

TrustNavGPT: Modeling Uncertainty to Improve Trustworthiness of Audio-Guided LLM-Based Robot Navigation

Xingpeng Sun¹, Yiran Zhang¹, Xindi Tang¹, Amrit Singh Bedi², Aniket Bera¹

Abstract—Large language models (LLMs) exhibit a wide range of promising capabilities – from step-by-step planning to commonsense reasoning –that provide utility for robot navigation. However, as humans communicate with robots in the real world, ambiguity and uncertainty may be embedded inside spoken instructions. While LLMs are proficient at processing text in human conversations, they often encounter difficulties with the nuances of verbal instructions and, thus, remain prone to hallucinate trust in human command. In this work, we present *TrustNavGPT*, an LLM-based audio-guided navigation agent that uses affective cues in spoken communication—elements such as tone and inflection that convey meaning beyond words—allowing it to assess the trustworthiness of human commands and make effective, safe decisions. Experiments across a variety of simulation and real-world setups show a 70.46% success rate in catching command uncertainty and an 80% success rate in finding the target, 48.30%, and 55% outperform existing LLM-based navigation methods, respectively. Additionally, *TrustNavGPT* shows remarkable resilience against adversarial attacks, highlighted by a 22%+ less decrease ratio than the existing LLM navigation method in success rate. Our approach provides a lightweight yet effective approach that extends existing LLMs to model audio vocal features embedded in the voice command and model uncertainty for safe robotic navigation. For more information, visit the TrustNav project page.

I. INTRODUCTION

Recent advances in Large Language Models (LLMs), such as GPT-4 [1] or Gemini [2], and Robotics have shown significant improvement for human-robot interactions (HRI) areas such as task planning [3], [4], [5], or social navigation [6], [7], [8]. It is crucial for robots to emulate how humans interact and form opinions about each other, including assessments of credibility and trust, and understand human uncertainty to ensure safe and efficient actions [9], [10]. When humans interact, they subconsciously form opinions about one another, including judgments about vocal credibility and trust. These perceptions impact their decision-making in collaborative settings. For instance, imagine a scenario in which two individuals, both unfamiliar with a theme park, are interacting. One asks for directions to an attraction entrance, and the responder, uncertain of the way, provides unclear instructions. The inquirer, drawing on extensive experiential knowledge, can discern the uncertainty not only from the words but also from the hesitant vocal nuances. Consequently, they choose not to rely solely on this dubious guidance. In contrast, a robot lacking this nuanced reasoning capability

¹ Purdue University, West Lafayette, IN 47907, USA sun1223, zhan5058, tang666, aniketbera@purdue.edu

² University of Central Florida, Orlando, FL 32816, USA amritbedi@ucf.edu

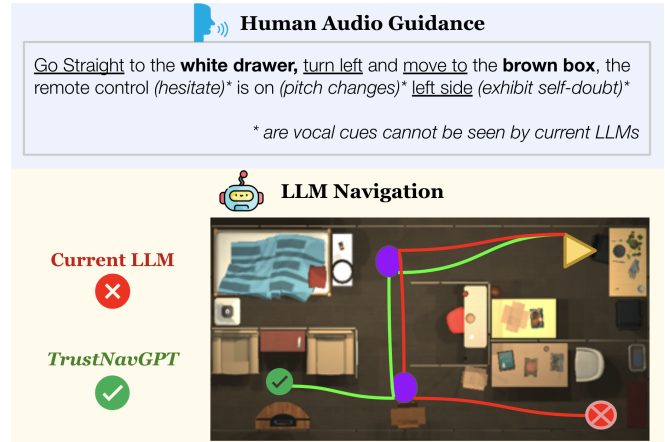


Fig. 1. The current navigation methods using Large Language Models (LLMs) struggle with making accurate decisions when faced with ambiguous audio instructions. Our strategy involves affective cues from spoken communication into LLMs, enabling them to evaluate the reliability of human instructions from the semantic and vocal uncertainty, thus allowing for safe and successful navigation.

would follow the instructions without question, potentially resulting in failure to reach the intended destination.

To model human uncertainty, KnowNo [5] proposes an LLM-based planner and asks humans for clarification when needed, but it is only built on analyzing semantic uncertainty. However, in intricate settings such as theme parks, due to unfamiliarity with the space and spatial anxiety [11], humans' guidance can be vague or uncertain, affected not just by the choice of words but also by the subtleties in their voice [12], [13]. Current LLMs [1], [2] provide capabilities for converting speech to text, but this process often omits important vocal characteristics, leading to a significant loss of information that could indicate uncertainty. This gap in capturing vocal nuances limits the LLMs' capacity to accurately judge the reliability of voice-based commands and successfully navigate to the target, underscoring the necessity for advancements that can interpret and leverage these vocal cues in the realm of human-robot cooperation.

Taking a step towards more human-like social navigation, we propose *TrustNavGPT*, a cognitive agent empowered by LLMs. The fundamental insight of our approach lies in the integration of both audio transcription and affective vocal features, including pitch, loudness, and speech rate, to improve robot ability in audio-guided navigation under uncertainty. Moreover, as LLM is good at high-level task planning but not good at low-level control and motion planning, we also propose a tool library that gives the LLM decision-making engine the ability to control the robot based on planning and

