unsalted butter• ... [9 more items] <p><i>Instructions:</i></p> <ol style="list-style-type: none">1. In a large pot, melt the butter over medium heat.2. ... [9 more steps] <p>Enjoy your homemade cream of mushroom soup! Feel free to add other spices to suit your taste.</p>	<p>To make cream of mushroom soup, you'll need mushrooms, onion, butter, flour, salt, and pepper.</p> <p>Are you ready to move on with the steps?</p>

Figure 1: Current instruction-tuned language models tend to be verbose and generate complex sentences and non-verbal writing devices, such as bullet lists or parentheses, that are not suitable for speech-based interactions (left). Speech is serial and transient, and therefore concise and simple yet informative sentences are preferred (right).

to recognize speech (Hori et al., 2018; Chiu et al., 2018), synthesize speech (Wang et al., 2017), or correct speech recognition errors (Mousavi et al., 2024), far less attention has been paid to how suitable the *content* of the response is when spoken. Identifying and increasing the "speech-suitability" of ITLM-generated content – the main focus of this work – is critical to the development of equitable digital assistants as ITLMs become the backbone of such systems today (Nass and Brave, 2005).

Speech as a second-class citizen stems from training ITLMs with textual instruction-following datasets (Peng et al., 2023; Chung et al., 2022; Wang et al., 2022a,b; Wei et al., 2021; Sanh et al., 2022; Zhou et al., 2023) and preference annotations between text pairs (Bai et al., 2022; Ethayarajh et al., 2022; Ouyang et al., 2022; Touvron et al., 2023) to optimize for human preferences.

060	Responses that maximize human preference in text	111
061	may—and intuitively, should—be different from	112
062	those that do the same for speech. For example,	113
063	verbosity and complex vocabulary may be less im-	114
064	portant in text because users can reread content that	115
065	is difficult to understand. Speech, however, is in-	116
066	herently serial and transient; users cannot as freely	117
067	or easily navigate spoken content. Additionally,	118
068	formats that are preferred in the text domain, such	
069	as bulleted lists, code blocks, and math formulas,	
070	are incomprehensible when spoken (Kern, 2008;	
071	Abel, 2015; Dowling and Miller, 2019; Kolodzy,	
072	2012).	
073	The current emphasis on text for ITLM de-	
074	velopment together with the unique constraints	
075	of the audio domain result in ITLMs that are	
076	poorly optimized for spoken dialogue. For in-	
077	stance, ITLMs that are optimized for helpfulness	
078	using RLHF consistently produce more verbose	
079	outputs (Prasann Singhal, 2023). As illustrated in	
080	Figure 1, in which the top response is detailed,	
081	lengthy, and contains abundant lists and paren-	
082	theses, this verbosity is likely to have a negative	
083	impact on listener preference. To create “speech-	
084	suitable” ITLMs—i.e. those optimized for listener	
085	preference—such as the one shown in bottom re-	
086	sponse of Figure 1, we develop alignment efforts	
087	specifically for speech.	
088	First, we conduct an extensive survey of spoken	
089	ITLM responses, which verifies that the responses	
090	of current ITLMs are ill-suited to the spoken do-	
091	main. Our survey additionally characterizes vari-	
092	ation in multiple dimensions of response quality	
093	between audio and text responses, finding current	
094	text responses are considered to be too lengthy,	
095	contain too much information, and are generally	
096	difficult to understand when spoken.	
097	Next, we explore aligning ITLMs for speech	
098	interactions through prompt engineering (Lester	
099	et al., 2021) and RLHF (Stiennon et al., 2020;	
100	Bai et al., 2022). To develop a reward model	
101	for RLHF, we collect extensive preference anno-	
102	tations by presenting annotators with spoken re-	
103	sponse pairs, a first to the best of our knowledge.	
104	In addition, for efficient RLHF, we use a set of	
105	system prompts inspired by rules-of-thumb from	
106	the radio industry (Kern, 2008; Abel, 2015), pod-	
107	casting (Dowling and Miller, 2019), multimedia	
108	journalism (Kolodzy, 2012), and relevant literature	
109	on listenability (Chall and Dial, 1948; Fang, 1966;	
110	Messerklinger, 2006; Kotani et al., 2014) to gener-	
	ate sample responses that span varying degrees of	111
	speech-suitability for our preference annotations.	112
	Lastly, we conduct human and automatic evalua-	113
	tions on our speech-aligned model and find that it	114
	outperforms both the base model and its prompt	115
	engineered version.	116
	In summary, our main contributions are the fol-	117
	lowing:	118
	• To the best of our knowledge, we are the	119
	first work to examine the suitability of ITLM-	120
	generated content for speech-based interac-	121
	tions.	122
	• We establish that current ITLM responses are	123
	not optimized for speech and that this adapta-	124
	tion is not trivially achieved through prompt	125
	engineering.	126
	• We collect 20K speech-based preference an-	127
	notations that span varying degrees of speech-	128
	suitability.	129
	• We finetune a ITLM to generate speech-	130
	suitable responses via RLHF using sampled	131
	responses.	132
	2 Instruction-tuned Language Model	133
	Responses Are Not Suitable for Speech	134
	2.1 Lessons from audio-based media	135
	We initiate our research with the key questions	136
	“Are ITLM responses that were suitable in text also	137
	suitable when spoken?” and “If not, how should	138
	spoken responses be different?”. Kern (2008) and	139
	Abel (2015) provide a glimpse to the answers for	140
	these questions by illustrating numerous examples	141
	of how news reporting and storytelling in text dif-	142
	fers from audio. Through decades of radio experi-	143
	ence, they establish a set of rules-of-thumb to abide	144
	by when delivering information through audio as	145
	opposed to text.	146
	Some of these rules-of-thumb that generalize	147
	beyond news reporting include (i) use simple sen-	148
	tence structures: allotting a sentence to each idea	149
	and putting the subject at the beginning as much	150
	as possible; (ii) do not use atypical syntax, such	151
	as “President Bush today told members of congress”	152
	and “I today went shopping”; (iii) avoid hyphen-	153
	ated adjectives (e.g. mineral-rich, tech-heavy); (iv)	154
	avoid too many names and numbers; and other mi-	155
	nor ones such as (v) avoid tongue twisters and (vi)	156
	avoid too much alliteration. Similar principles are	157

	Category	Prompt
Kept	Factual QA	<i>Why can camels survive for long without water?</i>
	Brainstorming	<i>Give me some ideas to manage my manager</i>
	Advice	<i>How do I start running?</i>
Removed	Creative Writing	<i>Write a scene between two actors discussing movie Inception.</i>
	Summarization	<i>Summarize in one sentence this article about a famous song. {article}</i>
	Info. Extraction	<i>Provide a list of all numbers mentioned and what they represent: {context}</i>

Table 1: Examples of speech-suitable prompts that are kept and removed from Dolly-15K for evaluation and response sampling. Removed user prompts are those that are unlikely to be asked in a speech-based setting.

