# LLMs在股票评级中的应用分析

### 🧑‍💻 User

可以告诉我在篇文章是如何使用LLMs实现股票评级的

──────────────────────────────────────────────────

### 🤖 Assistant

这篇文章的主要内容是探索如何使用大型语言模型（LLMs）来预测股票表现和生成股票评级。以下是关键步骤和发现的总结：

方法：

**•** 数据收集：

**•** 财务基本面数据： 使用 finagg API 收集 S&P 500 成分股公司 2022 年 1 月至 2024 年 3 月的季度财务报表数据 (10-Q 和 10-K 表格)。

**•** 市场数据： 收集每日股票价格，并计算历史回报率、市场相对回报率、板块相对回报率、52 周价格区间和 90 天波动率。

**•** 新闻数据： 收集 S&P 500 成分股的新闻文章，并使用 GPT-4-32k (v0613) 进行摘要，提取关键事件和趋势。

**•** 新闻情绪数据： 使用 GPT-4-32k (v0613) 对新闻摘要进行情绪评分，范围为 -5（极度负面）到 5（极度正面）。

**•** 分析师股票评级： 收集 126 家公司对 S&P 500 成分股的 45,000 个分析师评级数据。

**•** 模型选择：

**•** 使用 GPT-4-32k (v0613)，该模型具有 32,000 个 token 的上下文窗口，并在 2021 年 9 月之前的数据上进行训练，以防止信息泄漏。

**•** 提示工程：

**•** 系统提示： 指示 LLM 扮演财务分析师的角色，并定义股票评级等级及其定义。

**•** 用户提示： 采用思维链和少样本提示方法，引导 LLM 在进行最终预测之前进行推理，并提供一个输出示例。以结构化格式提供公司特定输入数据，首先是文本信息，然后是表格中的数字数据。并执行基本的 CoVE（验证链）来检测它是否在预测正确的日期。

**•** 实验设置：

**•** 要求 LLM 在 2022 年 1 月至 2024 年 6 月的每个月初为 S&P 500 中的每家公司生成 1 个月、3 个月、6 个月、12 个月和 18 个月的股票评级。

**•** 进行五项实验，每项实验都使用 GPT-4-32k (v0613)，但向模型提供的输入数据不同：

**•** Vanilla： 仅包含公司历史数据的快照（回报率、价格等）。

**•** News： 在 Vanilla 的基础上，添加前一个月公司新闻和行业新闻的摘要。

**•** Sentiment： 在 Vanilla 的基础上，添加前一个月公司新闻情绪和行业新闻情绪的评分。

**•** Fundamentals： 在 Vanilla 的基础上，添加季度财务基本面数据。

**•** Fundamentals + Sentiment： 在 Fundamentals 的基础上，添加 Sentiment 中的情绪评分。

**•** 评估：

**•** 基于 1 个月、3 个月、6 个月、12 个月和 18 个月的远期收益来评估评级，包括市场相对和板块相对收益的评估。

**•** 如果真实远期收益的分位数与评级等级一致，则认为评级是正确的。

**•** 使用平均绝对误差 (MAE) 来衡量预测的准确性。

主要发现：

**•** 基准 Vanilla LLM 模型的表现优于传统的分析师评估。

**•** Fundamentals 和 Fundamentals + Sentiment 模型的表现最好，突出了财务基本面对预测准确性的重要影响。 将情绪评分与基本面数据相结合，进一步提高了预测准确性。

**•** 整合新闻摘要和情绪分析提供了一些短期预测优势，但与 Vanilla 模型相比，并没有显著提高长期预测准确性。

**•** 将新闻作为文本添加到 LLM 与添加新闻情绪之间的性能差异非常小， 这表明这两种方法提供了类似的好处。

**•** LLM 在短期预测中表现更好。

**•** 新闻摘要对短期预测更有帮助，而传统分析师在长期预测中表现更好。

**•** 使用新闻衍生数据会使模型偏向于做出更积极的评级。

总结：

这篇文章证明了 LLM 在预测股票评级方面的潜力。通过整合各种类型的信息，LLM 可以生成比传统分析师更准确的预测，特别是当利用财务基本面数据时。研究结果强调了 LLM 在金融分析中的应用价值，并为未来的研究方向提供了见解。

