REPORT: LLAVA-BASED IMAGE RATING TASK AND ANALYSIS

Yanheng Li
Renmin University of China
chrislee1067@outlook.com

1 METHODOLOGY

I select llava-v1.6-vicuna-13b (Liu et al. (2024)) as the model to execute the image rating task because of its availability and capability to process multi-modal tasks. Below I will introduce the methods used in the task, including the data, prompts design, implementation and analysis.

Dataset and Benchmark As provided in the given paper (Lehman et al. (2019)), the dataset is available online[1]. The dataset including 320 images to be processed. Besides, I used their uploaded sheet file containing average human ratings of relevance, arousal, and valence for each image as the benchmark to assess the performance of LLaVa rating.

Prompts Design. Combining the previous prompts of the task and some other more popular templates used in many ACL conference works (e.g., Zhou et al. (2024)), I design three structured prompts for the LLaVa to output the rating as well as the model CoT (Chain of Thoughts) of "Relevance", "Arousal" and "Valence" in a relatively rigid template for easier parsing process. See prompts in Appendix A.

Implementation Details and Pseudocode. I first deploy and test the LLaVA model locally. Then, I integrate the image dataset with designated prompts for each rating aspect, formatting the prompts into a "jsonl" structure as model input. Each image input consists of the image itself and the corresponding prompt. Next, I run the model to perform the rating task, generating ten outputs per image to ensure more effective results and provide additional analysis options (e.g., evaluating model robustness). To analyze the results, I extract all ratings from the output "jsonl" files and compile them into a new ".xlsx" file. I then conduct statistical and empirical analyses, as detailed in the accompanying ".ipynb" files. For further reference, see the pseudocode in Appendix B.

Analysis. To evaluate the performance of LLaVa image encoder, I set two parts for the analysis. Firstly, **the standard deviation, the Pearson correlation and the Intra-Class Correlation (ICC)** of all LLaVa ratings (10 ratings for each aspect per image) are computed to evaluate the robustness of the model itself. Secondly, to compare the average ratings between "LLaVa" and "Human", I design a three-phases evaluation. See the statistical details in Appendix B.

- Test whether there is significant difference as well as the direction of the difference between Human ratings and LLaVa ratings, using **Wilcoxon Signed-Rank Directional Test**.
- If there is no significant difference between some rating results, test whether there is statistical equivalence by **TOST**.
- Calculate **the errors and the variances** between LLaVa and Human ratings across different rating aspects (i.e., relevance, arousal, valence) to quantitatively further analyze the difference between LLaVa and Human encoder.

2 RESULTS AND DISCUSSION

LLaVa. Collectively, the relatively high standard deviation (Figure 1) and the low ICC (Figure 2) in LLaVA's ratings for each aspect per image indicate suboptimal model robustness. This suggests

[1] https://affectiveclimateimages.weebly.com/

significant variability in LLaVA's ratings, which may necessitate either a more capable model or a fine-tuning process before encoding. Additionally, the correlation matrix (Figure 3) demonstrates the independence of LLaVA's ratings across the three aspects, potentially reflecting the model's ability to distinguish between different rating criteria.

LLaVa Versus Human Benchmark. I set the average human rating as the benchmark for comparison against LLaVA's average ratings across the three aspects per image. As shown in Table 1, LLaVA's *Relevance* ratings are significantly lower than human ratings, whereas its *Valence* ratings are significantly higher, both at the 0.05 significance level. In contrast, the *Arousal* rating exhibits no significant difference between LLaVA and human ratings, as confirmed by the TOST analysis in Table 2, suggesting statistical equivalence. Consistently in Figure 4, the variance of *Arousal* among human and LLaVA ratings is the smallest, approaching $zero$ despite the presence of some outliers. In comparison, the variances of *Relevance* and *Valence* deviate further from $zero$, indicating greater discrepancies between LLaVA and human ratings in the two aspects.

