2025 10th International Conference on Computer Science and Engineering (UBMK) | 979-8-3315-9975-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/UBMK67458.2025.11206985

Constructing a Turkish Complaint Dataset with Images and Benchmarking Vision Language Models

Kasım Murat Karakaya
TED University
Department of Software Engineering
Ankara, Türkiye
kmkarakaya@gmail.com

Elif Bengü Saraç
TED University
Department of Software Engineering
Ankara, Türkiye
bengu.sarac@tedu.edu.tr

Elif Ünal Çayır
TED University
Department of Software Engineering
Ankara, Türkiye
elif.unal @tedu.edu.tr

***Abstract*— This study aims to benchmark the capabilities of open-source Vision Language Models (VLMs) in understanding Turkish textual data within a multimodal framework, using e-commerce platforms as the evaluation domain. With the rapid growth of global e-commerce platforms, there has been a notable increase in customer complaints containing both text and images; however, the extent to which VLMs can understand Turkish and extract meaningful insights from combined text-image inputs remains unclear, as existing models have primarily been tested on English datasets. To address this gap, we constructed a new Turkish multimodal benchmark dataset by pairing real user complaint texts with product images collected from Turkish e-commerce platforms. We evaluated five open-source VLMs (Gemma-3-27b-it, Llama-3.2-11B-Vision-Instruct-bnb-4bit, Llava-v1.6-mistral-7b-hf-bnb-4bit, Qwen2.5VL:7b, and Llava:13b) to assess their ability to align, interpret, and reason over Turkish text and image data for tasks such as visual verification and complaint type classification. Experimental results showed that Gemma-3-27b-it achieved the highest overall accuracy of 60.8%, indicating strong performance in processing multimodal Turkish data, while Llava and Qwen demonstrated moderate results with overall accuracy scores of 37.8% and 47.3%, respectively. In contrast, the lowest performance was observed with 30.5%, highlighting the challenges in processing and aligning Turkish text-image data in low-resource multimodal settings. This study highlights how model architecture, size, and the level of visual-textual alignment critically influence VLM performance in understanding Turkish within multimodal contexts and contributes a new benchmark dataset to advance research in low-resource multimodal AI.**

Keywords — Vision Language Models (VLMs), Multimodal Benchmark, Turkish Language Processing, Text-Image Alignment, Low-Resource Languages

I. INTRODUCTION

Today, e-commerce platforms are widely used [1]. These sites, which provide services in many areas such as electronic products, food, cosmetics and clothing, play an important role in meeting the shopping needs of users. People now prefer e-commerce platforms, which are more easily accessible, instead of face-to-face shopping. This change in user preferences has led to the development of e-commerce platforms and an increase in product sales [2,3]. E-commerce platforms provide many advantages such as easy accessibility, time saving, and ease of payment. In addition to these advantages, e-commerce platforms also have various disadvantages such as sending wrong or damaged products, delays in the shipping process, sending products to the wrong address, making payments more than once, and difficulties in return processes. On e-commerce platforms, users can share these problems they experience in text and also by using images. User comments and shared product images on e-commerce platforms play an essential role in determining customer satisfaction. Images are important in terms of verifying users' complaints. However, there have not been enough studies, particularly for the Turkish context, on whether the damaged product images and comments are consistent with each other and how these images are evaluated by artificial intelligence systems. Vision Language Models (VLMs), which use images and text together, are helpful in evaluations.

Vision Language Models (VLMs) are a combination of visual and language models. These models can process and analyze data consisting of images and text simultaneously [4]. VLMs have many features such as image captioning, analyzing images with questions (VQA) [5], object or text detection in documents (grounding) [4,5,6]. VLMs are usually composed of three main parts: a large language model for understanding text, a visual encoder that converts the images into numerical data that the computer can understand, and an intermediate layer that converts the output of the visual encoder into something that the language model can understand. Thanks to this model structure, both text and visual information can be processed together to achieve better results. There are many open-source VLMs available today. These models differ in terms of parameter size, language model used, and visual encoder. Open-source VLMs such as Gemma, Llava, Llama, Mistral, and Qwen can be accessed on the Ollama website [7]. Each model is available in different sizes and is used in multimodal tasks. Depending on the model chosen, the success rate in the tasks also varies. Vision Language Models (VLMs) are mostly trained with English datasets. How VLMs perform in low-resource languages such as Turkish is a topic that needs to be studied.