Criteria	Score		
	0	1	2
Helpfulness	Not helpful	Somewhat helpful	Helpful
Relevance	Not relevant	Somewhat relevant	Relevant
Accuracy	Not accurate	Contains minor errors	Accurate
Informative.	Too little/much	Fair	Good
Length	Too short/long	Short/Long	Adequate
Understand.	Difficult	Fair	Easy

Table 2: We evaluate each response with various dimensions of speech-suitability using a 3 point Likert scale mapped from 0 to 2. Evaluation results with these dimensions for responses delivered in text and audio are shown in Table 3.

	Rel.	Help.	Correct.	Inform.	Len.	Under.
Text	1.98	1.95	1.97	1.72	1.70	1.84
Audio	2.00	1.95	1.99	1.59 [†]	1.44 [‡]	1.77

Table 3: Comparison between text and audio responses indicate the need to reduce excess information, decrease length, and enhance how easy the response is to comprehend when spoken. [†] and [‡] indicates a statistically significant drop compared to text results with a pairwise t-test at $p < 0.05$ and $p < 0.01$, respectively.

158 echoed for podcasting (Dowling and Miller, 2019),
159 multimedia journalism (Kolodzy, 2012), and liter-
160 ature on listenability (Chall and Dial, 1948; Fang,
161 1966; Messerklinger, 2006).

2.2 Text vs. speech for ITLMs

162 Optimization of speech-based interactions with
163 ITLMs may not share the same principles as those
164 suggested for radio news or podcasts. Therefore,
165 we verify whether such discrepancies in text versus
166 speech also apply for ITLM interactions.

167 First, we filtered the Dolly-15K dataset (Conover
168 et al., 2023) for user prompts that are relevant to
169 a speech-based interaction. We remove prompts
170 that require an accompanying reference text as
171 well as those that explicitly ask for code or sim-
172 ilarly unlikely-to-be-spoken requests. Categories
173 of user prompts that are kept and removed and
174 their respective examples are shown in Table 1.
175 This results in 11K user-prompt and response pairs,
176 from which we randomly sampled 80 prompts. Re-
177 sponses to these prompts were generated using
178 GPT-4 (gpt-4-0613) and converted to spoken
179 language using Amazon Polly¹, a text-to-speech
180 (TTS) API. Each responses was evaluated by three
181 different annotators, resulting in a total of 240 an-
182 notations completed by 22 unique annotators. Since
183

¹<https://aws.amazon.com/polly/>

184 we are measuring speech-based preferences, the an-
185 notators can only *listen* to the response. We share
186 more details and illustrations of the annotation in-
187 terface in Appendix A.

188 Following current practice in evaluating a
189 ITLM’s response, we ask annotators to indicate
190 the helpfulness, accuracy, and relevance fo the re-
191 sponse (Ouyang et al., 2022; Zhao et al., 2022).
192 In addition, we ask how suitable the responses are
193 in terms of its verbosity (length), amount of infor-
194 mation (informativeness), and ease of comprehen-
195 sion (understandability). Each of these criteria is
196 measured using a three-point Likert scale that is
197 mapped to scores ranging from 0 to 2, which is
198 shown in Table 2. Note that informativeness and
199 length have multiple options such as ‘too little’ or
200 ‘too much’ correspond to the same score as we treat
201 them as equally undesirable.

202 From the annotation results in Table 3, we ob-
203 serve that preferences for relevance, helpfulness,
204 and correctness do not significantly differ depend-
205 ing on the modality. However, the amount of infor-
206 mation and length are considered less suitable for
207 speech. We additionally observe a mild decrease in
208 the understandability of spoken responses. We hy-
209 pothesize that making LM responses terser, contain
210 less information in general, and easier to compre-
211 hend in general will increase the suitability of LMs
212 for speech.

System Prompt	You are a helpful, respectful and honest <i>speech</i> assistant. Keep your response <i>compact</i> without missing key information and make it suitable for speech such that it can be <i>easily verbalized</i> by a text-to-speech system. Your response should use <i>simple vocabulary and sentence structures</i> . Avoid <i>jargon, hyphenated adjectives, excessive alliteration, and tongue twisters</i> . <i>It should not contain parentheses, numbered or bullet lists, and anything else that cannot be verbalized</i> . If the answer requires multiple steps or a list, <i>provide only the first 3-5</i> and ask if the user is ready to move on to the next steps or know more.
User Prompt	Why can camels survive for long without water?
GPT-4 Response	Camels can survive for a long time without water because of several unique adaptations. 1. They can lose a large amount of their body weight in water without harm, while most animals can only lose a small percentage. 2. Camels can rehydrate very quickly, drinking up to 30 gallons of water in just 13 minutes. 3. Their bodies have a special way to store fat in their humps, which helps to minimize water loss. Would you like to know more about any of these points?

Table 4: Relying on GPT-4 with a detailed system prompt iteratively determined to get most consistent speech-suitable results according to findings from Section 2 can still lead to verbose outputs and non-verbal artifacts such as lists (in red). Even if this prompt gets consistent results, occupying many tokens as part of the input is impractical.

3 Adapting ITLMs for Speech

Current ITLMs are trained with supervised fine-tuning using instruction datasets collected as text input-output pairs and further fine-tuned through RLHF with preference data collected using pairs of text responses. ITLMs that result from this procedure already follow instructions well, but as illustrated in Section 2.2, they generate suboptimal responses for spoken interactions. Therefore, we conjecture that adapting to the spoken domain requires a primarily stylistic change, rather than a fundamental improvement in the capabilities of the underlying model.

3.1 Is prompt engineering sufficient?

Improving speech-suitability by prompting is non-trivial because speech-suitability covers multiple dimensions and lacks a one-size-fits-all definition. Even if such a definition could be composed in a simple prompt, even SOTA ITLMs like GPT-4 do not consistently follow these guidelines. For example, even with a detailed prompt shown in Table 4 that embeds principles from Section 2, GPT-4 reverts to generating verbose responses and non-verbal artifacts (e.g. lists and parentheses), especially in subsequent conversational turns. In addition, it is inefficient to prepend such a long sequence of text each time a response is generated.

Given the ambiguities of speech-suitability, we elected to use RLHF with a reward model trained on preference data collected using spoken response pairs. An overview of our adaptation methodology is shown in Figure 2. Our approach is similar to

standard RLHF with the exception that we 1) skip supervised finetuning to minimize annotation costs and compute demand and 2) use fewer preference annotations than comparable works to minimize the cost of developing our reward model.

3.2 Sampling responses

As we would like to solely rely on preference data, it is important to carefully compile the samples we will collect preferences with. Trying to sample responses to collect preference data to train a reward model that captures speech-suitability from the base ITLM would be inefficient as both sampled responses paired for comparison will likely be similarly poor. Therefore, the resulting preference annotation will provide little signal for the reward model to learn from. Instead, we need a reward model that observes a wider variety of responses of style to pick up various indications of which is suitable for speech, given the user prompt.

