Table 16: **Results on the datasets concerning knowledge and understanding.** Specifically, we report the results on BoolQ, CommonsenseQA, NaturalQuestions, and LAMBADA.

Model	Params	BoolQ	CommonsenseQA	NaturalQuestions	LAMBADA
MPT	7B	75.0	61.8	11.6	70.0
Falcon	7B	67.5	20.8	15.7	-
ChatGLM2	6B	79.0	65.4	9.7	54.3
InternLM	7B	64.1	59.8	8.9	67.0
	20B	87.5	70.6	25.2	71.8
XVERSE	13B	64.2	62.2	0.3	48.2
Baichuan2	7B	63.2	63.0	9.4	73.3
	13B	67.0	65.6	16.3	74.0
LLaMA	7B	76.5	64.9	16.8	73.3
	13B	78.7	67.4	20.2	75.2
	33B	84.4	72.5	30.9	77.2
	65B	86.6	74.1	33.4	77.7
Llama 2	7B	77.4	66.5	19.1	73.3
	13B	82.4	67.3	**24.9**	**76.5**
	70B	87.7	78.5	34.2	78.9
StableBeluga2	70B	89.4	72.6	25.1	71.3
Qwen	1.8B	68.0	60.1	3.2	58.4
	7B	76.4	66.8	17.4	67.9
	14B	**86.2**	**70.3**	23.9	71.1

Table 17: **Results on the datasets related to natural language reasoning.** Specifically, we report the results on HellaSwag, PIQA, SIQA, and OCNLI.

Model	Params	HellaSwag	PIQA	SIQA	OCNLI
MPT	7B	76.4	**80.6**	48.5	30.0
Falcon	7B	74.1	76.7	47.2	-
ChatGLM2	6B	57.0	69.6	64.3	33.1
InternLM	7B	70.6	77.9	60.5	37.5
	20B	78.1	80.3	72.8	42.5
Baichuan2	7B	67.0	76.2	44.4	30.3
	13B	70.8	78.1	44.3	30.0
LLaMA	7B	76.1	79.8	48.9	33.6
	13B	79.2	80.1	52.5	32.1
	33B	82.8	82.3	57.8	30.7
	65B	84.2	82.8	61.2	44.9
Llama 2	7B	77.2	78.8	48.5	32.1
	13B	**80.7**	80.5	54.8	34.1
	70B	85.3	82.8	64.8	46.5
StableBeluga2	70B	84.1	83.3	78.1	48.3
Qwen	1.8B	56.7	73.3	56.1	39.0
	7B	75.1	77.9	69.9	47.4
	14B	80.2	79.9	**77.9**	**57.9**

- **SIQA** (Sap et al., 2019) This is an NLI dataset evaluating social commonsense intelligence. We report zero-shot results.
- **OCNLI** (Hu et al., 2020) This is an NLI dataset focusing on Chinese. We report zero-shot results.