Exploring Human-AI Perception Alignment in Sensory Experiences: Do LLMs Understand Textile Hand?

Anonymous ACL submission

Abstract

Aligning LLMs behaviour with human intent is critical for future AI. An important yet often overlooked aspect of this alignment is the perceptual alignment. Perceptual modalities like touch are more multifaceted and nuanced compared to other sensory modalities such as vision. This work investigates how well LLMs align with human touch experiences. We created an interaction in which participants were given two textile samples to handle without seeing them and describe the differences between them to the LLM. Using these descriptions, the LLM attempted to identify the target textile by assessing similarity within its high-dimensional embedding space. Our results suggest that a degree of perceptual alignment exists, however varies significantly among different textile samples. Moreover, participants didn't perceive their textile experiences closely matched by the LLM predictions. We discuss possible sources of this alignment variance, and how better human-AI perceptual alignment can benefit future everyday tasks.

1 Introduction

Several studies in human-AI alignment research have discussed the imperative for AI models to "align" by having robustness, interpretability, controllability, and ethicality (Askell et al., 2021; Hendrycks et al., 2020; Dafoe et al., 2020, 2021). Those are important requirements especially as AI products and services are increasingly embedded in everyday life interactions, such as self-driving cars, smart home applications, and online shopping solutions (Elliott, 2019; Shneiderman, 2020).

A critical but often overlooked necessity to ensuring AI models general alignment with human lies in their *perceptual alignment*. Here, we define perceptual alignment as the agreement between AI assessments and human subjective judgments across different sensory modalities, such as vision, hearing, taste, touch, and smell. However, perceptual modalities vary in their explicitness and ease of evaluation (Dienes and Berry, 1997; Lynott and Connell, 2013). For instance, vision, relying on the human retina, can be effectively captured by cameras and is more straightforward to evaluate and quantify. In contrast, the sense of touch poses greater challenges, both with regards to measuring and describing touch sensations (Lynott and Connell, 2013). In this work, we are exploring how well LLMs can achieve perceptual alignment with humans in a textile hand task. In other words, we explored how well LLMs can predict the textile the human is handling based on the descriptions of their touch experiences.

To investigate the perceptual alignment between humans and AI for touch, We designed a "Guess What Textile" interaction experience and conducted an in-person user study. We focused on the concept of "textiles hand" – the describing the feel of textiles through touch (Behery, 2005), because it reflects an integral everyday task, where a good perceptual alignment would be very desirable. To the best of our knowledge, this study is the ***first exploration of the level of alignment between human touch experiences and LLMs***. We analyzed the model accuracy (i.e., success rate), as well as the participants' validity and similarity ratings. Our approach emphasizes the importance of using comparative measures like validity and similarity, which encompass human subjective judgment in an interactive experience. Our observations indicate that ***LLMs exhibit perceptual biases across various textiles***–showing significantly greater alignment with human perception for certain textiles compared to others (e.g., silk satin being better aligned than cotton denim). While we focused on the sensory experience of textiles, this interactive task can also be used in other everyday sensory interactions, such as selection of foods (e.g. choice between sweet fruits) or perfumes (e.g. different fragrances).

2 Related Work

Representation alignment refers to the extent to which the internal representations of two or more information processing systems are aligned (Sucholutsky et al., 2023). While we see advances in human-AI perceptual alignment, they are still mostly limited to vision, such as recent efforts by Boggust et. al (Boggust et al., 2022), Lee et.al (Lee et al., 2023), Kawakita (Kawakita et al., 2023). Extending the exploration of Human-AI alignment into sensory judgments, Marjieh et al. (Marjieh et al., 2023) displayed the same pair of colours (red and blue) to both humans and GPT models[1], requesting each to rate the similarity score, and then comparing the resulting scores, i.e. inter-rater reliability (IRR). They demonstrated that GPT-4 can effectively interpret certain human sensory judgments (e.g., colour, sound and taste). Despite IRRs can measure agreement levels, they do not adequately capture deeper nuances in the perceptual alignment. Our research extends beyond these studies by integrating semantic embeddings with textual sensory information, focused on touch.