──────────────────────────────────────────────────

### 🧑‍💻 User

我想复现这篇论文的LLMs预测功能我该如何实现，你认为有什么可以提升的方案

──────────────────────────────────────────────────

### 🤖 Assistant

复现这篇论文的 LLMs 预测功能，你需要按照以下步骤进行操作。我会根据论文中描述的方法，提供一个详细的实现指南，并提出一些可能的提升方案。

实现步骤：

**•** 数据准备：

**•** 财务基本面数据：

**•** 工具： 使用 finagg API [22] (或其他类似工具) 访问 SEC EDGAR API [24]，获取 S&P 500 成分股公司 2022 年 1 月至 2024 年 3 月的季度财务报表数据 (10-Q 和 10-K 表格)。

**•** 数据： 提取 Figure 1 中列出的财务指标，例如：每股收益 (EPS)、流通股、净收入、现金、资产、负债、股东权益、资产回报率、股本回报率、资产覆盖率、每股账面价值 (BVPS)、债务股本比 (D/E)、速动比率和营运资本比率。

**•** 市场数据：

**•** 工具： 使用 yfinance [1] (或其他类似工具) 获取每日股票价格。

**•** 数据： 计算 1 个月、3 个月和 12 个月的历史回报率、市场相对回报率 (相对于 S&P 500 指数)、板块相对回报率 (相对于公司所在板块)、52 周价格区间和 90 天波动率 (每日回报率的标准差)。

**•** 新闻数据：

**•** 工具： 使用 SerpAPI [25] (或其他新闻 API) 收集 S&P 500 成分股的新闻文章。

**•** 数据： 使用公司名称和别名进行实体识别 (NER) 过滤相关新闻。论文中提到用[8] LangChain可以作为参考.

**•** 新闻摘要： 使用 GPT-4-32k (v0613) 对每个公司和板块的月度新闻进行摘要，提取关键事件和趋势。

**•** 新闻情绪数据：

**•** 工具： 使用 GPT-4-32k (v0613) 对新闻摘要进行情绪评分。

**•** 数据： 对每个公司和板块的新闻摘要进行情绪评分，范围为 -5 (极度负面) 到 5 (极度正面)。

**•** 分析师股票评级：

**•** 工具： 论文中使用了[1] yfinance，你可以尝试使用它或其他数据源。

**•** 数据： 收集分析师股票评级数据，包括评级机构、评级日期和评级本身。

**•** 模型选择：

**•** 使用 GPT-4-32k (v0613)，该模型具有 32,000 个 token 的上下文窗口，并在 2021 年 9 月之前的数据上进行训练。

**•** 注意： 由于模型训练数据截止日期和 API 访问限制，你可能需要使用更新的模型或其他具有类似功能的模型。

**•** 提示工程：

**•** 系统提示：

你是一位专业的财务分析师，负责根据提供的财务数据、市场数据和新闻信息预测股票评级。   
股票评级等级如下：   
-2: 强烈卖出 (Strong Sell)   
-1: 适度卖出 (Moderate Sell/Underperform/Underweight)   
0: 持有 (Hold)   
1: 适度买入 (Moderate Buy/Outperform/Overweight)   
2: 强烈买入 (Strong Buy/Buy)   
   
以下是财务指标的定义：   
[在此处插入 Figure 1 中列出的财务指标及其定义]

**•** 用户提示模板：

公司：{公司名称}   
评级日期：{评级日期}   
   
请预测以下时间范围的股票评级：   
1 个月后：{评级日期 + 1 个月}   
3 个月后：{评级日期 + 3 个月}   
6 个月后：{评级日期 + 6 个月}   
12 个月后：{评级日期 + 12 个月}   
18 个月后：{评级日期 + 18 个月}   
   