979-8-3315-9975-1/25/$31.00 ©2025 IEEE

In this study, we aim to measure the performance of open-source VLMs based on the Turkish dataset. We investigate the ability of these models to (i) detect the compatibility and verifiability of a written Turkish complaint with its accompanying image, (ii) accurately classify the type of complaint (e.g., damaged product, wrong product, color difference) when both modalities are provided, and (iii) analyze the relative performance of different dimensions open-source VLMs on e-commerce data while assessing the impact of model size on their effectiveness. This paper addresses these research points and presents a structured benchmark and analysis of the readiness of existing open-source VLMs for practical use on Turkish e-commerce platforms.

II. Methods

A. Related Works

Vision Language Models (VLMs) have demonstrated significant success in multimodal tasks by unifying visual and textual data into a shared representation space. Radford et al. [8] introduced CLIP, which leveraged contrastive learning with natural language supervision to achieve strong zero-shot classification and retrieval performance while providing flexibility in multimodal representation. VisualBERT, developed by Li et al. [9], utilized a transformer-based architecture to align and fuse visual and textual representations, achieving effective results in tasks such as VQA and retrieval. Li et al. [10] further advanced this line of research with BLIP, which improved zero-shot performance in image-text matching tasks through a bootstrapped pre-training approach. Additionally, Flamingo, proposed by Alayrac et al. [11], demonstrated strong multimodal analysis capabilities under few-shot scenarios.

In recent years, multimodal learning research focusing on e-commerce has gained considerable momentum. Jin et al. [12] developed ECLIP, which learned instance-level representations of product images for retrieval and classification tasks, achieving notable results in e-commerce settings. Chen et al. [13] proposed a unified VLM architecture that integrates textual and visual attributes into a joint embedding space to improve similar product retrieval processes. Zheng et al. [14] successfully deployed a VLM-based retrieval system on large-scale e-commerce platforms, serving hundreds of millions of users in real-world scenarios. Gong et al. [15] introduced a CLIP-based zero-shot framework enabling attribute extraction for e-commerce products directly from images. Hu et al. [16] proposed a de-noised multimodal fusion method guided by visual cues to enhance retrieval stability on noisy e-commerce datasets.

However, most existing studies have been evaluated on English-centric, clean, and balanced datasets, leaving the performance of VLMs on low-resource and real-world e-commerce complaint data largely unexplored. To address this gap, Alayrac et al. [17] proposed the CLAIM method, which applies cross-lingual attention intervention to reduce visual-textual inconsistencies and hallucinations in VLMs, contributing to alignment improvement in low-resource languages. To further advance this area, our study constructs a new benchmark dataset composed of real user complaint texts and corresponding images collected from Turkish e-commerce platforms, focusing exclusively on multimodal (text + image) inputs. We analyze the performance of VLMs on multimodal verification and classification tasks using this dataset. Open-source VLMs, including Llama-3.2-11B-Vision-Instruct, Llava-v1.6-mistral, Gemma-3-27b-it, and Qwen-VL-Chat, are evaluated under zero-shot scenarios, revealing that multimodal inputs significantly enhance classification and reasoning performance. Our work provides a reliable benchmark for multimodal AI research in low-resource languages and demonstrates the practical potential of VLMs in the visual-supported analysis of Turkish e-commerce complaints.

B. Research Questions

In our study, we collected visual and text-based complaints of users commenting on e-commerce platforms. Using this data, we analyzed the performance of VLMs, focusing on the success of VLMs on Turkish data. We identified 3 research questions in the study:

- **RQ1:** Can open-source VLMs detect the compatibility and verifiability of a written Turkish customer complaint and the image of the complaint?
- **RQ2:** Can open-source VLMs correctly classify the type of complaint (damaged product, wrong product, color difference, etc.) in cases where complaint text and image are provided together?
- **RQ3:** Does the size of open-source VLMs affect the success rate in Turkish datasets?

