Technical Write-Up

Project: Image Text Extraction & Quality Reasoning Pipeline

1. Overview

This project implements an end-to-end image analysis pipeline that combines **OCR**, **classical computer vision**, and **LLM-based reasoning** to produce a structured, explainable assessment of image quality.

The system is designed to separate **perception** and **reasoning**: deterministic visual features are extracted using classical methods, while a lightweight LLM reasons over these structured features to generate a clean JSON output.

2. System Architecture

```
Image
 → Preprocessing
 → OCR Text Extraction
 → Classical CV Feature Extraction
 → LLM-Based Reasoning (Gemini Flash)
 → Structured JSON Output
```

A key design principle is that **the LLM never sees the image directly**. It operates only on explicit, structured features, which improves explainability, controllability, and robustness.

3. Feature Extraction

OCR: Text is extracted and treated as a supporting signal rather than ground truth, as OCR accuracy varies with lighting, font style, and background complexity.

Classical Computer Vision Features:

- **Blur detection** using Laplacian variance to estimate sharpness
- **Brightness analysis** to detect under/over-exposure
- **Edge density** as a proxy for visual detail
- **Text presence metadata**

These features are intentionally simple, fast, and interpretable.

4. LLM-Based Reasoning

The system uses **Gemini 2.5 Flash**, selected for its availability on the free tier, low latency, and suitability for structured reasoning tasks.

The LLM is prompted with:

- A strict JSON schema

- Explicit instructions to return only valid JSON
- Structured features serialized as input

Markdown-wrapped outputs are sanitized before parsing to ensure reliable downstream processing.

5. Output

The final output is a machine-readable and human-interpretable JSON object containing:

- An image quality score
- Detected issues
- Positive signals
- A final verdict
- Confidence score
- Reasoning summary

This format is suitable for both analysis and integration into downstream systems.

6. Design Rationale

Rather than using an end-to-end vision-language model, this approach leverages **classical CV for deterministic signals** and an **LLM for higher-level reasoning and explanation**.

During development, real-world constraints such as **API quota limits, model deprecations, and strict free-tier usage** influenced architectural decisions. As a result, the system prioritizes robustness and reproducibility over reliance on large or expensive models.

7. Limitations & Future Work

Limitations:

- OCR degrades on noisy images and stylized fonts
- Heuristic thresholds may not generalize across domains
- Gemini Flash has strict rate limits

Future Improvements:

- Integrate a multimodal vision-language model
- Improve OCR preprocessing
- Add confidence calibration and batching
- Package the pipeline as a service for production use

8. Conclusion

This project demonstrates a practical and explainable approach to image quality assessment by combining classical computer vision with LLM-based reasoning. The focus is on **engineering clarity, realistic constraints, and structured outputs**, rather than relying solely on large opaque models.