Gemini - Research Proposal Translation Request

Trading-R1：通过强化学习实现具备推理能力的大语言模型金融交易

Yijia Xiao [1], Edward Sun [1], Tong Chen [2], Fang Wu[3], Di Luo [1], Wei Wang [1]

[1] 加州大学洛杉矶分校 [2] 华盛顿大学 [3] 斯坦福大学 Tauric Research

摘要

【原文翻译】 在金融人工智能领域，发展与人类金融分析师和交易员相媲美的专业化、结构化推理能力仍然是一个核心挑战，因为市场要求可解释性和信任。传统的时间序列模型缺乏可解释性，而大语言模型（LLMs）在将自然语言分析转化为纪律严明、可执行的交易方面面临挑战。尽管具备推理能力的LLMs在分步规划和验证方面取得了进展，但它们在风险敏感的金融决策中的应用仍未得到充分探索。我们提出了TRADING-R1，一个具备金融意识的模型，它融合了战略性思考和规划，以实现全面的论点构建、基于事实的分析以及经波动性调整的决策制定。TRADING-R1通过监督微调和强化学习，并采用一个从易到难的三阶段课程，使其推理与交易原则保持一致。训练使用了Tauric-TR1-DB，这是一个包含10万个样本的语料库，时间跨度18个月，涵盖14只股票和五种异构金融数据源。在六种主要股票和ETF上的评估表明，与开源和专有的指令遵循模型以及推理模型相比，TRADING-R1展示了更高的风险调整后回报和更低的回撤。该系统能够生成结构化的、有证据支持的投资论点，从而支持纪律严明且可解释的交易决策。TRADING-R1终端将在 https://github.com/TauricResearch/Trading-R1 发布。

【深度解读】 这篇摘要为你打开了一扇通往金融AI前沿领域的大门。想象一下，金融世界就像一个巨大且瞬息万变的棋局。过去的AI，就像只会计算棋子价值的简单程序（**传统的时间序列模型**），它们能告诉你过去的数据规律，但无法解释“为什么”要这么走，这在动辄数百万美元的交易中是无法让人信任的。而现在，我们有了像人类一样能理解语言和逻辑的**大语言模型（LLMs）**，比如你可能听说过的ChatGPT。但问题是，这些通用LLM虽然能写诗、聊天，却像一个夸夸其谈的“理论家”，无法将它们的分析转化为精准、果断的交易操作，因为金融市场充满了“噪音”和不确定性。

这篇论文的核心贡献，就是打造了一个名为**TRADING-R1**的“专家级AI交易员”。它不仅仅是一个模型，更是一个决策框架。作者们教它像一个华尔街的顶级分析师那样思考：

1. **构建投资论点 (Comprehensive thesis composition)**：不是简单地给出“买”或“卖”的指令，而是先写一份详细的分析报告，说明为什么看好或看空某只股票。
2. **基于事实分析 (Facts-grounded analysis)**：报告中的每一个观点都必须有数据支持，比如公司财报、新闻报道等，杜绝“张口就来”的猜测。
3. **波动性调整决策 (Volatility-adjusted decision making)**：在做决策时，会考虑市场的风险水平。在高风险时期（市场波动大）会更谨慎，这就像开车在暴风雨中会减速一样。

为了实现这个目标，他们设计了一套“魔鬼训练营”：一个**从易到难的三阶段课程**。首先通过**监督微调（Supervised Fine-Tuning）**，像教学生做范文一样，让模型学习海量专业报告的格式和风格。然后通过**强化学习（Reinforcement Learning）**，像训练宠物一样，如果模型根据数据做出了正确的决策（比如预测股价上涨并且真的涨了），就给它“奖励”，反之则“惩罚”，让它在不断的试错中学会真正的交易技巧。

这个训练营的教材也非同一般，是一个名为**Tauric-TR1-DB**的庞大数据库，包含了18个月内14家大公司（如苹果、英伟达）的股价、财报、新闻、高管交易和宏观经济数据等，总计10万份学习材料。最终的“毕业考试”结果令人振奋：TRADING-R1在真实历史数据回测中，不仅比其他AI模型赚了更多的钱，而且在市场下跌时亏得更少（**更高的风险调整后回报和更低的回撤**）。最关键的是，它的每一个决策都附带一份逻辑清晰、证据确凿的分析报告，解决了金融AI最核心的“信任”问题。

1. 引言

【原文翻译】 金融交易的历史早于文字记载，但阿姆斯特丹证券交易所的成立通常被视为现代证券市场的诞生 [Petram (2014)]。从那时起，学者和从业者提出了各种各样的理论来解释价格形成和指导交易，涵盖了社会学和心理学解释、计量经济学模型以及诸如艾略特波浪理论、道氏理论和价格行为等技术范式。随着时间的推移，可用的市场数据量、计算能力和技术发生了巨大变化并急剧增加，量化方法蓬勃发展，自然语言处理的进步使得利用情感分析等工具对新闻、财报披露和宏观经济报告等各种形式的非结构化来源进行大规模分析成为可能。然而，这些信号很少被整合到一个连贯的决策框架中，而是被分析师和公司作为独立的工具来使用。相反，定制的因子通常是孤立地被设计出来，留给人类交易员去解释和组合。

【深度解读】 这一段是在为我们铺陈历史背景，告诉你为什么TRADING-R1这项研究如此重要。作者从四百多年前的**阿姆斯特丹证券交易所**讲起，那里是现代股票交易的摇篮。从那时起，无数聪明人就试图破解市场的密码。

最早的尝试五花八门，有些像心理学家，试图分析大众的恐慌与贪婪（**社会学和心理学解释**）；有些像数学家，用复杂的公式建立模型（**计量经济学模型**）；还有一些像艺术家，试图在股价图表中寻找优美的形态和韵律，比如**艾略特波浪理论**和**道氏理论**，他们相信市场的波动像潮汐一样有规律可循。

然而，时代变了。今天我们拥有的工具是几百年前的人无法想象的：海量的市场数据（**可用市场数据量...急剧增加**）、超级计算机的强大算力（**计算能力和技术...巨大变化**）。这催生了“**量化交易**”——用数学和统计模型在市场中寻找微小的获利机会。同时，**自然语言处理（NLP）**技术的发展，让计算机能够“阅读”和理解成千上万篇新闻报道、公司财报，并判断其中的情绪是积极还是消极（**情感分析**）。

但这里出现了一个核心问题，也是这篇论文要解决的痛点：我们拥有了无数的“武器”（各种数据信号），但它们却各自为战，散落一地。分析师们看着股价图、读着新闻、算着估值，最后还是得靠自己的大脑这个“中央处理器”把所有信息整合起来，做出最终决策。这个过程高度依赖个人经验，效率低下且难以复制。作者指出，这些信号“**很少被整合到一个连贯的决策框架中**”，而是“**留给人类交易员去解释和组合**”。这正是AI可以大显身手的地方——创建一个能自动整合所有信息、像专家一样思考的“超级大脑”。

【原文翻译】 最近大型语言模型（LLMs）的突破性进展改变了各个领域的自动化推理能力。自然语言处理（NLP）已经从单一用途的模型转变为可提示的智能体：即通过思维链、自我验证和强化学习增强的通用系统，现在能够以越来越高的可靠性和范围处理复杂的推理任务。然而，它们在金融领域的应用仍处于起步阶段。市场是动态、嘈杂且多因素的，要求在不确定性下进行适应性强且可解释的推理，这些要求与近期LLM优化所主导的数学、编码和科学任务显著不同 [Hendrycks et al. (2021); Lu et al. (2024)]。虽然存在其他选择，特别是纯量化方法 [Ericson et al. (2024); Fjellström (2022)]，但它们通常不透明且在不同市场环境下表现脆弱；与此同时，通用的推理型LLM难以将其推论与金融背景相结合，提供可验证的逻辑，并产生可追溯的决策。此外，稀疏、碎片化的公开金融数据使模型训练变得复杂；而且，开放式的金融问答基准测试 [Liu et al. (2025b); Qian et al. (2025)] 与交易所要求的结构化、序列化推理之间存在不匹配。与问答不同，市场决策本质上是不确定的且具有路径依赖性：即使是合理的选择也可能产生截然不同、无法预料的结果。

【深度解读】 这一段将焦点从历史转向了当下最热门的技术——**大型语言模型（LLMs）**。作者告诉我们，AI的能力已经发生了质的飞跃。过去的AI模型通常是“专才”，比如一个模型专门做翻译，另一个专门做情感分析。而现在的LLM是“通才”，就像一个可以被“提示”的智能大脑（**promptable intelligence**）。通过**思维链（Chain-of-Thought）**，它能像人一样一步步思考问题；通过**自我验证（self-verification）**，它能检查自己的答案是否正确；通过**强化学习（reinforcement learning）**，它能在与环境的互动中不断进步。这使得LLM在数学、编程等逻辑性强的任务上表现出色。

然而，作者紧接着指出了一个关键问题：金融市场这个“考场”和数学、编程的考场完全不同。为什么LLM在金融领域“水土不服”？

1. **动态、嘈杂、多因素**：金融市场充满了真假难辨的信息（嘈杂），影响因素成千上万（多因素），并且规则和环境总在变化（动态）。这不像解数学题，有固定的公式和唯一的正确答案。
2. **要求可解释性**：在金融领域，一个错误的决策可能导致巨额亏损。所以，你不仅要知道AI的建议，更需要知道它“为什么”这么建议。纯粹的数字模型（**纯量化方法**）往往像个“黑箱子”，无法解释其决策逻辑，因此在市场风格切换时容易失效（**表现脆弱**）。
3. **决策的风险和路径依赖**：金融决策的后果是连锁反应。今天的一个买入决定，会影响你明天的资金和持仓，进而影响你未来的所有决策（**路径依赖性**）。更重要的是，一个当下看起来非常“合理”的决策，可能因为一个意想不到的事件（比如突发新闻）而导致亏损。这与问答任务截然不同，问答的答案通常是固定且客观的。

最后，作者还提到了现实的困难：高质量的金融数据稀少且分散，这给模型训练带来了麻烦。现有的金融AI评测标准（**金融问答基准测试**）也过于简单，无法模拟真实交易中需要的严谨、连续的推理过程。所有这些挑战，都为TRADING-R1的诞生提供了舞台。

【原文翻译】 为了解决这些差距，我们提出了TRADING-R1，一个为金融，特别是面向交易的推理而量身定制的金融交易推理基础模型。我们策划了一个包含超过10万个公开来源的金融推理样本的高质量数据集，并通过监督微调以及一个从易到难的强化学习课程来训练TRADING-R1。这种设计直接解决了现有推理型LLM在金融领域的核心局限性，推动了模型的发展，使其既能植根于市场复杂性，又能在实践中用于交易。

我们的主要贡献如下：

- **Tauric-TR1-DB**：一个大规模、多样化的金融推理语料库。我们策划了一个全面的数据集，时间跨度为2024年1月1日至2025年5月31日的18个月，涵盖14个主要股票代码，整合了真实交易员使用的异构金融数据，如技术市场数据、基本面、新闻、内部人情绪和宏观经济指标。最终的语料库包含10万个经过筛选的高质量金融信息样本，并通过反向推理蒸馏和波动率感知奖励标注进行监督。
- **使用反向思维链蒸馏的监督微调**。由于专有LLM仅提供最终答案而没有中间推理过程，我们从高性能但不透明的API模型中重建隐藏的推理轨迹，并将其用作面向推理训练的监督信号。这种方法使我们的模型能够生成从最先进的推理系统洞察中提炼出的简洁、可解释且高质量的投资论点。
- **用于执行级别决策的强化学习**。除了撰写论点，我们还对模型进行优化，以做出可操作的决策。我们将交易推荐视为一个强化学习问题，在标准的五级投资量表（强烈买入、买入、持有、卖出、强烈卖出）上对资产进行标注。评级经过波动性调整，并用作奖励，以使模型输出与现实的交易目标保持一致。
- **Trading-R1**：一个用于交易的金融推理LLM。我们提出了Trading-R1，一个在多样化资产和市场条件（牛市和熊市）下训练的大规模金融推理LLM。该模型在交易场景中表现出强大的泛化能力，既能产生高质量的分析，也能提供盈利的交易建议。

【深度解读】 面对前面提到的种种挑战，作者团队提出了他们的解决方案——**TRADING-R1**。它不是一个普通的聊天机器人，而是一个专门为金融交易打造的“**推理基础模型**”。你可以把它理解为一个经过特殊训练，拥有金融博士水平的AI大脑。它的训练方法也很有特色，先通过**监督微调**（看标准答案学习）打好基础，再通过一个**从易到难的强化学习课程**（在模拟实战中试错和提高）进行强化。

接下来，作者清晰地列出了他们的四大核心贡献，这基本上是整篇论文的“武功秘籍”：

1. **Tauric-TR1-DB：一个强大的“教材库”**。他们没有随便在网上找些数据，而是精心打造了一个高质量的数据集。这个数据集的特点是：
    - **大规模**：10万个样本，足够模型学到各种情况。
    - **时间跨度长**：18个月，覆盖了市场的不同阶段。
    - **多样化**：不仅有股价（**技术市场数据**），还有公司财报（**基本面**）、新闻、公司高管的买卖情况（**内部人情绪**）和整体经济状况（**宏观经济指标**），模拟了真实分析师需要看的所有信息。
2. **反向思维链蒸馏：一种聪明的“学习方法”**。这是一个非常巧妙的技术创新。想象一下，你想向一位顶尖的数学家学习解题思路，但他只告诉你答案，不告诉你过程。怎么办？你可以找一个聪明的助教，让他看着题目和最终答案，反向推导出这位数学家可能的解题步骤。这里的“顶尖数学家”就是那些性能强大但不公开内部工作原理的商业AI模型（如OpenAI的GPT-4），而“助教”就是作者们用来**重建隐藏推理轨迹**的另一个模型。通过这种“**反向工程**”，他们凭空创造出了大量高质量的“解题过程”，用来作为监督微调的“标准答案”。
3. **强化学习：从“理论家”到“实干家”的转变**。仅仅会写分析报告还不够，最终要落实到具体的交易决策上。他们把交易决策变成了一个强化学习问题。模型会给出“**强烈买入、买入、持有、卖出、强烈卖出**”这五种操作建议。然后，研究人员会根据后续市场的真实走势来评判这个建议的好坏，并给予“奖励”或“惩罚”。这个评价标准还考虑了市场的风险（**波动性调整**），确保模型学会的是在风险可控的前提下赚钱，而不是盲目冒险。
4. **Trading-R1：最终的“成品”**。这是以上所有努力的结晶——一个既能写出专业分析报告，又能给出靠谱交易建议的金融AI。它在牛市和熊市中都接受过训练，因此具有很强的适应能力（**强大的泛化能力**）。

