AlphaFin：使用检索增强股票链框架对财务分析进行基准测试

Xiang Li[1]*, Zhenyu Li[1]*, Chen Shi[2]*,
Yong Xu[1], Qing Du[1]†, Mingkui Tan[1], Jun Huang[2], Wei Lin[2]
[1]South China University of Technology, China
[2]Alibaba Group, China
{ lixiangjacky, zhenyuli2148 } @gmail.com, deling.sc@alibaba-inc.com

Abstract

The task of financial analysis primarily encompasses two key areas: stock trend prediction and the corresponding financial question answering. Currently, machine learning and deep learning algorithms (ML & DL) have been widely applied for stock trend predictions, leading to significant progress. However, these methods fail to provide reasons for predictions, lacking interpretability and reasoning processes. Also, they can not integrate textual information such as financial news or reports. Meanwhile, large language models (LLMs) have remarkable textual understanding and generation ability. But due to the scarcity of financial training datasets and limited integration with real-time knowledge, LLMs still suffer from hallucinations and are unable to keep up with the latest information. To tackle these challenges, we first release AlphaFin datasets, combining traditional research datasets, real-time financial data, and handwritten chain-of-thought (CoT) data. It has a positive impact on training LLMs for completing financial analysis. We then use AlphaFin datasets to benchmark a state-of-the-art method, called Stock-Chain, for effectively tackling the financial analysis task, which integrates retrieval-augmented generation (RAG) techniques. Extensive experiments are conducted to demonstrate the effectiveness of our framework on financial analysis.

Keywords: Large Language Models, Retrieval-Augmented Generation, Chain-of-Thoughts, Finance, Stock Trend Prediction, Financial Question Answering

1. 介绍

随着金融业的进步，金融分析的重要性日益凸显。财务分析能力主要体现在股票趋势预测和相应的财务 Q & A 方面。LLM 的出现引起了金融界的关注，因为它们具有非凡的生成能力（**???**）。因此，人们强烈希望利用这些 LLM 来提高财务分析的准确性。[1] 最近的许多研究都试图使用 ML & DL 创建有效的算法来预测股票趋势（**??**）. 目前，ML & DL 已广泛用于基于时间序列数据的股票趋势预测，对行业产生了积极影响。然而，ML & DL 算法的性能有限，只能提供不确定的结果，无法处理复杂的文本数据。同时，他们未能为投资者提供有效的理由和分析根本原因，可能会破坏他们的投资信心。如图 **??** 所示，对于苹果公司来说，ML & DL（如 LSTM）可以根据以前的股价数据预测下个月的不确定股票趋势（”上涨”）。然而，它无法提供可靠的结果和预测分析。如果他们能提供有效的分析，将大大增强投资者对决策的信心。

幸运的是，LLM 在文本处理和生成方面具有出色的能力。为了利用 LLM 的能力，FinGPT（**?**）和 BloombergGPT（**?**）被专门设计为 FinLLM。它们可以应用于各种财务任务，以满足行业的需求。如图 **??** 所示，它们具有处理各种文本数据的潜力，包括新闻和报道（**?**）。这一进步使投资者能够做出精确的投资和交易决策。

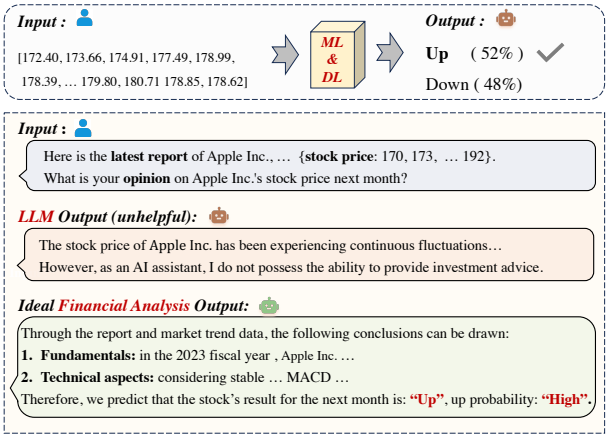

Figure 1: 传统的 ML & DL 方法仅提供不确定的预测（向上/向下），而原始 LLM 可以提供对预测的分析，但无济于事。

然而，构建 FinLLMS 并不是一项简单的任务。LLM 经常会出现幻觉和无意义输出（**?**）等现象，即使是高级 ChatGPT（**?**）。如图 1 所示，通用 LLM 生成的内容缺乏实用性，无法满足实时需求。这可以归因于两个原因。首先，生成内容的质量取决于数据的可用性。缺乏高质量的金融训练数据集（**?**）影响了生成的质量。其次，股票趋势预测依赖于精确和实时的信息，这些信息的缺失会导致 LLMs 幻觉。尽管 RAG（**?**）在其他领域得到了广泛的应用，但它们在金融领域的适应性仍然不足。

为了应对这些挑战，在本文中，我们将财务分析的任务正式化，并发布了 AlphaFin，用于微调

---

* Equal contribution. This work was conducted when Xiang Li and Zhenyu Li were interning at Alibaba.

† Corresponding author.