Comparison	Wilcoxon Stat.	p-value (Higher)	p-value (Lower)
Relevance	6,506.00	1.00	≈ 0.00
Arousal	24,479.00	0.70	0.30
Valence	34,116.50	≈ 0.00	1.00

Table 1: Wilcoxon Directional Test Results (Human ratings as comparison benchmark)

Comparison	Mean Diff.	p-value	Equivalent
Relevance	-0.89	1.75×10^{-19}	No
Arousal	-0.04	0.25	Yes
Valence	0.31	≈ 0.00	No

Table 2: TOST Equivalence Test Results

Discussion. The results suggest that while LLaVA captures *Arousal* in a manner closely aligned with human perception, its assessments of *Relevance* and *Valence* exhibit systematic differences probably due to limitations in its understanding of visual relevance and emotional intensity, as well as a lack of rating robustness. To better understand the performance of LLaVa, a case could be illustrated by two ratings of *Relevance* both from "pic001" with the same prompts, see Appendix B for the pic001 (Figure 5) and the prompts. Two answers of LLaVa are as below:

> $AnswerNo.1$ {Rating: 9. Reason: This image depicts two polar bears swimming across an area of water. Climate change is a major factor affecting the Arctic ecosystem, and as sea ice melts due to climate change, polar bears face increasing challenges in finding food and breeding. The image serves as a powerful visual reminder of the impacts of climate change on the environment and wildlife.}

> $AnswerNo.4$ {Rating: 5. Reason: The image shows two polar bears, which are animals directly affected by climate change. Their habitat is changing due to rising temperatures and melting ice, leading to challenges in their survival. However, the image does not explicitly depict any direct effects of climate change on the environment. It also doesn't show clear evidence of human activity or infrastructure that contributes significantly to climate change. Therefore, while it is relevant to climate change as an indicator species, its direct impact on human-induced climate change is less visible in this image alone.}

The above responses demonstrate significant inconsistencies in LLaVA's rating mechanism and reasoning. For example, $AnswerNo.1$ describes the image as a "powerful visual reminder" of climate change, emphasizing **the symbolic significance of polar bears** and justifying a high relevance score, whereas $AnswerNo.4$ downplays this symbolism and argues that the image does not provide direct **evidence of human activity**, leading to a much lower rating. This suggests that the model

applies stochastic evaluation criteria across different runs. To be considered, fine-tuning process or a more reliable prompt setting may be beneficial to the reliability of LLaVA's image rating system.

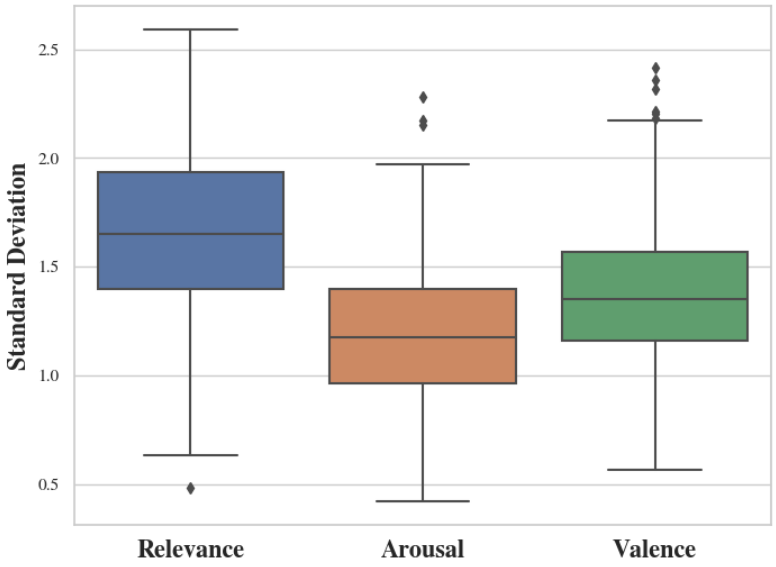

Figure 1: Standard deviation of ten LLaVA ratings per image across three aspects.

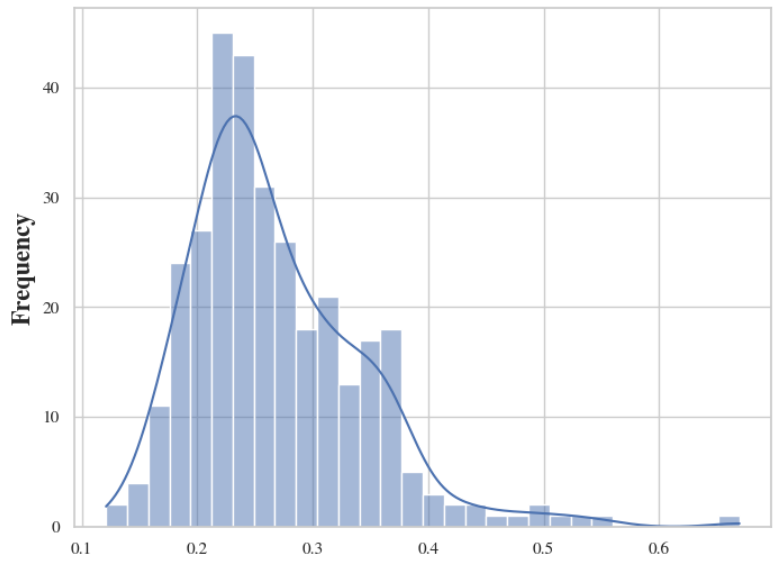

Figure 2: Consistency analysis (ICC) of ten LLaVA ratings per image across three aspects.

