Counterfeit Medicine Detection by Visual Inspection of Package Design Using Multimodal LLMs with Text and Image Prompt Engineering

Yona Zakaria[1,4], Eiki Ishidera[1,2], Rui Ishiyama[1,2], Tomokazu Matsui[1,3], Hiroiko Suwa[1,3], Yuki Matsuda[1,5], and Keiichi Yasumoto[1,3]

[1]Nara Institute of Science and Technology, Nara, Japan
[2]NEC Corporation, Tokyo, Japan
[3]RIKEN Center for Advanced Intelligence Project, Tokyo, Japan
[4]The University of Dodoma, Dodoma, Tanzania
[5]Okayama University, Okayama, Japan

ABSTRACT

This study explores the potential use of multimodal large language models (LLMs) in detecting counterfeit drugs through visual inspection of medicine packaging designs. Specifically, we investigate how can ChatGPT-4o provide clear explanations of design differences between genuine and counterfeit packaging. We combine structured textual prompts with three distinct image configurations: (1) query-only images, (2) query-plus-reference images, and (3) query-plus-reference-plus-difference. This setup allows for context-aware comparative analysis, helping the model to effectively identify and explain packaging design inconsistencies—key indicators of counterfeit or substandard medicines. Experimental results show that ChatGPT-4o achieves a binary classification accuracy of up to 74.6% in distinguishing authentic from counterfeit medicine packaging. Furthermore, user evaluations reveal that ChatGPT-4o delivers high levels of clarity, ease of understanding, reliability in identifying discrepancies, detail, and overall quality of analysis. These findings underscore the notable potential of ChatGPT-4o to enhance explainability and usability in counterfeit detection workflows, particularly by enabling accurate, actionable insights without requiring training on counterfeit-specific datasets, which are often challenging to collect.

Keywords: Fake drug, Counterfeit medicine, Packaging inspection, Large language models, ChatGPT, Vision language model, Prompt engineering, AI-assisted visual inspection

1. INTRODUCTION

Counterfeit medicines threaten public health, especially in low-resource countries where regulatory infrastructure and consumer-level verification tools are often lacking.[1] Detecting these counterfeit medicines is a multifaceted challenge. One of the most immediate and accessible screening methods for consumers is through visual inspection of pharmaceutical packaging.[2,3] Many counterfeit exhibit spelling errors, misaligned logos, inconsistent font or unverifiable manufacturer names.[4] In rare cases where they produce high-quality forgeries, visual inspection alone may not be sufficient. In such instances, chemical analysis or laboratory testing is required. WHO and drug regulatory authorities recommend visual inspection as the first-line method for quick screening of counterfeit medicines. Nevertheless, manual inspection can be challenging, especially when a counterfeit closely resembles its authentic counterpart. This underscores the need for advanced tools that can assist in the identification of subtle design discrepancies.[5]

Recent advances in large language models (LLMs) have led to the development of powerful multimodal models that integrate visual and textual data.[6] Among these developments, conversational agents like ChatGPT[7] stand out as versatile multimodal LLM that offers an intuitive natural language interface that simplifies user interactions

Further author information: (Send correspondence to Yona Zakaria)
Yona Zakaria.: E-mail: zakaria.yona.aa8@is.naist.jp

while leveraging extensive multimodal knowledge bases. ChatGPT has already been deployed with success in several real-world applications such as deepFake detection,[8–10] face verification,[11] fake image detection[10] and handwriting verification[12] offering explainable outputs that enhance trust and usability. Its adaptability and ability to generate clear, human-readable explanations make it particularly well-suited for tasks requiring detailed analysis, such as our task of visual inspection of medicine packages.

Building on our previous work in image retrieval and alignment using keypoint-based image matching,[13] we leverage ChatGPT-4o multimodal capabilities to the visual inspection of medicine packaging to identify subtle design inconsistencies such as logos, typography, or layout that often signal counterfeit medicines while providing user-friendly explanations of discrepancies.

The main contributions of this paper are:

- We introduce a novel application of a multimodal ChatGPT-4o to detect counterfeit medicines in real-world scenarios. Through structured text prompts and varied image inputs, our approach enables practical analysis of design discrepancies—key indicators of poor quality or counterfeit packaging—without relying on specialized training datasets, which are often challenging to collect.
- Comprehensive input configurations. We propose three image input approaches—query-only, query-plus-reference, and query-plus-reference-plus-difference— to guide the LLM in identifying discrepancies. These tailored configurations enhance accuracy and explainability in the visual inspection task.
- Explainability of visual inspection-based detection. We leverage LLMs to provide detailed reasoning alongside binary classification, which is essential to user understanding in real-world applications.