Based on the above research questions, our study analyzes how different open-source VLMs perform using Turkish user complaint texts and complaint images.

C. Dataset

For our study, we created a dataset by collecting user complaints and images of complaints on online Turkish shopping websites. While collecting the data, we identified clothing categories (such as sweaters, pants, dresses) and found user complaints within these categories. We applied filtering options on shopping websites to find the complaints and images that are suitable for our study. By activating the photo review option (reviews where users add a photo of the product) and filtering the score, we analyzed the complaints of products with a score of 1 and 2. We collected the product images in .jpeg format in a separate file. We transferred the user complaint texts to an excel file. The Excel file consists of the columns *image_ID* which contains the id of the image, *complaint_text_tr* which contains the complaint text, *complaint_type* which specifies the type of complaint and *is_verifiable* which checks whether the complaint is compatible with the image. The labeling process in the *is_verifiable* and *complaint_type* sections was carried out by us. For the *complaint_type* attribute, we defined 7 categories; damaged/defective product, product with wrong size, product with color mismatch, product with missing parts, product with packaging problem, product that came wrong and others option. The "*Other*" category includes various types of complaints that do not clearly fit into the predefined labels. For example, complaints about low product quality (such as thin fabric, flimsy materials, or poor workmanship) or receiving previously used items fall into this category. An illustrative example is the user comment: “The quality is not good, I wouldn’t recommend it, I returned it,” which reflects general dissatisfaction with the product’s quality. We also labeled the compatibility of the complained product text with the complaint image as *is_verifiable* in the dataset. If there is compatibility, we labeled it as ‘*yes*’, otherwise as ‘*no*’, and if the complaint text is not fully visible in the image, we labeled it as ‘*partially*’. Figure 1 displays the distribution of labels within the ‘*is_verifiable*’ field in our dataset, with ‘*Yes*’ at 36.1%, ‘*No*’ at 32.5%, and ‘*Partially*’ at 31.5%. This figure presents the proportion of verifiability statuses included in our study.

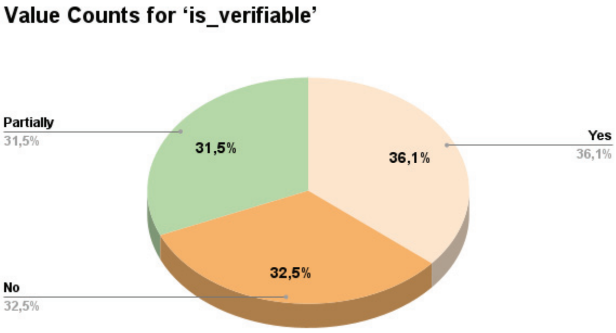

Fig. 1. Value counts for ‘*is_verifiable*’.

Figure 2 illustrates the distribution of complaint types within our dataset, showing that *Defect/Damaged* complaints constitute the largest portion at 42.7%, followed by *Other* at 25.2%. This figure presents the proportion of each complaint type considered in our study.

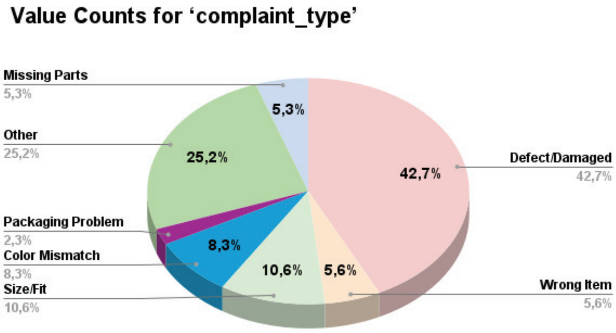

Fig. 2. Value counts for ‘*complaint_type*’

Table I and Figures 3 and 4 present sample text-image pairs from our Turkish multimodal dataset. Both complaints relate to defective products, with Figure 3 showing a case where the model incorrectly predicted verifiability, while Figure 4 was correctly identified. In both cases, the models accurately classified the complaint type as “*Defect/Damaged*”. These examples highlight the dataset’s complexity and the models’ performance in processing Turkish e-commerce complaints. We created the dataset by collecting 300 visual and text-based complaint data. The created dataset will be shared on the Kaggle platform after the publication of the paper.

