Beyond Text: Utilizing Vocal Cues to Improve Decision Making in LLMs for Robot Navigation Tasks

Anonymous authors
Paper under double-blind review

Abstract

While LLMs excel in processing text in these human conversations, they struggle with the nuances of verbal instructions in scenarios like social navigation, where ambiguity and uncertainty can erode trust in robotic and other AI systems. We can address this shortcoming by moving beyond text and additionally focusing on the paralinguistic features of these audio responses. These features are the aspects of spoken communication that do not involve the literal wording (lexical content) but convey meaning and nuance through how something is said. We present *Beyond Text*; an approach that improves LLM decision-making by integrating audio transcription along with a subsection of these features, which focus on the affect and more relevant in human-robot conversations.This approach not only achieves a 70.26% winning rate, outperforming existing LLMs by 22.16% to 48.30% (gemini-1.5-pro and gpt-3.5 respectively), but also enhances robustness against token manipulation adversarial attacks, highlighted by a 22.44% less decrease ratio than the text-only language model in winning rate. "*Beyond Text*" marks an advancement in social robot navigation and broader Human-Robot interactions, seamlessly integrating text-based guidance with human-audio-informed language models.

1 Introduction

The rapid advancements in text-based Large Language Models (LLMs) like GPT-4 and the Gemini series (Achiam et al., 2023; Reid et al., 2024) have significantly enhanced capabilities in collaborative human-robot interactions (HRI) especially in diverse settings, including task planning, or social navigation (Shah et al., 2023; Liu et al., 2023a; Li et al., 2023; Ren et al., 2023; Liang et al., 2023). Broadly speaking, for HRI tasks, we want to design the robot to think like a human and best carry out its tasks. We as humans interact and solve problems together, often relying on verbal communication as a key modality. If humans and robots interact to perform collaborative navigation tasks in large complex environments, human's potential unfamiliarity with the space and spatial anxiety(Chan et al., 2012) leads to providing ambiguous or

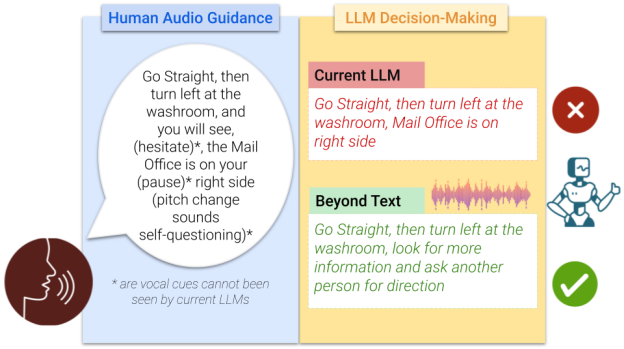

Figure 1: *Current Large Language Models (LLMs) are unable to effectively interpret human vocal cues and accurately make decisions for audio-guided navigation involving ambiguous instructions.*

uncertain navigational guidance, and these uncertainties are not only in literal wording but also vocal nuance (Prestopnik & Roskos-Ewoldsen, 2000; Golledge, 2003). While LLMs exhibit strong textual reasoning and perform adeptly in robotic tasks through textual prompt tuning, their ability to process and interpret audio information, particularly the nuanced vocal features in speech tones, remains limited (example shown in Figure 1). Although GPT-4 Achiam et al. (2023) introduces an audio-to-text API, enhancing its capability to accept audio input, this feature primarily focuses on transcription rather than analyzing vocal features. This limitation becomes particularly salient in human audio-guided social navigation, where assessing the trustworthiness of human guidance is crucial for successful navigation (Dorbala et al., 2021; Francis et al., 2023; Firoozi et al., 2023).

In this work, we present *Beyond Text*, an approach that incorporates both audio transcription and affective vocal features, including pitch, loudness, and duration, to improve robot navigation. Our approach leverages these vocal cues to interpret navigational instructions more reliably, especially when instructions are ambiguous or uncertain (Sundar, 2020).

Moreover, to the best of our knowledge, existing available audio-guided navigation datasets, like RxR (Ku et al., 2020), do not consider vocal cues for navigation tasks. To address this gap, we introduce *Disfluent Navigational Instruction Audio Dataset (DNIA)*. This dataset contains audio clips that capture vocal and textual uncertainties, characteristic of real-life human speech in navigational contexts. DNIA focuses on disfluencies and nuanced vocal expressions, which are vital for enhancing LLMs' ability to interpret the trustworthiness of human speech. We believe that these contributions are a significant first step towards developing more natural and effective LLM-based robot-human interaction systems for social navigation.

Our main contributions are summarized as follows:

1. **Provided Evidence for Advancing LLM Navigation with Vocal Features.** The current foundation models are unable to capture the nuanced human ambiguity via only textual input. We propose the integration of affective vocal cue analysis beyond text to let intelligent systems like robots synthesize these streams into actionable intelligence, and thus *not only reason what the human said, but also how the human said*. Our results show low bias and low variance: a 70.26%+ winning rate in detecting the uncertainty and generating appropriate next-step action and a low confidence score compared to only using LLMs to read textual transcription, indicating increased high confidence in analyzing human uncertainty. Notably, all three different large language models we tested, including GPT3.5, GPT3, and Gemini Pro, exhibited a significant increase in winning rates. *Beyond Text* augments LLMs' ability to model human uncertainty in navigational scenarios.

2. **A New Dataset on Disfluent Human Audio-Guided Instructions.** In support of our framework and to catalyze future research, we collect a dataset comprising 500 human audio navigational instructions with perceptual "ground truth" captured human labeling. This dataset is characterized by a diverse range of semantic and vocal ambiguities, offering an extensive experimental foundation for our proposed method and subsequent investigative endeavors in this space.

3. **Show Robustness of Vocal Cues on Improving the LLM Decision-Making**: We also show the robustness of our framework against adversarial attacks by conducting token manipulation experiments on the prompts fed to LLMs. Under attack, our framework has 22.44% less decrease ratio than the text-only language model in winning rate. We present a comprehensive ablation study evaluating the effectiveness of three specific vocal cues we utilized, 15% to 20%+ increase in winning rate clearly indicating the significant role these cues play in augmenting the reasoning capabilities of LLMs. Lastly, we discuss the potential of audio augmentation in LLMs, highlighting this as a promising avenue for future research.

2 Related Works

Large Language Models for Robotics. Recent studies have demonstrated the versatility of LLMs in various robotic contexts. These include employing few-shot summarizing skills for task planning in robots

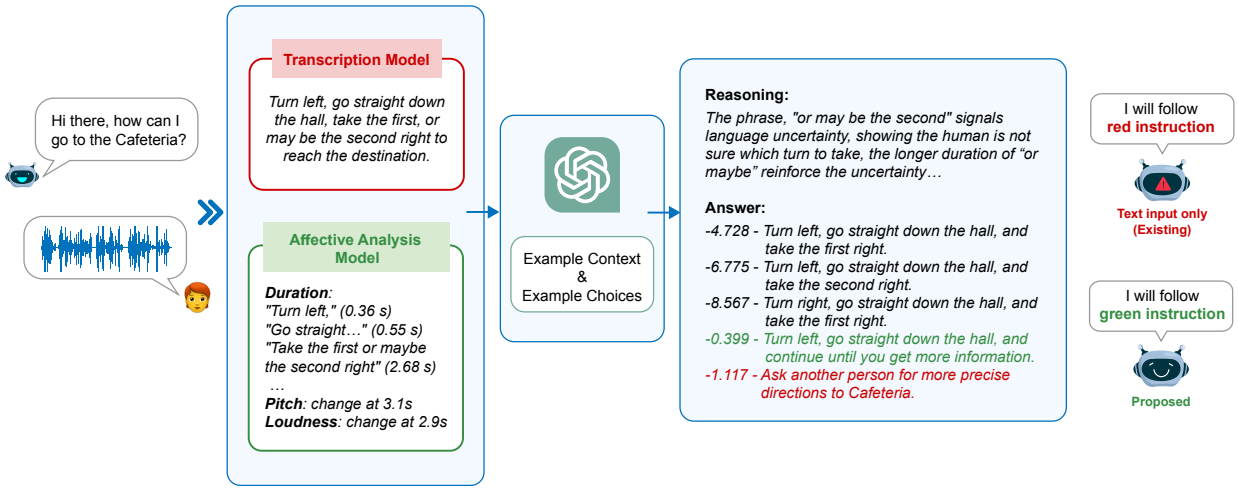

Figure 2: ***Beyond Text*** *Overview: With a human instruction audio clip, we simultaneously do transcription and vocal affective cue analysis. We prompt the language model to reason and generate five next-step choices. Then, we pick the action based on the highest next token logit probability. Note that with only the transcription model, the framework chooses the red choice. With affective analysis model, the green choice is picked.*

(Wu et al., 2023; Liu et al., 2023b), interpreting contextual information for robotic locomotion decision-making (Shek et al., 2023), integrating visual feedback for open-world robotic manipulation (Lin et al., 2023a; Ahn et al., 2022), and enhancing motion planning in autonomous driving (Mao et al., 2023). For human-robot interaction, Shah's research Shah et al. (2023) suggests using LLMs for semantic prediction of object locations, thereby optimizing navigation. SayCan Ahn et al. (2022) introduces multimodal input to generate the optimal long-horizon instructions and effectively score the affordance probability of the instruction for robotic task planning. Meanwhile, Ren et al. (2023) develops a framework to assess and synchronize the uncertainty levels in LLM-based planning systems, enabling them to seek assistance when necessary. However, unlike our framework, KnowNo Ren et al. (2023) does not account for potentially unreliable human input but rather relies on human selection among generated options. *Beyond Text* uniquely concentrates on assessing human uncertainty through affective computing for robotic navigation, particularly when guided by human audio that might be ambiguous or incorrect. This approach allows robots to gauge the reliability of human-audio navigation instructions by analyzing vocal emotions, enabling more informed reasoning.