As a result, we empirically compile a set of system prompts that will lead to a diverse set of responses, rather than just directly letting the model generate response without a system prompt or a constant system prompt. The system prompts ask the ITLM to generate responses that follow the insights from Section 2.2 with varying levels of detail on desiderata of an speech-suitable response.

We share the full set of system prompts that we use and the process for devising them in Appendix B. Lastly, to further diversify the set of responses and thus enhance the generalizability of the reward model, we pair these system prompts with various

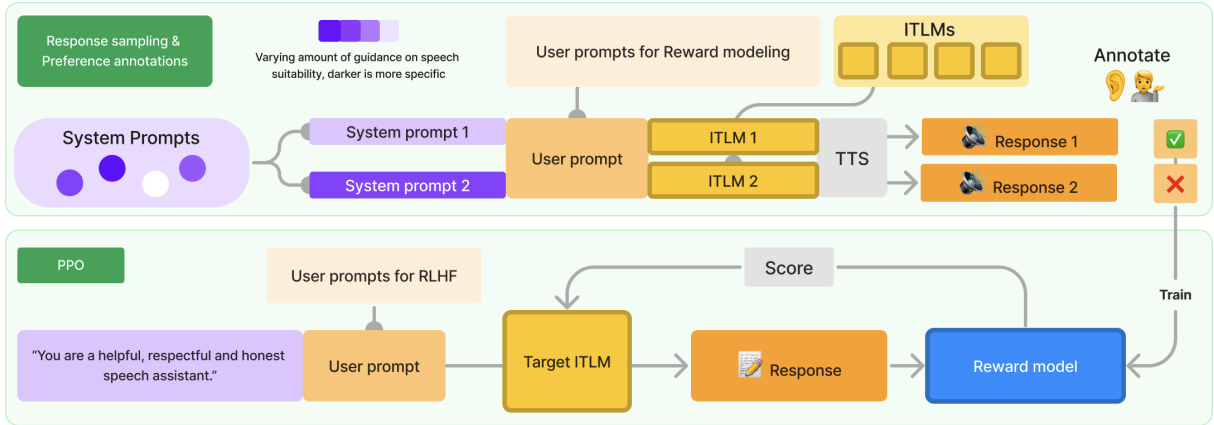


Figure 2: **Method overview.** Since we only have an approximate idea of what makes a good spoken response, we first compile a set of system prompts intended to vary the speech suitability of generated responses. We sample a pair to generate responses from various ITLMs to further diversify responses, transform them to speech with a TTS service, and human annotators rank their preferences after listening to them. The preference data is used to train a reward model, which is used for the reinforcement learning step, where we use PPO.

base ITLMs, which are Falcon Instruct 7B (Almazrouei et al., 2023), GPT-4, and GPT-3.5. In addition, we also adjust the decoding temperature from 0.7, 1.0, and 1.3. We include the reference response from the original dataset as well, and this yields a total of 37 configurations ($4 \times 3 \times 3 + 1$). By choosing two out of these 37 configurations, we have 666 possible combinations.

Using these configurations, we generate sample responses for the filtered user prompts from Dolly-15K described in Section 2.2. To collect a diverse set of both user prompts that the reward model will see, we use 9K user prompts and cycle through them with the 666 combinations to generate 40K responses, which result in 20K preference pairs. The remaining 2K user prompts are later used for evaluation.

3.3 Annotating speech-based preferences

Annotators are only able to listen to the user prompt and the responses. They are only able to see the survey to indicate their preferences only after listening to everything. The order in which the responses are shown are randomized so that we minimize any bias that may be introduced due to the order in which the responses are listened to. The preference annotation interface is shown in Figure 7 and the guidelines for annotations are shown in Figure 6.

Similar to Touvron et al. (2023), annotators are forced to choose one response over another with the following choices: significantly better, slightly better, and negligibly better. In addition, we ask for a brief explanation for their choice, mainly for

quality control purposes. An interesting future avenue would be leverage these explanations to train either a more accurate reward model or one more efficiently with fewer samples.

Before inviting annotators to a larger batch of annotations, each annotator was asked to complete 10 annotations, which were manually evaluated for consistency and accuracy according to the given guidelines. If their annotations were reliable, they were invited to a larger batch for which quality was monitored on a small sample for every 100 annotations that were completed. If not, we provided feedback to the annotators and asked them to do another 10, repeating this process until the annotations met our standards.

We pay our annotators an hourly wage that exceeds the minimum wage from where this study was conducted.² We had a total of 71 unique annotators participate in these annotations.

3.4 Reward modeling and reinforcement learning with human feedback

In order to perform RLHF, we train a reward model that generates a scalar score for speech-suitability given a single user prompt and response pair. We add a single projection layer to a GPT-J 6B model (Wang and Komatsuzaki, 2021) and train with the pairwise binary ranking loss (Ouyang et al., 2022):

$$\mathcal{L}_{\text{ranking}} = -\log(\sigma(r_{\theta}(x, y_c)) - r_{\theta}(x, y_r))$$

²Details hidden for anonymity.

where $r_\theta(x, y)$ is the reward model’s score for the user prompt x and the generated response y , given model weight θ . y_c is the response chosen by the annotator and y_r is the rejected one. We take a 9:1 split with our preference data for training and evaluation and use early stopping on the evaluation set to determine the best model to use for reinforcement learning. Our best performing model achieved an accuracy of 64.5% on the evaluation set, which is similar to the reward model accuracies reported by Touvron et al. (2023) and highlights that determining speech-based preference is not an easy task.

For the RLHF step, we use proximal policy optimization (PPO) (Schulman et al., 2017). The only difference with regular PPO is that we give it a constant system prompt “*You are a helpful, respectful and honest speech assistant.*”

4 Experimental Setup

4.1 Model

We use the Falcon Instruct 7B model (Almazrouei et al., 2023) as our base model to perform our adaptation process described in Section 3. We choose this model because at the time of this study, it was the ITLM with the most permissive license and had one of the strongest performance on various LLM evaluation benchmarks. We denote the resulting model that is finetuned from Falcon Instruct 7B with RLHF using our custom audio-preference reward model as **SpeechFalcon**.

4.2 Data

The user prompts that we use for sampling responses for preference annotations and the RLHF step are from the filtered version of Dolly-15K (Conover et al., 2023), described in Section 2.2. This instruction dataset does not have a predefined train-test split. Instead, it partitions the user prompts into eight categories, which make it convenient to filter out user prompts that are unrealistic in a speech-based interactions. It is one of the first open source, human-generated instruction dataset with a permissive creative commons license.

4.3 Baselines

We compare our speech-adapted model from Section 4.1 with the following baselines:

- Base ITLM: directly generating response from LM without any specific system prompts.

- Prompted ITLM: responses generated using system prompts that provide detailed guidance on what a suitable speech prompt is. We use our most detailed system prompt shown in Table 4.

