IFSCC 35TH
CONGRESS
FRANCE_CANNES 15-18 SEPT 25
THE FUTURE IS SCIENCE
ifscc2025.com
IFSCC
SOCIÉTÉ FRANÇAISE DE COSMÉTOLOGIE

IFSCC 2025 full paper (IFSCC2025-640)

“A step beyond self-rating of sensory attributes with artificial intelligence”

Laura Molina [1], Belén Aguirre[1], Mireia Almela[1], Marine Vincendet[1], Gemma Mola[1], Raquel Delgado[1], Cristina Amezcua[1]

[1] Lubrizol Life Science, Lipotec S.A.U, Gavà, Spain

1. Introduction

Perceived skin-feel during and after application of skin care products is highly important to the consumer and, therefore, to cosmetic formulators. Panelist product testing processes are arduous, and time consuming, and have traditionally relied on written questionnaires to capture panelists perceptions. Furthermore, panelist assessments are, by nature, a subjective process that introduces human error into an already complex process.

More recently, tools such as measurements of nervous system signals [1,2], virtual reality systems [3], and physical skin parameters [4] have been explored. While these methods have scientific foundations, their analysis tends to be complex, expensive, and often uncomfortable for participants. To move beyond these limitations and thanks to the advancement of artificial intelligence (AI), video analysis has emerged as a powerful alternative. Following this new path, significant efforts have been made to analyze emotions and consumer reactions to products, both for product development and marketing purposes [[5], 6]. Some frameworks now integrate emotion detection with ChatGPT [7], enabling emotionally intelligent and responsive interactions. However, AI systems, including those used for generating content, often struggle to recognize and adapt to the nuanced emotional context that humans naturally navigate.

These limitations have propelled the use of reasoning models and generative AI. Until now, many applications in the cosmetics industry have relied on multimodal models capable of analyzing video frames, audio, text, and metadata. These models are typically powered by neural networks such as convolutional neural networks (CNNs) or transformers, which focus on pattern detection and statistical correlations. While effective, they provide predictions without offering insights into the reasoning behind them. In contrast, reasoning models are designed to think logically as they deduce, infer, chain thoughts, and form contextual relationships. For example, when analyzing a subject's emotion in a video, a non-reasoning model might simply state: “the person is sad” whereas a reasoning model could explain: “the person is sad because they are looking down while speaking, and their voice tone is very low.”. In this study, we aim to integrate video analysis with AI in the context of sensory product evaluation. Our goal was to explore the possibilities AI can offer in providing a rating for different sensory attributes in a cosmetic product, without the interruption of stopping to fill out questionnaires, resulting in more genuine and accurate behavioral data.

Additionally, we compared the AI outcome capturing the participants’ sensory experience from verbal and non-verbal analysis to the self-reported questionnaires rating and the experts’ rating for each attribute. Agreeability correlations were also established, and the results were compared to the physico-mechanical characterization of the tested products. This tridimensional analysis allowed us to explore disparities in consumer responses, eliminating subjective biases that can occur with self-reported data, providing more objective and consistent results without human interpretation.

2. Materials and Methods

2.1. Participants

100 male and female Spanish speaking volunteers aged between 18 and 70 were recruited as participants for the study. Only 21 of them had prior training experience in sensory studies with cosmetic products.

2.2. Sensory attributes and materials

Three different sensory attributes were considered for this proof-of-concept study: spreadability, smoothness and stickiness.

Spreadability can be defined as the ease of moving of the product over the skin during its application; softness can be related to skin slipperiness after the product has been applied, leaving a smooth and pleasant-to-the-touch skin feel; and stickiness can be defined as the degree which the fingers and palm adhere to the skin after a product has been applied.

For each attribute, three different materials were selected to be assessed in the study. Samples to be tested for each attribute had been previously rated in a scale from 0 to 5 by the expert panelists as shown in Table I.

