构建合成指令数据集以提升领域特定大语言模型的推理能力：日本金融领域案例研究

Yuma Okochi[1], Fabio Milentiansen Sim[2],Tomoyasu Okada[1]

[1]野村综合研究所， [2]NRI Indonesia**通信作者：**

t3-okada@nri.co.jp

arXiv:2603.01353v1 [cs.LG] 2 Mar 2026

摘要

在将大语言模型适配到特定领域时，同时实现领域专业知识与推理能力仍是一项紧迫挑战。本研究提出一种通用方法，从领域特定词汇出发，为任意领域构建高质量合成指令数据。作为示范，我们将该方法应用于金融领域，构建了包含思维链推理轨迹、总计约95亿词元的大规模指令数据集。评估结果证实，相较于基线模型，本方法在金融基准测试上实现了性能提升，从而证明了其有效性。我们还报告了推理轨迹长度对性能的影响及其局限性。最后，我们在HuggingFace上开源了模型与数据集。

1 引言

随着大语言模型（LLM）在社会中的应用日益广泛，如何将其适配到金融和法律等需要高度专业知识的领域，已成为一项活跃的研究课题。传统上，持续预训练被广泛用于适配这些领域，并在获取领域知识方面取得了显著成果（Wu 等人，2023年）。此外，近期研究重点已从单纯获取知识转向应对需要高级逻辑推理的实际任务，例如复杂事件分析和未来预测（如欺诈检测和收益预测（Sugiura 等人，2025年））。

近年来，能够在生成答案前先生成思维链（CoT）推理轨迹，从而提升性能的推理模型受到了关注（OpenAI 等人，2024 年）。尽管采用推理轨迹进行监督微调（SFT）有助于获得这一能力，且已有通用数据集（Nathawani 等人，2025年；Bercovich 等人，2025年），但构建既具备大规模、领域特定性（尤其是日语）且包含推理轨迹的数据的方法，尚未得到充分确立。

因此，本研究提出了一套数据构建流程，旨在获取特定领域的推理能力，并作出如下贡献：

- 我们建立了一种方法，从特定领域的主题词出发，使用推理大语言模型合成并筛选带有推理轨迹的指令数据。该方法也适用于金融领域以外的多种领域。

- 我们将所提出的方法应用于日本金融领域，构建了一个总计约95亿词元的数据集。在该数据集上进行训练后，模型在金融基准测试中的表现超过了官方指令微调模型。

- 我们分析了受控推理轨迹长度对领域特定任务性能的影响，并讨论了其局限性。

2 相关工作

2.1 领域适应与指令数据集

使用领域专属数据集进行指令微调，作为让大语言模型适应特定领域的一种方法，已得到广泛研究。在金融领域，Ke 等人(2025年)和Lee 等人(2024 年)提出了大规模指令数据集，但这些数据集主要聚焦于英语和中文。针对日语，Tanabe 等人(2024 年)提出了一个金融领域专属指令数据集，但其主要采用不包含推理轨迹（思维链）的格式。本研究采用合成数据生成技术，尝试弥合“领域知识”与“推理轨迹”之间的差距，而这一问题此前

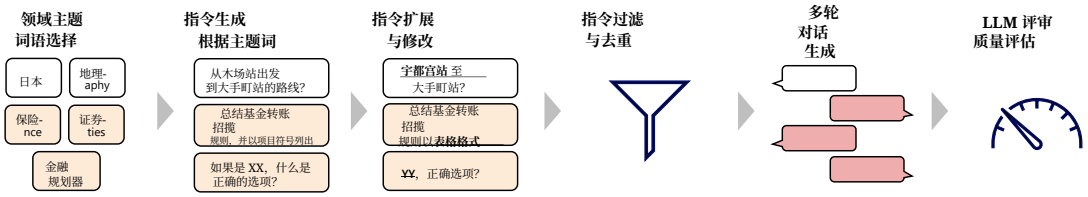


图 1：所提出的数据集构建流程

尚未得到解决。我们采用基于主题词的生成方法，提出了一个可扩展至任意领域的框架，这是本研究的一大特色。

2.2 推理轨迹长度的验证

对于复杂任务，在输出答案之前生成推理轨迹的方法是有效的，并且已经提出了许多会输出推理轨迹的推理模型（OpenAI等人，2024 年；DeepSeek-AI等人，2025年；Guo等人，2025年；Yang 等人，2025年）。先前关于数学基准测试的研究报告指出，随着推理轨迹长度的增加，性能有所提升（Muennighoff等人，2025年），同时也有研究指出，超过一定长度后性能可能会下降Ghosal等人（2025年）。这些研究主要局限于数学任务，而对于需要专业知识与监管理解的领域特定任务，相关验证尚未充分开展。本研究利用金融基准测试验证了推理轨迹长度与性能之间的关系。

