XuanYuan 2.0: A Large Chinese Financial Chat Model with Hundreds of Billions Parameters

Xuanyu Zhang, **Qing Yang** and **Dongliang Xu**

Du Xiaoman Financial

Abstract

In recent years, pre-trained language models have undergone rapid development with the emergence of large-scale models. However, there is a lack of open-sourced chat models specifically designed for the Chinese language, especially in the field of Chinese finance, at the scale of hundreds of billions. To address this gap, we introduce **XuanYuan 2.0** (轩辕 2.0), the largest Chinese chat model to date, built upon the BLOOM-176B architecture. Additionally, we propose a novel training method called hybrid-tuning to mitigate catastrophic forgetting. By combining general-domain with domain-specific knowledge and integrating the stages of pre-training and fine-tuning, XuanYuan 2.0 is capable of providing accurate and contextually appropriate responses in the Chinese financial domain.

1 Introduction

In recent years, pre-trained language models have witnessed rapid development. Broadly speaking, they can be categorized into three main architectures: the Encoder architecture represented by BERT (Devlin et al., 2018), the Decoder architecture represented by GPT (Radford et al., 2018), and the Encoder-Decoder architecture represented by T5 (Raffel et al., 2020). Each architecture has its unique characteristics and advantages, catering to different NLP requirements.

The GPT series, with GPT-4 (OpenAI, 2023) being the latest addition, has gained considerable attention due to its remarkable performance in natural language generation tasks, including dialogue generation. The ChatGPT (OpenAI, 2022) model, in particular, has impressed researchers and practitioners with its ability to generate coherent and contextually relevant responses in conversational settings. As a result, the GPT series has become a focal point of research and development in the NLP community.

Moreover, the emergence of large-scale pre-trained models has further fueled the advancements in language modeling. Models such as OPT (Zhang et al., 2022), BLOOM (Scao et al., 2022), and LLaMA (Touvron et al., 2023), with parameter sizes reaching billions, have recently been open-sourced, enabling researchers and developers to explore the potential of these massive models. These models have demonstrated superior performance on various tasks, pushing the boundaries of what is possible in NLP.

While the general-purpose large models mentioned above have garnered significant attention, the importance of domain-specific models cannot be overlooked. In many domains, the distribution of language and the specific linguistic nuances require models that are fine-tuned or specifically trained for that particular domain. Consequently, a range of domain-specific large models has been proposed to cater to the unique needs of various fields. For example, BioBERT (Lee et al., 2020) and PubMedBERT (Gu et al., 2021) are proposed for the biomedical field, and BloombergGPT (Wu et al., 2023) are proposed for financial scenarios. These models have shown promising results in their respective domains, leveraging the domain-specific knowledge learned during pre-training.

Within the Chinese financial domain, there has been considerable progress in the development of pre-trained language models. Researchers have introduced models such as FinBERT (Araci, 2019; Yang et al., 2020; Liu et al., 2021), Mengzi (Zhang et al., 2021), and FinT5 (Lu et al., 2023), which have been tailored for financial text analysis and understanding. These models, though valuable for certain applications, have parameter sizes below one billion, limiting their ability to handle the increasing demands of the Chinese financial NLP landscape. As the volume of financial data and the complexity of language usage continue to grow, there is a pressing need for more powerful models

Model	Type	Parameter	Corpus Content
FinBERT (Araci, 2019)	PLM	110M	News filtered by financial keywords
FinBERT (Yang et al., 2020)	PLM	110M	Corporate Reports, Earnings Call Transcripts, Analyst Reports
Mengzi-BERT-base-fin (Zhang et al., 2021)	PLM	110M	News, Analyse reports, Company announcements
FinT5 (Lu et al., 2023)	PLM	220M, 1B	Corporate Reports, Analyst Reports, Social media and Financial News
XuanYuan 2.0	ChatLM	176B	Corporate Reports, Analyst Reports, Social media and Financial News

Table 1: Comparison of different financial language models.

that can effectively process and understand Chinese financial text.

Despite significant advancements in chat models, there is currently no open-sourced chat model at the scale of hundreds of billions specifically designed for the Chinese language, let alone in the field of Chinese finance. To address this gap, we propose **XuanYuan 2.0** (轩辕 2.0), the largest Chinese chat model to date, based on BLOOM-176B. XuanYuan 2.0 not only surpasses its predecessor, **XuanYuan 1.0** (轩辕 1.0), which achieved first place at the leaderboard of CLUE classification in 2021, but also addresses the need for a large-scale chat model specifically designed for the Chinese financial domain.

