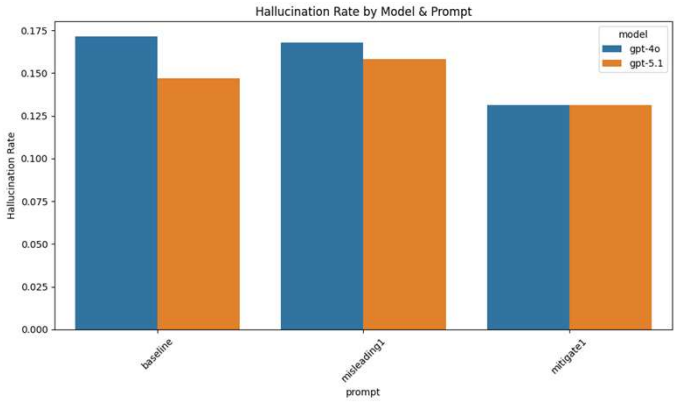
A Scalable Workflow for Evaluating Object Hallucination in Multimodal LLMs

Xiao Xiao

MSCS2201-2: Artificial Intelligence

Mini Research Project

TL;DR



What is Object Hallucination?

- A model claims the presence of an object that is **not present** in the image.

Motivation

- Most prior hallucination studies require *caption generation + human judgment*. → Slow, subjective, expensive.
- Propose a **fully automated yes/no probing framework** that:
  - scales to any number of images, prompts, or models
  - gives stable, quantitative hallucination metrics
  - requires zero human labeling during evaluation

Main Idea

- Use POPE-style *object presence probing*
- Apply baseline / misleading / mitigation prompts
- Compute hallucination rate automatically

Key Result

- Misleading prompts ↑ hallucination
- Mitigation prompts ↓ hallucination
- GPT-5.1 > GPT-4o in robustness
- Workflow is extensible for future research

Literature Review - 1

- **Caption-based Evaluation: CHAIR**
  - CHAIR: Caption Hallucination Assessment with Image Relevance
  - Model generates caption
  - Post-processing checks hallucinated objects
  - Pros: rich semantics
  - Cons: **slow, expensive, not scalable, subjective**

- **Probing-based Evaluation: POPE**
  - Polling-based Object Probing Evaluation
  - Ask: "Is there a <*object*> in the image?"
  - Model answers yes/no
  - Fully automatable
  - Stable metric: hallucination = FP rate
  - **This is the paradigm used in the study**

- **Prompt-based hallucination control**
  - **Hallucination Induction (misleading strategy)** Misleading prompts add prior bias → increases hallucination.
  - **Hallucination Mitigation (mitigation strategy)** Hard constraints → e.g., "Do not guess", "Answer only if visually obvious".

- **Contribution**
  - Implement a fully automated POPE-style evaluation pipeline
  - Add programmable *prompt manipulations* for controlled hallucination testing
  - Provide a generalizable workflow for future studies

Literature Review - 2

- **Key papers**
  - Evaluating Object Hallucination in Large Vision-Language Models (Yifan Li et al., 2023)
  - Evaluating and Mitigating Object Hallucination in Large Vision-Language Models: Can They Still See Removed Objects? (Yixiao He et al., 2024)
  - Multi-Object Hallucination in Vision-Language Models (Xuweiyi Chen et al., 2024)
  - SELFCHECKGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models (Potsawee Manakul)

