SIMULATING PERCEPTION: EXPLORING AN AI-ASSISTED APPROACH TO KANSEI ENGINEERING IN EMOTIONAL PRODUCT DESIGN

Sigrid ÅKERSKOG[a], Linus FROSTELL[a], Nuno VALVERDE[b], Simon SCHÜTTE[a], Lluís MARCO-ALMAGRO[1 c]

[a] *Linköping University, Sweden*
[b] *Logitech Europe SA, Switzerland*
[c] *Universitat Politecnica de Catalunya - BarcelonaTech, Spain*

ABSTRACT

Kansei Engineering translates emotional user responses into product design decisions, normally using self-reported questionnaires where users rate products on different perceptions (Kansei words). This paper investigates whether a large language model can support this data collection step by approximating users' emotional ratings of product images. A browser-based tool called KansAI has been developed, in different versions, to assess to what extent human ratings can be simulated by an engine powered by the Claude API. The study uses AI-generated computer mouse images as product stimuli and ten kansei words in the semantic space: Ergonomic, Sturdy, Sleek, Organic, Premium, Quick, Cheerful, Trustworthy, Grippy and Futuristic.

Two Kansei studies with humans were conducted: a first one for calibration of the system (with 71 respondents rating nine images of computer mice), and a second one for validation of the system (with 30 respondents rating five new images of computer mice). Four versions of the KansAI system were compared with progressively richer calibration information: no human data, demographic profiles, mean calibration ratings and mean ratings with few-shot individual examples. Both the aggregated and individual accuracy was evaluated using the mean absolute error between real human ratings and KansAI ratings.

Calibration with real human ratings improved aggregate prediction substantially. There were also differences depending on the type of Kansei word: KansAI performed best on Kansei Words with clearer sensory or visual cues, such as ergonomic, sturdy and sleek, and least accurately on symbolic or affective descriptors, such as futuristic, trustworthy and cheerful. However, the system was not able to correctly simulate the individual variability among human responses. The findings suggest that calibrated LLMs can complement human-centered Kansei Engineering as an early-stage exploratory support tool.

Keywords: *emotional product design; large language models; AI-assisted evaluation; computer mice*

[1] Corresponding author: lluis.marco@upc.edu

1 INTRODUCTION

Kansei Engineering (KE) is a method for translating users' subjective impressions and affective and emotional needs into design-relevant information. It is especially relevant when product differentiation depends not only on functional performance, but also on how a product is perceived (Nagamachi, 1995; Norman, 2003; Schütte et al., 2023). Traditional KE studies rely on developing the semantic space, with a selection of kansei words (KWs), the space of properties, with representative product samples, and linking both the semantic space and the space of properties using descriptive and statistical tools based on human respondent ratings. This human-centered foundation is essential, but it can make early-stage design evaluation slow and costly.

At the same time, multimodal large language models (LLMs) are becoming capable of interpreting both text and images. Recent work has explored LLMs as tools for building semantic spaces in Kansei studies, while broader research on AI agents suggests that calibrated models may approximate aspects of human survey behavior under specific conditions (Alcaide-Marzal & Diego-Mas, 2025; Park et al., 2025). These developments raise a practical question for affective design: can an AI tool be calibrated to approximate the aggregate emotional perception of a target group? And if so, can it also work at an individual level? And will it work equally well no matter the type of Kansei word used?

This paper presents the development of a tool, based on the Claude AI API, called KansAI, to simulate Kansei evaluations of product images. The study, developed in collaboration with the consumer electronics company Logitech, focuses on computer mice.

2 HUMAN KANSEI ENGINEERING STUDIES

This section describes how the traditional Kansei engineering study was prepared: selection of Kansei words, creation of products to be rated, and conduct of the data collection.

2.1 Selection of Kansei words

To gather Kansei words for the study, a hybrid approach combining manual work and AI was used. Relevant Kansei Words (KWs) for computer mice were first collected by analyzing review blogs, online discussions, and three academic papers on mouse design, with Gemini used to support text mining, synonym generation, and preliminary clustering. The expanded list was then manually reduced by removing semantically overlapping terms and reorganizing the remaining words into distinct semantic groups.

