

Toward a Stable, Fair, and Comprehensive Evaluation of Object Hallucination in Large Vision-Language Models

Hongliang Wei, Xingtao Wang, Xianqi Zhang, Xiaopeng Fan, Debin Zhao



Harbin Institute of Technology
Harbin, China

Challenge In Hallucination Evaluation of LVLMs



Describe this image

Gemini Pro

The image shows a busy street in Ho Chi Minh City, Vietnam. There are many motorbikes, cars, and buses on the street. The motorbikes are mostly driven by men, and they are wearing helmets. The cars and buses are driven by both men and women. The street is lined with trees, and there are a few buildings in the background. The image is taken from the perspective of a passenger on a motorbike.

Owen-VL

The image depicts a group of people riding motorcycles and mopeds down a city street, surrounded by various vehicles such as cars, trucks, and a bus. The street is busy with a total of 13 motorcycles and mopeds, with some of them being closer to the foreground and others further in the background. In addition to the motorcyclists, there are four cars and two trucks visible on the street. One of the trucks can be seen closer to the right side of the scene, while the other vehicle, a car, is positioned more towards the center of the street. A person on foot is also present near the center of the scene, likely walking between the vehicles or waiting to cross the street.

What is described in the image

Gemini Pro

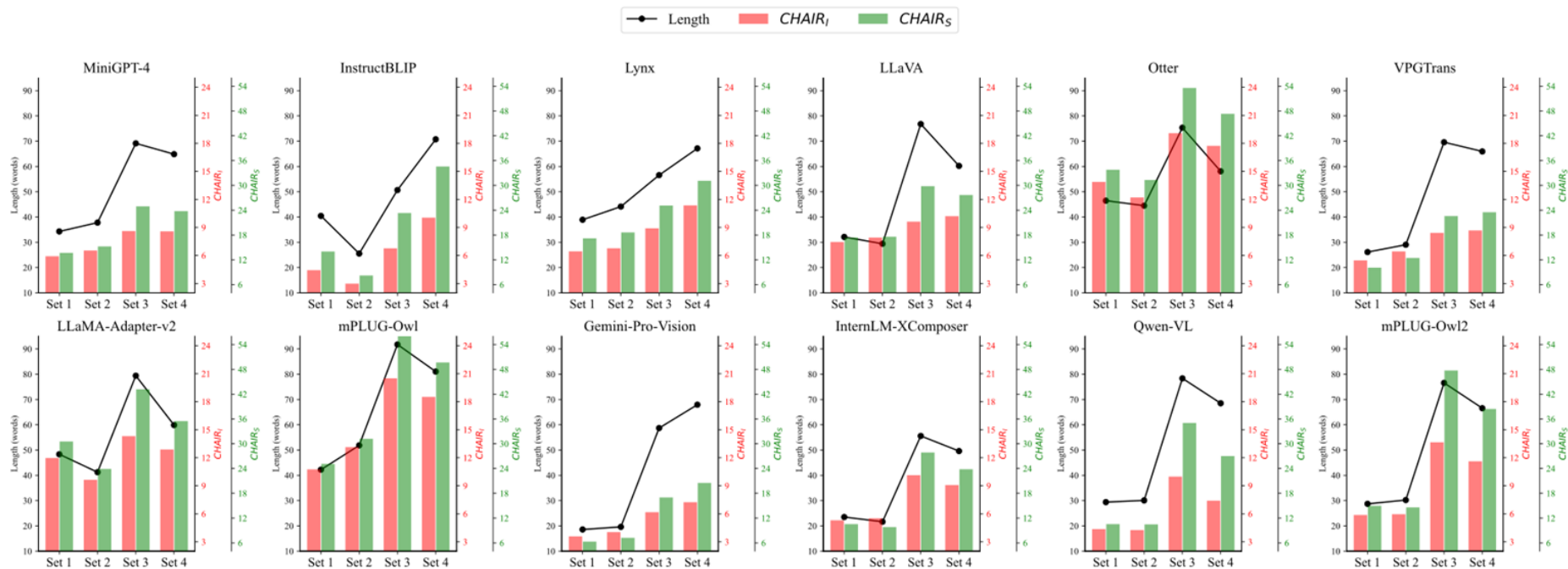
The image shows a busy street in Vietnam with many motorbikes, cars, and people crossing the road.

