Project Instructions: Multi-Modal Evaluation (Multilingual)

1. Project Overview

Field	Value
Project Name	Gemini-LegalEval-2024
Task Type	Multi-Modal Evaluation (Multilingual)
Client	Google
Platform	Airtable
Domains	Medicine
Start Date	2024-10-22

Taskers evaluate AI responses to prompts that involve both text and images. Tasks may include image description, visual question answering, chart interpretation, or image-based reasoning. Taskers assess whether the model correctly understands and responds to the visual content.

2. Task Types

Type	Description	Evaluation Focus
Image Description	Model describes what's in an image	Evaluate accuracy, completeness, and level of detail
Visual QA	User asks a question about an image, model answers	Evaluate whether the answer correctly addresses what's visible
Chart/Graph Reading	Model interprets data from charts or graphs	Verify numerical accuracy and trend interpretation
OCR Verification	Model reads text from an image	Check character-level accuracy and formatting
Spatial Reasoning	Model reasons about object positions and relationships	Verify spatial claims (left/right, above/below, size comparisons)

3. Scoring Rubric

Dimension	Scale	Criteria
Visual Accuracy	1–5	Does the model correctly identify objects, people, text, and scenes?
Completeness	1–5	Does the response address all relevant visual elements?
Hallucination	Yes/No	Does the model describe things that are NOT in the image?

Text Accuracy	1–5	If the image contains text, does the model read it correctly?
Reasoning	1–5	If the task requires inference, is the reasoning sound?

4. Hallucination Guidelines

Hallucination is the most critical failure mode in multi-modal tasks. A hallucination occurs when the model describes something not present in the image. Examples:

- Describing a person wearing a hat when no hat is visible
- Stating a chart shows an upward trend when it shows a downward trend
- Reading text as 'January' when the image says 'June'
- Claiming there are 5 objects when there are only 3

If ANY hallucination is detected, the maximum overall score is 2/5 regardless of other dimensions.

5. Requirements

- Domain expertise: Medicine
- Platform: Airtable
- Must have a high-resolution display for image evaluation
- Quality threshold: ≥ 75% agreement rate
- For escalations or questions, contact your assigned Project Lead