3 Method

We chose textile hand task to evaluate the perceptual alignment between Human and LLMs. First, it is an everyday activity that people are familiar with. Second, textile descriptions are widely used in fashion retail websites, catalogues and books, serving as training data for web-scale models like GPTs (Radford et al., 2019). We examine the LLM's behaviors by analysing their embeddings. In our particular case, similar textiles based on their descriptors should cluster in proximity. In other words, we examine whether LLMs grasp the concept of a "softer textile" by checking whether such textiles cluster closely in the LLM's embedding space.

To facilitate human-in-the-loop evaluation, we have developed an interactive AI guessing task named "Guess What Textile" (Figure 1, Sec 3.1) for our in-person user study. The study is designed as a *comparative description task* that requires *continuous verbal Human-AI interaction*. The system is embedded into an user interface to deliver feedback in both text and audio. The text feedback serves as a backup for participants to double-check the replies while rate the validity score (see Sec 3.2).

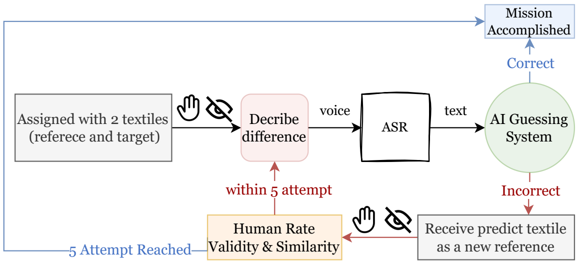
Mission Accomplished
Correct
Assigned with 2 textiles (referece and target)
Decribe difference
voice
ASR
text
AI Guessing System
within 5 attempt
Incorrect
Human Rate Validity & Similarity
Receive predict textile as a new reference
5 Attempt Reached

Figure 1: The overall design of the "Guess What Textile" task. Participants touch two assigned textiles (a target and a reference textile) placed inside a box to hide any visual influences. The AI guessing system knows only the reference textile and is required to make a prediction of the target textile based on participants' descriptions. The task is iterative, and stops only when a correct prediction is made or when the maximum number of five attempts is reached.

3.1 "Guess What Textile" System Design

We developed an interactive AI guessing task, "Guess What Textile", as shown in Figure 1. For the AI to make a prediction of the target textile sample, we first encode the possible textiles into LLM embeddings. The study selected 20 textile samples based on the TextileNet taxonomy (Zhong et al., 2023). Each sample, made from 100% single material, was selected from a diverse collection of over 100 samples to represent a broad range of textures and physical properties. The descriptions of these samples were developed based on Textilepedia (Fashionary, 2020) and commercial sample books, following consultations with domain experts. This expert involvement ensured that our descriptions adhered to industry standards and comprehensively represented the textile properties. We provide a full list of descriptions in our Appendix.

We used the 20 textiles descriptions to generate 20 unique vectors ($\mathcal{E}_{textiles}$) encoded by OpenAI's `text-embedding-3-small` (Neelakantan et al., 2022) to create our embeddings. These embeddings are generated once and used repeatedly in the vector search process during each human input throughout the user study as illustrated in Figure 2. In the study, participant provide comparative descriptions by articulating the differences they perceive while handling two textiles. These descriptions were processed by ASR (Radford et al., 2023) and then feed into encoder to form a vector (v_{query}) that joins the vector search to predict target textile. The user interface is then provide an informed prediction of the target textile from the output of the vector search. It is important to note that the (v_{query}) is not replaced but appended with subse-

[1] Given that GPT models lack the ability to "see" colour hex codes are provided as textual inputs for LLM's "vision".