Owen-VL

The image shows a group of people riding motorcycles and mopeds down a city street, surrounded by various vehicles such as cars, trucks, and buses. There are multiple motorcycles and mopeds visible in the scene, with riders wearing helmets for safety. The street appears to be busy and bustling with activity.

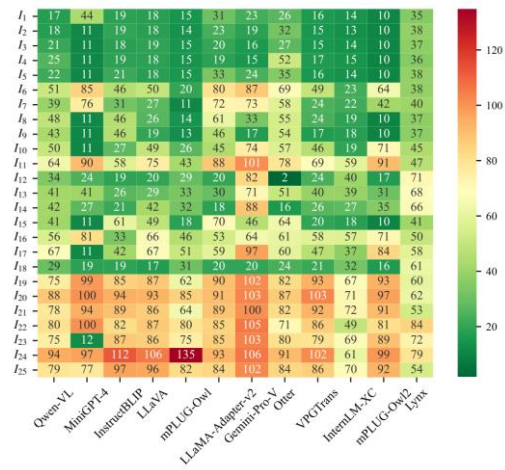
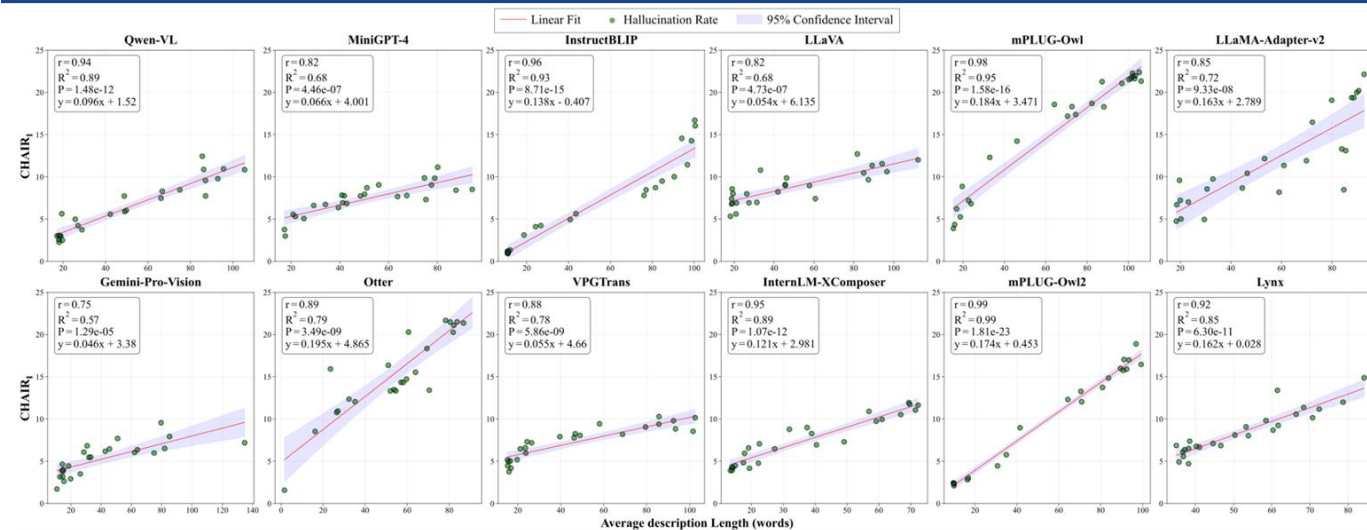
- **Definition of Object Hallucination:**
Occurs when Large Vision-Language Models (LVLMs) describe objects that are not present in the visual content.
- **Key Challenge:**
Variability in Instructions: Different instructions lead to varying hallucination degrees, making consistent evaluation difficult.
- **Current Methods:**
Average-Based Frameworks: They average hallucination rates across different instructions to mitigate inconsistencies.