Table I. Materials used as testing samples for each attribute and its rating by experts where: PTID (Triisostearoyl Polyglyceryl-3 Dimer Dilinoleate); DISM (Diisostearyl Malate); NGDO (Neopentyl Glycol Diethylhexanoate) and GMIS (Diisostearyl Malate), all supplied by Lubrizol Life Science (Brecksville, Ohio, US). IVDL (Isododecane, Vinyl Dimethicone/Lauryl Dimethicone Crosspolymer, Lauryl Dimethicone (and) Dimethicone was supplied by Koda Corporation (New York, US).

Attribute/ Experts Rating score	0	2.5	5
Spreadability	PTID	DISM	NGDO
Smoothness	PTID	GMIS	IVDL
Stickiness	IVDL	DISM	PTID

2.3. Sensory evaluation set up and procedure

Recruited participants were instructed to keep their hands and forearms free of any skin care products the day of testing. Sensory evaluation was carried out in a controlled room (temperature (23 ºC ± 2 ºC) and humidity (45% ± 5% R.H.)), where testing environment was free of distractions.

Upon entering the room, participants were shown a video with instructions on how to apply the materials, including the appropriate amount to use, and how and where to apply it. They were then presented with a detailed explanation of the sensory attributes they were to evaluate. The instructive videos were recorded by specialists from the company's Applications laboratory to ensure clear and consistent descriptions, preventing any external influence on the participants' perceptions and or interpretation.

In addition to the initial explanatory video, a second informative video was played to instruct the volunteers on when to begin the sensory evaluation and look at the camera while briefly describing their perception of the sample. They were also asked to rate the sensory attributes on a scale from 0 to 5 (0-2.5-5), and provide a rationale for their ratings in front of the camara. At the end of the recording, participants were also asked to give their rating in a written questionnaire for further comparison.

Each participant tested two different samples and therefore, a total of 200 videos were recorded. Samples' distribution among participants according to experts rating and attribute designation is depicted in Table II.

Table II. Samples' distribution among participants.

Number of subjects assessing an attribute	**Experts rating score for the sample**		
Attribute	**0**	**2.5**	**5**
Spreadability	22	20	23
Smoothness	23	21	22
Stickiness	22	23	24

2.4. Physico-mechanical characterization

Spreadability and stickiness were characterized with a texture analyzer model TA XT Plus from Micro Stable Systems. Expert Exceed software (version 2.6) was used to collect and display the data. As softness is related to skin slipperiness after the product has been applied onto the skin, it was characterized as a frictional property [8] measuring the coefficient of friction (COF) with a CETR UMT-2 Tribometer (Bruker Nano Surfaces Division) with a 5N load sensor.

2.4.1. Spreadability test

The test measures the force needed to penetrate a fluid sample held in a cone-like vessel (TTC Spreadability Rig (HDP/SR)), translating into how well the sample is spread when applied on the skin. Tests conditions were defined using the compression test mode and a penetration distance of 23 mm. The test speed was set at 3.0 mm/s and post-test speed at 10 mm/s. Three replicates were done per sample.

2.4.2. Coefficient of friction test

The method measures the coefficient of friction (COF) and uses a Teflon ball probe (15.8 mm diameter), EVA foam substrate, 1N load and 20 mm/s speed. The pass length is 40 mm, and the number of passes is 6 (non-reciprocal, each on an untested area of the substrate). The substrate is EVA craft foam, cut to 6.35 x 11.43 cm and 0.65 to 0.75 grams of material were spread over the substrate with a finger cot. Three replicates were done per sample.

2.4.3. Stickiness test

The test measures the force required to separate 2 mm thick ethylene-vinyl acetate (EVA) craft foam sheets (50 x 25 mm) previously impregnated with the test material. Tests conditions were defined using the compression test mode with 10 repetitions and a force limit of 2,500 grams. The test speed was set at 2.0 mm/s and post-test speed at 30 mm/s. Three replicates were done per sample to evaluate the tackiness or stickiness over the 10 compressions.

2.5. Machine learning methodology with AI

The main goal was to assess how well the models could predict the parameters for the tested products using different input modalities: either relying solely on the visual cues from the video without audio (non-verbal approach) or using the speech content through audio transcription (verbal approach) as shown in Figure 1.

