TacTalk: Personalizing Haptics Through Conversation

Anchit Mishra
University of Waterloo
Waterloo, Ontario, Canada
amishra@uwaterloo.ca

Oliver Schneider
University of Waterloo
Waterloo, Ontario, Canada
oliver.schneider@uwaterloo.ca

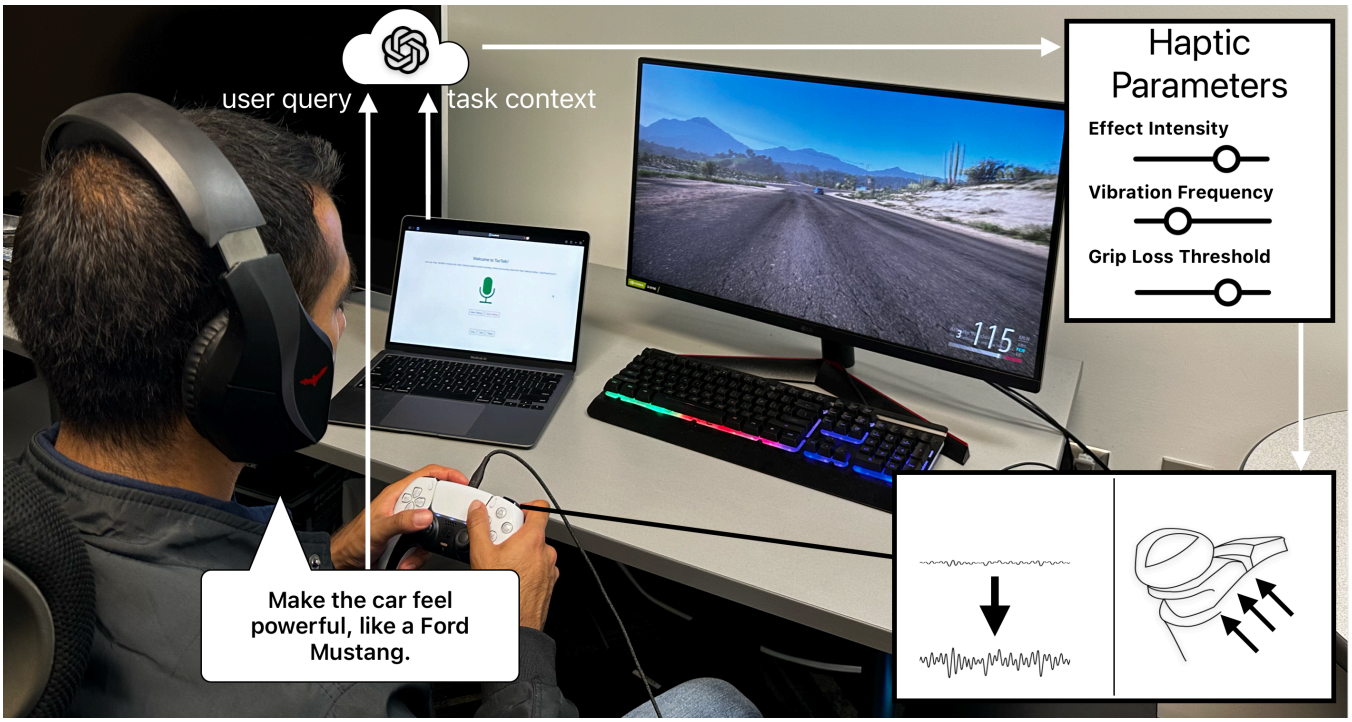


Figure 1: TacTalk is a conversational system for personalizing haptic feedback. Using LLMs, it maps natural-language user queries to low-level software parameters, enabling users to personalize haptic experiences in real time.

Abstract

Haptic experiences are highly personal, but despite prior work exploring interfaces enabling personalization, we don't know what process drives the personalization of haptics. To enable a study of this process, including users' mental models and vocabularies, we introduce TacTalk, a conversational system enabling real time tuning of virtual haptic experiences. We present an application using TacTalk in a popular racing video game, Forza Horizon 5. Through an empirical study, we find that tracking user preference profiles may improve TacTalk's ability to cater to individual differences, and that TacTalk is more usable than an existing slider-based personalization tool. A thematic analysis of participant interviews reveals an archetypal process of conversational personalization - starting with real-world experiences and domain-specific metaphors, then subsequently inspecting specific aspects of the experience including in-game events and the game controller.

CUI '25, Waterloo, ON, Canada

ACM ISBN 979-8-4007-1527-3/25/07
https://doi.org/10.1145/3719160.3736638

CCS Concepts

• **Human-centered computing** → **Interactive systems and tools**; **Haptic devices**; *Natural language interfaces.*

Keywords

Haptics, Mental Models, Vocabulary, Personalization, Conversational Interfaces

ACM Reference Format:
Anchit Mishra and Oliver Schneider. 2025. TacTalk: Personalizing Haptics Through Conversation. In *Proceedings of the 7th ACM Conference on Conversational User Interfaces (CUI '25), July 08–10, 2025, Waterloo, ON, Canada.* ACM, New York, NY, USA, 19 pages. https://doi.org/10.1145/3719160.3736638

1 INTRODUCTION

Researchers have established that individual differences exist in the perception and interpretation of haptic feedback [17, 40, 73, 92, 93, 104]. While certain experiences aim to simulate reality and thus cannot cater to individual preferences, personalized haptic feedback can benefit applications ranging from mobile applications to education environments and multisensory video games.

Prior work has explored different strategies enabling the personalization of haptic feedback, including design tools that allow users to visualize and design their own haptic effects [86, 89, 98], as well as morphing algorithms that allow users to draw upon pre-existing libraries of haptic effects to produce predictable and perceivable compound effects [18]. Seifi [87] classifies personalization approaches into three categories: choosing (selecting haptic effects from a larger collection), tuning (manipulating a given haptic effect based on perceivable parameters), and chaining (composing multiple smaller haptic effects to obtain a compound sensation). While choosing is the most common, it is least preferred; tuning is most preferred [88].

While single parameter tuning (e.g., volume, intensity) is easily deployed, navigating more complex haptic parameter spaces can be difficult. Previous work has explored language to define semantic dimensions mapping to technical parameters, including psychophysics studies aiming to understand perceived tactile dimensions for various materials [30, 66] and semantic meaning of effects [34] and higher level descriptions of the experience of sensations [33, 65]. Zheng and Morrell [111] found a negative correlation between user attention and affect, recommending an intensity control to account for individual preferences. Personalization interfaces have been explored [88], with haptic effects organized by different organizational schema [90] but progress has been limited. Some question the existence of a constant tactile language altogether [35]. Ultimately, the language around tactile feedback is multi-faceted, varied, and often vague.

To support vague, varied language, we turn to Large Language Models (LLMs). Recently, LLMs have shown strong capabilities of producing high-quality natural language outputs [36, 53, 67, 69, 100], which can be extended to multimodal application contexts through the concept of user interface transducers [8]. LLMs can keep track of context and can even understand the meaning of seemingly vague natural language queries [4, 29, 53]. This presents the opportunity of using LLMs to map user descriptions of haptic feedback derived from a diverse and non-standardized vocabulary to underlying control parameters. By interacting entirely through natural language, LLMs further enable opportunities to study individual language surrounding touch and store profiles for individualized feedback.

In this paper, we introduce TacTalk, a novel application of LLMs as a natural language transducer to study conversational personalization of haptic experiences. We explore TacTalk in the context of a car-racing video game, Forza Horizon 5, with 17 haptic parameters for PS5 DualSense controller. We first conduct a technical evaluation of TacTalk to tune the system and check for consistency. We then conduct a user study (N=11) comparing TacTalk and a slider-based visual interface, finding TacTalk has significantly higher usability scores and lower perceived mental demand, success, effort and frustration ratings. TacTalk showed consistency in terms of its ability to map technical parameters to a given world model, which we demonstrate may be formed using user profiles or even the baseline GPT-4o model [67] without providing additional context. Our primary contribution is a thematic analysis of user interactions with TacTalk that reveals insights into how people personalize haptic experiences. We find that personalization follows a unidirectional, few-shot (less than five queries) process; users employ real-world knowledge of haptics both when describing preferences and when evaluating TacTalk's outputs; domain-specific metaphors are specifically favoured over generic language when describing preferences. We conclude with guidelines for conversational and personalization interfaces involving haptics.

2 RELATED WORK

2.1 User Experience and Haptic Experience

User experience (UX) depends on individual context, including the user's internal state of mind, the environment, individual differences [28]. It also varies over time - a user's perception of UX may be shaped based on previous interactions with a product, ongoing use, and can even change after the interaction concludes [7, 49]. There are various approaches towards modeling user preferences for UX personalization, including recommender systems for personalizing website widgets [39], conceptual user models constructed using long-term usage data for augmented reality training systems [70], and domain- and device-independent UX models [32].

Recently, researchers have started to connect UX research to haptics. Studies have found that measures of UX can increased with haptic feedback, e.g., with movies [59] and video games [91]. The personal nature of UX and haptics is increasingly recognized in this line of work. The Haptic Experience (HX) model specifically aims to address experiences involving haptic technologies, and highlights personalization as a critical factor [41]. Although the HX model has been further developed with the goal of measuring HX [5, 83], they focus on the experiential factors, not personalization. Other efforts have looked into the design parameters specifically to support customization, identifying spatial density and intensity as independent parameters to enable personalization [82]. While an important development, this work focuses on technical design parameters rather than semantic or natural language from users.

2.2 Personalizing Haptic Feedback

Parallel to work formalizing the concept of HX, researchers have studied personalization of haptic feedback. Tools supporting such personalization make use of a wide variety of methods, including relative ranking and refinement using machine learning [56], choosing from existing libraries of effects [90, 98], drawing [89, 98], morphing and chaining [18] and even voice-based design involving sound-symbolic language [21, 60].

In addition to tools enabling haptic feedback personalization, previous work has also attempted to characterize individual differences

in the perception of touch. These differences can be physiological or involve information processing. For instance, haptic perception declines with age, with some regions of the body more affected than others [92, 93]. Similarly, physiological age- and sex-related factors such as differences in physical structure and stiffness of the skin, and hand size also influence the perception of touch [52, 104]. Some of these findings have been embedded into adaptive systems that use age in their model, predicting whether a user would detect a wearable vibration [9].

We find similar results when looking at higher level perception based on physiological and information processing factors. For instance, the Need For Touch (NFT) scale [73] consists of two dimensions: autotelic and instrumental need for touch, based on information processing. Physiological factors have been reported on for vibrations on the hand [55], and the skin's vibration characteristics [40].

2.3 Haptic Vocabularies

Language is a key factor of interest in the study of individual differences in haptic perception. Works focusing on both personalized haptics [87] as well as HX from a design perspective [85] acknowledge the importance of language, especially affective language in describing haptic experiences.

Studies have attempted to analyze the different perceivable dimensions of haptic stimuli based on psychophysical methodologies. For instance, Hollins et al. [30] ran a study to isolate common perceivable dimensions of tactile surface texture, finding a three-dimensional space with dimensions roughness-smoothness, hardness-softness, and compressional elasticity. Okamoto et al. [66] conducted a similar analysis with texture perception, finding five overall dimensions of tactile perception: roughness-smoothness (both macro and fine), hardness-softness, coldness-warmness, and friction. Kaneda et al. [38] investigated the relationship between physical attributes and onomatopoeic language when pressing a soft object, finding interactions between the visuo-tactile properties of an object and the onomatopoeia used to describe it. Tools such as VibViz [90] enable choosing vibrotactile effects based on different facets of language, including metaphors, physical filters, emotional filters, and usage example filters. Voodle [60] enables user customization of haptic feedback in the form of the motion of a 1-degree-of-freedom robot through sound-symbolic language like vocalizations and onomatopoeia like. boom, woof, and ding. Weirding Haptics [21] enables designers to vocally sketch vibrations to rapidly prototype haptics when designing in VR. However, the use of haptic vocabularies is still not perfectly understood, and there is a lack of consensus about the existence of a tactile language [35].

2.4 Large-Language Models and UI Transducers

Recent advances have resulted in the introduction of LLMs, autoregressive language models built using decoder-only transformer architectures [103]. LLMs possess language understanding and text generation capabilities, and are hypothesized to possess emergent knowledge about domain-specific language as well as 'world models' based on the datasets they are trained on [10, 96], even though this isn't yet fully understood [84]. As a consequence of their data sources, they also possess biases and inaccuracies [37, 64] which can be reduced using techniques such as data augmentation [75, 105], filtering [25, 97], and data generation [94, 101].

Despite their limitations, LLMs are widely used today for conversational interaction [22, 67], text classification [15, 74], text summarization [95, 110], and math [112]. Moreover, with methods like prompt engineering and associated support tools [6, 26], LLMs have been used for increasingly complex tasks, such as social simulacra [71], web design [3] and virtual reality (VR) scene generation [20]. Using LLMs as transducers for user interfaces (UI Transducers) [8] shows the potential of using LLMs' for modalities beyond text. For example, when acting as digital characters in social simulacra [71], LLMs were provided text descriptions of the environment and the characters they played, which they used as context to inform their actions. When used for VR scene generation, a similar framework enabled both scene understanding and inter-LLM communication [20], which allowed the system to control the visuals, physics and sensor integration for scenes in Unity3D [102]. LLMs are also being explored for personalization tools, with techniques like using misaligned responses [42] and smartphone sensor data [109] enabling personalization without having to retrain separate models for different users.