[1]资源可在以下网址公开获取：https://github.com/AlphaFin-proj/AlphaFin

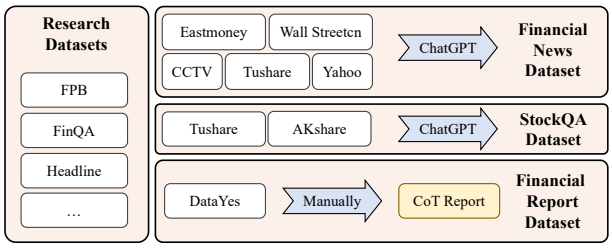

Figure 2: 所提出的 AlphaFin 数据集的数据源和预处理。

FinLLM，其中包含传统的研究数据集、实时财务数据和手写的 CoT 数据。此外，我们提出了一个与 RAG 集成的 Stock-Chain 框架。Stock-Chain 不仅为投资者提供股票趋势预测，还通过 RAG 整合实时市场数据和宏观经济新闻，在与投资者的互动中实现准确的股票分析。

实验结果表明，Stock-Chain 能够以最先进的精度和超过 30 % 的年化收益率（ARR）完成股票趋势预测的任务。同时，Stock-Chain 可以提供全面的金融 Q & A 分析，增强投资者的决策信心，为他们的投资选择提供坚实的基础。我们进行了广泛的补充实验，如消融研究、GPT4 & 人类偏好评估和案例研究。

总而言之，贡献在于四个方面：

- We formally define the task of financial analysis, which aims to accomplish stock trend prediction and the corresponding financial Q & A.
- We propose AlphaFin datasets, which contains traditional research datasets, real-time financial data, and handwritten CoT data, enhancing LLMs's ability in financial analysis.
- We fine-tune a StockGPT based on AlphaFin datasets and integrate it to a Stock-Chain framework, which is further integrated with a real-time financial database through RAG. By integrating with RAG, we address the issue of the hallucination of LLMs's output and LLMs's inability to generate real-time content.
- We conduct extensive experiments on the AlphaFin datasets, to reveal that Stock-Chain outperforms all the baseline methods, and shows effectiveness for financial analysis.

2. AlphaFin 数据集

我们发布 AlphaFin 数据集，如图 **??** 所示，包括四个部分，研究数据集、StockQA、财经新闻和财报。AlphaFin 来自十几个数据源。从表 **??** 可以看出，传统研究数据集的标签长度相对较短，这阻碍了 FinLLMs 的训练。因此，AlphaFin 解决了传统研究数据集中质量低和长度低的问题。在本节中，我们提供了其来源和构建过程的详细信息。

2.1. 数据源

- ***Research datasets*** : This part includes traditional financial datasets from the academic, including FPB (**?**), FinQA (**?**), convFinQA (**?**) and Headline (**?**), etc. which enhance the information extraction and summarization ability for LLMs.
- ***StockQA dataset*** : This part encompasses stock price and other financial data from Tushare (**?**) and AKshare (**?**). It utilizes sequential data format, such as the real-world stock price trend (e.g. $\{..., 170, 173, 171, 175, 173, 170, ...\}$).
- ***Financial News dataset*** : To provide real-world financial knowledge for LLMs, we incorporate online news sources, such as the financial sections of CCTV, and Wall Street CN.
- ***Financial reports dataset*** : We build financial report datasets via DataYes (**?**), including professional analysis and knowledge of companies conducted by institutions.

2.2. 数据预处理

如图 **??** 所示，我们探讨了 AlphaFin 预处理的细节：

- ***Research dataset*** : Traditional research datasets are primarily in English and of substantial quantity. To enhance the LLMs's ability in Chinese and ensure quality fine-tuning, we only sample a portion from the source.
- ***StockQA dataset*** : Given the source data is presented in sequential format, we utilize ChatGPT with the following prompt, to generate financial questions upon it. `Based on the ..., give me a good financial question. Input: <sequential data>, Output: <Question>.` This can facilitate training and enhance the diversity of questions. Subsequently, we use ChatGPT to generate responses and obtain Q & A pairs for training LLMs.

Dataset	Size	Input	Label	Type
Research	42,373	712.8	5.6	en
StockQA	21,000	1313.6	40.8	zh
Fin. News	79,000	497.8	64.2	zh
Fin. Reports	120,000	2203.0	17.2	zh
Fin. Reports CoT	200	2184.8	407.8	zh

Table 1: AlphaFin 数据集的详细信息。"Input" 和"Label" 表示其文本长度。

- ***Financial News dataset*** : We leverage ChatGPT to extract a summary for each news, and construct the financial news dataset. This process improves LLMs's ability to generate summaries for financial news.
- ***Financial reports dataset*** : We manually align the financial reports for the companies and their stock price on the day of report publication, and use the following template to generate the final data. `According to ... give a clear answer up or down. Input: <reports & stock price>, Output: <Up/down>.`

  Furthermore, we manually create 200 financial reports CoT data with professional financial knowledge and longer labels, to provide the LLMs with progressive analytical ability. The output format is:

  ```
  According to ... conclusions can be drawn: 1. Fundamentals: ... 2. Technical aspects: ... Therefore, we predict the ... is <up/down>, probability: <Prob>
  ```