Limitation of The Average-based Framework



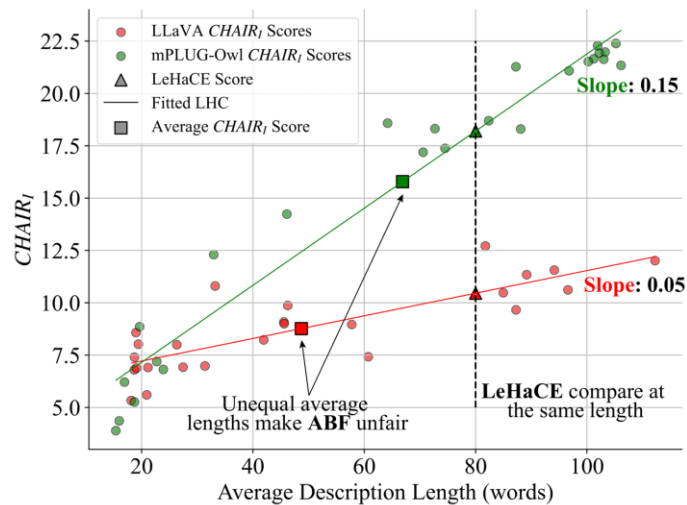
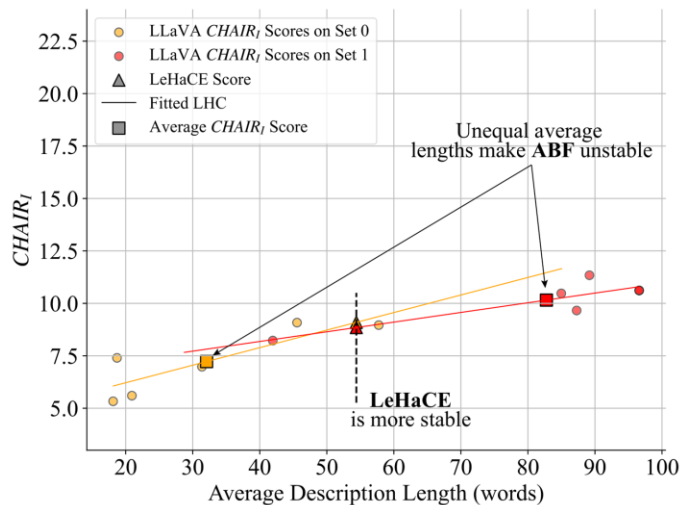
The average-based framework's evaluation is affected by description length and fails to provide consistent results when lengths differ significantly.

Key Observations



- The degree of object hallucinations is primarily influenced by the length of image descriptions, with instructions only indirectly affecting hallucinations through their effect on image description lengths. The longer the image description, the higher the object hallucination degree, and there is a clear linear relation between them.
- Different models produce outputs of significantly varying lengths under the same instruction, making it difficult for averaging to eliminate the impact of length.

Length-Hallucination Curve based Evaluation Framework (LeHaCE)



- LeHaCE performs a linear fit of a length-hallucination curve to evaluate hallucinations at any given description length to reduce inconsistencies.
- LeHaCE ensures fair evaluation by evaluating different LVLMs at the same description length.
- LeHaCE introduces the "curve slope" metric to measure how description length influences hallucination rates.