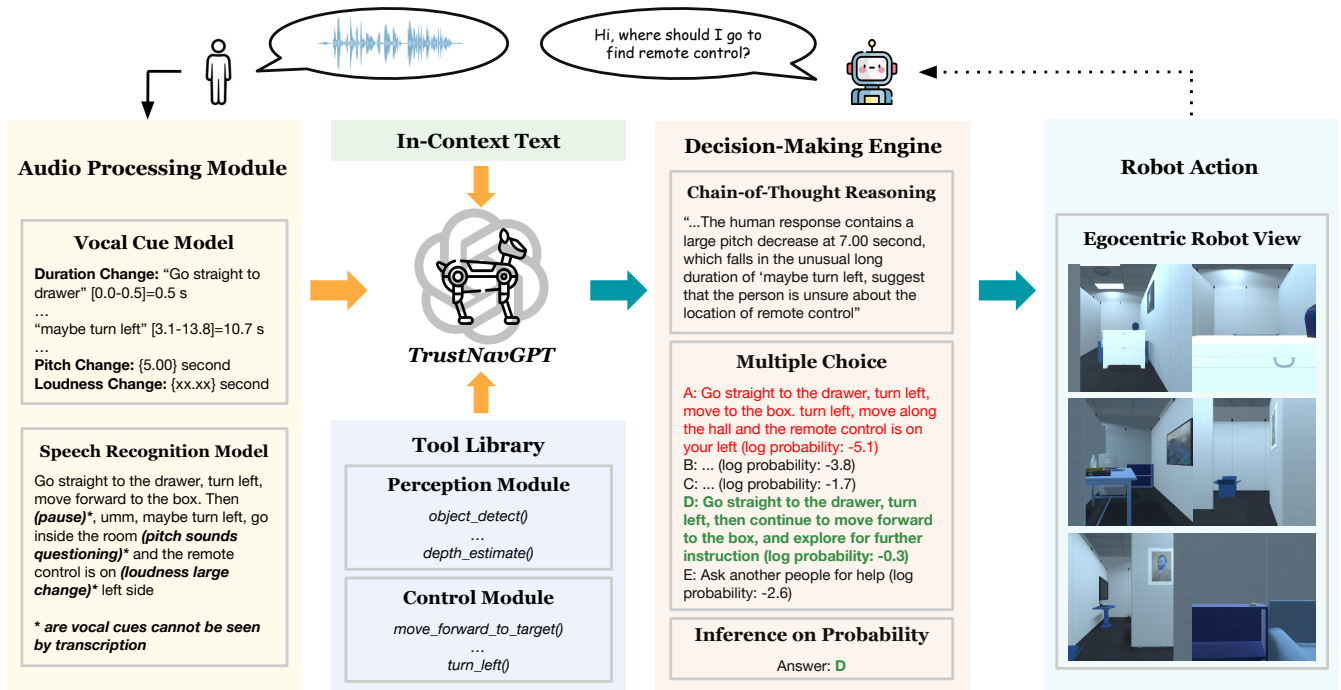


Fig. 2. Overview: Human audio goes through an audio-processing module that transcribes it, while a vocal cue model identifies three essential affective cues. We then prompt a language model to generate five possible next-step actions, selecting the choice based on the next token logit probability. Notably, semantic transcription alone leads to the red choice, but incorporating the vocal cue results in the green choice being selected. Finally, a tool library translates the chosen language instruction into agent actions for navigation.

visual perception. When the confidence of human guidance is reasoned as low, the robot undertakes scene exploration by categorizing objects in each direction and formulating conjectures based on each group (e.g. when the robot is tasked to find the microwave and sees a dishwasher on the left, it will infer left as kitchen and explore left side for target, instead of following human’s uncertain command to “turn right”). These components, coordinated by LLMs, create an automatic robotic system that navigates human audio uncertainty. Our main contributions are summarized as follows:

- 1) Our work introduces a layer of interpretation by examining not just the content of human speech but also the manner in which it is conveyed. The *TrustNavGPT* approach significantly refines LLMs’ proficiency in interpreting human uncertainty within navigational contexts, evidenced by achieving over an 80% success rate in robot navigation tasks. This integration allows for a more nuanced synthesis of information, paving the way for intelligent conversational systems, to better understand and act upon ambiguous human instructions.
- 2) Integrate a motion planning tool library that translates high-level LLM language commands into “robot actions, dynamic perception, and prediction, which can be accessed through function calls, to facilitate a human-like, audio-guided navigational capability in robots.
- 3) We conduct experiments on a large-scale Disfluent Navigational Instruction Audio Dataset [14], RoboTHOR simulation environment [15], and also real-world setup,

to show that *TrustNavGPT* significantly surpasses existing LLM-based navigation techniques, by a 55% improvement in achieving successful target arrival under conditions of human navigational uncertainty with 70%+ closer to the target, indicating a substantial enhancement in navigational efficiency and precision. Detailed ablation studies on heterogeneous parts of our architecture are also provided, pointing to areas for future works.

II. RELATED WORK

A. Large Language Model for Robotic Navigation

With remarkable proficiency in commonsense reasoning and planning, Large Language Models (LLMs) have been utilized for navigation-related contexts. Recent scholarly work has explored the integration of LLMs with visual inputs to map landmarks and subgoals mentioned in navigational commands [7], [16], the application of LLMs in facilitating sequential decision-making for zero-shot robot navigation [17], [8], and also the investigation of LLMs for the semantic prediction of object locations, thereby enhancing navigational efficiency [6]. Despite these advancements, including NavGPT [8], the current body of research predominantly considers only textual instruction and assumes the reliability of human input commands, overlooking scenarios where such instructions might be ambiguous or incorrect. Our study distinctively addresses this gap by evaluating human uncertainty through the analysis of both textual and vocal emotions in audio-based navigation instructions.

B. Large Language Model Agent

Inspired by strong emergent capabilities of LLMs, such as zero-shot prompting and complex reasoning, LLM agent, a system with complex reasoning capabilities, planning skills, and the means to execute tasks, becomes popular [18]. Voyager [19] is an LLM-embodied gaming agent that plays Minecraft without human intervention through lifelong learning. Agent Driver [20] and Inner Monologue [21] integrate LLM into autonomous driving systems and robot planning by incorporating environment feedback and making the LLM able to execute action through a versatile function library. However, to the best of our knowledge, current LLM agent works do not take into account the affective emotion of human command, especially extracting vocal uncertainty from speech in audio-guided navigation scenarios.