3. 存量链框架

我们将财务分析任务视为两个对应任务，股票趋势预测和相应的财务 Q & A。因此，我们提出的 Stock-Chain 框架分为两个阶段，如图 **??** 所示。在本节中，我们首先将任务形式化，然后介绍两个阶段的细节。

3.1. 问题定义

对于第一阶段，给定一组公司 $C = \{c_i\}_{i=1}^{N}$ 和相应的知识文档 $D = \{d_j\}_{j=1}^{M}$ ，我们可以预测股票趋势：

$$Pred_i = \phi(c_i, d_j),\ \ Pred_i \in \{up, down\} \tag{1}$$

其中 ϕ 表示股票预测系统，d_j 作为公司 c_i 的相关文档进行检索。目标是选择预计会上涨的公司子集 C_{chosen} 。

$$C_{chosen} = \{c_i | c_i \in C \wedge Pred_i = up\} \tag{2}$$

在第二阶段，我们将多轮对话会话视为两个对话者之间的几个查询-响应对的序列。我们将 Q_t 和 R_t 表示为当前时间步长 t 的用户查询和代理响应，$H_t = [Q_0, R_0, ..., Q_{t-1}, R_{t-1}]$ 表示对话历史记录。然后，我们将财务 Q & A 任务形式化为根据当前查询、对话历史记录和相应文档获取响应：

$$R_t = \pi(d_k, H_t, Q_t) \tag{3}$$

其中 π 表示会话系统，d_k 是检索到的与 Q_t 相关的文档。

3.2. 第一阶段：股票趋势预测

如图 **??** 左图所示，我们第一阶段是股票趋势预测。给定一个公司 c_i 和相应的文档 d_j ，这个阶段通过结合 LLM 和 AlphaFin 数据集来维护一个股票预测系统 ϕ ，为 c_i 提供股票趋势预测 $Pred_i$ 。

3.2.1. 知识处理

如图 **??** ① 所示，给定一个公司 c_i ，我们首先检索 d_j 它的相关文档。然后，我们设计一个提示模板 $Prompt_1$ 如下：

```
Please predict the rise and fall of the stock next month based on the research reports and data provided below. Please provide a clear answer, either ``up" or ``down". <report><market data>
```

其中 `<report>` 和 `<market data>` 组成 d_j 。最终，我们将提示与文档连接起来，以获取 LLM 的输入 I_i

$$I_i = concat(Prompt_1, d_j) \tag{4}$$

3.2.2. StockGPT 微调

如图 **??** ② 所示，我们设计了一个叫做 StockGPT 的 LLM 的微调过程，它包括两个步骤。首先，我们利用 AlphaFin 的所有财务报告数据集进行训练。在第二步中，我们利用手动创建的报表 CoT 数据集来指导模型逐步思考。StockGPT 的所有微调过程都使用低秩适应（LoRA）(**?**) 方法。通过对这两个步骤的微调，我们得到了 $StockGPT_{stage1}$ ，能够更准确地预测基于 d_j 的 c_i 趋势，并提供详细的分析和解释。

3.2.3. 预测和后处理

如图 **??** ③ 所示，给定输入 I_i ，我们利用 StockGPT 来预测股票的涨跌，可以看作是二元分类任务。通过输入 $StockGPT_{stage1}$ 输入 I_i ，我们得到 Res_i 大约 c_i 的响应文本。

Res_i 的格式可以称为第 2.2 节。`<Prob>` 是以下类别：$\{very\ large,\ large,\ medium\ to\ upper,\ average\}$ ，这可能是投资者决策的补充信息。

$$Res_i = StockGPT_{stage1}(I_i) \tag{5}$$

然后，我们手动提取预测结果 $Pred_i$ Res_i 。最后，我们选择所有预测为 “上涨” 的股票作为 C_{chosen} 。

$$Pred_i = \begin{cases} up, & if\ \text{“}up\text{”} \in Res_i \\ down, & else \end{cases} \tag{6}$$

$$C_{chosen} = \{c_i | Pred_i = \text{“}up\text{”}\} \tag{7}$$

此外，我们实施这种投资策略是按月滚动的。每个月，对于 C_{chosen} 的所有 c_i ，我们在整个月都会举行。投资组合中每只股票的比例是通过市值加权法计算的。

