

Table R1: Model performance on factuality evaluation with POPE-based evaluation method.

Dataset	LLaVA-Med	Med-Flamingo	MedVInT	RadFM	LLaVA-v1.6	Qwen-VL-Chat
MIMIC-CXR	71.13	61.27	66.06	69.30	63.70	60.43
Harvard-FairVLMed	61.37	42.06	35.92	52.47	48.52	38.06

Table R3: Model performance in OL3I on factuality evaluation.

Table R2: Label distribution of OL3I dataset.

	Number
Positive	84
Negative	759

Model	Acc	F1	AUROC	Yes Ratio
LLaVA-Med	34.70	26.70	56.73	35.13
Med-Flamingo	61.90	55.32	67.67	25.48
MedVInT	61.90	54.38	68.20	23.75
RadFM	20.50	17.62	51.25	38.69
LLaVA-v1.6	31.54	23.43	58.92	24.74
Qwen-VL-Chat	61.80	54.85	67.70	22.05

Table R4: Model performance on factuality evaluation.

Dataset	LLaVA-Med-v1.0	Med-Flamingo	MedVInT	RadFM	LLaVA-v1.6	Qwen-VL-Chat	LLaVA-Med-v1.5	VILA	InternLM-XComposer2	InternVL2
MIMIC-CXR	71.13	61.27	66.06	69.30	63.70	60.43	75.79	62.50	60.17	59.47
Harvard-FairVLMed	61.37	42.06	35.92	52.47	48.52	38.06	63.03	54.08	47.61	44.38

Table R5: Model performance on uncertainty estimation based on probability-based testing method.

Dataset	Model	AUROC↑	ECE↓	MSE↓
IU-Xray	LLaVA-Med	0.6001	0.1346	0.1908
	Med-Flamingo	0.6407	0.1225	0.1768
	MedVInT	0.5543	0.1659	0.2479
	RadFM	0.5813	0.1481	0.2130
	LLaVA-v1.6	0.6630	0.1122	0.1637
	Qwen-VL-Chat	0.6911	0.1018	0.1511
MIMIC-CXR	LLaVA-Med	0.5975	0.1421	0.1949
	Med-Flamingo	0.5768	0.1542	0.2053
	MedVInT	0.5430	0.1762	0.2546
	RadFM	0.5879	0.1464	0.1986
	LLaVA-v1.6	0.6270	0.1321	0.1870
	Qwen-VL-Chat	0.6148	0.1276	0.1765

Table R7: Model performance in OL3I and IU-Xray on factuality evaluation.

Dataset	Model	Acc	F1	AUROC	Yes Ratio
OL3I	LLaVA-Med	34.70	26.70	56.73	35.13
	Med-Flamingo	61.90	55.32	67.67	25.48
	MedVInT	61.90	54.38	68.20	23.75
	RadFM	20.50	17.62	51.25	38.69
	LLaVA-v1.6	31.54	23.43	58.92	24.74
	Qwen-VL-Chat	61.80	54.85	67.70	22.05
IU-Xray	LLaVA-Med	66.61	54.66	66.70	44.73
	Med-Flamingo	26.74	30.17	55.36	32.59
	MedVInT	73.34	69.24	74.24	31.56
	RadFM	26.67	30.36	55.31	54.28
	LLaVA-v1.6	48.39	39.40	60.70	35.84
	Qwen-VL-Chat	31.17	33.50	56.86	39.79

Table R6: Label distribution of OL3I and IU-Xray dataset.

	OL3I	IU-Xray
Yes	84	639
No	759	1814

Table R8: Results tested on noisy data. We report the model’s performance when confronted with noisy data, as well as the extent of the decline compared to the original performance.

Dataset	Model	Acc	AUROC
OL3I	LLaVA-Med	28.49 ↓ 6.21	52.30 ↓ 4.43
	Med-Flamingo	51.70 ↓ 10.20	59.47 ↓ 8.2
	MedVInT	51.47 ↓ 10.43	58.82 ↓ 9.38
	RadFM	19.04 ↓ 1.46	50.56 ↓ 0.69
IU-Xray	LLaVA-Med	57.28 ↓ 9.33	59.83 ↓ 6.87
	Med-Flamingo	23.29 ↓ 3.45	52.14 ↓ 3.22
	MedVInT	64.38 ↓ 8.96	65.82 ↓ 8.42
	RadFM	25.29 ↓ 1.38	53.69 ↓ 1.62