C. Uncertainty Quantification for Large Language Model

A growing body of research investigated quantifying uncertainty due to LLM’s hallucinations [22], [23]. Entropy has been introduced as a method to model uncertainty in the large language model [24], [25], while conformal prediction [26] is another method applied to quantify uncertainty for next-token prediction in Multiple Choice Question Answering(MCQA) setups [27], [5]. In our approach, we take advantage of these works and define a confidence score $\mathcal{C}(\rho)$ inspired by entropy, which builds on the MCQA setup and shows effectiveness in gauging the LLM confidence.

D. Affective Analysis in Social 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 [28], while deep reinforcement learning methods [29] and cognition model [30] has been used to understand textual ambiguities in natural languages. While, for speech, not only does textual transcription offer insights, but vocal cues also hold substantial information that can reveal human emotions and ambiguities. However, there has been limited research on the role of vocal nuances in audio-guided social navigation. *TrustNavGPT* proposes an LLM agent that analyzes both textual and vocal affective cues embedded within human audio commands, creating a more human-like robot system for social navigation.

III. METHODOLOGY

A. Problem Formulation

In this section, we mathematically formalize the navigation problem towards a designated target location τ under uncertainties within navigational instruction, utilizing Large Language Models (LLMs). In the depicted scenario (Figure 2), a robot seeks navigational commands from a human, communicated through auditory means. This framework is adapted by seminal works in robot social navigation [30], [31], [32], [33]. Upon the robot’s inquiry, a human articulates an auditory instruction $v \in \mathcal{V}$, where \mathcal{V} represents the ensemble of auditory commands. This vocal input is transcribed into a

textual format \mathcal{W} through a pre-trained transcription model $\mathcal{T} : \mathcal{V} \rightarrow \mathcal{W}$, with \mathcal{W} embodying the set of all feasible textual instructions. 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:

$$\mathcal{P}(\mathcal{V}) = \mathcal{W} \oplus \mathcal{K}, \quad (1)$$

which constitutes the prompt for the LLM. The LLM (denoted as F_{LLM}) thus elucidates a response planning sequence S conditioned on chain-of-thought reasoning D :

$$S = \{s_1, s_2, \dots, s_k\} = F_{LLM}(\mathcal{P}(\mathcal{V})|D), \quad (2)$$

where each intermediate action step s_i is generated sequentially. We define the joint probability distribution of generating the sequence S from $\mathcal{P}(\mathcal{V})$ as:

$$P_{\Theta}(S|\mathcal{P}(\mathcal{V})) = \prod_{i=1}^k P_{\Theta}(s_i|s_1, s_2, \dots, s_{i-1}), \quad (3)$$

where Θ denotes the parameter set of the LM. In the case at time k the response s_k is ambiguous, the robot leverages help from its decision-making engine (details in section III-C) based on the visual exploration of the current state environment M and thus inference target location from the surrounding objects. The high-level planning sequence is translated into low-level executable commands $\mathcal{A} = \{\alpha^1, \alpha^2, \dots, \alpha^k\}$, where $\alpha^i = \varphi(s_i)$ based on a tool library φ (details in section III-D). The objective is to successfully navigate to target τ , represented by:

$$\begin{cases} \max P(S', \tau, M) \\ \min \|R(\mathcal{A}, M) - \tau\|, \end{cases} \quad (4)$$

to maximize success rate for a robot R to arrive at target τ in environment M utilizing a LLM-reasoned step sequence S' with minimum distance to target τ after applying the execution sequence \mathcal{A} . Detailed overview is shown in Fig. 2.

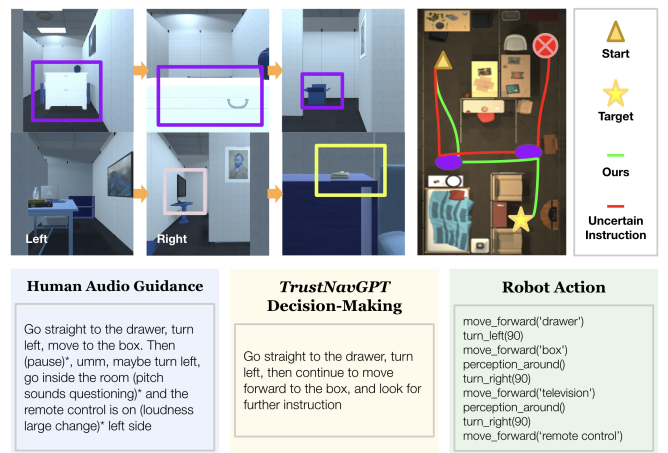


Fig. 3. Illustration of action sequences. The purple box shows the reference object. At the point the human is ambiguous, the robot sees a television on the right-hand side (pick box), and thus reasons that the television is near to the remote control, then moves to the right side instead of following the human instruction. Notably, without uncertainty analysis, the LLM navigation path is shown in red, leading in the wrong direction. The navigation result of our method is shown in green, arriving at the target (yellow box) successfully.

B. Audio Command Uncertainty Modeling

In this section, we introduce the process of quantifying semantic and vocal uncertainty embedded in audio input.

Semantic Uncertainty: With a human navigational guidance audio, we convert it into text using the open-source Whisper model [34]. Our analysis focuses on speech disfluencies as indicators of uncertainty, impacting the confidence measures of Large Language Models (LLMs). We identify three types of language uncertainty: *Ambiguous Word Choice*, *Speech Repair*, and *Hesitation Signs*.

