Experimental Results

PhanVinhPhu - 45747989

April 3, 2024

Domain	Level	Vanila	Cot	Link	Content
Vicuna					
All	seg.	22.3/17.7/30.1	34.0/20.7/94.9	35.3/21.5/98.8	34.7/21.3/93.7
WK	seg.	40.0/31.1/55.9	44.5/29.0/94.6	**45.3**/29.2/100	43.1/28.3/90.3
Sci/Tech	seg.	20.9/14.6/36.8	28.8/16.9/98.2	28.9/16.9/98.2	28.9/17.0/96.5
Wri/rec	seg.	9.3/29.5/5.5	38.5/24.3/93.2	40.7/25.6/99.6	40.5/25.9/92.8
Math	seg.	27.9/20.0/46.5	37.2/23.0/97.2	-	-
Reasoning	seg.	17.2/9.9/63.6	19.0/10.5/96.4	-	-
LLama					
All	seg.	40.0/20.9/67.5	29.3/20.2/53.2	31.3/20.3/68.5	29.5/19.2/63.2
WK	seg.	42.4/29.8/73.1	30.8/28.7/33.3	41.6/28.9/74.2	**43.4**/29.6/81.7
Sci/Tech	seg.	28.7/18.0/70.2	25.8/23.9/28.1	29.0/17.8/78.9	21.5/13.6/50.9
Wri/rec	seg.	37.6/27.2/60.8	33.8/25.2/51.1	35.9/10.3/85.5	33.8/24.2/56.2
Math	seg.	33.0/21.8/67.6	35.4/22.7/80.3	-	-
Reasoning	seg.	18.2/10.2/83.6	19.2/10.8/87.3	-	-
Mistralai					
All	seg.	22.3/17.7/30.1	23.3/19.3/29.5	24.5/19.1/34.0	21.4/16.8/29.5
WK	seg.	40.0/31.1/55.9	37.6/29.9/50.5	**42.6**/32.9/60.2	39.7/30.2/58.1
Sci/Tech	seg.	20.9/14.6/36.8	19.6/14.1/31.6	25.2/17.7/43.9	15.8/11.0/28.1
Wri/rec	seg.	9.3/29.5/5.5	13.5/24.4/9.36	17.2/10.0/63.6	8.9/22.4/5.5
Math	seg.	28.0/20.0/46.5	32.5/21.7/64.8	-	-
Reasoning	seg.	17.2/10.0/63.6	14.3/9.2/32.7	-	-

Table 1: Segment-level results of factual error detectors powered by Vicuna, LLama and Mistralai on FELM, numbers are arranged according to F1/Precsion/Recall. We do not involve claim-based methods for math and reasoning domains cause it is often difficult to extract self-contained, atomic claims from these two domains. There is no reference for math and reasoning either

Method	All	WK	Sci/Tech	Writing/Rec	Math	Reasoning
Vicuna						
Vanila	50.2	50.5	48.6	50.2	**53**	48.4
Cot	48.7	50.2	49.8	48.1	50.1	48.4
Link	50.8	50.6	50	51.4	-	-
Doc	50.3	48.7	50.2	52.1	-	-
LLama						
Vanila	49.5	51.5	52.6	**53.2**	47.7	47.1
Cot	48.3	49.8	54.9	50.2	49.4	49.7
Link	48	49.8	52.3	50.3	**-**	-
Doc	45.6	51.2	42.6	48.7	-	-
Mistralai						
Vanila	46.2	52.7	46.5	50.6	45.5	46.9
Cot	48.1	51.1	46.3	49.8	47.5	46.7
Link	47.7	**55.1**	51.2	51.3	**-**	**-**
Doc	45.1	51.6	41	49.6	**-**	**-**

Table 2: Segment-level balanced accuracy of factual error detectors powered by Vicuna, LLama and Mistralai on FELM.