DADA-π(PAI) Technology Report: Less Parameters, More Capabilities

Winter Yu[1*] *Rui Cao*[2] *Limin Zhu*[2]

DMS AI[1]

Sichuan DMS Engineering Design Co.Ltd[2]

winter741258@gmail.com

Abstract

Large Language Models (LLMs) like ChatGPT[1] have sparked discussions in the Artificial General Intelligence(AGI) era. Chinese open-source large language models also develop rapidly, such as ChatGLM[2], Baichuan[3], Qwen[4], Deepseek[5], etc. Although several works have made notable strides in several domains, including medicine[7], finance[8], and law[9], the specific domain of process industry remains unexplored. This paper introduces a domain specific language model, DADA-π(DADA for Plant Artificial Intelligence, PAI for short), which is designed for the intricate field of process industry. The continual pre-training process of PAI is meticulously detailed, highlighting the preparation of diverse data types including law, book, design instructions, operation handbooks and paper related to process industry . Following continual pre-training, we utilized supervised fine-tuning (SFT) to ensure the model adheres well to human instructions. Then we utilized reinforcement learning from human feedback (RLHF)[16] to align human preference. Notably, we meticulously construct the first process industry dataset, DADA, during these processes to further improve the processing capability of PAI. Furthermore, via thorough experiment on diverse datasets, we argue that DADA-PAI not only has domain-specific functionalities but also preserves its generality.

1 Introdution

In the age of artificial intelligence, the development of large-scale language models such as GPT4[6] has triggered impressive changes in NLP domain. Chinese open-source large language models also grow rapidly,such as ChatGLM[2], Baichuan[3], Qwen[4], Deepseek[5], etc. Although several works have made notable strides in several domains, including medicine[7], finance[8], and law[9], the specific domain of process industry remains unexplored.

The process industry plays a vital role in the economy by producing a wide range of goods that are essential for daily life. It involves complex processes and requires careful management to ensure the efficient production of high-quality products. This industry typically involves the use of raw materials, energy, and labor to create finished products. The process industry can include various sectors such as chemical, petroleum, nuclear power, and pipeline design. However, mainstream general models often fail to generate domain-specific content with high fidelity and accuracy, despite their general capabilities in various tasks have outperformed human. Therefore, to enhance the capabilities of domain-specific there is an urgent need for a specialized language model tailored to process industry.

This study introduces DADA-π(PAI), a language model specifically designed for process industry. Our approach consists of two steps: first, countinual pretraining a generic model using a new curated process industry dataset, DADA, encompassing law, book, design instructions, operation handbooks and paper related to process industry. Second, implementing an instruction-tuning strategy with a dataset of question-answer pairs, generated using related prompts. This research aims to show that continual pretraining and supervised fine-tuning large language models will improve their performance in specific domains. Furthermore, via thorough experiment on diverse evaluations , we argue that DADA-PAI not only has domain-specific functionalities but also preserves its generality, which will be described in detail in subsequent sections. In light of these efforts, the contributions can be summarized three factors as follow:

- In the pretraining stage, we provide detailed insights into the data process recipes and statistics.
- In the SFT stage, we present our converged approach to data improvement. We established DADA-π(PAI), a fine-tuned LLM focused on the process industry domain. Then we utilized reinforcement learning from human (RLHF) to align human preference.
- We meticulously construct a domain dataset, DADA, and conduct comprehensive experiments on both universal and domain benchmarks to verify the effectiveness of our model.

2 Data Preprocessing

Pretraining is the first and essential stage of developing LLMs, providing them with the capability to learn general linguistic representations from huge text corpora. Prior works have shown that language models can benefit from the knowledge acquired through domain-specific datesets. We meticulously compile a domain-specific corpus to enhance the models with knowledge about process industry.Additionally, we explore a new method to balance general knowledge and domain knowledge. This section shows a comprehensive study of data recipes, including cleaning and deduplication methods. We present the model training, as well as evaluation results on benchmarks.