请按照以下格式输出：   
{预测日期}：{预测评级}，目标价格：{预测价格}, 理由: {你的推理过程}   
   
以下是该公司过去四个季度的财务数据：   
{以 HTML 表格格式提供财务数据，参考 Figure 1}   
   
以下是该公司的市场数据：   
当前股价：{当前股价}   
52 周最低价：{52 周最低价}   
52 周最高价：{52 周最高价}   
90 天波动率：{90 天波动率}   
1 个月回报率：{1 个月回报率}   
3 个月回报率：{3 个月回报率}   
12 个月回报率：{12 个月回报率}   
1 个月市场相对回报率：{1 个月市场相对回报率}   
3 个月市场相对回报率：{3 个月市场相对回报率}   
12 个月市场相对回报率：{12 个月市场相对回报率}   
1 个月板块相对回报率：{1 个月板块相对回报率}   
3 个月板块相对回报率：{3 个月板块相对回报率}   
12 个月板块相对回报率：{12 个月板块相对回报率}   
   
[根据不同的实验，选择性地提供以下信息]   
以下是该公司上个月的新闻摘要：   
{公司新闻摘要}   
   
以下是该公司所在板块上个月的新闻摘要：   
{板块新闻摘要}   
   
以下是该公司上个月的新闻情绪评分：{公司新闻情绪评分}   
以下是该公司所在板块上个月的新闻情绪评分：{板块新闻情绪评分}   
   
以下是一个示例：   
输入:   
...[示例输入数据]...   
输出:   
...[示例输出数据]...

**•** 注意： 根据不同的实验 (Vanilla, News, Sentiment, Fundamentals, Fundamentals + Sentiment)，你需要调整用户提示中提供的数据。

**•** 代码实现：

**•** 使用 Python 和你选择的 LLM API (例如 OpenAI API) 编写代码。

**•** 代码应包括数据加载、预处理、提示构建、LLM 调用、输出解析和评估等步骤。

**•** 你需要编写函数来计算市场相对回报率和板块相对回报率。

**•** 评估：

**•** 根据论文中描述的方法，使用 1 个月、3 个月、6 个月、12 个月和 18 个月的远期收益来评估预测的准确性。

**•** 计算每个预测时间范围的市场相对 MAE 和板块相对 MAE。

**•** 计算 3 个月、6 个月和 12 个月预测的平均 MAE 作为复合误差。

提升方案：

**•** 更强大的模型： 尝试使用更新、更强大的 LLM，例如 GPT-4 Turbo 或其他具有更大上下文窗口和更强推理能力的模型。

**•** 更精细的提示工程：

**•** 更详细的财务指标解释： 在系统提示中提供更详细的财务指标解释和计算公式。

**•** 更具指导性的推理步骤： 在用户提示中引导 LLM 进行更具体的推理，例如要求其分析特定财务指标的变化趋势或比较公司与同行业其他公司的表现。

**•** 动态示例选择： 根据输入数据的特征，动态选择最相关的示例提供给 LLM。

**•** 更丰富的数据：

**•** 更长的历史数据： 使用更长时间范围的历史财务数据和市场数据。

**•** 更细粒度的新闻数据： 使用每日新闻数据而不是月度摘要，并尝试使用更先进的 NER 技术来提高新闻数据的相关性。

**•** 其他数据源： 考虑纳入其他数据源，例如分析师报告、社交媒体情绪、宏观经济指标等。

**•** 财报电话会议记录： 使用财报电话会议记录，这是分析师经常使用的信息来源。

**•** 模型微调： 如果你有足够的标记数据 (例如分析师评级数据)，可以考虑对 LLM 进行微调，以提高其在股票评级任务上的性能。

**•** 集成其他机器学习模型： 可以考虑将 LLM 与其他机器学习模型 (例如时间序列模型) 集成，以结合不同模型的优势。

**•** 更全面的评估：

**•** 考虑交易成本： 在评估中考虑交易成本，以更真实地反映投资策略的盈利能力。

**•** 风险调整后的回报： 使用风险调整后的回报指标 (例如夏普比率) 来评估投资策略的表现。

**•** 回测： 使用更长的历史数据进行回测，以验证模型的稳健性。

**•** 解释性： 增强模型的可解释性，例如使用 attention 机制可视化 LLM 关注的输入数据部分，或生成更详细的推理过程说明。

总结：

复现这篇论文需要仔细的数据准备、模型选择和提示工程。通过采用上述提升方案，你可以尝试构建一个更准确、更可靠的股票评级预测系统。记住，金融市场预测是一个非常具有挑战性的任务，没有模型能够保证完美的预测。