TABLE I
SELECTED COMPLAINT CASES WITH VERIFIABILITY AND CLASSIFICATION OUTCOMES

Fig. No	**Complaint Text (TR)**	**Complaint Type**	**Is Verifiable**	**Predicted Verifiable**	**Predicted Complaint Type**
Fig.3	Utanmıyor musunuz defolu ürün göndermeye hem ipliği kaçmış hem belini dar dikmişler uyması için pile koymuşlar	Defect/ Damaged	No	Yes	Defect/ Damaged
Fig.4	paça boyları eşit değil üstelik yamuk yumuk kesilmiş	Defect/ Damaged	Yes	Yes	Defect/ Damaged

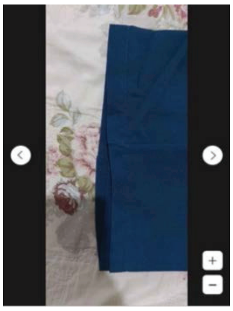

Fig. 3. Example Complaint Image of Blue Trousers with Alleged Sizing Issues.

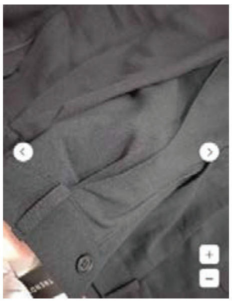

Fig. 4. Example Complaint Image of Black Trousers with Alleged Defects.

III. Experiments

A. Experimental Setup

In this study, we evaluate the performance of open-source vision language models, which are detailed in Table II with their abbreviations and main specifications. We chose these models because they are among the most popular open-source VLMs on the Ollama [7] website and they fit the VRAM capacity provided by Kaggle. We did not fine-tune the models and adopted a zero-shot learning approach in the experiment. We conducted our experimental evaluation on a dataset consisting of 300 customer complaint samples tested with 5 different VLMs. We gave the collected data and images to the VLM together with the system prompt. In the system prompt, we asked the model to first analyze the image, then consider the complaint text together with the image to determine whether the complaint is visually verifiable from the image. We asked the model to classify the results and return them in a structured JSON output. The JSON output consists of 2 fields, *predicted_verifiable* and *predicted_complaint_type.* We used the default values of each language model and did not change parameters such as top-k, temperature, etc. We performed the testing phase on the Kaggle platform. We gave each model up to 5 rights to produce a valid JSON output for each sample. Each "right" corresponds to a single response generated by the model, regardless of whether the output is valid or not. If the model failed to produce a valid JSON or missed one of the required fields, we automatically triggered a retry. We continued this process for up to 5 rights. We repeated this process until the model produced a valid JSON or used all 5 rights. If the model failed to produce a valid JSON in all 5 rights, we excluded that sample from evaluation. We only scored and analyzed responses that were valid JSON and contained both required fields. The number of malformed JSON responses and hallucinated class labels is reported, showing that smaller models are not always successful. For example, Mistral failed to produce 110 valid JSON outputs and generated 40 non-existent class labels, indicating hallucinations in classification.

TABLE II
VISION LANGUAGE MODELS: ABBREVIATIONS, PARAMETERS AND MEMORY USAGE

Model Abbreviation	Full Model Name	Parameter Size (B)	Approx. VRAM Requirement (GB)
Gemma	Gemma-3-27b-it	27B	22–24 GB
Llama	Llama-3.2-11B-Vision-Instruct-bnb-4bit	11B	14–16 GB
Llava	Llava:13b	13B	16–18 GB
Mistral	Llava-v1.6-mistral-7b-hf-bnb-4bit	7B	10–12 GB
Qwen	Qwen2.5vl:7b	7B	10–12 GB

B. *Experimental Results*

Our experimental results show that open-source VLMs can partially verify and assess the compatibility of Turkish customer complaints when paired with product images. The models achieved up to 60.8% accuracy with Gemma, demonstrating strong sensitivity in identifying visually verifiable complaints. In complaint type classification, the models consistently produced outputs aligned with the complaint categories in the inputs. In contrast, Mistral showed the lowest overall model accuracy (30.5%), demonstrating significant performance variation across different model architectures.