2.1 Data Collection

- **National Standards.** We collect several National Standards for design and operation related to process industry, such as SH-T 3005.
- **Domain Books and Papers.** We collect textbooks in the field of raw materials and energy, and piping process, which published in the past decade. We also gather some papers related to process indusry, which contribute to model training.

2.2 Data Statistic

We have classified and analyzed the data used for pre-training, with national standards accounting for over 85% of the total. We have also compiled statistics on the tokens in each category. The data statistics are shown in Figure 1.

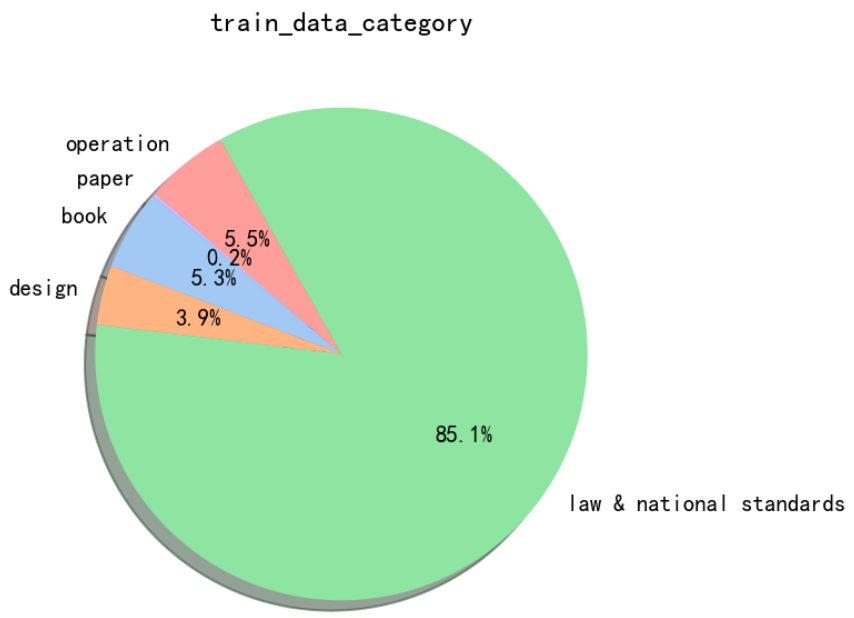

Figure 1 Pretraining Data Category Statistics

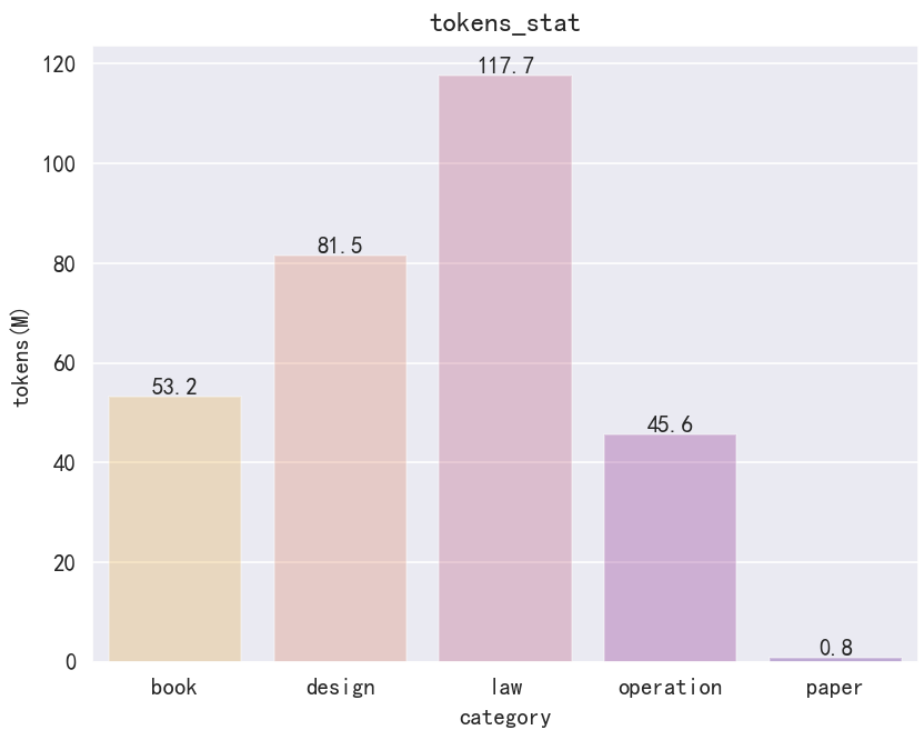

Figure 2 Pretraining Data Token Statistics

2.3 Pre-training Corpus

The pipeline for pretraining data preparation includes: 1) Heuristics rules filter; 2) Minhash deduplication; and 3) Language filtering. All actions performed in this pipeline are document-level filtering, totally removing 5% of the data. Furthermore, all data preprocessing work is accelerated using the Spark framework.

During text extraction, we discard pictures, tables, and URLs, only storing the relevant text. During quality filtering, we ensure that each data is usable through heuristic methods like sensitive word filtering, language filtering, and text length filtering. We aim to minimize the influence of duplicate data on model training by minhash deduplicating at document level. Ultimately, we acquire approximately 300M tokens from the domain corpus. The data statistics are depicted in Figure 2.

2.4 Supervised Fine-Tuning Corpus

In order to enhance the performance of the model in specific fields, we manually construct approximately 8k question-answering pairs data from high-quality process industry books and national standards to assist the model adapting to specific domain knowledge.

3 Methodology

We developed the DADA-(π)PAI model using Alibaba's publicly accessible Qwen2.5 model[4] , which use Transformers with the structure of the decoder only. Despite its relatively modest 7 billion parameters, the PAI model exhibits comparable performance to the much larger model across several NLP benchmarks. This performance enhancement was achieved by diversifying the training data rather than increasing network parameters. Specifically, PAI was continual trained on 300M tokens from our freshly process industry dataset, DADA. We used conversations from 8k instruction data to fine-tune the PAI model in line with Sharegpt training methodology.