- Original response: the original human-written response from the Dolly-15K dataset.

For Base ITLM and Prompted ITLM, we use the same Falcon Instruct 7B model from Section 4.1. We also compare against GPT-3.5 (gpt-3.5-turbo-0613) and GPT-4 (gpt-4-0613) and their Prompted versions to compare with state-of-the-art ITLMs. Our goal is for SpeechFalcon trained with the methodology described in Section 3 to generate responses that are more frequently preferred than those by its base (Falcon Instruct) and prompted counterparts (Falcon Instruct + Prompt) in a speech setting.

4.4 Evaluation

We evaluate SpeechFalcon against baselines with both human evaluation and automatic evaluation. Similar to the human evaluation in Zhou et al. (2023), we perform a head-to-head comparison of responses from SpeechFalcon to those of each baseline with the same setup as the preference annotations, described in Section 3.3. Evaluators listen to two responses in random order without knowing their corresponding models and then indicate which they prefer and briefly explain why. ‘Significantly better’ and ‘Slightly better’ choices are used to determine win or loss, while ‘Negligibly better’ is considered a tie. They are also asked to indicate if each response contains false information, which is later used to control for differences in base model’s performance when comparing speech suitability (Section 5.1). SpeechFalcon is compared to each baseline with 140 response pairs generated with prompts not seen during RLHF training and validation, which are from the test set described in Section 3.2.

For automatic evaluation, we examine factors related to listenability, sentence complexity, and length. For listenability, we measure the Flesch Reading Ease score, which examines the readability of some text. It is a function of the number of syllables within each sentence and it is known to be highly correlated with listenability (Chall and Dial, 1948; Fang, 1966; Messerklinger, 2006; Kotani et al., 2014). To quantify sentence complex-

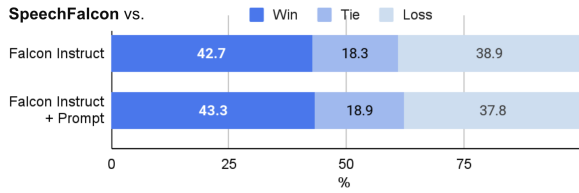


Figure 3: SpeechFalcon’s head-to-head human evaluation results with its directly comparable baselines listed in Section 4.3. A higher win rate vs. loss rate means that SpeechFalcon more frequently generates responses preferred in the speech setting.

ity, we use SpaCy’s dependency parser³ and measure the depth of the resulting dependency graph. For length, we simply use white space to split sentences into words and count the total number of words.

4.5 Implementation details

For training our reward model and finetuning Falcon Instruct 7B with RLHF, we use 8 A100 GPUs. We use the RLHF implementation from trlx repo⁴ and use the default hyperparameters for both reward modeling and PPO.

5 Experimental Results and Analysis

5.1 Human evaluation

SpeechFalcon improves over in generating speech-suitable responses. Human evaluation results are shown in Figure 3, which show the win, loss, and tie rate when comparing SpeechFalcon to Falcon Instruct and a prompted Falcon). The most important comparisons in Figure 3 are those with Falcon Instruct, the base model, and Prompted Falcon Instruct, which is the counterpart prompted with the detailed system prompt from Table 4. Compared to both of these baselines, we see that the win rate is higher than the loss rate, meaning that SpeechFalcon is able to attain meaningful improvements for speech-based responses.

Another important comparison is compared to the reference response, the original response contained in the instruction dataset that was collected with in text. SpeechFalcon falls short compared to the reference response, but not as much compared to the GPT baselines. Upon examining the comparison results with the GPT baselines, we found that in many cases the Falcon-based models frequently generated false information (hallucination)

³<https://spacy.io/api/dependencyparser>

⁴<https://github.com/CarperAI/trlx>

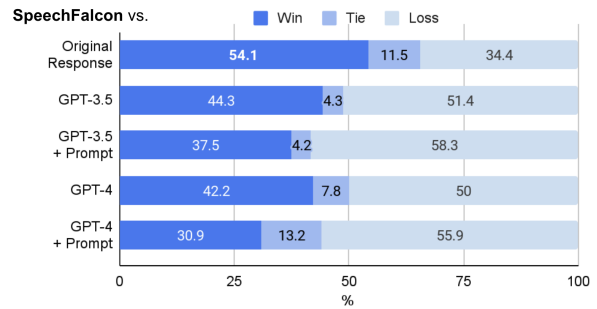


Figure 4: SpeechFalcon’s head-to-head human evaluation results against the original human-written response and GPT-based models, controlled for response factuality. Only cases where responses were both true or both false were considered to make a fairer comparison between different base ITLM models to isolate the analysis to speech-suitability. This table takes the same format as Figure 3.

Model	Length	Comprehensibility	
		FRE \uparrow	DD \downarrow
Reference	58.6 _{6.1}	61.7 _{2.5}	5.8 _{0.2}
Falcon Instruct	38.3 _{2.6}	65.2 _{1.8}	5.4 _{0.1}
Prompted Falcon Instruct	40.8 _{2.3}	64.2 _{1.7}	5.6 _{0.1}
SpeechFalcon	31.9 _{1.4}	70.9 _{1.8}	5.9 _{0.1}
GPT-3.5	125.4 _{8.8}	58.8 _{1.4}	6.1 _{0.1}
Prompted GPT-3.5	38.0 _{2.3}	71.3 _{1.3}	5.8 _{0.1}
GPT-4	118.8 _{8.0}	60.7 _{1.2}	5.9 _{0.1}
Prompted GPT-4	49.8 _{1.8}	68.7 _{1.1}	6.0 _{0.1}

Table 5: **Automatic evaluation results.** FRE: Flesch Reading Ease, DD: Dependency Depth. SpeechFalcon generate shorter responses with higher readability and sentence structure complexity than other Falcon baselines.

and this was indicated as the reasons for losing in a head-to-head comparison with other baselines.

SpeechFalcon is preferred to the original response but falls short of responses from larger black-box models. Since factuality significantly influences the annotator’s preference, we control for this factor that varies between models and strongly favors models (Ouyang et al., 2022) by filtering out comparisons results where either one of the compared responses contain false information. This filter results in about 80 comparisons per pair. The factuality-controlled results are shown in Figure 4 and we see significant boosts in the win rate for SpeechFalcon against the GPT baselines and also the original response. SpeechFalcon is able to achieve a higher win rate than loss rate against the original response, and the win rate against GPT baselines increase on average by 13.5%, but it is still lower than the loss rate.