Verifiability Classification Performance: Table III summarizes the verifiability classification performance of the evaluated open-source VLMs within our multimodal complaint verification pipeline, where verifiability classification refers to the automated assessment of whether a Turkish customer complaint can be visually verified using the associated product image. In our implementation, this classification guides downstream complaint resolution processes by determining if image evidence supports the user's claim, categorizing complaints as "*Yes*", "*No*", or "*Partially*" verifiable. Among the models, Gemma achieved the highest accuracy (47%) and demonstrated relatively balanced precision (48%) and recall (45%), resulting in the highest F1-score (40%), indicating strong overall performance in detecting visually verifiable complaints. Qwen showed moderate accuracy (44%) but lower precision (34%) and recall (44%), reflecting fair but less balanced performance. Llama and Llava models yielded similar moderate accuracies around 36–38%, with relatively balanced precision and recall values but lower F1-scores, indicating average effectiveness. Mistral had the lowest accuracy (33%) and the weakest precision (23%), recall (33%), and F1-score (26%), highlighting challenges in handling verifiability detection. The precision, recall, and F1-score values reported in the table are calculated based on macro averages.

Overall, all models struggled with partial verifiability cases, with low F1-scores across the board, underscoring the need for further fine-tuning to improve handling of nuanced, partially verifiable complaints in Turkish e-commerce settings.

TABLE III
VERIFIABILITY CLASSIFICATION METRICS ACROSS MODELS

Model Name	Accuracy	Precision	Recall	F1-Score
Llama	0.38	0.38	0.38	0.36
Mistral	0.33	0.23	0.33	0.26
Gemma	0.47	0.48	0.45	0.40
Qwen	0.44	0.34	0.44	0.34
Llava	0.36	0.38	0.36	0.28

Figure 5 illustrates the accuracy comparison of open-source VLMs in predicting the verifiability of Turkish e-commerce complaints using multimodal inputs.

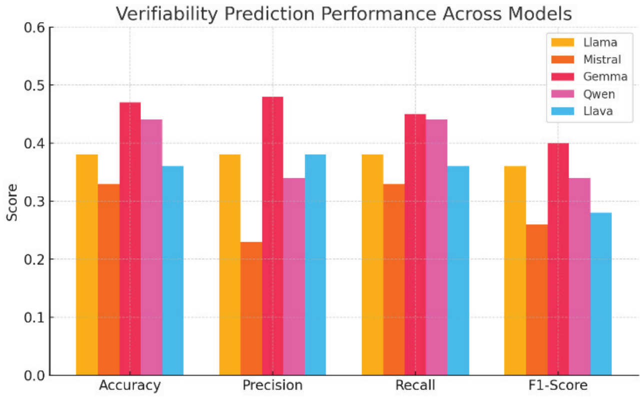

Fig. 5. Comparison of Verifiability Prediction Accuracy Across Open-Source VLMs on Turkish E-commerce Complaint Data

Complaint Classification Performance: Complaint classification in our application refers to the automated categorization of Turkish customer complaints into predefined types (e.g., *Defect/Damaged, Wrong Item, Color Mismatch*) by leveraging both the complaint text and its corresponding product image within a structured multimodal pipeline. This classification enables downstream automation in customer support, returns processing, and quality control by aligning user-reported issues with operational categories. Table IV summarizes the classification performance across models, showing that Gemma achieved the highest accuracy (74%) and F1-score (69%), demonstrating strong and balanced capabilities across categories. In contrast, Qwen attained moderate performance (accuracy: 51%, F1: 49%), while Llava and Mistral displayed lower accuracies (40% and 28%, respectively) and Mistral produced hallucinated class outputs absent from the ground truth labels, indicating stability issues during classification. Notably, all performed well in categories with clear visual cues (e.g., *Color Mismatch, Defect/Damaged*) but struggled in less visually explicit or imbalanced categories, with "*Other*" and "*Missing Parts*" yielding lower F1-scores across models. The F1-scores are calculated based on macro averages. These findings highlight that while open-source VLMs can effectively classify clear multimodal complaint types, they require targeted fine-tuning to improve consistency across nuanced and low-resource categories for robust e-commerce complaint automation.