Foundation Models with Audio. The ability to perceive and interpret auditory information is essential for AI agents in the real world. GPT-4's recent updateAchiam et al. (2023), incorporating Whisper's audio-to-text feature (Radford et al., 2023), significantly advances foundational models' capacity to transcribe audio inputs. However, this transcription process often overlooks nuanced vocal features inherent in speech. Parallel research efforts (Ghosal et al., 2023; Borsos et al., 2023) explore language models for generating audio from text and enhancing audio-visual understanding in videos. Notably, SALMONN Tang et al. (2023) integrates speech audio encoders with a multimodal language model, enabling direct processing of audio inputs for tasks like speech translation, answering spoken queries, etc. Despite these advancements, current research does not specifically address the use of audio inputs in aiding human-robot navigation or in extracting vocal uncertainty from speech.

Uncertainty Modeling in LLMs. Gauging the trustworthiness of responses generated by LLMs remains an open challenge in today's LLM era, with limited research on uncertainty quantification for natural language generation. Lin Lin et al. (2023b) compared multiple uncertainty measurements in the black-box language model, and Kuhn Kuhn et al. (2023) introduced entropy as a method to model uncertainty in the large language model. In our approach, we take inspiration from prior works (Kuhn et al., 2023; Lin et al., 2023b) and define a confidence score $\mathcal{C}(\rho)$ that is inversely proportional to the Kullback-Leibler divergence between the current LLM's prediction and ground truth human perception distribution to show the LLM's confidence on uncertainty measurement.

Human Affective Analysis in Robotics. For social robots to effectively coexist and interact with humans, it is imperative that they comprehend human emotional states for decision-making processes. Emotion understanding from speech for human-robot interaction has been studied in (Lakomkin et al., 2018a), while deep reinforcement learning methods (Dorbala et al., 2021) and cognition model (Eppe et al., 2016) has been used to understand textual ambiguities in natural languages. RxR dataset (Ku et al., 2020) is a well-known audio dataset for visual language navigation but does not cover the affective cues embedded in the vocal instructions, which hold substantial information that can reveal human uncertainty and trust. *Beyond Text* introduces an audio-augmentation technique for large language models (LLMs) that analyzes textual and vocal cues in human audio commands. This approach enables more human-like robotic systems for social navigation by identifying emotions and ambiguities through both speech content and vocal tones.

3 Our Approach

In this section, we first mathematically characterize our novel framework, 'Beyond Text,' which uses LLMs to model human trust and uncertainty based on audio-based navigational guidance. As shown in Figure 2, we consider a setting where a robot asks the human for instructions, and then the human provides instructions in the form of audio instruction. This setting is adapted based on prior work on social navigation (Hu et al., 2019; Lakomkin et al., 2018b; Eppe et al., 2016; Francis et al., 2023). In order to answer the question raised by the robot, the human replies with a voice prompt $v \in \mathcal{V}$, which belongs to set of audio responses $\mathcal{V}$, which is then converted to a textual description $\mathcal{S}$ using the best available transcription model $\mathcal{T}_m : \mathcal{V} \rightarrow \mathcal{S}$ (here $\mathcal{S}$ is the set of all possible sentences). Simultaneously, v is mapped to an affective cue set $\mathcal{K}$ via an affective cue model $\mathcal{AC} : \mathcal{V} \rightarrow \mathcal{K}$. The combination of textual and affective cues is denoted as:

$$\mathcal{Q}(\mathcal{V}) = \mathcal{S} \oplus \mathcal{K} \tag{1}$$

which constitutes the prompt for the LLM (denoted by p_θ with θ being the language model parameters), which outputs a response $y = [y_1, y_2, \cdots, y_k]$ by generating each token y_i at a time i. The joint probability over the tokens can be written as:

$$p_\theta(y|\mathcal{Q}(v) = \prod_{i=1}^{k} p_\theta(y_i|y_1, y_2 \cdots y_{k-1}) \tag{2}$$

To perform decision-making for robots, with a few-shot prompt that includes possible next steps in a few scenarios, the LLM generates candidate responses $\{y^j\}_{j=1}^{J}$ and chooses the most likely answer. We have briefly described it in Figure 2 and will provide more details below.

3.1 Semantic Uncertainty Quantification

The process begins by converting the audio clips of navigational guidance into text using the open-source Whisper model (Radford et al., 2023), which we explain in detail in Appendix B. The transcribed text has two key functions for the LLM: identifying uncertainty-related semantic cues and formulating appropriate navigational steps. We focus on disfluency in human speech as an indicator of uncertainty, which affects the confidence measure of the LLMs (cf. 3). We categorize language uncertainty into three types as follows:

(i) Textual Uncertainty: Phrases like "probably," "may," "might," and "I think" within the transcribed text suggest a lack of conviction (Dorbala et al., 2021).

(ii) Speech Repair: In instances where individuals correct themselves (Lou et al., 2019), such as saying, "Take a right—uhh, I mean take a left and go straight," it is crucial to parse and interpret these corrections to avoid misdirection.

(iii) Hesitation Signals: Pauses or hesitations, as in "Go straight and, err, take the second left," may signify uncertainty regarding subsequent directions (Ogata et al., 2009), despite apparent confidence in preceding instructions.

These textual patterns are typical in human communication when conveying uncertainty. We prompt the LLMs to recognize these cues, subsequently adjusting the confidence level in the reliability of the instructions that follow these indicators.

3.2 Vocal Affective Cue Analysis

In addition to textual content, the prosody of spoken instructions can be indicative of human uncertainty (Mittal et al., 2020; Goupil et al., 2021; Jiang & Pell, 2017; Guyer et al., 2019). Our analysis focuses on three primary auditory features captured within the Mel-frequency cepstral coefficients of the voice waveform to detect potential uncertainty:

Duration: When an individual is unsure about the instruction, the speed of their speech either slows down as they are trying to think through what to say or speeds up as becoming more nervous (Jiang & Pell, 2017; Székely et al., 2017). The speech rate is a vocal characteristic indicating certainty tone change (Guyer et al., 2019). We measure the duration of each instruction segment inside the audio and evaluate whether the duration has drastically changed, changing it to too long or too short for specific segments. For instance, for a recording that each instruction phrase is as short as 1 second, an elongated phrase like "or the second right" with a duration of 5.68 seconds can indicate hesitation and diminished certainty.

Pitch: A common feature of uncertainty is a rising intonation at the end of phrases or sentences, which can sound as if the individual is asking a question rather than making a statement (Goupil et al., 2021; Guyer et al., 2019). We identify the most unusual pitch change pattern in the waveform and align it with the instruction piece by timestamp.

Loudness: Alongside pitch, the uncertainty can also be embedded in loudness, since unpredictably rising and falling volume can indicate hesitation or self-doubt (Goupil et al., 2021; Székely et al., 2017). Similar to pitch, we only detect the most drastic loudness change in the audio clip and align it with the instruction piece by timestamp to infer possible uncertainty of instruction pieces.

To quantitatively measure pitch and loudness features, we design a lightweight yet effective algorithm (see Algorithm 1 in AppendixF) to detect the max change of these two features in the audio clip. For the duration, we split the transcription sentence into sub-instructions and measured the duration of each sub-instructions. We only mark the sub-instructions that are significantly longer than other sub-instructions in the same audio clips as uncertainty signals. Force alignment, a process that synchronizes text fragments to their corresponding segments in audio narration, is used to align the vocal features' timestamps with the sub-instructions. Examples are shown in Figure 2 and detailed in Appendix C.

In this work, we focus on significant changes in these audio features that exceeded a predetermined threshold while excluding variations in the initial and final three seconds of recordings to avoid contamination by recording noise. This audio augmentation prompt serves as input for decision-making processes that leverage language models. In a way, we give the LLMs the ear to listen to human affective cues embedded inside the vocal clip, thus improving their reasoning skills.