487	5.2 Automatic evaluation		
488	RLHF with speech-based preference data lead		
489	to shorter responses and higher listenability.		
490	The automatic evaluation results are shown in		
491	Table 5. One of the most notable changes for		
492	SpeechFalcon compared to other Falcon baselines		
493	is the shorter response length and higher readabil-		
494	ity, which corresponds to higher listenability. It		
495	achieves a high listenability score that is only sec-		
496	ond to Prompted GPT-3.5 by a statistically insignif-		
497	icant margin. Sentence complexity, measured by		
498	dependency depth, varies only little for all baselines		
499	except for the base Falcon Instruct, which achieves		
500	the lowest average depth. Beyond comparisons		
501	between Falcon-based models, we observe that		
502	length and readability are important factors as GPT-		
503	baselines with higher win rates generate shorter		
504	responses and have higher readability scores. None		
505	of the rankings provided by each of these automatic		
506	metrics match with the win rate ranking against		
507	SpeechFalcon from Figure 3 or Figure 4. There-		
508	fore, these metrics mostly serve as an analytical		
509	tool that show that determining responses preferred		
510	in speech is a complex task that must consider vari-		
511	ous factors.		
512	6 Related work		
513	6.1 Language models and speech		
514	Language models have been widely adopted for		
515	modular components of a voice assistant, mainly		
516	for automatic speech recognition (Yu and Deng,		
517	2016; Wang et al., 2020; Chiu et al., 2018), re-		
518	sponse generation (Cho and May, 2020; Zhou et al.,		
519	2022; Liu et al., 2023), response selection (Humeau		
520	et al., 2019; Gao et al., 2020; Cho et al., 2021), and		
521	speech synthesis (Tan et al., 2022; Wang et al.,		
522	2017; Le et al., 2023). While speech synthesis fo-		
523	cus on how to translate text to speech such that		
524	it <i>sounds</i> natural, our work explores how to best		
525	compose the response itself for speech-based in-		
526	teractions. More similar in context to our work,		
527	Mousavi et al. (2024) explores whether LMs are		
528	robust to speech-based interactions by analyzing		
529	their robustness to input that include ASR errors.		
530	Another line of research is concerned with de-		
531	veloping multi-modal ITLMs that can process both		
532	speech and text input (Huang et al., 2023). Zhang		
533	et al. (2023) trains a GPT-based model called		
534	SpeechGPT with both speech data and text data		
535	such that it does not requiring additional ASR or		
536	TTS systems. However, this line of work also over-		
		looks on how we should compose responses that	537
		are delivered as speech.	538
	6.2 Fine-tuning LMs		539
	While the paradigm of pre-training and then fine-		540
	tuning has become the defacto status quo, there		541
	is still active research in how to go about fine-		542
	tuning to get the best results and do it efficiently.		543
	Among these efforts, fine-tuning pretrained LMs to		544
	become ITLM s has been one of the most active		545
	research areas recently. One of the central methods		546
	for training ITLM s is RLHF, which is a process		547
	of finetuning models with reinforcement learning		548
	via scores from reward models trained on human		549
	preference data (Bai et al., 2022; Ethayarajh et al.,		550
	2022; Ouyang et al., 2022; Touvron et al., 2023).		551
	RLHF is a relatively recent algorithm for which		552
	there are still many open questions, such as which		553
	reinforcement learning algorithm is the most effec-		554
	tive, how much SFT data and preference data is		555
	necessary for RLHF to work, and whether other		556
	forms of data can help train more generalizable		557
	reward models. Our work sheds some light into		558
	the minimum scale of data required for RLHF and		559
	how to train a generalizable reward model by more		560
	aggressively sampling for a diverse set of responses		561
	that go beyond decoding temperature variations.		562
	7 Conclusion		563
	We explore an important yet overlooked challenge		564
	of adapting instruction-following language models		565
	to compose responses that are specifically designed		566
	to be verbalized, i.e. speech-suitable. With rules-		567
	of-thumb of the radio industry and through our sur-		568
	veys that ask compare the suitability of a response		569
	for both text and audio, we establish that adap-		570
	tations for speech-suitable responses is necessary.		571
	Therefore, we fine-tune models to generate speech-		572
	suitable responses by aligning them with human		573
	preferences collected with responses directly pre-		574
	sented through speech. We show through both		575
	human and automatic evaluations that our model		576
	fine-tuned with RLHF using our speech-based pref-		577
	erence data outperforms the reference responses		578
	and its carefully prompted counterparts. Our work		579
	has focused on single-turn interactions and the con-		580
	tent of the responses, and we hope to expand this		581
	investigation for multi-turn interactions and the		582
	acoustic aspect of speech-suitability.		583

584 Limitations

585 In this work, we focused on *what* ITLMs should
586 generate for responses that will be delivered via
587 speech. However, we recognize that another inter-
588 esting line of research is *how* the response should
589 be delivered, where factors related to speech, such
590 as timber, pitch, and speed, are important. In ad-
591 dition, our examination is focused on single-turn
592 interactions, but another intriguing realistic dimen-
593 sion to suitability of a response in speech is multi-
594 turn interactions. It would be interesting to com-
595 pare the effect of delivering information at various
596 granularities and how different types of follow-up
597 questions on the user experience with a ITLMs in
598 speech-based interactions. We leave these lines of
599 research to future work.

600 Broader Impact

601 Since those who cannot read due to illiteracy or
602 blindness rely on voice assistants to interface with
603 modern technology, expanding their capabilities
604 can directly lead to improvements in their standard
605 of living. However, current voice assistants are
606 not as generally useful as current state-of-the-art
607 ITLMs. The likes of Siri and Alexa tend to fulfill
608 simple routine tasks and are brittle when facing
609 complex requests. Therefore, adopting ITLMs as
610 the main backbones of voice assistants and adapt-
611 ing them to become suitable to interface through
612 voice and can significantly increase the accessibil-
613 ity for these people to the powerful capabilities of
614 ITLMs.

615 References

616 Jessica Abel. 2015. *Out on the wire: The storytelling*
617 *secrets of the new masters of radio*. Crown.

618 Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Al-
619 shamsi, Alessandro Cappelli, Ruxandra Cojocaru,
620 Merouane Debbah, Etienne Goffinet, Daniel Hes-
621 low, Julien Launay, Quentin Malartic, Badreddine
622 Noune, Baptiste Pannier, and Guilherme Penedo.
623 2023. Falcon-40B: an open large language model
624 with state-of-the-art performance.

625 Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda
626 Askell, Anna Chen, Nova DasSarma, Dawn Drain,
627 Stanislav Fort, Deep Ganguli, Tom Henighan, et al.
628 2022. Training a helpful and harmless assistant with
629 reinforcement learning from human feedback. *arXiv*
630 *preprint arXiv:2204.05862*.