2. METHODOLOGY

This study employs ChatGPT-4o, a multimodal Large Language Model (LLM), to enhance visual inspection workflows for identifying design inconsistencies in pharmaceutical packaging designs. We use an input query-reference-assisted visual inspection setup, where a query image is analyzed alongside a reference image. By exploring prompt engineering strategies and input image configurations, we examine how the reference images and difference images influence the detection accuracy and interpretability of LLM-generated descriptions. With their ability to process various input modalities, LLMs are particularly well suited to generate detailed explanations when provided with structured image and text data.[14] This strength positions them as valuable tools for detecting subtle packaging discrepancies, such as logo misalignments, font irregularities, or color mismatches, which resemble fine-grained semantic indicators used in media forensics.[8] The complete process of using ChatGPT-4o to improve visual inspections is illustrated in Fig. 1.

2.1 Text prompt

Prompt engineering involves designing text prompts that incorporate context, instructions, and examples to guide LLMs in understanding tasks and generating accurate, context-specific, and interpretable responses without retraining. Techniques such as zero-shot, few-shot, and Chain-of-Thought (CoT) prompting have been successfully applied in practical real-world tasks like deepFake detection,[8–10] face verification,[11] and handwritingverification.[12]

In this work, we adopt CoT-based prompting[14] to enhance ChatGPT-4o's visual analysis for counterfeit detection in pharmaceutical packaging. Our approach goes beyond binary classification by eliciting detailed reasoning focused on key design elements such as spelling errors, logo alignment, and font consistency to identify subtle discrepancies. Structured prompts not only yield actionable and interpretable outputs but also reduce response variability, thereby improving the model's consistency. This demonstrates the potential of CoT-based prompting to enhance visual inspection and descriptions in counterfeit medicine detection.