Despite promising capabilities of systems like LLMR [20], researchers have only recently started using LLMs for haptic experiences [54, 79]. This presents an opportunity to explore the use of conversational LLM systems as an interaction technique for users to personalize haptic experiences. Such systems may enable end-user-driven personalization, which has been shown to improve user experience in conversational interactions [77], or to probe language use when customizing haptic effects. Prior work comparing the usability of voice interfaces to graphical alternatives shows that they may provide advantages in certain contexts, but aren't preferred in others. For instance, greater performance improvements were seen for both literate and semi-literate users for a ticket reservation task when using a visual interface instead of voice interaction [16] , although the voice interface involved fewer intermediate steps, allowing users to jump straight to what they wanted. For smart-home control with robots, Luria et al. [57] found that users perceived lower control and situational awareness when using their voice instead of a visual display. More recently, Reicherts et al. [78] showed that participants prefer voice interfaces over visual ones in human-human collaborative tasks mediated by virtual agents. Even as voice interfaces become increasingly common in modern applications, issues such as learnability, error correction, and feedback are yet to be thoroughly resolved [63].

3 TACTALK SYSTEM DESIGN

TacTalk is a conversational assistant using the GPT-4o model for language processing. We use Forza Horizon 5 [2], a popular racing game as our study context due to the similarity between our interaction loop and a driver communicating with their race engineer, along with the potential for rich, dynamic haptic sensations while driving. Forza Horizon 5 is accessible to a variety of audiences, has easy-to-learn controls, and arcade-like physics engine which does not necessarily require realistic physical sensations, but still avoids a fantasy setting like Mario Kart.

In order to provide high-fidelity haptic feedback to users, we use the DualSense PS5 controller [1] which offers triggers with force feedback. However, the DualSense controller's trigger feedback is not supported in Forza Horizon 5 by default. To work around this issue, we use the ForzaDSX mod [19] (shown in Figure 2) with the DualSenseX mod application [24]. This allows us to render haptic effects on the DualSense controller's triggers using telemetry data from Forza Horizon 5. Table 1 shows a complete list of the telemetry-based parameters exposed by ForzaDSX, and Figure 3 shows how some of these parameters correspond to in-game elements. Similar methods have been used previously to synthesize haptic feedback on custom devices from telemetry data [51].

3.1 Interaction Loop

Users only see the TacTalk front-end, which activates voice recording and processes queries. Figures 4 and 5 illustrate how a user might interact with TacTalk. During gameplay, a user submits a query via voice commands and waits for the system to apply changes to the game. Our implementation enables users to perform different tasks:

- Exploring haptic feedback settings freely, with no prior idea of the underlying system parameters.
- Modifying specific haptic feedback parameters, such as vibration frequency or resistance intensity.
- Reverting to a previous configuration.
- Saving/retrieving a configuration.

3.2 System Architecture

The system architecture (Figure 6) consists of four layers:

Input Processing Saying 'Hey, TacTalk!' or pressing a button activates recording. A button-press stops recording. Recorded queries are then passed to a speech-to-text engine; we use OpenAI's Whisper API [68].

Context TacTalk performs two tasks - (1) change the haptic feedback configuration of the game controller and (2) maintain a user profile of haptic feedback preferences. Four component prompts are relevant:

(1) A system prompt specifying the task context, consisting of the task description, the name of the game, and controller being used.
(2) The user's existing preference profile. This helps align the system to individual preferences by summarizing past interactions.
(3) The existing parameter list, along with example query-response pairs.
(4) A short sliding window of the conversation history, enabling short-term memory. We include the last 3 exchanges.

The exact prompts are provided as supplementary material.

Response Two queries are sent - one for updating haptic feedback, and one updating the user preference profile based on the response to the first. Between queries, the parameters from the first query's response are validated to prevent invalid configurations.

Update In the context of ForzaDSX, we modify the `config.xml` file for the mod and restart it to apply modified settings without disrupting gameplay. We also update the controller setting prompt and user preference profile for continuity.

3.3 Deciding Which Settings to Change

The first query passed to TacTalk includes a system prompt, providing context about the LLM's task along with the user's query. The context prompt includes the task description, past conversation history, current haptic feedback configuration, and a user preference profile indicating user preferences from past interactions. It also lists all parameters, acceptable ranges of values, and explanations for how each one influences haptic feedback. The LLM's response is a JSON object containing a list of modified parameter names, their previous and new values, and a justification for each change.

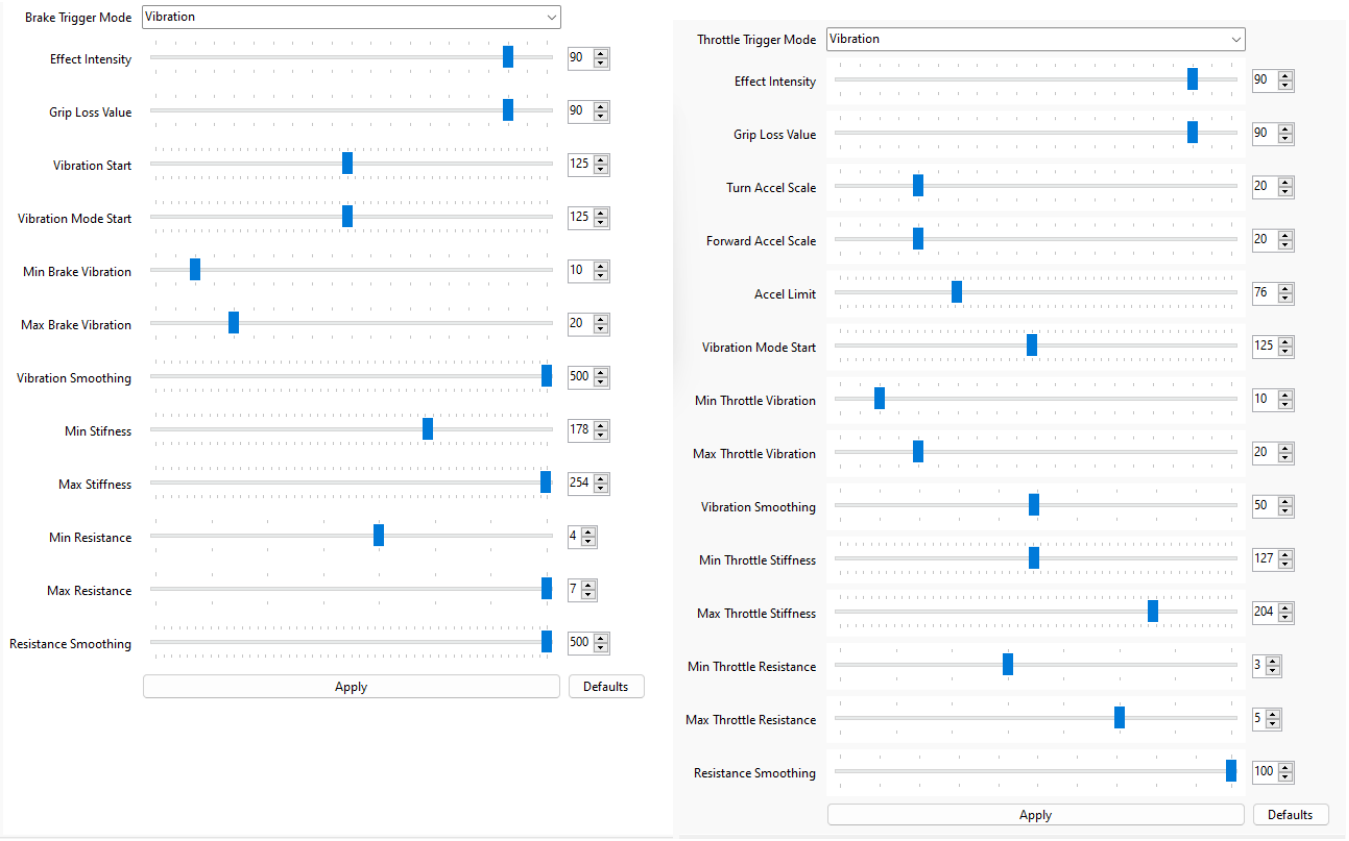
While LLMs struggle with complex control tasks [62], prompting techniques like Chain-of-Thought prompting [106] and In-Context Learning [23] can help. This lets us explore LLMs to render haptic feedback without requiring large datasets, as needed for other deep learning approaches. The first prompt thus includes examples to illustrate the role of different parameters, and the justification field helps make parameter modifications more predictable and consistent. We choose relatively abstract parameters compared to low-/hardware-level haptic feedback parameters. This means that instead of generating haptic rendering parameters for each frame (e.g., a PID controller), TacTalk manipulates hyperparameters influencing the quality of haptic feedback. For instance, 'Grip Loss Sensitivity' sets a traction loss threshold for the triggers to start vibrating instead of generating vibration parameters each frame. Moreover, mapping vague user queries to a predefined, 17-dimensional space allows us to look closely at user interactions and whether user queries actually resulted in corresponding parameter changes. Such hyper-parameters are commonly used for customization, whether in commercial APIs (e.g., "sharpness" in Apple's AHAP format) or through semantic mappings in research tools (e.g., [34, 90]). Further, the scarcity of large datasets required for specialized training makes implementing a fully neural haptic rendering algorithm difficult in practice.

3.4 Updating the User Preference Profile

Once the LLM responds to the first query, the second query updates the user preference profile. LLMs are capable of text summarization and sentiment analysis, and thus this component was a logical step towards enabling an 'understanding' of individual haptic vocabularies. The summary produced by TacTalk (1) helps retain information about user preferences beyond the default 3-exchange conversation history, and (2) isolates the various adjectives, onomatopoeia, metaphors etc. mentioned by users. Such profiles may also help analyze user sentiments or to infer common attributes across user bases in future applications.

4 TECHNICAL EVALUATION

We conducted (1) an error analysis and (2) a dimensionality reduction visualization of responses to multiple queries to evaluate TacTalk's consistency. A 51-prompt dataset was used for the evaluations, based on the notion of car performance levels.



(a) (b)

Figure 2: The ForzaDSX visual interface enabling customized force feedback for the DualSense controller. (a) The 13 parameters for the brake/left trigger. (b) The 15 parameters for the throttle/right trigger. 17 total parameters were identified as viable for use with TacTalk, shown in Table 1. All other parameters were found to behave inconsistently, likely due to the complete query approaching the GPT-4o model's context-tracking limits.

Parameter Name	Targeted Control	Acceptable Values	Significance
Trigger Mode	Throttle + Brake	OFF, RESISTANCE, VIBRATION	Set triggers to passive, render only uniform resistance or resistance and vibrations.
Effect Intensity	Throttle + Brake	0-100	Overall intensity of rendered haptic feedback.
Grip Loss Sensitivity	Throttle + Brake	0-100	Threshold traction loss required to start vibrations on triggers.
Forward/Turning Acceleration Scale	Throttle	0-100	Forward acceleration's contribution towards throttle vibration.
Acceleration Limit	Throttle	0-100	Proportion of acceleration beyond which haptic feedback plateaus.
Minimum/maximum Vibration	Throttle + Brake	0-100	Vibration frequency controls.
Minimum/maximum Resistance	Throttle + Brake	0-100	Resistance magnitude controls.

Table 1: The parameters exposed by ForzaDSX. All numeric parameters were modified to the 0-100 range since this resulted in fewer LLM-generated out-of-range errors.

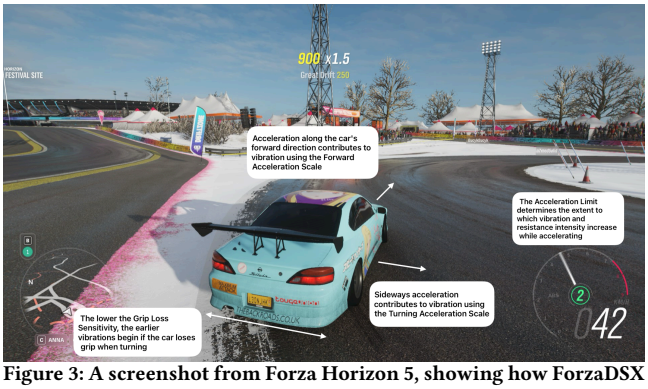


Figure 3: A screenshot from Forza Horizon 5, showing how ForzaDSX settings apply in-game. Refer to Table 1 for more details. This camera is only used for illustrative purposes; a first person view was used in our study, reducing visual indicators.

4.1 Prompt Dataset

Large datasets for haptics have only recently emerged [90, 98], and datasets for specialized contexts are even scarcer. Thus, we constructed a small 51-query dataset, each of the form 'I want the car to feel like a...' followed by the name or type of a car. Such queries allow TacTalk to freely manipulate parameters, as opposed to queries specifically mentioning technical terms. Based on each car's perceived performance levels online, we divided the dataset into three groups - high, average, and low performance, with 17 cars belonging to each group. It is important to note that performance-based classification is somewhat arbitrary and based on existing biases in public perceptions of performance. Nevertheless the classification, although imperfect, still provides clarity to distinguish between cars. See the supplementary materials for the complete dataset.

4.2 Error Analysis

4.2.1 Methods. We used the following criteria to identify output errors:

(1) The validity of the returned JSON, including syntax and the presence of all required fields.
(2) The validity of categorical attributes (the trigger mode must be set to 'Off', 'Resistance' or 'Vibration').
(3) The validity of numeric attributes, including data type and range.
(4) The consistency of minimum-maximum pairs (e.g., is the minimum stiffness less than the maximum stiffness?).

