Checkbox State Classification using Vision-Language Models

AI Engineer Intern (R&D) – Assignment Submission

Submitted by:
Priyanshu Maheshwari
MCA, Manav Rachna International Institute of Research and Studies

Date: 21/01/2026

1. Introduction

Document understanding is a core component of many real-world applications such as form processing, surveys, KYC systems, banking applications, and OCR-based automation pipelines. One common yet challenging task in such systems is determining the state of checkboxes.

While some checkboxes are clearly marked as checked or unchecked, real-world documents often contain ambiguous cases caused by partial ticks, faded ink, scanning noise, or overlapping marks. Traditional image classification approaches may struggle to handle these cases effectively.

This assignment focuses on designing a vision-language based solution to classify checkbox images into checked, unchecked, and ambiguous categories.

2. Problem Statement

Given an image containing a single checkbox, the task is to automatically classify its state into one of the following categories:

- Checked
- Unchecked
- Ambiguous

The solution must be robust to noise, blur, and partial markings commonly found in scanned documents.

3. Dataset Preparation

3.1 Motivation

Publicly available datasets containing checkbox-level annotations are limited. Additionally, ambiguous checkbox cases are rarely labeled explicitly. To overcome this limitation, a synthetic dataset was created programmatically.

3.2 Dataset Description

- Image resolution: 224 × 224
- Classes: Checked, Unchecked, Ambiguous
- Augmentations: Rotation, Gaussian blur
- Environment: Python (PIL)

Synthetic data allowed precise control over ambiguity and ensured reproducibility.

4. Model Selection

A Vision-Language Model (VLM) was selected to jointly reason over visual input and textual instructions.

Model Used:

- **Qwen2-VL-2B-Instruct**

Reason for Selection:

- Open-source and reproducible
- Multimodal image–text support
- Suitable for local Jupyter Notebook execution

5. Methodology

The system follows an instruction-based multimodal inference pipeline:

1. Input image is provided to the model
2. A natural language instruction guides the classification task
3. Image features and text tokens are aligned using the model's chat template
4. A one-word output is generated

Correct alignment between image and text is critical for accurate inference.

6. Evaluation

Metrics Used:

- Precision
- Recall
- F1-score
- Confusion Matrix

Observations:

- Clear cases were classified accurately
- Ambiguous cases showed expected difficulty
- Model behavior aligned with real-world document challenges

7. Challenges and Learnings

- Gated model access restrictions
- CPU-only hardware constraints
- Image-token mismatch during early inference

All challenges were resolved through proper configuration and understanding of multimodal model requirements.

 Importance of image-text alignment in VLMs

 Effectiveness of synthetic data

 Handling ambiguity is essential in document AI

 Practical constraints matter in real deployments

8. Conclusion

This project demonstrates a practical vision-language approach to checkbox state classification. The system effectively handles both clear and ambiguous cases and reflects a solid understanding of multimodal AI systems.

9. Future Work

- Fine-tuning with GPU resources
- Testing on real scanned documents
- OCR and form-level integration