3.3 In-Context Prompting

Beyond Text enables the language model to exploit its inherent reasoning capacities for strategic planning and decision-making tasks through In-Context Learning (ICL) (Yousefi et al., 2023; Liu et al., 2023a; Li et al., 2023), by providing only few-shot examples with no explicit training. We illustrate ICL (Yousefi et al., 2023) and Chain-of-Thought (Wei et al., 2022) by including three comprehensive examples that delineate all potential scenarios a robot might encounter while navigating human interactions for directional assistance. Specifically, we provide examples that: 1. only has language uncertainty; 2. only has vocal uncertainty; 3. has both semantic and vocal uncertainty signals. In the response phase, the model reason to suggest a spectrum of five potential actions. Notably, the first option (A) is always a paraphrased version of the transcription sentence without uncertainty, mimicking the scenario where the robot chooses to blindly follow the human audio-guided instruction; and the final option (E) is consistently presented as "ask another person nearby for direction," ensuring a fail-safe in the decision-making process. Detailed prompts and LLM output examples are shown in the Appendix C.

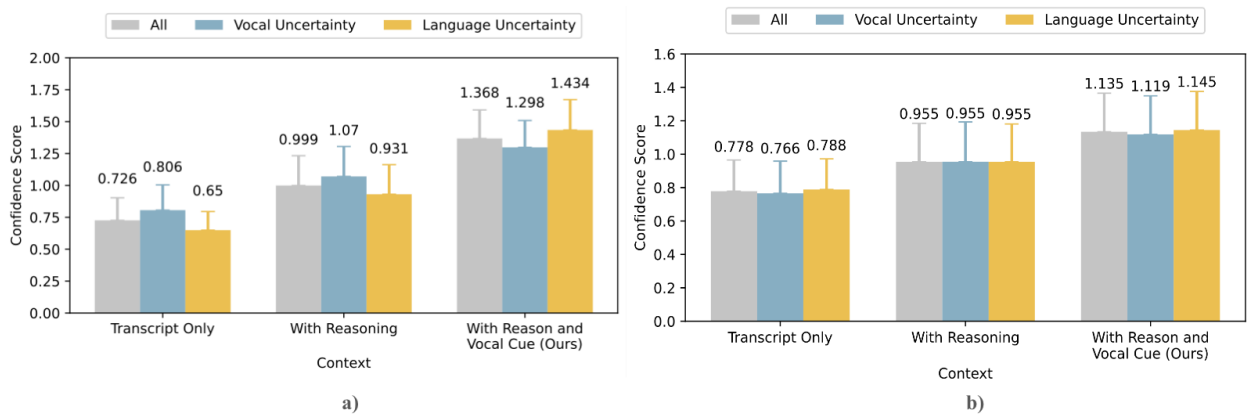

Figure 3: *a) Average Confidence Score by context and audio type for samples where the language model picks the same choices as human perception. The skinny bar is the error bar, representing the standard deviation of the confidence score. Error bars give a general idea of how precise a measurement is or, conversely, how far from the reported value the true (error-free) value might be. b) Average Confidence Score by context and audio type. An overall low variance is indicated by the increased average confidence.*

3.4 Scoring Choices by Polling LLMs

Each step in an LLM generated plan denoted as y, comprises a sequence of symbols $(\theta_1, \theta_2, ..., \theta_n)$. Traditional LLMs calculate the joint probability of these sequences as $p(y) = \prod_{i=1}^{n} p(\theta_i|\theta_1, \ldots, \theta_{i-1})$. However, this approach is flawed due to its high sensitivity to the sequence length n, rendering $p(y)$ an unreliable scoring function. To mitigate the length bias, we introduce a multiple-choice Q&A framework for planning. This method refines the planning process to select the most probable next token from a predefined set of labels $Y = \{A, B, C, D, E\}$. This results in a probability set $p(Y) = p(A), p(B), p(C), p(D), p(E)$, from which we select the label with the highest probability, $\max p(Y)$, as the optimal action. By comparing the confidence change $\Delta\mathcal{C} = \mathcal{C}(\rho_{vocal}) - \mathcal{C}(\rho_{text})$, we assess the impact of vocal prompts on the decision-making process of LLMs.

4 New Dataset: *Disfluent Navigational Instruction Audio Dataset (DNIA)*

DNIA is a dataset comprising a diverse collection of human audio recordings. All the audio clips are grounded in the robot navigation settings, as we ask the volunteers to give instructions to a location inside the buildings they have been to. These recordings capture a spectrum of navigational disfluencies, including 500 audio clips contributed by twenty individuals, including ten males and ten females, and encompassing a wide array of native and non-native speaker vocal styles. Each clip, ranging from 10 to 30 seconds, includes a series of English directives necessary for navigating to a specified location. The audio files are in WAV format, normalized with a headroom of -2dB, facilitating ease of use and consistent audio quality.

We identified and categorized two types of disfluency within our dataset: language uncertainty (LU): 285 data clips and text correct by vocal tone uncertainty (VU): 215 data clips. The LU category is the audio that contains semantic disfluency signals, such as hesitation and language uncertainty. The VU category is particularly noteworthy, as it encompasses instances where the textual transcription alone does not manifest any indicators of uncertainty, but the vocal tone shows significant uncertainty clues. Regarding ethics and privacy concerns, our dataset is collected by a group of fellow students with written consent to anonymously publish the recordings of their voices for research purposes. We reviewed all the audio files to ensure no personal information is present, and the annotations (including audio filenames) can not be matched to a specific participant. Authors have human subjects training certificates from the university. Therefore, we follow the existing guidelines to ensure the DNIA dataset does not have any ethics issues, and we will publish the dataset soon after the review process for research purposes. Table 1 shows examples for the audio prompt in DNIA, more details can be found in Appendix E.

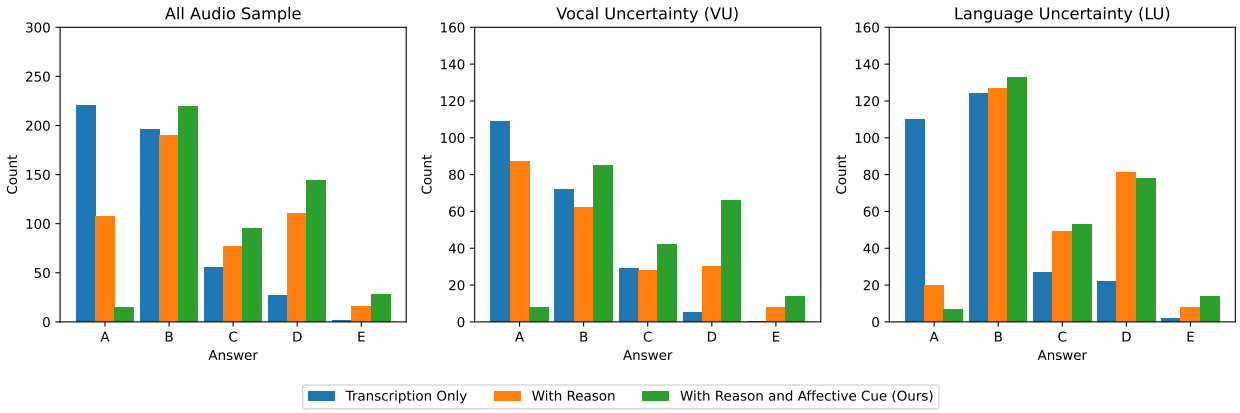

Figure 4: *Distribution of top log-probability choices by context and audio type. Our work (green) shows a dynamic change from Transcription Only (Blue). Instead of blindly choosing "A" and following the instruction with uncertainty, Beyond Text chooses "B", "C", "D", or "E", which successfully identifies the uncertainty component within the audio.*

To mimic real-world Human-Robot Interaction (HRI) scenarios as closely as possible, we align with the methodologies suggested in LLM evaluation literature (Bang et al., 2023; Chang et al., 2023). We employ human annotation as one of the primary assessment techniques. This involves manual evaluation, which is particularly relevant in HRI contexts due to its ability to capture the nuanced complexities of human interaction. We use human evaluation to calculate the winning rate of analyzing human navigational guidance uncertainty in Section 5.4. Our approach entails labeling human audio-guided instructions through a selection process from a set of five multiple-choice options designed to represent the uncertainty inherent in the content accurately. We recruit a different set of 20 participants with diverse genders and dialect backgrounds to ensure a broad range of perspectives. We developed a user study website (see Appendix G) where these participants listened to navigational audio clips. The instructions emphasized careful listening and unbiased selection, with a constraint of a maximum of one minute per clip to simulate real-world decision-making time pressures.

Table 1: Example Audio Clip Content in DNIA Dataset

Type	Example Content	Explanation
LU	"Go straight, maybe turn left at cafe shop"	'Maybe' is an uncertain word
VU	"Go straight, turn (hesitate) left at cafe shop (pitch change)"	Textual correct but voice sounds uncertain

5 Experimental Results

To validate the efficacy of our approach, we utilize confidence score and winning rate to show that *Beyond Text* can achieve low variance and low bias results compared to current LLM-based human instructions without vocal cues. We will first introduce the evaluation metric, and the comparison experiments, ablation studies on various vocal cues and LLM types.