We compute error rates in two conditions:

(1) Using a single prompt to both modify haptic feedback settings and update the preference profile.
(2) Using two separate prompts for haptic feedback settings and updating the preference profile. In this condition, the prompt for haptic feedback settings also includes example query-response pairs.

4.2.2 Results. We repeated dataset queries 50 times, for 2550 total queries. The initial settings we used are specified in the supplementary material. Overall, using two prompts resulted in 54 errors compared to 1932 with one prompt - a difference of 73.6%. All 1932 errors involved issues with output structure - for instance, inconsistencies in the returned JSON string (missing fields, inconsistently including or omitting quotation marks around the output, etc.). Of these, 1154 errors also involved value range violations. The 54 errors for the two-prompt setup were also related to inconsistencies in output JSON structure, and of these 14 errors included value range violations.

Our proposed design for TacTalk is robust, causing output errors only 2.1% of the time. Since the time of writing this paper, new API features enable structured JSON output, and thus we believe these issues will only become rarer over time. TacTalk took an average of 6.15 seconds per query (minimum = 4.18s, maximum = 15.17s, SD = 2.63s).

4.3 t-SNE Dimensionality Reduction

4.3.1 Methods. We conducted an evaluation of TacTalk's response generation consistency. Using our 51-query dataset, we obtained system-generated responses to each query and ran t-SNE dimensionality reduction [58] on the generated parameter vectors (learning rate = 10, perplexity = 20, random initialization).

4.3.2 Results. Figure 7 shows the t-SNE plot. The TacTalk-generated parameter configurations belong to relatively distinct clusters for high, average, and low performance vehicles, with little overlap between the average and low performance vehicle groups. Some outliers were also observed, such as the Mini Cooper being classified as a high performance vehicle, or the Fiat 500 being closer to average than low performance. While these can be labeled as misclassifications, there are sports variants of the Mini Cooper, and the Fiat 500 also has higher performance variants, making the boundaries between categories fuzzier. Note that distances do not convey much information beyond identifying clusters in t-SNE. Even so, alternative dimensionality reduction methods for non-linear data such as UMAP [61] have been shown to differ from it only in their initialization strategies [44, 45].

Our technical evaluation provides evidence that TacTalk behaves predictably, inferring implicit patterns from user queries and changing values based on these inferences.

5 USER STUDY

Our study addresses the following research questions:

RQ1 Can an LLM effectively function as a UI transducer for mapping natural language user queries to technical haptic feedback parameters?
RQ2 How do people personalize haptic feedback using language, without any assistance in the form of visualization?
RQ3 How does a voice-based personalization interface compare with an existing slider-based tool (ForzaDSX) offering the same functionality in terms of usability and perceived cognitive load?

As discussed in section 2, prior work studying voice interfaces in different contexts shows they may be preferred for certain applications but problematic elsewhere, and issues like learnability and error correction persist. Unlike most interfaces reported on previously, however, TacTalk does not provide any conversational

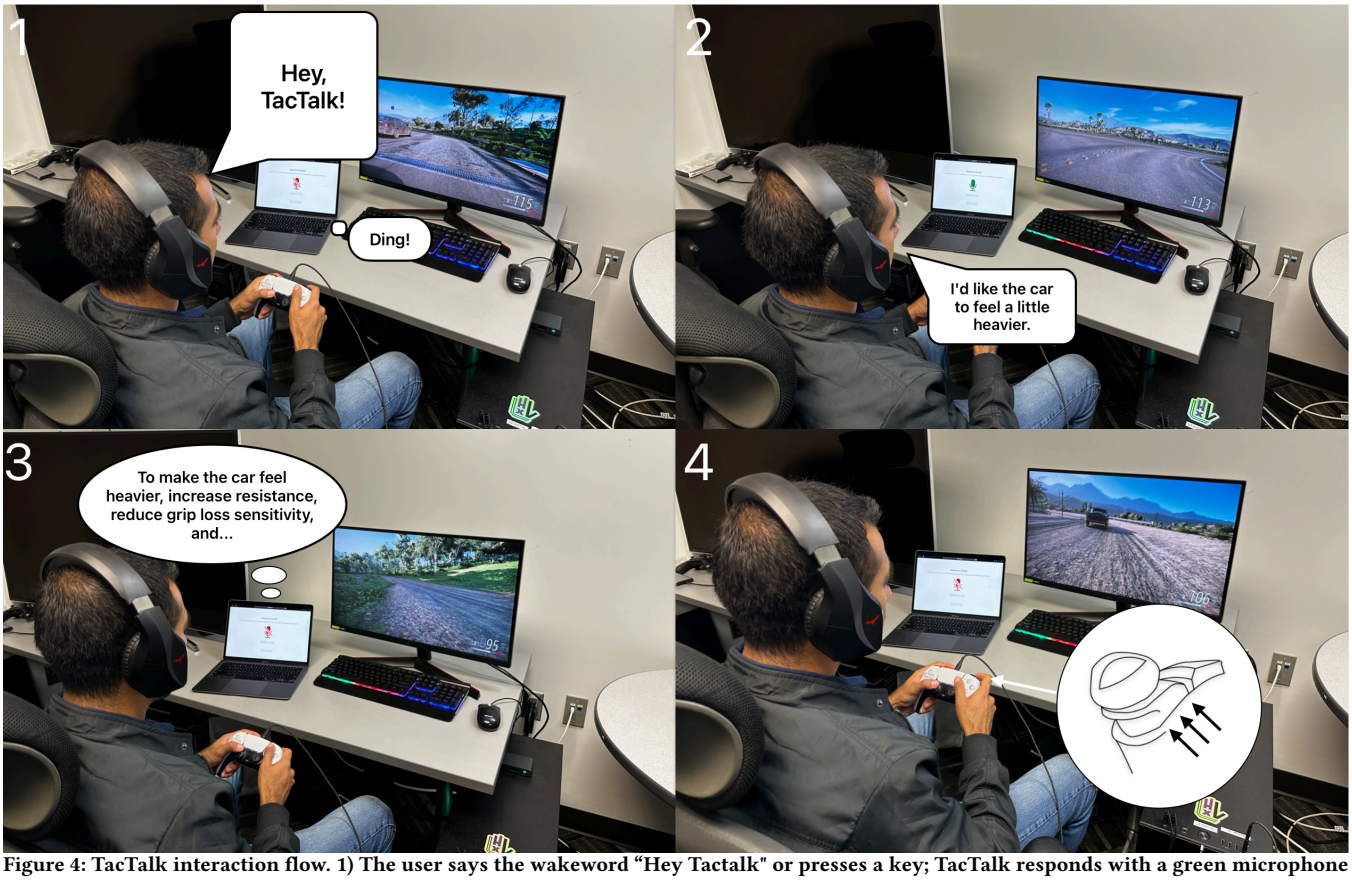


Figure 4: TacTalk interaction flow. 1) The user says the wakeword "Hey Tactalk" or presses a key; TacTalk responds with a green microphone and sound effect. 2) The user poses their query. 3) TacTalk processes the request. 4) TacTalk updates the haptic feedback.

Figure 5: A user may choose to follow up with the system in one of three ways. (A) If the response aligns with their preferences, they may only ask for minor adjustments. (B) If not, they may look towards other attributes to achieve a preferable setting. (C) They may also revert to a previous configuration.

feedback. Instead, it receives voice commands and outputs an updated haptic feedback configuration, only playing alert sounds to indicate when it is listening. We did this to allow the study of naturalistic user queries without introducing vocabulary bias; at the same time it allows us to evaluate the voice interface's usability purely in terms of its ability to convert user queries to desirable haptic feedback configurations.

5.1 Participants

We recruited 11 participants (10 male, 1 female, mean age = 24.18, min age = 18, max age = 34, standard deviation = 4.30, all self-reported) . All participants completed a screening questionnaire outlining their experience with driving and motorsport (see the supplementary material). We only included people who had some racing experience - real-world racing, virtual driving simulators or even arcade racing games (highly unrealistic games such as Mario Kart did not count). This ensured that participants (1) could get familiar with the game relatively easily, and (2) would have an idea

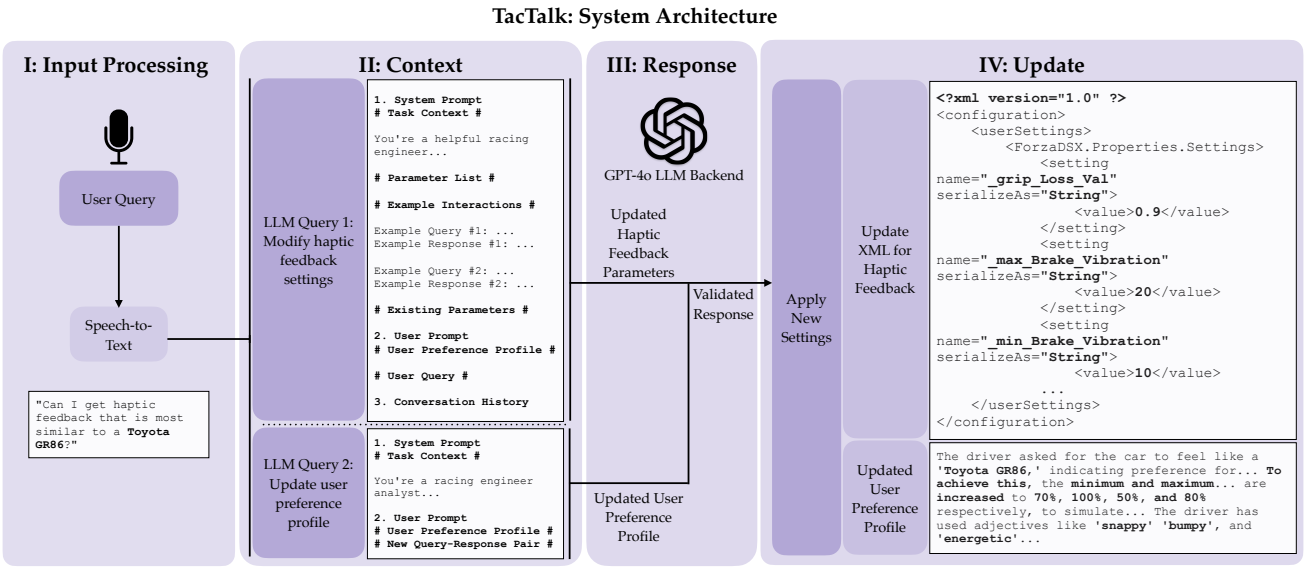


Figure 6: An illustration of TacTalk's system architecture, divided into four layers: Input Processing, Context, Response and Update. An example user query and its associated prompt and configuration strings illustrate how these conceptual layers are implemented. The Response layer involves making queries to an LLM backend (GPT-4o in our implementation).

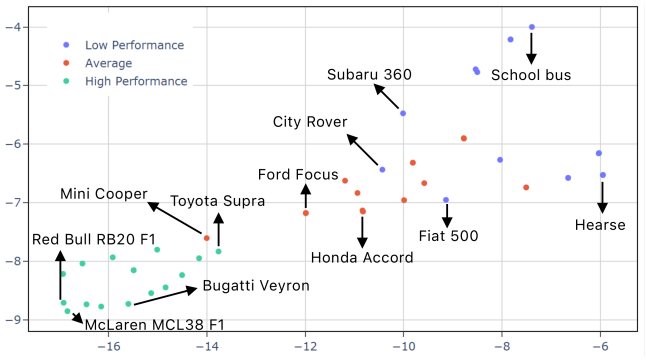


Figure 7: A t-SNE visualization of the 51-query dataset. The plot shows that haptic feedback configurations generated by TacTalk possess some structure, with vehicles of similar performance clustered together. Some visible outliers are highlighted.

Figure 8: A participant seated for a user study session, with the various devices used in the study labelled.

of their desired haptic experience and some associated vocabulary. Participants had a variety of experience levels with driving and motorsport, ranging from people who had played Need For Speed games to people who had professional go-kart racing experience.

Each study session took approximately one hour, and participants were remunerated with CA$20 Amazon gift cards for their time. Written informed consent was obtained prior to each session. The study was approved by the [anonymized for review] Research Ethics Board.

5.2 Apparatus

All study sessions were conducted indoors, with participants seated in front of a PC monitor where the game was displayed. The PC had an AMD ThreadRipper Pro 5975WX 32-core CPU, an RTX 4080 GPU and 64GB of RAM. A second laptop (an M1 MacBook Air) displaying the TacTalk web interface was placed on the left of the PC monitor. Finally, participants were provided headphones for in-game audio and the DualSense controller for gameplay. Figure 8 shows a participant seated for the study. For the entire study, we used the in-game bumper camera which hid the car from the player, removing visual bias.

5.3 Tasks

Study sessions consisted of two phases, which we refer to as Free-Personalization and Car-Classification. The study session timeline is shown in Figure 9. To further address **RQ2** and **RQ3**, we first collected participant responses to the Big Five Inventory [76]

and Player Traits [99] questionnaires. After familiarization with both the controller and game, participants commenced phase one.

