XuanYuan 2.0：一个拥有数千亿参数的中文金融聊天模型

徐轩宇，杨清，徐东亮

度小满金融

摘要

近年来，预训练语言模型随着大规模模型的出现而迅速发展。然而，在中文领域，尤其是在中文金融领域，缺乏专门设计的、具有数千亿参数规模的开源聊天模型。为了解决这个问题，我们推出了 XuanYuan 2.0 (轩辕 2.0)，这是迄今为止最大的中文聊天模型，建立在 BLOOM-176B 架构之上。此外，我们提出了一种名为混合调优的新型训练方法，以减轻灾难性遗忘。通过将通用领域知识与特定领域知识相结合，并整合预训练和微调阶段，XuanYuan 2.0 能够在中文金融领域提供准确且符合语境的响应。

1 介绍

近年来，预训练语言模型发展迅速。广义上讲，它们可以分为三种主要架构：以 BERT（Devlin et al., 2018）为代表的 Encoder 架构，以 GPT（Radford et al., 2018）为代表的 Decoder 架构，以及以 T5（Raffel et al., 2020）为代表的 Encoder-Decoder 架构。每种架构都有其独特的特点和优势，以满足不同的 NLP 需求。

GPT 系列，以 GPT-4（OpenAI, 2023）为最新成员，因其在自然语言生成任务（包括对话生成）中的卓越表现而备受关注。特别是 ChatGPT（OpenAI, 2022）模型，以其在对话环境中生成连贯且与上下文相关的响应的能力，给研究人员和从业者留下了深刻的印象。因此，GPT 系列已成为 NLP 社区研究和开发的重点。

此外，大规模预训练模型的出现进一步推动了语言建模的进步。诸如 OPT（Zhang et al., 2022）、BLOOM（Scao et al., 2022）和 LLaMA（Touvron et al., 2023）等模型，其参数规模达到数十亿，最近已开源，使研究人员和开发人员能够探索这些大型模型的潜力。这些模型在各种任务中都表现出卓越的性能，突破了 NLP 领域的可能性界限。

虽然上述通用大型模型受到了广泛关注，但特定领域模型的重要性也不容忽视。在许多领域中，语言的分布和特定的语言细微差别需要针对该特定领域进行微调或专门训练的模型。因此，已经提出了一系列特定领域的大型模型，以满足各个领域的独特需求。例如，BioBERT（Lee et al., 2020）和 PubMedBERT（Gu et al., 2021）被提议用于生物医学领域，而 BloombergGPT（Wu et al., 2023）被提议用于金融场景。这些模型在其各自的领域中显示出令人鼓舞的结果，利用了在预训练期间学到的特定领域知识。

在中国金融领域，预训练语言模型的发展取得了相当大的进展。研究人员推出了 FinBERT（Araci, 2019; Yang et al., 2020; Liu et al., 2021）、Mengzi（Zhang et al., 2021）和 FinT5（Lu et al., 2023）等模型，这些模型是为金融文本分析和理解量身定制的。这些模型虽然对于某些应用很有价值，但其参数规模均低于 10 亿，限制了它们处理日益增长的中国金融 NLP 领域需求的能力。随着金融数据的数量和语言使用的复杂性持续增长，迫切需要更强大的模型，能够有效地处理和理解中文金融文本。

模型	类型	参数量	语料内容
FinBERT (Araci, 2019)	PLM	110M	通过金融关键词过滤的新闻
FinBERT (Yang et al., 2020)	PLM	110M	公司报告、盈利电话会议记录、 分析师报告
Mengzi-BERT-base-fin (Zhang et al.,2021)	PLM	110M	新闻、分析报告、公司公告
FinT5 (Lu et al., 2023)	PLM	220M,1B	公司报告、分析师报告、 社交媒体和金融新闻
Xuan Yuan 2.0	ChatLM	176B	公司报告、分析师报告、 社交媒体和金融新闻