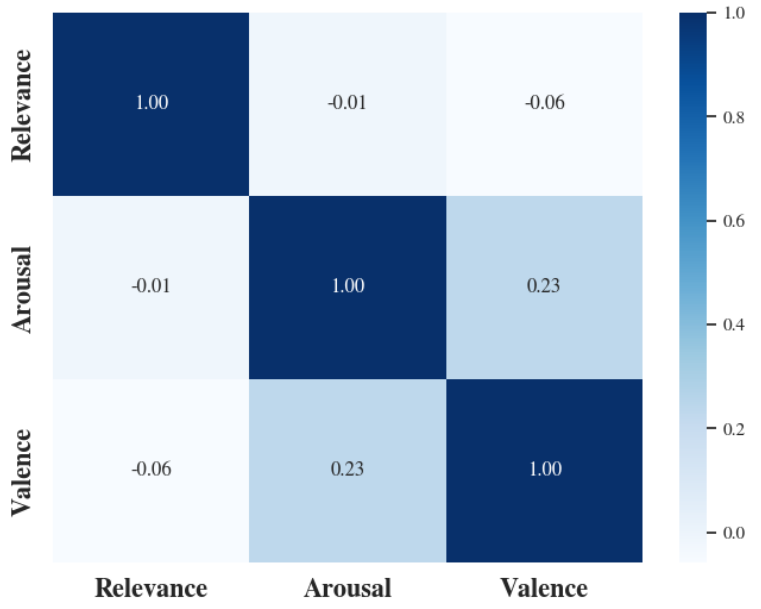

Figure 3: Pearson correlation coefficients of all LLaVA ratings per image among three aspects.

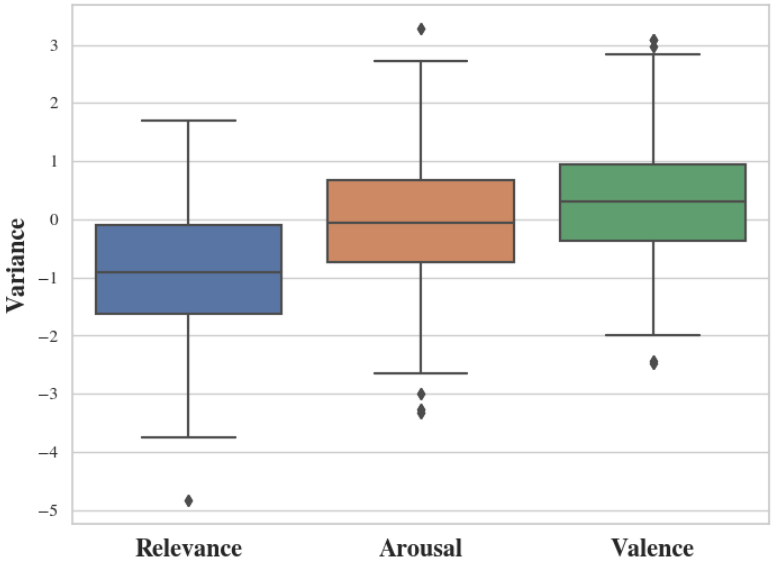

Figure 4: Variance among Human and LLaVA ratings per image across three aspects. (Variance = LLaVA Rating − Human Rating)

REFERENCES

Daniël Lakens. Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, 8:194855061769717, 05 2017. doi: 10.1177/1948550617697177.

Betsy Lehman, Jessica Thompson, Shawn Davis, and Joshua M. Carlson. Affective images of climate change. *Frontiers in Psychology*, 10, 2019. ISSN 1664-1078. doi: 10.3389/fpsyg.2019.00960. URL https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2019.00960.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

Ruiyang Zhou, Lu Chen, and Kai Yu. Is LLM a reliable reviewer? a comprehensive evaluation of LLM on automatic paper reviewing tasks. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 9340–9351. ELRA and ICCL, May 2024. URL https://aclanthology.org/2024.lrec-main.816/.