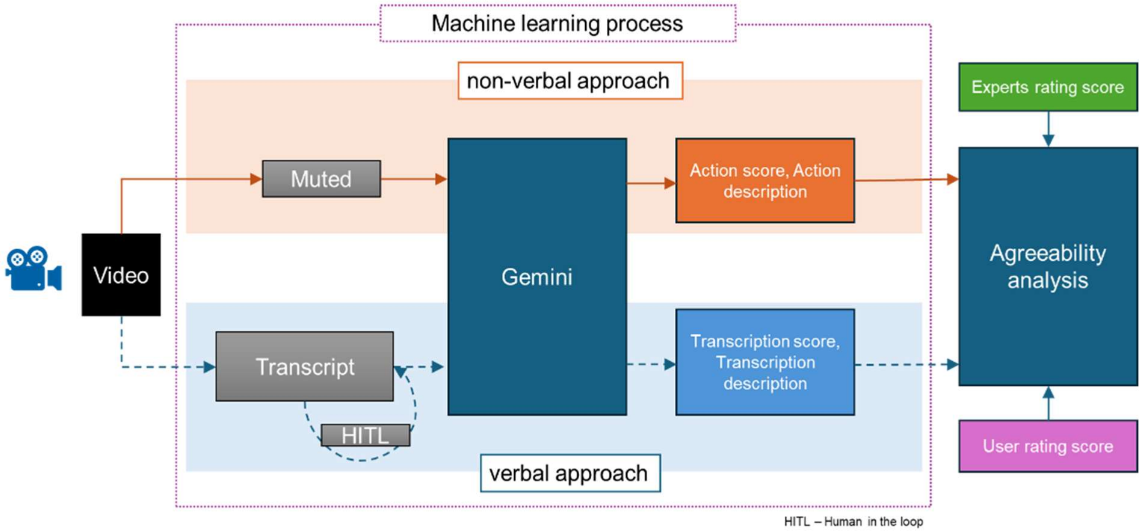


Figure 1. Scheme on the methodology followed during the machine learning process and the agreeability analysis.

Before starting the analysis, careful data preparation was required. While video-based action analysis needed minimal data cleaning, analyzing user speech involved transcribing videos to text for improved accuracy compared to direct speech analysis by models. Several transcription services were evaluated: Whisper 3 and Whisper 3 Large from OpenAI (San Francisco, California, US); LeMUR from AssemblyAI (San Francisco, California, US); and Gemini 2.0 Flash Thinking from Google (Mountain View, California, US). Gemini 2.0 Flash Thinking was selected for its superior performance and further enhancement was carried out by native Spanish speakers (human in the loop) to ensure accuracy while preserving the volunteers' natural language and expressions.

The machine learning approach included advanced reasoning models, specifically Gemini 2.0, to gain deeper insights beyond simple benchmarking. Reasoning models allowed us to understand the model's decision-making, enabling iterative prompt refinement for better output quality. For strategic purposes, no fine tuning was done since it posed the risk of biasing the model towards the data. This aimed to preserve the model's ability to generalize and avoid constraining its outputs, thereby supporting the overarching goal of discovering novel insights rather than reinforcing known patterns.

Video batches were then processed through the chosen model, outputs were reviewed and prompts adjusted based on any identified shortcoming. This cycle of evaluation and refinement allowed to progressively enhance the quality and accuracy of the model's results. To optimize efficiency, videos were compressed for easier handling and automated processing steps. Finally, different APIs were developed for streamlined data exchange and to support future scalability. This overall process focused on robust data preparation and the strategic application of reasoning models to extract valuable insights.

During model prediction, two different scores (ranging from 0 to 5) were used: (i) the action score (from the non-verbal approach) corresponding to the prediction of the model relying only on the action, gestures and movements recorded in the videos for each participant, and (ii) the transcription score (from the verbal approach) associated to the score predicted by the model after participants speech transcription and analysis.