2. 相关工作

2.1. 金融领域的大型语言模型

【原文翻译】 大型语言模型（LLMs）已在许多领域展现出令人印象深刻的能力，包括金融领域。为了使它们适应金融任务，研究人员通常依赖三种策略：在领域特定数据上进行预训练、在任务特定数据集上进行微调，以及应用强化学习来使模型行为与期望结果对齐。另一条研究路线是探索直接将预训练模型作为多智能体系统中的专门专家来使用。在这些设置中，LLMs被分配不同的角色，它们协调互动旨在引出更明确的金融推理，并增强系统的整体推理能力。

【深度解读】 这一段是文献综述的开篇，作者在告诉我们，在他们之前，学术界和工业界已经尝试了哪些方法来让LLM成为“金融专家”。这就像是在说：“在我们发明自己的武功之前，先看看江湖上已有的几大门派。” 作者将这些方法归纳为三大主流策略，外加一个新兴的“多智能体”流派。

1. **预训练 (Pretraining)**：这相当于从零开始培养一个金融领域的“神童”。研究人员收集海量的金融文本（比如几十年的华尔街日报、所有上市公司的财报），让一个模型从头开始学习，使其天生就带有“金融基因”。这种方法成本极高，但基础非常扎实。
2. **微调 (Fine-tuning)**：这更像是“因材施教”。拿一个已经很聪明的通用LLM（比如LLaMA），它已经懂得了人类的语言和基本逻辑，然后用专门的金融数据集对它进行“补课”，让它快速掌握金融领域的专业知识和术语。这是目前最常见的方法，性价比高。

3. **强化学习 (Reinforcement Learning)**：这是“实战演练”。让模型在模拟的金融环境中进行交易，做对了就奖励，做错了就惩罚。通过不断的试错，模型会自己摸索出一套最优的决策策略。这种方法最接近真实交易，但训练过程非常复杂且难以控制。

除了这三大主流方法，还有一个很有趣的新方向：**多智能体系统 (Multi-agent systems)**。这就像组建一个“AI投研团队”。一个LLM扮演“宏观分析师”，负责分析经济大势；另一个扮演“行业分析师”，负责研究特定行业；还有一个扮演“技术分析师”，负责看图表。它们分工合作，互相讨论，最终形成一个综合的投资决策。这种方法试图模仿真实金融机构的决策流程，通过协作来提升决策的质量和深度。

【原文翻译】 为金融领域预训练和微调LLMs 金融领域LLM的领域适应主要有两种方法：在金融语料库上从头开始预训练，以及在金融数据上微调现有模型。像BloombergGPT (Wu et al., 2023)、轩辕2.0 (Zhang et al., 2023b) 和Fin-T5 (Lu et al., 2023) 这样的模型是在公共数据集和金融特定数据集的组合上进行训练的。BloombergGPT利用了彭博的专有数据，在市场情绪分类和摘要任务上优于通用的同类模型如BLOOM-176B，同时在通用语言理解方面与类似规模的开源模型相比保持了竞争力。微调方法的代表模型有PIXIU (FinMA) (Xie et al., 2023)，它在13.6万条金融相关指令上对LLAMA进行了微调；FinGPT (Yang et al., 2023)，它使用LORA技术，用大约5万个金融特定样本来适配LLaMA和ChatGLM；以及Instruct-FinGPT (Zhang et al., 2023a)，它在来自金融情感分析数据集的1万个指令样本上进行了微调。这些模型在金融分类任务上表现出比其基础版本和其他开源LLM（如BLOOM和OPT (Zhang et al., 2022)）更强的性能，有时甚至超过了BloombergGPT。然而，它们在生成任务上的表现仍然落后于强大的通用模型如GPT-4，这表明需要更高质量的、领域特定的数据集。

【深度解读】 这里作者详细介绍了前面提到的“预训练”和“微调”两大门派的具体案例。

预训练派（从零培养）：

- **BloombergGPT**：这是这个门派的“掌门人”。它由全球最大的金融数据提供商彭博社打造，用其独有的、高质量的内部金融数据进行训练。因此，它在理解金融新闻、判断市场情绪等方面，比那些只学习了通用网络文本的“外行”模型（如BLOOM-176B）要强得多。这说明了“出身”的重要性——用专业数据喂大的模型，自然更懂专业。
- **轩辕2.0** 和 **Fin-T5**：这些是国内团队开发的类似模型，同样专注于中文金融领域。

微调派（半路出家）： 这个门派的策略是“站在巨人的肩膀上”。它们选择一个已经很强大的通用大模型（比如Meta的LLAMA），然后用特定的金融“教材”对它进行强化训练。

- **PIXIU (FinMA)**：它用13.6万条金融领域的“指令-回答”数据对LLAMA进行了微调，教它如何像金融专家一样回答问题。
- **FinGPT** 和 **Instruct-FinGPT**：它们使用了更少的数据（几万条），采用了更高效的微调技术（如**LORA**），这是一种“轻量化”的微调方法，只调整模型的一小部分参数，既省时又省力。

作者总结道，这些“半路出家”的模型在某些特定任务上（比如判断一段财经新闻是利好还是利空）表现非常出色，有时甚至能超过“科班出身”的BloombergGPT。但有一个关键的弱点：当需要它们像分析师一样写一份详细、有深度的分析报告时（**生成任务**），它们的能力就比不上像GPT-4这样的顶级通用模型了。这揭示了一个深刻的道理：仅仅灌输知识是不够的，底层的推理和生成能力同样重要。这也暗示了，要打造一个真正强大的金融AI，需要更高质量、更具启发性的训练数据，而不仅仅是简单的分类标签。

【原文翻译】 用于LLM的强化学习 来自人类反馈的强化学习（Kaufmann et al., 2023; Ouyang et al., 2022, RLHF）已成为使LLM与人类偏好对齐的基石技术（Lambert et al., 2024）。这种方法涵盖了从近端策略优化（Schulman et al., 2017, PPO）到直接偏好优化（Rafailov et al., 2023, DPO）和简单偏好优化（Meng et al., 2024, SimPO），后者消除了对显式奖励建模的需求，并有助于稳定训练。最近的创新，如组相对策略优化（Shao et al., 2024a, GRPO），通过优化分组比较和实施批量归一化奖励来解决计算挑战。显著的进展包括DeepSeek-R1的多阶段强化学习训练（DeepSeek-AI et al., 2025）和通过变分偏好学习实现的个性化对齐（Poddar et al., 2024）。尽管取得了重大进展，但仍然存在一些根本性的局限性，包括奖励“黑客”（reward hacking）、离策略不稳定性，以及需要多元化对齐以适应不同人类偏好的风险（Casper et al., 2023; Kaufmann et al., 2024）。

【深度解读】 这一段深入探讨了“强化学习”这个门派。强化学习的核心思想是让AI通过试错来学习，而**来自人类反馈的强化学习（RLHF）** 则是让这个试错过程更高效、更符合人类期望的关键技术。

想象一下，你在训练一个AI写摘要。它写了两个版本，A和B。你作为一个人类老师，觉得版本A写得更好。这个“A比B好”的反馈，就是RLHF的核心。通过收集大量这样的人类偏好数据，我们可以训练一个“奖励模型”，这个模型就能像人类一样，给AI生成的任何内容打分。然后，AI就会努力调整自己，以生成能获得更高分数的内容，从而使其行为越来越符合人类的偏好。

作者在这里列举了一系列技术名词，它们就像是强化学习工具箱里的各种工具：

- **PPO (近端策略优化)**：这是最经典、最常用的RL算法之一。它的特点是“稳健”，每次只对AI的策略做小幅调整，避免AI“学跑偏”了。
- **DPO (直接偏好优化)** 和 **SimPO (简单偏好优化)**：这些是更新、更高效的算法。它们巧妙地绕过了训练一个独立“奖励模型”的步骤，直接从人类偏好数据中学习，简化了流程并提高了训练的稳定性。
- **GRPO (组相对策略优化)**：这是本文后面会用到的一个更前沿的算法。它的创新之处在于，通过比较一组不同输出的好坏来评估策略，进一步提高了效率和稳定性，尤其适合大模型训练 。

然而，作者也坦诚地指出了强化学习的“阿喀琉斯之踵”：

- **奖励“黑客” (Reward Hacking)**：AI非常聪明，它可能会找到奖励规则的漏洞来“作弊”。比如，你奖励它写长摘要，它可能会写出一篇又长又没内容的废话来骗取高分。
- **不稳定性 (Off-policy instability)**：训练过程可能像过山车一样，时好时坏，难以控制。
- **对齐的多元性 (Pluralistic alignment)**：人类的偏好是多种多样的。一个摘要，有人喜欢简洁，有人喜欢详细。如何让AI满足不同用户的偏好，是一个巨大的挑战。

【原文翻译】 **用于金融的多智能体LLMs** 尽管在金融数据上训练模型可以提高性能，但有限的资源和数据可用性常常使现成的LLM成为一个有吸引力的替代方案。虽然大型通用模型并非专门针对金融领域，但它们在推理和指令遵循方面表现出色，这推动了智能体系统的兴起。智能体系统是一种框架，它为LLM配备了记忆、工具和角色专业化，以实现复杂的目标。这种范式已迅速扩展到从编码到AI4Science再到计算机使用智能体的各个领域 [Gottweis et al. (2025); Hong et al. (2024); Liu et al. (2025a)]。在金融领域，多智能体系统通常被设计用来复制真实的决策过程，例如对冲基金的结构，通过为智能体分配不同的角色和工具（例如，新闻检索、指标计算）。最近的框架，如Trading Agents，明确地模拟金融机构，结合结构化的沟通和辩论，以生成按信息来源分段的详细推理报告 [Xiao et al. (2025)]。

【深度解读】 这一段介绍了前面提到的第四个流派——**多智能体LLMs**。这个想法非常直观且强大：既然一个AI的能力有限，为什么不组建一个AI团队呢？

作者指出，对于很多没有海量专有数据（像彭博社那样）或巨大计算资源的研究者来说，从头训练或微调一个金融大模型是不现实的。一个更聪明的办法是直接使用那些现成的、非常强大的通用大模型（**off-the-shelf LLMs**），比如GPT-4，然后把它们“武装”起来。

智能体系统（Agent systems）就是这种“武装”的框架。它给一个普通的LLM赋予了三样法宝：

1. **记忆 (Memory)**：让AI能记住过去的对话和信息，从而进行长期规划。
2. **工具 (Tools)**：让AI能使用外部工具，比如连接互联网搜索最新新闻、调用计算器进行数学运算，或者访问数据库查询股价。
3. **角色专业化 (Role specialization)**：给不同的AI分配不同的角色。

在金融领域的应用就非常形象了。你可以创建一个模拟**对冲基金**的AI团队：

- **AI分析师A**：角色是“新闻检索员”，负责24小时监控全球新闻，并提取与某公司相关的关键信息。
- **AI分析师B**：角色是“量化分析师”，负责调用工具计算各种技术指标，如移动平均线、相对强弱指数等。
- **AI分析师C**：角色是“基本面分析师”，负责阅读公司的财务报表，分析其盈利能力和健康状况。
- **AI基金经理**：负责汇总所有AI分析师的报告，组织一场“辩论”，并根据辩论结果做出最终的投资决策。

像**Trading Agents**这样的框架，就是这种思想的具体实现。它通过结构化的沟通（比如要求分析师必须提交标准格式的报告）和辩论机制，来确保最终生成的决策报告逻辑严密、考虑周全。这种方法不直接提升单个AI的能力，而是通过设计一个高效的协作流程来提升整个系统的智能水平。

2.2. 金融交易中的大型语言模型

【原文翻译】 LLMs在金融交易中主要应用于四种范式：信息处理、基于推理的决策、强化学习优化和阿尔法因子生成。

信息驱动交易 信息驱动方法处理新闻和市场数据以生成交易信号。评估封闭源代码模型（如GPT-4.1, Claude 3.7）和开源LLM（如Qwen (Bai et al., 2023)）的研究表明，基于情感分数的简单多空策略是有效的。像FinGPT这样的微调LLM通过领域特定的对齐显示出改进的性能。更先进的方法包括总结金融新闻并推理它们与股价之间的关系。

推理增强交易 推理增强方法通过反思和多智能体辩论来利用LLM的分析能力。基于反思的系统，如FinMem (Yu et al., 2023) 和FinAgent (Zhang et al., 2024b)，采用分层记忆和多模态数据来总结输入、为决策提供信息并整合技术指标，从而在回测中取得更好的性能，同时减轻幻觉问题 [Ji et al., 2023]。多智能体框架 (Xiao et al., 2025; Xing, 2024) 通过让专业智能体之间进行LLM辩论来增强推理和事实有效性。像TradingGPT (Li et al., 2023) 这样的系统通过这种协作方法展示了改进的情感分类和增强的交易决策稳健性。

强化学习优化 强化学习优化的交易系统使用回测性能作为奖励来完善决策过程。SEP (Koa et al., 2024) 采用带有记忆和反思的强化学习，根据市场历史来优化LLM的预测。经典RL方法也被整合到框架中，这些框架将LLM生成的嵌入与股票特征相结合，通过近端策略优化（PPO）等算法进行训练。这些方法通过迭代反馈循环系统地提高LLM的交易能力。