5.3.1 Free-Personalization. Here, participants were allowed to freely customize in-game haptic feedback, first using TacTalk and then using the visual ForzaDSX interface, following a within-participants design. We decided to control interface order (TacTalk first, then ForzaDSX) rather than counterbalance to reduce priming for user queries and interaction. Our priorities were **RQ1** and **RQ2** over a quantitative analysis of **RQ3**, targeting users' preferences regarding voice-based personalization . Interacting with ForzaDSX's visual interface first would expose participants to haptic parameter names (e.g., Effect Intensity, Minimum Frequency, Trigger Mode, etc.), which could bias their queries to TacTalk. On the other hand, TacTalk did not produce any visual or auditory outputs that would induce vocabulary bias. We believe this potential order effect trade-off improves the ecological validity of our study of personalization using TacTalk.

When using TacTalk, participants described their desired haptic feedback, and based on the system's response, iterated until they arrived at a configuration that they felt comfortable with. Similarly, when using ForzaDSX, participants switched back-and-forth between the mod's interface and gameplay, manipulating settings until they found a comfortable haptic feedback configuration. In both cases, semi-structured interviews were conducted to evaluate whether the participants were able to obtain a desirable haptic feedback configuration (addressing **RQ1**), and to understand the intent behind their queries (**RQ2**). After using each interface, participants completed the NASA Task Load Index (NASA-TLX) [27] and System Usability Scale (SUS) [12] questionnaires (**RQ3**) before proceeding to the next phase.

5.3.2 Car-Classification. In the second phase, participants tried three randomly-ordered haptic feedback configurations generated using TacTalk and matched each configuration to one of the following vehicles: a "Formula One car", a "double-decker bus", and an "everyday sedan". This task was performed twice: once without any added context, and once with added context (order counterbalanced). The game audio omitted engine noise for this stage to avoid biasing participants through modalities other than haptics. Participants first described what they expected each vehicle to feel like (this was added as context). They then tried out each configuration, freely switching between them until they had an answer. We included this task to better answer **RQ1**, more specifically the following sub-questions:

RQ1.1 Can TacTalk generate perceivably distinct haptic feedback settings?

RQ1.2 Are TacTalk's generated settings consistent with user expectations even without additional context?

RQ1.3 If TacTalk's responses do not match user expectations, can additional context correct this misalignment?

6 QUANTITATIVE RESULTS

6.1 Free-Personalization Results

Due to the failure of the assumption of normality, we ran the Wilcoxon Signed Rank Test to compare the NASA-TLX and SUS scores. Overall, we found that participants perceived significantly lower mental demand ($p = 0.003$), perceived success ($p = 0.043$), perceived effort ($p = 0.003$) and perceived frustration ($p = 0.012$) when using TacTalk instead of ForzaDSX. SUS scores (TacTalk: $\mu = 66.36$, $SD = 9.51$, ForzaDSX: $\mu = 47.72$, $SD = 11.32$) also yielded significant differences with a large effect size ($p < 0.01$, Cohen's $d = 1.78$). No significant differences were observed for Big-Five Inventory and Player Traits responses. TLX score plots can be found in the **appendix**.

Without Preference Context			
Vehicle	Formula One Car	Double-Decker Bus	Average Sedan
Sensitivity	0.636	0.636	0.545
Specificity	0.818	0.818	0.772
Accuracy: 60.61%; 95% CI = (42.14%, 77.09%)			
With Preference Context			
Vehicle	Formula One Car	Double-Decker Bus	Average Sedan
Sensitivity	0.818	0.909	0.818
Specificity	0.909	0.954	0.909
Accuracy: 84.85%; 95% CI = (68.10%, 94.89%)			

Table 2: The specificity and sensitivity values measured for the Car-Classification task, both with and without added user preference context.

6.2 Car-Classification Results

The results of the Car-Classification task are summarized in Table 2 and Figure 10. In general, participants achieved a higher classification accuracy when they provided additional preference context to the system (84.85% vs. 60.61%), although 95% confidence intervals indicated this was not significant. Interestingly, 6/11 participants were able to guess all three classifications accurately even when the system did not track their preferences, indicating that the 'world model' derived from GPT-4o's training dataset was able to capture signals related to haptic feedback attributes for different vehicles. When profiles were considered, these 6 participants also accurately classified all three cars. Of the remaining 5 participants, 3 increased their accuracy to 100% after profiles were applied. This suggests that profiles might help with some participants without hindering those who align with the 'world model'.

7 THEMATIC ANALYSIS

We collected 5h 53m audio recordings of participants' study sessions, including interview responses and interactions with TacTalk and ForzaDSX. Video recordings (2h 12m) were also collected for consenting participants. We also analyzed user queries and TacTalk's corresponding generated responses.

We conducted a reflexive thematic analysis [11], adopting an inductive style to obtain new insights from the data. Audio was transcribed using Microsoft Word's transcription tool then manually cleaned. Initial codes were developed by the first author by exploring the entire dataset once, then refined by clustering codes and associated data items in Freeform on macOS. Both authors participated in refinement of codes and development of themes. Our work is informed by our experiences and prior knowledge,

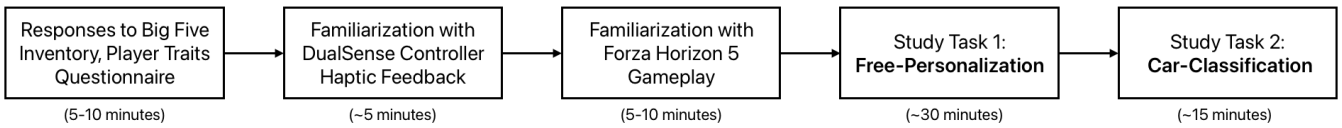


Figure 9: A timeline depicting the flow of a single user study session. The order of tasks was kept the same across all participants, since our aim was to first understand their approach towards personalization and then reveal the notion of TacTalk keeping track of individual user preferences.

for example, the first author actively follows motorsport and enjoys playing simulation racing games. Overall, we identified the following 5 themes:

7.1 T1: Users favour a few-shot, unidirectional approach towards personalizing haptics through conversation

We observed that haptic feedback personalization with TacTalk typically followed a few-shot process, with only one participant (P3) posing more than five queries. On average, participants posed 3.5 queries to the system, with four participants (P0, P1, P6, P9) only making one query. When asked about their perception of the haptic feedback generated by TacTalk, P0 said *'I think I'm very content... I'm pretty accurate.'* This confidence was partially due to their own ability to phrase a detailed query:

> *'I need the accelerator to be a little more firm, so less sensitive. I also need it to feel like I am accelerating, so more rumble as I accelerate and as I go faster, as I hit the accelerator harder, I want it to feel a lot more rough. For the brake - I want it to be a little less sensitive and brake a little more smoothly, so I don't want to feel as many vibrations, unless I'm hitting the handbrake.'* (P0)

Similarly, P9 called the first configuration TacTalk generated *'comfortable'*. However, their query was less direct - they were unsure about the expected haptic feedback due to the disparity between the controller and the full-body sensation of driving a car (*'I don't know how much haptic feedback I would expect from an accelerator...'*). They simply asked TacTalk for *'haptic feedback that is most similar to a Toyota GR86.'* Hence, while different users have different initial ideas of desired haptic feedback, TacTalk can navigate the parameter space efficiently, requiring few interactions before reaching a desirable setting. TacTalk's 'undo' function was not used much - none of the participants navigated to previous settings, even when told about the feature. This suggests participants found it more natural to continue the conversation without reverting to past configurations.

These trends in participants' usage of TacTalk and their suggestions reveal an archetypal process of conversational personalization. This process is unidirectional, has multiple points of entry and goes from users describing intended haptic feedback using more abstract metaphors (e.g., P9's usage of a Toyota GR86) to specifying minute details (e.g., P3's query to make both controller triggers equal in stiffness and vibration characteristics). We saw that more experienced participants often started this process further downstream, using fewer abstract metaphors and using more low-level car- or controller-related vocabulary.

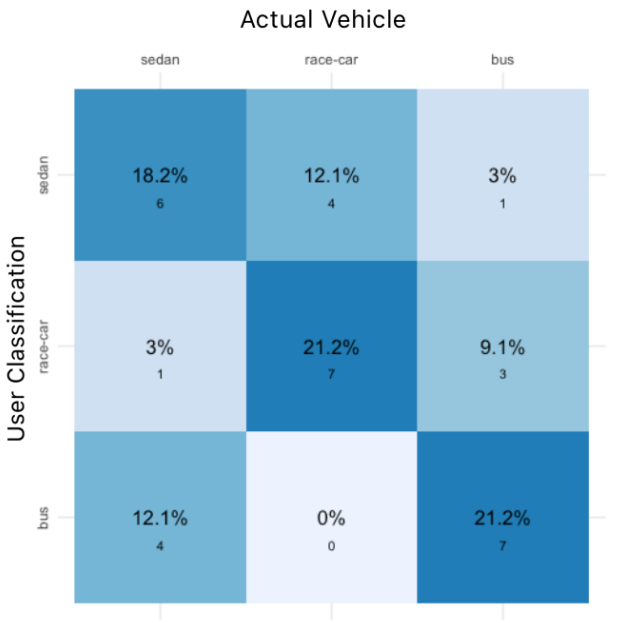
When asked for feedback on possible interfaces for TacTalk, including a physical 'push-to-talk' button on the controller was favoured over voice-based activation by all participants except two - P5 (*'I would prefer if there is an inbuilt voice assistant and I can say "Hey!", like we speak to Siri or Alexa.'*) and P6. Participants also proposed different combinations of TacTalk with ForzaDSX. For instance, P1 proposed initially setting up haptic feedback using visual sliders and using TacTalk for subsequent fine-tuning *'because it's just a little bit changed, compared to the beginning'*; P2 echoed this point too, citing prior experience with racing simulator games and real-life motorsport: *'... I can set up the parameters roughly before I start. Then I'd talk to the voice interface to adjust more precisely.'* In contrast, other participants (P4, P9) felt it would be easier to begin with TacTalk for parameter space exploration using more generic queries and subsequently fine-tuning using ForzaDSX. P5 even suggested allowing users to set custom keywords as slider parameters and manipulate them to explore different sensations, essentially creating a custom slider tool using LLMs. They also proposed an educational application for children to demonstrate how different haptic sensations feel while introducing new language.

7.2 T2: A realistic world suggests using knowledge of real-world haptics

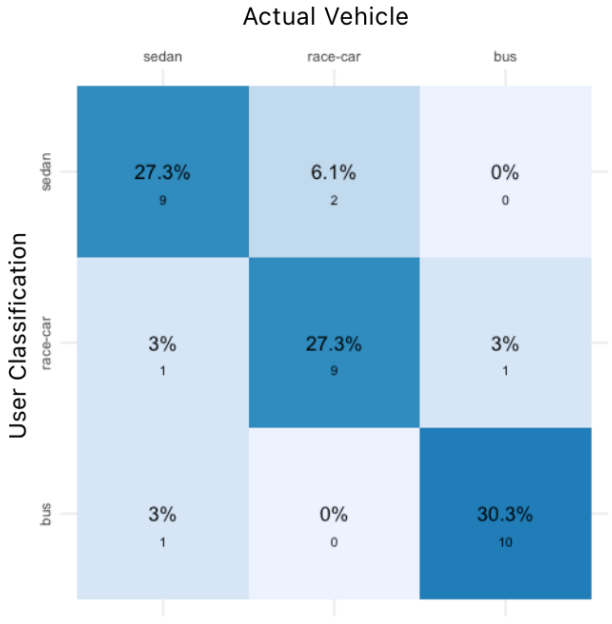
Forza Horizon 5 was described as a game with realistic visuals by multiple participants (P4, P5, P9, P10). This perceived realism in turn led to participants drawing from their real-world knowledge. Depending on their individual experiences, participants referred to both real-world driving and simulator experiences when describing their haptic feedback preferences. P10, for instance, said *'I think it's more like what I have felt when I was in a racing simulator... in the go-kart, that's the kind of the sensation you have accelerating'* when talking about their preference for haptics on the throttle trigger. Similarly, some participants (P0, P2, P5) alluded to real-world driving experience, either from go-karting (P2) or day-to-day driving (P0, P5) when describing the haptic feedback they expected.

In addition to helping construct queries to TacTalk, real-world experience also confused participants, seen in the case of P9's interactions described in T1. They noted form-factor disparity between the controller and a real car' steering and pedals as the primary reason behind their confusion. In this case, P9 resorted to their *'Toyota GR86'* query as previously mentioned. This ties back into the unidirectional nature of conversational personalization: starting from a lack of clarity about their desired feedback, P9 drew upon a realistic metaphor to explore sensations.

Just as real-world knowledge was used to describe preferences, participants often used real-world knowledge to justify the outputs generated by TacTalk. Of course, these included simple comments about the capabilities of the DualSense controller, such as when



(a) Without preference context



(b) With preference context

Figure 10: Confusion matrices visualizing performance at matching haptic feedback to vehicles. Users' overall accuracy increased from 60.6% in (a) to 84.9% in (b). Percentages consider the total number of guesses across all users.

P4 mentioned they wanted force feedback on the analog sticks (*'I don't think PlayStation controllers have that feature yet'*) or when P2 said that the intensity of the force feedback was hardware-limited (*'I would like to get some more feedback from both the brake and accelerator pedals, which may be due to the limitation of the controller'*). However, participants also mentioned personal driving experiences, such as the observation that the brake pedal in a real car feels stiffer than the accelerator (P8).

Another interesting observation was the existence of individual perceptions of realism. We as authors recognize our own ideas of realism may differ from the participants, adding to the subjectivity of our analysis. While some participants (P6, P8) referred to their idea of a realistic driving experience (driving a muscle car, feeling both resistance and vibrations on the triggers), others referred to the in-game terrain they were driving on (P5, P7) and one even referred to their belief that real-world cars have brakes that are less stiff so as to stop faster (P1).