For further agreeability and comparison analysis, two more scores were considered: (iii) the experts rating score of the material for the specific attribute established by the specialists from the company's Applications laboratory; and (iv) the user rating, explicitly stated by the volunteers in written questionnaires.

3. Results

3.1. Sensory evaluation video results

A total of 200 videos of participants following the given instructions on how to proceed for each attribute evaluation were successfully recorded.
At first glance, no extreme positive or negative reactions were detected in most of the videos when the volunteers were assessing the samples.

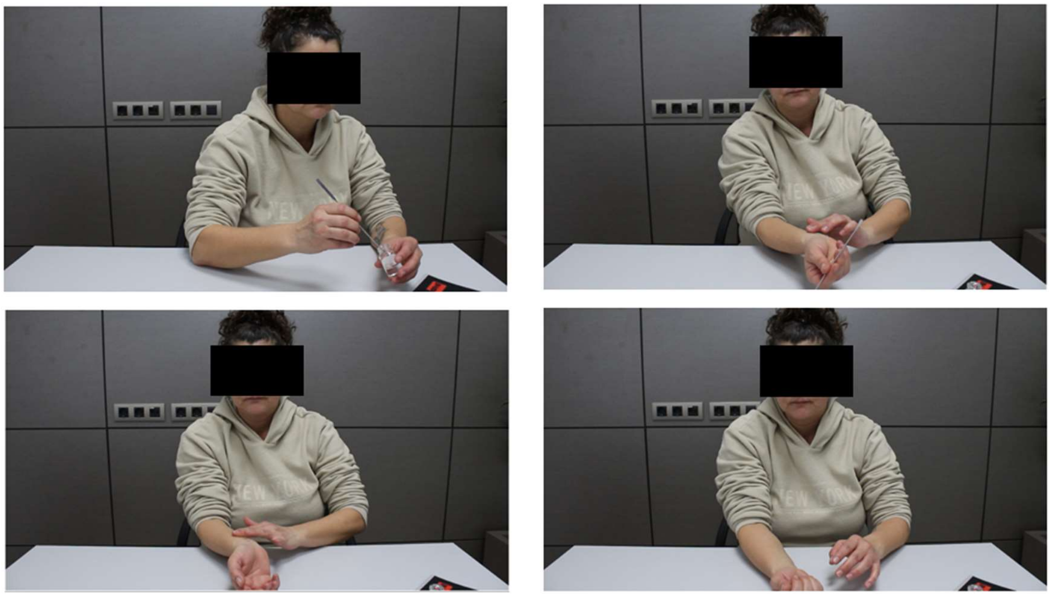

[redacted] [redacted] [redacted] [redacted]

Figure 2. Example of video frames captured from the recording of a volunteer evaluating a sample.

3.2. Physico-mechanical characterization results

The physico-mechanical characterization results for spreadability, smoothness and stickiness for each reference are shown in Figure 3.

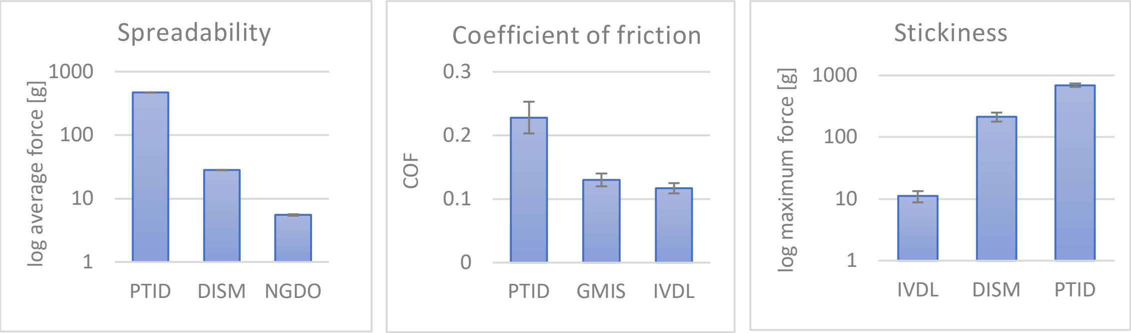


Figure 3. Physico-mechanical characterization tests: Spreadability (left); Coefficient of friction (middle); and Stickiness (right).

