

Table 4: Test Accuracy of QWEN PMP and reward model on diverse human preference benchmark datasets.

Model	QWEN Helpful-base	QWEN Helpful-online	Anthropic Helpful-base	Anthropic Helpful-online	OpenAI Summ.	Stanford SHP	OpenAI PRM800K
PMP	62.68	61.62	76.52	65.43	69.60	60.05	70.59
RM	74.78	69.71	73.98	64.57	69.99	60.10	70.52

3.2.2 REINFORCEMENT LEARNING

Our Proximal Policy Optimization (PPO) process involves four models: the policy model, value model, reference model, and reward model. Before starting the PPO procedure, we pause the policy model’s updates and focus solely on updating the value model for 50 steps. This approach ensures that the value model can adapt to different reward models effectively.

During the PPO operation, we use a strategy of sampling two responses for each query simultaneously. This strategy has proven to be more effective based on our internal benchmarking evaluations. We set the KL divergence coefficient to 0.04 and normalize the reward based on the running mean.

The policy and value models have learning rates of 1×10^{-6} and 5×10^{-6} , respectively. To enhance training stability, we utilize value loss clipping with a clip value of 0.15. For inference, the policy top-p is set to 0.9. Our findings indicate that although the entropy is slightly lower than when top-p is set to 1.0, there is a faster increase in reward, ultimately resulting in consistently higher evaluation rewards under similar conditions.

Additionally, we have implemented a pre-trained gradient to mitigate the alignment tax. Empirical findings indicate that, with this specific reward model, the KL penalty is adequately robust to counteract the alignment tax in benchmarks that are not strictly code or math in nature, such as those that test common sense knowledge and reading comprehension. It is imperative to utilize a significantly larger volume of the pretrained data in comparison to the PPO data to ensure the effectiveness of the pretrained gradient. Additionally, our empirical study suggests that an overly large value for this coefficient can considerably impede the alignment to the reward model, eventually compromising the ultimate alignment, while an overly small value would only have a marginal effect on alignment tax reduction.

3.3 AUTOMATIC AND HUMAN EVALUATION OF ALIGNED MODELS

To showcase the effectiveness of our aligned models, we conduct a comparison with other aligned models on well-established benchmarks, including MMLU (Hendrycks et al., 2020), C-Eval (Huang et al., 2023), GSM8K (Cobbe et al., 2021), HumanEval (Chen et al., 2021b), and BBH (Suzgun et al., 2022). Besides the widely used few-shot setting, we test our aligned models in the zero-shot setting to demonstrate how well the models follow instructions. The prompt in a zero-shot setting consists of an instruction and a question without any previous examples in the context. The results of the baselines are collected from their official reports and OpenCompass (OpenCompass Team, 2023).

The results in Table 5 demonstrate the effectiveness of our aligned models in understanding human instructions and generating appropriate responses. QWEN-14B-Chat outperforms all other models except ChatGPT (OpenAI, 2022) and LLAMA 2-CHAT-70B (Touvron et al., 2023b) in all datasets, including MMLU (Hendrycks et al., 2020), C-Eval (Huang et al., 2023), GSM8K (Cobbe et al., 2021), HumanEval (Chen et al., 2021b), and BBH (Suzgun et al., 2022). In particular, QWEN’s performance in HumanEval, which measures the quality of generated codes, is significantly higher than that of other open-source models.

Moreover, QWEN’s performance is consistently better than that of open-source models of similar size, such as LLaMA2 (Touvron et al., 2023b), ChatGLM2 (ChatGLM2 Team, 2023), InternLM (InternLM Team, 2023), and Baichuan2 (Yang et al., 2023). This suggests that our alignment approach, which involves fine-tuning the model on a large dataset of human conversations, has been effective in improving the model’s ability to understand and generate human-like language.