TABLE IV
COMPLAINT CLASSIFICATION METRICS ACROSS MODELS

Model Name	Accuracy	Precision	Recall	F1-Score
Llama	0.57	0.50	0.48	0.47
Mistral	0.28	0.21	0.16	0.14
Gemma	0.74	0.76	0.70	0.69
Qwen	0.51	0.66	0.55	0.49
Llava	0.40	0.27	0.26	0.25

Figure 6 illustrates the comparative performance of the evaluated open-source VLMs on complaint type classification using Turkish e-commerce complaint data.

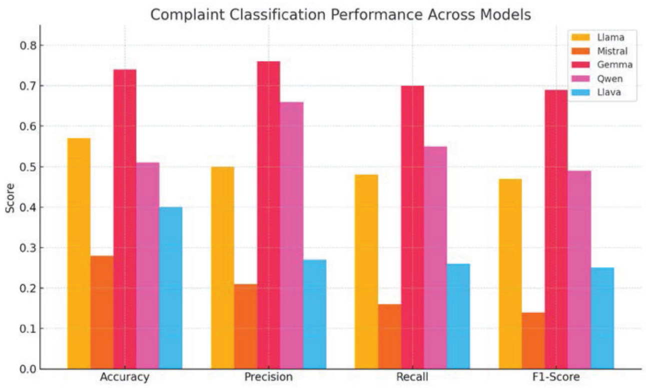

Fig. 6. Performance of Open-Source VLMs on Complaint Type Classification

Comparative Analysis of Model: Table V presents the overall performance ranking of the evaluated open-source VLMs, integrating their verifiability and complaint type classification results to assess their suitability for multimodal complaint analysis in Turkish e-commerce data. Gemma outperformed other models with the highest overall performance score (60.8%), driven by strong complaint classification accuracy (74.3%) and relatively higher verifiability accuracy (47.3%), reflecting its balanced capability in both clear visual verification and category classification. Llama followed with a moderate overall score (47.7%), while Qwen demonstrated comparable overall performance (47.3%), showing strength in verifiability recall but lower classification stability. Llava and Mistral exhibited lower overall performance scores (37.8 % and 30.5%, respectively), with high error rates in both verifiability and complaint type tasks, indicating limitations in stability and consistency across categories and visual verification challenges. These results collectively highlight that while larger and well-aligned models like Gemma provide strong multimodal analysis capabilities, model architecture, training alignment, and structured prompting are crucial for achieving high performance across both verifiability and complaint classification tasks in real-world e-commerce complaint processing pipelines.

TABLE V
OVERALL MODEL PERFORMANCE RANKING

Model Name	Sample Size	Verify Error Rate	Complaint Error Rate	Verify Accuracy	Complaint Accuracy	Overall Performance
Llama	300	0.617	0.430	0.383	0.570	0.477
Mistral	190	0.674	0.716	0.326	0.284	0.305
Gemma	300	0.527	0.257	0.473	0.743	0.608
Qwen	300	0.560	0.493	0.440	0.507	0.473
Llava	300	0.640	0.603	0.360	0.397	0.378

Figure 7 presents a comparative analysis of the F1-scores achieved by the evaluated open-source VLMs on verifiability prediction and complaint type classification tasks. The results demonstrate that Gemma consistently outperforms the other models across both tasks, reflecting its stronger capability in aligning and interpreting multimodal Turkish complaint data. Llama and Qwen exhibit moderate F1-scores in both tasks, indicating reasonable but less consistent performance, while Llava and Mistral show lower F1-scores, particularly in complaint classification, highlighting their limitations in processing and reasoning over combined text-image inputs in low-resource Turkish datasets.