3 数据集构建方法

本研究提出的数据集构建流程如图1所示。该方法扩展了 Nemotron-4 340B（Nvidia 等，2024 年）所采用的合成方法，并针对高效学习领域专属知识和推理能力进行了优化。尽管该方法是一个与领域无关的通用流程，但本实验以金融领域为目标，验证其有效性。除非另有说明，合成过程均使用 OpenAI 的 gpt-oss-120b（OpenAI 等，2025年）作为大语言模型。

此外，附录中汇总了该流程各步骤的详细参数设置，包括生成数量、筛选阈值及其他规范。A。

3.1 领域主题词语选择

该方法的第一阶段是选择能够全面覆盖目标领域的主题词。我们选择了金融领域中的行业类别、金融产品和相关技术等种子词。此外，为了减轻领域专门化导致的通用能力退化（灾难性遗忘），我们还混入了来自非目标领域的种子词，并为每个词生成相关的子主题词。

3.2 根据主题词生成指令

接下来，我们让大语言模型针对生成的每个子主题词生成相关的用户提示（指令）。为确保任务多样性，我们定义了四种类型：开放式问题、数学推理（计算题）、创意写作和选择题。对于每种类型，我们都生成了足够数量的指令，以确保多样性。

3.3 指令扩展与修改

为了扩展用户提示的多样性，我们对上一阶段生成的指令进行了扩展与修改。具体而言，我们添加了上下文，将指令转换为不同的格式和风格，针对特定情况对其进行详细阐述，并改写指令以涵盖相关主题。我们允许大语言模型选择修改类型，并根据各任务类型中的指令数量按比例生成变体。

3.4 指令过滤与去重

为确保生成数据的质量，我们按照 Swallow Corpus v2（Hattori 等人，2025年）的方法，基于 n元语法和词数进行筛选，并使用 MinHash 和 LSH（模糊去重）进行去重。对于n元语法筛选，我们采用了与先前研究相同的参数；对于词数筛选，我们使用 MeCab（Kudo 等人，2004年）进行分词，并移除了极短的指令。

3.5 多轮对话生成

利用上一阶段筛选得到的指令，我们生成了多轮对话，以提升交互式任务推理能力。第一轮直接输入指令；后续轮次则提供此前的对话内容，并交替生成包含推理轨迹的用户提示和大语言模型回复。

3.6 通过 LLM 评审进行质量过滤

最后，为了确定生成的数据是否准确反映了领域知识并包含逻辑推理，我们使用 LLM 评审进行了质量过滤。我们使用 gpt-oss-120b 进行评审，并根据多个评估标准的平均分排除了低质量数据。评估方法和阈值设置的详情见附录B。

3.7 公开数据集的使用

除了领域特定能力外，我们还利用公开指令数据集来强化基础数学能力和指令遵循能力。数学方面，我们使用了 NuminaMath-CoT（LI 等人，2024 年）和 Nemotron-Post-Training-Dataset-v1（Nathawani 等人，2025年；Bercovich 等人，2025年），指令遵循方面则使用了 smol-constraints（Allal 等人，2025年）。我们仅提取这些数据集中的问题部分，并按照图1所示，从筛选/去重步骤开始进行处理。请注意，这些数据仅限于单轮对话。

最终数据集包含约 144 万个样本数，总计约 95 亿个词元。数据集详情见附录C。

4 训练与评估

我们使用构建的数据集对大语言模型进行了训练和评估，以验证所提出的方法在目标领域的有效性。

4.1 金融语料上的持续预训练

在大语言模型领域适应中，通常使用特定领域语料库对预训练模型进行持续预训练（持续预训练）（Shi 等人,2025年）。在本研究中，我们对 CommonCrawl 等公开语料库应用了金融领域分类和质量过滤，并构建了日语金融语料库。利用该语料库，我们进行了对 Qwen3-14B-Base 的持续预训练（Yang 等人, 2025年）和 gpt-oss-20b（OpenAI 等人, 2025年）。

