Table 15: **Results on the other datasets of examination**. Specifically, we report the results on CMMLU, AGIEval, ARC-e, and ARC-c.

| Model | Params | CMMLU | AGIEval | Gaokao-Bench | ARC-e | ARC-c |
|---|---|---|---|---|---|---|
| MPT | 7B | 25.9 | 21.3 | 19.8 | 70.2 | 42.6 |
| Falcon | 7B | - | - | - | 70.0 | 42.4 |
| ChatGLM2 | 6B | 49.3 | 39.0 | 46.4 | 73.0 | 61.0 |
| InternLM | 7B | 51.8 | 36.9 | 43.0 | 78.7 | 69.5 |
| | 20B | 59.0 | 44.6 | 45.5 | 86.1 | 81.7 |
| Baichuan2 | 7B | 57.1 | 42.7 | 47.5 | 54.7 | 32.5 |
| | 13B | 62.0 | 48.2 | 54.3 | 61.9 | 38.0 |
| LLaMA | 7B | 26.8 | 20.6 | 21.3 | 72.8 | 47.6 |
| | 13B | 31.5 | 22.0 | 20.4 | 74.8 | 52.7 |
| | 33B | 36.0 | 33.5 | 18.9 | 80.0 | 67.5 |
| | 65B | 40.6 | 33.9 | 19.1 | 80.6 | 69.5 |
| LLAMA 2 | 7B | 31.8 | 21.8 | 18.9 | 75.2 | 45.9 |
| | 13B | 38.4 | 30.9 | 18.2 | 77.3 | 60.3 |
| | 70B | 53.6 | 40.2 | 23.3 | 85.9 | 78.3 |
| StableBeluga2 | 70B | 51.8 | 41.6 | 40.9 | 91.2 | 86.1 |
| QWEN | 1.8B | 49.3 | 36.9 | 44.9 | 71.6 | 53.2 |
| | 7B | 62.2 | 45.8 | 52.5 | 84.0 | 75.3 |
| | 14B | **71.0** | **52.3** | **61.9** | **90.3** | **84.4** |

the parts of Chinese and English, while LLAMA 2 only reported the results in the English part, so we use the results on OpenCompass. Additionally, while CMMLU, AGIEval, and Gaokao-Bench are related to Chinese, and MPT, Falcon, and the LLaMA series were not optimized for Chinese, these models achieved low performance on the datasets.

**Knowledge and Understanding** Here we evaluate the models on a series of datasets relevant to knowledge and natural language understanding. The datasets include

- **BoolQ** (Clark et al., 2019) This is a QA dataset, where the questions are about passages of Wikipedia, and the model should answer yes or no to the given possible answer. We report zero-shot results.
- **CommonsenseQA** (Talmor et al., 2019) This is a dataset of multiple-choice question answering that asseses the understanding of commonsense knowledge. We report 8-shot results.
- **NaturalQuestions** (Kwiatkowski et al., 2019) It is a dataset of QA where the questions are from users and the answers are verified by experts. We report zero-shot results.
- **LAMBADA** (Paperno et al., 2016) This is dataset to evaluate language understanding by word prediction. It consists of passages related to human subjects. We report zero-shot results.

We report the results in Table 16.

**Reasoning** We report the evaluation results on the datasets concerning reasoning, focusing on natural language reasoning. For the others, such as mathematics and coding, as we have illustrated detailed results, here we do not report those results repeatedly. The datasets for evaluation include:

- **HellaSwag** (Zellers et al., 2019) This is a commonsense natural language inference (NLI) dataset, where the questions are easy for humans but struggling for previous language models. We report zero-shot results.
- **PIQA** (Bisk et al., 2020) This is an NLI dataset assessing the physical knowledge. We report zero-shot results.