──────────────────────────────────────────────────

### 🧑‍💻 User

可以帮我基于这个方案结合提升建议将研究对象改为中国A股市场，进行完整提案

──────────────────────────────────────────────────

### 🤖 Assistant

好的，以下是一份基于上述方案，并结合提升建议，将研究对象改为中国 A 股市场的完整提案。

提案：基于大型语言模型 (LLMs) 的中国 A 股市场股票评级预测

一、 项目背景与意义

投资分析是金融服务业的基石，对于金融市场的运作至关重要。它提供推动投资决策、市场趋势和经济政策的重要见解。近年来，先进的机器学习技术，特别是大型语言模型（LLMs），为增强股票评级流程提供了新的机遇。

本研究旨在探索如何利用 LLMs 预测中国 A 股市场的股票表现并生成股票评级。中国 A 股市场作为全球第二大股票市场，具有独特的市场特征和数据环境，为 LLMs 的应用提供了新的挑战和机遇。通过将 LLMs 与 A 股市场的特定数据相结合，并借鉴国际研究的先进经验，本项目有望开发出更准确、更可靠的股票评级预测系统，为投资者提供更有价值的投资参考，并推动 AI 技术在金融领域的应用。

二、 研究目标

本研究的主要目标是开发一个基于 LLMs 的中国 A 股市场股票评级预测系统，该系统能够：

**•** 处理多模态金融数据： 有效整合和处理包括财务报表数据、市场行情数据、新闻文本数据、分析师报告等在内的多模态金融数据。

**•** 生成准确的股票评级： 基于多模态数据和 LLMs 的强大推理能力，生成准确的股票评级预测，包括“强烈买入”、“买入”、“持有”、“卖出”和“强烈卖出”五个等级。

**•** 提供可解释的预测结果： 解释 LLMs 做出预测的依据和推理过程，提高模型的可解释性和可信度。

**•** 评估并优化模型性能： 使用历史数据对模型进行回测和评估，并根据评估结果不断优化模型性能。

三、 研究方法

本研究将采用以下方法：

**•** 数据采集与预处理：

**•** 财务基本面数据：

**•** 数据源： 使用 Wind、Choice 等国内金融数据供应商提供的 API，获取 A 股上市公司 2015 年 1 月至 2024 年 12 月 (或其他时间段) 的季度财务报表数据。

**•** 数据内容： 包括资产负债表、利润表和现金流量表中的关键指标，例如：每股收益 (EPS)、流通股、净利润、现金、资产、负债、股东权益、资产回报率、股本回报率、资产覆盖率、每股账面价值 (BVPS)、债务股本比 (D/E)、速动比率和营运资本比率等。

**•** 预处理： 对数据进行清洗、缺失值处理、异常值处理和标准化等操作。

**•** 市场数据：

**•** 数据源： 使用 Wind、Choice 等金融数据供应商提供的 API，获取 A 股上市公司的每日股票价格数据。

**•** 数据内容： 计算 1 个月、3 个月、6 个月、12 个月和 24 个月的历史回报率、市场相对回报率 (相对于沪深 300 指数)、行业相对回报率 (相对于公司所在行业指数)、52 周价格区间和 90 天波动率 (每日回报率的标准差)。

**•** 预处理： 计算各种技术指标。

**•** 新闻数据：

**•** 数据源： 使用八爪鱼，后羿采集器等爬虫工具或新闻数据供应商提供的 API，收集 A 股上市公司和行业相关的新闻数据。

**•** 数据内容： 包括新闻标题、正文、发布时间和来源等信息。

**•** 预处理： 使用公司名称和别名进行实体识别 (NER)，筛选出与目标公司相关的新闻；使用 Jieba 等分词工具进行中文分词；使用停用词表去除停用词；使用 LLMs (例如 ChatGLM, Wenxin, Tongyi) 对新闻进行摘要，提取关键事件和趋势。