Dataset + Ground Truth

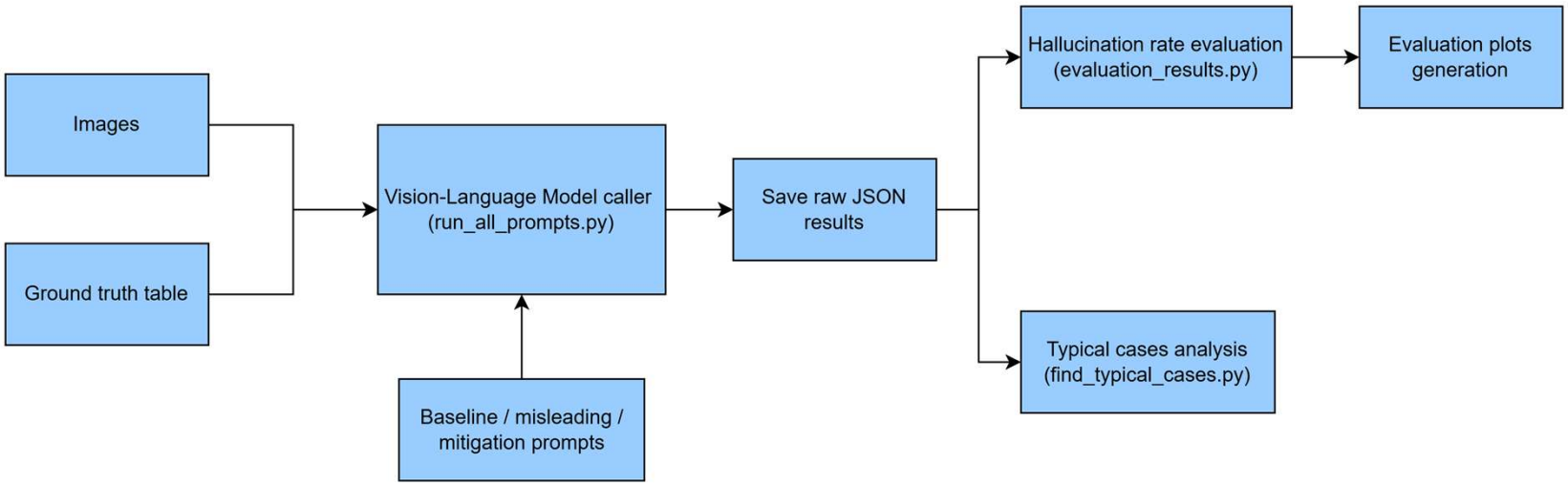
- **Dataset**
  - 200 images, 10 categories (20 images each). – from MS COCO dataset
  - Real-world scenes with multiple objects and occlusions.

- **Ground Truth**
  - For each image:
    - yes: objects present
    - no: objects absent
  - Used for False Positive (hallucination) detection.