3.1 Pretraining Stage

Due to catastrophic forgetting during model training, our goal is to incorporate domain-specific knowledge into the model with minimal loss of generalization ability. Therefore, we utilize LoRA[11] for continual pre-training. Subsequent experiments have demonstrated the effectiveness of our approach, leading us to propose the philosophy of **"less parameters, more capabilities"**. By training the model with a small number of parameters, we aim to enhance its generalization ability and ability to address domain-specific problems.

We use BPE[12] as tokenization and vocabulary size is approximately 152K. During the training phase, we employed the packing technique to accelerate the training speed of the GPU. Additionally, we use RoPE as the same with Qwen[4] and internLM[10]. But We adjust the base frequency (RoPE ABF)[13] to support long context windows up to 200K where the base model itself is trained on 4K context length. To enhance the model's generalization capability, we introduced noise in the embedding layer. Moreover, we utilized the LoRA[11] with rank stabilization scaling factor method to ensure the stability of the training process for LoRA[11].

Further, we follow MIP (Multi-Task Instruction PreTraining)[2] strategy, which is beneficial, given the conditions of the current domain dataset and base models. Owing to the constraints in training resources and time, our training data consist of domain pre-training data and domain instruction data exclusively, with no addition of a general corpus.

We pre-train the foundation model to follow the causal language modelling paradigm. Given the input token sequence x = ($x_0, x_1, x_2, \ldots, x_n$) from the above corpra, the next token autoregressively predicted by minimizing the negative log-likelihood:

$\mathcal{L}_{\mathrm{CPT}}(\theta, \mathcal{D}_{cpt}) = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_{cpt}}\left[-\sum_{i=1}^{n}\log p(x_i \mid x_0, x_1, \ldots, x_{i-1}; \theta)\right]$, where θ is the model parameter and the input context consists of $x_0, x_1, \ldots, x_{i-1}$.

3.2 Supervised Fine-tuning Stage

In the supervised fine-tuning (SFT) stage, we activate the model's domain-specific ability to follow curated instructions by the LoRA[11] Supervised Fine-tuning, using continual pretraining model. Specially, we use a dataset of 8k instruction data instances, which have been screened to ensure their helpfulness and harmlessness. The dataset encompasses a diverse range of topics, including national standards, pipeline process, construction and operation handbooks. The data statistics are depicted in Figure 3.