7.3 T3: Using haptic feedback to represent changing environment and state is preferred

Similar to participants' expectations being influenced by the realistic visuals of the game, the haptic feedback was also expected to follow real-world causal relationships. Participants naturally expected certain haptic cues to correspond to specific in-game events. This behaviour was initially seen in participants' justifications for different haptic sensations - for instance, P4, P5, P9 and P10 stated preferences for realistic haptic feedback for Forza Horizon 5 due to its visual fidelity. This was observed for both the vibrotactile (*'there is some vibration - screeching tires, I think'*) and force feedback (*'when it lost traction this time, there's actually a lot more feedback on the brake'*, *'I'm on a slope, I can feel some sort of resistance here'*) sensations rendered.

Just as participants justified haptic feedback using in-game events, their preferences for haptic feedback were also shaped around these events. Some participants (P5, P7) paid attention to environmental factors when describing their desired haptic feedback, mentioning mud, snow, and rain. P7 even commented on the lack of controls in ForzaDSX for manipulating weather-related sensations, mentioning they *'wanted to configure how they* (i.e., the triggers) *should behave. Like for example in winter you need winter tires'*. In other cases (P5, P6, P7), track-related factors were mentioned, referring to the incline (*'I do feel some kind of resistance also because this is an incline'*) and on/off-road conditions (*'I want the accelerator to be smoother when I'm off road'*). Notably, none of the participants commented on the haptic feedback combined with audio - references were mainly drawn from visual in-game events. Even so, the sounds made by the controller when the triggers vibrated were noticed by participants, who either liked (P2, P3, P4) or disliked them (P5, P7, P10). P10 even derived preferences from in-game telemetry (*'I want my index finger from the left to have a lot of shake when I'm accelerating, especially between when I start and until I go to 100km/h'*). Combined, all of these sub-themes inform the notion of haptic feedback that represents changes in an experience's environment and state. Specifically, the visual modality is favoured over audio in participants' framing of preferences for haptic feedback, involving both UI elements and actual in-game objects. Haptic feedback should thus also be personalized by accounting for changes in the environment and state of the experience.

7.4 T4: Knowing about 'reality' does not imply a preference for it

Although we saw that all participants drew from real-world knowledge both when describing preferences and justifying TacTalk's outputs, this did not necessarily correspond to a preference for realistic haptic feedback. In P7 and P8's interactions, real-world experiences were used as a reference to highlight sources of frustration, such as increased throttle stiffness when driving off-road, intense vibrations indicating grip loss and increased resistance when driving up an incline. Similarly, P3 distinguished their preferences from real life, mentioning that they were looking for something that *'feels good for the game'*. P5, after converging on their preferred parameter configuration was then intrigued by TacTalk's capabilities and attempted to experiment with different, unrealistic sensations, asking for the accelerator to feel *'like a spring'* and *'as light as air.'* These two queries were the only observed instances where a participant made a query to TacTalk that directly referred to unrealistic metaphors for driving. Even so, this indicates that the open-ended nature of voice-based interactions might invite users to play around with different haptic feedback configurations, irrespective of how easily they can be defined in terms of technical parameters. Further, this indicates that at a global scale, users prefer a large, varied haptic parameter space, allowing for both realistic and unrealistic haptic feedback.

7.5 T5: Domain-specific metaphors are a valuable tool

Five participants (P1, P3, P4, P8, P9) indicated difficulties composing queries for the system due to various reasons, from not knowing what language TacTalk would be able to understand to not having a clear sense of the haptic feedback they wanted. In such circumstances, metaphors proved useful. Similar to P5's use of *'spring'* and *'air'* as metaphors for the accelerator's behaviour, other participants used metaphors to describe different scenarios. P6 asked TacTalk to replicate the sensation of driving *'a sturdy muscle car'*, stating that they wanted to experience something similar to an old Mustang, similar to P9's metaphor of a Toyota GR86.

Other notable cases of metaphor usage included P3 and P9's descriptions of the haptic feedback they felt - they used onomatopoeic vocabulary, such as *'thuk-thuk-thuk-thuk'*, *'buzzy'*, and *'dhuk-dhuk-dhuk-dhuk'*. However, such sounds were not used when talking to TacTalk, but rather when responding to interview questions. In fact, P3, P8, and P9 explicitly pointed out not being able to ascertain whether or not TacTalk could pick up on such descriptions. Thus, we found that metaphors were valuable to participants when it came to describing their expectations, even if they did not feel that a computer would be able to understand their intention. Participants that did not make use of metaphors instead focused on basic haptic adjectives, such as stiffness and vibration and directly asked for them to be modified, as discussed earlier in T1.

We also saw trends in participants' use of adjectives. The adjectives used referred to various qualities of both vibrotactile and force feedback including *smoothness* (*'Make the accelerator vibration smoother'*), *hardness/softness* (*'Make accelerator vibrations softer'*, *'Increase the brake vibrations to be harder than accelerator vibrations'*) and *stiffness* (*'Make the throttle less stiff'*). This aligns with the perceptual dimensions discussed in prior work by Okamoto et al. [66] and Hollins et al.[30] excluding the coldness-warmness dimension, since the DualSense controller does not have thermal feedback capabilities.

Overall, domain-related metaphors were brought up by participants most frequently, and more generic, unrelated metaphors were rare. More specifically, by domain-related we refer to metaphors commonly associated with cars and motorsport. In fact, P5's description of making triggers feel like 'air' or like a 'spring' were the only such observed instances. Moreover, when prompted to describe how they came up with their queries, all participants except P7 mentioned metaphors related to cars. P7 mentioned the experience of riding a bicycle uphill when describing their preferred stiffness on an incline, an activity slightly related to driving a car.

7.6 Summary

Overall, users consistently favour the use of generic metaphors in the early stages of personalization, subsequently moving towards low-level details until a desirable configuration is achieved. Along the way, they draw upon real-world knowledge, even if they may not necessarily be looking for 'realistic' haptic feedback - the realism of the game world plays a key role in influencing this. Domain-specific metaphors were the most commonly used tool for describing preferences (e.g., *'aquaplaning'*, *'like a Toyota GR86'*, *'drifting around corners'* etc.), with only two recorded instances of domain-unrelated metaphors being used. In cases such as P5 (*'as light as air'*, *'like a spring'*), a user might wish to play around with the interface's capabilities further.

8 DISCUSSION

We now discuss the implications of our study results, combining both our quantitative and thematic analyses.

8.1 LLMs can make predictable and valid parameter adjustments for haptics

Our technical evaluations show that TacTalk, an LLM-based application, can make valid parameter adjustments for haptic feedback as shown in the context of a video game. In particular, using a multi-prompt structure leads to valid results much more often than a single, monolithic query. This aligns with commonly used techniques such as Chain-of-Thought [106] and few-shot prompting [42] improve LLM responses.

Through our t-SNE evaluation, we verify the predictability and consistency of TacTalk's generated responses. As observed using a small test dataset, TacTalk generates configurations aligning with performance expectations. Although the classification between high, average and low performance vehicles is somewhat arbitrary, a visible distinction between classes was still observed. TacTalk's 'misclassifications' are also interesting. For instance, the Mini Cooper was placed closer to the high performance cluster than the Toyota Supra. However, variants of the Mini Cooper are used for racing, which may have contributed to TacTalk's response. Similarly, a Fiat 500 and a Subaru 360 may be considered average instead of low performance. Our user study further validates our claims, as most users submitted fewer than 5 queries before finding a satisfactory haptic feedback configuration. Thus, we observe that

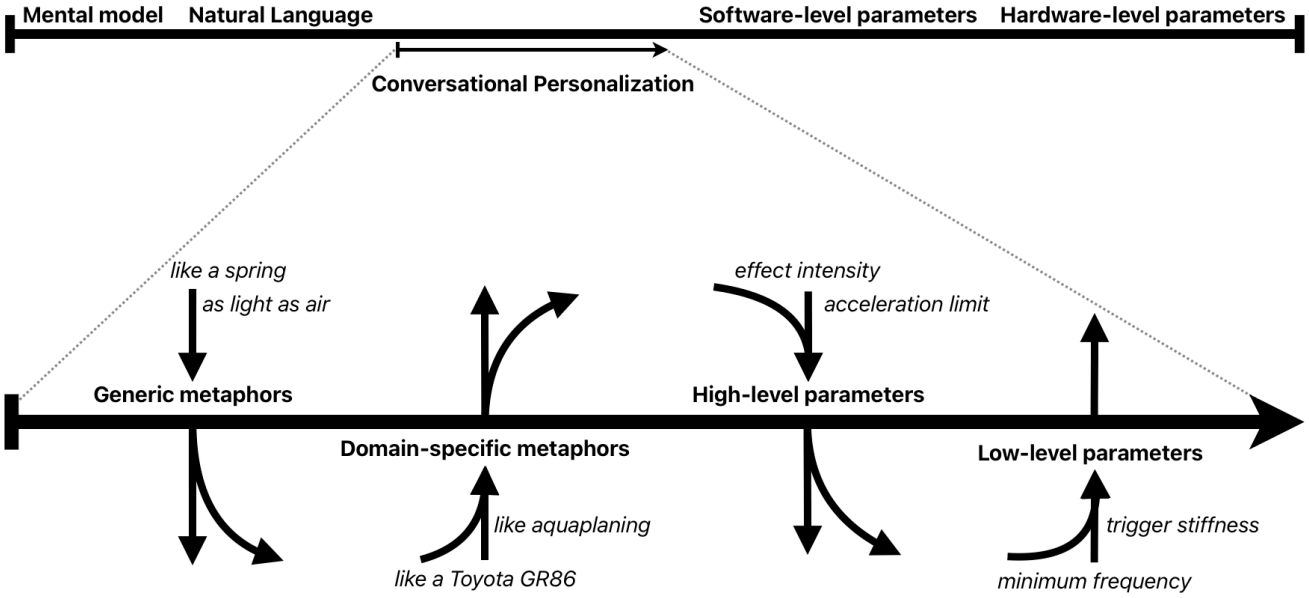


Figure 11: The gap between mental models of preference and low-level engineering parameters. LLMs, with their ability to understand vague user language, help bridge this gap, revealing an archetypal few-shot, unidirectional process underlying the personalization of haptic feedback. The short vertical and curved arrows represent users' ability to freely enter and exit the process at different stages either converging on a desirable configuration (shown by the straight arrows exiting the process completely) or honing in on lower levels of abstraction (shown by curved arrows continuing further down the axis of abstraction).

LLMs can effectively function as UI transducers for haptic feedback, bridging the gap between natural language descriptions of preferences and software-level haptic parameters (Figure 11).

8.2 Conversational personalization may be more usable than a slider-based interface

As observed through the SUS scores, users rated TacTalk as significantly more usable than ForzaDSX. Users noted that TacTalk was easier to use as it did not require users to learn system-specific language. Moreover, users highlighted not having to switch contexts with TacTalk (unlike ForzaDSX) as something that improved ease-of-use. The NASA-TLX responses align with this, with users reporting lower mental demand, perceived effort and perceived frustration, and higher perceived success when using TacTalk compared to ForzaDSX. Thus, we infer that when personalizing real-time, interactive applications, conversational interactions may help reduce cognitive load.

8.3 Preference context may improve alignment with user preferences

A key feature in TacTalk is the user preference profile keeping track of specific aspects of the haptic experience mentioned by the user, especially if they differ from TacTalk's base 'knowledge'. As seen in the results of the Car-Classification task, 6 of 11 participants correctly identified all three vehicle categories without profile adjustments. After profile adjustments, all 6 of these participants still correctly identified all three vehicle categories, and 3 of the others had higher accuracy (all of whom correctly identified all three vehicle categories), while 2 participants performed the same in both conditions. Overall, TacTalk seems to produce output that can be interpreted by some users using the default "world" model and others by adding profiles. Future quantitative-leaning studies could study this in more detail with a higher sample size and more complex task to reduce the risk of saturation.

8.4 Conversational interfaces should use signposting to make users' lives easier

As seen in our thematic analysis, users often look to their real-world knowledge to inform their preferences for haptic feedback in realistic-looking settings. This may result in confusion due to haptic modality differences. For example a controller's joysticks do not render force feedback, while users may expect a centering force on the steering wheel. For visual interfaces, using example libraries can prompt creative user-driven exploration of haptic feedback [90, 98].

In the context of voice-based interfaces, it is known that interactions range from the use of natural language to humming and other utterances [80, 81]. Voice-based interfaces for haptic feedback like Voodle [60] also offer the ability to communicate intended haptic feedback using the tone and rhythm of utterances. However, people have trouble gauging the capabilities of voice-based systems unless explicitly shown examples. Signposting a TacTalk-like system's capabilities may, for example, involve visual cues (e.g., diegetic tire smoke or other visual elements indicating trigger vibrations when

braking) drawing from users' tendency to notice in-game track-/environmental/telemetry factors. The Google Assistant illustrates examples through advertisements and allowing users to ask for example queries. In such cases, it is important to ensure the system's responses set realistic expectations - saying *'I can change your haptic experience'* is broad and vague, whereas examples like *'I can make your brake pedal shake less if that feels better'* or *'I can make the car feel less sluggish when driving on sand'* ground the system's abilities [72]. Another strategy may involve onboarding tutorials. For TacTalk, no tutorials were conducted, since the goal of the study was to study users' unbiased personalization strategies. Even so, two users (P6, P9) used TacTalk for suggestions through mimicking certain cars. Another user (P5) attempted explorations using different types of vocabulary and metaphors like *'light as air'* and *'like a spring'*. However, most users did not naturally consider asking for recommendations. In general, with conversational interfaces, users tend to trust the system more when voice-based feedback is also included [14]; such outputs may thus be used as another way to nudge the user towards exploring system capabilties.