To validate this hybrid approach, a workshop was conducted with five LiU participants who interacted with five different computer mice and described the emotions and impressions each product evoked. The participants were instructed to avoid obvious physical descriptions, such as color, and the collected words were translated into English and semantically grouped for comparison with the generated KW list.

A final set of 10 Kansei words was selected for use in the studies: Ergonomic, Sturdy, Sleek, Organic, Premium, Quick, Cheerful, Trustworthy, Grippy and Futuristic.

2.2 Generation of product images

Images of computer mice were created using AI. An evaluation of state-of-the-art generative AI image tools was conducted to select the most suitable engine for product visualization in the project. The engines were evaluated according to photorealism and quality, prompt adherence, design coherence, aesthetic consistency, ease of use, and editing possibilities. These criteria are especially important for Kansei Engineering, where unrealistic or inconsistent visuals could influence participants' emotional perceptions. Gemini's Nano Banana 2 was selected as the primary image generation engine because of its strong overall performance and suitability for iterative prompt refinement. ChatGPT's GPT Image was used as a complementary tool during final refinement, as it showed stronger instruction-following when editing existing generated images. The two models were therefore used in parallel, leveraging their respective strengths.

To ensure a consistent baseline across all created images in terms of lighting, camera angle, and composition, an existing Logitech product photograph was used as a reference image. Following known methods (Barros, 2024), utilizing an input image in text-to-image models provides a structural foundation, allowing the designer to control the deviation from the original design by balancing the prompt against the reference image's geometry.

From the 14 different mouse images created, nine of them (shown in Figure 1) were used iteratively as calibration images in the AI study.



Figure 1. Mouse images (mice 1 to 9) used for calibration in the AI study

The other five images (shown in Figure 2) were set aside to be used for validation.



Figure 2. Mouse images (mice 10 to 14) used for validation

2.3 Data collection

Two questionnaires were used to collect user ratings for all Kansei words for each of the presented mouse images. The two studies serve two different purposes: the first one, with mice 1 to 9 (Figure 1), as calibration data for the AI tool; the second one, with mice 10 to 14 (Figure 2), as data for comparison with the results given by the AI – that is, as validation data.

The two surveys were built and distributed using KESo, a web-based tool developed specifically for conducting KE studies (Schütte et al., 2006). Besides presenting stimuli to be rated on all Kansei words in a randomized way, KESo allows the inclusion of demographic questions. Ratings ranged from 0 to 100, based on a visual-analogue scale.

The survey data were cleaned by removing demographic outliers, duplicate entries, and likely careless responses. Careless responses were identified through straight-lining, defined as leaving more than three consecutive ratings unchanged, and through extreme-response patterns where most Kansei ratings were at the scale extremes.

Table 1 summarizes the participants profiles for both studies.

Table 1. Calibration and validation Kansei engineering studies, done with KESo

Study	Stimuli used	Participants	Gender	Age
Calibration survey	9 (mice 1 to 9)	71	37 male, 34 female	31 aged 18–24; 40 aged 25–29
Validation survey	5 (mice 10 to 14)	30	15 male, 15 female	5 aged 18–24; 25 aged 25–29

A hierarchical cluster analysis was conducted with data from the calibration study, where five different groups were characterized based on demographic and behavioural characteristics.

However, there were no significant differences in the profile of ratings for mice 1 to 9 among these groups. Therefore, all participants' data were jointly analyzed in the following steps of the procedure.

3 KANSEI ENGINEERING STUDY BASED ON AI: KANSAI

This section describes the development and implementation of the KansAI tool, done in four different versions. We will describe these four versions and discuss how each of the versions compares with the results obtained from human ratings, using the validation survey.

3.1 Incremental development of KansAI versions

The KansAI tool was developed through an iterative process, using Anthropic Claude API (claude-sonnet-4-20250514). Claude was selected primarily because a Pro license was provided by Logitech for the purpose of this study, and because Claude was not involved in generating any of the product images used in the evaluation.