Given the input instruction x = $(x_0, x_1, x_2, \ldots, x_n)$ and corresponding ground truth y = $(y_0, y_1, y_2, \ldots, y_m)$ from the above fine-tuning dataset, the optimization objective can be formulated as follows:

$\mathcal{L}_{\mathrm{SFT}}(\theta, \mathcal{D}_{sft}) = \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathcal{D}_{sft}}\left[-\sum_{i=1}^{m}\log p(y_i \mid \boldsymbol{x}, y_{<i}; \theta)\right].$

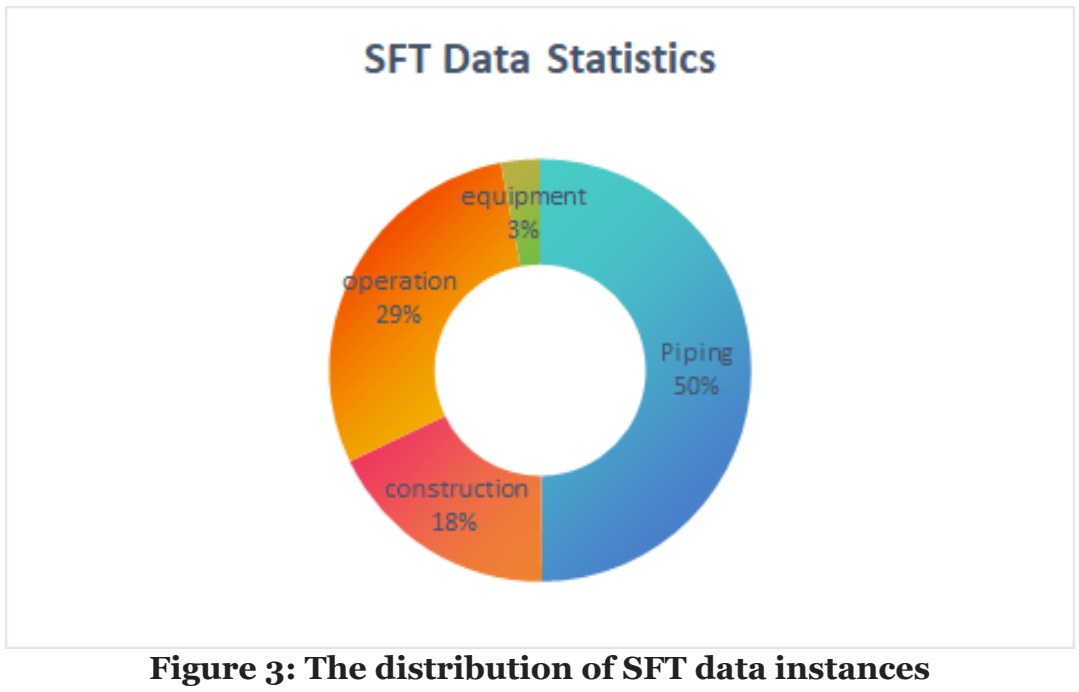

Figure 3: The distribution of SFT data instances

Howerve SFT largely aligns the base models with human preferences, reinforcement learning from human feedback (RLHF)[16] can further help mitigate issues of response rejection, unsafety, and harmfulness generated.

4 Implementation Details

Our DADA-PAI is developed upon the Qwen model with range from7 billion to 72 bilion parameters. The model training is accomplished through the PyTorch platform with Accelerate and DeepSpeed packages using 8 Nvidia RTX4090 GPUs. Various Zero Redundancy Optimizer (ZeRO)[14] strategies are employed to alleviate the memory overhead during overall training. The AdamW optimizer is adopted for network optimization, and the bf16 data accuracy is chosen. To facilitate private deployment, we have adopted various quantization methods such as GPTQ and WAQ for quantization.