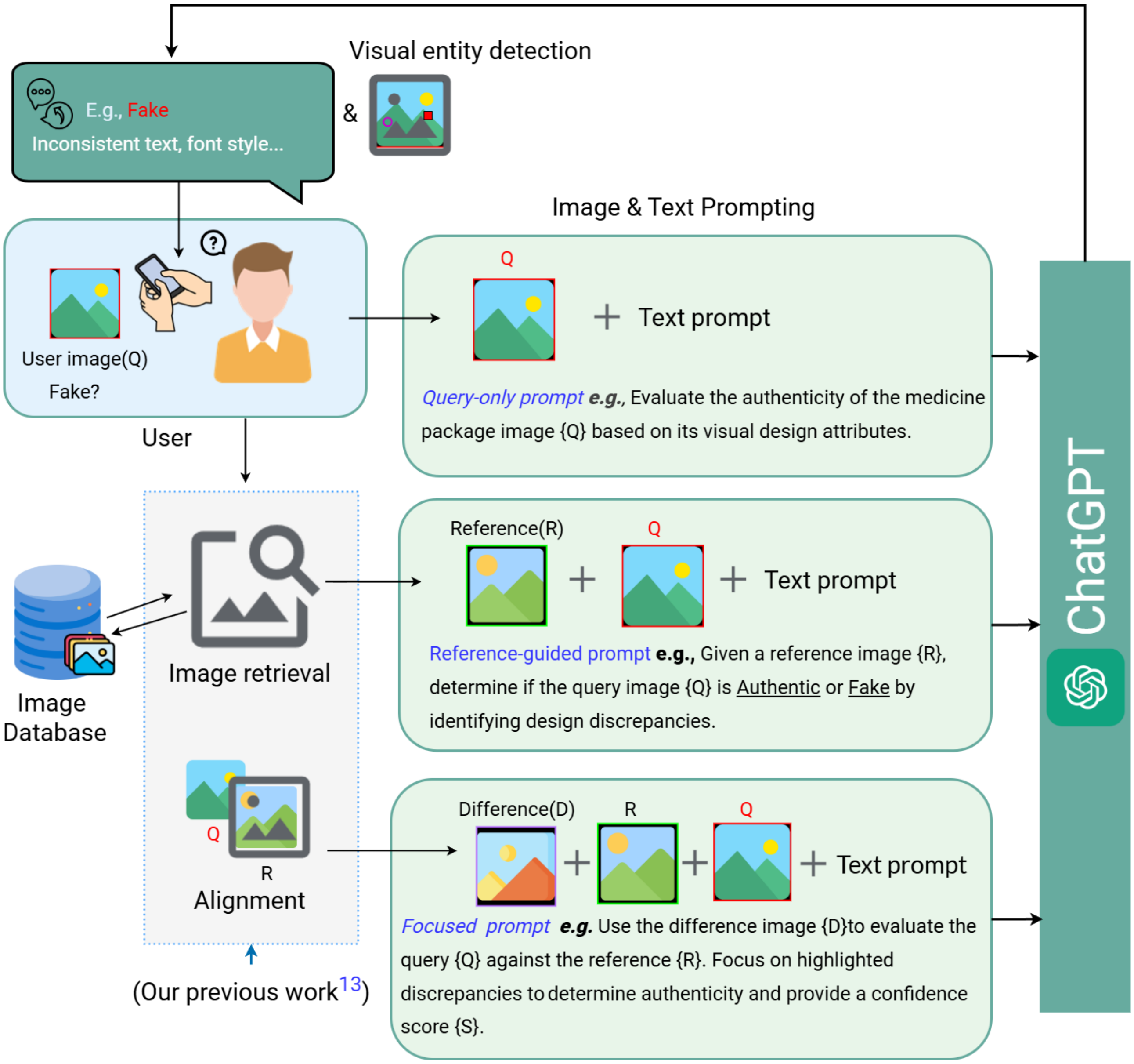

Visual entity detection
E.g., Fake
Inconsistent text, font style...
&
Image & Text Prompting
User image(Q)
Fake?
User
Q
+ Text prompt
Query-only prompt e.g., Evaluate the authenticity of the medicine package image {Q} based on its visual design attributes.
Image Database
Image retrieval
Reference(R)
Q
+
+ Text prompt
Reference-guided prompt e.g., Given a reference image {R}, determine if the query image {Q} is Authentic or Fake by identifying design discrepancies.
Q
R
Alignment
(Our previous work13)
Difference(D)
R
Q
+
+
+ Text prompt
Focused prompt e.g. Use the difference image {D}to evaluate the query {Q} against the reference {R}. Focus on highlighted discrepancies to determine authenticity and provide a confidence score {S}.
ChatGPT

Figure 1. The complete process of leveraging ChatGPT-4o for visual inspections of medicine packages. The three configurations differ by the number of images inputted: the query image (captured by the user for inspection), the query + reference images (2-image configuration), and the query + reference + difference images (3-image configuration). Structured text prompts, guided by the Chain-of-Thought (CoT) reasoning framework, direct the model to detect and explain design inconsistencies in pharmaceutical packaging.

2.2 Image configurations

Text and image prompts form the backbone of visual question-answering (VQA) tasks for multimodal Large Language Models (LLMs). This study introduces three structured prompt configurations to evaluate ChatGPT-4o's ability to perform visual inspections of pharmaceutical packaging. Each configuration employs textual guidance combined with a distinct number of image inputs to guide the model in identifying design discrepancies, which are critical indicators of counterfeit or substandard packaging. These configurations are categorized as follows:

- **Query-only prompt(1-image).** In the query-only prompt, the ChatGPT-4o is tasked to analyze the query image only, without the aid of reference images. As illustrated by the left image in Fig. 2, this represents the simplest configuration, where the ChatGPT-4o evaluates visual attributes such as logos, text alignment, layout, and color schemes in isolation. The ChatGPT-4o is required to provide a classification on whether the packaging is authentic or fake, accompanied by a confidence score that reflects the certainty of its assessment. This prompt simulates typical interaction between the user and the ChatGPT.
- **Reference-guided prompt(2-images).** The reference-guided prompt incorporates a reference image, depicted as the middle image in Fig. 2. The ChatGPT-4o is tasked to compare the query image (left) against the reference image (middle) to identify design inconsistencies, such as logo misalignment or incorrect text. This direct comparison enables the ChatGPT-4o to detect discrepancies that might not be evident in the baseline prompt.
- **Focused prompt(3-images).** The focused prompt builds upon the reference-guided approach by introducing a difference image, depicted as the rightmost image in Fig. 2. The difference image is image alignment and created through pixel-wise comparison, as introduced in our previous work,[13] of the query and reference images, highlighting discrepancies such as font irregularities, color mismatches, and misaligned design elements.