Ambiguous Word Choice encompasses phrases like “probably,” “maybe,” “might,” and “I assume,” indicating hesitancy or a lack of conviction [29]. Speech Repair involves self-correction instances, such as “Take a left turn, no no no, I mean take a right turn,” reflecting errors in thought expression and potentially leading to confusion [35]. Hesitation Signs are pauses in speech, evident in expressions like “Err, turn umm left,” signaling uncertainty about forthcoming directions [36]. These patterns are prevalent in human communication when expressing doubt. We instruct LLMs to detect these cues, allowing for an adjustment in the confidence level regarding the reliability of instructions following these indicators.

Vocal Uncertainty: Beyond textual analysis, the prosody of spoken instructions—specifically *Pitch*, *Loudness*, and *Speech Rate*—also serves as a marker of human uncertainty. Pitch is the rising intonation observed at the conclusion of phrases or sentences, resembling a questioning tone rather than an assertive statement [37], [38]. Loudness is shown as fluctuations in volume levels, as it can suggest wavering of confidence [37], [39]. In terms of speech rate, people might either slow down as they ponder their words or accelerate their speaking speed due to nervousness [40], [39]. Variations in speech rate serve as vocal markers reflecting changes in certainty levels [38]. Our method measures the speech rate of each instruction segment within the audio and assesses whether notable duration has occurred. For instance, in a recording where each instruction phrase typically spans around one second, an elongated phrase extending for more than 6 seconds may signify hesitation and reduced confidence.

Algorithm 1 Vocal Uncertainty Modeling

Require: audio v , loudness threshold θ_l , pitch threshold θ_p

Ensure: Timestamps of maximum loudness change (t_l) and pitch shift (t_p), speech rate of each instruction segment r_s

$$t_l \leftarrow \operatorname{argmax}_t \{ \delta_{l_t} \mid (t, \delta_{l_t}) \in v, \forall \delta_{l_t} \geq \theta_l \}$$

$$t_p \leftarrow \operatorname{argmax}_t \{ \delta_{p_t} \mid (t, \delta_{p_t}) \in v, \forall \delta_{p_t} \geq \theta_p \}$$

$$r_s \leftarrow \operatorname{time}(s_i), \forall s_i \in S$$

As detailed in Algorithm1, we detect the max change of loudness and pitch features δ_{l_t} , δ_{p_t} in the audio clip. For the speech rate, we measure the time of each sub-instruction s_i . We then use force alignment, a technique that aligns text fragments with δ_{l_t} , δ_{p_t} , r_s , to synchronize the timestamps of vocal features with the literal instructions.

C. Decision-Making Engine

The decision-making engine takes audio transcription and vocal analysis information as inputs, performs MCQA task planning, and scene direction conjecture, and eventually generates execution commands to navigate the robot to the target based on the potential ambiguous audio guidance.

MCQA task planning: To generate possible next steps based on $\mathcal{P}(\mathcal{V})$, the prompt consists both textual and vocal cue information of human audio command, we use Chain-of-Thought reasoning [41], a prompting mechanism aims to emulate the human reasoning process, together with In-Context Learning [42] by providing only few-shot examples with no explicit training. The process can be formulated as:

$$\mathcal{D} = F_{LLM}(\mathcal{P}(\mathcal{V}), E), \quad (5)$$

For in-context examples E , we fine-grained three examples that cover all human interaction scenarios: 1) textual uncertainty, 2) vocal uncertainty, and 3) both textual and vocal uncertainty. We ask the LLM to first identify potential uncertainty signals and then assess how this notable uncertainty will influence the subsequent decision-making process.

In the answering phase, the model reason to suggest a spectrum of five potential actions follows the Equation 2 in MCQA settings. Particularly, the first option (A) is a direct paraphrase of the transcription, excluding any uncertainty, representing the robot’s choice to unconditionally accept the human audio-guided instruction. Options (B), (C), and (D) reflect various actions acknowledging uncertainty, while the last option (E) is always “ask another person nearby for direction”, providing a reliable fallback in the decision-making process. Then, we predict the next-token log probability, the most commonly-used pre-training objective for causal language models, for the set of options $Y = \{‘A’, ‘B’, ‘C’, ‘D’, ‘E’\}$ to select the label with the highest probability as the optimal planning sequence S .

Visual Scene Direction Conjecture: Given optimal planning sequence $S = \{s_1, s_2, \dots, s_k\}$, if the audio v sounds uncertain, then s_k is always an ambiguous action, like “look for more information at this location to plan navigation to target”, due to learning in-context examples. We employ the semantic knowledge embedded in language models in a tailored manner, utilizing it not just as a heuristic for search [6], but also as a way to visual grounding the trust of human guidance. Our decision-making engine deduces k th action based on the visual exploration of the current state environment M denoted as $\Lambda(s_k, M)$. Specifically, the robot will segment M into left, right, front, and three directions and categorize objects in each direction to hypotheses about their locations relative to the target. As illustrated in Figure 3, when the target is a “Remote Control” and the robot, upon detecting a “Television” to the right at a point of ambiguous human instruction, it deduces that the “Television” is likely near the “Remote Control”. This inference leads it to prioritize its own decision-making directions s'_k over less reliable human directions s_k . On the other hand, if a limited number of objects can be detected in M , the decision-making

engine will resort to any other supervisor agent, such as asking help from a human, denoted as Γ , to yield a clarified action $S' = \{s_1, s_2, \dots, s'_k\}$, where $s'_k = \Lambda(s_k, M) \oplus \Gamma(s_k)$. *TrustNavGPT*, therefore, leverages vision to better infer the target location, minimizing the attempt to ask humans and mimic a human-like social navigation process.