表 1：不同金融语言模型的比较。

尽管聊天模型取得了显著进展，但目前还没有专门为中文设计的、具有数千亿参数规模的开源聊天模型，更不用说是在中国金融领域了。为了解决这一差距，我们提出了 XuanYuan 2.0 (轩辕 2.0)，这是迄今为止最大的中文聊天模型，基于 BLOOM-176B。XuanYuan 2.0 不仅超越了其前身 XuanYuan 1.0 (轩辕 1.0)（在 2021 年 CLUE 分类排行榜上名列第一），而且满足了对专门为中文金融领域设计的大规模聊天模型的需求。

此外，与通用领域模型相比，特定领域语言模型和聊天模型对数据分布和训练方法提出了更高的要求。特定领域模型需要捕获特定领域的独特语言特征、术语和上下文，以实现最佳性能。然而，仅在特定领域数据上训练这些模型可能会导致灾难性遗忘，即模型会丢失先前从通用领域学习的知识，从而影响其整体性能。为了缓解这个问题，我们提出了一种新颖的训练方法，即混合调优，它结合了预训练和微调阶段。通过整合这两个阶段，我们的方法保证了使用金融特定指令微调模型不会阻碍其在预训练期间获

得的一般生成能力。因此，XuanYuan 2.0 可以有效地利用其通用领域知识和特定领域金融知识，从而在中国金融领域提供准确且符合语境的响应。

2 相关工作

预训练语言模型的进步已导致各种 NLP 任务取得显著进展，吸引了广泛的研究工作。在众多值得关注的贡献中，BERT（Devlin et al., 2018）系列是预训练模型领域的一项突破性发展。在 BERT 成功之后，GPT（Radford et al., 2018）系列作为一条重要的研究方向出现，专注于语言建模的解码方面。与 BERT 的双向方法相比，GPT 模型利用了自回归语言建模。通过在大量未标记的文本数据上进行训练，GPT 模型获得了对语言的丰富理解，并在生成连贯且与上下文相关的文本方面表现出令人印象深刻的能力。GPT 系列的后续迭代，例如 GPT-4（OpenAI, 2023），在各种语言生成任务中表现出卓越的性能。ChatGPT（OpenAI, 2022）是 GPT 系列的扩展，展示了参与交互式和上下文连贯对话的能力。这一突破激发了人们对开发能够模拟类人对话的会话 AI 代理的极大兴趣。

除了通用的 BERT 和 GPT 模型之外，人们对特定领域的预训练越来越感兴趣。研究人员已经认识到，在预训练期间结合特定领域的知识可以显着提高这些领域内下游任务的性能。特定领域的预训练模型旨在捕获特定领域的细微差别，使它们能够在与目标领域相关的任务中表现出色。例如，在生物医学领域，提出了 BioBERT（Lee et al., 2020）和 PubMedBERT（Gu et al., 2021）以在预训练期间利用大规模生物医学语料库。同样，在金融领域，开发了诸如 BloombergGPT（Wu et al., 2023）之类的模型，以解决金融文本的独特挑战和复杂性。