Reference images are retrieved from an authenticated pharmaceutical packaging database using our prior method[13] for efficient matching. While deep learning-based approaches [15] are viable, we prioritize speed. Additionally, in the second and third configurations, images are merged into a single composite before input into ChatGPT-4o, optimizing computational cost and comparison time.[11]

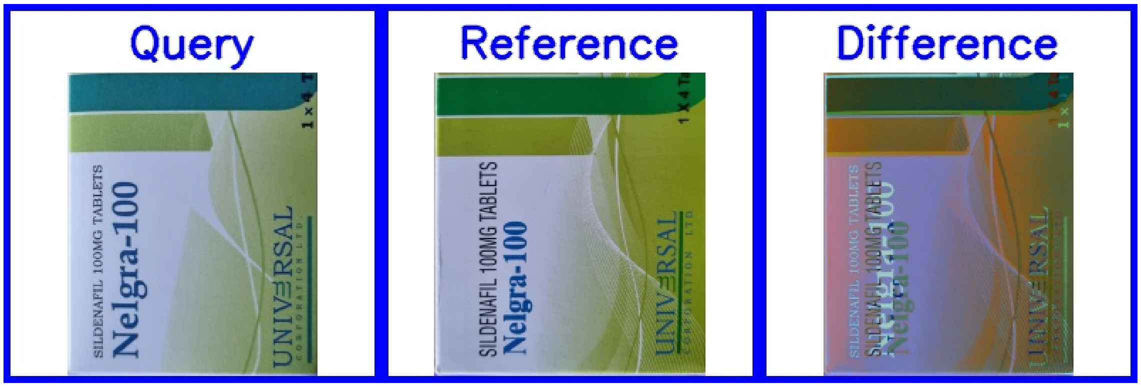


Figure 2. Graphical representation of the image inputs provided to ChatGPT: (left) Query image: the primary input representing pharmaceutical packaging potentially at risk of counterfeiting, serving as the focal point for visual inspection. (middle) Reference image: sourced from a database of verified pharmaceutical packaging using a keypoint-based image matching algorithm, providing a reliable baseline for identifying critical design discrepancies. (right) Difference image: produced via pixel-wise comparison between the query and reference images, highlighting visual inconsistencies such as misaligned logos, irregular fonts, or color deviations. The difference image strategically focuses the ChatGPT's analysis on relevant areas, enhancing the precision and efficiency of the visual inspection process.

2.3 Evaluation

We evaluate the performance of ChatGPT-4o using two primary metrics: accuracy, and rejection rate similar to prior studies.[8,11] Accuracy measures the proportion of correct classifications—whether the packaging is authentic or fake—relative to the total number of valid responses provided by the ChatGPT-4o. This metric highlights the model's capability to identify discrepancies effectively and make accurate classifications. Rejection rate quantifies the proportion of prompts where the ChatGPT-4o refrains from providing a definitive response, instead stating limitations such as "I cannot assist with this request." This metric evaluates the model's ability to handle challenging queries while avoiding potentially misleading answers. The experimental setup adhered to the parameter configurations and best practices outlined by DeAndres-Tame et al.[11]

2.4 Dataset

The data set used in this study to evaluate the ability of ChatGPT-4o to detect design discrepancies in pharmaceutical packaging was originally collected by ourselves[13] in a real African market on-site, where the problem of fake medicine is serious. The data set consists of images of drug packages collected using smartphone cameras from Tanzanian retail outlets and is further enriched with the data set of mobile captured drug packages,[16] resulting in 245 unique categories and a total of 4,581 images. For this study, a subset of 20 different package designs was selected as reference images. To simulate real-world counterfeit scenarios, replica packages of the selected reference images were created by mimicking the authentic designs while deliberately introducing variations and discrepancies. These discrepancies include adjustments to logo placement, font style, text alignment, color tones, and layout structures elements commonly exploited in counterfeit packaging. This controlled dataset allows for a robust evaluation of ChatGPT-4o's capability to detect and explain subtle design inconsistencies, thereby providing actionable insights for visual inspection workflows.