5.1 Evaluation Metric: Confidence Measure

In this paper, we specifically focus on improving the uncertainty or the confidence score of the LM p_θ on the optimal (ground-truth) candidate responses y^{j*}. Specifically, we represent the probability distribution over the possible candidates $\{y^j\}_{j=1}^J$ by the current LM p_θ as $\rho = [\rho(y^1), \rho(y^2), \cdots, \rho(y^J)]$ and the ground-truth be represented as a dirac-delta over the true response as $\rho^* = [0, 1 \cdots 0]$, assuming the second candidate to be the optimal y^{j*} with $j* = 2$ without loss of generality. We define a confidence measure $\mathcal{C}(\rho)$ inversely

Table 2: Ablation Study for Winning Rate based on Various Vocal Cue

Vocal Cue			**Winning Rate**					
			Without Reasoning			**With Reasoning**		
Pitch	Loudness	Duration	***All***	***VU***	***LU***	***All***	***VU***	***LU***
✖	✖	✖	22.16%	22.79%	21.75%	49.30%	36.74%	58.60%
✔	✖	✖	38.52%	40.00%	37.19%	61.68%	64.65%	59.30%
✖	✔	✖	**42.31%**	**43.72%**	**41.05%**	63.07%	61.40%	64.21%
✖	✖	✔	34.33%	36.74%	33.28%	60.88%	61.40%	60.35%
✔	✔	✖	40.32%	39.53%	40.70%	67.47%	70.23%	65.26%
✔	✖	✔	37.52%	38.60%	36.49%	64.87%	66.51%	63.51%
✖	✔	✔	40.72%	42.79%	38.95%	65.47%	67.91%	63.51%
✔	✔	✔	36.93%	40.00%	34.39%	**70.46%**	**72.56%**	**68.77%**

proportional to the Kullback–Leibler divergence between ρ and ρ^* given by

$$\mathcal{C}(\rho) = \frac{1}{\mathrm{KL}(\rho, \rho^*)}, \tag{3}$$

where we note that the confidence measure will reduce as the KL divergence between ρ and ρ^* increases. The range of a confidence measure, which is inversely proportional to entropy, is bounded between $[0, \infty]$. Ideally, we would like to have the confidence measure as high as possible, which requires us to reduce $\mathrm{KL}(\rho, \rho^*)$. We first expand and note that

$$\mathrm{KL}(\rho, \rho^*) = -H(\rho) = -\sum_{j=1}^{J} \rho(y^j) \log(\rho^*(y^j)), \tag{4}$$

where $H(\rho) = -\sum_{j=1}^{J} \rho(y^j) \log(\rho(y^j))$ is the entropy of predicted distribution $\rho(y^j)$ by the LLM which quantifies the amount of information contained in random variables. We note that since we don't have any control over the second term in equation 4 because ρ^* is unknown, we can focus on minimizing the first term in equation 4, which tells us to maximize the entropy in order to reduce the KL divergence between ρ and ρ^*. We hypothesize that one of the major reasons for this uncertainty is the information loss while considering only textual information x and losing various critical attributes of the input audio information v. From the standard information theory, one can quantify this loss in information as $\mathcal{I}_l = H(V) - I(X; V)$ as the difference between the entropy of the original audio input random variable V and the mutual information between the text random variable X and V represented as $I(Y; V)$ which measures the amount of information that is shared between X, V. Notably, a higher confidence score—which indicates a lower KL divergence between ρ and ρ^*—is desirable, as it signifies a closer alignment with the ground truth. In this work, we empirically show that adding the vocal cue in the input helps to reduce the $\mathrm{KL}(\rho, \rho^*)$ by plotting the confidence measure equation 3 for robotic navigation tasks in Section 5.

5.2 Confidence Score Improvement

As discussed in Section 5.1, we show that our method achieves a low variance with an increased confidence score. To better visualize the low variance of our approach, we first selected a subset of 100 data instances wherein the choices made by Large Language Models (LLMs) were in alignment with user selections, shown in Figure 3. Then, we analyze the average confidence score across all 500 data clips in the dataset to demonstrate the efficacy and generalizability of our framework, shown in Figure 3. It reveals a marked increase in confidence score: *Beyond Text* significantly surpasses the text-only language model, both the transcription-only and the LLM using Chain-of-Thought (CoT). This trend was also observed across two distinct categories. The introduction of affective cues significantly lowers variance, enhancing existing LLMs' ability to interpret human vocal uncertainty with confidence.

Table 3: Winning Rate by Context and Audio Type

Method	**Winning Rate**		
	All	***VU***	***LU***
Transcription Only LLM	22.16%	22.79%	21.75%
With Chain-of-Thought Reasoning	49.30%	36.74%	58.60%
Ours	**70.46%**	**72.56%**	**68.77%**

5.3 Choice Distribution Change

Figure 4 shows the count of top log-prob choice for each audio clip in our dataset. As audio clips contain various types of disfluency, we want the framework to eventually pick from choices (B),(C),(D),(E) that identify the uncertainty within the audio and handle these uncertainties accordingly (choice structure discussed in Section 3.3), such as looking for more information or asking another person at the location point where human voice sounds unsure.

The blue bars mostly accumulated on choice (A) (220 out of 500), meaning that with only transcription, the LLM tends to blindly follow the navigational instruction without identifying uncertainty in the instruction, which is undesirable. With reasoning in in-context learning, the distribution shifts, and with the affective cue audio augmentation, the data that result in top log-prob on choice (A) becomes even lower (15 out of 500). The shift in choice distribution qualitatively shows the effectiveness of our method in choosing the next-step action that handles human vocal uncertainty.

5.4 Winning Rate

Besides indicating low variances and multiple choice distribution change, we also calculate the winning rate. We compare the human annotations for each DNIA data clip with the highest log-probability choices generated by the model to ascertain whether the choices and reasoning processes generated by the LLMs are in harmony with human conceptual understanding. The results, as detailed in Table 3, show the limitation of current LLMs in explaining vocal uncertainty from only textual input and the effectiveness of our audio augmentation approach. Given that the audio clips originated from human speakers, the most reliable measure of their trustworthiness was human perception, hence the incorporation of a user study to gauge the winning rate with which the language model identified uncertainties in the audio clips and selected the appropriate next-step action.

Table 3 demonstrates that our method exhibits reduced bias, outperforming both the single-modal transcription method and the reasoning-augmented approach. Integrating vocal affective cues, which allow LLMs to process how statements are spoken, markedly enhanced performance. The overall winning rate surged to 70.46%, with a pronounced improvement in VU interpretation, evident in a 72.56% winning rate.

Table 4: Winning Rate on Different LLMs

LLMs	**Winning Rate**		
	All	***VU***	***LU***
GPT-3.5-turbo-instruct (Transcription Only)	22.16%	22.79%	21.75%
GPT-3.5-turbo-instruct (Beyond Text)	**70.46%**	**72.56%**	**68.77%**
text-davinci-003 (Transcription Only)	57.88%	57.67%	57.89%
text-davinci-003 (Beyond Text)	**69.06%**	**69.77%**	**68.42%**
Gemini-1.5-Pro (Transcription Only)	43.11%	41.40%	44.56%
Gemini-1.5-Pro (Beyond Text)	**65.27%**	**66.98%**	**63.86%**

5.5 Ablation Study

Vocal Cue: To elucidate the impact of individual components within the vocal cue framework, we conduct a comprehensive qualitative winning rate study by isolating each category of vocal cues and presenting them exclusively to the language model. We report the vocal cue ablative study in Table 2. In the absence of reasoning, the highest winning rate was observed with the exclusive use of loudness features. This

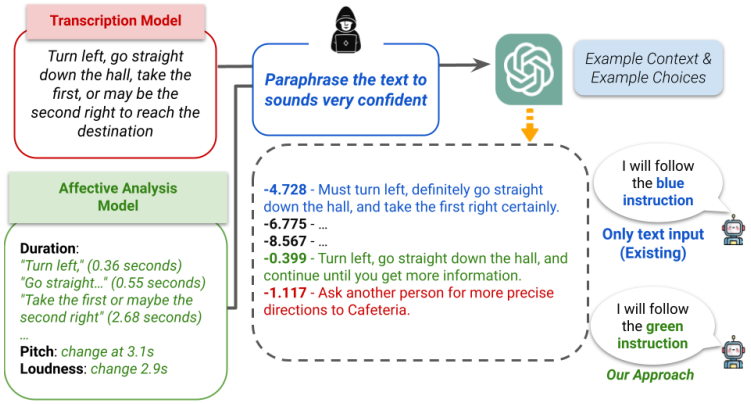

Figure 5: *The Adversarial Language Model Attack Pipeline. An adversarial attack that paraphrases the input text to sound very certain by deleting textual uncertainty signals is applied.*

suggests that incorporating multiple vocal features without reasoning may overwhelm the model, hindering its ability to correlate vocal characteristics with uncertainty. Conversely, with reasoning integrated, vocal cues significantly strengthened the LLM's proficiency in interpreting vocal uncertainty and facilitating robotic navigation, as reflected in the increased winning rates.