Experimental Results

Model	L _{C_l} (20)	L _{C_l} (40)	L _{C_l} (60)	L _{C_l} (80)	L _{C_lGR}	L _{C_s} (20)	L _{C_s} (40)	L _{C_s} (60)	L _{C_s} (80)	L _{C_sGR}
MSCOCO										
MiniGPT-4	5.33	6.66	7.98	9.31	0.07	9.27	15.71	22.15	28.59	0.32
InstructBLIP	2.35	5.10	7.86	10.61	0.14	5.61	16.24	26.87	37.50	0.53
Lynx	<u>3.26</u>	6.49	9.72	12.95	0.16	8.00	17.48	26.97	36.46	0.47
LLaVA	7.22	8.30	9.38	10.46	0.05	14.48	20.31	26.14	31.97	<u>0.29</u>
Otter	8.76	12.66	16.56	20.45	0.19	15.31	29.88	44.45	59.02	0.73
VPGLTrans	5.77	6.87	7.97	<u>9.08</u>	0.06	9.08	<u>15.01</u>	<u>20.94</u>	<u>26.86</u>	0.30
LLaMA-Adapter-v2	6.04	9.29	12.54	15.80	0.16	11.31	22.99	34.66	46.34	0.58
mPLUG-Owl	7.15	10.84	14.52	18.20	0.18	11.18	23.71	36.25	48.79	0.63
Gemini-Pro-Vision	4.30	<u>5.22</u>	6.15	7.07	0.05	8.00	12.61	17.22	21.83	0.23
InternLM-XComposer	5.40	7.82	10.25	12.67	0.12	9.48	19.18	28.88	38.58	0.48
Qwen-VL	3.44	5.36	<u>7.28</u>	9.20	0.10	<u>6.15</u>	15.31	24.47	33.63	0.46
mPLUG-Owl2	3.92	7.39	<u>10.86</u>	14.33	0.17	8.19	21.66	35.12	48.59	0.67
NoCaps										
MiniGPT-4	14.53	16.79	19.05	21.30	0.11	23.75	35.75	47.76	59.77	0.60
InstructBLIP	6.52	10.20	<u>13.88</u>	17.56	0.18	<u>13.33</u>	26.39	39.45	<u>52.50</u>	0.65
Lynx	13.79	17.18	20.57	23.96	0.17	36.07	46.11	56.16	66.21	<u>0.50</u>
LLaVA	12.68	14.48	16.29	18.09	0.09	24.15	33.90	43.66	53.42	0.49
Otter	15.49	19.03	22.58	26.12	0.18	25.38	38.89	52.40	65.91	0.68
VPGLTrans	12.51	14.39	16.26	18.14	0.09	20.39	31.95	43.51	55.07	0.58
LLaMA-Adapter-v2	12.52	16.07	19.62	23.17	0.18	22.44	35.31	48.18	61.04	0.64
mPLUG-Owl	12.85	15.84	18.84	21.83	0.15	19.77	30.68	41.60	52.52	0.55
Gemini-Pro-Vision	12.76	15.17	17.57	19.98	0.12	22.63	34.56	46.50	58.44	0.60
InternLM-XComposer	10.93	12.74	14.54	<u>16.34</u>	0.09	20.12	31.22	42.33	53.44	0.56
Qwen-VL	8.37	<u>10.69</u>	13.01	15.33	0.12	14.15	25.00	35.85	46.71	0.54
mPLUG-Owl2	<u>6.91</u>	10.82	14.72	18.63	0.20	11.72	<u>25.45</u>	<u>39.17</u>	52.90	0.69

- The performance of different LVLMs varies at different output lengths, and LeHaCE is able to capture this variation.
- LeHaCE can evaluate the extent to which the object hallucination degree is influenced by the ioutput lengths.

Experimental Results

MSCOCO																
# of Ins	Gemini-Pro-Vision				Qwen-VL				MiniGPT-4				LLaVA			
	C ₁	L _{C₁}	C _S	L _{C_S}	C ₁	L _{C₁}	C _S	L _{C_S}	C ₁	L _{C₁}	C _S	L _{C_S}	C ₁	L _{C₁}	C _S	L _{C_S}
3	0.16	0.25	0.32	0.18	0.29	0.11	0.41	0.09	0.14	0.09	0.22	0.07	0.16	0.23	0.29	0.23
4	0.14	0.16	0.30	0.13	0.27	0.11	0.38	0.15	0.12	0.08	0.21	0.06	0.13	0.20	0.23	0.18
5	0.16	0.15	0.27	0.12	0.22	0.08	0.34	0.05	0.08	0.06	0.15	0.05	0.10	0.08	0.20	0.08
6	0.13	0.09	0.24	0.10	0.21	0.07	0.32	0.06	0.10	0.04	0.16	0.04	0.10	0.07	0.19	0.06
7	0.15	0.12	0.22	0.11	0.18	0.06	0.25	0.05	0.10	0.06	0.15	0.06	0.10	0.06	0.18	0.07
8	0.12	0.09	0.20	0.08	0.15	0.06	0.22	0.04	0.06	0.04	0.12	0.03	0.08	0.06	0.15	0.06
# of Ins	InternLM				Otter				LLaMA-Adapter-v2				mPLUG-Owl			
	C ₁	L _{C₁}	C _S	L _{C_S}	C ₁	L _{C₁}	C _S	L _{C_S}	C ₁	L _{C₁}	C _S	L _{C_S}	C ₁	L _{C₁}	C _S	L _{C_S}
3	0.19	0.13	0.31	0.13	0.24	0.13	0.31	0.14	0.27	0.75	0.32	0.80	0.19	0.08	0.25	0.06
4	0.18	0.07	0.27	0.06	0.19	0.07	0.24	0.07	0.20	0.13	0.25	0.11	0.20	0.07	0.26	0.03
5	0.17	0.08	0.28	0.08	0.15	0.08	0.20	0.09	0.22	0.10	0.27	0.09	0.16	0.06	0.23	0.02
6	0.16	0.05	0.27	0.06	0.18	0.09	0.23	0.09	0.17	0.10	0.21	0.09	0.14	0.04	0.18	0.02
7	0.15	0.05	0.22	0.05	0.14	0.05	0.19	0.05	0.15	0.08	0.19	0.07	0.14	0.04	0.17	0.02
8	0.09	0.05	0.16	0.05	0.10	0.06	0.14	0.06	0.17	0.08	0.21	0.07	0.15	0.04	0.20	0.02
# of Ins	InstructBLIP				mPLUG-Owl2				Lynx				VPGTrans			
	C ₁	L _{C₁}	C _S	L _{C_S}	C ₁	L _{C₁}	C _S	L _{C_S}	C ₁	L _{C₁}	C _S	L _{C_S}	C ₁	L _{C₁}	C _S	L _{C_S}
3	0.43	0.28	0.53	0.20	0.39	0.34	0.50	0.27	0.13	0.19	0.14	0.14	0.14	0.17	0.32	0.14
4	0.37	0.12	0.44	0.10	0.32	0.05	0.41	0.09	0.13	0.07	0.14	0.06	0.13	0.14	0.30	0.13
5	0.43	0.15	0.50	0.11	0.33	0.04	0.44	0.06	0.12	0.07	0.13	0.05	0.10	0.10	0.26	0.09
6	0.28	0.11	0.34	0.10	0.29	0.03	0.37	0.04	0.10	0.05	0.10	0.03	0.11	0.05	0.26	0.07
7	0.30	0.09	0.35	0.10	0.23	0.03	0.28	0.05	0.12	0.04	0.11	0.02	0.10	0.06	0.21	0.05
8	0.36	0.10	0.41	0.09	0.26	0.04	0.31	0.04	0.11	0.05	0.10	0.03	0.08	0.05	0.17	0.04