阿尔法因子生成 LLM不直接做出交易决策，而是可以生成阿尔法因子——即预测股票回报的信号。QuantAgent (Wang et al., 2023) 采用双循环架构：内循环中，一个编写者智能体根据交易思想生成代码，并获得一个裁判智能体的反馈；外循环中，代码在真实市场中进行测试，以增强裁判智能体的能力。类似地，AlphaGPT (Wang et al., 2023) 提出了一个用于阿尔法挖掘的人在环路框架。这些方法利用LLM的能力，通过系统地生成和完善预测信号来自动化和加速交易策略的开发。

【深度解读】 这一部分将LLM在金融交易中的具体应用场景，划分成了四个清晰的类别。这就像是分析一位武林高手，不仅要看他属于哪个门派，还要看他具体擅长哪几种武功。

1. **信息驱动交易 (Information-Driven Trading)**：这是最基础的应用，相当于AI的“信息搜集和初步判断”能力。
    - **核心任务**：让LLM阅读新闻，然后给出一个**情感分数**（比如，这篇新闻对苹果公司是+0.8的利好，还是-0.5的利空）。
    - **策略**：根据这个分数制定简单的交易规则，比如“分数高于0.7就买入，低于-0.7就卖出”（**多空策略**）。
    - **进阶**：更高级的方法是让LLM不只给分数，而是写一段**摘要**，并**推理**新闻内容与股价可能的关系。
2. **推理增强交易 (Reasoning-Enhanced Trading)**：这更进了一步，强调AI的“深度思考”能力。
    - **核心任务**：不仅仅是处理信息，还要进行逻辑推理和分析。
    - **方法一：反思 (Reflection)**。系统如**FinMem**和**FinAgent**，会让AI在做出决策后，回顾历史上的相似情况和决策结果，进行“复盘”，从而在下一次决策时做得更好。这模仿了人类的经验积累过程。
    - **方法二：多智能体辩论 (Multi-agent debate)**。正如前面提到的，让多个扮演不同角色的AI进行辩论，通过思想的碰撞来得出更全面、更可靠的结论。系统如**TradingGPT**就是这么做的。
3. **强化学习优化 (Reinforcement Learning Optimization)**：这是最直接的“实战训练”。
    - **核心任务**：让AI直接在模拟市场中进行交易，并根据最终的盈亏来学习。
    - **奖励机制**：这里的“奖励”非常直接，就是**回测的收益率**。赚钱了就给正奖励，亏钱了就给负奖励。
    - **方法**：系统如**SEP**让AI在交易中加入“记忆”和“反思”机制。其他方法则将LLM对信息的理解（**嵌入**）与其他量化数据（**股票特征**）结合，然后用经典的RL算法（如**PPO**）来训练一个端到端的交易模型。
4. **阿尔法因子生成 (Alpha Factor Generation)**：这是一种更“高级”和间接的应用，AI不直接当“交易员”，而是当“策略研究员”。
    - **核心任务**：在金融领域，“**阿尔法因子**”指的是能够预测股票未来收益的有效信号。例如，“一家公司过去三个月的盈利超预期程度”可能就是一个阿尔法因子。
    - **方法**：系统如**QuantAgent**和**AlphaGPT**，让LLM根据市场数据和金融理论，自动地提出新的交易思想，并把这些思想写成代码（即阿尔法因子），然后在历史数据上进行测试。这个过程就像是让AI 7x24小时不间断地进行量化策略研究，极大地加速了策略的发现和迭代过程。

这四种范式展示了LLM在金融交易中从简单到复杂的应用层次，而本文的TRADING-R1，实际上是深度融合了前三种范式的一种综合性解决方案。

3. Trading-R1 方法论

3.1. 动机

【原文翻译】 与其他领域相比，为金融交易训练一个推理模型是独一无二的挑战。金融决策风险高、涉及面广、依赖市场，并且对噪声高度敏感。仅仅通过标准的推理训练来延长思维链并不一定能提高模型质量；相反，它可能会放大幻觉并降低生成的交易决策的可靠性。由于语言模型是条件自回归生成器，最终行动的质量取决于两个相互耦合的先验：(i) **外部先验**，由启动生成的输入上下文给出；以及 (ii) **内部先验**，由模型在展开过程中自己先前生成的词元所塑造。这些动态导致了两个实践上的必要性：

- **输入质量（外部先验）** 如果提示上下文充满噪声、不一致或信噪比低，模型的分析就会锚定在劣质证据上，无论解码或提示如何，都会降低下游的推理质量。
- **推理脚手架（内部先验）** 在生成过程中，结构不良的中间思想会累积起来，导致脆弱的论点和不可靠的决策。交易需要一个纪律严明的投资论点，具有清晰的结构、可辩护的主张、明确的证据和对风险的仔细关注。提供这样的脚手架可以确保推理过程保持连贯，并且最终行动基于合理的逻辑。

这些挑战激发了我们为Trading-R1设计的方法论。我们同时控制模型所依赖的条件以及它如何推理出决策。具体来说，我们 (1) 实施了严格的数据收集、清洗和组装流程，以在训练时提供高信噪比、基于金融的上下文，以及 (2) 采用了一个多阶段、从易到难的课程，用于监督微调和强化学习，该课程首先教模型如何构建投资论点，然后构建逻辑上、基于证据的论证，最后做出基于市场动态的决策。这种设计使Trading-R1能够像专业交易员一样推理，生成有根据且透明的分析，从而得出连贯、可操作的决策，而不仅仅是产生更长的文本。

【深度解读】 这一节是整篇论文方法论的灵魂，解释了作者设计思路的根本出发点。他们开宗明义地指出，金融交易对AI来说是一个“**独一无二的挑战**”，因为它**风险高、因素多、充满噪音**。

作者提出了一个非常深刻的观点：对于一个语言模型来说，想让它做出高质量的决策，必须控制好两个关键的“**先验**”（可以理解为“前提条件”）。

1. **外部先验 (External Prior) - 输入质量**：这很好理解，就是“**垃圾进，垃圾出**”。如果喂给模型的是一堆充满错误、谣言和无关紧要信息的“垃圾数据”，那么无论模型本身多聪明，它做出的分析也必然是“垃圾”。这就解释了为什么他们要花费巨大精力去构建Tauric-TR1-DB这个高质量数据集。他们要确保喂给模型的每一口“粮食”都是营养丰富、干净卫生的。

2. **内部先验 (Internal Prior) - 推理脚手架**：这个概念更深入，也更关键。语言模型生成文本，就像我们说话一样，是一个一个词往外蹦的（**自回归生成器**）。如果你在思考一个复杂问题时，脑子里的思路是混乱的、东一榔头西一棒子，那么你最终得出的结论很可能也是错乱的。AI也是如此。如果在推理过程中，它的中间“想法”（**中间思想**）结构混乱，那么错误就会像滚雪球一样越滚越大，最终导致一个看似头头是道、实则不堪一击的结论（**脆弱的论点和不可靠的决策**）。

因此，作者认为必须给模型的“思维”过程搭一个“**脚手架**”（**scaffolding**）。这个脚手架，就是专业投资论证的固定结构：**清晰的结构、可辩护的主张、明确的证据、风险评估**。这个结构强迫模型必须按照一个纪律严明的逻辑流程来思考，防止其“思维发散”导致“胡言乱语”。

基于这两个核心洞察，作者提出了他们的解决方案：

- **针对外部先验**：建立一套严格的数据处理流水线，确保输入信息的**高信噪比**。
- **针对内部先验**：设计一个**从易到难的三阶段训练课程**，先教模型搭好“脚手架”（构建论点结构），再教它往脚手架上填充“砖块”（组织证据和论证），最后教它根据盖好的“大楼”做出最终判断（决策）。

这个设计的最终目标，是让TRADING-R1不仅仅是生成更长的文本，而是能像一个真正的专业人士一样，进行**有根据、透明、连贯且可操作的推理**。

3.2. 规模化输入数据收集

【原文翻译】 为了管理外部先验并确保高质量的训练输入，我们实施了一个严格的数据收集过程，该过程跨越了不同的时间段、市场条件、行业和分析模式。在金融交易研究中，主要挑战不是获取数据，而是选择能够提高信噪比并产生可操作见解的信息。我们的数据集建立在可靠的来源之上，捕捉了市场动态、公司基本面和公众情绪。我们广泛地定义输入数据，涵盖了对宏观经济趋势和公司特定状况的整体看法——公司做什么，它们表现如何，以及它们如何被看待。为了建立可泛化的市场情报并为模型提供最强的先验知识以生成高质量的论点，我们的收集过程遵循三个核心目标：

公司的广度 我们包含了来自不同行业和市值的各种股票的数据。通过关注十多家广受关注的公司（例如，NVDA、AAPL、JNJ）在2024年1月1日至2025年5月31日的18个月期间，我们捕捉了广泛的市场状况和公司发展。

信息的深度 对于每一天和每一个给定的资产，我们聚合了涵盖技术数据、基本面、新闻、情绪和宏观经济因素的特征。来源包括Finnhub、SimFin、谷歌新闻抓取和stockstats，从而为每家公司提供了一个密集的、多视角的快照。

对变化的鲁棒性 真实世界的数据通常是不完整或不平衡的。为了增强弹性，我们在生成标签时通过从市场数据、新闻、情绪、基本面和宏观经济因素中随机抽样来改变输入组成。这种方法训练模型即使在信息有限的情况下也能有效推理。它还使模型能够检测跨上下文的复杂模式，同时保持对现实世界变异性的适应性，这对于在动态金融环境中取得成功至关重要。更多方法论细节在附录S1中提供。

【深度解读】 这一节详细阐述了他们如何解决“外部先验”的问题，即如何构建一个高质量的数据集。作者强调，在金融研究中，难点不在于找到数据，而在于从海量信息中**筛选出真正有用的信号，提高信噪比**。他们的数据收集策略围绕三个核心目标展开，旨在为模型提供最全面、最真实的“世界观”。

1. **广度 (Breadth)**：为了让模型不偏科，他们挑选了14家来自不同行业的巨头公司，比如科技行业的英伟达（NVDA）、苹果（AAPL），和医药行业的强生（JNJ）。时间跨度长达18个月，覆盖了市场的牛熊转换，确保模型见过各种“大风大浪”。这就像让一个学生不仅学习数学，还要学习物理、化学、历史，培养其全面的知识体系。
2. **深度 (Depth)**：对于每一家公司在某一天的情况，他们都力求做到“信息全覆盖”。他们从多个专业数据源（如Finnhub, SimFin）收集了五大类信息：
    - **技术数据**：股价、成交量等图表信息。
    - **基本面**：公司的财务报表，如收入、利润等。
    - **新闻**：与公司相关的最新动态。
    - **情绪**：分析师评级、社交媒体讨论等。
    - **宏观经济因素**：利率、通货膨胀等大环境信息。 这就像为侦探提供一个案件的所有线索，让他能从不同角度进行推理，而不是只看片面的证据。
3. **鲁棒性 (Robustness)**：这是非常关键的一点。真实世界的信息往往是**不完整或不平衡的**。比如，某一天可能没有关于某家公司的大新闻，或者财报还没发布。为了让模型能适应这种情况，他们在训练时会“故意捣乱”——**随机抽样一部分信息**喂给模型。有时给它看新闻和技术数据，有时给它看财报和宏观数据。这样训练出来的模型，就不会因为缺少某一种信息而“束手无策”，它学会了在信息不完整的情况下也能做出最合理的判断，大大增强了其在现实世界中的适应能力。

总而言之，这个数据收集策略的目标是打造一个既广博又深刻，还能适应信息残缺的“全能型”数据集，为训练一个强大的金融AI打下最坚实的基础。

3.3. Trading-R1 训练概览

【图1描述】 图1的标题是“Trading-R1 训练示意图”（Trading-R1 Training Schema）。这张图展示了一个复杂但结构清晰的三阶段训练流程，旨在将一个基础的语言模型（Reasoning Model）逐步训练成专业的金融交易模型（Trading-R1）。

整个流程图从左到右分为三个大阶段（Stage I, II, III），每个阶段都包含监督微调（SFT）和强化微调（RFT）两个步骤，并且有数据增强（Self Augmentation）环节。

- **左侧输入**：流程的起点是“推理模型”（Reasoning Model），它接收来自“Tauric DB”数据库的各种金融数据，如基本面（fundamentals）、情绪（sentiment）、结论（conclusion）等。这些数据被用来生成投资论点。
- **中间的三个阶段**：
    1. **第一阶段：结构 (STRUCTURE)**
        - **目标**：增强论点结构和主张格式（Enhance Thesis Structure & Claim Format）。
        - **流程**：首先进行**SFT**，使用规划好的章节（Planning sections）、示例（Examples）和记忆（Memories）来教模型如何组织专业的报告结构。然后进行**RFT**，通过“结构化奖励”（Structured Reward）来强化这种结构。最后通过“自增强”（Self Augmentation）筛选出结构清晰的案例，进一步巩固学习。
    2. **第二阶段：主张 (CLAIMS)**
        - **目标**：在第一阶段的基础上，让模型学会提出有证据支持的主张。
        - **流程**：同样先进行**SFT**，然后进行**RFT**。这里的奖励是“主张奖励”（Claim Reward），专门奖励那些观点明确、引用了证据（Evidence）、并标明了来源（Source）的论述。同样，通过自增强筛选出专业的、忠于事实的主张。
    3. **第三阶段：决策 (DECISION)**
        - **目标**：在前两阶段的基础上，让模型做出最终的、市场化的交易决策。
        - **流程**：SFT之后是RFT。这里的奖励是“预测奖励”（Prediction Reward），它基于市场的真实走向（Truth）来判断模型的“买入/卖出”建议是否正确。自增强环节会筛选出那些方向预测正确的案例。
- **右侧输出**：经过这三个阶段的“千锤百炼”，最终产出的就是**Trading-R1**模型。
- **图例**：图的右下角解释了SFT代表监督微调（Supervised Fine-Tuning），RFT代表强化微调（Reinforcement Fine-Tuning）。整个图用箭头和图标生动地展示了数据流和训练步骤的递进关系，强调了从“学习格式”到“学习论证”再到“学习决策”的渐进式、从易到难的课程设计。