A PROMPTS

1. **Relevance Rating Prompt:**
   Now you are an image encoder for some images. You will be given an image. You must observe and analyze the image carefully, then produce an answer considering:

   Rating:
   Please provide a rating of this image to describe ḧow relevant or irrelevant it was to climate change, on a scale of 1 to 9 (1 means the most irrelevant, 9 means the most relevant.)

   Your answer must rigidly follow this template:
   Rating: an integer on a scale to 1 to 9.
   Reason: a short reason of the given rating.

2. **Arousal Rating Prompt:**
   Now you are an image encoder for some images. You will be given an image. You must observe and analyze the image carefully, then produce an answer considering:

   Rating:
   Please provide a rating of this image to describe ḧow calming or exciting it made participants feel" on a scale of 1 to 9 (1 means the most calm, 9 means the most excited.)

   Your answer must rigidly follow this template:
   Rating: an integer on a scale to 1 to 9.
   Reason: a short reason of the given rating.

3. **Valence Rating Prompt:**
   Now you are an image encoder for some images. You will be given an image. You must observe and analyze the image carefully, then produce an answer considering:

   Rating:
   Please provide a rating of this image to describe ḧow negative or positive the images appeared, on a scale of 1 to 9 (1 means the most negative, 9 means the most positive.)

   Your answer must rigidly follow this template:
   Rating: an integer on a scale to 1 to 9.
   Reason: a short reason of the given rating.

B SUPPLEMENTS

Intra-Class Correlation The ICC is computed as:

$$ICC = \frac{\sigma_B^2}{\sigma_B^2 + \sigma_W^2}$$

where: $\sigma_B^2$is the **between-group variance**, representing variability among different groups (e.g., different images). σ_W^2 is the **within-group variance**, representing variability within each group (e.g., different model ratings for the same image).

The between-group variance σ_B^2 is given by:

$$\sigma_B^2 = \frac{1}{k}\sum_{i=1}^{k} n_i(\bar{X}_i - \bar{X})^2$$

where: k is the number of groups (images), n_i is the number of ratings in group i, $\bar{X}_i$ is the mean of ratings in group i, $\bar{X}$ is the grand mean (overall mean across all groups).

The within-group variance σ_W^2 is given by:

$$\sigma_W^2 = \frac{1}{N-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)^2$$

where: N is the total number of ratings, X_{ij} is the rating of the jth observation in the ith group.

$ICC \approx 0$ indicates low agreement (random variation dominates).
$ICC \approx 0$ indicates high agreement (most variation is between groups).

Wilcoxon Signed-Rank Test. Let (b_i, p_i) be the baseline and LLaVa ratings for instance i, and define $d_i = b_i - p_i$. We discard any i where $d_i = 0$. For $d_i \neq 0$, let r_i be the rank of $|d_i|$ among these nonzero differences, with $r_i = 1$ for the smallest and $r_i = N$ for the largest (where N is the number of nonzero differences). We then define:

$$W^+ = \sum_{i:d_i>0} r_i, \quad W^- = \sum_{i:d_i<0} r_i.$$

We use W^+ to test $H_0 : \text{median}(d_i) \leq 0$ against $H_A : \text{median}(d_i) > 0$, and an analogous procedure tests $H_0 : \text{median}(d_i) \geq 0$. Under the null, W^+ follows a known distribution for small N, and an asymptotic normal approximation for large N:

$$z = \frac{W^+ - \frac{N(N+1)}{4}}{\sqrt{\frac{N(N+1)(2N+1)}{24}}}.$$

If z exceeds a critical value z_α, we reject H_0, indicating a significant directional shift.

Equivalence (TOST). If the Wilcoxon Signed-Rank Test indicates no significant difference, we turn to determine whether $|\bar{d}|$ is sufficiently small for practical equivalence. Specifically,

$$\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i,$$

$$\sigma_d = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(d_i - \bar{d}\right)^2},$$

$$\text{SE} = \frac{\sigma_d}{\sqrt{n}}.$$

and an equivalence margin δ is defined. We compute

$$t_\ell = \frac{\bar{d} + \delta}{\text{SE}}, \quad t_u = \frac{\bar{d} - \delta}{\text{SE}}.$$

We reject $H_0^-(\delta) : \bar{d} \leq -\delta$ if $t_\ell > t_{\alpha,n-1}$, and $H_0^+(\delta) : \bar{d} \geq +\delta$ if $t_u < -t_{\alpha,n-1}$. Rejecting both implies $\left|\bar{d}\right| < \delta$—no meaningful difference (Lakens, 2017).