$$AR_m = AR_{m-1} + \sum_{c_i \in C_{chosen}} w * R_{c_i} \tag{8}$$

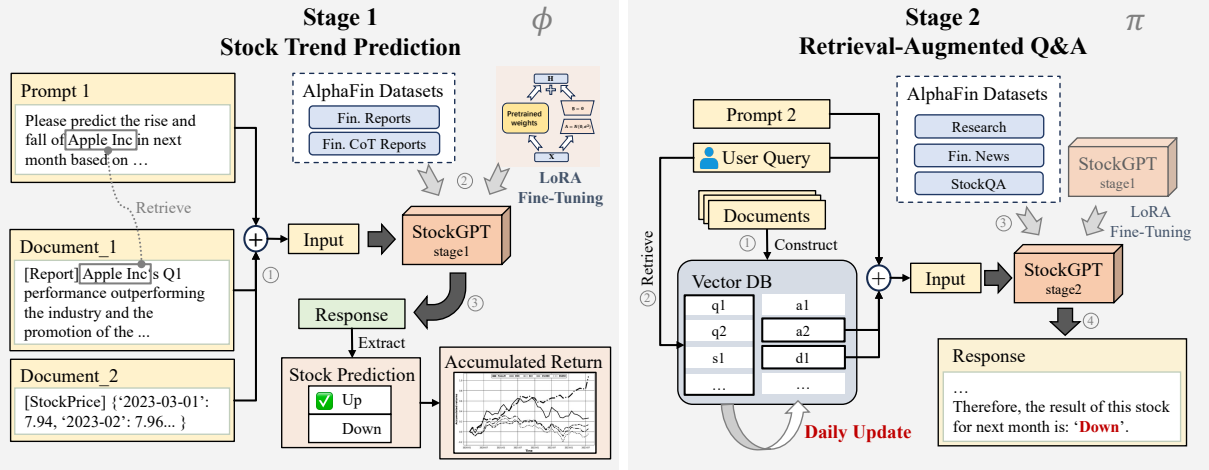

Figure 3: 财务分析中两个阶段的股票链框架的图示。

其中 AR_m 是第 m 个月的累计回报，R_{c_i} 是股票 c_i 的回报。w 表示股票 c_i 在投资组合中的比例。v_i 是公司 c_i 的市值。

$$w = \frac{v_i}{\sum_{n=1}^{N} v_n} \tag{9}$$

3.3. 第 2 阶段：金融 Q & A

除了股票趋势预测外，拟议的 Stock-Chain 还具有金融 Q & A 的能力，这对投资者来说可能更具建设性。

给定对话历史 H_t 、用户查询 Q_t 和检索到的文档 d_j 与 Q_t 相关，对话系统 π 可以给出响应 R_t 。我们采用 RAG 来增强 LLM 的 Q & A 能力，通常包括向量数据库构建、知识检索和响应生成三个部分。

3.3.1. 矢量数据库构建

如图 **??** 阶段 2 ① 所示，vector DB 是 RAG 的重要组成部分，用于知识文档的高效存储和检索。

知识提取 为了提高文档检索的准确性和效率，我们从文档中提取关键知识。我们采用两种提取策略：使用 ChatGPT 进行粗粒度文档级总结，以及通过 RefGPT (**?**) 生成细粒度的实体级对话。对于文档 d_k ，两种策略的提取过程如下：

$$s_k = ChatGPT(d_k) \tag{10}$$

$$(q_{k0}, a_{k0}), (q_{k1}, a_{k1}), ... = RefGPT(d_k) \tag{11}$$

其中 s_k 表示 d_k 的摘要，$(q_{k\_}, a_{k\_})$ 是生成的对话的查询-答案对。例如，对于关于 k 线 d_k ，$q_{k\_}$ 可以是 "k 线的含义是什么?

知识嵌入 我们以 ChatGPT 的提取策略为例。给定摘要 s_k ，我们通过句子嵌入模型获得嵌入向量 e_{s_k} 。该向量将存储在数据库中以供后续检索。

$$e_{s_k} = SentEmbed(s_k) \tag{12}$$

其中 $SentEmbed$ 是句子嵌入模型，例如 BGE (**?**) 和 SGPT (**?**) 。我们采用 BGE 作为框架中的嵌入模型。

持续更新 最后，我们构建了一个包含报告、市场数据和财经新闻的向量数据库。数据库中的知识文档可以通过在线数据回流不断更新，以保持知识的实时性。

3.3.2. 知识检索

为了检索向量数据库中的知识，用户查询 Q 也会被输入到同一个句子嵌入模型中，以获得嵌入向量 e_Q 。

$$e_Q = SentEmbed(Q) \tag{13}$$

我们选择与查询具有最高余弦相似度的文档，因为外部知识有助于 StockGPT 生成响应。

$$d^* = \arg\max_{d_k} \frac{e_Q^\top \cdot e_{s_k}}{|e_Q||e_{s_k}|} \tag{14}$$

s_k 和 d_k 可以被 $q_{k\_}$ 和 $a_{k\_}$ 取代，用于 RefGPT 的提取策略。

3.3.3. LLMs 微调

我们继承 $StockGPT_{stage1}$ 作为本部分的基础 LLM，然后继续在 AlphaFin 的研究数据集、财经新闻和 StockQA 数据集上训练 $StockGPT_{stage1}$ ，以获得 $StockGPT_{stage2}$ 。

3.3.4. 响应生成

给定对话历史 H_t 、用户查询 Q_t 和检索到的与 Q_t 相关的文档 d^* ，目标是在回合 t 的对话中给出响应 R_t 。我们提供提示模板 $Prompt_2$ 如下：

```
You are an intelligent assistant,
please answer my question. To help
you ... local knowledge base is
provided as follows: <knowledge>
```