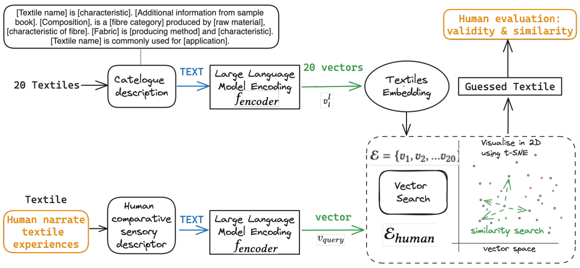
[Textile name] is [characteristic]. [Additional information from sample book]. [Composition], is a [fibre category] produced by [raw material], [characteristic of fibre]. [Fabric] is [producing method] and [characteristic]. [Textile name] is commonly used for [application].
20 Textiles
Catelogue description
TEXT
Large Language Model Encoding
$f_{encoder}$
20 vectors
v_i^l
Textiles Embedding
Guessed Textile
Human evaluation: validity & similarity
$\mathcal{E} = \{v_1, v_2, ...v_{20}\}$
Vector Search
$\mathcal{E}_{human}$
Visualise in 2D using t-SNE
similarity search
vector space
Textile
Human narrate textile experiences
Human comparative sensory descriptor
TEXT
Large Language Model Encoding
$f_{encoder}$
vector
v_{query}

Figure 2: An overview of the AI Guessing System, i.e. "Guess What Textile?". The vector search process uses pre-built embeddings for 20 textile samples and compares them with a user query-generated vector to identify the best matching textile ID.

quent trial within a single task round. This happens when the AI system makes an incorrect prediction and the user proceeds to another trial. Here, previous queries are retained and the new query is appended, along with an added prompt stating, *“[previous query] You were asked to guess with the following additional information because your previous answer was wrong. [new query]"*. This design enables the AI model to maintain awareness of past information within the interactive structure. This iterative process continues until the AI correctly identifies the textile or reaches the maximum of five attempts. This approach necessitates the AI system to possess a more nuanced understanding of how humans describe sensory perceptions than is needed for direct textile identification.

3.2 Measuring Matrix

To measure the degree of perceptual alignment between the human and LLM, we used the following three evaluation metrics (1) AI Success Rate in the "Guess What Textile?" task; (2) Validity Score by participants as a subjective assessment (3) Similarity Score by participants as a subjective assessment. If the AI correctly identified the target textile within five attempts, we considered it as a success. The two additional subjective measures (validity and similarity scores) captured the participants’ subjective judgement of the AI’s performance, as the accuracy (i.e., success rate) alone does not fully capture the human-AI alignment. For instance, AI might make an incorrect prediction, yet it is still closely aligned with human input. In such cases, we rely on human judgments to gauge the degree of error. A slightly incorrect answer could still indicate strong alignment if the human judges it to be valid and similar. This combination of objective and subjective metrics is unique to our approach, as prior works mainly rely on AI accuracy without human validation.

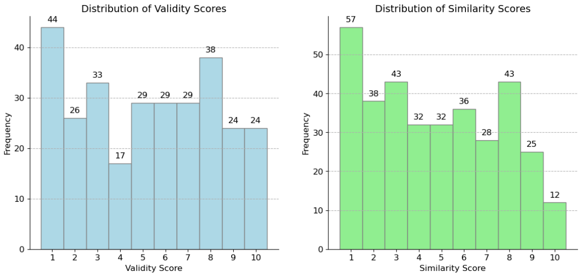

Figure 3: Distribution of Similarity and Validity Scores.

4 Results and Discussion

We analyzed 80 "Guess What Textile" tasks with 362 attempts (avg 4.53 attempts per task, std = 1.41) completed by 40 participants.

4.1 Overall Alignment Performance

The primary performance indicators for alignment are based on the AI’s success rates across 80 completed tasks. The success rate is measured by counting the number of AI’s successful predictions of a textile and dividing this number by the total tasks completed. The AI correctly predicted the target textile in 18 out of the 80 completed tasks, resulting in an overall accuracy rate of 22.5%. For the tasks where the AI succeed, an average of 3 attempts (std=1.20) is needed to make a correct prediction.