Prompts were developed iteratively. The tool was run on all calibration images and the outputs were inspected for failure modes such as product misidentification, insufficient score variation, and unrealistic rating distributions. When a failure mode was identified, the prompt was adjusted to address it, and the process was repeated until outputs consistently met the intended criteria on natural-language instructions.

Each version of KansAI represents the same KansAI tool, but with progressively richer information added to the prompt. The core task, rating scale, Kansei words, and output format were kept constant so the effect of added calibration information could be compared. Table 2 shows the incremental data added in each version.

Table 2. Overview of KansAI versions

Version	Summary	Description
V1	Baseline prompt design	Baseline prompt only. No human-study data, no demographics
V2	Demographic-informed	Adds demographic profiles from the human study: age group, gender, occupation, education level, mouse-use context. No human Kansei ratings are shown.
V3a	KW-rating-informed	Adds human calibration data from 5 calibration mice. Averages of Kansei ratings are given for the 5 mice. Mice are presented with detailed textual visual descriptions
V3b	KW-rating-informed	Same calibration data as V3a, but instead of textual descriptions of the 5 calibration mice, the actual images are embedded and sent as base64 images in the prompt
V4	KW-rating-tuned	Gives human calibration data in the same way as V3a, but including data from the 9 calibration mice (instead of only 5). In addition, it includes some (3) individual ratings from participants as examples

Version V1 is an uncalibrated baseline, testing if Claude could rate product images using Kansei words with only general instructions. Version V2 adds demographic persona information. Version V3a adds averages of human ratings from the calibration study, giving product "images" in a detailed textual form. Version V3b tested the same calibration strategy as V3a but supplied the calibration images directly as base64-encoded visual inputs. It was discontinued due to increased computational cost without observable performance gains. Version V4 expands calibration using all available product images, and giving some individual ratings as examples.

3.2 Comparative Analysis of KansAI Versions

This section presents a comparison of results from the four different versions of the KansAI tool. The indicator chosen to evaluate how well the KansAI tool simulates the ratings of a mouse for each of the 10 kansei words is the mean absolute error. This computation is done in the following manner:

1. Compute the human mean rating ($\bar{H}_{i,k}$) for a given mouse i and Kansei word k.
2. Compute the KansAI mean rating ($\bar{A}_{i,k}$) for the same mouse i and Kansei word k.
3. Take the absolute difference
4. Average those absolute differences across the five validation mice (mice 10 to 14)

So, for one Kansei word:

$$MAE_k = \frac{1}{5}\sum_{i=1}^{5}\left|\bar{H}_{i,k} - \bar{A}_{i,k}\right|$$

Table 3 shows the mean absolute error for each Kansei word in each of the four versions of KansAI.