Furthermore, domain-specific language models and chat models impose higher requirements on data distribution and training approaches compared to general-domain models. Domain-specific models need to capture the unique linguistic characteristics, terminologies, and contexts of a particular field to achieve optimal performance. However, training these models solely on domain-specific data may lead to catastrophic forgetting, where the model loses previously learned knowledge from the general domain, impacting its overall performance. To mitigate this issue, we propose a novel training method, hybrid-tuning, that combines the stages of pre-training and fine-tuning. By integrating the two stages, our approach guarantees that fine-tuning the model with financial-specific instructions does not impede its general generation capabilities acquired during pre-training. As a result, XuanYuan 2.0 can effectively leverage both its general-domain knowledge and domain-specific financial knowledge to provide accurate and contextually appropriate responses in the Chinese financial domain.

2 Related Work

The advancements in pre-trained language models have led to remarkable progress in various NLP tasks, attracting extensive research efforts. Among the notable contributions, the BERT (Devlin et al., 2018) series stands out as a groundbreaking development in the field of pre-trained models. Following the success of BERT, the GPT (Radford et al., 2018) series emerged as a prominent line of research, focusing on the decoding aspect of language modeling. GPT models, in contrast to BERT's bidirectional approach, leveraged autoregressive language modeling. By training on large amounts of unlabeled text data, GPT models acquired a rich understanding of language and demonstrated impressive capabilities in generating coherent and contextually relevant text. Subsequent iterations of the GPT series, such as GPT-4 (OpenAI, 2023), showcased superior performance in various language generation tasks. And ChatGPT (OpenAI, 2022), an extension of the GPT series, demonstrated the ability to engage in interactive and contextually coherent conversations. This breakthrough sparked considerable interest in developing conversational AI agents capable of simulating human-like dialogue.

In addition to the general-purpose BERT and GPT models, there has been a growing interest in domain-specific pre-training. Researchers have recognized that incorporating domain-specific knowledge during pre-training can lead to substantial performance gains in downstream tasks within those domains. Domain-specific pre-trained models aim to capture domain-specific nuances, enabling them to excel in tasks relevant to the target domain. For instance, in the biomedical domain, BioBERT (Lee et al., 2020) and PubMedBERT (Gu et al., 2021) are proposed to leverage large-scale biomedical

