Table 1: **Performance Evaluation of SpikingBrain Chat Models.** All models are tested with the vLLM framework and evaluated using a generation-based method. Except for Qwen2.5, the other baselines are trained on limited Chinese data, resulting in clear disadvantages on CMMLU and C-Eval. For QuALITY and IFEval, we report results from the non-CoT model (after SFT stage 2) to avoid chain-of-thought interference.

| | **SpikingBrain-7B** | **SpikingBrain-76B** | Llama3 dubey2024llama | Qwen2.5 Yang2024Qwen2TR | Mixtral jiang2024mixtral |
|---|---|---|---|---|---|
| Params | 7B | 12B/76B | 8B | 7B | 13B/47B |
| Complexity Type | Linear | Hybrid | Quadratic | Quadratic | Quadratic |
| **Benchmarks** | | | | | |
| MMLU hendrycks2020mmlu | **65.57** | **73.71** | 68.69 | 75.17 | 71.03 |
| CMMLU li2023cmmlu | **68.76** | **77.41** | 55.17 | 79.14 | 51.03 |
| HS zellers2019hellaswag | **68.95** | **86.63** | 76.80 | 85.39 | 75.63 |
| Ceval huang2023ceval | **69.07** | **76.32** | 55.01 | 77.93 | 50.88 |
| NQ kwiatkowski2019nq | **21.47** | **21.55** | 30.97 | 17.67 | 28.48 |
| TrQ joshi2017triviaqa | **57.03** | **55.13** | 65.78 | 55.72 | 71.00 |
| QuALITY pang-etal-2022-quality | **60.12** | **69.56** | 66.25 | 73.63 | 51.34 |
| IFEval zhou2023instruction | **42.70** | **49.72** | 73.01 | 73.20 | 48.06 |