- LeHaCE demonstrates superior stability compared to the average-based framework

Experimental Results

#Ins	MiniGPT-4									InstructBLIP									LLaVA								
	CHAIR _S			CHAIR _I			CHAIR _S			CHAIR _I			CHAIR _S			CHAIR _I											
	L ₁	L ₂	L ₃	L ₁	L ₂	L ₃	L ₁	L ₂	L ₃	L ₁	L ₂	L ₃	L ₁	L ₂	L ₃	L ₁	L ₂	L ₃									
3	0.07	0.50	0.58	0.09	0.58	0.70	0.20	2.30	2.81	0.28	1.62	2.61	0.23	1.51	1.53	0.23	1.15	1.62									
4	0.06	0.18	0.62	0.08	0.19	0.67	0.10	3.88	1.79	0.12	2.00	1.64	0.18	0.41	3.31	0.20	0.46	2.06									
5	0.05	0.04	0.08	0.06	0.06	0.15	0.11	1.00	2.56	0.15	1.31	2.56	0.08	0.14	0.42	0.08	0.14	0.38									
6	0.04	0.04	0.06	0.04	0.05	0.06	0.10	0.41	0.70	0.11	0.47	0.54	0.06	0.12	0.24	0.07	0.10	0.25									
7	0.06	0.05	0.07	0.06	0.06	0.08	0.10	0.86	1.56	0.09	1.08	0.80	0.07	0.12	0.26	0.06	0.09	0.15									
8	0.03	0.04	0.05	0.04	0.04	0.04	0.09	0.23	0.19	0.10	0.28	0.18	0.06	0.12	0.11	0.06	0.10	0.09									