The distribution of validity and similarity scores are shown in Figure 3. If AI made a correct prediction, this means the prediction is completely valid (10) and completely similar (10). Therefore, we focus only on attempts where the AI failed to make correct predictions, and ask human to provide their subjective judgements on validity and similarity of the AI’s answers. In essence, these two metrics function as a gauge through which humans assess the degree of inaccuracy in the AI’s predictions.

The AI’s predictions received an average validity score of 5.25 (std=1.71), indicating a moderate level of validity as evaluated by the participants. Regarding the similarity ratings, the average similarity score across all comparisons was 4.77 (std=1.67). There appears to be a correlation between validity and similarity scores, with the highest frequency at a score of 1 and the second-highest at a score of 8. A significant number of participants rated both validity and similarity at 1 in the failed attempts, indicating that the ***AI’s guesses were perceived as highly inaccurate***.

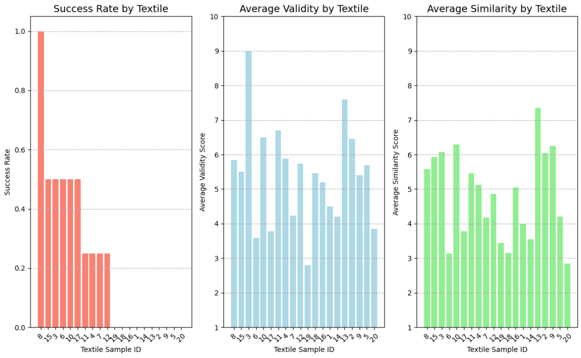

Figure 4: Textile-specific success rates; average validity and similarity scores per textile.

4.2 Textile-Specific Performance

Figure 4 shows the success rate, validity, and similarity scores for each textile. We arranged the textile samples in descending order of success rate and plotted their corresponding average validity and similarity scores. The results suggest that there is a significant perceptual bias across various textiles on all metrics evaluated.

Textile-specific Success Rate The success rate varies significantly across textiles, suggesting that the AI found some textiles easier to guess than others. Some textiles yield significantly higher success rates: 100% for silk satin, and 0% for many other textiles, resulting in a highly skewed distribution for success rates. This could be due to various factors, on both the AI and the user sides, such as the specificity of user descriptions for certain items or inherent characteristics of the items make them more distinguishable. This result supported the claim that there exhibits a significant bias in perceptual alignment across textiles.

Validity and Similarity Scores The average validity scores also show considerable variation across textiles. This suggests that the context or the relevance of the AI's guesses fluctuated, with some guesses being more contextually valid than others, when the humans are making judgements. Similar to the validity score, the average similarity varies by textiles. This indicates that for some items, the AI's guesses were closer to the target in terms of similarity, possibly because these items had more distinctive features or were described more accurately and precisely by users.

Additionally, while the distributions of average validity and similarity scores are less skewed, they still exhibit considerable variance across different textiles. This is even true for the textiles which have a success rate value of zeros. For instance, in Figure 4, linen plain (id 3) shows significantly larger validity and similarity scores compared to pu faux leather (id 19).

4.3 Variables Influencing LLM Alignment

Although measured differently, Marjieh et al (Marjieh et al., 2023) suggest that LLMs have good perceptual alignment in common modalities such as vision. For instance, they experimented on colors and observed high alignment measured by interrater reliability scores. We, however, observed the exact opposite for the sense of touch on textile experience. We hypothesize that this significant difference *origins from the training data*. We thus conducted a simple experiment that is to traverse the common training datasets WikiText-103 (Merity et al., 2016) and BookCorpus (Zhu et al., 2015). The former is a collection articles on Wikipedia, and the later is a large collection of novel books (Zhu et al., 2015).

We then take a list of keywords for textiles, which are basically words and subwords from the 20 textile sample names. We also built another list of keywords that contains common colors [2]. We observed that 0.15% and 0.04% of the words in WikiText-103 and BookCorpus respectively contain color keywords, while only 0.0033% and 0.0018% are observed for textiles. It is therefore reasonable to suggest that variations in the training data could contribute to the varying levels of alignment observed.