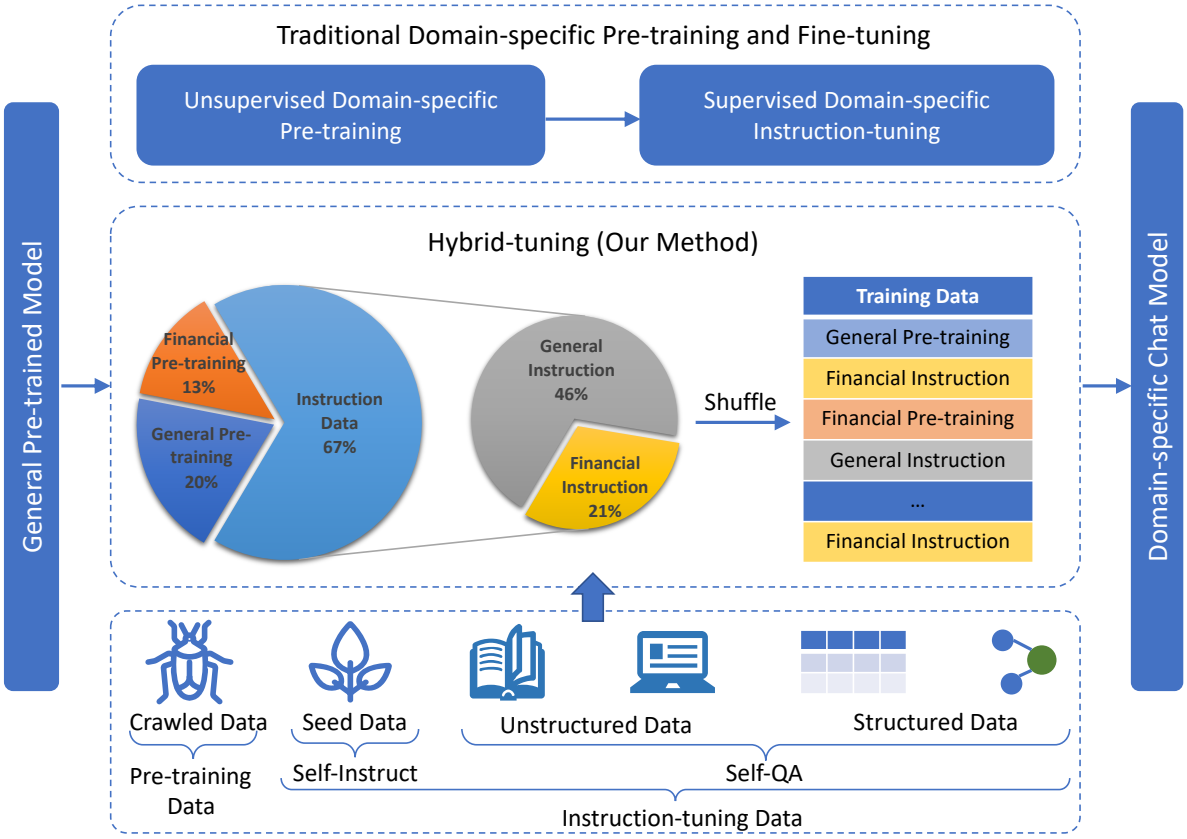

Figure 1: Our proposed hybrid-tuning.

corpora during pre-training. Similarly, in the financial domain, models such as BloombergGPT (Wu et al., 2023) were developed to address the unique challenges and intricacies of the financial text.

Despite the advancements in domain-specific pre-training, the availability of large-scale open-source chat models specifically tailored for the Chinese language and the Chinese financial domain has remained limited. This gap motivates our work in proposing XuanYuan 2.0, a model built upon BLOOM-176B (Scao et al., 2022) with hundreds of billions parameters, to address the unique requirements of the Chinese financial domain and facilitate the development of sophisticated conversational AI systems.

3 XuanYuan 2.0

3.1 Model Architecture

We adopted the original BLOOM (Scao et al., 2022) architecture, which is a decoder-only architecture. The joint probability of tokens in a text can be represented as:

$$p(w) = p(w_1, \ldots, w_T) = \prod_{t=1}^{T} p(w_t | w_{<t}) \quad (1)$$

where w represents a sequence of tokens, w_t is the t^{th} token, and $w_{<t}$ is the sequence of tokens preceding w_t. This method is called autoregressive language modeling, where we predict the probability of the next token in an iterative manner. And following BLOOM, we utilize ALiBi positional embeddings (Press et al., 2021) and embedding LayerNorm (Dettmers et al., 2022) in the traditional decoder structure of Transformer (Vaswani et al., 2017).

3.2 Hybrid-tuning

To alleviate the problem of catastrophic forgetting, we propose a novel domain-specific training framework, hybrid-tuning. In terms of the training stage, it integrates the pre-training stage and instruction fine-tuning stage that are previously split together. In terms of the field of data, it integrates data from both general and financial domains.

As shown in Figure 1, different from traditional two-stage domain-specific training, our proposed hybrid-tuning randomly shuffles pre-training data (general pre-training, financial pre-training) and instruction data (general instruction, financial instruction) into one training data. And all the training

process is done in one stage. In this way, the model can accurately handle instructions in the financial domain, while retaining general conversational capabilities.

For unsupervised pre-training data, we crawl them from the Internet and clean and filter them. For Instruction-tuning data, we use human-written seed instructions to collect general data by Self-Instruct (Wang et al., 2022) and utilize unstructured and structured data in the financial field to gather domain-specific instruction data by Self-QA (Zhang and Yang, 2023). Unstructured financial data comprises a wide range of textual information, such as financial news articles, market reports, analyst commentary, and social media discussions. And structured financial data includes company information and so on. These sources offer valuable insights into market trends, investment strategies, and economic situations.

