Table 2: **Overall performance on widely-used benchmarks compared to open-source base models**. Our largest Qwen model with 14 billion parameters outperforms previous 13B SoTA models on all datasets.

Model	**Params**	**MMLU**	**C-Eval**	**GSM8K**	**MATH**	**HumanEval**	**MBPP**	**BBH**
		5-shot	5-shot	8-shot	4-shot	0-shot	3-shot	3-shot
MPT	7B	30.8	23.5	9.1	3.0	18.3	22.8	35.6
	30B	47.9	-	15.2	3.1	25.0	32.8	38.0
Falcon	7B	27.8	-	6.8	2.3	-	11.2	28.0
	40B	57.0	-	19.6	5.5	-	29.8	37.1
ChatGLM2	6B	47.9	51.7	32.4	6.5	-	-	33.7
InternLM	7B	51.0	53.4	31.2	6.3	10.4	14.0	37.0
	20B	62.1	58.8	52.6	7.9	25.6	35.6	52.5
Baichuan2	7B	54.7	56.3	24.6	5.6	18.3	24.2	41.6
	13B	59.5	59.0	52.8	10.1	17.1	30.2	49.0
LLaMA	7B	35.6	27.3	11.0	2.9	12.8	17.7	33.5
	13B	47.7	31.8	20.3	4.2	15.8	22.0	37.9
	33B	58.7	37.5	42.3	7.1	21.7	30.2	50.0
	65B	63.7	40.4	54.4	10.6	23.7	37.7	58.4
Llama 2	7B	46.8	32.5	16.7	3.3	12.8	20.8	38.2
	13B	55.0	41.4	29.6	5.0	18.9	30.3	45.6
	34B	62.6	-	42.2	6.2	22.6	33.0	44.1
	70B	69.8	50.1	63.3	13.5	29.9	45.0	64.9
StableBeluga2	70B	68.6	51.4	69.6	14.6	28.0	11.4	69.3
Qwen	1.8B	44.6	54.7	21.2	5.6	17.1	14.8	28.2
	7B	58.2	63.5	51.7	11.6	29.9	31.6	45.0
	14B	**66.3**	**72.1**	**61.3**	**24.8**	**32.3**	**40.8**	**53.4**

2.4 Training

To train Qwen, we follow the standard approach of autoregressive language modeling, as described in Radford et al. (2018). This involves training the model to predict the next token based on the context provided by the previous tokens. We train models with context lengths of 2048. To create batches of data, we shuffle and merge the documents, and then truncate them to the specified context lengths. To improve computational efficiency and reduce memory usage, we employ Flash Attention in the attention modules (Dao et al., 2022). We adopt the standard optimizer AdamW (Kingma & Ba, 2014; Loshchilov & Hutter, 2017) for pretraining optimization. We set the hyperparameters $\beta_1 = 0.9$, $\beta_2 = 0.95$, and $\epsilon = 10^{-8}$. We use a cosine learning rate schedule with a specified peak learning rate for each model size. The learning rate is decayed to a minimum learning rate of 10% of the peak learning rate. All the models are trained with BFloat16 mixed precision for training stability.

2.5 Experimental Results

To evaluate the zero-shot and few-shot learning capabilities of our models, we conduct a thorough benchmark assessment using a series of datasets. We compare Qwen with the most recent open-source base models, including LLaMA (Touvron et al., 2023a), LLaMA2 (Touvron et al., 2023b), MPT (Mosaic ML, 2023), Falcon (Almazrouei et al., 2023), Baichuan2 (Yang et al., 2023), ChatGLM2 (ChatGLM2 Team, 2023), InternLM (InternLM Team, 2023), XVERSE (Inc., 2023b), and StableBeluga2 (Stability AI, 2023). Our evaluation covers a total of 7 popular benchmarks, which are MMLU (5-shot) (Hendrycks et al., 2020), C-Eval (5-shot) (Huang et al., 2023), GSM8K (8-shot) (Cobbe et al., 2021), MATH (4-shot) (Hendrycks et al., 2021), HumanEval (0-shot) (Chen et al., 2021b), MBPP (0-shot) (Austin et al., 2021), and BBH (Big Bench Hard) (3 shot) (Suzgun et al., 2022). We aim to provide a comprehensive summary of the overall performance of our models across these benchmarks.