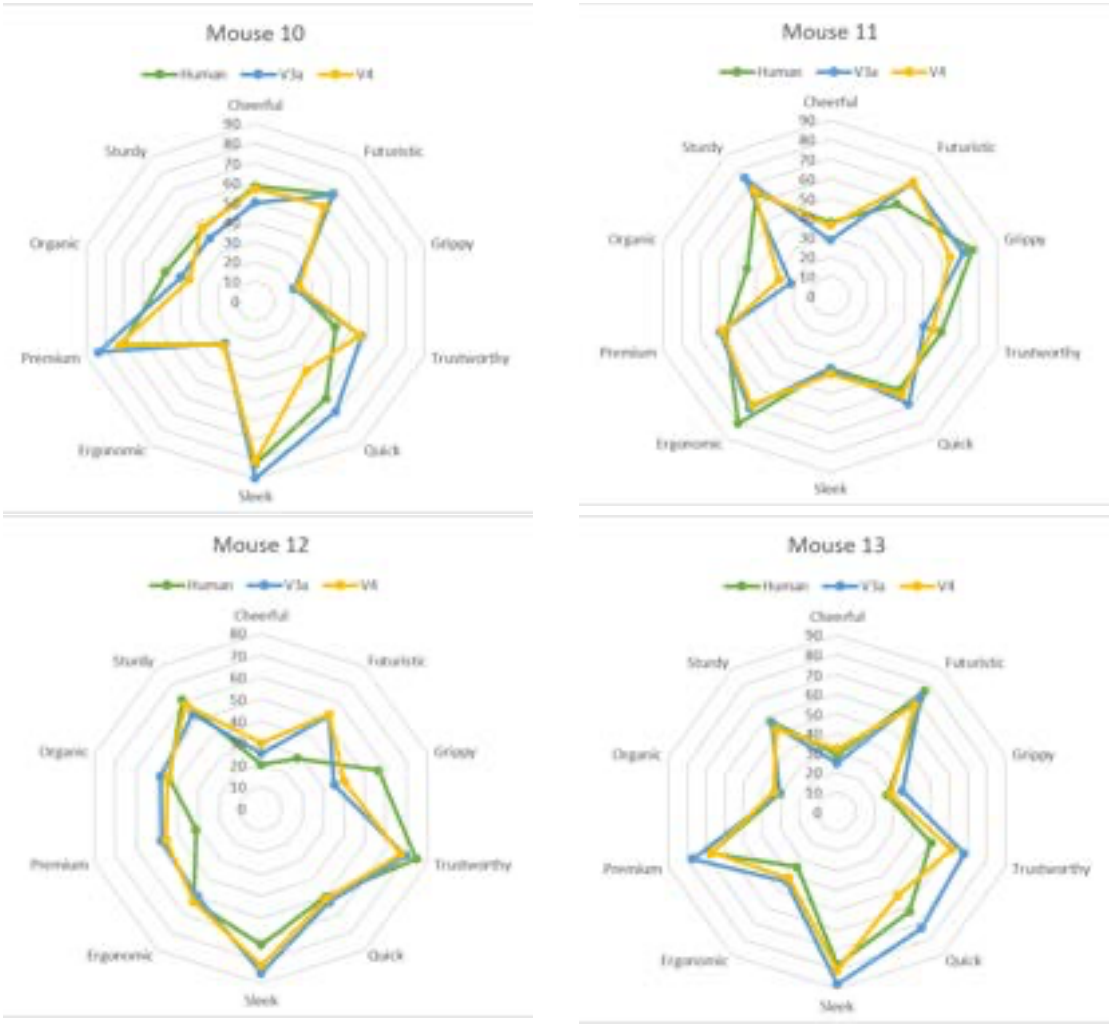
Table 3. Mean absolute error by Kansei word and KansAI version

Kansei word	V1	V2	V3a	V4
Ergonomic	8.09	7.37	6.34	4.46
Sturdy	5.58	5.13	6.65	4.85
Sleek	18.70	18.52	9.57	5.55
Organic	10.35	8.53	7.55	7.05
Premium	17.91	17.31	10.16	7.08
Quick	15.05	12.80	6.92	7.35
Cheerful	21.81	20.36	15.24	9.59
Trustworthy	13.93	12.21	12.79	9.88
Grippy	12.64	13.61	9.03	10.49
Futuristic	17.43	16.90	15.26	17.14
Overall	14.15	13.27	9.95	8.34

Looking at the overall MAE indicator for each version, the largest reduction occurs between V2 and V3a, where the tool was informed with human ratings from the calibration study. Versions V1 and V2 (where no human ratings are informed) have a similar performance. Version V4 is in general slightly better than V3a, but the MAE reduction is modest.

Looking at individual KWs, V4 has the lowest MAE on seven of the ten words. Cheerful and Trustworthy show a large improvement from V3a to V4. Sleek and Premium also improve from V3a to V4, but their main improvement comes from V2 to V3a. Futuristic is a word that shows no improvement over the versions, and is the word where KansAI does a worse job in predicting the rating.

Figure 3 shows radar plots with the profile of average ratings for each mouse and Kansei word for the human ratings (green), and versions V3a (blue) and V4 (yellow). The profiles are, in general, very similar, showing how well the KansAI tool works, especially version V4. However, discrepancies are not negligible in some Kansei words and mice.