4.2 通过监督微调进行指令微调

为验证本研究构建的数据集的有效性，我们对采用的基础模型以及第4.1节构建的持续预训练模型进行了监督微调。训练轮次设为 2，并在一台 AWS p5en.48xlarge 实例（NVIDIA H200 Tensor Core GPU × 8）上进行训练。每个模型约需 240 小时的训练时间。

4.3 金融基准测试评估 s

我们使用 japanese-lm-fin-harness（Hirano、2024 年）和 pfmt-bench-fin-ja（Hirano 和 Imajo、2025年）评估了训练模型在金融领域的能力。japanese-lm-fin-harness 是一个基准测试，涵盖证券报告的情感分析和金融资格考试题目，我们使用该基准来验证金融知识获取的程度。

pfmt-bench-fin-ja 是一个基准测试，用于比较模型能否在金融领域针对写作、信息抽取和创意生成等任务，在多轮交互中恰当地响应用户查询；我们使用该基准来验证金融对话能力。

各基准测试的评估方法均根据公共代码仓库中的实现进行了调整，具体细节见附录 D。

各基准测试的评估结果如表 1 所示。仅进行监督微调时，Qwen3-14B-Base 模型的得分低于官方指令微调模型。进一步分析表明，得分较低主要是由于答案提取失败所致，例如重复输出或不遵守指定的答案格式。然而，在最终结果所对应的训练过程中，可以观察到，随着训练步数增加，这些错误呈减少趋势，因此预计继续训练将带来性能提升。

在 CPT 模型上使用我们的数据集进行监督微调（SFT）后，两项基准测试的平均分数均有所提升。gpt-oss-20b 在所有子任务上的表现都超过了官方指令微调模型，Qwen3-14B 也呈现出类似趋势，但部分任务除外。这表明，所提出的数据构建方法能够有效衔接领域知识注入（CPT）与推理能力

模型	CoT	japanese-lm-fin-harness						pfmt-bench-fin-ja		
		Avg.	chabsa	cma	cpa	fp2	ss1	Avg.	第1轮	第2轮
Qwen3-14B	No	66.50	91.73	87.58	41.83	47.03	64.31	7.781	7.819	7.743
Qwen3-14B	Yes	71.04	**91.96**	**93.26**	**49.37**	53.37	67.22	8.104	8.211	7.997
Qwen3-14B-Base ＋监督微调（我们的方法）	Yes	70.69	91.48	91.20	47.96	55.00	**67.81**	8.415	8.472	8.358
Qwen3-14B-Base ＋持续预训练 ＋监督微调（我们的方法）	Yes	**71.78**	91.62	91.45	48.59	**60.00**	67.27	**8.455**	**8.514**	**8.395**
gpt-oss-20b	No	61.17	91.09	81.33	33.45	41.82	58.17	7.480	7.425	7.534
gpt-oss-20b	Yes	66.93	91.80	90.46	38.51	49.74	64.15	7.883	7.858	7.908
gpt-oss-20b ＋监督微调（本文）	Yes	69.56	91.87	90.87	43.65	52.63	**68.80**	**8.432**	**8.439**	**8.424**
gpt-oss-20b ＋持续预训练 ＋监督微调（本文）	Yes	**72.50**	**91.89**	**94.24**	**45.51**	**62.71**	68.15	8.209	7.992	8.425

表1：各模型在金融基准测试上的评估结果（思维链 = 推理轨迹是否存在）

获取（SFT），从而最大限度地提升特定领域的任务表现。

我们还比较了基线模型在有推理和无推理情况下的性能。我们发现，在所有模型中，使用推理轨迹进行推断使 japanese-lm-fin-harness 上的性能提升了约 4.5–5.7 分，使 pfmt-bench-fin-ja 上的性能提升了约 0.4 分，从而证实了启用推理的推断在金融领域基准测试上的有效性。

4.4 通过推理轨迹长度验证性能提升

从第4.3节的评估结果来看，已确认输出推理轨迹有助于提升金融任务上的性能。为验证推理轨迹长度对模型性能的影响，我们将推理轨迹长度固定为特定值，并分阶段改变这些固定值。