D. Tool Library

While Large Language Models (LLMs) have demonstrated impressive proficiency in high-level task planning [43], [6], bridging the divide from strategic planning to practical execution remains a significant hurdle. Recent studies [44] have explored the use of foundation models to directly generate executable action codes from linguistic instructions. However, this approach often encounters limitations in the speed of translating instructions to actions, and the resultant code accuracy is not assured. To address these issues and enhance the precision of action codes, we follow the idea of [19], which creates a skill library to store and retrieve behaviors for gaming agents and introduces a customized robot navigation tool library. This library comprises a collection of functions specifically tailored to parse environmental data based on textual instructions and decompose complex language into robot-executable actions through dynamic function calls. This framework establishes a robust loop encompassing perception, planning, and action, thereby facilitating robots to execute tasks more reliably and efficiently.

Tool Functions. We developed functions for dynamic perception and control to enhance our system’s interaction with its environment. For the perception module, we implemented object detection and depth estimation using off-the-shelf pre-trained models, enabling the system to dynamically detect the object and provide information for robot action planning. For the control module, we employ few-shot learning techniques, enabling the language model to convert textual navigation instructions into pairs of actions and locations through function calls. This approach is operated through custom functions such as `move_forward_to_target()`, `turn_left()`, which guide the robot’s movement through each step of the instruction sequence. Fig.2&3 illustrate this integrated process.

IV. EXPERIMENTS AND RESULTS

In this section, we demonstrate the effectiveness, few-shot learning, and characteristics of *TrustNavGPT* through extensive experiments on Disfluent Navigational Instruction Audio Dataset (DNIA) [14], RoboTHOR simulation environment [15], and real-world scenarios. First, we introduce the evaluation metrics and setup; then, we compare our methods with other LLM-based navigation methods in simulation environments. Finally, we conduct ablation studies to investigate the reasoning ability, perception ability, and effectiveness of each vocal cue in analyzing human audio uncertainty. Finally, we show robustness to LLM adversarial attacks. To be specific about the evaluation conditions, Tables I, II, and IV present results under real-world setups, as our audio clips are recorded

by humans in a quiet environment. Table III presents results from the RoboTHOR simulation environment.

A. Uncertainty Detection

To demonstrate our agent’s capability in detecting human uncertainty, we utilize the DNIA dataset as detailed by [14]. DNIA dataset [14] encompasses a range of navigational disfluencies, comprising 500 audio clips divided into two categories: language uncertainty (LU) with 285 clips, and vocal tone uncertainty (VU) with 215 clips. LU clips feature semantic disfluencies, such as hesitations and language uncertainties, whereas VU clips include instances where the vocal tone, rather than the textual content, indicates uncertainty. For example, “*Go straight to the drawer, turn left and move to the garbage can, the vase (hesitate) is on (pitch changes) your left*” is a VU command, and “*Go straight to the drawer, turn left and move to the garbage can, the vase maybe is on your left*” is a LU command. Each clip is labeled with a user-study annotation (human annotation is a suggested method for LLM-based HRI evaluation [45], [46]), which chooses the best choices from a set of five multiple-choice options designed to represent the uncertainty inherent in the audio. We calculate the **Prompt Selection Success Rate(PSSR)** and the **Confidence Score(CS)** for evaluation. We defined PSSR as:

$$PSSR = \frac{p_{succ}}{p_{total}} \quad (6)$$

where p_{succ} denotes the number of instances where the LLM chooses a correct next-step action that aligns with the user study annotation, and p_{total} denotes the total number of audio clips that prompt the LLM.

Inspired from prior works[24], [25] which use entropy as a method to quantify uncertainty in large language models, we define confidence score $\mathcal{C}(\rho)$ based on the model’s p_θ probability distribution over the potential candidates $\{y^j\}_{j=1}^J$ against the ground-truth response distribution ρ^* as a dirac-delta over the true response as $\rho^* = [0, 1 \dots 0]$, assuming the second candidate to be the optimal y^{j^*} with $j^* = 2$ without loss of generality. This confidence score is inversely proportional to the Kullback–Leibler(KL) divergence between ρ and ρ^* , as shown in the equation:

$$\mathcal{C}(\rho) = \frac{1}{\text{KL}(\rho, \rho^*)}, \quad (7)$$

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. Together with PSSR and CS, we show low bias and low variance in our uncertainty measurement.

Table I demonstrates that our method exhibits reduced bias and reduced variance, 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 PSSR surged to 70.46%, with a pronounced improvement in VU interpretation, evident in a

TABLE I
UNCERTAINTY MEASUREMENT BY CONTEXT AND AUDIO TYPE

Method	Metric	Audio Category		
		All	VU	LU
Text-based LLM	PSSR	22.16%	22.79%	21.75%
	CS	0.7782	0.7656	0.7884
With CoT	PSSR	49.30%	36.74%	58.60%
	CS	0.9545	0.9553	0.9549
Ours	PSSR	70.46%	72.56%	68.77%
	CS	1.1354	1.1189	1.1445

TABLE II
ABLATION STUDY FOR VARIOUS VOCAL CUES

Vocal Cue			PSSR			CS
Pitch	Loudness	Speech Rate	All	VU	LU	All
✘	✘	✘	49.30	36.74	58.60	0.4965
✓	✘	✘	61.68	64.65	59.30	0.8322
✘	✓	✘	63.07	61.40	64.21	0.9683
✘	✘	✓	60.88	61.40	60.35	0.8467
✓	✓	✘	67.47	70.23	65.26	1.1282
✓	✘	✓	64.87	66.51	63.51	0.9076
✘	✓	✓	65.47	67.91	63.51	1.0115
✓	✓	✓	70.46	72.56	68.77	0.8227

72.56% PSSR. The overall confidence score surged to 1.135, improved by 45.8% in comparison to existing LLM methods.

To elucidate the impact of individual components, we conduct a comprehensive ablation study on each category of vocal cues and report in Table II. Each of the Pitch, loudness, and speech rate features can significantly enhance LLM’s ability to discern vocal uncertainty, increasing the PSSR by 20%+ and CS by 127%. Regarding CS, the inclusion of vocal features consistently outperforms scenarios lacking these features and also surpasses those without reasoning. The combination of pitch and loudness features achieves the highest confidence score. This suggests that while the presence of multiple vocal features across different time frames may slightly diminish the model’s confidence, it does not affect its accuracy in selecting the correct responses.

