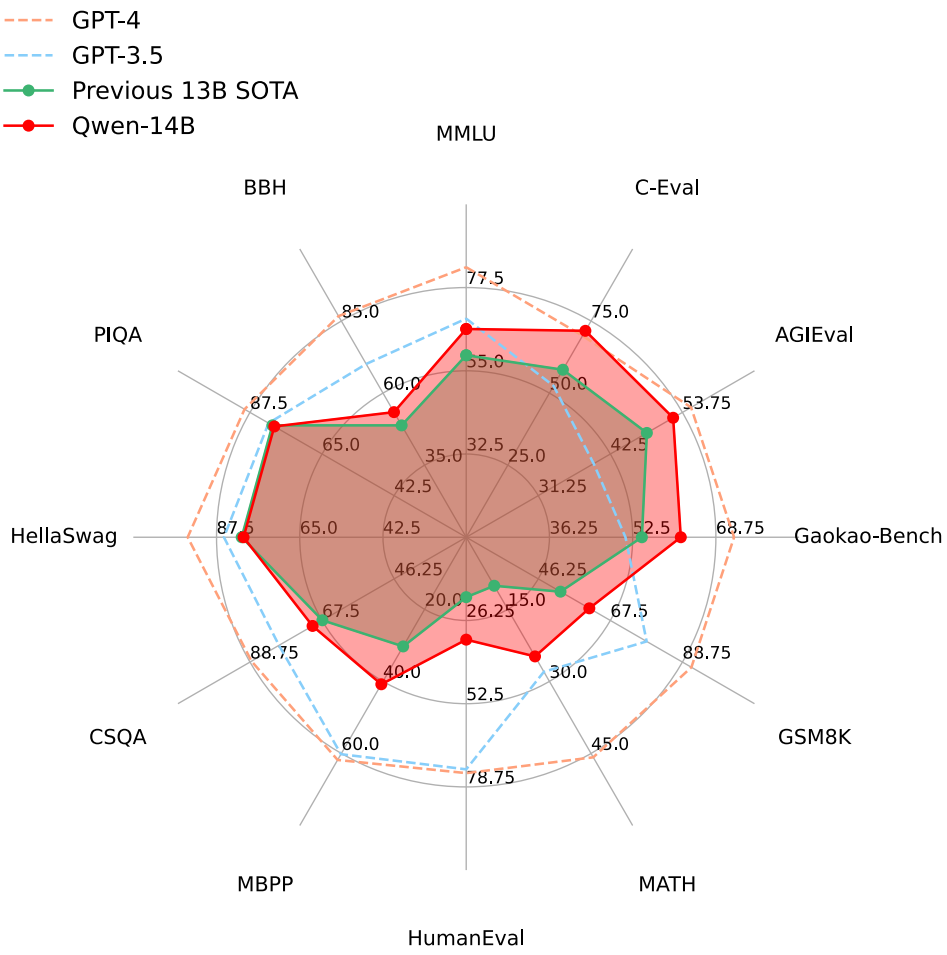

Figure 2: **Performance of GPT-4, GPT-3.5, the previous 13B SOTA, as well as QWEN-14B.** We demonstrate the results on 12 datasets covering multiple domains, including language understanding, knowledge, reasoning, etc. QWEN significantly outperforms the previous SOTA of similar model sizes, but still lag behind both GPT-3.5 and GPT-4.

of types, domains, and tasks. Our dataset is designed to meet these requirements and includes public web documents, encyclopedia, books, codes, etc. Additionally, our dataset is multilingual, with a significant portion of the data being in English and Chinese.

To ensure the quality of our pretraining data, we have developed a comprehensive data preprocessing procedure. For public web data, we extract text from HTML and use language identification tools to determine the language. To increase the diversity of our data, we employ deduplication techniques, including exact-match deduplication after normalization and fuzzy deduplication using MinHash and LSH algorithms. To filter out low-quality data, we employ a combination of rule-based and machine-learning-based methods. Specifically, we use multiple models to score the content, including language models, text-quality scoring models, and models for identifying potentially offensive or inappropriate content. We also manually sample texts from various sources and review them to ensure their quality. To further enhance the quality of our data, we selectively up-sample data from certain sources, to ensure that our models are trained on a diverse range of high-quality content. Finally, we have built a dataset of up to 3 trillion tokens.