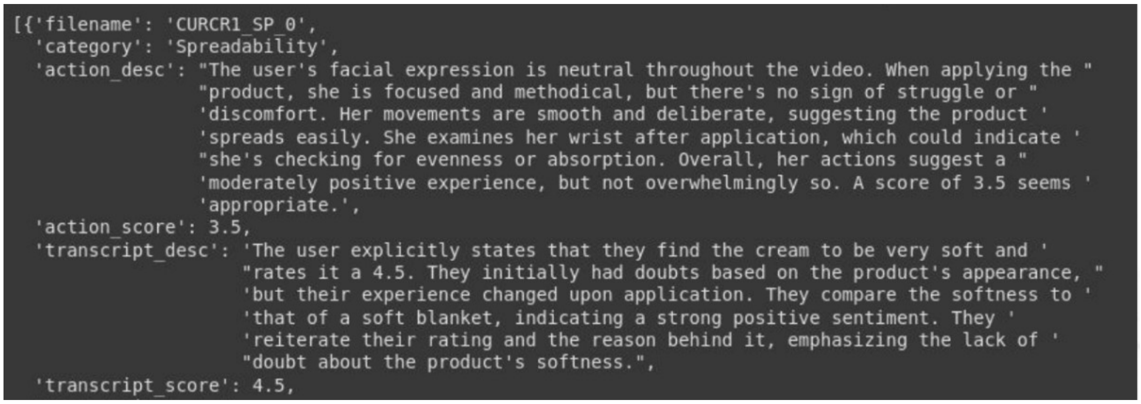
The average force needed to penetrate the PTID sample held in the cone-like vessel was around 468 grams, where the penetration of the same amount of a DISM sample only required 28 grams of force. The lowest force required corresponded to the NGDO sample, with only 5.5 grams of force to displace the material. These results agree with the stablished ratings from our expert panelists, where PTID was given a 0 due to the high difficulty for spreading the sample and NGDO was given a score of 5 due to the easiness in moving the product over the skin.

The smoothness characterization through the COF determination also showed a good correlation between the experts scoring and the test results. PTID sample was rated with a score of 0 and this is associated to the highest COF among the three references in this test, indicating the lowest smoothness among the three references. On the opposite side, the IVLD sample yielded the lowest COF, and this is in consonance with the highest scoring for smoothness from experts.

Stickiness was another attribute whose characterization correlated very well with the experts scoring. PTID, with an expert's scoring of 5 for stickiness, required a maximum force of around 700 grams to separate the two EVA foam sheets impregnated with the material. The lowest values recorded in the test were for IVDL, with only around 10 grams of force and a scoring of 0 from the experts group.

3.3. Machine learning results

An action score model was implemented to analyze non-verbal cues in the videos, with the goal of gauging user sentiment. In parallel, an analysis of verbal feedback was conducted, where the model was asked to interpret user's statements. A description for the reasoning and the assumptions behind each of the model's scores was provided, as shown in the example from Figure 4.



```
[{'filename': 'CURCR1_SP_0',
  'category': 'Spreadability',
  'action_desc': "The user's facial expression is neutral throughout the video. When applying the "
                 "product, she is focused and methodical, but there's no sign of struggle or "
                 'discomfort. Her movements are smooth and deliberate, suggesting the product '
                 'spreads easily. She examines her wrist after application, which could indicate '
                 "she's checking for evenness or absorption. Overall, her actions suggest a "
                 'moderately positive experience, but not overwhelmingly so. A score of 3.5 seems '
                 'appropriate.',
  'action_score': 3.5,
  'transcript_desc': 'The user explicitly states that they find the cream to be very soft and '
                     "rates it a 4.5. They initially had doubts based on the product's appearance, "
                     'but their experience changed upon application. They compare the softness to '
                     'that of a soft blanket, indicating a strong positive sentiment. They '
                     'reiterate their rating and the reason behind it, emphasizing the lack of '
                     "doubt about the product's softness.",
  'transcript_score': 4.5,
```

Figure 4. Example of the advanced reasoning behind the model for the analysis of a video file named named 'CURCR1_SP_0', where the spreadability of a sample was assessed by a panelist.