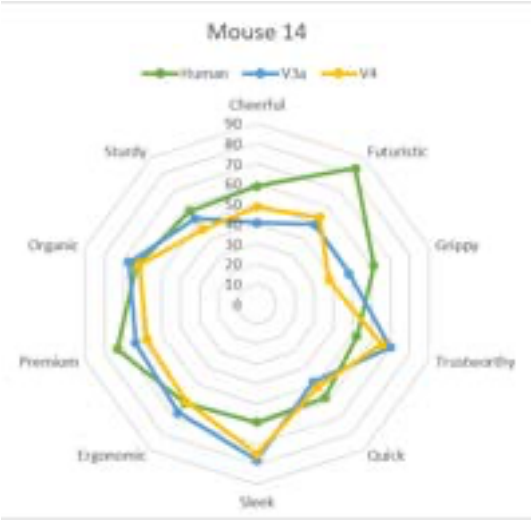


Figure 3. Radar plots showing the profile of ratings given by humans and KansAI (V3a and V4)

3.3 Individual Rating Analysis

As seen in the results from Table 3 and Figure 3, KansAI V3a and V4 are quite good in predicting mean ratings. However, they fail when we want to use them to predict individual ratings. As an example, we can look at Figure 4. For all five mice, the variability of ratings given by KansAI V3a and V4 is much lower than the real variability shown in human ratings. In general, the AI tool correctly locates the level of the response (the average), but, when asked to give individual ratings, it tends to always repeat the same or very similar values, not appropriately capturing the real human variability.

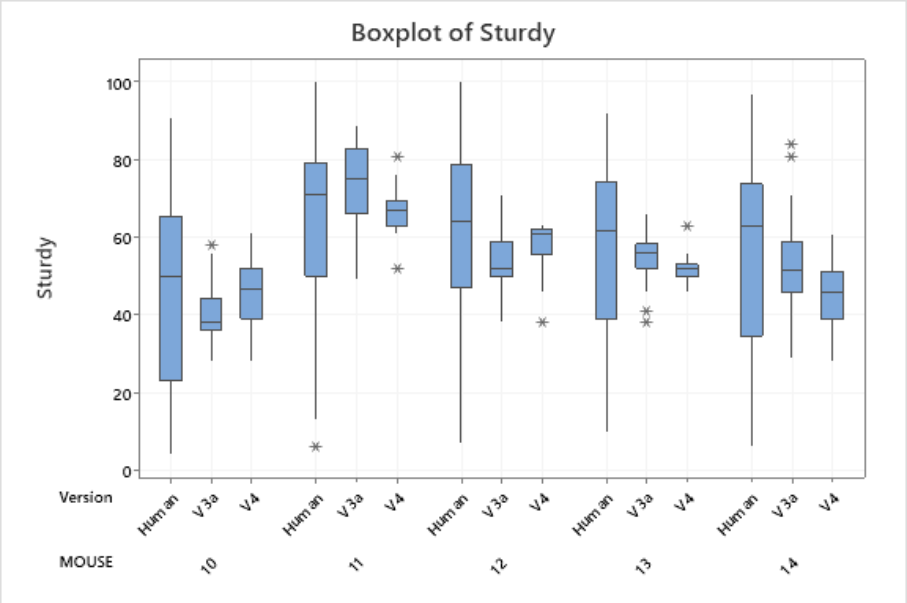


Figure 4. Boxplot of ratings for Sturdy, with Human ratings, and V3a and V4 of KansAI

4 DISCUSSION

The obtained results show that, across the KansAI versions, prediction accuracy generally improved as more human calibration data were incorporated, suggesting that the model could

use examples of human Kansei ratings to better approximate the central tendency of the target group. However, KansAI was not successful at reproducing the variability of individual human responses. This means the tool is better understood as an aggregate estimator.

The variation in accuracy across Kansei words provides important insight into where AI-assisted Kansei evaluation is most promising. Following the distinction proposed by Fenko et al. (2010) and Tavares et al. (2021), the Kansei words can be interpreted as sensory, symbolic, or affective descriptors. KansAI performed best for words with strong visual grounding, such as Ergonomic or Sleek. Although some of these terms involve tactile or use-related qualities, they are still strongly suggested by visible product features such as shape, curvature, and surface texture.