2.5 Qualitative results

To evaluate ChatGPT-4o's effectiveness in identifying design inconsistencies in pharmaceutical packaging, we present qualitative examples across three prompt configurations: Baseline, reference-guided, and focused configurations. Fig 3 showcases both successful detections (green) and failures (red), illustrating the model's reasoning and performance.

Successful cases demonstrate ChatGPT-4o ability to identify discrepancies such as text misalignment, font inconsistencies, and logo inaccuracies. The reference-guided configuration improves accuracy by leveraging reference images to highlight design differences, while the focused configuration further enhances clarity by directing the model's attention to the visual overlays between the query and reference images. These results underscore the value of structured multimodal inputs in guiding the model's attention to critical areas.

Failures, particularly in the query-only configuration, highlight the limitations of single-image inputs, which often led to generic or less actionable responses. Moreover, in some cases, over-reliance on pixel-level comparisons in the focused configuration occasionally caused false positives, underscoring the need for balanced contextual and visual inputs.

2.6 Quantitative results

Figure 4 presents the accuracy of ChatGPT-4o across three image prompt configurations demonstrating the importance of contextual inputs in enhancing model performance. The reference-guided configuration (2-images) achieved the highest accuracy of 74.6%, showcasing the effectiveness of reference images in comparative analysis and identifying design discrepancies. The focused configuration (3-images), incorporating a difference image, showed slightly lower accuracy at 68.7%, likely due to overemphasis on minor, inconsequential variations in the overlay. The baseline configuration (1-image), using only a query image, had the lowest accuracy at 54.1%, reflecting the limitations of single-image inputs and a higher rejection rate due to insufficient contextual information.

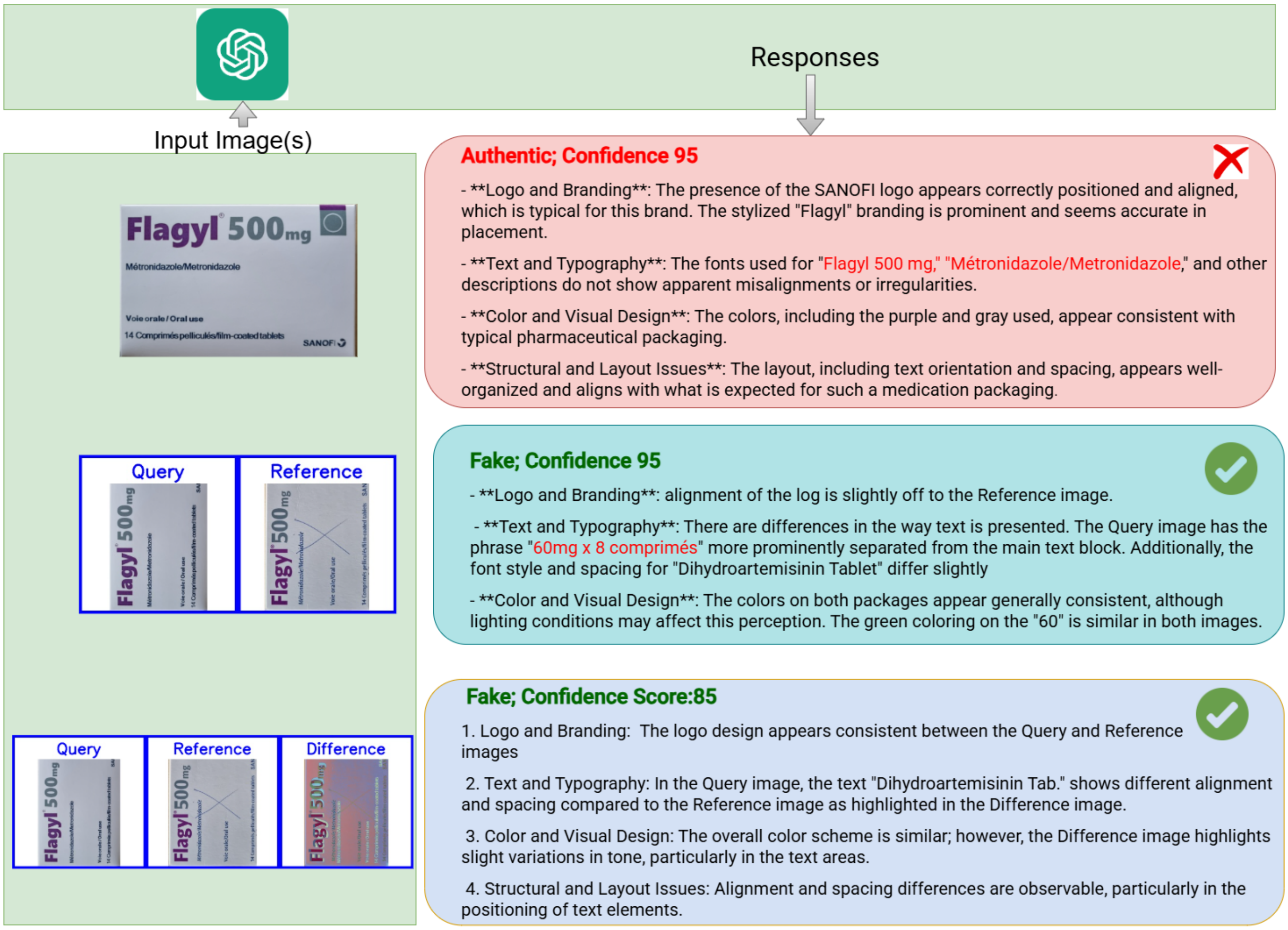
In addition to accuracy, Table 1 shows the rejection rate—instances where ChatGPT-4o declined to provide a definitive response. The baseline configuration had the highest rejection rate, underscoring the challenge of making decisions with minimal visual context.