8.5 Interfaces enabling personalization of haptic feedback should employ narrative framing effectively

In conjunction with signposting system capabilities, the narrative of the virtual experience is an important consideration. As mentioned in section 7, users employ metaphors when describing their desired haptic feedback. However, domain-specific metaphors are much more commonly used than generic vocabulary. Thus, an interface enabling personalization should focus on domain-specific narrative framing, as it can impact how users think about in-game events and their corresponding haptic feedback [13]. In the context of a racing game, for instance, it would make sense for a conversational interface to act as a race engineer, talking about in-game factors when describing changes made to the haptic feedback. Responding with *'I've tried increasing your grip loss sensitivity, try using the e-brake and you should notice the throttle feel more slippery'* not only describes the changes effected by the system, but also frames the explanation in a manner that the user can easily understand and evaluate.

8.6 Conversational interfaces can be a valuable learning tool

Domain-specific language and jargon often emerges in domains like motorsport to describe specific concepts. For example, picking up marbles, trail-braking and aquaplaning are terms that motorsport fans may know, but newcomers are unlikely to have heard, just as sommeliers use specialized vocabulary when describing wines. Prior work has explored the use of conversational interfaces for education [43], especially learning by teaching such interfaces [48]. However, such interfaces may also be used to actively teach users and help familiarize them with different aspects of an experience:

- How haptic feedback is rendered on their specific device
- How haptic feedback relates to in-game events and environmental factors

For example, if a user is uncertain about their desired haptic feedback, a conversational interface can help them learn about the parameters behind the haptic rendering and subsequently help identify what they would like to change. This may be direct, with the interface referring to the controller and its components. However, narrative framing may help too - using purely narrative-specific language to communicate with users can make it easier for them to focus on the virtual experience without switching contexts for personalization. This may positively affect user involvement/immersion, and may even extend to individuals' perception of realism. Since TacTalk's capabilities are influenced by the language of the system prompt, what users learn to perceive as realistic may reflect these biases too. This results in both the user and system being influenced by each others' biases, making the phrasing of parameter explanations important. For instance, we had to modify the explanation for the *acceleration limit* parameter since an earlier phrasing resulted in nonsensical outputs.

8.7 Conversational personalization seems to be a few-shot, unidirectional process

Prior work on HX highlights the importance of personalization, but end-users' modality-specific approaches towards personalization are poorly understood. Seifi [87] defined different approaches towards enabling end-user customization, namely choosing, tuning and chaining, but did not study the processes underlying personalization purely through language. In our study, we found that personalization takes the form of a unidirectional process, illustrated in Figure 11. This process has different entry and exit points, since different users begin with different notions of how in-depth they want to be and may end early if they achieve their preferred configuration. However, a clear unidirectional structure was observed, despite TacTalk supporting queries in any order.

More specifically, personalization commences from generic ideas of the intended experience. This is where metaphors are most valuable. In such situations, users expect example libraries or other suggestion mechanisms to better understand their preferences and the interface's capabilities. Asking TacTalk to mimic an existing vehicle or using other generic metaphors are examples of this. Moving downstream, we see that users hone in on specific aspects of the parameter space, attempting to fine-tune them. In our study, this involved queries directly addressing parameters exposed by ForzaDSX, such as frequencies. This process is iterative, and is often where users spend the most time. Overall, these findings align with prior works exploring personalization of vibrotactile effects [86]. Interestingly, despite being informed they could navigate to previous settings, none of the participants did so. Notably, the phenomenon of reverting to previous configurations differs from query-repair, which is common in users of voice interfaces [107]. Our findings also connect to Lakoff's idea that metaphors are inherently derived from our perception and sensorimotor experiences in the real world [47] - for instance, our experiences of spatial relationships in the real world help us agree that 'up' corresponds to an increase and 'down' to a decrease. We saw this carry over to the task of driving a car in a game, but different users have their own mental models of desirable haptic feedback given differences in their levels of experience with driving both virtually and in real life. TacTalk essentially

attempts to find the closest configuration to individual mental models, and retains individual preferences over time. We distinguish TacTalk's use for navigating haptic feedback settings from prior work on AI-powered co-creative systems [50], where a model of agentive flow was proposed - TacTalk was never viewed as a collaborator, although some interactions did support playful discovery. We also acknowledge that our findings are derived from a controlled lab study, and thus any limitations of human-participant studies (sample demographics, remuneration etc.) may have influenced our results.

8.8 Limitations

Our study focused on rich qualitative analysis testing viability of LLM-based approaches for personalized haptic feedback and users' approach towards personalizing haptics through conversation. As such, to reduce priming, we did not counterbalance interface order. Future studies may remove order effect confounds to compare conversational and visual slider-based interfaces. Our screening questionnaire asked participants about their familiarity with motorsport-related vocabulary, such as aquaplaning. This could have possibly primed participants to use similar terms in their queries. For example, we observed the use of the terms 'throttle', 'traction', and 'drifting', although other terms such as 'aquaplaning', 'brake bias', 'RPM' and 'trail braking' were not used in queries. Our participant sample had limited gender diversity, with 10 male and one female participants. This may have been influenced by the study task context (playing a racing game) and recruitment methods (advertising via university mailing lists). Prior surveys have found that only 6% of gamers in the racing genre were female [108] an issue prevalent in the broader racing landscape [31, 46]. Our focus on collecting a sample with diverse backgrounds with racing games and motorsport may have further limited our demographic diversity. While we considered gender diversity while developing TacTalk, recruiting 2 male and 2 female participants for pilot sessions, future studies may adopt techniques like purposive sampling to increase diversity of samples along dimensions like gender. Larger sample sizes for more quantitative studies may also have a wider range of individuals represented.

9 CONCLUSION

In this paper, we investigated the personalization of haptics through conversation. We presented a conversational interface, TacTalk, that enables customization of haptic feedback in real time (section 3). Technical evaluations and a user study were conducted to validate TacTalk's functionality and usability. We found that TacTalk is capable of producing consistent mappings of haptic feedback configurations, has higher usability and lower cognitive load than an existing alternative visual interface, and found an archetypal unidirectional process underlying personalization.

Future investigations into voice-based interactions for haptic feedback customization could examine different tasks, devices and interaction modalities to further understand whether the unidirectional nature of conversational personalization can be seen outside of a voice assistant and whether a single deployment of TacTalk is generalizable across multiple platforms and experiences. We believe TacTalk could scale across devices and experiences , making use of LLMs' pretrained knowledge base to reduce the amount of additional finetuning data required. Examples of other relevant contexts for future work include haptic feedback-enabled assistive technologies for people with motor control impairments. Further, longitudinal studies may examine how shifting user preferences over long-term interactions can be tracked, essentially collecting per-user prompt datasets and re-aligning haptic parameter mappings. Parallel to this, future studies may also explore alternative architectures to directly manipulate haptic feedback parameters in real time, based on relevant large scale training datasets.

Acknowledgments

We thank the University of Waterloo HCI community, especially the Haptic Experience Lab and Games Institute, for their continued guidance and feedback. We acknowledge the support of the Canada Foundation for Innovation (CFI), Ontario Research Foundation (ORF), and the Mitacs Globalink Graduate Fellowship. This work was partly supported by an unrestricted gift from Google intended to support "Software Tools for Rendering Haptics and Assessing Quality at Scale". LLM hosting and inferencing was supported by Microsoft Azure credits as part of a larger allocation to the University of Waterloo.

References

[1] Sony Interactive Entertainment LLC. 2020. DualSense wireless controller | The innovative new controller for PS5. https://www.playstation.com/en-ca/accessories/dualsense-wireless-controller/

[2] Xbox Game Studios. 2021. Forza Horizon 5: Play with Xbox Game Pass | Xbox. https://www.xbox.com/en-CA/games/forza-horizon-5

[3] Draw a UI. 2023. draw-a-ui. https://www.draw-a-ui.com/

[4] Laura Aina and Tal Linzen. 2021. The Language Model Understood the Prompt was Ambiguous: Probing Syntactic Uncertainty Through Generation. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, Jasmijn Bastings, Yonatan Belinkov, Emmanuel Dupoux, Mario Giulianelli, Dieuwke Hupkes, Yuval Pinter, and Hassan Sajjad (Eds.). Association for Computational Linguistics, Punta Cana, Dominican Republic, 42–57. https://doi.org/10.18653/v1/2021.blackboxnlp-1.4

[5] Ahmed Anwar, Tianzheng Shi, and Oliver Schneider. 2023. Factors of Haptic Experience across Multiple Haptic Modalities. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3544548.3581514

[6] Ian Arawjo, Chelse Swoopes, Priyan Vaithilingam, Martin Wattenberg, and Elena L. Glassman. 2024. ChainForge: A Visual Toolkit for Prompt Engineering and LLM Hypothesis Testing. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–18. https://doi.org/10.1145/3613904.3642016

[7] Javier A. Bargas-Avila and Kasper Hornbæk. 2011. Old wine in new bottles or novel challenges: a critical analysis of empirical studies of user experience. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. Association for Computing Machinery, New York, NY, USA, 2689–2698. https://doi.org/10.1145/1978942.1979336

[8] Michael S. Bernstein, Joon Sung Park, Meredith Ringel Morris, Saleema Amershi, Lydia B Chilton, and Mitchell L. Gordon. 2023. Architecting Novel Interactions with Generative AI Models. In *Adjunct Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 1–3. https://doi.org/10.1145/3586182.3617431

[9] Jeffrey R. Blum, Ilja Frissen, and Jeremy R. Cooperstock. 2015. Improving Haptic Feedback on Wearable Devices through Accelerometer Measurements. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology* (Charlotte, NC, USA) *(UIST '15)*. Association for Computing Machinery, New York, NY, USA, 31–36. https://doi.org/10.1145/2807442.2807474

[10] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John

Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2022. On the Opportunities and Risks of Foundation Models. https://doi.org/10.48550/arXiv.2108.07258 arXiv:2108.07258 [cs].

[11] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (Jan. 2006), 77–101. https://doi.org/10.1191/1478088706qp063oa Publisher: Routledge _eprint: https://www.tandfonline.com/doi/pdf/10.1191/1478088706qp063oa.

[12] John Brooke. 1995. SUS: A quick and dirty usability scale. *Usability Eval. Ind.* 189 (Nov. 1995), 1–7.

[13] Paul Bucci, Lotus Zhang, Xi Laura Cang, and Karon E. MacLean. 2018. Is it Happy? Behavioural and Narrative Frame Complexity Impact Perceptions of a Simple Furry Robot's Emotions. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3173574.3174083

[14] Ramón Burri. 2018. *Improving user trust towards conversational chatbot interfaces with voice output*. KTH Diva Portal, Stockholm, Sweden. https://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-240585

[15] Youngjin (YJ) Chae and Thomas Davidson. 2023. Large Language Models for Text Classification: From Zero-Shot Learning to Fine-Tuning. https://doi.org/10.31235/osf.io/sthwk

[16] Priyanka Chandel, Devanuj, and Pankaj Doke. 2013. A comparative study of voice and graphical user interfaces with respect to literacy levels. In *Proceedings of the 3rd ACM Symposium on Computing for Development (ACM DEV '13)*. Association for Computing Machinery, New York, NY, USA, 1–2. https://doi.org/10.1145/2442882.2442921

[17] Roger W. Cholewiak and Amy A. Collins. 1997. Individual differences in the vibrotactile perception of a "simple" pattern set. *Perception & Psychophysics* 59, 6 (Jan. 1997), 850–866. https://doi.org/10.3758/BF03205503

[18] Ben Clark, Oliver S. Schneider, Karon E. MacLean, and Hong Z. Tan. 2017. Predictable and distinguishable morphing of vibrotactile rhythm. In *2017 IEEE World Haptics Conference (WHC)*. IEEE, Munich, Germany, 84–89. https://doi.org/10.1109/WHC.2017.7989881

[19] cosmii02, zchenak, and patmagauran. 2024. cosmii02/RacingDSX. https://github.com/cosmii02/RacingDSX original-date: 2022-09-26T15:22:42Z.

[20] Fernanda De La Torre, Cathy Mengying Fang, Han Huang, Andrzej Banburski-Fahey, Judith Amores Fernandez, and Jaron Lanier. 2024. LLMR: Real-time Prompting of Interactive Worlds using Large Language Models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–22. https://doi.org/10.1145/3613904.3642579

[21] Donald Degraen, Bruno Fruchard, Frederik Smolders, Emmanouil Potetsianakis, Seref Güngör, Antonio Krüger, and Jürgen Steimle. 2021. Weirding haptics: In-situ prototyping of vibrotactile feedback in virtual reality through vocalization. In *The 34th Annual ACM symposium on user interface software and technology*. 936–953.