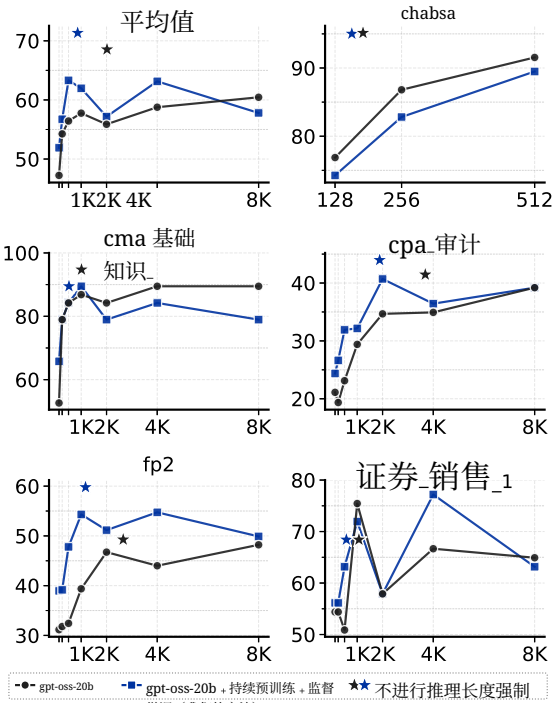


图2：任务准确率随推理轨迹长度的变化

按照以往研究（Muennighoff等人，2025年），我们采用以下流程，将推理轨迹长度固定为指定值：

结束位置早于指定长度时

移除推理结束标记（例如，</think>），并在末尾添加“等等”，以继续推理。

结束位置达到或超过指定长度时

删除超出指定长度的词元，插入推理结束标记和用于提示最终答案的文本（换行）“最终答案：”，然后重新生成最终答案。

目标基准测试为japanese-lm-fin-harness。对于除chabsa以外的四个子任务，我们设置了7个固定长度级别：128、256、512、1024、2048、4096和8192个词元。对于chabsa，由于执行时间限制，我们将实验限制为3

长度级别：128、256 和 512 个词元。

各项任务的执行结果如图所示2。除 chabsa 外，各子任务的性能均随着固定长度延长至 1024 个词元而有所提升。对于 chabsa，在验证范围内延长至 512 个词元时也观察到了类似趋势。然而，将固定长度设为 2048 个词元或更长时，并未确认性能有显著提升（收益递减）。对推理轨迹的分析揭示了以下行为因素：

推理提前终止 即使插入“等等，”，模型也会紧接着立即得出结论，例如“等等，**但是**答案是 D。因此，最终答案是 D。”，试图在没有进行更深入推理的情况下结束推理过程。

输出循环 尤其是在 4096 个词元或更多时，输出会不断重复确认结论，直到达到固定长度，导致没有进行实质性的推理。

此外，与未强制终止推理轨迹的推理相比，在所有任务中，强制终止推理轨迹均导致性能下降数个百分点（图中以⋆ 标记，见2）。即使推理轨迹长度大致相同，性能仍会下降，这表明终止推理轨迹的方法本身可能会影响性能。例如，可以考虑采用自然终止语句的方法，如 Qwen3 中所验证的做法（ Yang 等人，2025年），例如“考虑到用户时间有限，我现在必须直接根据思考给出解决方案\n</think>\n\n”，但未来还需研究包括其他终止方法在内的各种应对措施。

5 结论

本研究提出了一个通用框架，用于构建旨在提升特定领域大语言模型推理能力的合成指令数据集。为展示所提方法，我们构建了一个总计约 95 亿词元的数据集，其中包含针对日本金融领域的推理轨迹；评估结果表明，该数据集的表现优于官方指令微调模型。这说明，从主题词出发生成合成数据，对于领域适应极为有效。此外，对领域特定任务中的推理轨迹长度进行分析后发现，性能会随着长度增加而提升，直至约 1024 个词元，此后则趋于平稳。该方法很容易推广至金融以外的其他专业性较强的领域；我们希望这一方法能够成为未来构建领域专用大语言模型的基础。

