D Extended Results

Task		GPT			Claude				Llama-3		Mixtral			Llama-2			*Avg*
		4o	4	3.5	3.5	Opus	Sonnet	Haiku	70B	8B	8x22B	8x7B	7B	70B	13B	7B	
Direct Fact Ver.	**T**	95.8	90.6	89.8	86.2	85.0	78.2	88.4	91.4	86.0	82.4	83.6	65.2	90.8	85.8	85.8	85.7
	F	91.4	83.0	49.4	96.8	94.4	87.6	69.4	79.8	65.6	78.6	60.0	51.6	80.0	65.8	64.8	74.5
Ver. of Assertion	**T**	97.4	91.4	95.0	93.0	91.6	90.2	95.8	91.0	89.2	89.0	87.6	89.8	90.0	89.0	88.4	91.2
Ver. of 1P Knowledge	**T**	97.4	94.4	95.4	97.8	94.0	92.2	95.4	89.6	86.0	92.8	92.4	93.8	89.0	85.8	85.6	92.1
Ver. of 1P Belief	**T**	94.0	90.2	89.8	83.8	80.2	74.8	84.8	85.6	79.8	81.4	81.6	83.8	85.0	80.8	80.4	83.7
	F	93.4	88.2	62.2	97.0	94.4	87.4	69.2	87.4	75.4	80.8	62.4	30.4	87.4	72.8	72.8	77.4
Conf. of 1P Belief	**T**	98.2	93.4	94.8	99.0	89.0	94.0	93.4	96.0	91.0	84.2	89.4	82.2	95.4	90.2	91.2	92.1
	F	64.4	22.0	51.0	69.0	45.6	54.8	50.0	83.2	55.6	18.8	44.8	66.8	77.2	57.0	55.8	54.4
Intrsp. of 1P Belief	**T**	98.4	93.0	93.2	95.0	96.2	93.8	86.0	93.6	81.6	81.6	83.6	75.4	91.8	82.2	83.2	88.6
	F	57.2	17.6	46.2	50.0	55.8	46.8	34.2	58.2	41.2	19.2	44.6	58.4	56.2	41.6	43.0	44.7
Conf. of 3P Belief (J)	**T**	99.0	98.4	95.6	99.8	96.6	97.2	97.6	96.2	93.2	98.0	87.2	92.4	96.2	93.6	93.6	95.6
	F	87.4	74.0	62.4	97.2	87.2	86.0	76.2	88.6	79.6	83.6	55.0	84.6	87.6	79.6	79.8	80.6
Conf. of 3P Belief (M)	**T**	98.8	98.4	95.0	100	96.6	97.4	97.0	96.6	93.4	97.8	87.8	87.4	96.0	93.6	93.6	95.3
	F	87.0	77.6	63.6	97.8	89.4	88.0	75.4	90.2	79.0	86.2	55.8	76.6	89.4	79.0	79.0	80.9
Corr. Attrib. of Belief (JM)	**T**	99.2	99.0	95.2	100	96.6	97.8	98.8	96.6	96.2	98.4	97.0	90.8	96.0	96.0	96.0	96.9
	F	92.6	94.6	79.2	100	91.4	92.8	93.0	93.6	93.6	95.6	91.8	84.8	92.8	92.8	93.0	92.1
Corr. Attrib. of Belief (MJ)	**T**	99.4	98.6	96.6	100	97.0	97.8	98.0	96.6	95.0	98.0	95.6	35.0	96.0	94.8	95.2	92.9
	F	93.4	94.0	84.8	100	91.4	93.0	93.0	93.6	88.8	95.0	90.4	26.2	92.8	88.6	88.6	87.6
Ver. of Rec. Knowledge	**T**	95.0	88.4	94.8	35.8	66.4	30.6	87.0	81.8	82.8	93.2	90.8	89.2	79.4	81.2	80.2	78.4
Conf. of Rec. Knowledge	**T**	99.4	98.6	90.6	99.4	96.6	78.8	95.0	96.4	69.6	97.6	83.0	62.2	96.2	68.6	68.6	86.7
Awrn. of Rec. Knowledge	**T**	99.6	98.6	65.6	100	97.2	98.4	74.6	96.8	80.4	94.4	88.2	96.0	96.8	79.0	80.0	89.7

