

TECHNICAL APPENDIX

Insight in Sight: Complaint Detection and Aspect-based Reasoning through Visually-grounded Reviews with VLLMs

No Author Given

No Institute Given

A Experimental Setup

A.1 Metrics

For CD task, we report scores for Accuracy, Weighted-F1 (w-F1), and Macro-F1 (m-F1). To ensure comparability with existing benchmarks for complaint detection, such as the MCI framework [4], we prioritize the w-F1 metric as it was emphasized in the referenced work. Additionally, we report Macro-F1, as it is widely regarded as a more suitable metric for tasks involving class-imbalanced datasets.

For ARG task, we evaluate on two of the most popular and reliable metrics [3, 1, 5, 2] for assessing LLM-generated outputs: Consistency and Relevance. Consistency measures the factual alignment of the generated output (ARG) with the input content (original review), ensuring no factual inaccuracies are introduced. Relevance, on the other hand, evaluates whether the generated output captures all key and important content from the source (original review).

A.2 Model Descriptions and Prompts

- **LLaVA-1.5 13B**: LLaVA is an open-source chatbot fine-tuned from LLaMA/Vicuna on GPT-generated multimodal instruction-following data. It operates as an auto-regressive language model based on the transformer architecture.
- **LLaVA-1.6 13B**: Building on LLaVA-1.5, LLaVA-1.6 enhances performance by increasing input image resolution and leveraging an improved visual instruction tuning dataset. These upgrades significantly improve OCR capabilities and common-sense reasoning.
- **LLaMA-3.2 11B Vision-Instruct**: LLaMA-3.2 Vision extends the LLaMA-3.1 text-only model into a multimodal framework. It uses an optimized transformer architecture, with instruction tuning tailored for visual recognition and image reasoning tasks.
- **Mistral-7B Instruct-v0.2**: This is a fine-tuned 7-billion-parameter model designed for instruction-following tasks, including question answering and

text generation. It enhances the base Mistral-7B model by aligning more closely with user intent and improving context awareness, offering strong performance in a compact and efficient design.

The various prompts regarding our investigations in this study are detailed in Tables 1 and 2.

Table 1: Prompts for the Complaint Detection (CD) and the Aspect-based Rationale Generation (ARG) Tasks.

Task	Prompt
CD	<p>{IMAGE} TITLE: {TITLE} REVIEW: {REVIEW}</p> <p>Here is a product’s image, accompanied by the user’s review and the review title. The task is to classify the review as either ‘0’ or ‘1’ based on the provided image and the text. A classification of ‘0’ indicates the review is ‘non-complaint’, meaning the user does not express any dissatisfaction with the product. A classification of ‘1’ indicates the review is a ‘complaint’, where the user is expressing a grievance or issue with the product. The output should be either ‘0’ or ‘1’.</p> <p>COMPLAINT LABEL:</p>
ARG	<p>TITLE: {TITLE} REVIEW: {REVIEW} COMPLAINT LABEL: {LABEL}</p> <p>Here is a product’s review accompanied by the review’s title and the associated label (either ‘0’ or ‘1’). A label of ‘0’ indicates the review is ‘non-complaint’, meaning the user does not express any dissatisfaction with the product. A label of ‘1’ indicates the review is a ‘complaint’, where the user is expressing a grievance or issue with the product.</p> <p>Generate a detailed aspect-based rationale for why the review falls under the given label. Ensure the rationale is grounded for the provided label (complaint or non-complaint) and focusses on the key aspects discussed in the review. The rationale should combine the aspect, description, and impact into a cohesive explanation tied to the assigned label. The rationale must not include the repetition of the user’s review, the review’s title and the label. The rationale should be in a single paragraph.</p> <p>RATIONALE:</p>




Table 2: Prompts for the zero-shot baseline setup and the generation for the Visually-grounded Enhanced Review (VER) Generation of step 0.

Task	Prompt
Zero-shot Baseline (CD+ARG)	<p>{IMAGE} TITLE: {TITLE} REVIEW: {REVIEW}</p> <p>Here is a product’s image, accompanied by the user’s review and the review title. The task is to classify the review as either ‘0’ or ‘1’ and then generate a detailed aspect-based rationale for the review’s classification based on the provided image and the text. A classification of ‘0’ indicates the review is ‘non-complaint’, meaning the user does not express any dissatisfaction with the product. A classification of ‘1’ indicates the review is a ‘complaint’, where the user is expressing a grievance or issue with the product. The classification output should be either ‘0’ or ‘1’. Ensure the generated rationale is grounded for the classification label and focuses on the key aspects discussed in the review. The rationale should combine the aspect, description, and impact into a cohesive explanation tied to the label.</p> <p>OUTPUT: The review is classified as</p>
VER Generation	<p>{IMAGE} TITLE: {TITLE} REVIEW: {REVIEW}</p> <p>Examine the given product image with the user’s review and the review title. If the review is incomplete or missing details, refer the image to identify additional information about the product, such as issues or positive features.</p> <p>The task is to enhance the review by making it more descriptive and precise, ensuring it is grounded in both the review’s text and the visual elements of the image. Avoid adding any unrelated details, ensure that the enhanced review remains coherent and accurately portrays the user’s overall experience.</p> <p>ENHANCED REVIEW:</p>

B Results

We present some representative outputs of our proposed method in Table 3.