These findings confirm the significance of structured multimodal prompts in improving ChatGPT-4o's performance in visual inspection tasks. They also emphasize the need for refined techniques, particularly in leveraging difference images, to optimize both accuracy and confidence in intricate use cases.

2.7 User evaluation results

A user study with 25 participants was conducted to assess ChatGPT-4o's performance in inspecting pharmaceutical packaging designs. Participants evaluated outputs generated under three input configurations using a 5-point Likert scale for clarity, ease of understanding, reliability, detail, and overall quality.

The focused configuration received the highest ratings (clarity: 4.3; ease of understanding: 4.2; reliability: 4.3; detail: 4.2; overall quality: 4.3), with participants noting its effectiveness in highlighting critical design



Authentic; Confidence 95 ✗

- **Logo and Branding**: The presence of the SANOFI logo appears correctly positioned and aligned, which is typical for this brand. The stylized "Flagyl" branding is prominent and seems accurate in placement.

- **Text and Typography**: The fonts used for "Flagyl 500 mg," "Métronidazole/Metronidazole," and other descriptions do not show apparent misalignments or irregularities.

- **Color and Visual Design**: The colors, including the purple and gray used, appear consistent with typical pharmaceutical packaging.

- **Structural and Layout Issues**: The layout, including text orientation and spacing, appears well-organized and aligns with what is expected for such a medication packaging.

Fake; Confidence 95 ✓

- **Logo and Branding**: alignment of the log is slightly off to the Reference image.

- **Text and Typography**: There are differences in the way text is presented. The Query image has the phrase "60mg x 8 comprimés" more prominently separated from the main text block. Additionally, the font style and spacing for "Dihydroartemisinin Tablet" differ slightly

- **Color and Visual Design**: The colors on both packages appear generally consistent, although lighting conditions may affect this perception. The green coloring on the "60" is similar in both images.

Fake; Confidence Score:85 ✓

1. Logo and Branding: The logo design appears consistent between the Query and Reference images

2. Text and Typography: In the Query image, the text "Dihydroartemisinin Tab." shows different alignment and spacing compared to the Reference image as highlighted in the Difference image.

3. Color and Visual Design: The overall color scheme is similar; however, the Difference image highlights slight variations in tone, particularly in the text areas.

4. Structural and Layout Issues: Alignment and spacing differences are observable, particularly in the positioning of text elements.

Figure 3. presents qualitative examples of ChatGPT-4o for the visual inspection of medicine packaging to detect counterfeit products. Instances of successful detections are indicated by (✓), while failed cases are marked in mark (✗). The figure underscores the potential of AI for transparent and reliable real-world applications, enhancing user comprehension during visual inspection. For optimal interpretation, the figure should be viewed in color, with zoomed-in areas offering additional detail.