`Now, answer the question...: <history><query>`

然后，我们将提示模板、检索到的知识、对话历史记录和用户查询连接起来，以获取 LLM 的输入 I_t 。将 I_t 输入 StockGPT，我们可以得到 R_t 响应。

$$I_t = concat(Prompt_2, d^*, H_t, Q_t) \quad (15)$$

$$R_t = StockGPT_{stage2}(I_t) \quad (16)$$

4. 实验

在本节中，我们进行了实验，以验证 Stock-Chain 完成财务分析任务的能力。由于我们框架的结构，实验可以分为两部分。第一部分的实验主要考察了模型的年化收益率和准确率。在第二部分中，我们通过人类 & GPT-4 的偏好评估、消融研究和案例研究来展示我们的 Stock-Chain 的性能。

4.1. AlphaFin-Test 数据集

我们从数据源中选择一个数据子集，该子集从训练数据集中排除，作为我们的测试数据集。鉴于所有研究数据集都是英文的，我们的主要重点是从其他数据集中抽样。例如财务报告和 StockQA 数据集。对于第 1 阶段，我们从财务报告数据集中选择测试数据集。示例演示如下：`According to ..., please judge the trend of the company and give a clear answer up or down. Input: <reports & stock price>, Output: <Up/down>.` 至于第 2 阶段，测试数据集是从 StockQA 和研究数据集中抽取的。AlphaFin 检验数据集使我们能够评估模型在资本市场中的能力。

4.2. 基线

为了在测试数据集上充分验证我们的 Stock-Chain 的有效性，我们选择了四类模型：

- $\boldsymbol{Major\ Indices}$: We select indices in the Chinese capital market, including the SCI, CSI 300, SSE50, and CNX.
- $\boldsymbol{ML\&DL\ Algorithms}$: We employ ML algorithms such as Logistic and XGBoost, and DL models like LSTM and GRU, which are widely employed for time-series prediction.
- $\boldsymbol{General\ LLMs}$: We focus on the general-purpose LLMs like ChatGLM2 and ChatGPT. These LLMs have been chosen due to their ability and wide range of applications in NLP.
- $\boldsymbol{FinLLMs}$: In the financial domain, we focus on open-source FinLLM, such as FinGPT and FinMA, which have been trained for financial tasks like financial analysis and forecasting.

4.3. 设置

对于第一阶段，实验旨在预测下个月的股价走势，并观察模型在真实市场中的回报。

- ML & DL：由于它们的局限性，它们只能分析时间序列数据。因此，他们的投入仅限于股票价格。
- LLMs：相比之下，LLMs 具有生成能力，允许它们同时合并报告数据和股票价格系列数据。

在第 2 阶段，我们检查模型生成能力并使用 GPT4 & 人类作为评估器。所有 LLM 的生成策略都是贪婪搜索，以实现最佳和稳定的性能，并在我们的实验中集成了 RAG。

其中，超参数如下：批量大小 16、LoRA 等级 8、余弦 lr 调度器、学习率 5e-5、bf16 和 1 个 NVIDIA A800-80GB，用于所有训练过程。具体来说，在第 1 阶段，我们为第一步训练了 4 个 epoch，为第二步训练了 20 个 epoch。在第 2 阶段，我们使用 $StockGPT_{stage1}$ 作为基础模型，并在 AlphaFin 数据集上对 2 个时期进行了增量微调。

4.4. 指标

对于第一阶段，我们使用两类指标。第一类是核心指标，包括衡量盈利能力的 ARR 和 ACC。第二类是辅助指标，有助于分析不同的模型，例如最大回撤（MD）、卡尔马比率（CR）和夏普比率（SR），用于衡量风险评估。通过这些指标，我们对模型的能力进行了全面评估。对于第 2 阶段，我们使用 ROUGE (**?**) 作为评估指标，用于衡量生成的输出和参考信息之间的相似性。此外，我们使用 GPT4 & 人类作为评分的裁判。通过考虑这些指标，我们可以更好地评估模型的性能。

4.5. 比较结果

如图 **??** 所示，曲线表示每种方法的 AR。值得注意的是，从 2023 年开始，Stock-Chain 实现了最高的 AR，并保持了上升趋势。它表明了 Stock-Chain 在投资中的有效性。

参考表 **??** ，Stock-Chain 达到了最高的 30.8 % ARR 证明了其有效性。根据 Table **??** ，我们可以得出以下观察结果：

首先，ML & DL 在股票趋势预测方面具有一定的分析能力，它们取得了令人印象深刻的 ARR。其次，在将报告数据与市场数据整合后，LLM 普遍超过 ML & DL，从而增强了股票趋势预测能力。ChatGPT 的 ARR 为 14.3 %。虽然 LLM 在大量文本数据上进行了训练，但它们缺乏对金融领域的优化。因此，通过对金融领域的微调，FinLLMs 可以提高股票趋势预测能力。FinGPT 模型的 ARR 为 17.5 %。