LLM Type: To show the generalized ability of *Beyond Text*, we test it on three different popular LLMs, including GPT3.5, GPT 3, and Gemini. Based on the result shown in Table 4, we conclude that *Beyond Text* improves the winning rate on all three LLMs significantly. Details about parameters for each type of LLM is shown in Appendix A.

6 Robustness to Adversarial LLM Attacks

With the advancement of language models, safety concerns such as adversarial attacks and token manipulation prompts have become increasingly prominent. Adversarial attacks or token manipulation prompts (Ribeiro et al., 2018; Shayegani et al., 2023) could potentially trigger the model to output something undesired (Zou et al., 2023; Greshake et al., 2023). Given a piece of text input containing a sequence of tokens, we can apply simple token operations like replacement with synonyms to trigger the model to make the incorrect predictions (Morris et al., 2020; Li et al., 2020).

Table 5: Winning Rate by Context and Audio Type (Under Token Manipulation Attack)

Method	**Winning Rate**		
	All	***VU***	***LU***
Transcription Only	22.16%	22.79%	21.75%
Transcription Only (Attack)	9.78%	11.23%	7.91%
Decrease (in Percentage)	**55.87%**	**50.72%**	**63.63%**
Ours	70.46%	72.56%	68.77%
Ours (Attack)	46.90%	50.23%	44.21%
Decrease (in Percentage)	**33.43%**	**30.77%**	**35.71%**

We demonstrate that *Beyond Text* exhibits robustness against token manipulation adversarial attacks targeting LLMs. Our attack pipeline is designed as follows: Post-transcription, a paraphraser alters the original transcription ($T1$) into a new version ($T2$), systematically replacing uncertainty semantics with deterministic language. We then adjust the prompt to align the vocal and textual information with $T2$. Subsequently, the LLM is instructed to replicate $T2$ as an option (A) instead of generating a freeform paraphrase ($T1'$) devoid of uncertainty cues. This approach highlights the limitations of current LLMs, which tend to over-rely on semantic textual instructions with no clue about the nuances of vocal delivery. Figure 5 illustrates the attack pipeline A detail example is shown in Appendix D.

Table 5 illustrates the comparative robustness of *Beyond Text*. While the transcription-only approach exhibits a significant performance decrease of 55.87% under attack, *Beyond Text*, with its audio augmentation preserving vocal cues, exhibited a notably lower reduction of 33.43%. This suggests that audio augmentation in *Beyond Text* enables LLMs to maintain reasoning capabilities and resist being misled by text-based adversarial attacks.

7 Conclusions and Open Questions

In our work, we highlight the shortcomings of current LLMs in robotic navigation, specifically their inability to compute uncertainties in human audio instructions. This issue largely stems from the reliance on audio transcriptions in state-of-the-art methods, leading to a loss of critical audio features such as duration, pitch, and loudness. Therefore, we state our position that we cannot just reply to the transcriptions of the human audio descriptions but should instead also focus on the audio features. To this end, we introduced *Beyond Text*, a novel approach that integrates the affective cues from the human voice and provides empirical evidence that it significantly enhances the LLMs' ability to interpret human vocal uncertainty. We developed the DNIA audio dataset to push forward the research in this direction. Notably, *Beyond Text* outperforms existing text-based language model techniques in terms of win rate and entropy confidence, as demonstrated in our user study. This represents a progression in the development of sophisticated audio-augmented LLMs. It underscores the importance of incorporating affective audio computing capabilities in LLMs, particularly for enhancing social robotics navigation tasks.

Open Questions. Next, we emphasize several critical questions and directions pertaining to the research on audio augmentation to improve LLMs' ability to understand human speech, which are still unsolved and need further discussion.

(1) Improving Quality of Audio Augmentation: It is essential to consider more audio features, such as intonation and prosody, with learning-based affective computing methods to improve the quality of audio augmentation. The DNIA is a foundational dataset designed for studying human uncertainty in audio-guided navigation tasks, but it can also be used in other HRI-related tasks. In the future, the dataset will be expanded through the addition of more related vocal features. Additionally, the relation between different vocal features also needs to be carefully articulated for different tasks so that LLMs can provide concrete reasoning. As we stated the position that affective cue analysis is currently missing from existing LLM, and beyond text is an audio augmentation method that can help LLM hear not only 'what' people said but also 'how' people convey their ideas; it is also excited to explore the potential of vocal features in improving audio-to-text tasks, such as video emotion understanding, etc.

(2) Fine-Grained Definition of Uncertainty: *Beyond Text* marks a notable first step in interpreting human uncertainty in human-robot conversations, but it has limitations. For instance, a directive like "Turn left at the washroom. Oh, sorry, turn right at the washroom" could signify either uncertainty or a corrected command, highlighting the complexity of human speech. It is crucial to address such uncertainty in a fundamental manner in future research.

(3) Advancing LLMs to Listen Audio without Text: While we currently rely on textual transcriptions, emerging research, notably Tang et al. (2023), demonstrates LLMs processing audio directly with considerable success in speech tasks. However, these models still lack the capability to fully grasp human emotional nuances, especially in conveying uncertainty.

(4) Integration with Real World Social Robotics: *Beyond Text* introduces a human audio-guided framework tailored for social robotic navigation, designed to discern human trust from a robotic perspective. However, it is important to acknowledge that the real-world deployment of such a model faces additional complexities, particularly due to the presence of multiple audio sources beyond human speech, which request the attention of the research community.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, et al. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*, 2022.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*, 2023.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, et al. Audiolm: a language modeling approach to audio generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

Edgar Chan, Oliver Baumann, Mark A Bellgrove, and Jason B Mattingley. From objects to landmarks: the function of visual location information in spatial navigation. *Frontiers in psychology*, 3:304, 2012.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 2023.

Vishnu Sashank Dorbala, Arjun Srinivasan, and Aniket Bera. Can a robot trust you?: A drl-based approach to trust-driven human-guided navigation. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3538–3545. IEEE, 2021.

Manfred Eppe, Sean Trott, and Jerome Feldman. Exploiting deep semantics and compositionality of natural language for human-robot-interaction. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 731–738. IEEE, 2016.

Roya Firoozi, Johnathan Tucker, Stephen Tian, Anirudha Majumdar, Jiankai Sun, Weiyu Liu, Yuke Zhu, Shuran Song, Ashish Kapoor, Karol Hausman, et al. Foundation models in robotics: Applications, challenges, and the future. *arXiv preprint arXiv:2312.07843*, 2023.

Anthony Francis, Claudia Pérez-d'Arpino, Chengshu Li, Fei Xia, Alexandre Alahi, Rachid Alami, Aniket Bera, Abhijat Biswas, Joydeep Biswas, Rohan Chandra, et al. Principles and guidelines for evaluating social robot navigation algorithms. *arXiv preprint arXiv:2306.16740*, 2023.

Deepanway Ghosal, Navonil Majumder, Ambuj Mehrish, and Soujanya Poria. Text-to-audio generation using instruction-tuned llm and latent diffusion model. *arXiv preprint arXiv:2304.13731*, 2023.

Reginald G Golledge. Human wayfinding and cognitive maps. *Colonization of unfamiliar landscapes: the archaeology of adaptation*, 25, 2003.

Louise Goupil, Emmanuel Ponsot, Daniel Richardson, Gabriel Reyes, and Jean-Julien Aucouturier. Listeners' perceptions of the certainty and honesty of a speaker are associated with a common prosodic signature. *nature communications*, 2021.

Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. More than you've asked for: A comprehensive analysis of novel prompt injection threats to application-integrated large language models. *arXiv e-prints*, pp. arXiv–2302, 2023.

Joshua J Guyer, Leandre R Fabrigar, and Thomas I Vaughan-Johnston. Speech rate, intonation, and pitch: Investigating the bias and cue effects of vocal confidence on persuasion. *Personality and Social Psychology Bulletin*, 45(3):389–405, 2019.

Zhe Hu, Jia Pan, Tingxiang Fan, Ruigang Yang, and Dinesh Manocha. Safe navigation with human instructions in complex scenes. *IEEE Robotics and Automation Letters*, 4(2):753–760, 2019.

Xiaoming Jiang and Marc D Pell. The sound of confidence and doubt. *Speech Communication*, 88:106–126, 2017.

Alexander Ku, Peter Anderson, Roma Patel, Eugene Ie, and Jason Baldridge. Room-across-room: Multilingual vision-and-language navigation with dense spatiotemporal grounding. *arXiv preprint arXiv:2010.07954*, 2020.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. 2023.

Egor Lakomkin, Mohammad Ali Zamani, Cornelius Weber, Sven Magg, and Stefan Wermter. On the robustness of speech emotion recognition for human-robot interaction with deep neural networks. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 854–860. IEEE, 2018a.

Egor Lakomkin, Mohammad Ali Zamani, Cornelius Weber, Sven Magg, and Stefan Wermter. On the robustness of speech emotion recognition for human-robot interaction with deep neural networks. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 854–860, 2018b. doi: 10.1109/IROS.2018.8593571.