Table 1. Rejection rates for each image prompt configuration

Configuration type	**Number of input images**	**Rejection rate (%)**
Query-only	1	18.2
Reference-guided	2	0
Focused (Query + Reference + Difference)	3	0

discrepancies and providing actionable insights. In contrast, the query-only configuration scored lowest (clarity: 3.4; reliability: 3.5) due to insufficient contextual information, while the reference-guided configuration performed moderately (clarity: 4.1; reliability: 4.2; overall quality: 4.2).

Although the focused configuration was praised for its clarity and detail, some participants observed that it sometimes overemphasized pixel-level differences that did not always correspond to meaningful design flaws. In contrast, the query-only configuration was seen as overly generic and less informative.

These results underscore the importance of integrating prompt engineering and multimodal inputs to enhance ChatGPT-4o's effectiveness in visual inspection tasks.

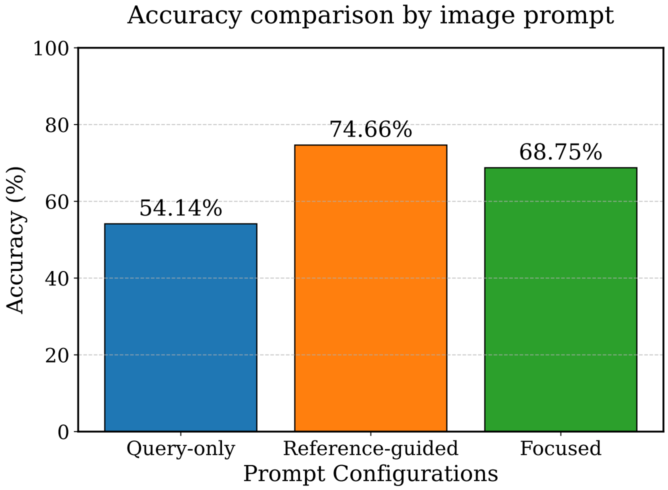


Figure 4. Performance of ChatGPT-4o for the three prompts based on input image configurations in terms of zero-shot binary classification accuracy. Notably, the reference-guided prompt achieved the highest accuracy, outperforming both the query-only and focused prompts.

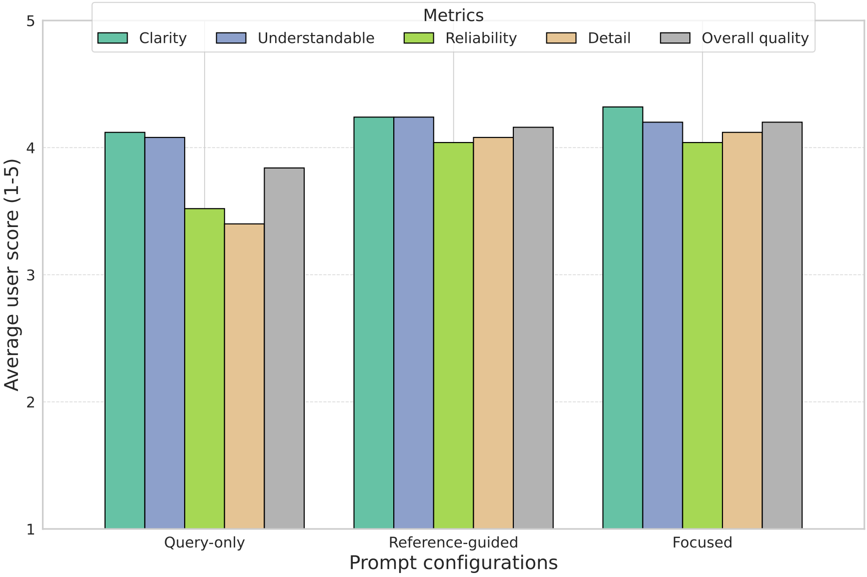


Figure 5. Overall ratings across prompt configurations (query-only, reference-guided, and focused) for acceptability and usefulness metrics. The focused configuration consistently outperformed the other configurations, demonstrating the benefits of integrating reference and difference images in enhancing user understanding.

3. CONCLUSION

This study investigated the use of multimodal LLMs for detecting counterfeit pharmaceutical packaging through visual inspection. By integrating structured text prompts with three image configurations (1-image, 2-images, and 3-images), we demonstrated that ChatGPT can effectively identify and explain design discrepancies with promising accuracy and explainability. Quantitative results confirmed that multimodal inputs enhance performance, while qualitative evaluations highlighted the importance of clarity, reliability, and contextual analysis—especially with the focused configuration (query, reference, and difference images) yielding the best outcomes.