#Ins	LLaMA-Adapter-v2									Lynx									InternLM-XC								
	CHAIR _S			CHAIR _I			CHAIR _S			CHAIR _I			CHAIR _S			CHAIR _I											
	L ₁	L ₂	L ₃	L ₁	L ₂	L ₃	L ₁	L ₂	L ₃	L ₁	L ₂	L ₃	L ₁	L ₂	L ₃	L ₁	L ₂	L ₃									
3	0.80	2.23	1.76	0.75	1.97	1.69	0.14	0.35	0.36	0.19	0.40	0.40	0.13	1.13	1.91	0.13	2.35	5.83									
4	0.11	0.75	1.03	0.13	1.02	1.06	0.06	0.12	12.90	0.07	0.15	0.84	0.06	0.20	1.78	0.07	0.20	1.13									
5	0.09	0.33	0.60	0.10	0.39	0.79	0.05	0.14	0.22	0.07	0.15	0.14	0.08	0.23	0.67	0.08	1.08	0.54									
6	0.09	0.19	0.43	0.10	0.27	0.51	0.03	0.04	0.08	0.05	0.07	0.13	0.06	0.12	0.14	0.05	0.10	0.21									
7	0.07	0.09	0.10	0.08	0.12	0.15	0.02	0.05	0.13	0.04	0.08	0.14	0.05	0.12	0.12	0.05	0.09	0.08									
8	0.07	0.10	0.20	0.08	0.14	0.27	0.03	0.06	0.05	0.05	0.10	0.08	0.05	0.10	0.12	0.05	0.10	0.10									

#Ins	mPLUG-Owl									Otter									VPGTrans								
	CHAIR _S			CHAIR _I			CHAIR _S			CHAIR _I			CHAIR _S			CHAIR _I											
	L ₁	L ₂	L ₃	L ₁	L ₂	L ₃	L ₁	L ₂	L ₃	L ₁	L ₂	L ₃	L ₁	L ₂	L ₃	L ₁	L ₂	L ₃									
3	0.06	0.93	1.22	0.08	0.67	0.80	0.14	0.49	0.45	0.13	0.59	0.49	0.14	3.27	1.84	0.17	16.25	5.50									
4	0.03	0.12	1.12	0.07	0.10	2.80	0.07	0.13	0.83	0.07	0.12	0.35	0.13	0.87	2.66	0.14	0.85	5.69									
5	0.02	0.10	0.53	0.06	0.08	0.77	0.09	0.12	0.17	0.08	0.12	0.16	0.09	0.24	1.06	0.10	0.16	0.96									
6	0.02	0.05	0.05	0.04	0.07	0.08	0.09	0.10	0.09	0.09	0.11	0.08	0.07	0.14	0.45	0.05	0.10	0.25									
7	0.02	0.04	0.07	0.04	0.06	0.08	0.05	0.08	0.10	0.05	0.09	0.09	0.05	0.13	0.22	0.06	0.05	0.10									
8	0.02	0.05	0.06	0.04	0.06	0.05	0.06	0.09	0.10	0.06	0.09	0.09	0.04	0.11	0.14	0.05	0.06	0.07									

#Ins	Qwen-VL									mPLUG-Owl2									Gemini-Pro-V								
	CHAIR _S			CHAIR _I			CHAIR _S			CHAIR _I			CHAIR _S			CHAIR _I											
	L ₁	L ₂	L ₃	L ₁	L ₂	L ₃	L ₁	L ₂	L ₃	L ₁	L ₂	L ₃	L ₁	L ₂	L ₃	L ₁	L ₂	L ₃									
3	0.09	1.11	0.99	0.11	2.26	1.67	0.27	0.88	0.81	0.34	0.91	0.91	0.18	1.02	1.03	0.25	3.92	1.24									
4	0.15	0.40	4.94	0.11	0.43	19.84	0.09	0.67	4.33	0.05	0.39	1.76	0.13	0.19	1.67	0.16	0.24	1.87									
5	0.05	0.05	3.40	0.08	0.14	2.23	0.06	0.20	4.27	0.04	0.24	0.84	0.12	2.28	0.51	0.15	1.27	1.18									
6	0.06	0.09	0.74	0.07	0.12	0.44	0.04	0.08	0.33	0.03	0.10	0.85	0.10	1.00	0.84	0.09	0.38	0.47									
7	0.05	0.08	0.93	0.06	0.08	0.89	0.05	0.09	0.30	0.03	0.15	0.23	0.11	0.09	0.15	0.12	0.13	0.30									
8	0.04	0.06	0.16	0.06	0.08	0.19	0.04	0.08	33.42	0.04	0.11	0.20	0.08	0.08	0.13	0.09	0.11	0.12									

- Linear fitting performs better than quadratic fitting and cubic fitting.

Conclusion

- Systematic study linking description length to hallucination degree.
- LeHaCE framework provides a Stable, Fair, and Comprehensive Evaluation of Object Hallucination in Large Vision-Language Models



NEURAL INFORMATION PROCESSING SYSTEMS

THANKS