**•** 新闻情绪数据：

**•** 数据源： 使用 LLMs (例如 ChatGLM, Wenxin, Tongyi) 对新闻摘要进行情绪评分。

**•** 数据内容： 对每个公司和行业的新闻摘要进行情绪评分，范围为 -5 (极度负面) 到 5 (极度正面)。

**•** 预处理： 对情绪评分进行标准化处理。

**•** 分析师报告：

**•** 数据源： 使用 Wind、Choice等数据供应商。

**•** 数据内容： 收集分析师发布的股票评级报告，包括评级机构、评级日期、评级和目标价等信息。

**•** 预处理： 将文本形式的分析师报告转化为结构化数据。

**•** 模型选择与改进：

**•** 基础模型： 选择 ChatGLM, Wenxin, Tongyi 等国内开发的、针对中文金融领域进行过优化的 LLMs 作为基础模型。这些模型具有较强的中文理解能力和一定的金融知识。

**•** 模型改进：

**•** 更大的上下文窗口： 尝试使用支持更大上下文窗口的模型，以处理更长的文本输入和更复杂的推理任务。

**•** 模型微调： 使用收集到的分析师评级数据对 LLMs 进行微调，以提高其在股票评级任务上的性能。

**•** LoRA/QLoRA: 尝试使用 LoRA/QLoRA 等轻量级微调方法，在保证模型性能的同时降低训练成本。

**•** 提示工程：

**•** 系统提示：

你是一位专业的金融分析师，负责根据提供的财务数据、市场数据和新闻信息预测中国 A 股市场股票的评级。   
股票评级等级如下：   
-2: 强烈卖出 (Strong Sell)   
-1: 适度卖出 (Moderate Sell/Underperform/Underweight)   
0: 持有 (Hold)   
1: 适度买入 (Moderate Buy/Outperform/Overweight)   
2: 强烈买入 (Strong Buy/Buy)   
   
以下是财务指标的定义：   
[在此处插入详细的财务指标及其定义，例如：每股收益 (EPS) = 净利润 / 流通股]

**•** 用户提示模板：

公司：{公司名称} ({股票代码})   
评级日期：{评级日期}   
   
请预测以下时间范围的股票评级：   
1 个月后：{评级日期 + 1 个月}   
3 个月后：{评级日期 + 3 个月}   
6 个月后：{评级日期 + 6 个月}   
12 个月后：{评级日期 + 12 个月}   
24 个月后：{评级日期 + 24 个月}   
   
请按照以下格式输出：   
{预测日期}：{预测评级}，目标价格：{预测价格}, 理由: {你的推理过程}   
   
以下是该公司过去四个季度的财务数据：   
{以 HTML 表格格式提供财务数据}   
   
以下是该公司的市场数据：   
当前股价：{当前股价}   
52 周最低价：{52 周最低价}   
52 周最高价：{52 周最高价}   
90 天波动率：{90 天波动率}   
1 个月回报率：{1 个月回报率}   
3 个月回报率：{3 个月回报率}   
12 个月回报率：{12 个月回报率}   
1 个月市场相对回报率：{1 个月市场相对回报率}   
3 个月市场相对回报率：{3 个月市场相对回报率}   
12 个月市场相对回报率：{12 个月市场相对回报率}   
1 个月行业相对回报率：{1 个月行业相对回报率}   
3 个月行业相对回报率：{3 个月行业相对回报率}   
12 个月行业相对回报率：{12 个月行业相对回报率}   
   
[根据不同的实验，选择性地提供以下信息]   
以下是该公司最近一个月的新闻摘要：   
{公司新闻摘要}   
   
以下是该公司所在行业最近一个月的新闻摘要：   
{行业新闻摘要}   
   
以下是该公司最近一个月的新闻情绪评分：{公司新闻情绪评分}   
以下是该公司所在行业最近一个月的新闻情绪评分：{行业新闻情绪评分}   
   
以下是一个示例：   
输入:   
...[示例输入数据]...   
输出:   
...[示例输出数据]...