致谢

本研究得到经济产业省（METI）和新能源与产业技术综合开发机构（NEDO）通过“面向后5G信息通信系统的强化基础设施研发项目”的支持，

该项目由生成式人工智能加速器挑战赛（GENIAC）资助。

我们还要感谢斯坦福大学的 TatsunoriHashimoto 助理教授，感谢他在研究构想形成过程中给予的宝贵建议。

参考文献

Loubna Ben Allal、Anton Lozhkov、Elie Bak- ouch、Gabriel Martín Blázquez、Guilherme Penedo、Lewis Tunstall、Andrés Marafioti、Hynek Kydlí˘cek、Agustín Piqueres Lajarín、Vaibhav Srivastav、Joshua Lochner、Caleb Fahlgren、Xuan-Son Nguyen、Clé- mentine Fourrier、Ben Burtenshaw、Hugo Larcher、Haojun Zhao、Cyril Zakka、Mathieu Morlon 及另外 3 位作者。2025年。 SmolLM2：当 Smol 变大——以数据为中心的训练小语言模型。预印本，arXiv:2502.02737。

Akhiad Bercovich、Itay Levy、Izik Golan、Mohammad Dabbah、Ran El-Yaniv、Omri Puny、Ido Galil、Zach Moshe、Tomer Ronen、Najeeb Nabwani、Ido Sha- haf、Oren Tropp、Ehud Karpas、Ran Zilberstein、Jiaqi Zeng、Soumye Singhal、Alexander Bukharin、Yian Zhang、Tugrul Konuk 及另外 117 位作者。2025年。 Llama-Nemotron：高效推理模型。 预印本，arXiv:2505.00949。

DeepSeek-AI、Aixin Liu、Bei Feng、Bing Xue、Bingx- uan Wang、Bochao Wu、Chengda Lu、Chenggang Zhao、Chengqi Deng、Chenyu Zhang、Chong Ruan、Damai Dai、Daya Guo、Dejian Yang、Deli Chen、Dongjie Ji、Erhang Li、Fangyun Lin、Fucong Dai及其他181位作者。2025年。 DeepSeek-V3 技术报告。预印本，arXiv:2412.19437。

Soumya Suvra Ghosal、Souradip Chakraborty、Avinash Reddy、Yifu Lu、Mengdi Wang、Dinesh Manocha、Furong Huang、Mohammad Ghavamzadeh及Am- rit Singh Bedi。2025年。多思考总是有帮助吗？推理模型测试时扩展的幻象。载于第三十九届神经信息处理系统年会（*NeurIPS 2025*）。NeurIPS。

Daya Guo、Dejian Yang、Haowei Zhang、Junxiao Song、Peiyi Wang、Qihao Zhu、Runxin Xu、Ruoyu Zhang、Shirong Ma、Xiao Bi、Xiaokang Zhang、Xingkai Yu、Yu Wu、Z. F. Wu、Zhibin Gou、Zhihong Shao、Zhuoshu Li、Ziyi Gao、Aixin Liu 等 175位作者。2025年 DeepSeek-R1 通过强化学习激励大语言模型进行推理。 《自然》，645(8081):633–638。

Kakeru Hattori、Naoaki Okazaki、Sakae Mizuki、Kazuki Fujii、Taishi Nakamura、Masanari Oi、Taihei Shiotani、Koshiro Saito、Youmi Ma、Koki Maeda、Takumi Okamoto、Shigeki Ishida、Rio Yokota 和 Hiroya Takamura。2025年。Swallow Corpus v2：构建教育日语网络语料库。载于《第*31*届自然语言处理学会年会论文集（*NLP2025*）》。

Masanori Hirano。2024 年。面向大语言模型的日语金融基准构建。载于第七届金融科技与自然语言处理联合研讨会、第五届金融服务非结构化数据知识发现研讨会及第四届经济学与自然语言处理研讨会论文集，第 1–9 页，都灵，意大利。计算语言学协会。

Masanori Hirano 和 Kentaro Imajo。2025年。Pfmt-bench-fin-ja：日语金融偏好多轮基准。载于第*31*届自然语言处理年会论文集（*NLP2025*）。

Zixuan Ke、Yifei Ming、Xuan-Phi Nguyen、Caiming Xiong 和 Shafiq Joty。2025年。揭秘金融大语言模型的领域自适应后训练。载于*2025*年自然语言处理实证方法会议论文集，第31033–31059页，中国苏州。计算语言学协会。

Taku Kudo、Kaoru Yamamoto 和 Yuji Matsumoto。2004年。将条件随机场应用于日语形态素分析。载于*2004*年自然语言处理实证方法会议论文集，第230–237页，西班牙巴塞罗那。计算语言学协会。