【原文翻译】 管理LLM生成的内部先验对于交易至关重要。没有适当的结构，中间的推理步骤可能会累积错误，产生脆弱的论点和不可靠的最终决策。为了解决这个问题，我们设计了一个多阶段、从易到难的课程，将监督微调（SFT）与强化学习微调（RFT）交错进行，如图1所示。这个课程逐步教模型 (i) 像专业投资论点一样构建其输出，(ii) 构建逻辑上且有证据支持的论证，以及 (iii) 做出基于真实市场动态的决策。该课程分三个阶段展开，每个阶段都以SFT（图3）进行热启动，以建立结构和风格的先验知识，并用RFT通过特定任务的奖励来优化行为。这种交错进行确保模型首先掌握专业分析的一般形式，然后被引导向基于证据的推理，并最终实现与市场对齐的决策。分阶段的进展稳定了中间推理，减轻了错误的累积，并建立了连贯和可操作的交易所输出所需的内部纪律。在格式化阶段，模型因遵循投资论点的专业结构而获得奖励，系统地组织技术、基本面和基于情绪的分析。XML标签格式化在此阶段得到加强，以促进一致的推理模式和稳定的结构化输出。

【深度解读】 这一段是方法论的核心，详细解释了他们如何解决“内部先验”问题，即如何为模型的思维过程搭建“脚手架”。作者指出，如果任由AI自由发挥，它的“思路”很容易跑偏，导致最终结论不可靠。因此，他们设计了一套精密的“**多阶段、从易到难的课程**”，这套课程是整个TRADING-R1训练的精髓。

这个课程最大的特点是**交错进行（interleaves）**了两种训练方法：

1. **监督微调 (SFT)**：这相当于“理论课”。在每个阶段开始时，都先用SFT进行“**热启动**”（warm-started）。研究人员会给模型看大量高质量的范例，让它先模仿和学习专业分析报告的**结构和风格**。这就像教学生写作文，先让他背诵几篇满分作文，让他对好文章的样子有个基本概念。
2. **强化学习微调 (RFT)**：这相当于“实践课”。在模型掌握了基本格式后，再用RFT进行“**精炼**”（refined）。模型开始自己尝试写分析报告和做决策，然后系统会根据它做得好不好（比如结构是否清晰、论据是否充分、决策是否赚钱）给予**奖励或惩罚**。这就像老师批改学生的作文，写得好的地方画圈，写得不好的地方打叉，通过反馈来提升学生的能力。

这种“理论”与“实践”相结合的交错训练方式，确保了模型首先建立起正确的“思维框架”，然后再在这个框架内学习如何填充有价值的内容，并最终做出与市场一致的决策。作者强调，这种**分阶段的进展**（staged progression）有几个关键好处：

- **稳定中间推理**：防止模型在思考过程中“跑偏”。
- **减轻错误累积**：避免小错误滚雪成大雪球。
- **建立内部纪律**：让模型养成一种严谨、结构化的“思维习惯”。

特别地，在第一阶段（格式化阶段），他们还强调使用**XML标签**（比如`<fundamentals>`、`<technicals>`）来强制模型对不同类型的信息进行分类讨论，这进一步强化了输出的结构性和一致性。

【原文翻译】 在证据-基础阶段，奖励鼓励模型用输入上下文中的直接引文和引述来支持其主张，从而减少幻觉并培养纪律严明、基于证据的推理。最后，在决策阶段，模型使用第3.5节中基于波动率感知标签得出的基于结果的奖励进行训练，惩罚糟糕的预测并激励与可验证的市场结果一致的决策，如图4所示。通过这一进展，Trading-R1首先学会产生专业分析的正确形式，然后将其推理锚定在证据上，并最终生成连贯的、由市场驱动的交易决策。

【深度解读】 这里继续详细描述了训练课程的后两个阶段，以及每个阶段RFT（强化学习微调）的核心奖励机制。

- **第二阶段：证据-基础阶段 (Evidence-grounding stage)**
    - **目标**：解决LLM最大的通病之一——**幻觉 (hallucinations)**，也就是“一本正经地胡说八道”。
    - **奖励机制**：在这个阶段，系统会像一个严格的论文导师一样检查模型写的分析报告。如果模型提出了一个观点（**主张**），比如“苹果公司的最新财报显示其盈利能力强劲”，它就必须用**直接的引文和引述**来支撑这个观点，比如“*根据财报原文，其净利润同比增长了15%*”，并且要注明信息来源。凡是能做到“言必有据”的，就给予高额奖励。这种训练强迫模型养成**纪律严明、基于证据的推理**习惯，而不是凭空想象。
- **第三阶段：决策阶段 (Decision stage)**
    - **目标**：让模型的最终交易建议真正能够经受市场的检验。
    - **奖励机制**：这是最“残酷”也是最核心的阶段。模型的“买入/卖出”建议会被拿去和**真实的市场后续走势**进行比较。这个“标准答案”是经过精心设计的（即3.5节提到的**波动率感知标签**），它不仅考虑了股价是否上涨，还考虑了上涨的幅度和期间的风险。如果模型的预测与市场结果一致，就给予奖励；如果预测错误（**糟糕的预测**），就给予惩罚。这种**基于结果的奖励**机制，确保了模型的最终目标是与现实世界中的盈利目标对齐。

通过这三个循序渐进的阶段，TRADING-R1的成长路径被清晰地规划出来：

1. **学“形”**：先学会专业报告的**形式**和结构。
2. **学“神”**：再学会在这个形式下进行有**证据**支持的、有逻辑的**推理**。
3. **学“用”**：最后学会将这种推理能力转化为能适应真实**市场**、创造价值的**决策**。

3.4. 监督式投资推理蒸馏

【图2描述】 图2的标题是“Trading-R1蒸馏概览”（Overview of Trading-R1 distillation），包含(a)和(b)两个子图，详细解释了如何为监督微调（SFT）创造高质量的训练数据。

- **(a) 投资论点蒸馏 (Investment Thesis Distillation from OpenAI Reasoning Models)**：
    - 这个流程图展示了如何从一个强大的商业模型（如OpenAI-O3/O4 mini, Qwen 3）生成初步的交易建议。
    - **输入**：左侧是“Tauric DB”，代表包含新闻、公司简介、图表、财务数据、卖方评级、宏观经济和内部人交易等多种分类采样数据的数据库。
    - **过程**：从数据库中随机选择一部分数据作为输入，喂给一个强大的推理模型（如OpenAI-O3/O4 mini）。
    - **输出**：该模型会生成一个“交易提案”（Trading Proposal），包含一个交易建议（Recommendation）。
    - **筛选**：这个提案会经过一个“拒绝采样”（Rejected Sampling）的环节。如果提案的预测是正确的（if Correct），它就会被保留下来，进入下一步。
- **(b) 反向推理蒸馏 (Reverse Reasoning Distillation)**：
    - 这个流程图展示了本文的核心创新之一：如何从一个简单的“交易建议”反向工程出一个完整的、结构化的“推理过程”。
    - **输入**：输入是(a)中筛选出的正确“交易提案”（包含建议）和原始的“分类采样数据”。
    - **分解 (Decomposition)**：一个“规划者”模型（GPT-4.1）会接收这个交易建议，并将其分解成几个关键的“推理视角”（Reasoning Perspectives）。例如，要得出“买入”的结论，可能需要从竞争对手、技术分析、内部人交易等几个方面来论证。
    - **详述 (Elaboration)**：然后，一个更轻量级的模型（GPT-4.1 nano）会针对每一个“推理视角”，结合原始数据，生成详细的论述。比如，它会专门写一段关于技术分析的文字，解释为什么图表看起来是涨势。
    - **合并 (Merge)**：最后，所有这些分段的详细论述会被合并（Merge）成一个完整的、结构化的推理报告（Structured Reasoning），这个报告就是最终的“蒸馏推理”（Distilled Reasoning）结果。

总的来说，图2生动地展示了一个两步走的“数据制造工厂”：第一步，用一个强大的模型生成“答案”；第二步，用另一个（或一组）模型根据“答案”反向推导出详细的“解题步骤”。这个过程就是所谓的“**反向推理蒸馏**”。

【原文翻译】 为了支持Trading-R1交错的从易到难课程中的SFT热启动阶段（图1），需要高质量的推理轨迹作为监督目标。然而，为大型语言模型（LLMs）获取此类标签是众所周知的昂贵，而在金融领域，由于真实情况往往模糊不清、无法获得或生产成本过高，这一困难被放大了。为了解决这个问题，我们利用第3.5节中介绍的波动率驱动的标注方法，自动构建用于SFT的“输入-投资论点”对。每个投资论点都提供了一个详细的推理轨迹，该轨迹在逻辑上支持一个与当天分配的波动率感知标签一致的交易决策。

【深度解读】 这一段解释了作者面临的一个核心难题以及他们创新的解决方案。在机器学习中，**监督微调（SFT）**就像是教学生做题，你需要提供大量的“题目”和对应的“标准答案”。在TRADING-R1的训练中，“题目”就是某一天关于某只股票的所有金融数据，“标准答案”则是一份基于这些数据写出的、专业且逻辑严谨的投资分析报告（**投资论点**）。

难题在于：去哪里找成千上万份这样的高质量“标准答案”？

- **人工撰写？** 找金融专家来写，成本极高（**众所周知的昂贵**），而且耗时漫长。
- **寻找现成的？** 在金融领域，一份好的投资报告并没有绝对的“正确答案”（**真实情况往往模糊不清**），市场的走向是概率性的，这使得制作标准答案异常困难。

解决方案：作者提出了一种“**自动化生产标准答案**”的方法。他们利用了自己设计的**波动率驱动的标注方法**（将在3.5节详述），这个方法能根据未来的市场走势，自动为每一天的“题目”（输入数据）打上一个相对客观的标签，比如“应该买入”或“应该卖出”。然后，他们的目标就变成了：围绕这个“应该买入”的结论，自动地生成一篇看起来像是专家写的、逻辑通顺的分析报告。这个过程就是下一段要讲的“反向推理蒸馏”。这就像是，你先知道了考试的最终答案是“A”，然后让你反过来写出详细的解题步骤，证明为什么答案是“A”。

【原文翻译】 **反向推理蒸馏** 为了克服获取如此详细推理轨迹的困难，我们引入了一种我们称之为**反向推理蒸馏**的新技术。虽然通过API访问的商业LLM（例如，OpenAI的O1, O3）在推理质量上持续优于大多数开源模型，但它们通常不公开其完整的思维链（CoT）输出以便于进行完全蒸馏，只返回最终结论而没有解释步骤。为了在不自己托管大型模型的情况下提取高质量、长篇的推理，我们提出了一种从这些黑箱模型中综合重建推理路径的方法。

如图2a所示，我们首先将特定日期特定股票的结构化金融数据输入到专有的推理模型中，例如O3-mini或O4-mini，并检索其最终的交易建议（即前端响应）。接下来，如图2b所示，我们将这个最终响应连同原始输入一起，传递给一个专门的规划者LLM，该LLM的任务是推断出得出给定结论所需的关键推理步骤。为了模拟完整的推理过程，我们然后使用一个轻量级的LLM（例如，GPT-4.1-nano）来详细阐述每种数据模式（例如，市场数据、新闻、社交媒体、基本面）如何对投资决策做出贡献。然后，这些片段被程序化地拼接成一个连贯的推理轨迹。最终结果是一个高质量的、合成的数据集，其中结构化的金融输入与合理的、分步的投资论点配对，适合在SFT流程中使用。

【深度解读】 这一段详细解释了论文中最具创造性的技术之一：**反向推理蒸馏 (Reverse Reasoning Distillation)**。这个技术的出发点是为了解决一个实际问题：市面上最强大的AI模型（比如OpenAI的GPT系列）通常是**黑箱**的，你向它提问，它只给你最终答案，却不告诉你它是如何一步步思考得到这个答案的（**不公开其完整的思维链输出**）。这就像一个绝世高手，只展示最终的招式效果，却不传授内功心法。这对于想学习其“思维过程”的研究者来说是个巨大的障碍。

作者们发明的“反向推理蒸馏”就是一种“偷师学艺”的绝技。整个过程分为两步，正如**图2**所示：

1. **第一步：获取“答案” (图2a)**。他们先把某只股票的所有数据喂给一个强大的商业模型（比如**O3-mini**）。这个模型会给出一个简单的交易建议，比如“买入”。这个建议就是他们需要的“最终答案”。他们还会用一个筛选机制，只保留那些事后被证明是“正确”的建议。
2. **第二步：反向推导“解题过程” (图2b)**。这是最关键的一步。
    - **规划 (Planning)**：他们用另一个强大的模型（**规划者LLM**，如GPT-4.1）来扮演“侦探”的角色。这个“侦探”会看着原始数据和“买入”这个结论，然后反向推断：“要得出‘买入’的结论，一个理性的分析师可能会从哪些角度进行论证呢？哦，他可能会分析**基本面**、**技术面**和**新闻面**。” 这就定下了一份分析报告的“写作大纲”。
    - **详述 (Elaboration)**：然后，他们用一些更小、更高效的模型（**轻量级LLM**，如GPT-4.1-nano）来当“写手”。每个“写手”负责一个章节，比如写手A专门根据原始数据撰写“基本面分析”章节，写手B专门撰写“技术面分析”章节。
    - **拼接 (Stitching)**：最后，程序会自动把这些独立撰写的章节**拼接**成一篇完整的、逻辑连贯的投资分析报告。

通过这个巧妙的流程，他们成功地“凭空”制造出了大量高质量的训练数据——每一份数据都包含“输入数据”和一篇与之匹配的、看似由专家撰写的、分步清晰的“投资论点”。这为后续的监督微调提供了完美的“教材”，而且整个过程是自动化的，成本远低于人工制作。