631 Tom Brown, Benjamin Mann, Nick Ryder, Melanie
632 Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind

Neelakantan, Pranav Shyam, Girish Sastry, Amanda
Askell, et al. 2020. Language models are few-shot
learners. *Advances in neural information processing*
systems, 33:1877–1901. 633
634
635
636

Jeanne S Chall and Harold E Dial. 1948. Predicting
listener understanding and interest in newscasts. *Ed-*
ucational Research Bulletin, pages 141–168. 637
638
639

Chung-Cheng Chiu, Tara N Sainath, Yonghui Wu, Rohit
Prabhavalkar, Patrick Nguyen, Zhifeng Chen, An-
juli Kannan, Ron J Weiss, Kanishka Rao, Ekaterina
Gonina, et al. 2018. State-of-the-art speech recog-
nition with sequence-to-sequence models. In *2018*
IEEE international conference on acoustics, speech
and signal processing (ICASSP), pages 4774–4778.
IEEE. 640
641
642
643
644
645
646
647

Hyundong Cho and Jonathan May. 2020. [Grounding](#)
[conversations with improvised dialogues](#). In *Proceed-*
ings of the 58th Annual Meeting of the Association
for Computational Linguistics, pages 2398–2413, On-
line. Association for Computational Linguistics. 648
649
650
651
652

Hyundong Cho, Basel Shbita, Kartik Shenoy, Shuai
Liu, Nikhil Patel, Hitesh Pindikanti, Jennifer Lee,
and Jonathan May. 2021. Viola: A topic agnostic
generate-and-rank dialogue system. *arXiv preprint*
arXiv:2108.11063. 653
654
655
656
657

Hyung Won Chung, Le Hou, Shayne Longpre, Bar-
ret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi
Wang, Mostafa Dehghani, Siddhartha Brahma, et al.
2022. Scaling instruction-finetuned language models.
arXiv preprint arXiv:2210.11416. 658
659
660
661
662

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie,
Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell,
Matei Zaharia, and Reynold Xin. 2023. [Free dolly:](#)
[Introducing the world’s first truly open instruction-](#)
[tuned llm](#). 663
664
665
666
667

David O Dowling and Kyle J Miller. 2019. Immersive
audio storytelling: Podcasting and serial documen-
tary in the digital publishing industry. *Journal of*
radio & audio media, 26(1):167–184. 668
669
670
671

Kawin Ethayarajh, Yejin Choi, and Swabha
Swayamdipta. 2022. [Understanding dataset](#)
[difficulty with \$\mathcal{V}\$ -usable information](#). In *Proceedings*
of the 39th International Conference on Machine
Learning, volume 162 of *Proceedings of Machine*
Learning Research, pages 5988–6008. PMLR. 672
673
674
675
676
677

Irving E Fang. 1966. The “easy listening for-
mula”. *Journal of Broadcasting & Electronic Media*,
11(1):63–68. 678
679
680

Xiang Gao, Yizhe Zhang, Michel Galley, Chris Brock-
ett, and Bill Dolan. 2020. [Dialogue response ranking](#)
[training with large-scale human feedback data](#). In
Proceedings of the 2020 Conference on Empirical
Methods in Natural Language Processing (EMNLP),
pages 386–395, Online. Association for Computa-
tional Linguistics. 681
682
683
684
685
686
687

688	Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training compute-optimal large language models. <i>arXiv preprint arXiv:2203.15556</i> .	<i>Computational Linguistics (Volume 1: Long Papers)</i> , pages 8404–8419, Toronto, Canada. Association for Computational Linguistics.	744 745 746
694	Takaaki Hori, Jaejin Cho, and Shinji Watanabe. 2018. End-to-end speech recognition with word-based rnn language models. In <i>2018 IEEE Spoken Language Technology Workshop (SLT)</i> , pages 389–396. IEEE.	Josef Messerklinger. 2006. Listenability. <i>Center for English Language Education Journal</i> , 14:56.	747 748
698	Rongjie Huang, Mingze Li, Dongchao Yang, Jia-tong Shi, Xuankai Chang, Zhenhui Ye, Yuning Wu, Zhiqing Hong, Jiawei Huang, Jinglin Liu, et al. 2023. Audiogpt: Understanding and generating speech, music, sound, and talking head. <i>arXiv preprint arXiv:2304.12995</i> .	Seyed Mahed Mousavi, Gabriel Roccabruna, Simone Alghisi, Massimo Rizzoli, Mirco Ravanelli, and Giuseppe Riccardi. 2024. Are llms robust for spoken dialogues?	749 750 751 752
704	Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2019. Poly-encoders: Architectures and pre-training strategies for fast and accurate multi-sentence scoring. In <i>International Conference on Learning Representations</i> .	Clifford Ivar Nass and Scott Brave. 2005. <i>Wired for speech: How voice activates and advances the human-computer relationship</i> . MIT press Cambridge.	753 754 755 756
709	Jonathan Kern. 2008. <i>Sound reporting: The NPR guide to audio journalism and production</i> . University of Chicago Press.	Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. <i>Advances in Neural Information Processing Systems</i> , 35:27730–27744.	757 758 759 760 761 762
712	Janet Kolodzy. 2012. <i>Practicing convergence journalism: An introduction to cross-media storytelling</i> . Routledge.	Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with gpt-4. <i>arXiv preprint arXiv:2304.03277</i> .	763 764 765
715	Katsunori Kotani, Shota Ueda, Takehiko Yoshimi, and Hiroaki Nanjo. 2014. A listenability measuring method for an adaptive computer-assisted language learning and teaching system. In <i>Proceedings of the 28th Pacific Asia Conference on Language, Information and Computing</i> , pages 387–394, Phuket, Thailand. Department of Linguistics, Chulalongkorn University.	Jiacheng Xu Greg Durrett Prasann Singhal, Tanya Goyal. 2023. A long way to go- investigating length correlations in rlhf. <i>arXiv preprint arXiv:2310.03716</i> .	766 767 768
723	Matthew Le, Apoorv Vyas, Bowen Shi, Brian Karrer, Leda Sari, Rashel Moritz, Mary Williamson, Vimal Manohar, Yossi Adi, Jay Mahadeokar, et al. 2023. Voicebox: Text-guided multilingual universal speech generation at scale. In <i>Thirty-seventh Conference on Neural Information Processing Systems</i> .	Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. 2021. Scaling language models: Methods, analysis & insights from training gopher. <i>arXiv preprint arXiv:2112.11446</i> .	769 770 771 772 773 774
729	Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In <i>Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing</i> , pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.	Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. 2022. Multitask prompted training enables zero-shot task generalization. In <i>ICLR 2022-Tenth International Conference on Learning Representations</i> .	775 776 777 778 779 780 781
736	Opher Lieber, Or Sharir, Barak Lenz, and Yoav Shoham. 2021. Jurassic-1: Technical details and evaluation. <i>White Paper: AI21 Labs</i> , 1.	John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. <i>arXiv preprint arXiv:1707.06347</i> .	782 783 784 785
739	Shuai Liu, Hyundong Cho, Marjorie Freedman, Xuezhe Ma, and Jonathan May. 2023. RECAP: Retrieval-enhanced context-aware prefix encoder for personalized dialogue response generation. In <i>Proceedings of the 61st Annual Meeting of the Association for</i>	Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. <i>Advances in Neural Information Processing Systems</i> , 33:3008–3021.	786 787 788 789 790 791
		Xu Tan, Jiawei Chen, Haohe Liu, Jian Cong, Chen Zhang, Yanqing Liu, Xi Wang, Yichong Leng, Yuanhao Yi, Lei He, et al. 2022. NaturalSpeech: End-to-end text to speech synthesis with human-level quality. <i>arXiv preprint arXiv:2205.04421</i> .	792 793 794 795 796