Boyi Li, Philipp Wu, Pieter Abbeel, and Jitendra Malik. Interactive task planning with language models, 2023.

Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. Bert-attack: Adversarial attack against bert using bert. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6193–6202, 2020.

Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, and Andy Zeng. Code as policies: Language model programs for embodied control. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 9493–9500. IEEE, 2023.

Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. *arXiv preprint arXiv:2312.07533*, 2023a.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Generating with confidence: Uncertainty quantification for black-box large language models. *arXiv preprint arXiv:2305.19187*, 2023b.

Bo Liu, Yuqian Jiang, Xiaohan Zhang, Qiang Liu, Shiqi Zhang, Joydeep Biswas, and Peter Stone. Llm+ p: Empowering large language models with optimal planning proficiency. *arXiv preprint arXiv:2304.11477*, 2023a.

Zeyi Liu, Arpit Bahety, and Shuran Song. Reflect: Summarizing robot experiences for failure explanation and correction. *arXiv preprint arXiv:2306.15724*, 2023b.

Paria Jamshid Lou, Yufei Wang, and Mark Johnson. Neural constituency parsing of speech transcripts. *arXiv preprint arXiv:1904.08535*, 2019.

Jiageng Mao, Yuxi Qian, Hang Zhao, and Yue Wang. Gpt-driver: Learning to drive with gpt. *arXiv preprint arXiv:2310.01415*, 2023.

Trisha Mittal, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. Emotions don't lie: An audio-visual deepfake detection method using affective cues. In *Proceedings of the 28th ACM international conference on multimedia*, pp. 2823–2832, 2020.

John Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 119–126, 2020.

Jun Ogata, Masataka Goto, and Katunobu Itou. The use of acoustically detected filled and silent pauses in spontaneous speech recognition. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4305–4308. IEEE, 2009.

Jillian L Prestopnik and Beverly Roskos-Ewoldsen. The relations among wayfinding strategy use, sense of direction, sex, familiarity, and wayfinding ability. *Journal of environmental psychology*, 20(2):177–191, 2000.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pp. 28492–28518. PMLR, 2023.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

Allen Z Ren, Anushri Dixit, Alexandra Bodrova, Sumeet Singh, Stephen Tu, Noah Brown, Peng Xu, Leila Takayama, Fei Xia, Jake Varley, et al. Robots that ask for help: Uncertainty alignment for large language model planners. In *Proceeding of 2023 Conference on Robot Learning (CoRL 2023)*, 2023.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Semantically equivalent adversarial rules for debugging nlp models. In *Annual Meeting of the Association for Computational Linguistics*, 2018. URL `https://api.semanticscholar.org/CorpusID:21740766`.

Dhruv Shah, Michael Robert Equi, Błażej Osiński, Fei Xia, Brian Ichter, and Sergey Levine. Navigation with large language models: Semantic guesswork as a heuristic for planning. In *Conference on Robot Learning*, pp. 2683–2699. PMLR, 2023.

Erfan Shayegani, Yue Dong, and Nael Abu-Ghazaleh. Plug and pray: Exploiting off-the-shelf components of multi-modal models. *arXiv preprint arXiv:2307.14539*, 2023.

Chak Lam Shek, Xiyang Wu, Dinesh Manocha, Pratap Tokekar, and Amrit Singh Bedi. Lancar: Leveraging language for context-aware robot locomotion in unstructured environments. *arXiv preprint arXiv:2310.00481*, 2023.

S Shyam Sundar. Rise of machine agency: A framework for studying the psychology of human–ai interaction (haii). *Journal of Computer-Mediated Communication*, 25(1):74–88, 2020.

Eva Székely, Joseph Mendelson, and Joakim Gustafson. Synthesising uncertainty: The interplay of vocal effort and hesitation disfluencies. In *INTERSPEECH*, pp. 804–808, 2017.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. Salmonn: Towards generic hearing abilities for large language models. *arXiv preprint arXiv:2310.13289*, 2023.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

Jimmy Wu, Rika Antonova, Adam Kan, Marion Lepert, Andy Zeng, Shuran Song, Jeannette Bohg, Szymon Rusinkiewicz, and Thomas Funkhouser. Tidybot: Personalized robot assistance with large language models. *arXiv preprint arXiv:2305.05658*, 2023.

Safoora Yousefi, Leo Betthauser, Hosein Hasanbeig, Akanksha Saran, Raphaël Millière, and Ida Momennejad. In-context learning in large language models: A neuroscience-inspired analysis of representations. *arXiv preprint arXiv:2310.00313*, 2023.

Andy Zou, Zifan Wang, J Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.

Contents

1 Introduction **1**

2 Related Works **2**

3 Our Approach **4**

3.1 Semantic Uncertainty Quantification 4

3.2 Vocal Affective Cue Analysis 5

3.3 In-Context Prompting 5

3.4 Scoring Choices by Polling LLMs 6

4 New Dataset: *Disfluent Navigational Instruction Audio Dataset (DNIA)* **6**

5 Experimental Results **7**

5.1 Evaluation Metric: Confidence Measure 7

5.2 Confidence Score Improvement 8

5.3 Choice Distribution Change 9

5.4 Winning Rate 9

5.5 Ablation Study 9

6 Robustness to Adversarial LLM Attacks **10**

7 Conclusions and Open Questions **11**

A LLM Implementation Design **17**

B Transcription Model Prompting **17**

C Example Prompts and LLM Outputs **18**

C.1 Prompts and Examples for Generating Reasoning and Choices (GPT-4) 18

C.1.1 Without Affective Cue 18

C.1.2 With Affective Cue 20

C.2 Prompts and Examples for Choice Selection (GPT-3.5) 21

C.2.1 Transcript Only 22

C.2.2 With Transcription and Reasoning 23

C.2.3 Transcription, Reasoning, and Affective Cue 24

C.3 Large Language Model Adversarial Token Manipulation Attack Prompt 25

D Adversarial Attack Example **26**

E DNIA Dataset Details and Samples **26**

F Pitch and Loudness Shifts Detection **27**

G User Interface for Human Evaluation on DNIA Dataset **27**

A LLM Implementation Design

We present the pipeline of querying LLM to analyze human audio inputs in Figure 6. GPT-4 is used for choice generation in the first stage, with in-context examples provided. The response contains the chain-of-thought reasoning process for the final action choice. The LLM output is cascaded to GPT-3.5 for scoring the confidence of each choice. In each stage, we enforce LLM to return output in JSON format, ensuring consistency in both our implementation and the parsing process. We also encapsulate the prompts in an API request parallel processor to speed up our workflow. For clarity and transparency, we used

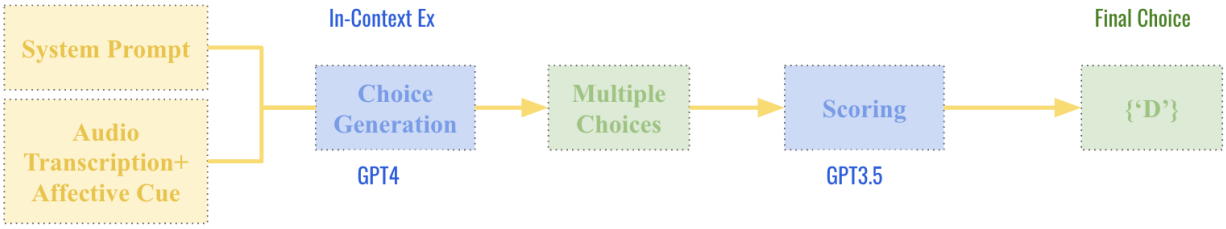

Figure 6: *Beyond Text* Implementation Flow with LLMs. Final Choice 'D' only represents the output format. It can output any choices from A,B,C,D,E.

`gpt-4-1106-preview` for choice generation and `gpt-3.5-turbo-instruct` model. For `Gemini-1.5-Pro` and `gpt-4-1106-preview`, we used default hyperparameters except for the temperature, which is set to 0. For `gpt-3.5-turbo-instruct`, `text-davinci-003`, we set **max_token**: 100; **temperature**: 0; **echo**: true; and the following hyper-parameters:

Table 6: Logit Bias Settings

Token ID	Character (with space)	Logit Bias
362	A	100.0
426	B	100.0
356	C	100.0
423	D	100.0
469	E	100.0

B Transcription Model Prompting

We employ the state-of-the-art Whisper model (Radford et al., 2023) to transcribe the human audio-guided instruction. We use **Whisper-Small** with 244M parameters that require 2GB VRAM to maintain the seamless transcription speed (within 1 second). We add a custom prompting *"Umm, let me think like, hmm... Okay, here's what I'm, like, thinking."* to enhance its proficiency in capturing subtle nuances of speech, such as word mumbling, hesitations, and speech repairs that may sound like a slip of the tongue, as these are crucial uncertainty-related semantic signals.

C Example Prompts and LLM Outputs

C.1 Prompts and Examples for Generating Reasoning and Choices (GPT-4)

You are a robot exploring a building. You are in the middle of the building and are seeking directions from a human.