3.5. 用于标签生成的波动率驱动离散化

【原文翻译】 一旦收集了广泛而多样的输入语料库，下一步就是定义一个可靠的目标标签。这可以作为一个明确的指标，指示在给定时间点的最佳交易行为，并支持市场可验证的、由奖励驱动的强化学习。我们没有尝试预测确切的未来价格变动，因为这些变动充满噪声、不稳定，并且语言模型尤其难以捕捉，而是将输出空间离散化为五个直观的行动：**强烈卖出**、**卖出**、**持有**、**买入**和**强烈**

买入。这种设计有两个目的。首先，它模仿了真实世界的交易，在真实世界中，决策是以行动而不是精确的价格预测来表达的。其次，它提供了一个从输出到投资组合分配权重的自然映射，可以根据用户的特定风险偏好进行调整。

标签是使用一个有原则的、多周期的波动率感知程序生成的。对于每个训练实例，我们从所有模式（市场、新闻、情绪、基本面和宏观经济信息）中抽样输入，并从多个时间周期构建一个复合信号。具体来说，我们计算指数移动平均（EMA）价格，并计算未来3天、7天和15天期间的远期回报。每个回报序列都通过其滚动的20周期波动率进行归一化，以创建类似夏普比率的信号。然后，这些信号使用经验确定的权重（分别为0.3、0.5、0.2）进行组合，形成一个复合加权信号。最后，根据从有效加权信号分布中计算出的百分位阈值来分配标签，使用反映市场动态的非对称分位数（85%、53%、15%、3%）。完整的过程在算法S1和附录S2中描述。这种多周期波动率感知设计有四个优势。(i) 信号通过多个时间周期捕捉了短期动量和中期趋势。(ii) 波动率归一化确保了在不同市场制度下信号强度的一致性。(iii) 加权组合平衡了即时价格行为与更广泛的趋势信息。(iv) 非对称分位数截点在保留市场长期向上漂移的同时，为稳健的训练保持了类别多样性。由此产生的代理标签对于下游学习非常有价值。它们为强化学习提供了自然的奖励信号（第3.7节），并使得能够为监督微调（第3.6节）大规模创建高质量的目标。这显著降低了基于推理的监督成本，否则将依赖于手动或专家标注。

【深度解读】 这一节解释了他们如何为强化学习（RFT）阶段制造“标准答案”，也就是**目标标签 (target label)**。这是整个训练流程的“指挥棒”，告诉模型什么样的决策是“好”的。

作者首先指出，直接让AI预测**未来的确切股价**是非常困难的，因为股价波动充满了随机性（**噪声**），就像预测下一秒钟风会吹向哪里一样。因此，他们采取了一种更聪明、更贴近现实的做法：将问题简化。他们不要求AI预测价格，而是要求它做出**行动决策**，就像真实的交易员一样。他们把复杂的输出空间简化（**离散化**）为五个选项：**强烈卖出、卖出、持有、买入、强烈买入**。这样做有两个好处：一是更符合交易实践，二是可以方便地将这些指令转化为具体的投资仓位（比如，“强烈买入”就投入20%的资金）。

那么，如何判断在某一天，正确的行动应该是什么呢？这就是本节的核心创新：一个**多周期的波动率感知程序**。这个过程听起来复杂，但可以分解为以下几个步骤：

1. **看未来，但看多个未来**：他们同时观察一只股票未来**3天、7天和15天**的收益率。这就像看天气预报，既要看短期的（明天是否下雨），也要看中期的（下周是否晴朗），综合判断。
2. **用风险来衡量收益（波动率感知）**：直接看收益率是有欺骗性的。一只股票可能一天涨了10%，但第二天就跌了15%，风险极高。所以，他们将每个周期的收益率除以这段时间的**波动率**（价格波动的剧烈程度）。这样得到的信号，更像是**风险调整后的收益**（类似夏普比率），更能反映投资的“性价比”。
3. **加权组合**：他们认为不同周期的重要性不同，所以给3天、7天、15天的信号分配了不同的权重（分别为0.3, 0.5, 0.2），形成一个**复合信号**。这表明他们更看重中期（7天）的趋势。
4. **非对称划分等级**：最后，他们根据所有股票在所有时间点上的这个“复合信号”的分布，来划分五个等级的边界。这个划分是**非对称的**。例如，排名前15%的信号被定义为“强烈买入”，而排名后3%的才被定义为“强烈卖出”。这样做是基于一个现实：股票市场长期来看是向上涨的（**长期向上漂移**），所以“买入”的机会天然就比“卖出”多。

通过这个精巧的设计，他们创造出了一套高质量、符合市场规律的“标准答案”。这套标签不仅为强化学习提供了清晰的奖励信号，还为前一节提到的“反向推理蒸馏”提供了最终结论，极大地降低了对人工标注的依赖。

3.6. 用于结构化分析的监督微调

【图3描述】 图3的标题是“使用来自Tauric-TR1-DB的反向推理蒸馏数据进行监督微调”（Supervised finetuning with reverse reasoning distilled data from Tauric-TR1-DB）。这张图清晰地展示了监督微调（SFT）阶段的工作流程。

- **左侧输入**：
  - **Tauric DB**：代表包含各种分类采样数据的金融数据库。
  - **蒸馏推理 (Distilled Reasoning)**：这是从图2流程中产生的、通过反向推理蒸馏技术制造出的高质量“投资论点”样本。这些样本包含了结构化的分析，如对竞争对手、技术分析、内部人交易的论述。

- **中间过程**：
    - **Trading-R1 (Pre-trained Weights)**：代表一个已经经过大规模预训练、拥有基础语言能力的模型。
    - **使用LoRA方法进行监督微调 (Supervised Finetune Using LoRA Method)**：这是核心步骤。预训练模型会学习“蒸馏推理”数据中的范例。LoRA是一种高效的微调技术，它只调整模型的一小部分参数，从而节省计算资源。
    - **示例案例 (Example Case)**：图中展示了一个具体的训练样本是如何构成的。它由三部分组成：
        1. **系统提示 (System Prompt)**：给模型设定一个角色，比如“你是一个金融分析师，需要提交结构化的报告”。
        2. **用户输入 (User Input)**：模拟用户的请求，比如“请分析这只股票”，并附上相关的股票数据。
        3. **助手输出 (Assistant Output)**：这就是模型需要学习模仿的“标准答案”，即那篇蒸馏出来的投资论点。这篇论点是**结构化的**，分解为关键部分（如基本面、技术面、宏观面），并且其中的论据都来自于输入的分类采样数据。最终，它会给出一个明确的投资决策（如**强烈买入/卖出**）。
- **右侧输出**：

    经过SFT训练后，模型就掌握了生成结构化分析报告的能力，其输出的“论点”（Thesis）会包含一个基于标签的分析和最终的投资决策。

总的来说，图3描绘了SFT这个“理论学习”阶段。模型通过学习大量的“题目-标准答案”对（用户输入-助手输出），掌握了像专业分析师一样，根据原始数据撰写结构清晰、逻辑严谨的投资报告的“套路”和“风格”。

【原文翻译】 用于结构化推理的SFT热启动 利用通过反向推理蒸馏生成的高质量投资论点，我们执行SFT来热启动从易到难课程的每个阶段（图3）。每个训练实例都将结构化的市场数据与详细的投资论点配对，教导模型以模仿专家金融工作流程的方式进行分析、综合和决策。因为蒸馏过程是可控的，所以可以设计特定阶段的SFT目标，以便在RFT精炼之前建立正确的推理先验。

阶段性目标 在第一阶段（结构），SFT强调论点的专业组织，灌输结构化思维和系统化的数据组织。在第二阶段（主张），SFT引入基于证据的推理，引导模型构建基于数据的论点。在第三阶段（决策），SFT专注于投资建议模式，为模型构建围绕可操作决策的输出做好准备。这种分阶段的热启动稳定了中间推理，减少了累积错误，并确保RFT在强大的结构和证据先验上运行。

【深度解读】 这一部分详细解释了**监督微调（SFT）** 在整个训练流程中的具体作用。作者将其定位为每个阶段的“**热启动**”（Warm-Start），这是一个非常形象的比喻。就像冬天启动汽车引擎前需要预热一样，在进行高难度的强化学习（RFT）之前，先用SFT对模型进行“预热”，能让后续的训练更平稳、更高效。

这个“预热”过程，就是让模型学习上一节中通过“**反向推理蒸馏**”技术制造出来的大量高质量“**投资论点**”范文。每一个学习样本都是一个“**数据-报告**”对，模型通过学习这些范例，来模仿专家的分析、综合和决策流程。

作者强调，由于这些“范文”是他们自己生成的，所以**蒸馏过程是可控的**。这意味着他们可以为训练课程的**每一个阶段**量身定制不同的“教材”：

- **第一阶段（结构）的SFT教材**：重点突出报告的**专业组织结构**。这个阶段的范文格式都非常标准，章节清晰，目的是让模型先学会“搭骨架”，养成**结构化思维**的习惯。
- **第二阶段（主张）的SFT教材**：重点展示如何进行**基于证据的推理**。这个阶段的范文里，每一个观点后面都跟着引文和数据来源，目的是教模型学会“用事实说话”。
- **第三阶段（决策）的SFT教材**：重点模仿**投资建议的表述方式**。这个阶段的范文会清晰地展示如何从前面的分析过渡到最终的“买入/卖出”结论，让模型学会如何给出**可操作的决策**。

通过这种**分阶段的热启动**，作者确保了模型在进入更具挑战性的RFT实践课之前，已经打下了坚实的理论基础。这不仅**稳定了推理过程**，**减少了错误**，更重要的是，它为RFT提供了一个非常好的起点（**强大的结构和证据先验**），使得RFT可以专注于优化决策本身，而不用再操心基本的格式和逻辑问题。

【原文翻译】 骨干模型和稳定性 我们采用Qwen3-4B作为骨干模型，因为它已经为推理任务进行了优化。这个先验知识在SFT和RFT期间都加速了收敛，同时提高了模型生成结构化、可解释输出的能力。没有这个热启动，模型倾向于过度拟合肤浅的启发式方法，忘记早期阶段的结构，并产生脆弱、不连贯的论点。相反，分阶段的SFT初始化提供了纪律严明的脚手架，保留了先前的知识，并允许强化学习去精炼而不是覆盖模型的分析能力。

【深度解读】 这里作者提到了他们选择的**基础模型（骨干模型）**以及SFT“热启动”对于训练稳定性的重要性。

他们选择了**Qwen3-4B**作为出发点。这里的“4B”指的是模型有40亿个参数，这是一个中等规模但性能强大的模型。选择它的原因是，这个模型本身在设计时就已经针对**推理任务**进行过优化，相当于一个“理科”比较好的学生。让一个有推理天赋的模型来学习金融推理，自然事半功倍，能够**加速收敛**（学得更快），并且更容易生成**结构化、可解释**的分析报告。

接着，作者通过一个反例，强调了SFT“热启动”的必要性。他们说，如果**没有这个热启动**，直接用强化学习去训练模型，会发生什么？

- **过度拟合肤浅的启发式方法**：模型可能会学到一些简单的、错误的规律，比如“看到‘上涨’这个词就买入”，而不是进行深度分析。
- **忘记早期阶段的结构**：在RFT追求高奖励的过程中，模型可能会为了“赚钱”而不择手段，把它在SFT阶段学到的专业报告结构忘得一干二净。
- **产生脆弱、不连贯的论点**：最终的输出可能逻辑混乱，前后矛盾。

因此，**分阶段的SFT初始化**就像是给模型打下了坚实的“地基”和“框架”（**纪律严明的脚手架**）。它让模型先掌握了正确的分析方法论，然后强化学习（RFT）就可以在这个坚实的基础上进行“精装修”（**精炼**），而不是“推倒重建”（**覆盖**）。这确保了模型在学习交易技巧的同时，不会丢失其核心的分析和推理能力。

3.7. 用于市场对齐决策的强化学习微调

【图4描述】 图4的标题是“基于论点结构、陈述和决策的强化学习”（Reinforcement learning on Thesis Structure, Statement, and Decision）。这张图直观地展示了强化学习微调（RFT）阶段的核心机制。

- **左侧输入**：
  - **Tauric DB**：代表输入的金融数据，包括新闻、情绪等。
  - **Trading-R1**：代表已经经过SFT热启动的模型。
- **中间过程**：
  - 模型接收数据后，会生成一份包含“交易提案”（Transaction Proposals）的“论点”（Thesis）。这个提案会是五个等级之一：强烈买入、买入、持有、卖出、强烈卖出。
  - **强化学习 (Reinforcement Learning)**：这是核心的反馈循环。系统会对模型生成的论点和提案进行评估，并给出奖励或惩罚。
    - **奖励 (Reward)**：如果生成的论点**结构良好**（If Well Structured），就会得到结构上的奖励。更重要的是，如果最终的交易提案被证明是**正确的**（If Correct），比如预测上涨后股价真的涨了，模型就会得到一个大的奖励。
    - **惩罚 (Penalty)**：如果交易提案是**错误的**（If wrong），模型就会受到惩罚。
- **右侧奖励机制**：
  - **非对称决策矩阵 (Asymmetric Decision Matrix)**：这是一个关键细节。图的右侧展示了一个奖励矩阵，它表明奖励和惩罚不是对称的。例如，错误地预测上涨（而市场实际下跌）所受到的惩罚，可能比错误地预测下跌（而市场实际上涨）更重。这体现了风险控制的思想。
  - **综合奖励**：最终的奖励信号是三部分的加权和：α(Structure)+β(Claims)+γ(Decision)。这意味着模型不仅要为正确的**决策**负责，还要为报告的**结构**和**主张**的质量负责。

总而言之，图4描绘了RFT这个“实践出真知”的阶段。模型在模拟环境中不断地做出决策，并从市场的直接反馈（奖励或惩罚）中学习。这个反馈机制是多维度的，既看重最终结果（决策是否正确），也看重过程质量（报告写得好不好），从而训练出一个既能赚钱又会分析的AI交易员。