5 Conclusion and Future Work

In this paper, we explored human-AI, specifically human-LLM, perceptual alignment using the textile hand concept. We developed a "Guess What Textile" interactive task and conducted an in-person user study with 40 participants. Our results suggest some level of perceptual alignment, however we observed a bias of the LLM across various textiles. We observed that there is significantly greater alignment with human perception for certain textiles compared to others (e.g., silk satin versus cotton denim). In our discussion of this exploratory work, we highlight that LLMs are still in their infancy concerning sensory judgment, particularly in the realm of tactile perception.

[2] We considered "red", "orange", "yellow", "green", "blue", "purple", "pink", "brown", "black", "gray" and "white"

6 Limitations

While we contributed initial insights into the understanding of human-AI touch alignment, we have to also acknowledge some limitations. First, subjective sensory judgement inherently can vary widely among individuals (Stevens, 1960). Conveying and interpreting tactile experiences through language poses significant challenges due to the inherent ambiguity in semantic descriptions (Rosenkranz and Altinsoy, 2020; Atkinson et al., 2016). The ambiguity in conveying sensory experience stems from cultural, social, and linguistic differences that influence our sensory perception (Marques et al., 2022). This diversity has long presented challenges in standardizing evaluations and design metrics that precisely encapsulate the depth of subjective experiences. Hence, this also affects our alignment measures. We have added new measures, especially the validity score, to our study, extending prior works; yet additional qualitative measures would shed light on the quality differences in subjective touch experiences.

Second, our study was confined to a limited selection of 20 textile samples, focusing specifically on tasks related to the feel of textiles. While this sample was selected out of a set of originally 100 samples, there is still scope to extend the choices to enrich the embedding space.

Furthermore, the advent of Multimodal Large Language Models (MLLMs), such as KOSMOS-1 (Huang et al., 2023), represents a major leap in emerging multimodal learning—including multimodal dialogue, image captioning, visual question answering, and vision tasks. While our study explores perceptual alignment in foundational language models, MLLMs' ability to process multimodal inputs offers a richer information landscape. Future research can now investigate the potential of MLLMs to enhance human-AI perceptual alignment, exploring how these advanced models can enhance our understanding of multimodal human-AI interaction for everyday tasks, where AI products and services are increasingly embedded into many devices, beyond choosing clothing.

7 Ethics

A total of 40 participants (30 female, 10 male; aged 18-39, mean = 25.79, std = 4.12) were recruited for the in-person user study. None of the participants had any sensory or motor impairments that would affect their perception and handling of the textile samples. Participants had a diverse range of backgrounds, including psychology students, computer scientists, designers, artists, researchers, and university lecturers. All participants were either native English speakers or highly proficient in English. All participants provided written informed consent before participating in the study and were compensated with a gift voucher for their participation in a 30-minutes study. The study was approved by the local University Research Ethics Committee (Ethics number anonymous).

References

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, et al. 2021. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*.

Douglas Atkinson, Sharon Baurley, Bruna Beatriz Petreca, Nadia Bianchi-Berthouze, and Penelope Watkins. 2016. The tactile triangle: a design research framework demonstrated through tactile comparisons of textile materials. *Journal of Design Research*, 14(2):142–170.

Hassan Behery. 2005. *Effect of mechanical and physical properties on fabric hand*. Elsevier.

Angie Boggust, Benjamin Hoover, Arvind Satyanarayan, and Hendrik Strobelt. 2022. Shared interest: Measuring human-ai alignment to identify recurring patterns in model behavior. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–17.

Allan Dafoe, Yoram Bachrach, Gillian Hadfield, Eric Horvitz, Kate Larson, and Thore Graepel. 2021. Cooperative ai: machines must learn to find common ground. *Nature*, 593(7857):33–36.

Allan Dafoe, Edward Hughes, Yoram Bachrach, Tantum Collins, Kevin R. McKee, Joel Z. Leibo, Kate Larson, and Thore Graepel. 2020. Open Problems in Cooperative AI. *Preprint*, arxiv:2012.08630.