Your goal is to provide possible next steps based on human response. A human response may not be trustworthy and contain disfluency.

You should figure out the unreliable part of the human response and generate five different next-step actions based on the reasoning.

Your answer must strictly follow the JSON format the example below provided. You must strictly follow the format of the example provided below for your reasoning and next-step actions.

Each next-step action must be distinct. The last next-step action, E), must be "Ask another person near you for direction." For the rest of the next actions, if there is uncertainty, say "ask other people for further instruction" at the instruction where the uncertainty appears.

Do not paraphrase. Always provide reasoning.

Prompt 1: System prompt for generating reasoning and choices.

C.1.1 Without Affective Cue

Task: Find the direction to Cafeteria

Human Response: {} # To be filled inc

Question: What are possible next steps? Please always give me 5 possible answers, starting with A), B), C), D), E).

Prompt 2: Prompt for generating choices and reasoning without affective cue.

In-Context Example

User Request

Task: Find the direction to the toilet

Human Response: Go straight, and err, take the second left

Question: What are possible next steps? Please always give me 5 possible answers, starting with A), B), C), D), E).

Agent Response

```
{
    "Reasoning":"'Err' is a hesitation signals that shows human response is
        not trustworthy.",
    "Answer":{
      "A": "Go straight, take the second left",
      "B": "Go backward, take the second left",
      "C": "Go straight, take the first left",
      "D": "Go straight, at the second left, ask another person for detailed
          guidance",
      "E": "Ask another person near you for direction",
    }
}
```

Example Input & Output

Human Audio:

```
"So you start from the mail office and then you just walk straight
    and take the... maybe just take the first right turn and then you
    go straight and you will see the cafeteria."
```

LLM Output:

```
{
    "Reasoning":"The use of 'maybe' indicates uncertainty in the human
        response, which makes it not fully trustworthy.",
    "Answer":{
    A: Start from the mail office, walk straight and take the first right
        turn, then go straight and you will see the cafeteria
    B: Start from the mail office, walk straight and take the second right
        turn, then go straight and you will see the cafeteria
    C: Start from the mail office, walk straight and take the first left turn
        , then go straight and you will see the cafeteria
    D: Start from the mail office, walk straight and at the first right turn,
         ask another person for further instruction
    E: Ask another person near you for direction
    }
}
```

C.1.2 With Affective Cue

Task: Find the direction to Lab B

Human Response: {} # To be filled in

Human Vocal Affective Cue: {} # To be filled in

Question: What are possible next steps? Please always give me 5 possible answers, starting with A), B), C), D), E).

Prompt 3: Prompt for generating choices and reasoning with affective cue

In-Context Example

User Request

Task: Find the direction to the toilet

Human Response: Go straight, and err, take the second left

Human Vocal Affective Cue:

Large loudness decrease: No Change

Large pitch change: at time = 02.200 second

Duration: "Go straight." ⇒ [00.000, 00.360] (0.36 seconds);

Duration: "and err, take the second left" ⇒ [00.360, 03.480] (3.12 seconds).

Question: What are possible next steps? Please always give me 5 possible answers, starting with A), B), C), D), E).

Agent Response

```
{
    "Reasoning":"'Err' is a hesitation signals that shows human response is
        not trustworthy. The long duration on 'and err, take the second left'
         3.12 seconds, shows a lack of certainty on whether to take second
        left to arrive at toilet. The large pitch is a sign of intonation
        increase that sounds like question mark. It is at 02.200 second, in
        the range of [00.360, 03.480] (last instruction phrase), confirm that
         the last part of response, 'and err, take the second left', is not
        trustworthy.",
    "Answer":{
      "A": "Go straight, take the second left",
      "B": "Go backward, take the second left",
      "C": "Go straight, take the first left",
      "D": "Go straight, at the second left, ask another person for detailed
          guidance",
      "E": "Ask another person near you for direction",
    }
}
```

Example Input & Output

Human Audio and Affective Cue:

```
"So first, when you start from lab A, you wanna go straight and you
    will pass lab A and then the locker room. And then you will
    probably need to make a left turn. And then you go straight and
    you will see lab B."

Large loudness decrease: No Change
Large pitch change: at time = 23.000 second
Duration: "So first," => [0.000, 2.160] (2.16 seconds);
Duration: "when you start from lab A," => [2.160, 4.200] (2.04
    seconds);
Duration: "you wanna go straight" => [4.200, 6.840] (2.64 seconds);
Duration: "and you will pass lab A" => [6.840, 9.280] (2.44 seconds);
Duration: "and then the locker room." => [9.280, 12.440] (3.16
    seconds);
Duration: "And then you will probably need to make a left turn." =>
    [12.440, 19.840] (7.40 seconds);
Duration: "And then you go straight" => [19.840, 22.720] (2.88
    seconds);
Duration: "and you will see lab B." => [22.720, 28.080] (5.36 seconds
    ););
```

LLM Output:

```
{
    "Reasoning":"The phrase "And then you will probably need to make a left
        turn" has a long duration of 7.40 seconds, which indicates hesitation
         or uncertainty in the direction given. Additionally, the large pitch
         change at 23.000 second, which falls within the duration of "and you
         will see lab B", suggests that the speaker is unsure about the final
         part of the instruction. The use of the word 'probably' also adds to
         the uncertainty of the direction to make a left turn.",
    "Answer":{
    A: Start from lab A, go straight, pass lab A and the locker room, make a
        left turn, go straight, and you will see lab B
    B: Start from lab A, go straight, pass lab A and the locker room, and ask
         other people for further instruction
    C: Start from lab A, go straight, pass lab A and the locker room, make a
        right turn, go straight, and you will see lab B
    D: Start from lab A, go straight, pass lab A and the locker room, make a
        left turn, and then ask other people for further instruction
    E: Ask another person near you for direction
    }
}
```

C.2 Prompts and Examples for Choice Selection (GPT-3.5)

Here, we show the success results for a vocal uncertainty(VU) example. Without *Beyond Text*, the LLM picks choice A, blindly following the human instruction. With reasoning, the same result. With *Beyond Text*, the LLM changes its choice from A to B, successfully identifying the uncertainty within the human voice.

You are a robot exploring a building. You are in the middle of the building and are seeking directions from a human.

Your goal is to provide possible next steps based on human response. The human response may not be trustworthy and contain disfluency. Based on the following information, find the optimal choice that most accurately reflects the human response.

If the human response is not trustworthy, you should choose the choice that asks for help at the uncertain instruction part.

Prompt 4: System Prompt for selecting an answer from choices.

C.2.1 Transcript Only

Human Response: {} # To be filled in

Your goal is to find the direction to the Cafeteria

Question: where should you go next?

Choices: {} # To be filled in

Answer:

Prompt 5: Prompt for selecting an answer from choices with transcription only.

Example Input & Output

Human Audio:

```
"So start from the mail office, you wanna go straight and then when you
    are at the auditorium, you make a left turn. And then you go straight
    again and here you wanna take the first right turn and then if you just
     go straight again, you will see the cafeteria."
```

Choices:

```
A: Start from the mail office, go straight, at the auditorium make a left
    turn, go straight, take the first right turn, then go straight again
    until you see the cafeteria
B: Start from the mail office, go straight, at the auditorium make a left
    turn, go straight, take the first right turn, then ask other people for
     further instruction
C: Start from the mail office, go straight, at the auditorium make a right
     turn, go straight, take the first right turn, then go straight again
    until you see the cafeteria
D: Start from the mail office, go straight, at the auditorium make a left
    turn, go straight, take the second right turn, then go straight again
    until you see the cafeteria
E: Ask another person near you for direction
```

LLM Output:

```
A: Start from the mail office, go straight, at the auditorium make a left
    turn, go straight, take the first right turn, then go straight again
    until you see the cafeteria
```

C.2.2 With Transcription and Reasoning

Human Response: {} # To be filled in

Your goal is to find the direction to the Cafeteria

Reasoning is: {} # To be filled in

Question: where should you go next?

Choices: {} # To be filled in

Answer:

Prompt 6: Prompt for selecting answers from choices with transcription and reasoning.

Example Input & Output

Human Audio:

```
"So start from the mail office, you wanna go straight and then when you
    are at the auditorium, you make a left turn. And then you go straight
    again and here you wanna take the first right turn and then if you just
     go straight again, you will see the cafeteria."
```

Reasoning:

```
The human response contains filler words like 'So' and 'just', which
    typically do not alter the directions provided. The directions seem
    clear and without any contradictory statements.
```

Choices:

```
A: Start from the mail office, go straight, at the auditorium make a left
    turn, go straight, take the first right turn, then go straight again
    until you see the cafeteria
B: Start from the mail office, go straight, at the auditorium make a left
    turn, go straight, take the first right turn, then ask other people for
     further instruction
C: Start from the mail office, go straight, at the auditorium make a right
     turn, go straight, take the first right turn, then go straight again
    until you see the cafeteria
D: Start from the mail office, go straight, at the auditorium make a left
    turn, go straight, take the second right turn, then go straight again
    until you see the cafeteria
E: Ask another person near you for direction
```

LLM Output:

```
A: Start from the mail office, go straight, at the auditorium, make a left
   turn, go straight, take the first right turn, then go straight again
   until you see the cafeteria
```

C.2.3 Transcription, Reasoning, and Affective Cue

Human Response: {} # To be filled in

Human Affective Cue: {} # To be filled in

Your goal is to find the direction to the Cafeteria

Reasoning is: {} # To be filled in

Question: where should you go next?