B. Simulation Environment Robot Navigation

To evaluate the effectiveness of *TrustNavGPT* in navigation, we adopt the LoCoBot and extensively test the audio-guided navigation performance in 10 different RoboTHOR indoor environments. For each test, we provide a piece of audio instruction with either semantic or vocal uncertainty that navigates the LoCoBot to a unique target instance in the environment. We evaluate 5 common robot navigation metrics: **Success Rate (SR)**: The LoCoBot successfully found the target within its vision distance. The higher SR is better.

Steps: the number of robot movement actions.

Path Distance: the explore path length that takes the LoCoBot to find the target if it succeeds or execute navigation events if it fails.

Distance to Target: The shortest path distance from the LoCoBot’s final position to the target position. The smaller this metric is, the more successful the navigation method is.

Success weighted by Path Length (SPL):

$$SPL = \frac{1}{N} \sum_{i=1}^N \frac{S_i}{\max(p_i, l_i)}, \quad (8)$$

SPL ranges from [0,1], where N is the total number of evaluated tasks, $S_i \in \{0, 1\}$ is the binary indicator of success, l_i denotes the ground truth shortest path length, and p_i denotes the actual path length of the agent in navigation. This metric indicates the efficiency of the actual path compared to the ground truth shortest path when the navigation task is successfully completed.

Note in TableIII, our method significantly outperforms the random method and existing LLM methods, achieving 80% for SR, an overall 0.69-meter distance to target, and 33.76% in terms of SPL. We also run a comprehensive ablation study on each section of *TrustNavGPT* architecture, as illustrated in the last three columns in TableIII. With no vocal cue analysis, just using few-shot in-context learning to teach LLM to detect semantic uncertainty results in a low 27.5% success rate. The highest SPL of 34.81% is achieved using *TrustNavGPT* without a perception module due to the fact that the perception toolbox will lead robots to explore the environment and navigate to the related objects based on inference, thus resulting in longer total path distance and then lower SPL. The highest success rate and nearest distance to the target are observed using a combination of the perception module and vocal cue module.

C. Robustness toward Adversarial Language Model Attacks

With the progression of language models, concerns like adversarial attacks and prompt manipulation have gained prominence [47], [48]. These attacks often involve simple token operations such as synonym replacement and misleading models into making errors [49], [50].

We present the *TrustNavGPT* resistance to such adversarial tactics aimed at LLMs. Our attack involves initially paraphrasing a given transcript (T_1) into a new form (T_2), where uncertain terms are all swapped for more deterministic phrases. (T_2) is then used to realign both vocal and textual prompts and replicate as one option for LLM to choose. This experiment underscores the current LLMs’ dependency on textual semantics, overlooking the subtleties embedded in vocal expression. We compare the result between adding in-context examples of detecting textual uncertainty within navigation command to existing LLM navigation methods against our method and show the result in TableIV. After the token attack, our approach exhibited a notably lower reduction of 33.43% in PSSR, 0.22 in distance to target, and 7.55% in SPL, all notably smaller than that of the existing LLM navigation method. This suggests that audio augmentation in our approach enables LLMs to resist text-based adversarial attacks and maintain safe capabilities for robot navigation.

D. Real-World Exploration

TrustNavGPT underwent rigorous testing within real-world scenarios, employing YOLOv8 [51] for object detection, the Tesseract Open Source Engine [52] for letter/word detection,

TABLE III
ROBOTHOR NAVIGATION RESULT & ABLATION RESULT ON DIFFERENT MODULE

Method	SR		Steps		Path Distance		Distance to Target		SPL	
	LU	VU	LU	VU	LU	VU	LU	VU	LU	VU
Random Search	25%	25%	534.6	534.6	31.62	31.62	1.56	1.56	8.88%	8.88%
LM-Nav [7]	25%	25%	8.5	8.5	5.30	5.30	2.35	2.35	9.57%	9.57%
<i>TrustNavGPT w/o Vocal (Ours)</i>	25%	50%	8.5	9.0	4.53	5.99	2.55	2.39	13.46%	30.24%
<i>TrustNavGPT w/o Vision (Ours)</i>	75%	50%	10.0	13.5	6.42	6.79	1.84	1.75	33.30%	36.32%
<i>TrustNavGPT (Ours)</i>	80%	80%	13.2	14	10.24	9.94	0.56	0.82	32.05%	35.47%

TABLE IV
ROBUSTNESS TO LLM ATTACK

Method	PSSR	Distance to Target	SPL
LM-Nav[7] + Few-Shot	22.16%	2.47	21.85%
After Token Attack	9.78%	2.72	19.24%
Decrease	55.87%	0.25	13.57%
<i>TrustNavGPT (Ours)</i>	70.46%	0.69	33.76%
After Token Attack	46.90%	0.91	31.39%
Decrease	33.43%	0.22	7.55%



Fig. 4. Real-world Navigation with vocal direction to Starbucks Coffee Shop. Successfully arrived at the target.

and MiDas [53] for generating depth maps. At each time stamp, an image is captured and subsequently analyzed by YOLOv8 and Tesseract to ascertain the presence of target objects/words within the scene. If the target is not detected, the robot proceeds forward; otherwise, upon detection, the depth map for the identified target object is computed to determine if the robot should proceed forward or do the next turning task. If the object is close enough, a turning action is executed. A real-world demonstration illustrated in the accompanying Figure 4 shows how the robot does the mission of detected verbal instructions: “walking straight until you see the traffic light; you wanna then turn left. Then, when you see a Comfort Suite, consider executing a uhhh...maybe a left turn, yea, you’ll see a Starbucks drive-thru sign. Continue straight along this path; you’ll see Starbucks coffee shop.” Notably, uncertainty arose regarding the direction of the second left turn instruction, prompting the robot to analyze its surroundings by checking forward, left, and right perspectives to make sure it could arrive at the destination as expected. In this instance, the uncertain human instruction was rectified as the robot identified a drive-thru letter sign on its right-hand side, prompting a refined trajectory adjustment to turn right.