Sangmin Lee、Suzie Oh、Saeran Park、Guijin Son 和 Pilsung Kang。2024 年。FINALE：通过思维链提示生成高质量推理过程的金融领域指令微调数据集。载于第八届金融科技与自然语言处理会议暨首届智能体*AI*情景规划会议论文集，第 89–106 页，韩国济州。

Jia LI、Edward Beeching、Lewis Tunstall、Ben Lipkin、Roman Soletskyi、Shengyi Costa Huang、Kashif Rasul、Longhui Yu、Albert Jiang、Ziju Shen、Zihan Qin、Bin Dong、Li Zhou、Yann Fleureau、Guillaume Lample 和 Stanislas Polu。2024 年。NuminaMath。

Niklas Muennighoff、Zitong Yang、Weijia Shi、Xiang Lisa Li、李飞飞、Hannaneh Hajishirzi、Luke Zettlemoyer、Percy Liang、Emmanuel Candes 和 Tatsunori Hashimoto。2025年。S1：简单的测试时扩展。载于*2025*年自然语言处理实证方法会议论文集，第 20275–20321 页，中国苏州。计算语言学协会。

Dhruv Nathawani、Igor Gitman、Somshubra Majumdar、Evelina Bakhturina、Ameya SunilMahabaleshwarkar、Jian 和 Zhang，以及 Jane Polak Scowcroft。2025年。Nemotron-Post-Training-Dataset-v1。Nvidia。

Nvidia、Bo Adler、Niket Agarwal、Ashwath Aithal、Dong H. Anh、Pallab Bhattacharya、Annika Brun- dyn、Jared Casper、Bryan Catanzaro、Sharon Clay、Jonathan Cohen、Sirshak Das、Ayush Dattagupta、Olivier Delalleau、LeonDerczynski、Yi Dong、Daniel Egert、Ellie Evans、Aleksander Ficek，以及另外63位作者。2024 年。Nemotron-4 340B 技术报告。预印本，arXiv:2406.11704。

OpenAI、Sandhini Agarwal、Lama Ahmad、Jason Ai、Sam Altman、Andy Applebaum、Edwin Arbus、Rahul K. Arora、Yu Bai、BowenBaker、Haiming Bao、Boaz Barak、Ally Bennett、Tyler Bertao、Nivedita Brett、Eugene Brevdo、Greg Brockman、Sebastien Bubeck、Che Chang，以及另外107位作者。2025年Gpt- oss-120b 与 gpt-oss-20b 模型卡。预印本，arXiv:2508.10925。

OpenAI、Aaron Jaech、Adam Kalai、Adam Lerer、Adam Richardson、Ahmed El-Kishky、Aiden Low、Alec Helyar、Aleksander Madry、Alex Beutel、Alex Carney、Alex Ifti mie、Alex Karpenko、Alex Tachard Passos、Alexander Neitz、Alexander Prokofiev、Alexander Wei、Allison Tam、Ally Bennett，以及其他 243 人。2024 年。OpenAI o1 系统卡。预印本，arXiv:2412.16720。

Haizhou Shi、Zihao Xu、Hengyi Wang、Weiyi Qin、Wenyuan Wang、Yibin Wang、Zifeng Wang、Sayna Ebrahimi和Hao Wang。2025年。大语言模型的持续学习：全面综述。*ACM*计算调查，58(5):120:1–120:42。

Issa Sugiura、Takashi Ishida、Taro Makino、Chieko Tazuke、Takanori Nakagawa、Kosuke Nakago和David Ha。2025年。EDINET-Bench：使用日语财务报表评估大语言模型在复杂金融任务上的表现。预印本，arXiv:2506.08762。

Kota Tanabe、Masahiro Suzuki、Hiroki Sakaji和Itsuki Noda。2024 年。JaFIn：日语金融指令数据集。载于*2024* 年*IEEE*金融工程与经济学计算智能研讨会（*Cifer*），第1–10页。

Shijie Wu、Ozan Irsoy、Steven Lu、Vadim Dabravolski、MarkDredze、Sebastian Gehrmann、PrabhanjanKam- badur、David Rosenberg 和 Gideon Mann。2023年。BloombergGPT：面向金融领域的大型语言模型。预印本，arXiv:2303.17564。