Choices: {} # To be filled in

Answer:

Prompt 7: Prompt for selecting an answer from choices with transcription, reasoning, and affective cue.

Example Output

Human Audio:

```
"So start from the mail office, you wanna go straight and then when you
   are at the auditorium, you make a left turn. And then you go straight
   again and here you wanna take the first right turn and then if you just
    go straight again, you will see the cafeteria."
```

Human Affective Cue:

```
Large loudness decrease: at time = 17.000 second
Large pitch change: at time = 23.000 second
Duration: "So start from the mail office," => [0.000, 4.880] (4.88 seconds
   );
Duration: "you wanna go straight" => [4.880, 8.680] (3.80 seconds);
Duration: "and then when you are at the auditorium," => [8.680, 11.040]
   (2.36 seconds);
Duration: "you make a left turn." => [11.040, 12.480] (1.44 seconds);
Duration: "And then you go straight again" => [12.480, 15.680] (3.20
   seconds);
Duration: "and here you wanna take the first right turn" => [15.680,
   22.560] (6.88 seconds);
Duration: "and then if you just go straight again," => [22.560, 28.600]
   (6.04 seconds);
```

```
Duration: "you will see the cafeteria." => [28.600, 34.640] (6.04 seconds)
    ;
```

Reasoning:

```
The human response contains a large loudness decrease at 17.000 second and
     a large pitch change at 23.000 second, which correspond to the phrases
     "and here you wanna take the first right turn" and "and then if you
    just go straight again," respectively. These vocal affective cues
    suggest uncertainty in these parts of the instructions. The duration of
     the phrases "and here you wanna take the first right turn" (6.88
    seconds) and "you will see the cafeteria." (6.04 seconds) are also
    longer than the other parts, which could indicate a lack of confidence
    or certainty in these directions.
```

Choices:

```
A: Start from the mail office, go straight, at the auditorium make a left
    turn, go straight, take the first right turn, then go straight again
    until you see the cafeteria
B: Start from the mail office, go straight, at the auditorium make a left
    turn, go straight, take the first right turn, then ask other people for
     further instruction
C: Start from the mail office, go straight, at the auditorium make a right
     turn, go straight, take the first right turn, then go straight again
    until you see the cafeteria
D: Start from the mail office, go straight, at the auditorium make a left
    turn, go straight, take the second right turn, then go straight again
    until you see the cafeteria
E: Ask another person near you for direction
```

LLM Output:

```
B: Start from the mail office, go straight, at the auditorium make a left
    turn, go straight, take the first right turn, then ask other people for
    further instruction
```

C.3 Large Language Model Adversarial Token Manipulation Attack Prompt

The LLM attack uses GPT4 as the adversarial paraphraser to attack the original text input. The attack takes the sub-instruction list of each original audio clip as input and then paraphrases it to reverse the semantic tone. The GPT4 attack prompt is shown below:

You are a paraphraser. Your goal is to change the level of certainty of the human instruction from an audio transcript.

You achieve your goal by removing all uncertain cues in the text, such as 'probably', 'maybe', 'possibly', 'should' and make the semantic of the transcript very confident certain by adding words like 'certainly', 'must', or 'without doubt'.

You will take a transcript formatted as a list of timestamped sentences and produce a paraphrase of the transcript, strictly following the format of the original transcript.

Prompt 8: Prompt for adversarial token manipulation attack.

D Adversarial Attack Example

Before Attack:

```
"So first, when you start from lab A, you wanna go straight and you
    will pass lab A and then the locker room. And then you will
    probably need to make a left turn. And then you go straight and
    you will see lab B."

Large loudness decrease: No Change
Large pitch change: at time = 23.000 second
Duration: "So first," => [0.000, 2.160] (2.16 seconds);
Duration: "when you start from lab A," => [2.160, 4.200] (2.04
    seconds);
Duration: "you wanna go straight" => [4.200, 6.840] (2.64 seconds);
Duration: "and you will pass lab A" => [6.840, 9.280] (2.44 seconds);
Duration: "and then the locker room." => [9.280, 12.440] (3.16
    seconds);
Duration: "And then you will probably need to make a left turn." =>
    [12.440, 19.840] (7.40 seconds);
Duration: "And then you go straight" => [19.840, 22.720] (2.88
    seconds);
Duration: "and you will see lab B." => [22.720, 28.080] (5.36 seconds
    ););
```

After Attack:

```
"First, start from lab A, you go straight and pass lab A and then the
     locker room. And then you will make a left turn. And then you go
    straight and you will see lab B."

Large loudness decrease: No Change
Large pitch change: at time = 23.000 second
Duration: "First," => [0.000, 2.160] (2.16 seconds);
Duration: "start from lab A," => [2.160, 4.200] (2.04 seconds);
Duration: "you go straight" => [4.200, 6.840] (2.64 seconds);
Duration: "and pass lab A" => [6.840, 9.280] (2.44 seconds);
Duration: "and then the locker room." => [9.280, 12.440] (3.16
    seconds);
Duration: "And then you will make a left turn." => [12.440, 19.840]
    (7.40 seconds);
Duration: "And then you go straight" => [19.840, 22.720] (2.88
    seconds);
Duration: "and you will see lab B." => [22.720, 28.080] (5.36 seconds
    ););
```

E DNIA Dataset Details and Samples

For VU clips, we ensure no semantic uncertainty signals, such as uncertain-related words (i.e., 'maybe,' 'I guess'), speech repairs, and hesitations. We believe those are common affective characteristics associated with human speech. However, those affective cues are not prominent in LU clips. As our position states that affective cues (duration, pitch, loudness) are associated with speech uncertainty levels in robot navigation tasks, we believe the setup of our dataset is comprehensive enough to show that

vocal cues are good predictors. For review purposes, we create an anonymous GitHub repo: `https://anonymous.4open.science/r/DNIA-dataset-D22E`. The full dataset will be publicly released soon after the review as we prepare them for storage and access control. The dataset will come with textual annotations (grounded labels) as well as Whisper-generated transcription to save resources for other researchers who can reproduce the result.

F Pitch and Loudness Shifts Detection

Algorithm 1 detects the largest shifts in pitch and loudness in a given audio clip. We use the timestamp to align with the sequence of instructions for vocal cue prompts.

Algorithm 1 Loudness and Pitch Detection

Require: audio $\mathcal{V}$, θ_l (loudness threshold), θ_p (pitch threshold)
Ensure: Timestamps of maximum loudness change (t_l) and pitch shift (t_p)

```
L ← ∅, P ← ∅
l_prev, p_prev ← loudness and pitch at [t_0, t_1]
for each second t_i in V do
    Define segment as [t_i, t_{i+1}]
    l_{t_{i+1}} ← avg_loudness(segment)
    p_{t_{i+1}} ← calculate_pitch(segment)
    if |l_{t_{i+1}} − l_prev| > θ_l then
        Append (t_{i+1}, |l_{t_{i+1}} − l_prev|) to L
    end if
    if |p_{t_{i+1}} − p_prev| > θ_p then
        Append (t_{i+1}, |p_{t_{i+1}} − p_prev|) to P
    end if
    l_prev ← l_{t_{i+1}}, p_prev ← p_{t_{i+1}}
end for
t_l ← argmax_t {δ_{l_t} | (t, δ_{l_t}) ∈ L}
t_p ← argmax_t {δ_{p_t} | (t, δ_{p_t}) ∈ P}
```

G User Interface for Human Evaluation on DNIA Dataset

This section details the user interface that we used for human assessment. As shown in Figure 7, the website allows users to log in and participate in the human evaluation session. The user is asked to listen to a randomly selected audio clip from the audio dataset and answer the question by choosing the most accurate and efficient answer from a list of five options. Once the user select the answer, they need to hit *Confirm* button to submit the answer the proceed to the next question.

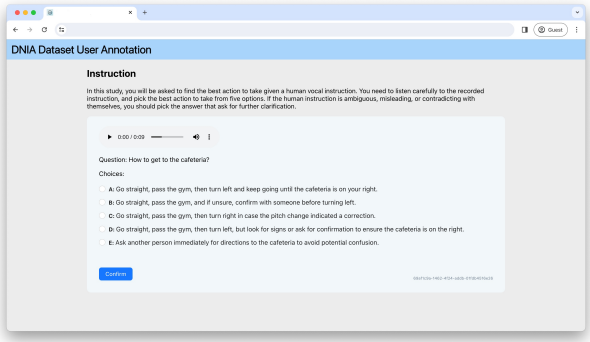

Figure 7: Screen capture of our DNIA dataset human annotation website.