Algorithm: Image Rating on LLaVa

1. **Input**:
   - Image data $\mathcal{P}$;
   - LLaVa Rating function $M(\cdot, \cdot)$;
   - Prompts settings $\mathcal{R}$.
2. **Implement**:
   - For each pair (p_i, r_i), let the LLaVa model output be $o_{\text{output}} = M(p_i, r_i)$;
   - Repeat 10 times, the model outputs $o_{\text{output}} = \{M_t(p_i, r_i)\}_{t=1}^{10}\}$.
3. **Benchmark Analysis** (Δo):
   - Use average human rating results as benchmark o_{base};
   - Use the average LLaVa rating Avg $o_{\text{output}} = \frac{1}{10}\sum_{i=1}^{10} o_{\text{output}_t}$ to be compared with benchmark o_{base};
   - Compute the variance $\Delta o =$ Avg $o_{\text{output}} - o_{\text{base}}$.
4. **Analyze Results**:
   - Compare Δo with 0;
   - Evaluate the performance of LLaVa by some test methods, e.g., directional test.

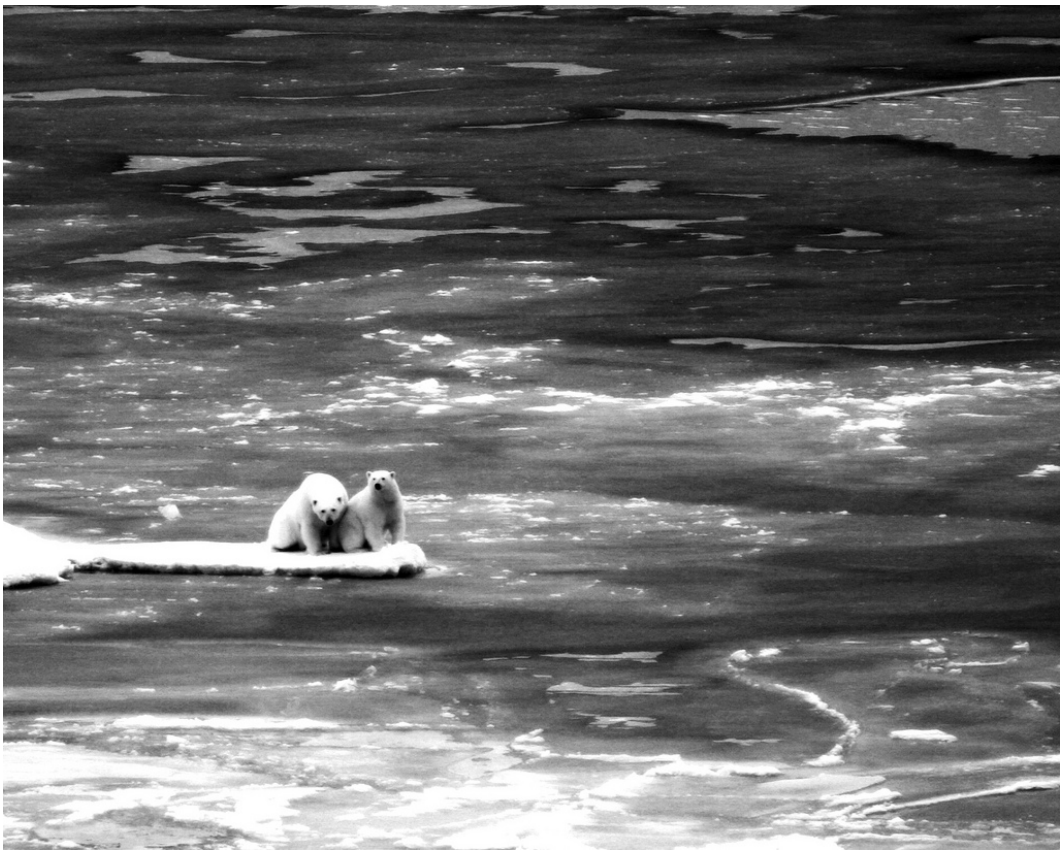

Figure 5: pic001