An Yang、AnfengLi、Baosong Yang、Beichen Zhang、Binyuan Hui、Bo Zheng、Bowen Yu、Chang Gao、Chengen Huang、Chenxu Lv、Chujie Zheng、Day- i heng Liu、Fan Zhou、Fei Huang、Feng Hu、Hao Ge、Haoran Wei、Huan Lin、Jialong Tang 等 41 位作者。2025年。Qwen3技术报告。预印本，arXiv:2505.09388。

数据集流水线参数详情

本研究数据集构建流程中各步骤的详细配置参数见表2。

B LLM 评审质量过滤详情

我们使用 LLM 评审对生成的多轮

Step	参数	值
1. 主题词选择	金融领域种子词 通用领域（用于混合）的种子词	135 20
2. 用户问题生成	生成数量 / 子主题（多项选择） 生成数量 / 子主题（其他）	8 10
3. 指令扩展	扩展变体 / 问题（多项选择） 扩展变体 / 问题（其他）	3 5
4. 筛选	最低词数（低于此数值的样本将被排除）	10
5. 多轮生成	最大轮数	3

表 2：数据集构建流程配置参数

交互式指令数据集进行了质量评估和筛选。我们使用 gpt-oss-120b 作为评审模型，并根据表3所示的五个评估视角，评估每个样本中的大语言模型回复是否恰当。我们针对每个视角进行了五级评估（1–5 分），并要求评估模型采用客观、严格的标准。

数据集评估结果显示，各视角获得 5 分的比例分别为：准确性 81.7%、相关性/指令遵循 97.3%、实用性/全面性 97.6%、推理/深度 91.2%，以及安全性/恰当性 99.95%。准确性的得分最低，证实了事实性错误和幻觉的存在。本研究为确保监督微调数据的最高质量，设定了严格的筛选标准，仅选取在全部 5 个视角上均获得 5 分的样本。该标准从 839,398 个样本中筛选出 632,636 个（75.4%），约 24.6% 的样本被排除。被排除的样本存在重复（重复相同短语）、对指令作出不恰当回应，以及推理过程崩溃等问题。

C 数据集详情

训练所用数据集的样本数量、词元数量以及是否包含多轮对话的信息见表4。

D 金融基准评测方法的修改

对于 japanese-lm-fin-harness，公共代码仓库中的实现采用与选项相对应的词元的对数似然进行判断，但这会使输出推理轨迹的指令模型的评估变得不稳定，因此我们将提示词修改为聊天格式，并要求模型在 \boxed{} 中作答。此外，由于近期模型往往会指定特定的温度和 top-p 采样参数以稳定深度推理，我们并未使用通过贪心方法进行的确定性解码，而是采用 pass@1 格式进行评估，并对多次尝试的结果取平均值。

对于 pfmt-bench-fin-ja，与公开仓库中的实现不同，我们使用性能更高的 GPT-5 mini（版本：2025-08-07）作为 LLM 评审模型。

视角	标准（5 个等级）
准确性（事实正确性、是否存在错误信息）	5：完全准确 / 3：存在一些错误 / 1：存在大量错误信息
回答的相关性/指令遵循	5：完全适当 / 3：部分未解决 / 1：偏离主题
回答/信息的实用性/全面性	5：极其有用 / 3：一般 / 1：不足
推理/深度质量、分析、逻辑一致性	5：推理出色 / 3：一般 / 1：逻辑崩溃
安全性/恰当性，表达伦理	5：完全安全 / 3：轻微问题 / 1：危险

表 3：质量评估视角与标准

数据集名称	类别	样本数	总token数	平均token数（/样本）				多轮
				user	助手推理		总计	
nri-fin-reasoning	金融/通用	632,636	6,352,341,377	191	8,996	855	10,041	Yes
NuminaMath-CoT-modified	Math	88,727	141,678,621	83	508	1,006	1,596	No
Nemotron-Post-Training-math-modified	Math	694,168	2,999,298,386	169	640	3,512	4,320	No
smol-constraints-modified	指令遵循	24,411	12,803,102	71	199	254	524	No
总计	—	**1,439,942 9**	**,506,121,486**	**171.7**	**4,295.4**	**2,134.6 6**	**,601.7**	—

表 4：训练中使用的各数据集统计信息