V. CONCLUSION

In our work, we present an LLM trust-driven audio-guided robot navigation agent *TrustNavGPT*, which effectively deals with potential uncertainty within human audio commands. Our findings highlight the improved planning, efficiency, and resilience achieved by integrating affective audio processing with large language models (LLMs) to improve navigation in social robots. As the integration of vocal and semantic analysis increases the computational overhead, which may limit the deployment in low-resource settings or in real-time applications; and system’s performance relies on the quality of audio input, Future works can include development of denoisy methods and more intelligent retrieval-augmented generation to improve the reliability and computation efficiency. We believe our work will encourage further exploration into aligning uncertainties with LLMs for the development of audio-directed robots.

REFERENCES

- [1] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, “Gpt-4 technical report,” *arXiv preprint arXiv:2303.08774*, 2023.
- [2] G. Team, R. Anil, S. Borgeaud, Y. Wu, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth *et al.*, “Gemini: a family of highly capable multimodal models,” *arXiv preprint arXiv:2312.11805*, 2023.
- [3] B. Liu, Y. Jiang, X. Zhang, Q. Liu, S. Zhang, J. Biswas, and P. Stone, “Llm+ p: Empowering large language models with optimal planning proficiency,” *arXiv preprint arXiv:2304.11477*, 2023.
- [4] B. Li, P. Wu, P. Abbeel, and J. Malik, “Interactive task planning with language models,” 2023.
- [5] A. Z. Ren, A. Dixit, A. Bodrova, S. Singh, S. Tu, N. Brown, P. Xu, L. Takayama, F. Xia, J. 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.
- [6] D. Shah, M. R. Equi, B. Osinski, F. Xia, B. Ichter, and S. Levine, “Navigation with large language models: Semantic guesswork as a heuristic for planning,” in *Conference on Robot Learning*. PMLR, 2023, pp. 2683–2699.
- [7] D. Shah, B. Osinski, B. Ichter, and S. Levine, “LM-nav: Robotic navigation with large pre-trained models of language, vision, and action,” in *6th Annual Conference on Robot Learning*, 2022.
- [8] G. Zhou, Y. Hong, and Q. Wu, “Navgpt: Explicit reasoning in vision-and-language navigation with large language models,” *arXiv preprint arXiv:2305.16986*, 2023.
- [9] M. Lewis, K. Sycara, and P. Walker, “The role of trust in human-robot interaction,” *Foundations of trusted autonomy*, pp. 135–159, 2018.
- [10] K. E. Schaefer, “Measuring trust in human robot interactions: Development of the “trust perception scale-hri”,” in *Robust intelligence and trust in autonomous systems*. Springer, 2016, pp. 191–218.
- [11] E. Chan, O. Baumann, M. A. Bellgrove, and J. B. Mattingley, “From objects to landmarks: the function of visual location information in spatial navigation,” *Frontiers in psychology*, vol. 3, p. 304, 2012.