Zoltan Dienes and Dianne Berry. 1997. Implicit learning: Below the subjective threshold. *Psychonomic bulletin & review*, 4:3–23.

Anthony Elliott. 2019. *The culture of AI: Everyday life and the digital revolution*. Routledge.

Fashionary. 2020. *Textilepedia: The Complete Fabric Guide*. Fashionary.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2020. Aligning ai with shared human values. In *International Conference on Learning Representations*.

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, Qiang Liu, Kriti Aggarwal, Zewen Chi, Johan Bjorck, Vishrav Chaudhary, Subhojit Som, Xia Song, and Furu Wei. 2023. Language Is Not All You Need: Aligning Perception with Language Models. *Preprint*, arxiv:2302.14045.

Genji Kawakita, Ariel Mikhael Zeleznikow-Johnston, Ken Takeda, Naotsugu Tsuchiya, and Masafumi Oizumi. 2023. Is my" red" your" red"?: Unsupervised alignment of qualia structures via optimal transport. In *ICLR 2024 Workshop on Representational Alignment*.

Jiyoung Lee, Seungho Kim, Seunghyun Won, Joonseok Lee, Marzyeh Ghassemi, James Thorne, Jaeseok Choi, O-Kil Kwon, and Edward Choi. 2023. Visalign: Dataset for measuring the alignment between ai and humans in visual perception. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Dermot Lynott and Louise Connell. 2013. Modality exclusivity norms for 400 nouns: The relationship between perceptual experience and surface word form. *Behavior research methods*, 45:516–526.

Raja Marjieh, Ilia Sucholutsky, P v Rijn, Nori Jacoby, and Thomas L Griffiths. 2023. Large language models predict human sensory judgments across six modalities. *arXiv preprint arXiv:2302.01308*.

Catarina Marques, Elisete Correia, Lia-Tânia Dinis, and Alice Vilela. 2022. An overview of sensory characterization techniques: From classical descriptive analysis to the emergence of novel profiling methods. *Foods*, 11(3):255.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. *Preprint*, arXiv:1609.07843.

Arvind Neelakantan, Tao Xu, Raul Puri, Alec Radford, Jesse Michael Han, Jerry Tworek, Qiming Yuan, Nikolas Tezak, Jong Wook Kim, Chris Hallacy, et al. 2022. Text and code embeddings by contrastive pretraining. *arXiv preprint arXiv:2201.10005*.

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

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Robert Rosenkranz and M Ercan Altinsoy. 2020. Mapping the sensory-perceptual space of vibration for user-centered intuitive tactile design. *IEEE Transactions on Haptics*, 14(1):95–108.

Ben Shneiderman. 2020. Bridging the gap between ethics and practice: guidelines for reliable, safe, and trustworthy human-centered ai systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 10(4):1–31.

Stanley S Stevens. 1960. The psychophysics of sensory function. *American scientist*, 48(2):226–253.

Ilia Sucholutsky, Lukas Muttenthaler, Adrian Weller, Andi Peng, Andreea Bobu, Been Kim, Bradley C. Love, Erin Grant, Iris Groen, Jascha Achterberg, Joshua B. Tenenbaum, Katherine M. Collins, Katherine L. Hermann, Kerem Oktar, Klaus Greff, Martin N. Hebart, Nori Jacoby, Qiuyi Zhang, Raja Marjieh, Robert Geirhos, Sherol Chen, Simon Kornblith, Sunayana Rane, Talia Konkle, Thomas P. O'Connell, Thomas Unterthiner, Andrew K. Lampinen, Klaus-Robert Müller, Mariya Toneva, and Thomas L. Griffiths. 2023. Getting aligned on representational alignment. *Preprint*, arxiv:2310.13018.

Shu Zhong, Miriam Ribul, Youngjun Cho, and Marianna Obrist. 2023. Textilenet: A material taxonomy-based fashion textile dataset. *arXiv preprint arXiv:2301.06160*.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *The IEEE International Conference on Computer Vision (ICCV)*.