Performance was weaker for words such as Quick and Organic, where the meaning is only partly visible in a static image. These words may depend on assumptions about material, weight, movement, or use experience. The lowest agreement was found for symbolic and affective words such as Futuristic, Cheerful or Trustworthy. These descriptors depend more strongly on cultural expectations, personal interpretation, and emotional response, making them harder to infer from product appearance alone.

Overall, the findings suggest that word category is an important factor in the performance of AI-assisted Kansei evaluation. KansAI appears most reliable when Kansei words are connected to visible product characteristics and less reliable when they depend on subjective, social, or emotional interpretation.

A proof of concept was created and presented to the Head of Industrial Design and the Senior User Experience Designer at Logitech. Both viewed KansAI mainly as a support tool for early-stage design, not as an autonomous replacement for designers. They considered it useful for quick “gut checks” and brainstorming. They also emphasized that the tool should be trustworthy, statistically robust and integrated into existing design workflows if it needs to be really adopted in practice. They said the tool “would need time to prove its usefulness”.

They noted that emotional words are often ambiguous and context-dependent, so Kansei ratings would be more useful if KansAI could explain which visual or physical product features convey each score. Another concern was related to privacy and control of company data.

5 CONCLUSIONS

To sum up, three main conclusions can be drawn:

- KansAI was able to approximate the target group’s general perception of AI-generated computer mouse images at an aggregate level, although it failed to reproduce realistic individual-level variation.
- Calibration with human Kansei rating data substantially improved performance, while demographic information alone produced little improvement.

- Agreement was highest for Kansei words with clear visual grounding, while symbolic and affective words were harder to predict due to their stronger dependence on cultural context, subjective interpretation, and emotional meaning.

REFERENCES

Alcaide-Marzal, J., & Diego-Mas, J. A. (2025). Exploring the use of large language models to build product Kansei semantic spaces. *International Journal of Industrial Ergonomics, 107*, Article 103709. https://doi.org/10.1016/j.ergon.2025.103709

Barros, M., & Ai, Q. (2024). Designing with words: Exploring the integration of text-to-image models in industrial design. *Digital Creativity*, 35(4), 378–391. https://doi.org/10.1080/14626268.2024.2411223

Fenko, A., Otten, J. J., & Schifferstein, H. N. J. (2010). Describing product experience in different languages: The role of sensory modalities. *Journal of Pragmatics*, 42(12), 3314–3327. https://doi.org/10.1016/j.pragma.2010.05.010

Nagamachi, M. (1995). Kansei engineering: A new ergonomic consumer-oriented technology for product development. *International Journal of Industrial Ergonomics*, 15(1), 3–11. https://doi.org/10.1016/0169-8141(94)00052-5

Norman, D. A. (2003). *Emotional design: Why we love (or hate) everyday things*. Basic Books.

Park, J. S., Zou, C. Q., Shaw, A., Hill, B. M., Cai, C. J., Morris, M. R., Willer, R., Liang, P., & Bernstein, M. S. (2025, May 20). *Simulating human behavior with AI agents*. Stanford Institute for Human-Centered Artificial Intelligence. https://hai.stanford.edu/policy/simulating-human-behavior-with-ai-agents

Schütte, S., Ayas Alikalfa, E., Schütte, R., & Eklund, J. (2006, August 9–11). Developing software tools for Kansei engineering processes: Kansei Engineering Software (KESo) and a design support system based on genetic algorithm [Conference paper]. *9th International Quality Management for Organizational Development (QMOD) Conference*, Liverpool, England.

Schütte, S., Lokman, A. M., Coleman, S., & Marco-Almagro, L. (2023). *Unlocking emotions in design: A comprehensive guide to Kansei engineering* (1st ed.). BoD – Books on Demand.

Tavares, D. R., Canciglieri Junior, O., Guimarães, L. B. de M., & Rudek, M. (2021). A systematic literature review of consumers' cognitive-affective needs in product design from 1999 to 2019. *Frontiers in Neuroergonomics*, 1, Article 617799. https://doi.org/10.3389/fnrgo.2020.617799