From the beginning, the results predicted by the model following the verbal approach (transcription score) were promising, whereas the non-verbal approach (action scores) failed to yield values in agreement with the volunteers' final scoring (user rating).

The non-verbal approach yielded neutral ratings in 95% of the dataset, as the model assigned action scores between 2 and 4 in most of the two hundred videos. Therefore, action scores correlated poorly with the experts' rating and the users rating for each of the references tested as depicted in Table III.

Table III. Correlation among the different scores established considering the whole dataset.

Correlation	Experts score	Action score	Transcription score	User score
Experts score	1.000000	0.056640	0.641033	0.730108
Action score	0.056640	1.000000	0.158605	0.224246
Transcription score	0.641033	0.158605	1.000000	0.787167
User score	0.730108	0.224246	0.787167	1.000000

The verbal model had a correlation of 0.78 with volunteers' ratings and 0.64 with experts' ratings. This showed that the model performed better with the transcripts, but since a higher correlation was initially expected, this seeded the basis for inconsistencies detection within the dataset. We identified a total of 25 videos with significant score discrepancies (>= 1.5 points) between the verbal AI model scoring and other scores. These divergences were then tagged as outliers and represented in Figure 5 in red dots.

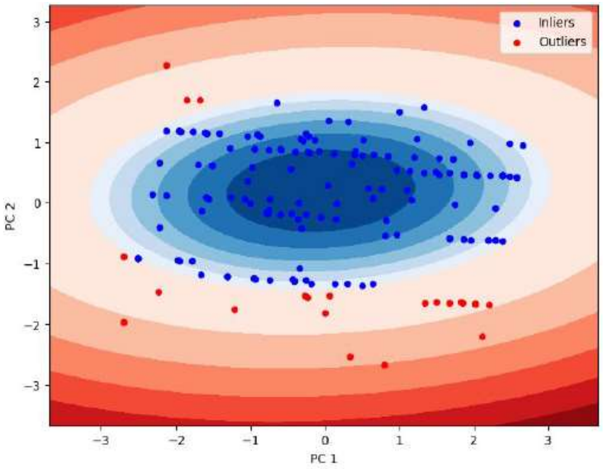


Figure 5. Outliers and decision boundaries.

After removing outlier videos and analyzing again the correlation between the different scores, the results significantly improved as shown in Table IV.

Table IV. Correlation among the different scorings after outliers' detection and removal from the dataset.

Correlation	Experts score	Action score	Transcription score	User score
Experts score	1.000000	0.155337	0.835549	0.857113
Action score	0.155337	1.000000	0.102040	0.119107
Transcription score	0.835549	0.102040	1.000000	0.958021
User score	0.857113	0.119107	0.958021	1.000000

Once the outliers were left out and when considering the different attributes separately, smoothness stood out with a significant difference between action and transcription scores. In this case, the non-verbal approach does not provide any clear indication of what the volunteer might be thinking or considering about the reference under test and therefore, the correlation with the transcription score is almost null. For spreadability and stickiness, the gap between both approaches was not so pronounced, but still yielding low correlation values of around 0.30 for the action-transcription scores.

4. Discussion

The analysis of the texture analyzer results allowed us to quickly establish a direct correlation between the physico-mechanical characterization results for spreadability, smoothness and stickiness and the specialists rating for each reference.

Another early finding was that the model did not lose quality when compressing the videos, which was helpful since videos take up a lot of space at processing time.
While recording the videos it was also noticed that there were not strong non-verbal signals in the participants' reactions to any of the references tested. This translated into video recordings where only subtle movements or gestures together with a neutral general demeanor could be identified in the volunteers. This is aligned with the action score model results for non-verbal cues. The model was only able to primarily identify strong emotional reactions, which were infrequent in the dataset. Therefore, action scoring is best suited for clear positive or negative reactions, as the model presented a limitation for the detection of nuanced sentiment.