797	Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. <i>arXiv preprint arXiv:2307.09288</i> .		
798			
799			
800			
801			
802			
803	Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. https://github.com/kingoflolz/mesh-transformer-jax .		
804			
805			
806			
807	Changhan Wang, Yun Tang, Xutai Ma, Anne Wu, Sravya Popuri, Dmytro Okhonko, and Juan Pino. 2020. Fairseq s2t: Fast speech-to-text modeling with fairseq. <i>arXiv preprint arXiv:2010.05171</i> .		
808			
809			
810			
811	Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hananeh Hajishirzi. 2022a. Self-instruct: Aligning language model with self generated instructions. <i>arXiv preprint arXiv:2212.10560</i> .		
812			
813			
814			
815			
816	Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krma Doshi, Kuntal Kumar Pal, Maitreya Patel, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Savan Doshi, Shailaja Keyur Sampat, Siddhartha Mishra, Sujun Reddy A, Sumanta Patro, Tanay Dixit, and Xudong Shen. 2022b. Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks. In <i>Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing</i> , pages 5085–5109, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.		
817			
818			
819			
820			
821			
822			
823			
824			
825			
826			
827			
828			
829			
830			
831			
832			
833			
834	Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. 2017. Tacotron: Towards end-to-end speech synthesis. <i>Interspeech 2017</i> .		
835			
836			
837			
838			
839	Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. <i>arXiv preprint arXiv:2109.01652</i> .		
840			
841			
842			
843			
844	Yin-Pin Yang. 2004. An innovative distributed speech recognition platform for portable, personalized and humanized wireless devices. In <i>International Journal of Computational Linguistics & Chinese Language Processing, Volume 9, Number 2, August 2004: Special Issue on New Trends of Speech and Language Processing</i> , pages 77–94.		
845			
846			
847			
848			
849			
850			
851	Dong Yu and Lin Deng. 2016. <i>Automatic speech recognition</i> , volume 1. Springer.		
852			
	Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023. SpeechGPT: Empowering large language models with intrinsic cross-modal conversational abilities. In <i>Findings of the Association for Computational Linguistics: EMNLP 2023</i> , pages 15757–15773, Singapore. Association for Computational Linguistics.	853	859
		854	
		855	
		856	
		857	
		858	
		859	
	Kaifa Zhao, Le Yu, Shiyao Zhou, Jing Li, Xiapu Luo, Yat Fei Aemon Chiu, and Yutong Liu. 2022. A fine-grained Chinese software privacy policy dataset for sequence labeling and regulation compliant identification. In <i>Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing</i> , pages 10266–10277, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.	860	867
		861	
		862	
		863	
		864	
		865	
		866	
		867	
	Chunting Zhou, Pengfei Liu, Puxin Xu, Srinu Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2023. Lima: Less is more for alignment. <i>arXiv preprint arXiv:2305.11206</i> .	868	871
		869	
		870	
		871	
	Pei Zhou, Hyundong Cho, Pegah Jandaghi, Dong-Ho Lee, Bill Yuchen Lin, Jay Pujara, and Xiang Ren. 2022. Reflect, not reflex: Inference-based common ground improves dialogue response quality. In <i>Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing</i> , pages 10450–10468, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.	872	879
		873	
		874	
		875	
		876	
		877	
		878	
		879	
	Appendix		880
	A Annotation interface and guidelines		881
	The annotation interface for text vs. voice annotations is shown in Figure 5. The instructions for annotating is shown in Figure 6. The interface for audio preference annotations are shown in Figure 7.	882	888
		883	
		884	
		885	
	B System prompts		886
	Our set of system prompts are presented in Table 6.	887	887

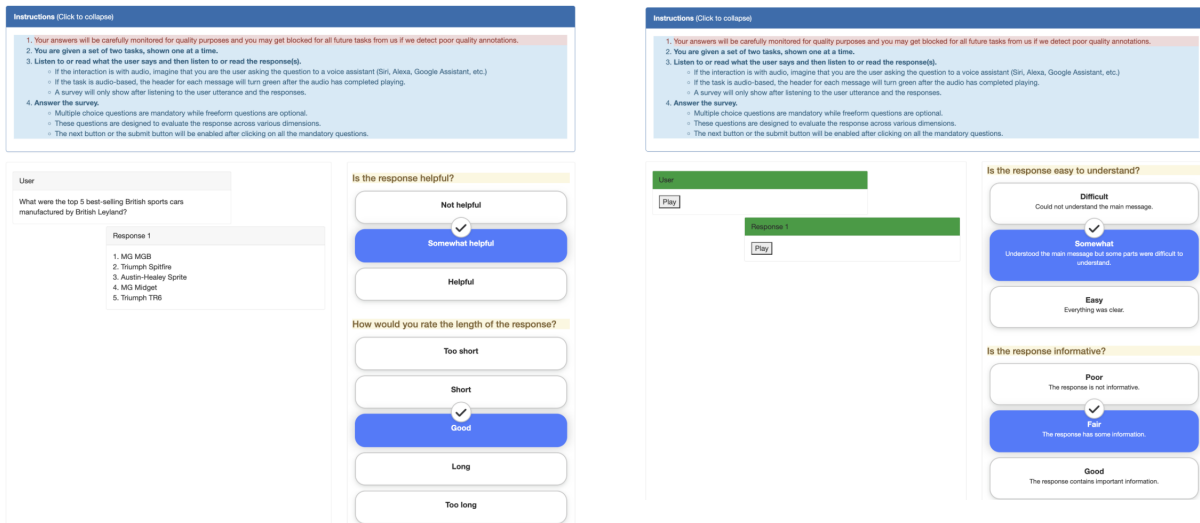


Figure 5: *Left*: suitability annotation interface for text. *Right*: suitability annotation interface for audio. The same set of survey questions are asked in both cases and only the modality of the user prompt and responses are different. Note that each annotator complete a pair of both text and audio tasks so that we collect results that are better calibrated.