**•** 提示工程优化：

**•** 更详细的财务指标解释： 在系统提示中提供更详细的财务指标解释、计算公式和行业对比信息。

**•** 更具指导性的推理步骤： 在用户提示中引导 LLMs 进行更具体的推理，例如要求其分析特定财务指标的变化趋势、比较公司与同行业其他公司的表现，并考虑行业发展前景和宏观经济因素。

**•** 动态示例选择： 根据输入数据的特征 (例如公司所处行业、市值规模等)，动态选择最相关的示例提供给 LLM。

**•** 模型训练与评估：

**•** 训练数据： 使用收集到的历史数据 (2015 年 1 月至 2023 年 12 月) 训练模型。

**•** 测试数据： 使用 2024 年 1 月至 2024 年 12 月的数据测试模型性能。

**•** 评估指标： 使用 1 个月、3 个月、6 个月、12 个月和 24 个月的远期收益来评估预测的准确性，并计算每个预测时间范围的市场相对 MAE 和行业相对 MAE。计算 3 个月、6 个月和 12 个月预测的平均 MAE 作为复合误差。同时，考虑使用夏普比率等风险调整后的回报指标来评估投资策略的表现。

**•** 回测： 使用更长的历史数据 (例如 2015 年 1 月至 2024 年 12 月) 进行回测，以验证模型的稳健性。在回测中，需要考虑交易成本 (例如佣金和印花税) 的影响。

四、 预期成果

本项目预期将取得以下成果：

**•** 一个基于 LLMs 的中国 A 股市场股票评级预测系统。

**•** 一套针对中国 A 股市场的多模态金融数据集。

**•** 一套优化的 LLMs 提示工程方案。

**•** 对模型性能的详细评估报告，包括 MAE、风险调整后的回报指标和回测结果。

**•** 关于 LLMs 在中国 A 股市场股票评级预测中应用的研究论文。

五、 创新点

本项目的主要创新点在于：

**•** 针对中国 A 股市场： 将 LLMs 应用于中国 A 股市场，并针对 A 股市场的特点进行模型优化和提示工程设计。

**•** 多模态数据融合： 融合财务报表数据、市场行情数据、新闻文本数据和分析师报告等多模态数据，提高预测的准确性和可靠性。

**•** 模型可解释性： 提供模型预测的依据和推理过程，增强模型的可解释性和可信度。

**•** 本土化 LLM 应用： 使用国产 LLMs，推动本土 AI 技术在金融领域的应用。

六、 项目进度安排

本项目计划在 12 个月内完成，具体进度安排如下：

**•** 第 1-3 个月： 数据采集与预处理。

**•** 第 4-6 个月： 模型选择与改进，提示工程设计。

**•** 第 7-9 个月： 模型训练与评估。

**•** 第 10-12 个月： 撰写研究论文，完善项目文档。

七、 团队成员与分工

[在此处列出团队成员及其分工]

八、 经费预算

[在此处列出经费预算，包括数据采购、计算资源、人员费用等]

九、 风险与对策

本项目可能面临的风险包括：

**•** 数据质量问题： 金融数据的质量和完整性可能会影响模型的性能。

**•** 对策： 使用多个数据源进行交叉验证，并进行严格的数据清洗和预处理。

**•** 模型过拟合： LLMs 可能会过拟合训练数据，导致在测试数据上表现不佳。

**•** 对策： 使用正则化方法，例如 Dropout 和权重衰减，并使用交叉验证来选择最佳模型。

**•** 市场波动风险： 金融市场的波动可能会影响模型的预测准确性。

**•** 对策： 使用更长的历史数据进行回测，并考虑使用风险管理技术来控制投资风险。

十、 总结

本项目旨在开发一个基于 LLMs 的中国 A 股市场股票评级预测系统，该系统能够处理多模态金融数据，生成准确的股票评级预测，并提供可解释的预测结果。本项目将推动 AI 技术在金融领域的应用，并为投资者提供更有价值的投资参考。

补充说明:

**•** 这个提案是一个框架，你可以根据你的具体情况和资源进行调整和完善。

**•** 你需要根据实际情况选择合适的 LLMs 和数据源。

**•** 你需要根据 A 股市场的特点，设计合适的评估指标和回测方法。

**•** 你需要关注最新的研究进展，并将其应用到你的项目中。

**•** 对于中国特有的政策变动和监管需要重点关注, 适当调整模型和数据。

希望这个提案对你有所帮助！

──────────────────────────────────────────────────

### 🧑‍💻 User

论文中是如何进行评估的

──────────────────────────────────────────────────

### 🤖 Assistant

论文中主要通过以下几个方面进行评估，核心是比较 LLM 生成的股票评级与基于未来收益的“真实”评级之间的差异：

1. 评估指标：平均绝对误差 (MAE)

**•** 计算方式： MAE 是一个衡量预测值与真实值之间差异的指标，对于顺序数据（如评级）特别适用，因为它考虑了误差的大小。

**•** 应用场景： 在本论文中，MAE 被用来衡量 LLM 预测的评级与基于公司未来收益分位数确定的“真实”评级之间的平均差异。

2. “真实”评级的确定：基于未来收益的分位数

**•** 核心思想： 一只股票未来的表现越好（收益越高），它的“真实”评级就应该越高。

**•** 具体步骤：

**•** 计算远期收益： 对于每只股票，计算其在特定时间段（例如 1 个月、3 个月、6 个月、12 个月、18 个月）后的远期收益。

**•** 分位数划分： 将所有股票在特定时间点的远期收益分成五个分位数（quintiles）。

**•** 确定“真实”评级：

**•** 远期收益最低的 20% 的股票（第一个分位数）被赋予“强烈卖出”(-2) 的“真实”评级。

**•** 倒数第二个分位数的股票被赋予“适度卖出”(-1) 的“真实”评级。

**•** 中间分位数的股票被赋予“持有”(0) 的“真实”评级。

**•** 倒数第二个最高分位数的股票被赋予“适度买入”(1) 的“真实”评级。

**•** 远期收益最高的 20% 的股票（第五个分位数）被赋予“强烈买入”(2) 的“真实”评级。

3. 评估流程：

**•** 计算每个评级 (rating(t, p)) 的 MAE：

**•** rating(t, p): LLM 在时间 t 对公司 c 在未来 p 个月后的评级。

**•** Qc(t, p): 公司 c 在 t+p 时间的真实收益所处的分位数对应的评级。

**•** MAE = |rating(t, p) - Qc(t, p)| (绝对值)

**•** 计算市场相对 MAE 和板块相对 MAE：

**•** 市场相对收益： 公司收益减去市场（例如 S&P 500 指数）在同一时间段的收益。

**•** 板块相对收益： 公司收益减去其所在板块在同一时间段的收益。

**•** 分别使用市场相对收益和板块相对收益来计算分位数，并据此确定“真实”评级，进而计算 MAE。

**•** 计算复合 MAE：

**•** 将 3 个月、6 个月和 12 个月预测的 MAE 取平均值，得到一个综合的评估指标。

4. 分析师评级的对比：

**•** 论文还将 LLM 生成的评级与真实世界中分析师发布的评级进行了比较。

**•** 使用相同的评估方法（基于未来收益的分位数和 MAE）来评估分析师评级的准确性。

总结：

论文的评估方法的核心是将 LLM 的预测评级与基于未来收益表现的“真实”评级进行比较。通过计算 MAE 来衡量预测的准确性，并分别考虑了市场相对表现和板块相对表现。此外，论文还将 LLM 的表现与真实世界的分析师进行了对比。

这种评估方法的优点：

**•** 客观性： 基于未来收益的“真实”评级是客观的，不受人为因素的影响。

**•** 可比性： MAE 提供了一个统一的指标来比较不同模型和分析师的预测准确性。

**•** 易于理解： MAE 的计算和解释都比较简单。

这种评估方法的局限性：

**•** 对市场波动的敏感性： 基于特定时间点收益的评估方法可能会受到市场波动的影响。

**•** 忽略了其他因素： 该方法主要关注收益，可能忽略了其他影响股票评级的因素，例如公司治理、行业前景等。

**•** “真实”评级并非完美： 基于分位数的“真实”评级只是一种近似，并不能完全代表股票的真实价值。

尽管存在一些局限性，但论文的评估方法仍然为衡量 LLM 在股票评级预测方面的性能提供了一个有效且可解释的框架。