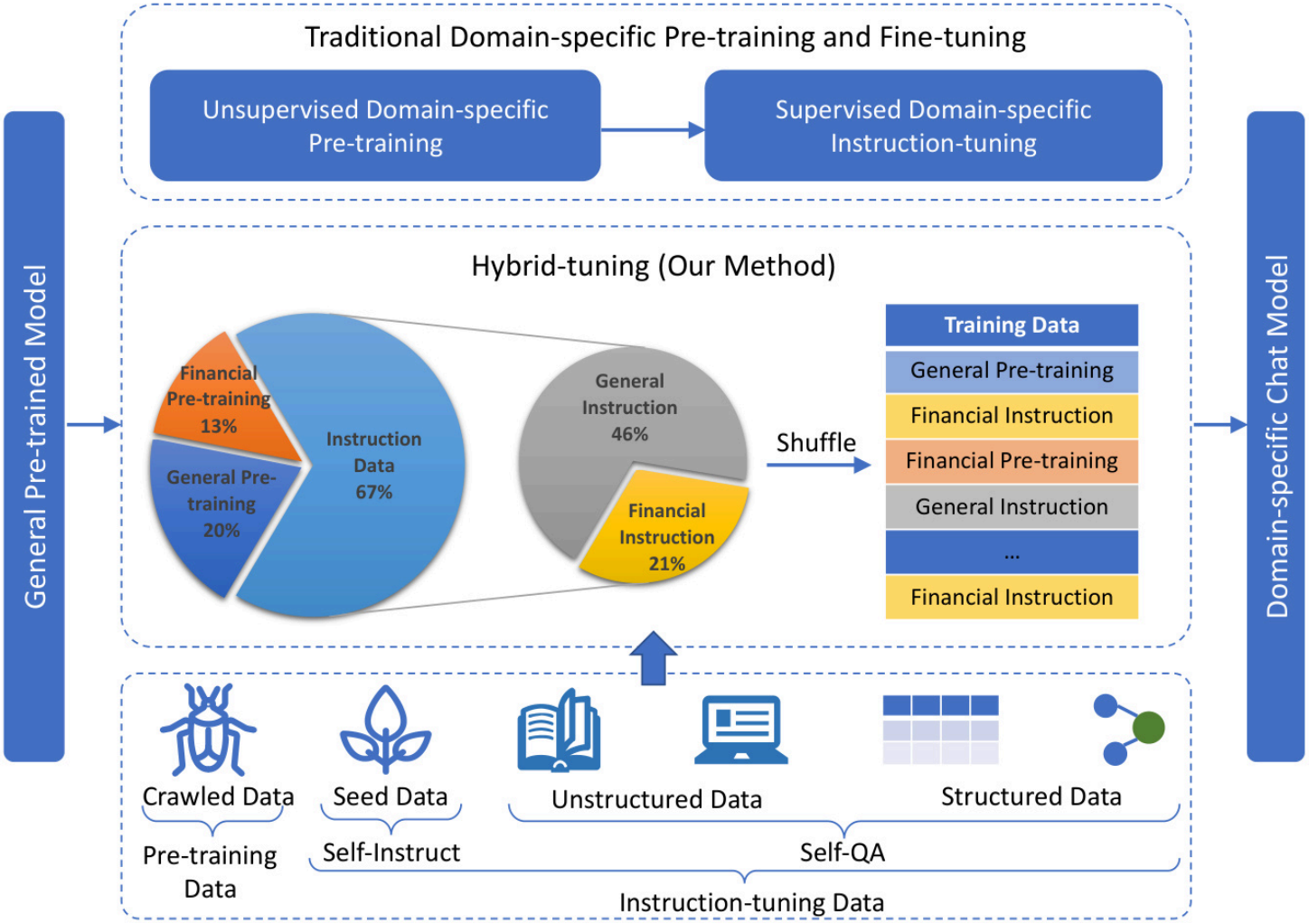

图 1：我们提出的混合调优。

尽管在特定领域的预训练方面取得了进步，但专门为中文和中国金融领域量身定制的大规模开源聊天模型的可用性仍然有限。这一差距促使我们提出了 XuanYuan 2.0，该模型建立在具有数千亿参数的 BLOOM-176B（Scao et al., 2022）之上，以满足中国金融领域的独特需求并促进复杂会话 AI 系统的开发。

3 XuanYuan 2.0

3.1 模型架构

我们采用了原始的 BLOOM（Scao et al., 2022）架构，它是一种仅解码器架构。文本中 tokens 的联合概率可以表示为：

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

其中 w 表示一个 tokens 序列，w_t 是第 t 个 token，$w_{<t}$ 是 w_t 之前的 tokens 序列。这种方法称为自回归语言建模，我们以迭代方式预测下一个 token 的概率。并且在 BLOOM 之后，我们在 Transformer 的传统解码器结构中使用 ALiBi 位置嵌入（Press et al., 2021）和嵌入 LayerNorm（Dettmers et al., 2022）（Vaswani et al., 2017）。