最后，在基于财务报告 cot 数据对 Stock-Chain 进行微调后，我们实现了 30.8 % 的 ARR，55.63 % 的 ACC。AlphaFin 数据集在 LLM 的训练中起着至关重要的作用。通过利用全面的财务数据进行微调，我们提高了预测的准确性和回报，从而验证了 Stock-Chain 的性能。

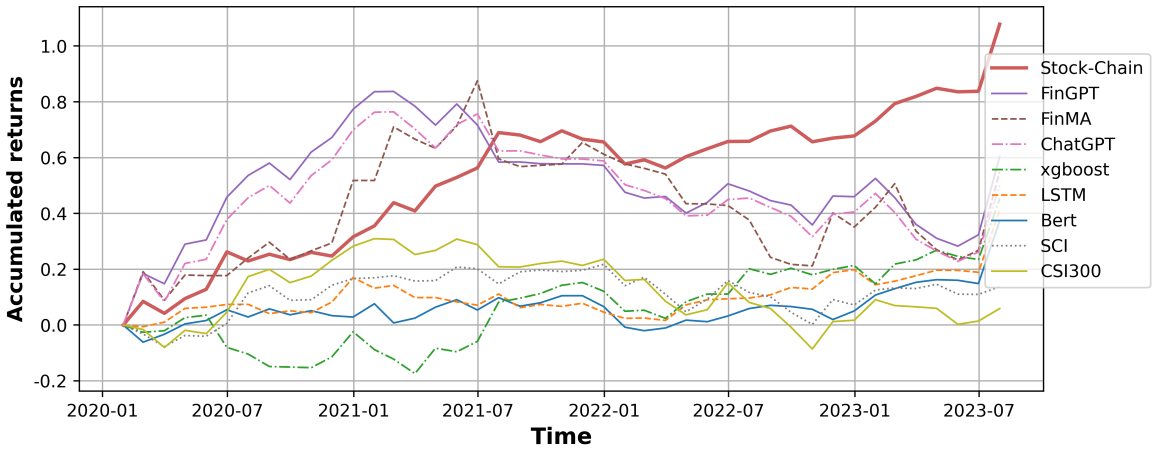

Figure 4: 2020 年 1 月至 2023 年 7 月财务报告数据集测试集下各基线的累计收益（AR）。该图显示了一些基线的曲线。

Model	ARR ↑	AERR ↑	ANVOL ↓	SR ↑	MD ↓	CR ↑	MDD ↓	ACC ↑
SSE50	-1.0 %	-2.7 %	19.3 %	-0.054	45.9 %	-0.023	29	-
CSI 300	1.7 %	0	18.2 %	0.092	39.5 %	0.043	30	-
SCI	3.9 %	2.2 %	14.8 %	0.266	21.5 %	0.183	19	-
CNX	7.6 %	5.9 %	26.5 %	0.287	41.3 %	0.185	20	-
Randomforest	9.8 %	8.1 %	19.5 %	0.501	16 %	0.608	22	55.5 %
RNN	8.1 %	6.4 %	10.9 %	0.742	15.7 %	0.515	12	54.1 %
BERT	10.7 %	9.0 %	16.1 %	0.664	13.5 %	0.852	14	51.4 %
GRU	11.2 %	9.5 %	13.7 %	0.814	14.6 %	0.765	21	54.7 %
LSTM	11.8 %	10.1 %	15.4 %	0.767	15.3 %	0.768	19	55.2 %
Logistic	12.5 %	10.8 %	27.1 %	0.463	32.5 %	0.385	18	54.8 %
XGBoost	13.1 %	11.4 %	20.5 %	0.633	20.9 %	0.619	17	55.9 %
Decision Tree	13.4 %	11.7 %	19.6 %	0.683	11.9 %	1.126	20	55.1 %
ChatGLM2	8.1 %	6.4 %	24.9 %	0.324	62.6 %	0.126	26	49.5 %
ChatGPT(3.5Turbo)	14.3 %	12.6 %	27.7 %	0.516	53.6 %	0.267	23	51.4 %
FinMa	15.7 %	14.0 %	37.1 %	0.422	66.3 %	0.236	25	49.1 %
FinGPT	17.5 %	15.8 %	28.9 %	0.605	55.5 %	0.312	24	50.5 %
Stock-Chain	30.8 %	29.1 %	19.6 %	1.573	13.3 %	2.314	10	55.7 %

Table 2: AlphaFin-Test 的主要实验结果。ARR（年化收益率）和 ACC（准确率）是核心指标，而中间指标（如 AERR、ANVOL 等）可以帮助投资者评估模型的有效性。由于回报率通常波动很大，为了保证性能的稳定性，我们运行每个模型 10 次并获得平均结果。