【原文翻译】 **SFT热启动后的RFT微调** 虽然SFT为模型配备了结构化和可解释的推理能力，但它常常过度拟合于表面模式，并且在产生既稳健又可操作的决策方面表现不佳。直接从SFT过渡到基于结果的RL是不稳定的，因为模型缺乏平衡结构化推理与可验证市场表现的纪律。为了解决这个问题，我们交错的从易到难课程在每次SFT热启动后都应用RFT，通过基于结果的反馈来加强特定阶段的先验知识。通过这种方式，RFT精炼了模型的推理，使得高质量的分析能够转化为连贯的、与市场对齐的行动。每个阶段的详细奖励规范在附录S4中提供。

【深度解读】 这一段解释了为什么在SFT之后，还必须要有RFT这个步骤。作者坦言，只经过SFT“理论学习”的模型，虽然能写出格式漂亮、看似专业的报告，但存在两个致命缺陷：

1. **过度拟合表面模式**：模型可能只是学会了“套话”和“模板”，并没有真正理解背后的逻辑。它的分析可能看起来很美，但实际上是空洞的。
2. **决策能力不足**：它生成的决策可能不够**稳健**（在不同市场情况下都有效）和**可操作**（能真正用于交易）。

作者接着指出了一个关键的技术难点：如果直接从SFT跳到完全基于最终盈亏的强化学习（**基于结果的RL**），训练过程会**不稳定**。因为模型还没有学会如何在“写好报告”（结构化推理）和“赚钱”（市场表现）之间找到平衡。它可能会为了追求短期的高奖励而放弃好不容易学来的严谨分析结构，导致行为崩溃。

因此，他们设计的**交错式课程**再次显示出其精妙之处。在每个SFT阶段之后，都紧跟着一个**与之匹配的RFT阶段**。这个RFT阶段的奖励机制，是专门用来**加强**（reinforcing）刚刚在SFT中学到的技能的。

- 在“结构”阶段后，RFT会奖励那些**结构清晰**的报告。
- 在“主张”阶段后，RFT会奖励那些**论据充分**的报告。
- 在“决策”阶段后，RFT会奖励那些**预测准确**的决策。

通过这种方式，RFT的作用是**精炼**（refines）模型的推理能力，而不是推倒重来。它教会模型如何将高质量的分析**转化**（translate into）为与市场表现一致的、连贯的行动。这确保了模型在追求高收益的同时，不会丢掉其作为“分析师”的核心素养。

【原文翻译】 **行动空间和标注** 我们定义了一个五类行动空间——**强烈卖出、卖出、持有、买入、强烈买入**——并映射到投资组合权重。这种设计反映了不同程度的信念，并与传统的三元组“买入/持有/卖出”相比，能够实现更精细的头寸控制。标签是特定于资产的，由第3.5节中描述的多周期波动率感知程序构建。为了更好地捕捉市场动态，我们将回报投射到非对称分位数上，产生一个偏斜的分布，该分布既反映了股票市场的长期向上漂移，也反映了我们蓝筹股训练宇宙的增长导向特征，同时为稳健的训练保持了足够的类别多样性：

【表格2：交易行动的目标类别分布】

强烈买入	买入	持有	卖出	强烈卖出
15%	32%	38%	12%	3%

这种分布鼓励模型学习现实的、与市场一致的策略，同时为稳健的训练保留了类别多样性。非对称的分配既反映了经验性的股票市场行为，也反映了既定的分析师实践，其中偏向积极行动的偏见在我们训练宇宙的构成中尤其合理。

【深度解读】 这一部分详细说明了RFT阶段的两个基本设定：**行动空间**和**标签分布**。

首先，**行动空间**被定义为五个离散的动作：**强烈卖出、卖出、持有、买入、强烈买入**。这比传统简单的“买入/卖出/持有”三分类要**更精细**。它允许模型表达其预测的“**信念强度**”（conviction）。例如，“强烈买入”和“买入”都表示看涨，但前者传达的信心更强，在实际应用中可以转化为更大的投资权重或仓位。

其次，也是更重要的一点，是**标签的分布**。如**表格2**所示，这个分布是**非对称的**，或者说是“**偏多**”的。

- **买入类（强烈买入+买入）** 总共占了 15%+32%=47%。
- **卖出类（强烈卖出+卖出）** 总共只占了 3%+12%=15%。
- **持有** 占了 38%。

为什么会这样设计？这背后有两个深刻的金融学和市场观察：

1. **股票市场的长期向上漂移**：从长远来看，由于经济增长和公司盈利，股票市场整体是上涨的。因此，在任何一个时间点，“买入”或“持有”是正确决策的概率天然就高于“卖出”。这个标签分布反映了这一客观规律。
2. **训练集的资产特征**：作者在下一段会提到，他们训练用的股票都是**蓝筹股**（blue-chip companies），比如英伟达、苹果、微软等。这些都是基本面非常优秀、具有长期增长潜力的龙头公司。对于这类公司，长期持有的策略本身就是有利的。因此，一个偏向于积极行动（买入/持有）的AI策略，是符合这些资产内在价值的。

这种**非对称的标签设计**，体现了作者将金融市场的先验知识融入到了AI的训练过程中。它鼓励模型学习一种与市场长期趋势和优质资产特性相符的、现实的投资策略，而不是一个在理论上多空完全对称的“理想化”模型。

【原文翻译】 我们的投资组合专注于大盘和超大盘蓝筹公司，包括科技领袖如英伟达、微软和苹果，成熟的金融公司如伯克希尔·哈撒韦和摩根大通，医疗保健巨头如礼来和强生，以及像SPY和QQQ这样的广泛市场ETF。这些公司共同代表了超过11万亿美元的市值，并以其坚实的基本面、稳健的现金流、主导的市场地位和各自行业内强大的竞争护城河为特征。鉴于这个蓝筹股宇宙的内在质量和增长导向，行动分布中的结构性看涨偏见与这些市场领先资产的长期增值潜力是一致的。重要的是，因为Trading-R1采用多空策略进行交易，一个卖出或强烈卖出的信号意味着发起空头头寸，而不仅仅是平掉多头头寸。虽然做空带来了实际可行性的挑战，但在训练中加入它提供了一个更丰富的行动空间和更清晰的信号辨别能力，以便围绕这些基本面强劲的公司进行战术性定位。

【深度解读】 这一段详细解释了上一段中提到的“**蓝筹股训练宇宙**”的具体构成，并进一步阐述了“**结构性看涨偏见**”的合理性。

作者明确列出了他们用于训练的资产类型，这些资产几乎囊括了美国股市中最具影响力的公司：

- **科技巨头**：英伟达（NVIDIA）、微软（Microsoft）、苹果（Apple）、Meta、亚马逊（Amazon）、特斯拉（Tesla）。这些公司是当前科技革命的驱动力。
- **金融巨头**：伯克希尔·哈撒韦（Berkshire Hathaway，巴菲特的公司）、摩根大通（JPMorgan Chase）。
- **医疗巨头**：礼来（Eli Lilly）、强生（Johnson & Johnson）。
- **能源巨头**：埃克森美孚（Exxon Mobil）、雪佛龙（Chevron）。

- **市场指数ETF**：SPY（追踪标普500指数）和QQQ（追踪纳斯达克100指数），代表了整个市场的走势。

这些公司的共同特点是：**市值巨大**（总计超过11万亿美元）、**基本面坚实**、**现金流充裕**、**市场地位主导**，并且拥有强大的“**竞争护城河**”（即难以被竞争对手模仿的优势）。

基于这些资产的“**内在质量和增长导向**”，作者认为，在AI的决策系统中内置一个“**看涨偏见**”是完全合理的。因为对于这些优质资产，长期来看，向上的潜力远大于向下的风险。

最后，作者还澄清了一个重要的技术细节：模型的“卖出”或“强烈卖出”信号，不仅仅是建议你卖掉持有的股票（**平掉多头头寸**），而是建议你**做空**（**发起空头头寸**），即借入股票卖出，等价格下跌后再买回以赚取差价。虽然在现实中做空操作更复杂，但在训练中引入这个概念，可以让模型学习到更丰富的交易策略。它不仅要判断“涨不涨”，还要判断“跌不跌”，从而能更灵活地进行**战术性仓位调整**。

【原文翻译】 时间范围 我们的目标是持有期约为一周的中期策略。这个范围在可操作性和可行性之间取得了平衡：排除了高频交易（受LLM推理延迟的限制），同时避免了需要超越当前语言模型能力的宏观经济远见的长期投资。中期交易提供了一个自然的环境，在这里，结构化推理、证据基础和结果对齐可以最有效地结合起来。

【深度解读】 这一段明确了TRADING-R1的**策略定位**：它是一个**中期交易**模型，目标持有时间大约是**一周**。这个时间范围的选择是经过深思熟虑的，旨在找到一个最能发挥LLM优势的“甜蜜点”。

- **为什么不是高频交易（HFT）？** 高频交易需要在微秒级别做出决策，争分夺秒。而大型语言模型（LLM）进行一次复杂的推理需要一定的时间（**推理延迟**），这对于高频交易来说太慢了。让LLM去做高频交易，就像让一位深思熟虑的战略家去参加百米短跑比赛，完全是扬短避长。
- **为什么不是长期投资？** 长期投资（比如持有数年）的成功，往往依赖于对未来几年**宏观经济**、**技术变革**和**产业格局**的深刻洞察和远见。这种超长周期的预测，超出了目前LLM的能力范围。它们可以分析现有数据，但无法像巴菲特那样“洞见未来”。

因此，**中期交易**成为了最佳选择。在一周左右的时间尺度上：

- **结构化推理**是有意义的：公司的财报、新闻、分析师评级等信息，其影响力通常会持续几天到几周。
- **证据基础**是可行的：可以基于最近发生的事实来进行分析和决策。
- **结果对齐**是有效的：可以在相对较短的时间内看到决策的结果，从而为强化学习提供及时的反馈。

这个定位非常务实，它找到了一个能将LLM的**语言理解和逻辑推理能力**与金融市场的**时间动态**完美结合的领域。

【原文翻译】 策略优化 为了在强化学习期间优化策略，我们采用了组相对策略优化（GRPO），这是近端策略优化（PPO）的一个最新变体，它消除了对单独价值模型的需求。PPO使用一个学习到的价值函数来估计每个词元的优势，而GRPO则直接从同一输入的多个采样轨迹组中导出基线。这种相对评分稳定了训练并减少了内存开销。具体来说，对于每个输入 q，我们从旧策略 πθold 中采样 G 个候选输出 {oi}i=1G，并从奖励模型中为每个输出分配一个奖励 ri。组相对优势通过其同伴来归一化每个候选者： A^i =std(r)ri−mean(r)然后GRPO的目标是： JGRPO(θ)=Eq,{oi}i=1G∼πθold[∑i=1G∑t=1∣oi∣min(rt(i)(θ)A^i,t,clip(rt(i)(θ),1−ϵ,1+ϵ)A^i,t)] −βEq其中 rt(i)(θ):=πθold(oi,t∣q,oi,<t)πθ(oi,t∣q,oi,<t)其中 ϵ 控制裁剪，β 缩放与参考SFT模型 πref 的KL散度惩罚，而 A^i,t 是归一化的组相对优势。对于Trading-R1的训练，奖励 ri 整合了结构、证据和决策部分（S4节）。因此，每个采样的输出不仅根据其交易决策的正确性来评判，还根据其论点结构的连贯性和其主张的基础性来评判。这种整体评分与GRPO的组相对框架自然对齐。总之，GRPO提供了稳定的优化，而无需评论家模型，而我们的三阶段奖励系统提供了特定任务的塑造信号，逐步精炼结构化推理、有证据支持的主张和与市场对齐的交易决策。

【深度解读】 这一段深入到了强化学习训练的“引擎室”，解释了他们使用的具体优化算法——**GRPO (组相对策略优化)**。对于高中生来说，这里的公式可能看起来很吓人，但其核心思想其实非常直观。

首先，我们需要理解传统的**PPO (近端策略优化)** 算法是如何工作的。PPO就像一个由“演员”（Policy Model）和“评论家”（Value Model）组成的二人组。

- **演员**：负责做出决策（比如生成一段文本或一个交易指令）。
- **评论家**：负责评估“演员”在某个状态下的表现有多好，给出一个“价值”分数。 “演员”会根据“评论家”的评分来调整自己的表演。问题在于，这个“评论家”模型本身也需要训练，而且通常和“演员”模型一样大，这会消耗大量的计算资源和内存。

而**GRPO** 找到了一种更聪明的方法，它**不需要“评论家”**。它的做法是：

1. **群体表演**：对于同一个问题（输入 q），让“演员”用旧的策略（πθold）一口气生成**一组**（G个）不同的答案（输出 {oi}）。
2. **相对评分**：然后，用奖励模型给这一组答案中的每一个都打分（ri）。
3. **计算优势 (A^i)**：一个答案的“优势”不再由“评论家”来评判，而是通过**和同组其他答案的平均分进行比较**来确定。如果它的得分远高于平均分，那么它的优势就是正的；如果低于平均分，优势就是负的。这个过程就是公式 A^i=std(r)ri−mean(r) 所表达的，它计算了一个标准化的相对得分。

GRPO的目标函数 JGRPO(θ) 看起来复杂，但我们可以理解它的两个主要部分：

- **第一部分 (min(...))**：这是核心的优化项。它的目标是，对于那些“优势”为正的（即好的）答案，提高生成它们的概率；对于“优势”为负的（即差的）答案，降低生成它们的概率。`min`和`clip`函数的作用是确保每次策略更新的步子不要迈得太大，从而保持训练的稳定，这也是PPO系列算法的精髓。

- **第二部分 (−βDKL(...))**：这是一个“**惩罚项**”或“**约束项**”。DKL 衡量的是新策略（πθ）和我们之前通过SFT训练好的参考策略（πref）之间的差异。这个惩罚项就像一根“**皮筋**”，它拉着模型，防止它在追求高奖励的过程中，变得和最初学到的、结构良好的SFT模型相差太远。β 就是这根皮筋的松紧度。