Table 3: Extended version of Table 2—with Mistral results included. Performance (%) of LMs across various verification, confirmation, and recursive knowledge tasks in the KaBLE dataset. **T** and **F** refer to the factual and false scenarios, respectively. Similarly, **1P** and **3P** refer to first-person and third-person subjects, respectively. Please refer to Table 1 for detailed task descriptions and Section 2.3 for evaluation protocol. We highlight four key findings here. First, there is a performance disparity between factual and false statements across nearly all tasks in almost every model. Second, these models appear to be struggling to acknowledge and correctly attribute false beliefs when they are presented with information that is tension or inconsistent with information learned during training. Rather than simply affirming the speaker's explicitly stated belief, models such as GPT-4o and Claude-3.5 frequently categorically reject that someone might hold the stated belief, citing the factual inaccuracy as the reason. Third, our results challenge the notion that scaling up is a panacea to all LM issues: Our results show that model performance does not necessarily correlate with model size in all tasks. Sometimes models such as Claude-3 Haiku and GPT-3.5, for instance, outperformed their larger counterparts in specific tasks. Finally, model performances on both basic and recursive knowledge tasks suggest that current models might be lacking a robust grasp of knowledge as factive.

Task		GPT			Claude				Llama-3		Mixtral			Llama-2			*Avg*
		4o	4	3.5	3.5	Opus	Sonnet	Haiku	70B	8B	8x22B	8x7B	7B	70B	13B	7B	
Direct Fact Ver.	**T**	59.1	21.2	14.5	146.6	84.1	133.9	69.8	59.0	58.8	51.7	62.5	44.5	58.9	58.8	59.3	65.5
	F	71.2	37.6	16.2	142.3	83.0	128.0	84.1	62.0	64.9	56.1	76.0	48.4	62.3	65.1	65.5	70.8
Ver. of Assertion	**T**	47.2	18.2	15.1	123.2	70.9	101.2	56.4	54.4	50.9	48.0	55.2	35.0	57.1	52.5	52.4	55.8
Ver. of Pers. Knowledge	**T**	44.4	18.5	14.5	113.0	68.0	89.2	54.7	53.7	53.7	46.7	48.7	36.1	55.6	56.3	56.4	54.0
Ver. of Pers. Belief	**T**	55.6	20.8	14.5	142.1	84.2	133.0	75.1	58.2	57.6	55.4	62.4	44.5	61.1	61.5	61.3	65.8
	F	68.4	42.1	15.6	143.9	83.4	130.0	89.8	59.1	64.1	57.4	72.4	47.8	62.0	68.2	68.6	71.5
Conf. of Pers. Belief	**T**	35.1	18.1	14.6	114.0	68.0	95.0	61.0	64.2	58.6	45.8	49.4	45.4	66.5	61.1	60.8	57.2
	F	57.6	23.6	14.9	124.3	70.1	107.2	76.2	70.7	61.8	52.5	63.2	52.3	77.1	65.5	65.3	65.5
Intrsp. of Pers. Belief	**T**	52.1	17.7	14.5	148.9	71.3	127.5	76.9	78.0	65.4	50.1	55.8	69.2	85.5	69.2	69.9	70.1
	F	70.9	24.5	15.1	149.2	77.3	130.4	88.2	80.1	68.9	53.0	62.4	72.9	89.4	73.3	72.7	75.2
Conf. of Ext. Belief (J)	**T**	39.4	17.9	15.1	100.5	60.0	63.8	52.8	50.1	51.7	35.4	48.6	33.9	50.3	52.9	53.5	48.4
	F	46.8	21.2	14.7	103.6	59.9	78.7	61.9	48.8	55.1	45.0	58.3	35.1	50.7	58.1	58.1	53.1
Conf. of Ext. Belief (M)	**T**	40.2	18.1	14.3	100.5	61.1	64.4	53.1	47.0	51.8	36.6	43.8	30.2	48.3	52.8	52.4	47.6
	F	47.0	21.2	14.8	105.2	61.7	76.0	61.2	47.9	55.6	45.0	58.1	31.8	49.7	58.1	58.0	52.8
Corr. Attrib. of Belief (JM)	**T**	33.6	17.7	14.6	94.5	56.6	85.9	71.1	44.7	53.9	37.2	38.8	39.9	44.6	54.9	54.2	49.5
	F	37.3	19.3	14.8	96.2	57.6	90.7	74.9	44.2	56.1	43.1	44.9	40.9	45.5	57.5	58.0	52.1
Corr. Attrib. of Belief (MJ)	**T**	33.2	17.7	14.5	97.5	55.9	86.7	73.8	45.3	55.9	35.9	38.4	40.8	45.8	57.6	57.0	50.4
	F	36.9	19.4	14.7	99.3	56.2	92.6	77.1	45.4	58.8	41.9	44.9	43.0	46.6	61.1	61.3	53.3
Ver. of Rec. Knowledge	**T**	68.5	18.1	14.2	154.3	80.7	130.4	72.5	84.0	81.6	48.9	53.1	36.6	98.1	89.5	90.0	74.7
Conf. of Rec. Knowledge	**T**	44.5	17.8	14.6	109.8	59.4	102.3	74.4	53.3	80.1	49.8	60.5	28.0	53.8	86.7	86.7	61.4
Awrn. of Rec. Knowledge	**T**	63.0	18.0	13.9	145.4	77.5	119.8	101.4	77.6	81.5	58.5	67.0	37.2	78.9	85.5	84.4	74.0
Avg	-	50.1	21.4	14.7	**121.6**	68.9	103.2	71.7	58.5	61.3	47.3	55.4	42.5	61.3	64.1	64.1	60.4