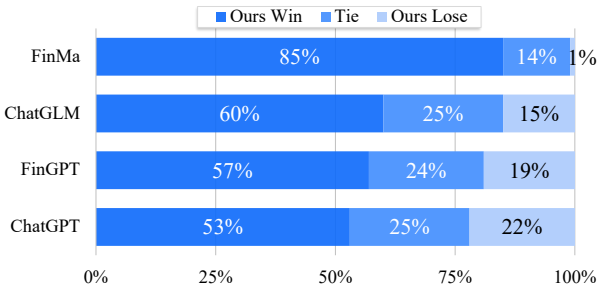

Figure 5: 通过人工进行偏好评估。

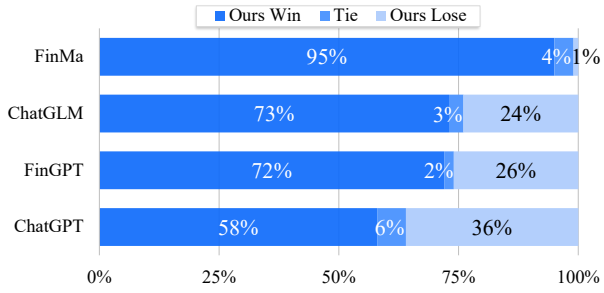

Figure 6: 通过 GPT-4 进行偏好评估。

4.6. 消融研究

我们进行了两个消融实验。首先，通过分析不同数据微调的效果，观察了股票链的股票趋势预测能

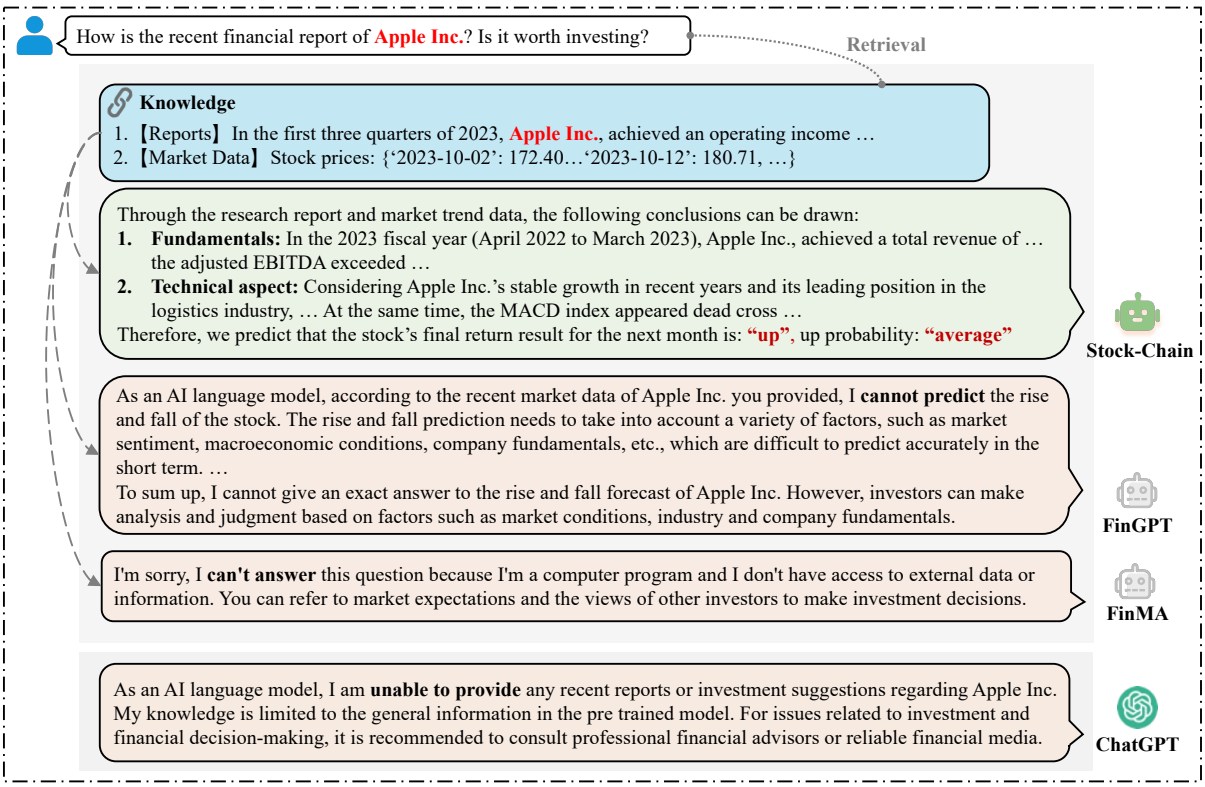

Figure 7: Stock-Chain、FinGPT、FinMA 和 ChatGPT 的输出案例

Model	ARR ↑	SR ↑	Out_len ↑	N/A ↓
ChatGLM2	8.1 %	0.324	228.1	52.3 %
w/ raw_data	15.8 %	0.636	17.2	-
w/ CoT_data	10.1 %	0.469	**476.1**	32.4 %
Stock-Chain	**30.8 %**	**1.573**	254.8	25.9 %

Table 3: 不同训练数据下第 1 阶段的消融结果。Out_len：LLM 输出的平均长度。N/A：无效答案率。

Model	ROUGE-1 ↑	ROUGE-2 ↑	ROUGE-L ↑
ChatGLM2	0.2794	0.1944	0.2642
w/ Fin. News	0.3477	0.2821	0.3445
w/ Fin. reports	0.2611	0.1603	0.2396
Stock-Chain	**0.4352**	**0.3056**	**0.4031**