最后，作者强调，用于计算优势的**奖励 ri** 是一个综合性的分数，它同时考虑了**结构、证据和决策**三个方面。这意味着，一个好的输出不仅要做出正确的交易决策，还必须附带一篇结构清晰、论据充分的分析报告。这种全面的评估方式与GRPO的相对评分机制完美结合，共同推动模型向着既专业又有效的方向进化。

4. 实验

4.1. 训练细节

【原文翻译】 训练Trading-R1涉及处理多维金融输入（2-3万词元）和生成全面的投资论点（6-8千词元）。使用LORA [Hu et al. (2021)] 的监督微调阶段在一台8×H100服务器（96GB）上进行，而强化学习阶段则使用一台8×H200服务器（141GB）。这个RL阶段增强了模型从分析推理过渡到高置信度交易决策的能力，完成了从洞察到行动的完整流程。

我们的训练投资组合包含一个经过战略性选择的大盘股资产宇宙，代表了不同的市场板块和投资工具。该投资组合专注于超大盘科技领袖，包括英伟达、微软和苹果，它们共同代表了超过11万亿美元的市值，并是现代股票市场动态的主要驱动力。除了科技股，选择范围还涵盖了通信服务（Meta）、非必需消费品（亚马逊、特斯拉）、金融（伯克希尔·哈撒韦、摩根大通）、医疗保健（礼来、强生）和能源（埃克森美孚、雪佛龙）。此外，两个主要的ETF（SPY和QQQ）分别提供了对更广泛市场贝塔和科技板块集中的敞口。这种精心策划的选择确保了Trading-R1能够遇到现代机构交易环境中特有的所有市场制度、板块动态和波动模式。按板块和市值的完整投资组合细分详见附录S3。

【深度解读】 这一节描述了实验的具体“硬件”和“软件”配置，让我们能一窥这项前沿研究所需的计算资源和数据范围。

首先，是**计算资源**。作者提到了他们使用的服务器型号：**H100**和**H200**。这些是NVIDIA公司生产的顶级AI计算芯片（GPU），是目前进行大规模AI训练的“标配”。一台服务器配备8张这样的显卡，可见训练的计算量之大。他们将训练分为两个阶段，分别在不同的硬件上进行：

- **监督微调（SFT）阶段**：在8张H100上进行。这个阶段处理的输入数据非常长（**2-3万个词元**，相当于一本小册子的长度），输出的分析报告也很长（**6-8千个词元**）。

- **强化学习（RL）阶段**：在更强大的8张H200上进行。这个阶段是训练的“攻坚”部分，需要大量的计算来模拟、评估和优化策略，目的是让模型学会从分析**过渡到高置信度的决策**。

其次，是**训练数据**，即他们选择的“**训练投资组合**”。这个组合经过了**战略性选择**，确保模型能学到足够广泛和有代表性的知识。它覆盖了美国经济的几大支柱行业：

- **科技**：英伟达、微软、苹果、Meta、亚马逊、特斯拉
- **金融**：伯克希尔·哈撒韦、摩根大通
- **医疗**：礼来、强生
- **能源**：埃克森美孚、雪佛龙
- **大盘指数**：SPY (标普500) 和 QQQ (纳斯达克100)

这个选择的目的是让模型在训练中接触到尽可能多的**市场情景**。科技股的高增长和高波动，金融股的周期性，医疗股的稳定性和政策影响，能源股与大宗商品价格的联动，以及ETF所代表的市场整体趋势，这些都为模型提供了丰富多样的学习素材。通过在这样一个“**微缩版**”的真实市场中学习，TRADING-R1才能更好地适应现实世界中复杂的交易环境。

4.2. 数据、提示和奖励结构

【原文翻译】 为了确保可比性，所有模型的输入都进行了标准化。每个提示都为给定的资产-日提供了一个结构化的市场数据、基本面、社交情绪和近期新闻标题的快照。模型被要求生成一个投资论点，然后是一个映射到五类离散行动空间（强烈卖出、卖出、持有、买入、强烈买出）的交易决策。

奖励来源于第3.5节中介绍的经波动率调整的、基于百分位的标注方案。标签根据每个资产的回报分布进行校准，反映了波动率和漂移的差异。由此产生的非对称目标分布（详见表2）既反映了经验性的股票市场行为，也反映了既定的分析师实践。通过为每个资产量身定制标签分布，我们确保了训练奖励和评估结果都是现实的、风险感知的，并与专业的金融分析保持一致。

【深度解读】 这一节阐述了实验的“游戏规则”，确保所有参赛的AI模型都在一个公平的环境下进行比赛。

首先，是**标准化的输入**。为了公平比较，所有模型收到的“考卷”（**提示**）都是一样的。这份考卷包含了在某一天（**资产-日**）可以获得的关于某只股票的所有关键信息：**市场数据**（股价图）、**基本面**（财报）、**社交情绪**和**新闻标题**。这就像一场开卷考试，所有考生都能看到同样的参考资料。

其次，是**标准化的输出要求**。所有模型都必须完成两项任务：

1. 写一篇**投资论点**（分析报告）。
2. 给出一个明确的**交易决策**，必须是**五选一**（强烈卖出、卖出、持有、买入、强烈买入）。

最后，是**标准化的评分体系**，即**奖励结构**。这个评分标准（**奖励**）来自于前面3.5节提到的那个复杂的**波动率调整标注方案**。这里的关键在于，这个评分标准是**因材施教**的（**为每个资产量身定制**）。

- **反映波动率差异**：对于像特斯拉这样波动性大的股票，和像强生这样波动性小的股票，评判“好”与“坏”的标准是不同的。
- **反映漂移差异**：对于长期增长趋势强劲的科技股，和增长平稳的公用事业股，其“向上漂移”的特性也不同。

通过这种精细化的、针对每个资产特性的奖励机制，作者确保了整个训练和评估过程是**现实的**、**考虑了风险的**，并且与专业金融分析师的思维方式**保持一致**。这避免了用一套“一刀切”的标准来衡量所有不同类型的资产，使得实验结果更具说服力。

4.3. 实验设计和评估方法

【原文翻译】 我们使用一个全面的历史回测框架，在一组精心策划的高交易量股票上评估TRADING-R1，包括苹果（AAPL）、谷歌（GOOGL）和亚马逊（AMZN），以及广泛交易的ETF如SPY。回测覆盖了2024年6月1日至8月31日，这是一个从未在训练数据中出现的持有期，反映了多样化的市场条件，并为评估泛化能力和稳健性提供了一个现实的基准。

【深度解读】 这一段描述了实验的“考场”和“考试时间”。

考场：他们选择了一些家喻户晓、交易量巨大的股票和ETF作为测试对象，比如**苹果（AAPL）**、**谷歌（GOOGL）**、**亚马逊（AMZN）**，以及代表美国大盘的**SPY**。选择这些资产是因为它们流动性好，市场关注度高，是检验交易策略有效性的理想标的。

考试时间：回测的时间段是**2024年6月1日至8月31日**。最关键的一点是，这是一个“**持有期**”（held-out period），意味着这段时间的数据**从未在训练过程中被模型看到过**。这就像高考用的是一套全新的、考生从未做过的试卷一样，是检验模型**泛化能力**（generalization）和**稳健性**（robustness）的黄金标准。

- **泛化能力**：模型是否只是“死记硬背”了训练数据中的规律，还是真正学会了能够应用于新情况的分析能力。
- **稳健性**：模型在面对与训练期间不同的**多样化市场条件**时，表现是否依然稳定。

这个严格的实验设计确保了评估结果的客观性和可信度，避免了所谓的“**过拟合**”（overfitting）——即模型在训练数据上表现完美，但在真实世界中一败涂地。

【原文翻译】 **基线模型** 我们将Trading-R1与广泛的基于LLM的分析工具进行比较，涵盖了小型、中型和大型模型类别。对于小型语言模型（SLMs），我们评估了QWEN-4B、GPT-4.1-NANO和GPT-4.1-MINI。对于较大型的LLM，我们包括了GPT-4.1、LLAMA-3.3、LLAMA-SCOUT和QWEN3-32B。对于经过强化学习增强的模型（RLMs），我们考虑了DEEPSEEK、O3-MINI和O4-MINI。此外，我们对Trading-R1的变体进行了消融研究：一个用SFT热启动进行初始化，另一个仅用RL进行训练，以更好地理解每个训练阶段对整体性能的贡献。

【深度解读】 这一段介绍了参加这次“交易大赛”的各位“选手”，也就是用来和TRADING-R1进行比较的**基线模型 (Baseline Models)**。为了全面地评估TRADING-R1的实力，作者邀请了来自不同“门派”和“重量级”的选手。

这些选手可以分为几类：

1. **小型语言模型 (SLMs)**：轻量级选手，如**QWEN-4B**、**GPT-4.1-NANO**。它们参数量较小，推理能力有限，相当于“初学者”。
2. **大型语言模型 (LLMs)**：重量级选手，如**GPT-4.1**、**LLAMA-3.3**。这些是强大的通用模型，知识渊博，但没有经过专门的金融训练，相当于“知识渊博的业余爱好者”。
3. **强化学习增强模型 (RLMs)**：专业选手，但专业不对口。如**DEEPSEEK**、**O3-MINI**。这些模型经过了强化学习的训练，在编码、数学等推理任务上非常强大，但它们的“专业”不是金融，相当于“跨界参赛的数学或编程冠军”。

除了这些“外部对手”，作者还进行了一场“内部比赛”，即**消融研究 (ablation study)**。这是一种科学研究中常用的方法，通过“去掉”系统的某个部分，来观察这个部分到底有多重要。他们测试了两个“残血版”的TRADING-R1：

- **仅SFT版**：只经过监督微调，没有经过强化学习。这相当于一个只会“纸上谈兵”的理论家。
- **仅RL版**：只经过强化学习，没有经过SFT的结构化训练。这相当于一个只顾“实战”但缺乏理论基础的莽夫。

通过将完整版的TRADING-R1与所有这些外部和内部的对手进行比较，作者可以非常清晰地证明：

- TRADING-R1比通用模型和专业不对口模型更强。

- TRADING-R1的成功，离不开SFT和RL这两个阶段的**协同作用**，缺一不可。

【原文翻译】 **评估指标** 我们使用标准的金融指标来评估模型性能，这些指标既捕捉了盈利能力也捕捉了风险特征。我们的评估框架包括**累积回报率（CR）**来衡量总回报，**夏普比率（SR）**来评估风险调整后的表现，**命中率（HR）**来评估预测准确性，以及最大回撤（MDD）**来量化下行风险。这些指标为评估交易策略在不同市场条件下的有效性提供了全面的评估。所有指标的详细数学定义和计算过程在附录S2中提供。

【深度解读】 这一段介绍了他们用来评判所有AI选手表现的“**评分标准**”，即**评估指标 (Evaluation Metrics)**。在金融领域，评价一个交易策略的好坏，绝不能只看它赚了多少钱，还必须看它承担了多大的风险。因此，作者选择了一套全面且专业的指标。

1. **累积回报率 (Cumulative Return, CR)**：这是最直观的指标，就是“**总共赚了（或亏了）多少钱**”。比如，CR为10%，意味着期初投入100元，期末变成了110元。
2. **夏普比率 (Sharpe Ratio, SR)**：这是**最重要的指标**，衡量的是“**性价比**”，即**每承担一单位风险，能获得多少超额回报**。一个策略可能回报率很高，但过程像过山车一样惊险（高风险）；另一个策略回报率稍低，但过程非常平稳（低风险）。夏普比率可以告诉你哪个策略更“聪明”。**夏普比率越高，说明策略在控制风险的同时获取回报的能力越强**。
3. **命中率 (Hit Rate, HR)**：这很好理解，就是“**预测的准确率**”。它衡量的是模型预测股价上涨或下跌的方向，猜对了多少次。比如，命中率为60%，意味着模型做出的100次方向性预测中，有60次是正确的。
4. **最大回撤 (Maximum Drawdown, MDD)**：这是一个衡量“**抗压能力**”或“**最惨时有多惨**”的指标。它记录了在整个回测期间，投资组合净值从最高点跌到最低点的最大跌幅。例如，MDD为-20%，意味着在最糟糕的情况下，你的账户曾经从最高时的100万跌到过80万。**这个值越小越好**，说明策略的下行风险控制得很好，不会让投资者经历巨大的心理压力。

通过这四个维度的综合评估，可以全面地判断一个交易策略的优劣，而不仅仅是片面地看重总收益。

【原文翻译】 **回测模拟** 我们采用一个基于历史市场数据的标准回测设置，为Trading-R1收集多维输入，如每日新闻、价格数据和派生指标，类似于Tauric-TR1-DB。交易仅使用截至每个交易日可用的信息来执行，消除了前视偏差，并确保在完全样本外的环境中进行严格的因果评估。这种受控设计隔离了模型质量对交易性能的影响。

【深度解读】 这一段描述了他们如何确保“考试”的公平性和严格性，即**回测模拟 (Backtesting Simulation)** 的具体设置。

核心原则是**杜绝一切形式的“作弊”**。

- **标准化的数据输入**：回测时提供给模型的数据，和训练时使用的Tauric-TR1-DB数据格式类似，包含了新闻、价格等多维度信息。
- **杜绝“未来函数”（前视偏差, look-ahead bias）**：这是回测中最关键、也最容易犯的错误。作者强调，在模拟的任何一个交易日，模型做决策时，**只能使用那个时间点之前（截至当天）可以获得的信息**。绝不能让模型用到未来的信息来做当下的决策。这就像你不能用明天的彩票中奖号码来买今天的彩票一样。这确保了回测模拟的是一个真实的、充满不确定性的决策环境。
- **严格的样本外测试**：整个回测是在一个模型从未见过的“**样本外**”（out-of-sample）时间段上进行的。这再次确认了实验的目的是测试模型的真实泛化能力。