5 Evaluation

5.1 Standard Benchmark

In this section, we evaluate the performance of PAI on general benchmarks and domain-specific benchmark(DADA-Eval) against other open-source models, such as Qwen2.5[4]. Evaluation experiments are conducted in zero-shot or few-shot formatting. For fairness in comparison, we use the official implementation of OpenCompass and calculated the average of the results from 5 experiments as the final outcome. The comparative results are showed in Table 1. PAI exhibits a slight decrease in general capability compared to models of similar size, but significantly outperforms all other models in its domain-specific abilities. It achieves top rankings in DADA benchmark assessments, showcasing strong Capabilities in both Chinese and English. These results highlight the model's superior capacity to handle diverse tasks scenarios and within the realms of process industry.

Benchmark	qwen1.5-32B	qwen2.5-32B	PAI-7B	PAI-14B	PAI-32B	PAI-72B	PAI-14B-Int8	PAI-32B-Int4	PAI-72B-Int4
C-Eval	82.50	-	68.71	80.20	88.70	88.39	78.32	83.58	84.30
GSM8k	72.71	95.90	64.90	66.3	68.40	83.51	64.61	67.14	78.52
NaturalQuestions	21.75	-	18.32	20.53	24.26	40.13	19.04	22.89	38.55
HumanEval	70.12	83.50	40.85	50.00	84.32	76.40	45.57	69.23	72.79
DADA-Eval	32.1	32.50	72.13	80.4	***84.1***	74.69	75.36	81.65	74.60

Table 1: Performance comparison of base models on different benchmarks. The best results are in bold.

5.2 Needle-in-the-Haystack

"**Needle-in-the-Haystack**" is a single-needle retrieval task, which is designed to validate the Large Language Models' ability to recall a single piece of key information. This is implemented by inserting a crucial piece of information into a Haystack text of a target length at various positions and then querying the model about this key information at the end of the prompt. This method precisely visualizes models' recall capabilities at different positions within long texts of varying lengths.

We utilize the ChineseDomainModelingEval dataset[15] as the same as internLM, ensuring diversity and quality in the sources of Chinese texts. The results presented in Figure 4 demonstrate PAI's capability for long context modeling.

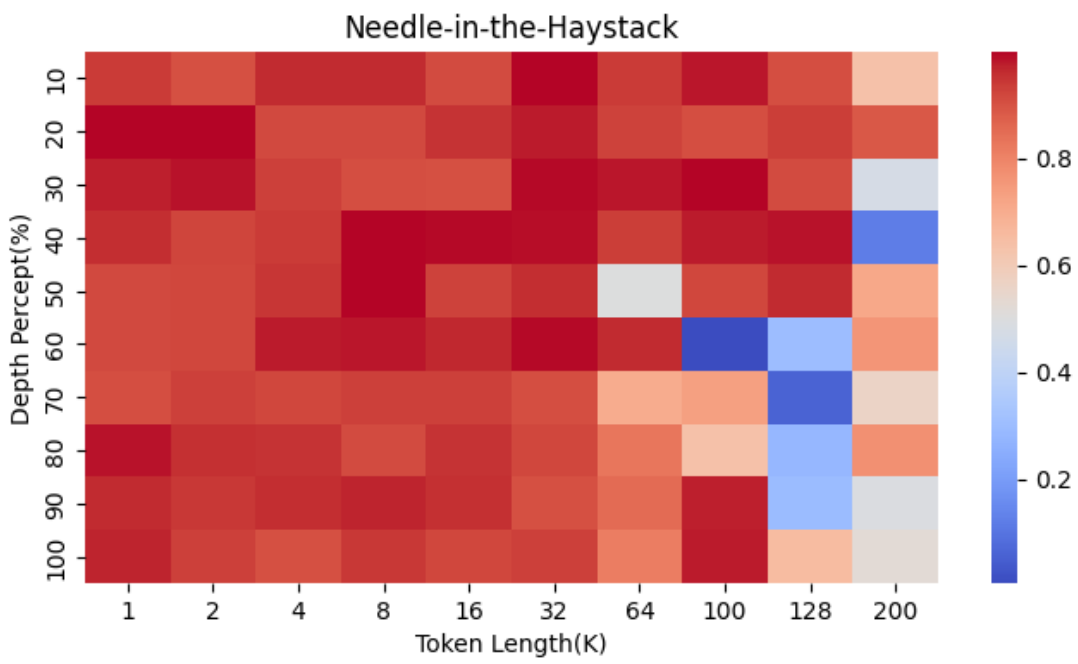

Figure 4: Results on Needle in the Haystack(Chinese).