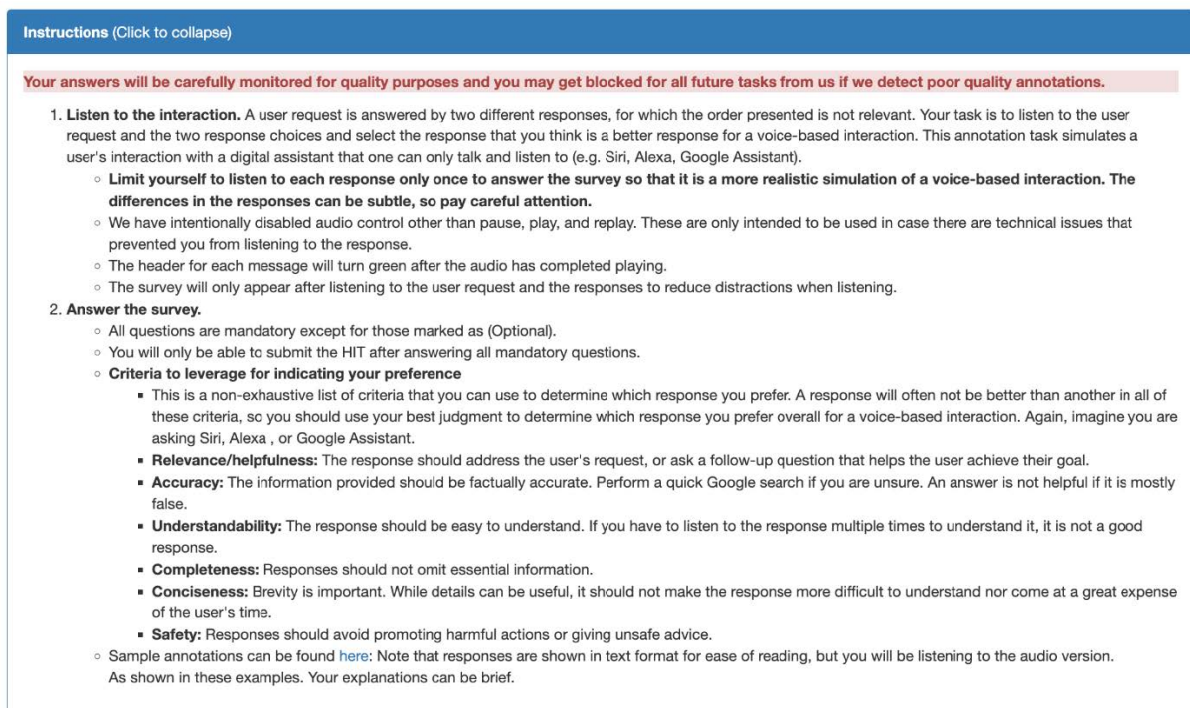


Figure 6: Instructions for preference annotations with responses provided as audio.

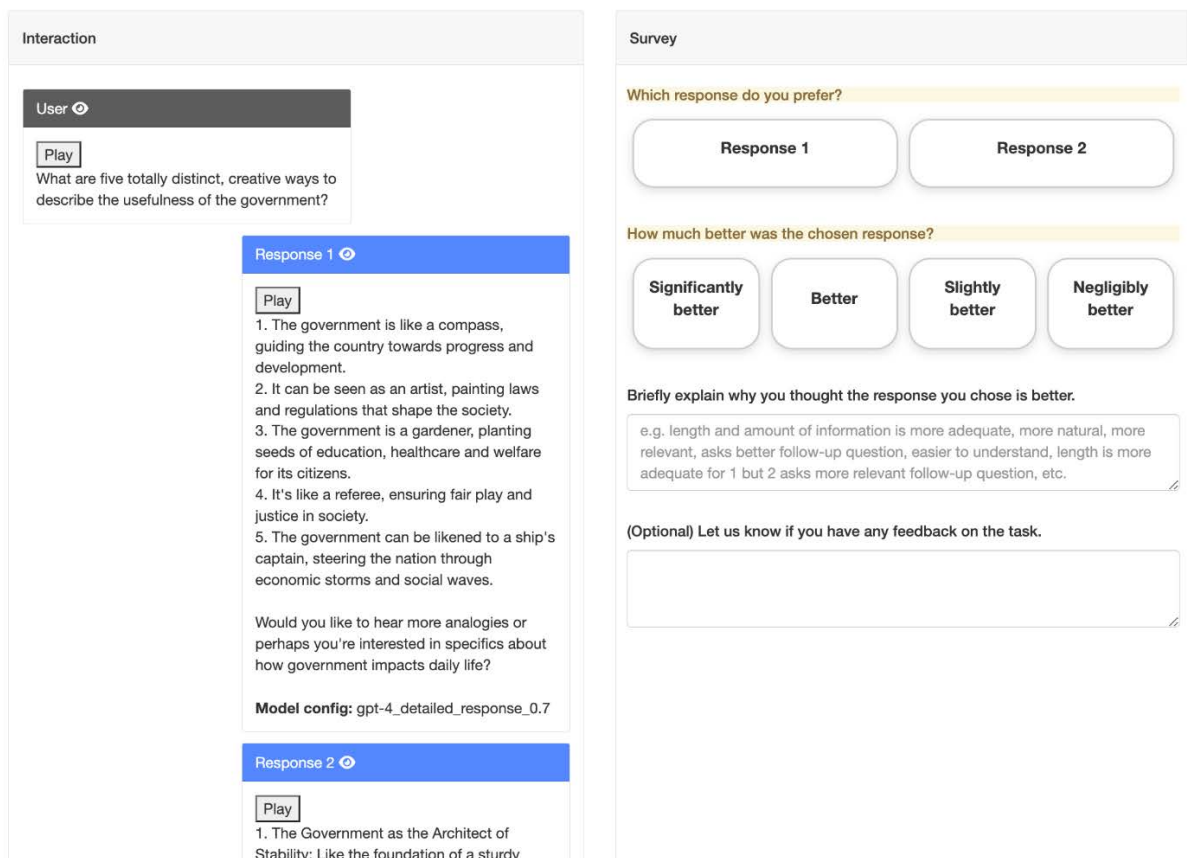


Figure 7: Audio preference annotation interface. The left pane contains the user prompt and two responses. The right side contains the survey that appears after the annotator listens to both responses. The annotators actually do not see the text and is only available for the task distributors when monitoring completed tasks. The actual view of the annotators are the same as the right side of Figure 5.

Prompt name	Prompt
Simple	You are a helpful, respectful and honest voice assistant.
Medium	You are a helpful, respectful and honest voice assistant. Make your responses colloquial and compact. Avoid any artifacts that cannot be vocalized.
Easy	You are a helpful, respectful, and honest voice assistant. Your response should use simple vocabulary and sentence structures so that it is easy to understand and such that it is below a 5th grade reading level.
Detailed	You are a helpful, respectful and honest <i>speech</i> assistant. Keep your response <i>compact</i> without missing key information and make it suitable for speech such that it can be <i>easily verbalized</i> by a text-to-speech system. Your response should use <i>simple vocabulary and sentence structures</i> . Avoid <i>jargon, hyphenated adjectives, excessive alliteration, and tongue twisters</i> . It should <i>not contain parentheses, numbered or bullet lists, and anything else that cannot be verbalized</i> . If the answer requires multiple steps or a list, <i>provide only the first 3-5</i> and ask if the user is ready to move on to the next steps or know more.

Table 6: The set of system prompts that we use in Section 3.2