A Detailed performance of MAW

We report the average and maximum MAW accuracy across different layers in Table 6. The average MAW of 6 layers significantly outperforms the random baseline, which indicates that the relevant question concept plays a highly important role in BERT encoding without fine-tuning. BERT-FT outperforms BERT in terms of both average MAW accuracy and maximum MAW accuracy, which shows that structured commonsense knowledge is enhanced by supervised training on commonsense tasks.

	BERT-FT			**BERT**			
L	Max	Avg	t	Max	Avg	t	Rand
11	34.11	19.78	✓	32.44	14.47	✓	10.53
10	39.09	26.10	✓	40.84	22.22	✓	10.53
9	46.31	25.59	✓	**46.82**	18.68	✓	10.53
8	**49.22**	13.71	✓	44.48	10.15	-	10.53
7	32.76	8.88	-	28.00	5.61	-	10.53
6	40.68	12.16	✓	41.99	9.01	-	10.53
5	33.30	14.41	✓	13.22	4.34	-	10.53
4	38.89	19.09	✓	24.10	10.46	-	10.53
3	37.30	14.59	✓	24.74	7.43	-	10.53
2	35.08	17.71	✓	31.96	12.14	✓	10.53
1	29.01	15.08	✓	27.64	11.09	✓	10.53
0	45.55	23.05	✓	46.16	22.95	✓	10.53
All	49.22	17.35	✓	46.82	12.38	✓	10.53

Table 6: The average and maximum MAW accuracy across different layers. ✓indicates p-value <0.01.

B Performance of MAT

Table 7 shows the MAT and MAS performance for each attention head across five turns. Noted that the standard derivations are only 1.17% and 1.76% for MAT and MAS, respectively, which demonstrates the robustness of our methods.

C Implementation Details

We adopt the huggingface BERT-base implementation for multiple-choice on CommonsenseQA. We conduct fine-tuning experiments using GeForce GTX 2080Ti. For BERT-FT and BERT-probing, we optimize the parameters with grid search: training epochs 3, learning rate $\{5e-4, 3e-5, 5e-5, 5e-6\}$, training batch size $\{8, 16, 32\}$, gradient accumulation steps $\{2, 4, 8\}$. To demonstrate the robustness of our analysis method, we repeat the experiment 5 times with the same hyperparameter, and report the experiment results based on one random model .

We calculate the attribution score to interpret BERT using captum, which is an extensible library for model interpret ability built on Pytorch.

	MAT						MAS					
Layer	M1	M2	M3	M4	M5	mean±std	M1	M2	M3	M4	M5	mean±std
0	20.10	20.26	20.32	20.35	20.58	20.32±0.17	18.86	18.38	18.40	18.07	18.19	18.38±0.30
1	19.35	19.36	19.11	19.39	19.51	19.34±0.15	21.00	20.09	20.37	20.16	20.23	20.37±0.37
2	20.25	19.57	19.97	20.13	19.85	19.95±0.26	19.20	19.61	20.19	19.03	19.01	19.41±0.50
3	21.98	21.15	21.80	22.05	23.16	22.03±0.73	18.88	19.46	18.39	18.12	19.01	18.77±0.53
4	19.78	19.19	19.91	20.15	19.27	19.66±0.42	19.99	19.50	20.05	22.28	20.76	20.52±1.08
5	19.90	19.74	19.89	20.33	20.55	20.08±0.34	16.98	16.90	17.90	17.56	17.15	17.30±0.42
6	19.29	19.31	19.02	19.25	18.51	19.08±0.34	15.77	16.25	15.72	18.27	17.15	16.63±1.08
7	20.71	20.47	19.92	20.14	19.30	20.11±0.54	18.40	19.70	16.57	22.09	20.66	19.48±2.11
8	19.48	21.20	21.00	20.65	19.13	20.29±0.93	18.16	19.70	17.28	21.89	19.12	19.23±1.75
9	21.95	21.85	22.87	21.80	22.88	22.27±0.55	12.54	14.94	13.07	14.85	13.75	13.83±1.06
10	25.59	24.95	25.61	24.96	25.07	25.24±0.34	16.60	17.80	15.28	18.12	16.68	16.90±1.12
11	45.91	45.87	44.61	42.28	41.88	44.11±1.93	10.32	18.21	6.04	22.73	22.80	16.02±7.55

Table 7: MAT and MAS overlapping rate for each attention head across five models, as well as their average value with a standard deviation. M - Model.