6 Related Work

The training of a LLM usually includes two stages: pre-training and alignment(supervised fine-tuning and preference optimization). Via pre-training on a large-scale corpus, the LLM can obtain basic generation capabilities and language representation. The alignment stage is to enable the model to have the ability to understand and follow human instructions, and can also improve the generalization ability of the model on OOD(Out Of Distribution) tasks. However, domain-specific tasks often involve complex concepts, technical terminology, and complex relationships between entities. Recently, several works related to PEFT(Parameter Effienct FineTuning) has emerged in the fields of legal, medical, and financial field. Using the retrieval augmented generation(RAG), LLM can be used in professional fields without updating parameters.

7 Conclusion and Future Work

In this technical report, we introduce DADA-π(PAI), a Chinese and English knowledge-enhanced large language model specifically designed for process industry applications , leading us to propose the philosophy of **"less parameters, more capabilities"**. We introduce the process industry pretraining and supervised fine-tuning to incorporate the domain knowledge and enhance the model's performance on downstream tasks, respectively. Our experimental results demonstrate that DADA-PAI not only has domain-specific functionalities but also preserves its generality.

However, our model still exhibits hallucination defects sometimes, and enhancing its professional capabilities remains an ongoing area for improvement. In the future, we are about to conduct further experimental explorations, including the implementation of knowledge graph, long context, multiple agents and reinforcement learning techniques, to enhance DADA-PAI's performance. Our work provides invaluable insights for further research on leveraging AI to enhance process industry domain.

Contributions and Acknowledgments

We thank all those who have contributed to DADA-PAI models including but not limited to the DMS AI team led by Yichen Li, annotation team led by Jiaxi Zhao, AI product manager Binbin Wang , and all those who provide assistance related to model training.

References

[1] OpenAI. 2023a. Chatgpt: Optimizing language models for dialogue. Blog post

[2] Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. Glm-130b: An open bilingual pre-trained model.

[3] Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. Baichuan 2: Open large-scale language models.

[4] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report.

[5] Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, et al. Deepseek llm: Scaling open-source language models with longtermism.

[6] OpenAI. Gpt-4 technical report.

[7] Haochun Wang, Chi Liu, Nuwa Xi, Zewen Qiang, Sendong Zhao, Bing Qin, and Ting Liu. 2023b. Huatuo: Tuning llama model with chinese medical knowledge.

[8] Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. 2023. Fingpt: Open-source financial large language models.

[9] Quzhe Huang, Mingxu Tao, Zhenwei An, Chen Zhang, Cong Jiang, Zhibin Chen,Zirui Wu, and Yansong Feng. 2023a. Lawyer llama technical report.

[10] InternLM2 Technical Report.

[11] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models.

[12] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units.

[13] Xiaoran Liu, Hang Yan, Shuo Zhang, Chenxin An, Xipeng Qiu, and Dahua Lin. Scaling laws of rope-based extrapolation.

[14] Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. Zero: Memory optimizations toward training trillion parameter models.

[15] Tianwen Wei, Liang Zhao, Lichang Zhang, Bo Zhu, Lijie Wang, Haihua Yang, Biye Li, Cheng Cheng, Weiwei Lu, Rui Hu, Chenxia Li, Liu Yang, Xilin Luo, Xuejie Wu, Lunan Liu, Wenjun Cheng, Peng Cheng, Jianhao Zhang, Xiaoyu Zhang, Lei Lin, Xiaokun Wang, Yutuan Ma, Chuanhai Dong, Yanqi Sun, Yifu Chen, Yongyi Peng, Xiaojuan Liang, Shuicheng Yan, Han Fang, and Yahui Zhou. Skywork: A more open bilingual foundation model, 2023.

[16] OpenAI Training language models to follow instructions with human feedback.