THRONE: An Object-based Hallucination Benchmark for Free-Form Generations of Large Vision-Language Models

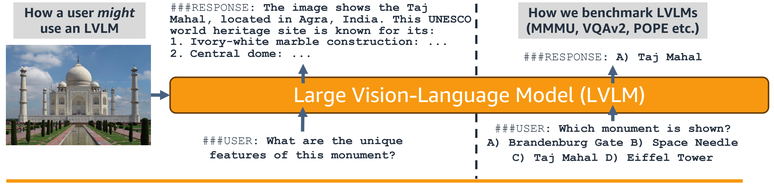
Prannay Kaul[1] Zhizhong Li[2]* Hao Yang[2] Yonatan Dukler[2] Ashwin Swaminathan[2] C. J. Taylor[2] Stefano Soatto[2]

[1]VGG, University of Oxford [2]AWS AI Labs

CVPR SEATTLE, WA JUNE 17-21, 2024

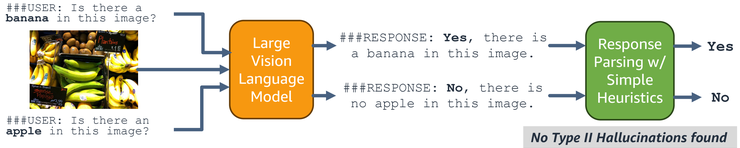
amazon | science

LVLM Usage vs Benchmarking

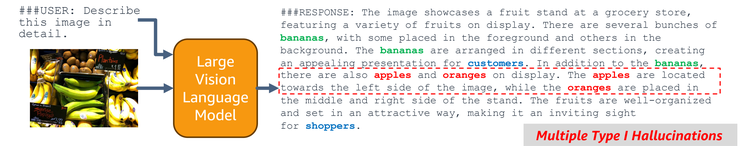


Measuring Hallucinations (POPE [1] vs THRONE)

- Current methods evaluate ***Type II Hallucinations*** in responses to yes/no questions



- THRONE evaluates ***Type I Hallucinations*** in free-form responses to *open-ended* prompt
- Type I and II performances are barely correlated: Spearman's $\rho = 0.2$



Measuring Hallucinations (CHAIR [2] vs THRONE)

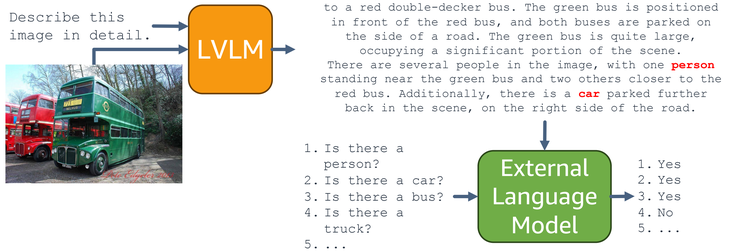
- CHAIR addresses Type I hallucinations *but uses simple text-match*
- Many abstractly referenced nouns + synonyms are incorrectly classified by CHAIR (**blue**)
- We estimate that THRONE makes ***half*** the errors of CHAIR (4% vs 9%)

The image features a green **motorcycle** parked on a street next to a building. The motorcycle is positioned on the left side of the scene, with its front end facing the viewer. The **bike** is parked on a curb, with its front wheel slightly raised. There are several **cars** parked nearby, with one car located on the right side of the scene and another one further to the left. A **person** can be seen standing near the motorcycle, possibly admiring it or preparing to ride it.