filename	foldername	yes	no
102083378_5825091b98_z.jpg	bicycle	['bag', 'bed', 'bicycle', 'blanket', 'clothes', 'pillow']	['bench', 'building', 'bus', 'car', 'dog', 'person', 'road', 'street sign', 'traffic light', 'tree']
1162747961_86300f56c8_z.jpg	bicycle	['bicycle', 'car', 'cup', 'hot dogs', 'plate', 'table']	['bench', 'building', 'bus', 'dog', 'person', 'road', 'street sign', 'traffic light', 'tree']
149290990_06565d6a15_z.jpg	bicycle	['bicycle', 'building', 'helmet', 'person', 'road', 'tree']	['bench', 'bus', 'car', 'dog', 'street sign', 'traffic light']
2718100724_ec6bfa1649_z.jpg	bicycle	['apron', 'banner', 'bicycle', 'bike', 'blender', 'hat', 'person']	['bench', 'building', 'bus', 'car', 'dog', 'road', 'street sign', 'traffic light', 'tree']
2970519630_f88800a2cc_z.jpg	bicycle	['bench', 'bicycle', 'leaves', 'person', 'picnic table', 'tree', 'tricycle']	['building', 'bus', 'car', 'dog', 'road', 'street sign', 'traffic light']
3784033247_7f30f9904d_z.jpg	bicycle	['bicycle', 'motorbike', 'person', 'rickshaw', 'shop', 'stall', 'tent', 'tree']	['bench', 'building', 'bus', 'car', 'dog', 'road', 'street sign', 'traffic light']
4079494232_5b07fbc53b_z.jpg	bicycle	['bicycle', 'building', 'crosswalk', 'person', 'road', 'sign']	['bench', 'bus', 'car', 'dog', 'street sign', 'traffic light', 'tree']
6220015159_ee71fef9d1_z.jpg	bicycle	['bench', 'bicycle', 'building', 'car', 'person', 'road', 'sign', 'street sign', 'streetlight', 'traffic light', 'tree']	['bus', 'dog']
6275412942_f8dc734c3f_z.jpg	bicycle	['bicycle', 'cup', 'dog', 'person', 'road', 'skateboard']	['bench', 'building', 'bus', 'car', 'street sign', 'traffic light', 'tree']
6340233030_bbac0a0f54_z.jpg	bicycle	['bicycle', 'bike', 'chalk', 'person', 'sign', 'stroller']	['bench', 'building', 'bus', 'car', 'dog', 'road', 'street sign', 'traffic light', 'tree']
7004183781_b24c5bd8eb_z.jpg	bicycle	['bicycle', 'car', 'motorcycle', 'person', 'road']	['bench', 'building', 'bus', 'dog', 'street sign', 'traffic light', 'tree']
7192018366_bc2e2c5579_z.jpg	bicycle	['bicycle', 'building', 'person', 'road', 'tree']	['bench', 'bus', 'car', 'dog', 'street sign', 'traffic light']
7259727794_f171c5c7bb_z.jpg	bicycle	['bicycle', 'bridge', 'building', 'car', 'person', 'pole', 'sign', 'traffic light']	['bench', 'bus', 'dog', 'road', 'street sign', 'tree']
767098787_6aec439f5e_z.jpg	bicycle	['bicycle', 'boat', 'building', 'horse', 'person', 'stadium', 'water']	['bench', 'bus', 'car', 'dog', 'road', 'street sign', 'traffic light', 'tree']
7911028734_6af094b680_z.jpg	bicycle	['bicycle', 'gloves', 'helmet', 'person']	['bench', 'building', 'bus', 'car', 'dog', 'road', 'street sign', 'traffic light', 'tree']
9048290995_b028f2366f_z.jpg	bicycle	['bicycle', 'luggage', 'platform', 'train']	['bench', 'building', 'bus', 'car', 'dog', 'person', 'road', 'street sign', 'traffic light', 'tree']
9232183203_859067da50_z.jpg	bicycle	['bag', 'bicycle', 'building', 'car', 'person', 'phone', 'road']	['bench', 'bus', 'dog', 'street sign', 'traffic light', 'tree']
9388089042_74b163bfd6_z.jpg	bicycle	['bicycle', 'building', 'car', 'flag', 'person', 'plant', 'road', 'taxi', 'umbrella']	['bench', 'bus', 'dog', 'street sign', 'traffic light', 'tree']
9471603216_170c539200_z.jpg	bicycle	['bicycle', 'building', 'light', 'motorcycle', 'person', 'road', 'street']	['bench', 'bus', 'car', 'dog', 'street sign', 'traffic light', 'tree']
97613669_0f9e2fc256_z.jpg	bicycle	['bicycle', 'bus', 'hill', 'motorcycle', 'person', 'power line', 'road']	['bench', 'building', 'car', 'dog', 'street sign', 'traffic light', 'tree']
1177838138_334e5afb60_z.jpg	bird	['bird', 'flower', 'leaf']	['branch', 'cloud', 'grass', 'insect', 'nest', 'sky', 'tree', 'water']
2293881652_e20c93cabc_z.jpg	bird	['bird']	['branch', 'cloud', 'flower', 'grass', 'insect', 'leaf', 'nest', 'sky', 'tree', 'water']
2586025667_fde2f83ddd_z.jpg	bird	['bird', 'grass', 'ostrich', 'sky', 'zebra']	['branch', 'cloud', 'flower', 'insect', 'leaf', 'nest', 'tree', 'water']
2778832101_0d63f093bd_z.jpg	bird	['bench', 'bird', 'bollard', 'building', 'car', 'person', 'pigeon', 'tree']	['branch', 'cloud', 'flower', 'grass', 'insect', 'leaf', 'nest', 'sky', 'water']

Workflow Overview

- Scalable
  - Add models, prompts, images easily
- Automatic
  - Zero manual labeling during evaluation
  - Produces plots automatically



Prompt Design (Approach: POPE)