Future work will address current limitations, such as sensitivity to minor variations, and will focus on developing user-friendly interfaces and scalable deployment strategies for real-world counterfeit detection.

REFERENCES

[1] Pathak, R., Gaur, V., Sankrityayan, H., and Gogtay, J., "Tackling counterfeit drugs: the challenges and possibilities," *Pharmaceutical Medicine* **37**(4), 281–290 (2023).

[2] Dégardin, K., Guillemain, A., Klespe, P., Hindelang, F., Zurbach, R., and Roggo, Y., "Packaging analysis of counterfeit medicines," *Forensic science international* **291**, 144–157 (2018).

[3] Bakker, I. M., Ohana, D., Venhuis, B. J., et al., "Current challenges in the detection and analysis of falsified medicines," *Journal of Pharmaceutical and Biomedical Analysis* **197**, 113948 (2021).

[4] Lächele, M., Gabel, J., Sunny-Abarikwu, N., Ohazulike, R. E., Ngene, J., Chioke, J. F., and Heide, L., "Screening for substandard and falsified medicines in nigeria using visual inspection and gphf-minilab analysis: lessons learnt for future training of health workers and pharmacy personnel," *Journal of Pharmaceutical Policy and Practice* **17**(1), 2432471 (2024).

[5] Islam, I. and Islam, M. N., "Digital intervention to reduce counterfeit and falsified medicines: A systematic review and future research agenda," *Journal of King Saud University-Computer and Information Sciences* **34**(9), 6699–6718 (2022).

[6] Zhao, C., Song, Y., Chen, J., RONG, K., Feng, H., Zhang, G., Ji, S., Wang, J., Ding, E., and Sun, Y., "Octopus: A multi-modal LLM with parallel recognition and sequential understanding," in [*The Thirty-eighth Annual Conference on Neural Information Processing Systems*], (2024).

[7] OpenAI, "https://openai.com/index/hello-gpt-4o/," (2023).

[8] Jia, S., Lyu, R., Zhao, K., Chen, Y., Yan, Z., Ju, Y., Hu, C., Li, X., Wu, B., and Lyu, S., "Can chatgpt detect deepfakes? a study of using multimodal large language models for media forensics," in [*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*], 4324–4333 (2024).

[9] Li, Y., Liu, X., Wang, X., Wang, S., and Lin, W., "Fakebench: Uncover the achilles' heels of fake images with large multimodal models," *arXiv preprint arXiv:2404.13306* (2024).

[10] Al-Janabi, O. M., Alyasiri, O. M., and Jebur, E. A., "Gpt-4 versus bard and bing: Llms for fake image detection," in [*2023 3rd International Conference on Intelligent Cybernetics Technology & Applications (ICICyTA)*], 249–254, IEEE (2023).

[11] DeAndres-Tame, I., Tolosana, R., Vera-Rodriguez, R., Morales, A., Fierrez, J., and Ortega-Garcia, J., "How good is chatgpt at face biometrics? a first look into recognition, soft biometrics, and explainability," *IEEE Access* (2024).

[12] Chauhan, M., Satbhai, A., Hashemi, M. A., Ali, M. B., Ramamurthy, B., Gao, M., Lyu, S., and Srihari, S., "Vision-language model based handwriting verification," in [*IET Conference Proceedings CP887*], **2024**(10), 343–346, IET (2024).

[13] Zakaria, Y., Ishiyama, R., Ishidera, E., Matsui, T., and Yasumoto, K., "Fast retrieval of pharmaceutical packaging images using keypoint matching with angle and scale voting for outlier rejection," in [*2024 IEEE International Conference on Visual Communications and Image Processing (VCIP)*], 1–5, IEEE (2024).

[14] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al., "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems* **35**, 24824–24837 (2022).

[15] Lindenberger, P., Sarlin, P.-E., and Pollefeys, M., "LightGlue: Local Feature Matching at Light Speed," in [*ICCV*], (2023).

[16] Abdelmaksoud, E., Gadallah, A., and Asad, A., "Mobile-captured pharmaceutical medication packages," (2022).