On the other hand, a better performance for the verbal model was identified. Gemini 2.0 model worked very well for transcribing audio to text, with minimal human corrections needed. Additionally, it was concluded that these human corrections had little to no impact on the model's reasoning when trying to understand the text. However, AI interpretation of user speech had less correlation than initially expected and this led us to further investigate outliers. During the analysis of the outliers, it was discovered that the model could not detect sarcasm expressed by volunteers. It also helped to detect quality issues in videos or in the experiment raw data collected. For example, it was detected that some volunteers inverted the rating scale: when providing their descriptions to the camara, their speech pointed at a low rating, but then, when writing down their score in the questionnaires, they gave a 5. The outlier detection also showed how many people maintained a neutral expression even when they verbally stated that they did not like the texture of the reference at all. This highlights the limitation of the model in capturing subtle micro expressions and nonverbal cues, which at the same time, are still difficult to be detected by a human being.

These insights open different new paths to enhance the non-verbal detection of the model: from improving data quality and quantity, to redesigning the test for more natural settings and open-ended tasks to elicit richer emotional responses.

5. Conclusion

This proof-of-concept study demonstrates that the use of AI with reasoning models such as Gemini 2.0 can be a valuable tool for panelist's verbal sentiment transcription and interpretation during cosmetic product evaluation. Score divergences between the verbal AI model and other scores were key indicators for identifying potential outliers in the dataset, allowing to differentiate true discrepancies from model errors.

However, the non-verbal model showed poor or limited capabilities for identifying subtle cues in panelists gestures or expressions, proving to be effective only for strong emotional reactions. Despite rapid advances in artificial intelligence, humans still maintain a significant edge when it comes to understanding underlying emotions, revealing fundamental limitations in AI's ability to interpret human behavior through videos.

6. Acknowledgments

We thank Lubrizol Advanced Materials for supporting this work. Special thanks to Alina González, Mireia Martín, Natalia Crespo, Felix Vázquez and the CloudFactory team for their contribution to this study.

7. References

1. Patrice Bellon, Céline Carrasco-Douroux, Julia Devismes. Can our emotions be used to evaluate the sensory feeling of cosmetic textures during application on skin? 25th IFSCC Conference **2019** – Milano, Italy.
2. Gabriel, D.; Merat, E.; Jeudy, A.; Cambos, S.; Chabin, T.; Giustiniani, J.; Haffen, E. Emotional Effects Induced by the Application of a Cosmetic Product: A Real-Time Electrophysiological Evaluation. Appl. Sci. **2021**, 11, 4766.
3. A.H.B. Zulkarnain et al. Assessment of a virtual sensory laboratory for consumer sensory evaluations. *Heliyon* 10 (**2024**) e25498
4. Saito, N.; Matsumori, K.; Kazama, T.; Arakawa, N.; Okamoto, S. Skin Sensory Assessors Highly Agree on the Appraisal of Skin Smoothness and Elasticity but Fairly on Softness and Moisturization. *Cosmetics* **2022**, *9*, 86.
5. "Dove's Emotional Connection: Beauty is for Everyone!" imentive ai. November 16, **2023**. https://imentiv.ai/blog/doves-emotional-connection-beauty-is-for-everyone/
6. "Understanding Emotion AI: Applications, Benefits, and Limitations" imentive ai. May 22, **2024**.https://imentiv.ai/blog/understanding-emotion-ai-applications-benefits-and-limitations/
7. Tomasi, Claudia "Revolutionizing AI Conversation with Emotional Understanding: Introducing MorphCast Facial Emotion AI for ChatGPT on July 2023" MorphCast. July 19, **2023**. https://www.morphcast.com/blog/revolutionizing-ai-conversation-with-emotional-understanding-introducing-morphcast-facial-emotion-ai-for-chatgpt-on-july-2023/
8. Smith, A. M., & Scott, S. H. (**1996**). Subjective scaling of smooth surface friction. Journal of Neurophysiology, 75, 1957–1962.