Table 5: **Performance of aligned models on widely-used benchmarks.** We report both zero-shot and few-shot performance of the models.

Model	Params	MMLU 0-shot / 5-shot	C-Eval 0-shot / 5-shot	GSM8K 0-shot / 8-shot	HumanEval 0-shot	BBH 0-shot / 3-shot
Proprietary models						
GPT-3.5	-	- / 69.1	- / 52.5	- / 78.2	73.2	- / 70.1
GPT-4	-	- / **83.0**	- / **69.9**	- / **91.4**	**86.6**	- / **86.7**
Open-source models						
ChatGLM2	6B	45.5 / 46.0	50.1 / 52.6	- / 28.8	11.0	- / 32.7
InternLM-Chat	7B	- / 51.1	- / 53.6	- / 33.0	14.6	- / 32.5
Baichuan2-Chat	7B	- / 52.9	- / 55.6	- / 32.8	13.4	- / 35.8
	13B	- / 57.3	- / 56.7	- / 55.3	17.7	- / 49.9
LLAMA 2-CHAT	7B	- / 46.2	- / 31.9	- / 26.3	12.2	- / 35.6
	13B	- / 54.6	- / 36.2	- / 37.1	18.9	- / 40.1
	70B	- / 63.8	- / 44.3	- / 59.3	32.3	- / 60.8
QWEN-CHAT	1.8B	42.4 / 43.9	50.7 / 50.3	27.8 / 19.5	14.6	27.1 / 25.0
	7B	55.8 / 57.0	59.7 / 59.3	50.3 / 54.1	37.2	39.6 / 46.7
	14B	64.6 / **66.5**	69.8 / **71.7**	**60.1** / 59.3	**43.9**	46.9 / **58.7**

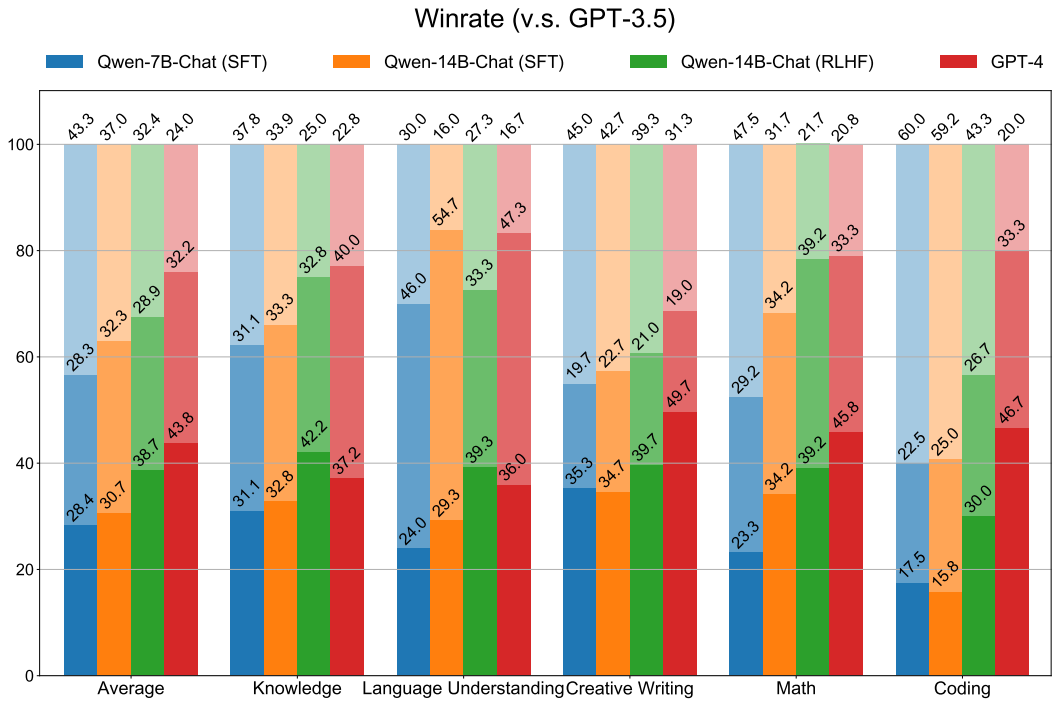

Figure 4: **Results of the human evaluation for chat models.** We compare Qwen-7B (SFT), Qwen-14B (SFT), Qwen-14B (RLHF), as well as GPT-4 against GPT-3.5. Each bar segment represents the percentage of wins, ties, and losses, from bottom to top. On average, the RLHF model outperforms the SFT model and falls behind GPT-4 by a relatively small margin.