Table 3: Sample results from the test set.

Inputs: <i>Image, Title, and Review</i>	Intermediate VER	Outputs: <i>CDO + ARG</i>
 <p>Title: Not worthy to buy</p> <p>Review: Not good. Stitching not good. Cloth worst, dirty material.</p>	<p>The black and white jacket in the image appears to be of poor quality. The stitching is not well done, and the material seems to be dirty or of low quality. The user's review suggests that the jacket is not worth buying due to its poor construction and unappealing appearance. The image does not provide any additional information about the product, but the review's title and content accurately describe the user's experience with the jacket.</p>	<p>CDO: 1</p> <p>ARG: In this review, the user explicitly expresses their displeasure with the quality of the black and white jacket's construction. The stitching on the jacket is poorly done, and the material appears to be of low quality and potentially dirty. The user's negative experience with the product is sufficient basis for labeling this review as a complaint. This assessment is in line with the aspects of the product discussed in the review, which reveals the user's dissatisfaction with the jacket's appearance and construction. Consequently, the impact of the review is to caution potential buyers against purchasing the jacket due to its perceived poor quality.</p>
 <p>Title: Very poor quality</p> <p>Review: Product is very cheap quality, very poor quality.</p>	<p>The product in the image is a black jacket, which the user has described as being of very poor quality. The jacket appears to be made of cheap materials, which may not provide the desired durability or comfort. The user's review suggests that the jacket may not be suitable for its intended purpose or may not meet their expectations in terms of quality and performance.</p>	<p>CDO: 1</p> <p>ARG: In this review, the user provides a critical assessment of the black jacket's quality. The user's description suggests that the jacket, which may have been intended to provide warmth or protection from the elements, is instead made of inferior materials. The user's disappointment with the product's quality is evident as they label it as 'very poor.' The impact of the jacket's poor quality is apparent, as the user implies that the jacket may not provide the necessary durability or comfort for its intended use. Given this information, the rationale for the complaint is clear. The user's review expresses their dissatisfaction with the product and highlights specific issues related to its quality.</p>
 <p>Title: Not good quality</p> <p>Review: The product was torn at the right leg which was hidden. Later, after trying could find the hole in the trouser.</p>	<p>The product in the image is a pair of camouflage pants, which the user found to be of poor quality. Upon closer inspection, the user noticed a tear in the pants, specifically at the right leg. This issue was not immediately apparent, but it was discovered upon trying the pants on. The hole in the trouser is a significant concern, as it may affect the pants' durability and overall functionality. The user's review highlights the importance of carefully examining the product before making a purchase to avoid potential issues like this one.</p>	<p>CDO: 1</p> <p>ARG: In this instance, the user's review highlights an issue with the product's quality, specifically with a pair of camouflage pants displaying a tear at the right leg. Though the tear was not initially visible, the impact on the pants' overall performance and longevity is significant. Consequently, the review falls under the "complaint" label as the user expresses their disappointment and concern regarding the product's deficiency. The aspect under scrutiny is the product's quality; the description reveals the presence of a hole, and the impact of this issue is its potential to impact the pants' durability and functionality.</p>

References

1. Fabbri, A.R., Kryscinski, W., McCann, B., Xiong, C., Socher, R., Radev, D.R.: Summeval: Re-evaluating summarization evaluation. *Trans. Assoc. Comput. Linguistics* **9**, 391–409 (2021). https://doi.org/10.1162/TACL_A_00373, https://doi.org/10.1162/tacl_a_00373
2. van Schaik, T.A., Pugh, B.: A field guide to automatic evaluation of llm-generated summaries. In: Yang, G.H., Wang, H., Han, S., Hauff, C., Zuccon, G., Zhang, Y. (eds.) *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14–18, 2024*. pp. 2832–2836. ACM (2024). <https://doi.org/10.1145/3626772.3661346>, <https://doi.org/10.1145/3626772.3661346>
3. Sheng, S., Xu, Y., Zhang, T., Shen, Z., Fu, L., Ding, J., Zhou, L., Wang, X., Zhou, C.: Repeval: Effective text evaluation with LLM representation. *CoRR* **abs/2404.19563** (2024). <https://doi.org/10.48550/ARXIV.2404.19563>, <https://doi.org/10.48550/arXiv.2404.19563>
4. Singh, A., Dey, S., Singha, A., Saha, S.: Sentiment and emotion-aware multi-modal complaint identification. In: *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*. pp. 12163–12171. AAAI Press (2022), <https://ojs.aaai.org/index.php/AAAI/article/view/21476>
5. Stureborg, R., Alikaniotis, D., Suhara, Y.: Large language models are inconsistent and biased evaluators. *CoRR* **abs/2405.01724** (2024). <https://doi.org/10.48550/ARXIV.2405.01724>, <https://doi.org/10.48550/arXiv.2405.01724>