3.2 混合调优

为了缓解灾难性遗忘的问题，我们提出了一种新颖的特定领域训练框架，即混合调优。在训练阶段，它集成了先前分开的预训练阶段和指令微调阶段。在数据领域方面，它集成了来自通用领域和金融领域的数据。

如图 1 所示，与传统的两阶段特定领域训练不同，我们提出的混合调优将预训练数据（通用预训练、金融预训练）和指令数据（通用指令、金融指令）随机混洗到一个训练数据中。并且所有的训练过程都在一个阶段完成。这样，模型可以准确地处理金融领域的指令，同时保留一般的对话能力。

对于无监督预训练数据，我们从互联网上抓取它们，并进行清理和过滤。对于指令调优数据，我们使用人工编写的种子指令，通过 SelfInstruct（Wang et al., 2022）收集通用数据，并利用金融领域中的非结构化和结构化数据，通过 Self-QA（Zhang and Yang, 2023）收集特定领域的指令数据。非结构化金融数据包括范围广泛的文本信息，例如金融新闻文章、市场报告、分析师评论和社交媒体讨论。结构化金融数据包括公司信息等。这些来源提供了对市场趋势、投资策略和经济形势的宝贵见解。

3.3 训练

为了训练我们复杂且计算密集型的模型，我们采用了强大的 NVIDIA A100 80GB GPU 和 DeepSpeed（Rasley et al., 2020）分布式训练框架。对于并行处理，我们主要依赖于流水线并行性，这涉及将我们模型的层分布在多个 GPU 上。这种方法确保每个 GPU 仅处理模型层的一部分，这种技术也称为垂直并行性。此外，我们采用 Zero Redundancy Optimizer（Rajbhandari et al., 2020）来使不同的进程仅存储一部分数据（参数、梯度和优化器状态）。具体来说，我们使用 ZeRO stage 1，这意味着只有优化器状态使用此方法进行划分。具体的超参数如表 2 所示。

4 实验

我们对我们的模型和其他开源中文会话模型进行了比较。同时，我们构建了包含通用领域和金融领域各个维度的评估数据集，随后对其进行了人工评估。结果表明，XuanYuan 在金融领域具有强大的知识库和会话能力。在发布评估排名后，将在本文的下一版本中介绍更多见解和其他发现。

表 2：XuanYuan 2.0 的训练超参数。

超参数	Xuan Yuan2-7B	Xuan Yuan2
架构超参数		
参数量 层数 隐藏层维度 注意力头数 词汇表大小	7,069M 30 4096 32 250,680	176,247M 70 14336 112
序列长度 精度 激活函数 位置嵌入 Tied 嵌入	2048 float16 GELU Alibi	
全局 BatchSize 学习率 总 tokens 数	True 预训练超参数 512	2048
	1.2e-4 341B	6e-5 366B
最小学习率 Warmup tokens 数	1e-5 375M	6e-6
梯度裁剪	(0.9, 0.95)	
衰减 tokens 数 衰减方式 Adam (β1, β2) 权重衰减	410B cosine	

5 结论

在本文中，我们提出了最大的中文金融聊天模型 XuanYuan 2.0 (轩辕 2.0)，以填补专门为中国金融领域设计的开源十亿级聊天模型的空白。此外，我们提出了一种名为混合调优的新型训练方法，以减轻灾难性遗忘。通过将通用领域知识与特定领域知识相结合，并整合预训练和微调阶段，XuanYuan 2.0 实现了在中文金融领域内提供精确且与上下文相关的响应的卓越能力。我们将继续收集更大规模的中文金融领域数据，以便进一步优化我们的模型。

参考文献

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. Domainspecific 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 Ilic, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176bparameter 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.