- [12] J. L. Prestopnik and B. Roskos-Ewoldsen, "The relations among wayfinding strategy use, sense of direction, sex, familiarity, and wayfinding ability," *Journal of environmental psychology*, vol. 20, no. 2, pp. 177–191, 2000.
- [13] R. G. Golledge, "Human wayfinding and cognitive maps," *Colonization of unfamiliar landscapes: the archaeology of adaptation*, vol. 25, 2003.
- [14] X. Sun, H. Meng, S. Chakraborty, A. S. Bedi, and A. Bera, "Beyond text: Improving llm's decision making for robot navigation via vocal cues," *arXiv preprint arXiv:2402.03494*, 2024.
- [15] M. Deitke, W. Han, A. Herrasti, A. Kembhavi, E. Kolve, R. Mottaghi, J. Salvador, D. Schwenk, E. VanderBilt, M. Wallingford, L. Weihs, M. Yatskar, and A. Farhadi, "Robothor: An open simulation-to-real embodied ai platform," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
- [16] B. Yu, H. Kasaei, and M. Cao, "L3mvm: Leveraging large language models for visual target navigation," *arXiv preprint arXiv:2304.05501*, 2023.
- [17] V. S. Dorbala, J. F. Mullen Jr, and D. Manocha, "Can an embodied agent find your "cat-shaped mug"? llm-based zero-shot object navigation," *IEEE Robotics and Automation Letters*, 2023.
- [18] L. Wang, C. Ma, X. Feng, Z. Zhang, H. Yang, J. Zhang, Z. Chen, J. Tang, X. Chen, Y. Lin *et al.*, "A survey on large language model based autonomous agents," *arXiv preprint arXiv:2308.11432*, 2023.
- [19] G. Wang, Y. Xie, Y. Jiang, A. Mandlekar, C. Xiao, Y. Zhu, L. Fan, and A. Anandkumar, "Voyager: An open-ended embodied agent with large language models," *arXiv preprint arXiv: Arxiv-2305.16291*, 2023.
- [20] J. Mao, J. Ye, Y. Qian, M. Pavone, and Y. Wang, "A language agent for autonomous driving," *arXiv preprint arXiv:2311.10813*, 2023.
- [21] W. Huang, F. Xia, T. Xiao, H. Chan, J. Liang, P. Florence, A. Zeng, J. Tompson, I. Mordatch, Y. Chebotar *et al.*, "Inner monologue: Embodied reasoning through planning with language models," *arXiv preprint arXiv:2207.05608*, 2022.
- [22] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin *et al.*, "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions," *arXiv preprint arXiv:2311.05232*, 2023.
- [23] Y. Xiao and W. Y. Wang, "Quantifying uncertainties in natural language processing tasks," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 7322–7329.
- [24] L. Kuhn, Y. Gal, and S. Farquhar, "Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation," 2023.
- [25] Z. Lin, S. Trivedi, and J. Sun, "Generating with confidence: Uncertainty quantification for black-box large language models," *arXiv preprint arXiv:2305.19187*, 2023.
- [26] V. Quach, A. Fisch, T. Schuster, A. Yala, J. H. Sohn, T. S. Jaakkola, and R. Barzilay, "Conformal language modeling," *arXiv preprint arXiv:2306.10193*, 2023.
- [27] B. Kumar, C. Lu, G. Gupta, A. Palepu, D. Bellamy, R. Raskar, and A. Beam, "Conformal prediction with large language models for multi-choice question answering," *arXiv preprint arXiv:2305.18404*, 2023.
- [28] E. Lakomkin, M. A. Zamani, C. Weber, S. Magg, and S. 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)*. IEEE, 2018, pp. 854–860.
- [29] V. S. Dorbala, A. Srinivasan, and A. 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)*. IEEE, 2021, pp. 3538–3545.
- [30] M. Eppe, S. Trott, and J. 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)*. IEEE, 2016, pp. 731–738.
- [31] Z. Hu, J. Pan, T. Fan, R. Yang, and D. Manocha, "Safe navigation with human instructions in complex scenes," *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 753–760, 2019.
- [32] E. Lakomkin, M. A. Zamani, C. Weber, S. Magg, and S. 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)*, 2018, pp. 854–860.
- [33] A. Francis, C. Pérez-d'Arpino, C. Li, F. Xia, A. Alahi, R. Alami, A. Bera, A. Biswas, J. Biswas, R. Chandra *et al.*, "Principles and guidelines for evaluating social robot navigation algorithms," *arXiv preprint arXiv:2306.16740*, 2023.
- [34] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning*. PMLR, 2023, pp. 28492–28518.
- [35] P. J. Lou, Y. Wang, and M. Johnson, "Neural constituency parsing of speech transcripts," *arXiv preprint arXiv:1904.08535*, 2019.
- [36] J. Ogata, M. Goto, and K. 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*. IEEE, 2009, pp. 4305–4308.
- [37] L. Goupil, E. Ponsot, D. Richardson, G. Reyes, and J.-J. Aucouturier, "Listeners' perceptions of the certainty and honesty of a speaker are associated with a common prosodic signature," *nature communications*, 2021.
- [38] J. J. Guyer, L. R. Fabrigar, and T. I. Vaughan-Johnston, "Speech rate, intonation, and pitch: Investigating the bias and cue effects of vocal confidence on persuasion," *Personality and Social Psychology Bulletin*, vol. 45, no. 3, pp. 389–405, 2019.
- [39] E. Székely, J. Mendelson, and J. Gustafson, "Synthesising uncertainty: The interplay of vocal effort and hesitation disfluencies," in *INTER-SPEECH*, 2017, pp. 804–808.
- [40] X. Jiang and M. D. Pell, "The sound of confidence and doubt," *Speech Communication*, vol. 88, pp. 106–126, 2017.
- [41] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 24824–24837, 2022.
- [42] S. Yousefi, L. Betthausen, H. Hasanbeig, A. Saran, R. Millière, and I. Momennejad, "In-context learning in large language models: A neuroscience-inspired analysis of representations," *arXiv preprint arXiv:2310.00313*, 2023.
- [43] J. Wu, R. Antonova, A. Kan, M. Lepert, A. Zeng, S. Song, J. Bohg, S. Rusinkiewicz, and T. Funkhouser, "Tidybot: Personalized robot assistance with large language models," *arXiv preprint arXiv:2305.05658*, 2023.
- [44] S. Huang, Z. Jiang, H. Dong, Y. Qiao, P. Gao, and H. Li, "Instruct2act: Mapping multi-modality instructions to robotic actions with large language model," *arXiv preprint arXiv:2305.11176*, 2023.
- [45] Y. Bang, S. Cahyawijaya, N. Lee, W. Dai, D. Su, B. Willie, H. Lovenia, Z. Ji, T. Yu, W. Chung *et al.*, "A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity," *arXiv preprint arXiv:2302.04023*, 2023.
- [46] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang *et al.*, "A survey on evaluation of large language models," *ACM Transactions on Intelligent Systems and Technology*, 2023.
- [47] M. T. Ribeiro, S. Singh, and C. Guestrin, "Semantically equivalent adversarial rules for debugging nlp models," in *Annual Meeting of the Association for Computational Linguistics*, 2018. [Online]. Available: <https://api.semanticscholar.org/CorpusID:21740766>
- [48] E. Shayegani, Y. Dong, and N. Abu-Ghazaleh, "Plug and pray: Exploiting off-the-shelf components of multi-modal models," *arXiv preprint arXiv:2307.14539*, 2023.
- [49] J. Morris, E. Lifland, J. Y. Yoo, J. Grigsby, D. Jin, and Y. 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*, 2020, pp. 119–126.
- [50] L. Li, R. Ma, Q. Guo, X. Xue, and X. Qiu, "Bert-attack: Adversarial attack against bert using bert," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 6193–6202.
- [51] D. Reis, J. Kupec, J. Hong, and A. Daoudi, "Real-time flying object detection with yolov8," *arXiv preprint arXiv:2305.09972*, 2023.
- [52] R. Smith, "An overview of the tesseract ocr engine," in *Ninth international conference on document analysis and recognition (ICDAR 2007)*, vol. 2. IEEE, 2007, pp. 629–633.
- [53] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun, "Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 3, 2022.