[22] Stavros Demetriadis and Yannis Dimitriadis. 2023. Conversational Agents and Language Models that Learn from Human Dialogues to Support Design Thinking. In *Augmented Intelligence and Intelligent Tutoring Systems*, Claude Frasson, Phivos Mylonas, and Christos Troussas (Eds.). Springer Nature Switzerland, Cham, 691–700. https://doi.org/10.1007/978-3-031-32883-1_60

[23] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024. A Survey on In-context Learning. https://doi.org/10.48550/arXiv.2301.00234 arXiv:2301.00234 [cs].

[24] DualSenseX. 2023. Game Mods with DualSenseX. https://dualsensex.com/game-mods/

[25] Aparna Garimella, Rada Mihalcea, and Akhash Amarnath. 2022. Demographic-Aware Language Model Fine-tuning as a Bias Mitigation Technique. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, Yulan He, Heng Ji, Sujian Li, Yang Liu, and Chua-Hui Chang (Eds.). Association for Computational Linguistics, Online only, 311–319. https://aclanthology.org/2022.aacl-short.38

[26] Louie Giray. 2023. Prompt Engineering with ChatGPT: A Guide for Academic Writers. *Annals of Biomedical Engineering* 51, 12 (Dec. 2023), 2629–2633. https://doi.org/10.1007/s10439-023-03272-4

[27] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Advances in Psychology*, Peter A. Hancock and Najmedin Meshkati (Eds.). Human Mental Workload, Vol. 52. North-Holland, 139–183. https://doi.org/10.1016/S0166-4115(08)62386-9

[28] Marc Hassenzahl. 2018. The thing and I: understanding the relationship between user and product. *Funology 2: from usability to enjoyment* (2018), 301–313.

[29] Qianyu He, Jie Zeng, Wenhao Huang, Lina Chen, Jin Xiao, Qianxi He, Xunzhe Zhou, Jiaqing Liang, and Yanghua Xiao. 2024. Can Large Language Models Understand Real-World Complex Instructions? *Proceedings of the AAAI Conference on Artificial Intelligence* 38, 16 (March 2024), 18188–18196. https://doi.org/10.1609/aaai.v38i16.29777 Number: 16.

[30] Mark Holliins, Richard Faldowski, Suman Rao, and Forrest Young. 1993. Perceptual dimensions of tactile surface texture: A multidimensional scaling analysis. *Perception & Psychophysics* 54, 6 (Nov. 1993), 697–705. https://doi.org/10.3758/BF03211795

[31] Olivia R. Howe. 2022. Hitting the barriers – Women in Formula 1 and W series racing. *European Journal of Women's Studies* 29, 3 (Aug. 2022), 454–469. https://doi.org/10.1177/13505068221094204 Publisher: SAGE Publications Ltd.

[32] Jamil Hussain, Anees Ul Hassan, Hafiz Syed Muhammad Bilal, Rahman Ali, Muhammad Afzal, Shujaat Hussain, Jaehun Bang, Oresti Banos, and Sungyoung Lee. 2018. Model-based adaptive user interface based on context and user experience evaluation. *Journal on Multimodal User Interfaces* 12, 1 (March 2018), 1–16. https://doi.org/10.1007/s12193-018-0258-2

[33] Inwook Hwang, K. E. MacLean, M. Brehmer, J. Hendy, A. Sotirakopoulos, and Seungmoon Choi. 2011. The haptic crayola effect: Exploring the role of naming in learning haptic stimuli. In *2011 IEEE World Haptics Conference*. IEEE, Istanbul, 385–390. https://doi.org/10.1109/WHC.2011.5945517

[34] Ali Israr, Siyan Zhao, Kaitlyn Schwalje, Roberta Klatzky, and Jill Lehman. 2014. Feel Effects: Enriching Storytelling with Haptic Feedback. *ACM Trans. Appl. Percept.* 11, 3, Article 11 (sep 2014), 17 pages. https://doi.org/10.1145/2641570

[35] Cathrine V. Jansson-Boyd. 2011. Touch matters: exploring the relationship between consumption and tactile interaction. *Social Semiotics* 21, 4 (Sept. 2011), 531–546. https://doi.org/10.1080/10350330.2011.591996 Publisher: Routledge _eprint: https://doi.org/10.1080/10350330.2011.591996.

[36] Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of Experts. https://doi.org/10.48550/arXiv.2401.04088 arXiv:2401.04088 [cs].

[37] Erik Jones and Jacob Steinhardt. 2022. Capturing Failures of Large Language Models via Human Cognitive Biases. *Advances in Neural Information Processing Systems* 35 (Dec. 2022), 11785–11799. https://proceedings.neurips.cc/paper_files/paper/2022/hash/4d13b2d99519c5415661dad44ab7edcd-Abstract-Conference.html

[38] Miku Kaneda, Masahiro Takeuchi, Seitaro Kaneko, and Hiroyuki Kajimoto. 2022. Relationship between onomatopoeia and physical properties when pressing a soft object. In *2022 IEEE Haptics Symposium (HAPTICS)*. IEEE, New York, USA, 1–6. https://doi.org/10.1109/HAPTICS52432.2022.9765623 ISSN: 2324-7355.

[39] Ioannis Kangas, Maud Schwoerer, and Lucas J Bernardi. 2021. Recommender Systems for Personalized User Experience: Lessons learned at Booking.com. In *Proceedings of the 15th ACM Conference on Recommender Systems (RecSys '21)*. Association for Computing Machinery, New York, NY, USA, 583–586. https://doi.org/10.1145/3460231.3474611

[40] Akifumi Kawai and Yoshihiro Tanaka. 2022. Individual Differences in Skin Vibration Characteristics and Vibrotactile Sensitivity at Fingertip. In *2022 IEEE Haptics Symposium (HAPTICS)*. IEEE, New York, USA, 1–6. https://doi.org/10.1109/HAPTICS52432.2022.9765612 ISSN: 2324-7355.

[41] Erin Kim and Oliver Schneider. 2020. Defining Haptic Experience: Foundations for Understanding, Communicating, and Evaluating HX. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376280

[42] Jaehyung Kim and Yiming Yang. 2024. Few-shot Personalization of LLMs with Mis-aligned Responses. https://doi.org/10.48550/arXiv.2406.18678 arXiv:2406.18678 [cs].

[43] Toshihiro Kita, Chikako Nagaoka, Naoshi Hiraoka, and Martin Dougiamas. 2019. Implementation of Voice User Interfaces to Enhance Users' Activities on Moodle. In *2019 4th International Conference on Information Technology (InCIT)*. IEEE, New York, USA, 104–107. https://doi.org/10.1109/INCIT.2019.8912086

[44] Dmitry Kobak and George C. Linderman. 2019. UMAP does not preserve global structure any better than t-SNE when using the same initialization. https://doi.org/10.1101/2019.12.19.877522 Pages: 2019.12.19.877522 Section: Contradictory Results.

[45] Dmitry Kobak and George C. Linderman. 2021. Initialization is critical for preserving global data structure in both t-SNE and UMAP. *Nature Biotechnology* 39, 2 (Feb. 2021), 156–157. https://doi.org/10.1038/s41587-020-00809-z Publisher: Nature Publishing Group.

[46] Jill Kochanek, Megan Davis, Karl Erickson, and David Ferguson. 2021. More than "just a driver": A study of professional women racecar drivers' agency in motorsport. *Psychology of Sport and Exercise* 52 (Jan. 2021), 101838. https://doi.org/10.1016/j.psychsport.2020.101838

[47] George Lakoff and Mark Johnson. 2008. *Metaphors we live by*. University of Chicago press.

[48] Edith Law, Parastoo Baghaei Ravari, Nalin Chhibber, Dana Kulic, Stephanie Lin, Kevin D. Pantasdo, Jessy Ceha, Sangho Suh, and Nicole Dillen. 2020. Curiosity Notebook: A Platform for Learning by Teaching Conversational Agents. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems (CHI EA '20)*. Association for Computing Machinery, New York, NY, USA, 1–9. https://doi.org/10.1145/3334480.3382783

[49] Effie Lai-Chong Law, Virpi Roto, Marc Hassenzahl, Arnold P.O.S. Vermeeren, and Joke Kort. 2009. Understanding, scoping and defining user experience: a survey approach. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '09)*. Association for Computing Machinery, New York, NY, USA, 719–728. https://doi.org/10.1145/1518701.1518813

[50] Tomas Lawton, Kazjon Grace, and Francisco J Ibarrola. 2023. When is a Tool a Tool? User Perceptions of System Agency in Human–AI Co-Creative Drawing. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference* (Pittsburgh, PA, USA) *(DIS '23)*. Association for Computing Machinery, New York, NY, USA, 1978–1996. https://doi.org/10.1145/3563657.3595977

[51] Jiwan Lee, Junwoo Kim, Jeonggoo Kang, Eunsoo Jo, Dong Chul Park, and Seungmoon Choi. 2024. Telemetry-Based Haptic Rendering for Racing Game Experience Improvement. *IEEE Transactions on Haptics* 17, 1 (Jan. 2024), 72–79. https://doi.org/10.1109/TOH.2024.3357885 Conference Name: IEEE Transactions on Haptics.

[52] Bingxu Li and Gregory J. Gerling. 2023. An individual's skin stiffness predicts their tactile discrimination of compliance. *The Journal of Physiology* 601, 24 (2023), 5777–5794. https://doi.org/10.1113/JP285271 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1113/JP285271.

[53] Chen Cecilia Liu, Fajri Koto, Timothy Baldwin, and Iryna Gurevych. 2024. Are Multilingual LLMs Culturally-Diverse Reasoners? An Investigation into Multicultural Proverbs and Sayings. https://doi.org/10.48550/arXiv.2309.08591 arXiv:2309.08591 [cs].

[54] Yang Liu, Haiwei Dong, and Abdulmotaleb El Saddik. 2025. Leveraging LLMs to Create a Haptic Devices' Recommendation System. *arXiv preprint arXiv:2501.12573* (2025).

[55] J. Lo¨fvenberg and R. S. Johansson. 1984. Regional differences and interindividual variability in sensitivity to vibration in the glabrous skin of the human hand. *Brain Research* 301, 1 (May 1984), 65–72. https://doi.org/10.1016/0006-8993(84)90403-7

[56] Shihan Lu, Mianlun Zheng, Matthew C. Fontaine, Stefanos Nikolaidis, and Heather Culbertson. 2022. Preference-Driven Texture Modeling Through Interactive Generation and Search. *IEEE Transactions on Haptics* 15, 3 (July 2022), 508–520. https://doi.org/10.1109/TOH.2022.3173935 Conference Name: IEEE Transactions on Haptics.

[57] Michal Luria, Guy Hoffman, and Oren Zuckerman. 2017. Comparing Social Robot, Screen and Voice Interfaces for Smart-Home Control. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. Association for Computing Machinery, New York, NY, USA, 580–628. https://doi.org/10.1145/3025453.3025786

[58] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research* 9, 86 (2008), 2579–2605. http://jmlr.org/papers/v9/vandermaaten08a.html

[59] Emanuela Maggioni, Erika Agostinelli, and Marianna Obrist. 2017. Measuring the added value of haptic feedback. In *2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, New York, USA, 1–6. https://doi.org/10.1109/QoMEX.2017.7965670 ISSN: 2472-7814.

[60] David Marino, Paul Bucci, Oliver S. Schneider, and Karon E. MacLean. 2017. Voodle: Vocal Doodling to Sketch Affective Robot Motion. In *Proceedings of the 2017 Conference on Designing Interactive Systems (DIS '17)*. Association for Computing Machinery, New York, NY, USA, 753–765. https://doi.org/10.1145/3064663.3064668

[61] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. 2018. UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software* 3, 29 (Sept. 2018), 861. https://doi.org/10.21105/joss.00861

[62] Mohammad Movahedi and Juyeong Choi. 2024. The Crossroads of LLM and Traffic Control: A Study on Large Language Models in Adaptive Traffic Signal Control. *IEEE Transactions on Intelligent Transportation Systems* (2024), 1–16. https://doi.org/10.1109/TITS.2024.3498735

[63] Christine Murad, Cosmin Munteanu, Benjamin R. Cowan, Leigh Clark, Martin Porcheron, Heloisa Candello, Stephan Schlögl, Matthew P. Aylett, Jaisie Sin, Robert J. Moore, Grace Hughes, and Andrew Ku. 2021. Let's Talk About CUIs: Putting Conversational User Interface Design Into Practice. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI EA '21)*. Association for Computing Machinery, New York, NY, USA, Article 98, 6 pages. https://doi.org/10.1145/3411763.3441336

[64] Roberto Navigli, Simone Conia, and Björn Ross. 2023. Biases in Large Language Models: Origins, Inventory, and Discussion. *Journal of Data and Information Quality* 15, 2 (June 2023), 10:1–10:21. https://doi.org/10.1145/3597307

[65] Marianna Obrist, Sue Ann Seah, and Sriram Subramanian. 2013. Talking about tactile experiences. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. Association for Computing Machinery, New York, NY, USA, 1659–1668. https://doi.org/10.1145/2470654.2466220

[66] Shogo Okamoto, Hikaru Nagano, and Yoji Yamada. 2013. Psychophysical Dimensions of Tactile Perception of Textures. *IEEE Transactions on Haptics* 6, 1 (2013), 81–93. https://doi.org/10.1109/TOH.2012.32 Conference Name: IEEE Transactions on Haptics.

[67] OpenAI. 2024. Hello GPT-4o. https://openai.com/index/hello-gpt-4o/

[68] OpenAI. 2024. OpenAI Platform. https://platform.openai.com

[69] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, C. J. Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. GPT-4 Technical Report. https://doi.org/10.48550/arXiv.2303.08774 arXiv:2303.08774 [cs].