通过这种**受控设计**，实验能够**隔离**出**模型质量**这唯一的变量。如果一个模型在回测中表现好，那就可以很有信心地说，这是因为它本身的能力强，而不是因为它运气好，或者是因为实验设计有漏洞让它“偷看”到了答案。这使得最终的评估结果具有很高的科学可信度。

5. 结果

【原文翻译】 我们的实验结果清晰地展示了不同模型类别在交易性能上的层次结构。表3和表4展示了TRADING-R1的性能指标。

【深度解读】 这是结果部分的开场白，预告了接下来将要展示的数据。作者在这里用了一个词——“**层次结构**”（hierarchy），这暗示了不同类型的AI模型在金融交易这个任务上的表现，存在着明显的、类似金字塔一样的等级差异。接下来的表格将会用数据来证明这个“金字塔”的结构，并展示TRADING-R1位于塔尖的位置。

【表格3：模型在NVDA, AAPL, MSFT上的性能。粗体绿色值表示最佳性能，下划线绿色值表示次佳性能】

类别	模型	CR(%)	SR	NVDA HR(%)	MDD(%)	CR(%)	SR	AAPL HR(%)	MDD(%)	CR(%)	SR	MSFT HR(%)	MDD(%)
SLM	Qwen-4B	-1.59	-1.62	52.2	2.80	-0.81	-0.92	41.7	3.76	-1.45	-1.28	50.0	4.38
	GPT-4.1-nano	0.76	-0.09	56.0	3.82	0.44	-0.31	51.9	3.52	-0.01	-0.95	39.3	1.60
	GPT-4.1-mini	0.29	-0.53	58.8	2.47	-2.14	-1.92	40.0	3.69	-2.34	-1.74	27.3	4.00
LLM	GPT-4.1	3.15	0.85	65.5	2.81	4.02	1.24	50.0	2.89	2.30	0.97	63.9	1.92

类别	模型	CR(%)	SR	NVDA HR(%)	MDD(%)	CR(%)	SR	AAPL HR(%)	MDD(%)	CR(%)	SR	MSFT HR(%)	MDD(%)
	LLAMA-3.3	0.65	-0.16	62.2	2.78	6.73	1.78	63.6	2.40	1.58	0.54	58.1	1.59
	LLAMA-Scout	-1.96	-1.64	31.8	2.90	2.03	0.58	59.4	3.21	-0.29	-1.33	36.8	1.44
	Qwen3-32B	1.74	0.27	64.5	2.80	0.62	-0.12	33.3	3.39	2.14	1.29	65.6	0.82
RLM	DeepSeek	-0.79	-0.66	50.0	3.66	0.68	-0.13	55.3	4.78	-0.38	-1.01	33.3	2.06
	O3-mini	-2.97	-1.48	46.9	5.33	-1.89	-1.13	50.0	3.72	1.19	0.15	47.4	1.19
	O4-mini	-0.99	-0.83	43.2	3.61	-3.19	-1.36	50.0	7.88	-1.72	-1.77	48.5	2.35
Ours	Supervise Finetune	7.42	2.72	72.5	2.01	-2.37	-1.27	45.2	5.20	-0.24	-0.64	56.1	3.87
	Reinforcement Learning	3.27	1.25	62.5	2.73	4.04	1.14	57.1	3.02	-0.18	-0.81	45.7	1.66
	TRADING-R1	**8.08**	**2.72**	70.0	3.80	**5.82**	**1.80**	63.6	3.68	**2.38**	**0.87**	60.4	1.90

【表格4：模型在AMZN, META, SPY上的性能。粗体绿色值表示最佳性能，下划线绿色值表示次佳性能】

类别	模型	CR(%)	SR	AMZN HR(%)	MDD(%)	CR(%)	SR	META HR(%)	MDD(%)	CR(%)	SR	SPY HR(%)	MDD(%)
SLM	Qwen-4B	-2.90	-1.13	46.2	6.05	1.32	0.14	51.7	3.80	-1.33	-3.37	42.3	1.71
	GPT-4.1-nano	-4.88	-2.34	40.7	6.20	-3.07	-1.69	47.8	5.19	0.04	-1.23	47.6	1.38
	GPT-4.1-mini	2.24	0.81	50.0	2.01	1.21	0.16	56.5	1.70	-1.03	-2.47	43.5	1.44
LLM	GPT-4.1	3.80	1.15	64.3	2.44	5.63	1.59	68.8	1.91	0.35	-0.74	43.3	1.21
	LLAMA-3.3	-0.89	-0.61	58.6	6.02	3.21	1.01	62.5	2.55	1.27	0.27	64.7	1.35
	LLaMA-Scout	-3.47	-1.48	35.7	5.95	3.51	0.92	53.1	2.78	-1.34	-3.36	36.0	1.65
	Qwen3-32B	5.61	2.12	64.3	1.89	-1.23	-0.58	46.2	6.61	2.32	1.87	70.4	0.65
RLM	DeepSeek	-1.15	-1.14	50.0	3.00	1.26	0.12	40.5	2.80	-1.15	-1.82	36.4	2.00
	O3-mini	-3.15	-1.37	38.2	5.50	2.05	0.53	73.1	2.64	0.80	-0.25	57.6	0.62
	O4-mini	-2.48	-1.28	51.6	4.83	-0.45	-0.80	53.6	2.68	-0.30	-1.34	36.8	1.72
Ours	Supervise Finetune	1.93	0.36	60.6	4.28	2.52	0.54	55.9	2.93	1.78	0.86	58.1	1.15
	Reinforcement Learning	-0.05	-0.29	52.5	4.84	-0.18	-0.36	44.4	5.11	1.85	1.00	67.6	0.69
	TRADING-R1	**5.39**	**1.72**	63.0	3.20	**5.12**	**0.86**	50.0	4.65	**3.34**	**1.60**	64.0	1.52

【原文翻译】 **小型语言模型（SLMs）** 表现最弱，由于其有限的参数容量和浅层推理，难以实现盈利，这导致了不稳定的分析、薄弱的论证和整体决策质量差。**推理语言模型（RLMs）** 相对于SLM取得了适度的改进，但面临重大挑战：它们有限的指令遵循能力有时会阻止它们以所需格式产生决策，而且它们冗长的推理路径常常偏离与市场相关的数据。**大型语言模型（LLMs）** 优于这两个类别，即使没有领域特定的训练，也表现出更强的连贯性和决策质量。有趣的是，尽管它们具有先进的推理能力，现成的RLM在交易任务上通常表现不如LLM。这种表现不佳源于它们无引导的推理过程，这些过程可能偏离金融分析，导致输出不集中。相比之下，Trading-R1系列（SFT、RFT和完整的Trading-R1）凸显了专业训练的重要性：SFT强制执行专业的输出格式和一致的决策模式，而RFT则逐步将推理与市场结果对齐。

【深度解读】 这一段是对表格数据的分析和解读，清晰地揭示了作者所说的“**性能层次结构**”。

1. **垫底层：小型语言模型 (SLMs)**
    - **表现**：几乎在所有股票上都是亏损的（CR为负），夏普比率（SR）惨不忍睹。
    - **原因**：它们的“脑容量”（**参数容量**）太小，**推理能力太浅**。面对复杂的金融数据，它们无法进行深入分析，做出的决策质量很差。这就像让一个小学生去做微积分，力不从心。

2. **中间层：推理语言模型 (RLMs)**
    - **表现**：比SLM稍好，但大多还是亏损或勉强持平。
    - **原因**：它们有两个主要问题。一是**指令遵循能力有限**，有时甚至无法按要求格式输出决策。二是它们的推理过程虽然强大，但由于是为数学、编程等领域优化的，在金融领域容易“跑偏”（**偏离与市场相关的数据**），导致分析不聚焦。这就像一个数学天才，你让他分析财报，他可能会陷入对数字的过度解读，而忽略了商业常识。
3. **上层：大型语言模型 (LLMs)**
    - **表现**：明显优于前两者，即使没有经过任何金融训练，也能取得正向的回报和夏普比率。比如GPT-4.1在NVDA、AAPL、MSFT上都表现不错。
    - **原因**：它们强大的通用知识和推理能力，使其能够更好地理解和处理复杂的金融信息，表现出更强的**连贯性和决策质量**。
4. **顶层：TRADING-R1 系列**
    - **表现**：全面超越所有其他模型。无论是只用SFT的版本，还是最终的完整版TRADING-R1，在关键指标（尤其是风险调整后的夏普比率SR）上都名列前茅。
    - **原因**：这证明了**专业训练**的巨大价值。SFT教会了模型专业的**输出格式**和**决策模式**，而RFT则将模型的推理能力与真实**市场结果**对齐。

一个特别值得注意的发现是，**专门为推理训练的RLM，在金融交易这个任务上的表现，竟然不如通用的LLM**。这揭示了一个深刻的道理：金融分析不是一个纯粹的、逻辑严密的推理任务，它充满了噪音和不确定性。RLM那种“一根筋”式的、为寻找唯一正确答案而优化的推理模式，在这种环境下反而不适应。而通用LLM更广泛的知识和更灵活的思维方式，使其表现得更好。TRADING-R1的成功，正是在通用LLM的强大基础上，通过专业的SFT和RFT训练，为其注入了金融领域的“灵魂”。

【原文翻译】 我们认为这种趋势反映了训练重点的差异。通用LLM在指令调优期间接触了大量多样的用户指令，因此在如何解决问题上保持了灵活性和开放性。相比之下，RLM最近被优化用于狭窄的领域，如编码、数学和科学推理。这种专业化在那些领域产生了强大的性能，但限制了泛化能力，使它们在金融推理任务上效果较差。尽管金融与数学和科学有重叠之处，但金融数据在关键方面有所不同：它充满噪声、模棱两可，并充满了混合信号，这使得定义逐步、可验证的奖励变得困难。因此，纯粹基于RFT的方法在金融LLM训练中无法直接使用。我们的Trading-R1通过结合两种范式的优势来解决这些挑战。通过SFT，它整合了结构化的论点写作和一致的决策模式，而RFT则逐步加强特定阶段的行为。这种设计稳定了推理，防止了漂移，并实现了连贯的、与市场对齐的交易决策。

【深度解读】 这一段深入剖析了上一段中那个有趣的发现：为什么为推理而生的RLM反而输给了通用的LLM？作者给出了一个非常精辟的解释，核心在于**训练焦点的差异**。

- **通用LLM (如GPT-4.1)**：它们的训练过程（**指令调优**）就像是和一个见过“三教九流”各种问题的“万事通”聊天。它们被训练来处理各种各样、千奇百怪的用户请求，因此它们的思维方式是**灵活的、开放的**。它们没有被预设一个固定的解题套路。
- **推理语言模型 (RLM, 如DeepSeek)**：它们是“**专才**”，被专门优化来解决**狭窄领域**的问题，比如数学、编程。在这些领域，问题通常有明确的定义、清晰的逻辑链条和唯一的正确答案。这种训练使得它们在这些特定领域表现超群，但也牺牲了**泛化能力**。

作者指出，金融领域虽然也需要数学和逻辑，但其本质与数学题完全不同。金融数据**充满噪声、模棱两可、信号混杂**。一个利好消息和一个利空消息可能同时出现，一个看似上涨的图表可能突然反转。在这种混乱的环境下，很难定义一个清晰的、一步一步的“正确”推理路径，并给予**可验证的奖励**。

因此，RLM那种为确定性问题设计的、严谨但僵化的推理模式，在金融这个充满不确定性的“沼泽”中反而寸步难行。而TRADING-R1的成功之道在于**取长补短**：

- 它利用**SFT**，吸收了通用LLM的灵活性，并为其套上了一个**结构化的分析框架**，防止其思维“天马行空”。
- 它利用**RFT**，但不是一步到位地追求最终结果，而是**逐步地**、分阶段地强化其行为，使其在保持结构的同时，慢慢向市场看齐。

这种设计**稳定了推理过程**，**防止了思维漂移**，最终实现了在混乱的金融市场中做出**连贯且与市场一致**的决策。

【原文翻译】 我们的Trading-R1方法通过结合SFT和RFT来有效捕捉市场动态，从而取得了最强的整体性能。在所有评估的资产上，TRADING-R1都比基线模型有所改进。它在NVDA上实现了1.88的夏普比率和8.08%的回报，在AAPL上以1.80的夏普比率胜过GPT-4.1的1.24，同时保持了更低的回撤（3.68% 对比 RLM的7.88%）。该模型还获得了领先的命中率，包括在NVDA上的70.0%和在SPY上的64.0%。相比之下，像Qwen-4B和GPT-4.1-nano这样的小型LLM通常产生负的夏普比率，而像O3-mini和O4-mini这样的RLM由于无引导的推理过程而招致重大损失。总体而言，性能层次结构 (SLM<RLM<LLM<Trading−SFT≈Trading−RFT<Trading−R1) 强调了模型规模和专业化推理在算法交易中的重要性，我们的方法在盈利能力和风险管理之间取得了最佳平衡。

【深度解读】 这一段用具体的数字总结了TRADING-R1的胜利。作者通过列举关键资产上的核心指标，直观地展示了其优势。

- **在英伟达（NVDA）上的表现**：TRADING-R1取得了**8.08%的回报率**，同时**夏普比率高达1.88**（在金融领域，夏普比率超过1就相当不错，接近2已经非常优秀了）。命中率也达到了**70.0%**。
- **在苹果（AAPL）上的表现**：TRADING-R1的**夏普比率为1.80**，显著高于表现最好的基线模型GPT-4.1的1.24。这意味着TRADING-R1在苹果这只股票上，以更小的风险获得了更高的回报。
- **风险控制能力**：TRADING-R1的最大回撤（MDD）控制得很好，例如在AAPL上是3.68%，而某些RLM模型的回撤高达7.88%，后者会让投资者体验到更剧烈的资金回撤。
- **在市场指数（SPY）上的表现**：命中率达到**64.0%**，显示了其对市场整体走势也有不错的判断力。

与此形成鲜