

Multimodal Models Experimentation

Rejoice Hu, 2024

OBJECTIVE

Given an input image, describe a screenshot of a webpage in as much detail as possible, including the text, position, layout, color, etc. of ALL its different webpage components such as headers, paragraphs, images, menu items, links, etc.

The goal is to yield highly accurate results in the most cost-efficient manner - in other words, ideally with **open source** models, while using as **minimal resources** as possible (ideally CPU).

In this PoC, screenshots of the Morgan Stanley homepage (morganstanley.com) were used.

IN SUMMARY (TOP MODELS RANKED)

- **1) BEST: GPT-4o / GPT-4o mini (ChatGPT)**
 - No hallucinations at all and describes the webpage just like a human would describe the page, able to read all the smaller text and very organized
 - Sometimes omits some text, but no longer an issue when add “Do not miss any text” to the prompt
 - Recognizes interactivity!—VERY GOOD FOR multi-image use cases / dynamic websites where users can click on links/dropdown menus!
- **2) LLaVA-Mistral (Huggingface)** — <https://huggingface.co/llava-hf/llava-v1.6-mistral-7b-hf>
 - Actually able to read/identify the smaller text!!!
- **3) LLaVA-v1.6 (Ollama)** — <https://ollama.com/library/llava>
 - Detailed description, but riddled with hallucinations—further prompt engineering improved results but only to a certain point. Root cause might be possibly be due to resolution curse (can’t recognize smaller text + can only read giant headers and images)
 - LLaVA-v1.6 in general supports multi-image inputs, which would be useful for few-shot prompting
- worst = BLIP-2, mapler/LLaVA-Mistral through Ollama

NEXT STEPS/RECOMMENDATIONS

- GPT-4o is still the best—WHY? Possibly because it is a **NATIVELY** multi-modal, and also the **LARGEST**/most complex model I tested
 - To find an alternative to GPT that performs better than LLaVA and the other models tested (which are **NOT** natively multi-modal and only fine-tuned on text-only LLMs), experiment with other **Native Multi-Modal Models**:
 - Gemini AI (Google Deepmind) (<https://www.nimblechapps.com/blog/google-deepminds-gemini-ai-a-truly-universal-ai-model>)

- Chameleon (Meta) (<https://medium.com/@TheSage/a-powerful-native-multimodal-llm-say-hello-to-metas-chameleon-7255e949162f>)
- Anole (**open source**) (<https://github.com/GAIR-NLP/anole#:~:text=Anole%20is%20the%20first%20open,of%20alternating%20text%20and%20images>)
- ALSO experiment with GPT-4o using actual API instead of ChatGPT
- **Try a larger model (13b? 34b?)**
- Resolution curse: VLMs are limited by the resolution of the vision encoder, which usually is not very large—as a result, the model might not be able to read/identify all the text that appear on the webpage screenshot
 - Possible solutions: Visual search, Visual cropping, MC-LLaVA (<https://huggingface.co/blog/visheratin/vlm-resolution-curse>)
 - <https://arxiv.org/abs/2312.14135>
 - LOOK INTO: how to increase resolution (LLaVA currently supports three aspect ratios, up to 672x672, 336x1344, 1344x336 resolution) — investigate resulting image quality when fed into the model to generate a response and how to possibly increase/address resolution problem through image processing or other techniques listed above
- Further Prompt Engineering
- Few-shot prompting with multi-image input
 - Resolve errors for multi-image input and prompt structure with LLaVA
 - How to incorporate Few-shot into Ollama syntax + structure—LLaVA thankfully accepts multi-image inputs, but figure out how to specify which one(s) are the examples (in other words, can you feed multi-image to a singular prompt (for few-shot) OR it simply just applies different prompts to each image fed)
 - Concerns: LLaVA was not pre-trained with several images interleaved in one prompt, so it still might not perform well
 - Try **LLaVA-NeXT-Interleave** since it supports multi-image and multi-prompt generation
- Incorporate Langchain? embeddings?

OTHER FINDINGS/ANALYSIS

- Rankings of several multimodal LLMs on various tasks/metrics - <https://github.com/BradyFU/Awesome-Multimodal-Large-Language-Models/tree/Evaluation?tab=readme-ov-file>
 - LLaVA vs. BLIP-2: LLaVA (with all dif parameters) is always ranked higher in Perception compared to BLIP-2 (existence, count, position, color, poster, celebrity, scene, landmark, artwork, and OCR)
- Reducing Hallucinations
 - What I tried:
 - reduce temperature — effective to a certain point (extremely high temperatures hallucinate more than lower temps, but effect of reducing temperature only does so much)
 - Prompt Engineering:

- increase prompt specify and detail — effective
- “step by step” — varies
- “Only describe what you can see” — varies, leads to response sometimes admitting that the image resolution is too blurry to be able to read
- break down prompt into bite-sized tasks and ask about features individually/separately + combine descriptions later (for example - instead of asking everything about text, images, etc. all at once, ask one prompt that focuses on the text, then another about the images) — not very effective
- BETTER models — LLaVA-1.5 with Vicuna base —> LLaVA-1.6 with Mistral base in both Ollama and Huggingface, LLaVA-Phi3, BakLLaVA, etc.
 - LLaVA-1.6 with Mistral in Huggingface was the most effective model change
- TO TRY:
 - play around with: system, template, prompt—instead of squeezing it all into prompt
 - try CoT prompt engineering—using all text extracted from the prev step, describe this [next component]
 - ask multiple vision language models to reach a consensus + combine their descriptions after
- Papers/Research
 - Mitigating hallucinations for image to text LLMs / VLMS / multi-modal models = still an ongoing, active area of research
 - Most techniques = instruction tuning or training/finetuning on additional datasets, and/or require human ground truth (<https://arxiv.org/html/2405.09589v1#S3>)
 - LVLm Hallucination Revisor (LURE) algorithm ?
 - “To mitigate object hallucination in LVLms without resorting to costly training or API reliance, Zhao et al. (2024) introduced **MARINE**, a training-free and API-free solution. MARINE enhances LVLms’ visual comprehension by combining existing open-source vision models and leveraging guidance without classifiers to incorporate object grounding features, thereby enhancing the precision of generated outputs” (<https://arxiv.org/html/2405.09589v1#S3>)
 - CLIP-Guided Decoding (CGD) approach- a straightforward but effective training-free approach to reduce object hallucination at decoding time. CGD uses CLIP to guide the model’s decoding process by enhancing visual grounding of generated text with the image ? (<https://github.com/d-ailin/CLIP-Guided-Decoding>)
 - Better models?
 - LLaVA: try greater parameter size? 13b? 34b?
 - natively multi-modal models vs. fine-tuned on text-only LLM
 - Image resolution / resolution curse: LLaVA can't seem to read/recognize the smaller text (only very large text/headings + images)
 - Try FEW-SHOT: The newer LLaVA 1.6 supports multi-image inputs — may be useful esp if we have multiple screenshots for capturing website interactions
- Few-Shot Prompting with Multi-Image Inputs
 - Try LLaVA-NeXT-Interleave!?
 - <https://github.com/haotian-liu/LLaVA/issues/197>

OTHER HELPFUL LINKS

- Collection of Multimodal Large Language Models: <https://github.com/BradyFU/Awesome-Multimodal-Large-Language-Models>
 - A Survey on Multimodal Large Language Models (<https://arxiv.org/pdf/2306.13549>)
- HALLUCINATIONS for VLMs / multi-modal models
 - PAPERS
 - Unveiling Hallucination in Text, Image, Video, and Audio Foundation Models: A Comprehensive Review (<https://arxiv.org/html/2405.09589v1#S3>)
 - Visual Hallucinations of Multi-modal Large Language Models (<https://arxiv.org/pdf/2402.14683>)
 - Detecting and Preventing Hallucinations in Large Vision Language Models (<https://arxiv.org/pdf/2308.06394>)
 - Multi-modal hallucination control by visual information grounding (<https://www.amazon.science/publications/multi-modal-hallucination-control-by-visual-information-grounding>)
 - Let there be a clock on the beach: Reducing Object Hallucination in Image Captioning (https://openaccess.thecvf.com/content/WACV2022/papers/Biten_Let_There_Be_a_Clock_on_the_Beach_Reducing_Object_WACV_2022_paper.pdf)
 - Object Hallucination in Image Captioning (<https://aclanthology.org/D18-1437.pdf>)
 - Mitigating Open-Vocabulary Caption Hallucinations (<https://assafbk.github.io/mocha/>)
 - POSSIBLE SOLUTIONS
 - Woodpecker: Hallucination Correction for Multimodal Large Language Models (<https://github.com/BradyFU/Woodpecker>, <https://arxiv.org/abs/2310.16045>)
 - Seeing is Believing: Mitigating Hallucination in Large Vision-Language Models via CLIP-Guided Decoding (<https://openreview.net/pdf/84e43111d901965aeb354a001699921796e8eaf0.pdf>, <https://github.com/d-ailin/CLIP-Guided-Decoding>)
 - ALOHa: A New Measure for Hallucination in Captioning Models (<https://arxiv.org/html/2404.02904v1>)
- FEW-SHOT LEARNING WITH MULTI-IMAGE
 - Self-Distillation for Few-Shot Image Captioning (https://openaccess.thecvf.com/content/WACV2021/papers/Chen_Self-Distillation_for_Few-Shot_Image_Captioning_WACV_2021_paper.pdf)
 - LMCap: Few-shot Multilingual Image Captioning by Retrieval Augmented Language Model Prompting (<https://arxiv.org/abs/2305.19821>)
 - Re-ViLM: Retrieval-Augmented Visual Language Model for Zero and Few-Shot Image Captioning (<https://arxiv.org/abs/2302.04858>)
 - PM2 : A New Prompting Multi-modal Model Paradigm for Few-shot Medical Image Classification (<https://arxiv.org/abs/2404.08915>)
 - <https://huggingface.co/llava-hf/llava-v1.6-mistral-7b-hf/discussions/19>
 - <https://medium.com/@suvasism/multimodal-few-shot-learning-with-frozen-language-models-focus-on-understanding-implementation-of-065fb8a602cc>

- Vision LLM Resources: https://github.com/jingyi0000/VLM_survey, <https://huggingface.co/blog/vlms>
- <https://github.com/OpenGVLab/VisionLLM/tree/main/VisionLLMv2>, <https://arxiv.org/pdf/2406.08394>