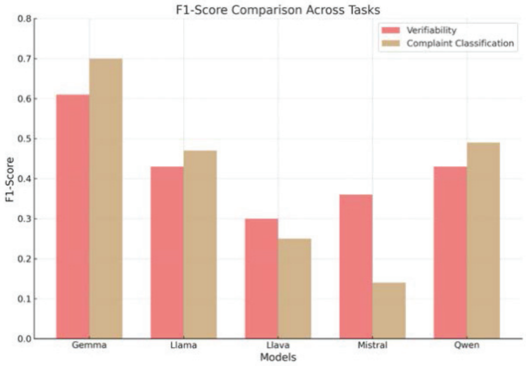

Fig. 7. F1-Score Comparison of open-source VLMs on Verifiability and Complaint Classification Tasks

IV. CONCLUSION

In this study, we tried to measure the performance of open-source Vision Language Models (VLMs) on Turkish dataset. Based on the RQ1 and RQ2 research questions, we tested whether the models were able to detect the correspondence between Turkish complaint texts and images and correctly classify the type of complaint. Based on the

RQ3 research question, we examined the effect of size on the success rate of open-source VLMs.

Among the evaluated models, Gemma achieved the highest overall performance (60.8%), with strong results in both complaint classification (74.3%) and verifiability accuracy (47.3%). Llama (47.7%) and Qwen (47.3%) showed moderate performance, while Llava (37.8%) and Mistral (30.5%) underperformed due to high error rates. To answer the RQ1 research question, we focused on the Verifiability Classification Report results and found that our Gemma model showed better results than other models. To answer our RQ2 research question, we focused on the Complaint Type Classification Report results and found that our Gemma model performed better in classification rate. While our Gemma model showed high F1 scores especially in *Defect/Damaged, Color Mismatch, Wrong Item* classes, our Mistral model produced 40 non-existent classes and hallucinated. Based on these results, the answer to the RQ3 research question is that size is a factor in the success rate of the model. Small models do not always succeed, while larger models are more likely to do so.

In the future, we plan to fine-tune open-source VLMs and compare their performance against zero-shot versions. Although our current dataset is limited due to time constraints, we aim to significantly increase its size. So far, we have tested five models, but we intend to expand the number of models and include larger models. We will improve the "*Other*" label by breaking it down into more specific subcategories such as Low Quality, Previously Used, and Ambiguous, informed by frequently observed complaint types. This will help reduce label noise and improve the clarity and consistency of the classification. Additionally, we aim to enable the model to generate context-aware and explanatory responses in Turkish, based on multimodal complaint data that will be provided both visually and textually. In this context, we will evaluate the generative capabilities of VLMs by examining whether they can produce meaningful and contextually appropriate outputs in Turkish.

V. Limitations

In this study, we were unable to utilize larger open-source language models due to hardware limitations. Since the experiments were conducted on the Kaggle platform, the available 15 GB of RAM restricted our model selection. Additionally, the dataset contains only 300 manually collected samples, but we plan to significantly increase its size in the future. We were also limited to testing five models, and working with very large models such as those with 70 billion parameters was not feasible due to memory constraints. Another limitation concerns the labeling process within the dataset. As illustrated in Figure 2, the "*Other*" category accounts for 25.2% of all samples, making it the second most frequent complaint type after "*Defect/Damaged"*. We first analyzed 50 pilot samples, examining both complaint texts and related images to create the category structure. Based on this analysis, we defined the core complaint categories and finalized the classification scheme. Then, we conducted a comprehensive annotation process on 300 samples using these categories. While the "*Other*" category had a low frequency in the initial 50 samples, it became the second most frequently labeled class after annotating the full 300 samples. While this approach helped minimize noise in the structured classes and preserved semantic diversity, the wide scope of the "*Other*" category may introduce ambiguity in model learning and evaluation.