Table 4: 不同训练数据下第 2 阶段的消融结果。

力; 基于 Table **??**，与 ChatGLM2 相比，LLMs 预测股价的能力在对原始数据和 CoT 数据进行微调后有所提高，分别实现了 15.8 % 和 10.1 % 的回报。

此外，无效答案的比例也有所提高。值得一提的是，在对原始数据进行微调后，LLM 的输出仅包括上升和下降，从而解决了无效答案的问题。经过两组数据的微调，我们的 Stock-Chain 以 30.8 % ARR 实现了最佳性能，无效答案的比例也有所下降，达到 25.9 %。

至于第二次消融实验，我们研究了在不同数据下微调 LLM 后输出质量是否有所提高。

根据表 **??**，我们观察到，在对新闻数据进行微调后，Stock-Chain 在 rouge1 和 rouge2 方面的得分分别达到 0.3477 和 0.2821。此外，值得注意的是，Stock-Chain 在对新闻和报道进行微调后达到了最佳性能。

4.7. 偏好评估

我们使用 GPT-4 和人类作为评委来评估每个 LLM 在测试数据集上的输出性能。所有 LLM 都在这个实验中集成了 RAG。在人类部分，Stock-Chain 在内容有效性方面优于其他 LLM。基于图 **??**，与 ChatGLM2 相比，Stock-Chain 的胜率超过 60 %，而与 FinGPT 等 FinLLMS 相比，胜率达到 62 %。基于图 **??**，当 GPT4 担任裁判时，得出了类似的结论。与人类评级相比，Stock-Chain 表现出更高的比率，对 ChatGPT 的胜率为 58 %，对 ChatGLM2 的胜率为 73 %。总体而言，Stock-Chain 的产出是有效的。

4.8. 个案研究

我们提出了 Stock-Chain 的部分输出以进行定性分析。如图 **??** 所示，当用户查询与 Appe Inc. 相关的最新报告和投资建议时，Stock-Chain 会从知识库中获取实时相关信息和市场数据，并将其作为输入提供给 LLM。LLM 的产出得到了增强，新闻保持最新，使投资者能够分析和获得建议。然而，对于 ChatGPT 和 FinGPT，我们观察到它在质量和实时响应方面存在很大的缺陷。因此，通过整合 RAG，Stock-Chain 解决了 LLM 中幻觉和实时输出不足的问题，增强了 LLM 的实用性和能力。

5. 相关工作

5.1. 金融数据集

一般金融数据集包括来自金融行业内各种来源的大量信息，例如互联网数据和专有数据。目前，主要的财务数据集集包括各种任务。FPB (**?**) 和 FiQA-SA (**?**) 主要用于情绪分析。Headline (**?**) 数据集主要用于新闻标题识别。对于问答，我们主要使用 FinQA (**?**) 和 ConvFinQA (**?**) 数据集。遗憾的是，金融领域仍然受到文本数据集稀缺的影响，阻碍了 FinLLM 的发展。为了弥合这一差距，我们提出了 AlphaFin，为金融业培训自己的 FinLLM 提供支持。

5.2. 金融领域的算法

传统的 ML & DL 算法，如 LSTM (**?**) 、Logistic (**?**) 和 BERT (**?**) 等，已被应用于股票趋势预测。然而，ML & DL 专注于最终结果，而没有分析推动市场趋势的潜在因素。至于 FinLLM，虽然 BloombergGPT (**?**) 、FinMA (**?**) 和 FinGPT (**?**) 在社区中发挥着重要作用，但它们主要基于英语数据集。相比之下，Stock-Chain 依赖于中文，专为股票趋势预测而设计。

至于 RAG，在社区中引起了越来越多的关注 (**?**) 。与传统方法相比，RAG 在各种 NLP 任务中都具有显著的性能 (**??**) 。然而，如果没有 RAG，FinLLM 通常会产生幻觉和无意义的输出。因此，我们将 LLM 与 RAG 集成以解决上述问题。

6. 结论

在这项工作中，我们将财务分析的任务形式化，并提出了 AlphaFin 数据集来增强 LLM 的能力，并在其上微调 StockGPT。然后，我们提出了通过 RAG 集成实时金融数据库的 Stock-Chain 框架，以解决 LLM 输出的幻觉和 LLM 无法生成实时内容的问题。我们对所提出的 AlphaFin 数据集进行了广泛的实验，以及一些补充实验，如消融研究、GPT4 & 人类偏好评估和案例研究，以揭示 Stock-Chain 优于所有基线方法，并显示出对财务分析任务的有效性。

7. 道德考量和局限性

我们断言，围绕本文的提交不存在道德困境，也没有已知的竞争性经济利益或个人关系可能对所提出的研究工作产生影响。

尽管这项研究做出了积极贡献，但我们认识到我们的工作仍有很大的发展空间。在我们未来的工作中，我们将进一步为开源 FinLLMs 做出贡献，提高其泛化性，增强其在其他金融任务中的能力，并创建一个更强大的开源 FinLLM。

8. 参考书目