THRONE	*CHAIR*
motorcycle	motorcycle
car	car
person	person
	bicycle

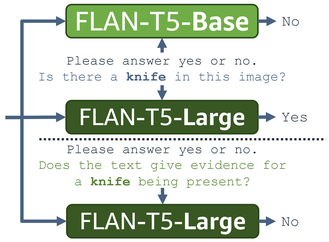
THRONE Method

1. **Extract long-form responses from a given LVLM**
   - We use a single prompt: `Describe this image in detail.`
2. **Abstractive question answering (AQA) on the LVLM response using a language model (LM)**
   - We only consider *object existence* in the responses w.r.t known classes (e.g. COCO classes)
3. **Collate LM responses and calculate informative metrics**
   - LM is prompted to give simple responses so no parsing required



Ensuring Robustness

The image features a dining table with a delicious slice of chocolate cake on a white plate. The cake is accompanied by a glass of beer, which is placed on the table as well. There are two bowls on the table, one near the cake and the other further away. A fork is also visible on the table, ready to be used for enjoying the cake. The scene creates a cozy and inviting atmosphere for a delightful dessert and beverage experience.



- No LM will perform AQA perfectly
  - Incorrect judgements will be made on an LVLM response regarding object existence
- Ensemble multiple *lightweight* LMs (FLAN-T5) – ***modeling ensembling***
- Ensemble multiple prompts *to the LMs* – ***prompt ensembling***

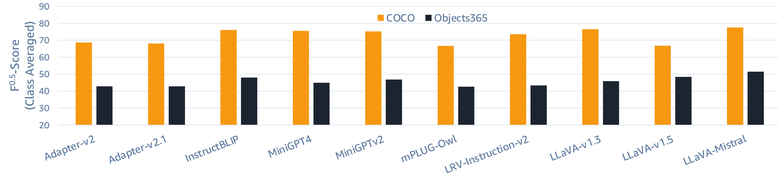
References

[1] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In Proceedings of the Conference on Empirical Methods in Natural Language, 2023

[2] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In Proceedings of the Conference on Empirical Methods in Natural Language, pages 4035–4045, 2018.

THRONE Results

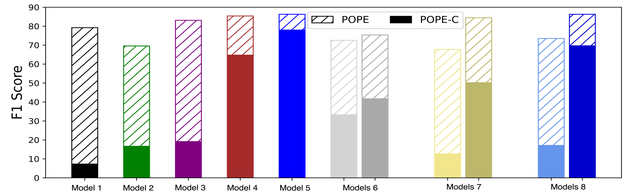
- THRONE performed for *object existence* of all classes, C, in a dataset across many images, $\mathcal{I}$
- Predicted object existence: $\hat{Y} \in \{0,1\}^{|\mathcal{I}|\times|\mathcal{C}|}$ GT object existence: $Y \in \{0,1\}^{|\mathcal{I}|\times|\mathcal{C}|}$
- This allows Overall and Classwise Precision/Recall to be calculated.
- $F^{0.5}$-Score calculated to prioritize precision over recall (Hallucinations ⊂ False Positives)



- **THRONE evaluates Type I hallucinations in free-form responses**
- **THRONE evaluation is automatic and does not use of subjective scoring by LLMs**
- **THRONE uses open-source lightweight language models (ensemble), improving comprehension, ease-of-use, and reproducibility**

Issues with gauging Type II Hallucinations

- Hallucinations ⊂ False Positives
- Need to search through many true negatives to reveal the complete picture
- Completing POPE for all classes reveals the true extent of Type II hallucinations (POPE-C)



Improved Baseline via Augmentation

- We introduce a simple data augmentation which forces LLaVA models to enumerate objects

Model	Object Enumeration Data	THRONE P_{CLS}	THRONE R_{CLS}	THRONE $F^{0.5}_{\text{CLS}}$	POPE P	POPE R	POPE F^1	POPE-C P	POPE-C R	POPE-C F^1
LLaVA-v1.5	✗	69.9	56.4	66.8	81.9	90.8	86.1	58.7	85.7	69.7
	COCO	**87.2**	76.6	**84.9**	88.6	**85.3**	**87.0**	58.9	**87.5**	70.4
	COCO + VG	86.1	**77.0**	84.1	**89.8**	83.7	86.7	**64.5**	86.1	**73.7**

*Corresponding author: lzhizhon@amazon.com