[70] Christos Papakostas, Christos Troussas, Akrivi Krouska, and Cleo Sgouropoulou. 2021. Measuring User Experience, Usability and Interactivity of a Personalized Mobile Augmented Reality Training System. *Sensors* 21, 11 (Jan. 2021), 3888. https://doi.org/10.3390/s21113888 Number: 11 Publisher: Multidisciplinary Digital Publishing Institute.

[71] Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. http://arxiv.org/abs/2304.03442 arXiv:2304.03442 [cs].

[72] Cathy Pearl. 2016. *Designing Voice User Interfaces: Principles of Conversational Experiences.* "O'Reilly Media, Inc.", 1005 Gravenstein Highway North, Sebastopol, CA. Google-Books-ID: MmnEDQAAQBAJ.

[73] Joann Peck and Terry L. Childers. 2003. Individual Differences in Haptic Information Processing: The "Need for Touch" Scale. *Journal of Consumer Research* 30, 3 (Dec. 2003), 430–442. https://doi.org/10.1086/378619

[74] Raul Puri and Bryan Catanzaro. 2019. Zero-shot Text Classification With Generative Language Models. https://doi.org/10.48550/arXiv.1912.10165 arXiv:1912.10165 [cs].

[75] Rebecca Qian, Candace Ross, Jude Fernandes, Eric Michael Smith, Douwe Kiela, and Adina Williams. 2022. Perturbation Augmentation for Fairer NLP. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 9496–9521. https://doi.org/10.18653/v1/2022.emnlp-main.646

[76] Beatrice Rammstedt and Oliver P. John. 2020. *Big Five Inventory.* Springer International Publishing, Cham, 469–471. https://doi.org/10.1007/978-3-319-24612-3_445

[77] Arpit Rana, Scott Sanner, Mohamed Reda Bouadjenek, Ron Dicarlantonio, and Gary Farmaner. 2023. User Experience and The Role of Personalization in Critiquing-Based Conversational Recommendation. *ACM Transactions on the Web* (May 2023), 111:1–111:21. https://doi.org/10.1145/3597499 Just Accepted.

[78] Leon Reicherts, Yvonne Rogers, Licia Capra, Ethan Wood, Tu Dinh Duong, and Neil Sebire. 2022. It's Good to Talk: A Comparison of Using Voice Versus Screen-Based Interactions for Agent-Assisted Tasks. *ACM Trans. Comput.-Hum. Interact.* 29, 3 (Jan. 2022), 25:1–25:41. https://doi.org/10.1145/3484221

[79] Qiaoqiao Ren and Tony Belpaeme. 2025. Touched by ChatGPT: Using an LLM to Drive Affective Tactile Interaction. *arXiv preprint arXiv:2501.07224* (2025).

[80] Google Research. 2018. Google's Next Generation Music Recognition. http://research.google/blog/googles-next-generation-music-recognition/

[81] Google Research. 2022. Look and Talk: Natural Conversations with Google Assistant. http://research.google/blog/look-and-talk-natural-conversations-with-google-assistant/

[82] Tom Roy, Yann Glémarec, Gurvan Lécuyer, Quentin Galvane, Philippe Guillotel, and Ferran Argelaguet. 2024. Towards End-User Customization of Haptic Experiences. In *EuroHaptics 2024 - 14th International Conference on Human Haptic Sensing and Touch-Enabled Computer Applications.* IEEE, New York, USA, 1. https://inria.hal.science/hal-04611300

[83] Suji Sathiyamurthy, Melody Lui, Erin Kim, and Oliver Schneider. 2021. Measuring Haptic Experience: Elaborating the HX model with scale development. In *2021 IEEE World Haptics Conference (WHC).* IEEE, Montreal, QC, Canada, 979–984. https://doi.org/10.1109/WHC49131.2021.9517220

[84] Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. 2023. Are Emergent Abilities of Large Language Models a Mirage? *Advances in Neural Information Processing Systems* 36 (Dec. 2023), 55565–55581. https://proceedings.neurips.cc/paper_files/paper/2023/hash/adc98a266f45005c403b8311ca7e8bd7-Abstract-Conference.html

[85] Oliver Schneider, Karon MacLean, Colin Swindells, and Kellogg Booth. 2017. Haptic experience design: What hapticians do and where they need help. *International Journal of Human-Computer Studies* 107 (Nov. 2017), 5–21. https://doi.org/10.1016/j.ijhcs.2017.04.004

[86] Oliver S. Schneider and Karon E. MacLean. 2016. Studying design process and example use with Macaron, a web-based vibrotactile effect editor. In *2016 IEEE Haptics Symposium (HAPTICS).* IEEE, Philadelphia, PA, USA, 52–58. https://doi.org/10.1109/HAPTICS.2016.7463155

[87] Hasti Seifi. 2019. *Personalizing Haptics: From Individuals' Sense-Making Schemas to End-User Haptic Tools.* Springer International Publishing, Cham. https://doi.org/10.1007/978-3-030-11379-7

[88] Hasti Seifi, Chamila Anthonypillai, and Karon E. MacLean. 2014. End-user customization of affective tactile messages: A qualitative examination of tool parameters. In *2014 IEEE Haptics Symposium (HAPTICS).* IEEE, Houston, TX, USA, 251–256. https://doi.org/10.1109/HAPTICS.2014.6775463

[89] Hasti Seifi, Sean Chew, Antony James Nascè, William Edward Lowther, William Frier, and Kasper Hornbæk. 2023. Feellustrator: A Design Tool for Ultrasound Mid-Air Haptics. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23).* Association for Computing Machinery, New York, NY, USA, 1–16. https://doi.org/10.1145/3544548.3580728

[90] Hasti Seifi, Kailun Zhang, and Karon E. MacLean. 2015. VibViz: Organizing, visualizing and navigating vibration libraries. In *2015 IEEE World Haptics Conference (WHC).* IEEE, Evanston, IL, 254–259. https://doi.org/10.1109/WHC.2015.7177722

[91] Tanay Singhal and Oliver Schneider. 2021. Juicy Haptic Design: Vibrotactile Embellishments Can Improve Player Experience in Games. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21).* Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3411764.3445463

[92] Joseph C. Stevens. 1992. Aging and Spatial Acuity of Touch. *Journal of Gerontology* 47, 1 (Jan. 1992), P35–P40. https://doi.org/10.1093/geronj/47.1.P35

[93] Joseph C. Stevens and Kenneth K. Choo. 1996. Spatial Acuity of the Body Surface over the Life Span. *Somatosensory & Motor Research* 13, 2 (Jan. 1996), 153–166. https://doi.org/10.3109/08990229609051403 Publisher: Taylor & Francis _eprint: https://doi.org/10.3109/08990229609051403.

[94] Hao Sun, Zhexin Zhang, Fei Mi, Yasheng Wang, Wei Liu, Jianwei Cui, Bin Wang, Qun Liu, and Minlie Huang. 2023. MoralDial: A Framework to Train and Evaluate Moral Dialogue Systems via Moral Discussions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 2213–2230. https://doi.org/10.18653/v1/2023.acl-long.123

[95] Liyan Tang, Zhaoyi Sun, Betina Idnay, Jordan G. Nestor, Ali Soroush, Pierre A. Elias, Ziyang Xu, Ying Ding, Greg Durrett, Justin F. Rousseau, Chunhua Weng, and Yifan Peng. 2023. Evaluating large language models on medical evidence summarization. *npj Digital Medicine* 6, 1 (Aug. 2023), 1–8. https://doi.org/10.1038/s41746-023-00896-7 Publisher: Nature Publishing Group.

[96] Ryan Teehan, Miruna Clinciu, Oleg Serikov, Eliza Szczechla, Natasha Seelam, Shachar Mirkin, and Aaron Gokaslan. 2022. Emergent Structures and Training Dynamics in Large Language Models. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, Angela Fan, Suzana Ilic, Thomas Wolf, and Matthias Gallé (Eds.). Association for Computational Linguistics, virtual+Dublin, 146–159. https://doi.org/10.18653/v1/2022.bigscience-1.11

[97] Himanshu Thakur, Atishay Jain, Praneetha Vaddamanu, Paul Pu Liang, and Louis-Philippe Morency. 2023. Language Models Get a Gender Makeover: Mitigating Gender Bias with Few-Shot Data Interventions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 340–351. https://doi.org/10.18653/v1/2023.acl-short.30

[98] Karthikan Theivendran, Andy Wu, William Frier, and Oliver Schneider. 2024. RecHap: An Interactive Recommender System For Navigating a Large Number of Mid-Air Haptic Designs. *IEEE Transactions on Haptics* 17, 2 (2024), 1–12. https://doi.org/10.1109/TOH.2023.3276812

[99] Gustavo F. Tondello, Karina Arrambide, Giovanni Ribeiro, Andrew Jian-lan Cen, and Lennart E. Nacke. 2019. "I Don't Fit into a Single Type": A Trait Model and Scale of Game Playing Preferences. In *Human-Computer Interaction – INTERACT 2019*, David Lamas, Fernando Loizides, Lennart Nacke, Helen Petrie, Marco Winckler, and Panayiotis Zaphiris (Eds.). Springer International Publishing, Cham, 375–395.

[100] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. https://doi.org/10.48550/arXiv.2307.09288 arXiv:2307.09288 [cs].

[101] Megan Ung, Jing Xu, and Y-Lan Boureau. 2022. SaFeRDialogues: Taking Feedback Gracefully after Conversational Safety Failures. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, Dublin, Ireland, 6462–6481. https://doi.org/10.18653/v1/2022.acl-long.447

[102] Unity. 2024. Unity Real-Time Development Platform | 3D, 2D, VR & AR Engine. https://unity.com/

[103] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, Vol. 30. Curran Associates, Inc., 57 Morehouse Ln, Red Hook, NY 12571, United States, 1–11. https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[104] Lalit Venkatesan, Steven M. Barlow, and Douglas Kieweg. 2015. Age- and sex-related changes in vibrotactile sensitivity of hand and face in neurotypical adults. *Somatosensory & Motor Research* 32, 1 (Jan. 2015), 44–50. https://doi.org/10.3109/08990220.2014.958216 Publisher: Taylor & Francis _eprint: https://doi.org/10.3109/08990220.2014.958216.

[105] Kellie Webster, Xuezhi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, Ed Chi, and Slav Petrov. 2021. Measuring and Reducing Gendered Correlations in Pre-trained Models. https://doi.org/10.48550/arXiv.2010.06032 arXiv:2010.06032 [cs].

[106] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. https://doi.org/10.48550/arXiv.2201.11903 arXiv:2201.11903 [cs].

[107] Jason Wu, Karan Ahuja, Richard Li, Victor Chen, and Jeffrey Bigham. 2019. ScratchThat: Supporting Command-Agnostic Speech Repair in Voice-Driven Assistants. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 2 (June 2019), 1–17. https://doi.org/10.1145/3328934

[108] Nick Yee. 2017. Beyond 50/50: Breaking Down The Percentage of Female Gamers By Genre. https://quanticfoundry.com/2017/01/19/female-gamers-by-genre/

[109] Shiquan Zhang, Ying Ma, Le Fang, Hong Jia, Simon D'Alfonso, and Vassilis Kostakos. 2024. Enabling On-Device LLMs Personalization with Smartphone Sensing. https://doi.org/10.48550/arXiv.2407.04418 arXiv:2407.04418 [cs].

[110] Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. 2024. Benchmarking Large Language Models for News Summarization. *Transactions of the Association for Computational Linguistics* 12 (Jan. 2024), 39–57. https://doi.org/10.1162/tacl_a_00632

[111] Ying Zheng and John B. Morrell. 2012. Haptic actuator design parameters that influence affect and attention. In *2012 IEEE Haptics Symposium (HAPTICS)*. IEEE, Vancouver, BC, Canada, 463–470. https://doi.org/10.1109/HAPTIC.2012.6183832

[112] Mingyu Zong and Bhaskar Krishnamachari. 2023. Solving math word problems concerning systems of equations with GPT models. *Machine Learning with Applications* 14 (Dec. 2023), 100506. https://doi.org/10.1016/j.mlwa.2023.100506

10 Appendix

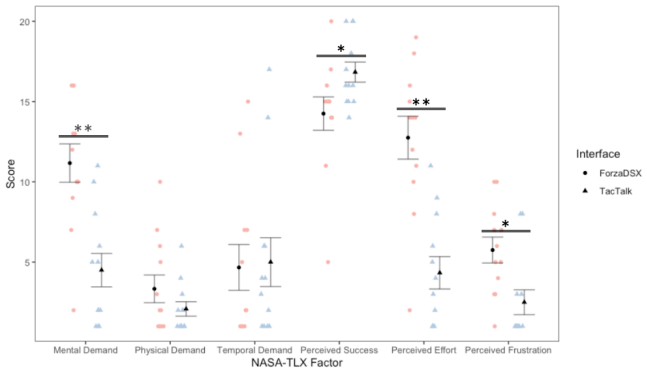


Figure 12: A scatter plot illustrating the NASA-TLX scores for TacTalk and the ForzaDSX visual interface, the mean scores for each NASA-TLX dimension for the two interfaces, and 95% confidence intervals. For *: $p < 0.05$, **: $p < 0.01$. Significantly lower Mental Demand, Perceived Effort and Perceived Success is observed, and significantly higher Perceived Success is observed when using TacTalk compared to ForzaDSX.