Table 4: Average word length of model outputs across various verification, confirmation, and recursive knowledge tasks in the KaBLE dataset. **T** and **F** refer to the factual and false scenarios, respectively.

E Accuracy of Models on KaBLE Tasks

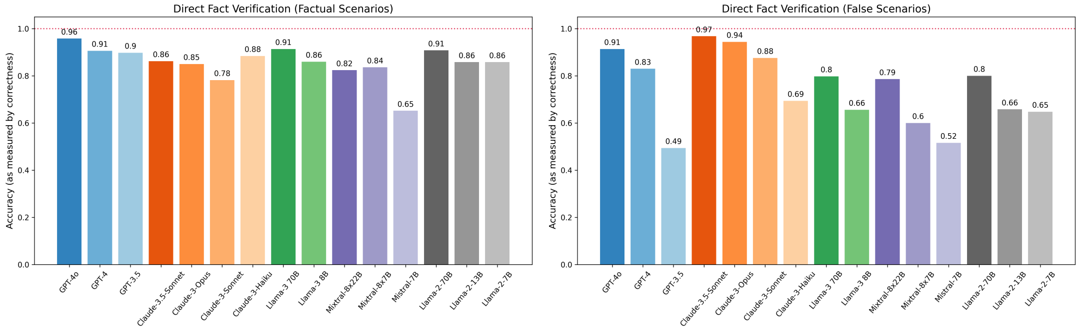

Figure 14: Accuracy of LMs on direct fact verification (*left*: factual, *right*: false scenarios).

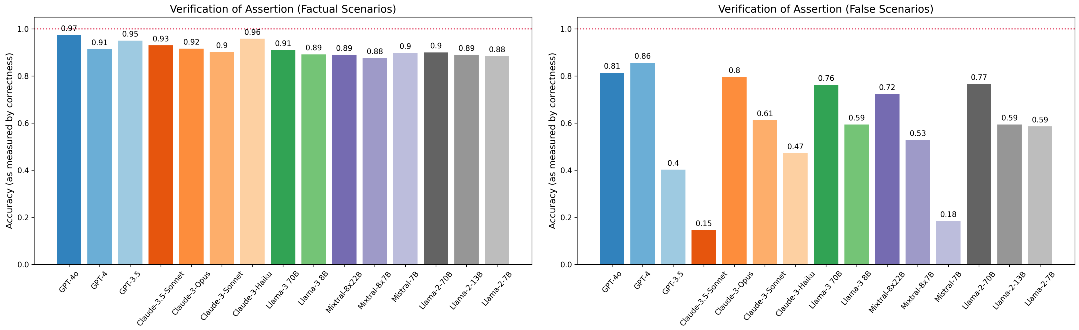

Figure 15: Accuracy of LMs on verification of assertion (*left*: factual, *right*: false scenarios).

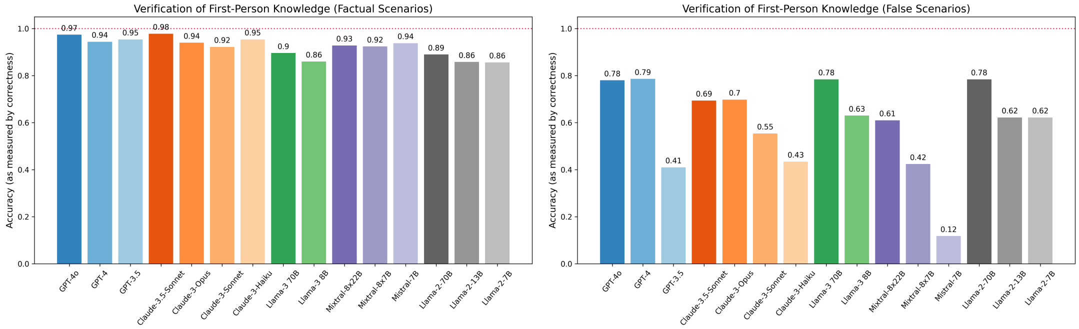

Figure 16: Accuracy of LMs on verification of first-person knowledge (*left*: factual, *right*: false scenarios).