References

[1] C. Wang, Y. Wang, and S. Zhong, "Analysis of Influencing Factors on the Development of Retail E-commerce in China," *2009 International Symposium on Information Engineering and Electronic Commerce*, May 2009, pp. 441–444.

[2] L. Wu and C. Li, "Explore the Application of Computer Network Technology in E-Commerce," *2022 International Conference on Computer Network, Electronic and Automation (ICCNEA)*, Sep. 2022, pp. 163–167.

[3] J. K. Shim *et al., The International Handbook of Electronic Commerce.* Chicago: Glenlake Publishing Company, Ltd., 2000.

[4] H. Li, H. F. Li, and Y. Shi, "Vision Language Models: Methods, Datasets and Training, and Bibliometric Analysis," *2025 4th International Conference on Artificial Intelligence, Internet and Digital Economy (ICAID)*, Apr. 2025, pp. 242–245.

[5] D. Das, D. Talon, M. Mancini, Y. Wang, and E. Ricci, "One VLM to Keep It Learning: Generation and Balancing for Data-Free Continual Visual Question Answering," *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, Feb. 2025, pp. 5635–5645.

[6] J. Huang, C. Limberg, S. M. N. Arshad, Q. Zhang, and Q. Li, "Combining VLM and LLM for Enhanced Semantic Object Perception in Robotic Handover Tasks," *2024 WRC Symposium on Advanced Robotics and Automation (WRC SARA)*, Aug. 2024, pp. 135–140.

[7] Ollama, "Ollama," 2024. [Online]. Available: https://ollama.com/. [Accessed: 29-Jun-2025].

[8] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proc. 38th Int. Conf. Mach. Learn. (ICML)*, Feb. 2021. [Online]. Available: https://github.com/openai/clip

[9] L. H. Li, M. Yatskar, D. Yin, C. J. Hsieh, and K. W. Chang, "VisualBERT: A simple and performant baseline for vision and language," *arXiv preprint* arXiv:1908.03557, Aug. 2019.

[10] J. Li, D. Li, C. Xiong, and S. C. H. Hoi, "BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation," *arXiv preprint* arXiv:2201.12086, Jan. 2022.

[11] J.-B. Alayrac *et al*., "Flamingo: A visual language model for few-shot learning," *arXiv preprint* arXiv:2204.14198, Apr. 2022.

[12] L. Jin, X. Liu, X. Wang, Z. Zhou, and H. Li, "ECLIP: An effective instance-level vision-language pretraining framework for e-commerce," in *Proc. 32nd ACM Int. Conf. on Information & Knowledge Management (CIKM)*, Oct. 2023, pp. 2783–2792.

[13] B. Chen, L. Jin, X. Wang, D. Gao, W. Jiang, and W. Ning, "Unified vision-language representation modeling for e-commerce same-style products retrieval," in *Companion Proc. of the ACM Web Conf. 2023 (WWW '23 Companion)*, Apr. 2023, pp. 1190–1195.

[14] X. Zheng, F. Lv, Z. Wang, Q. Liu, and X. Zeng, "Delving into e-commerce product retrieval with vision-language pre-training," in *Proc. 46th Int. ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR '23)*, Jul. 2023, pp. 2303–2308.

[15] J. Gong, M. Cheng, H. Shen, P.-Y. Vandenbussche, J. Jenq, and H. Eldardiry, "Visual zero-shot e-commerce product attribute value extraction," in *Proc. 2025 Conf. of the North American Chapter of the Assoc. for Computational Linguistics: Human Language Technologies (NAACL), Industry Track*, Apr. 2025, pp. 460–469.

[16] Z. Hu, L. Jin, D. Gao, and W. Ning, "De-noised vision-language fusion guided by visual cues for e-commerce product search," in *Proc. 2025 Int. Conf. on Multimedia Retrieval (ICMR)*, May 2025.

[17] Y. Ye, B. Liang, H. Wu, Z. Wang, and L. Li, "CLAIM: Mitigating multilingual object hallucination in large vision-language models with cross-lingual attention intervention," in *Proc. 2025 Conf. of the North American Chapter of the Assoc. for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, Jun. 2025, pp. 2350–2362.