3.3 Training

To train our complex and computationally intensive model, we employ the powerful NVIDIA A100 80GB GPU and the DeepSpeed (Rasley et al., 2020) distributed training framework. For parallel processing, we primarily rely on pipeline parallelism, which involves distributing the layers of our model across several GPUs. This approach ensures that each GPU only handles a portion of the model's layers, a technique also known as vertical parallelism. Additionally, we adopt the Zero Redundancy Optimizer (Rajbhandari et al., 2020) to enable different processes to store only a portion of the data (parameters, gradients, and optimizer states). Specifically, we use ZeRO stage 1, which means that only the optimizer states are divided using this method. The specific hyperparameters are presented in Table 2.

Hyperparameter	XuanYuan2-7B	XuanYuan2
Architecture hyperparameters		
Parameters	7,069M	176,247M
Layers	30	70
Hidden dim.	4096	14336
Attention heads	32	112
Vocab size	250,680	
Sequence length	2048	
Precision	`float16`	
Activation	`GELU`	
Position emb.	`Alibi`	
Tied emb.	`True`	
Pretraining hyperparameters		
Global Batch Size	512	2048
Learning rate	1.2e-4	6e-5
Total tokens	341B	366B
Min. learning rate	1e-5	6e-6
Warmup tokens	375M	
Decay tokens	410B	
Decay style	`cosine`	
Adam (β_1, β_2)	(0.9, 0.95)	
Weight decay	1e-1	
Gradient clipping	1.0	
Multitask finetuning hyperparameters		
Global Batch Size	2048	2048
Learning rate	2.0e-5	2.0e-5
Total tokens	13B	
Warmup tokens	0	
Decay style	`constant`	
Weight decay	1e-4	

Table 2: Training hyperparameters of XuanYuan 2.0.

4 Experiment

We conducted a comparison between our model and other open-source Chinese conversational models. Simultaneously, we constructed evaluation datasets encompassing various dimensions in both general and financial domains, which were subsequently subject to manual assessment. The results revealed XuanYuan's robust knowledge base and conversational capabilities in the financial domain. Further insights and additional findings will be presented in the next version of the paper after the release of the evaluation rankings.

5 Conclusion

In this paper, we propose the largest Chinese financial chat model, XuanYuan 2.0 (轩辕 2.0), to fill the gap of open-source billion-scale chat models specifically designed for the Chinese financial domain. Besides, we propose a novel training method called hybrid-tuning to mitigate catastrophic forgetting. By combining the general domain with domain-specific knowledge and integrating the stages of pre-training and finetuning, XuanYuan 2.0 achieves the remarkable ability to deliver precise and contextually relevant responses within the Chinese financial domain. We will continue to gather larger-scale Chinese financial domain data in order to further optimize our model.

References

Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022. Llm. int8 (): 8-bit matrix multiplication for transformers at scale. *arXiv preprint arXiv:2208.07339*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Zhuang Liu, Degen Huang, Kaiyu Huang, Zhuang Li, and Jun Zhao. 2021. Finbert: A pre-trained financial language representation model for financial text mining. In *Proceedings of the twenty-ninth international conference on international joint conferences on artificial intelligence*, pages 4513–4519.

Dakuan Lu, Jiaqing Liang, Yipei Xu, Qianyu He, Yipeng Geng, Mengkun Han, Yingsi Xin, Hengkui Wu, and Yanghua Xiao. 2023. Bbt-fin: Comprehensive construction of chinese financial domain pre-trained language model, corpus and benchmark. *arXiv preprint arXiv:2302.09432*.

OpenAI. 2022. Chatgpt.

OpenAI. 2023. Gpt-4 technical report.

Ofir Press, Noah A Smith, and Mike Lewis. 2021. Train short, test long: Attention with linear biases enables input length extrapolation. *arXiv preprint arXiv:2108.12409*.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–16. IEEE.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3505–3506.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. Self-instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560*.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*.

Yi Yang, Mark Christopher Siy Uy, and Allen Huang. 2020. Finbert: A pretrained language model for financial communications. *arXiv preprint arXiv:2006.08097*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Xuanyu Zhang and Qing Yang. 2023. Self-qa: Unsupervised knowledge guided language model alignment. *arXiv preprint*.

Zhuosheng Zhang, Hanqing Zhang, Keming Chen, Yuhang Guo, Jingyun Hua, Yulong Wang, and Ming Zhou. 2021. Mengzi: Towards lightweight yet ingenious pre-trained models for chinese. *arXiv preprint arXiv:2110.06696*.