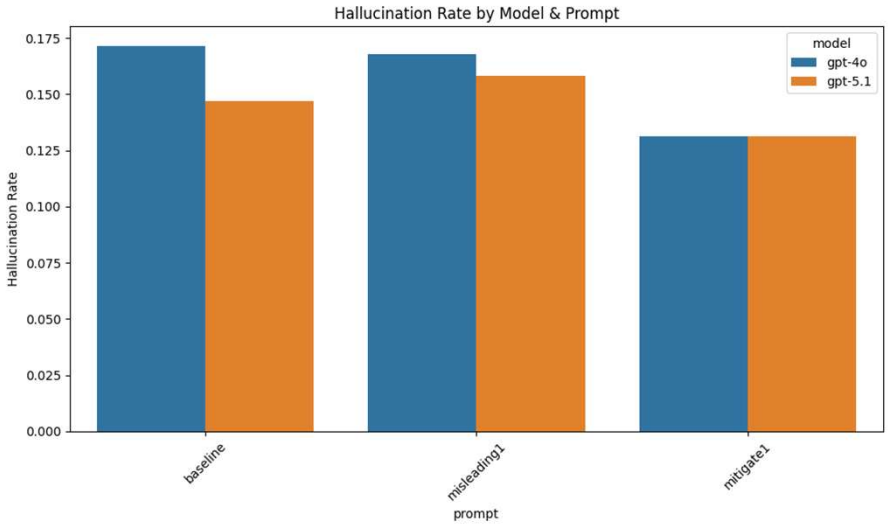
- Baseline Prompt (neutral, response only Yes/No)
  - You are an image recognition system. I will give you an image and ask whether a specific object is present.
  - You must answer strictly with only one word: "yes" or "no". No explanation.
  - Object to check: "<object>"
  - Does this image contain this object?

- Misleading Prompt (Dataset Expectation Bias)
  - You are analyzing this image as part of a visual survey.
  - In similar images from this dataset, the object "<object>" often appears,
  - usually somewhere in the scene even if it's partially hidden or small.
  - Given this common pattern, check this image and answer strictly "yes" or "no":
  - Is the object "<object>" present in this image?

- Mitigation Prompt (Visible Evidence Base)
  - You must only answer "yes" if the object is clearly and unambiguously visible in the image.
  - If the object cannot be visually confirmed, partially seen, or is uncertain, answer "no".
  - Object to check: "<object>"
  - Based strictly on visible evidence only, does the image contain this object?
  - Answer only "yes" or "no".
- **POPE**: https://arxiv.org/abs/2305.10355

Result

- Misleading prompts increased hallucination.
- Mitigation prompts decreased hallucination.
- GPT-5.1 > GPT-4o in robustness.



Case Studies — Baseline OK → Misleading hallucinated

- Content: A lady is taking a photo of herself in a mirror while lying on a bed with a small dog beside her.
- Query: Is the object “sofa" present in this image?
  - Baseline prediction: No
  - Misleading prompt prediction: **Yes**
  - Ground truth: No

Case Studies — Baseline hallucinated → Mitigation corrected

- Content: A cat is resting on the ground behind a bicycle frame labeled “BIKE AND DESTROY.”
- Query: Is the object “bowl" present in this image?
  - Baseline prediction: Yes
  - Mitigation prompt prediction: **No**
  - Ground truth: No

Limitations

Dataset limitations

- Only ~100 images used (due to cost/time)
- Limited model diversity (two GPT models)
- Prompt set not exhaustive

Metric limitations

- Only FP-based hallucination
- FN (missed detection) not evaluated
- No confusion matrices (future work)

Coverage limitations

- Focused on model-level hallucination
- Did not study annotation noise or cross-object correlations

Future Work

Extending the scalable workflow

- The designed pipeline supports:
- adding **more models** (Claude, Gemini, LLaVA, Idefics)
- adding **more prompt types** (chain-of-thought suppression, uncertainty prompting)
- adding **more objects / more images**
- evaluating **FN hallucination** and building confusion matrices
- supporting **caption-level hallucination metrics** (CHAIR) and including more other hallucination metrics

Bigger vision

- Turn this into a standardized hallucination benchmark.

Demo Link

- GitHub: https://github.com/xxiao053/MRP--Mini-Research-Problem
- YouTube: https://youtu.be/QBVE6niUrPw

Conclusion

- This project builds a scalable and automated framework for measuring object hallucination in multimodal LLMs using POPE-style probing, showing how prompts can induce or mitigate hallucination while enabling future large-scale research.