“GPT 如海”：RAG 与代码复现

华泰研究

研究员 林晓明
SAC No. S0570516010001 linxiaoming@htsc.com
SFC No. BPY421 +(86) 755 8208 0134

研究员 何康，PhD
SAC No. S0570520080004 hekang@htsc.com
SFC No. BRB318 +(86) 21 2897 2039

联系人 沈洋
SAC No. S0570123070271 shenyang023029@htsc.com
+(86) 21 2897 2228

2024 年 5 月 06 日 | 中国内地 深度研究

人工智能系列之 77：基于 GPT 和 RAG 技术的代码复现系统

本文探索大语言模型在量化研究领域中进行代码复现的实际应用。基于 GPT-4 系列模型和 RAG 技术，本文构建了一个完善且易用的代码复现框架，我们称之为“GPT 如海”。在框架内部，GPT 多模态模型提供了提取图片语义信息的能力，RAG 模块则有效支持海量文档的切分与检索，为大模型提供外部知识库，基于此，“GPT 如海”能够根据输入的图片或 PDF 文档，准确提取与代码复现任务相关的信息，并进行代码自动化生成。测试结果显示，“GPT 如海”能较准确地复现因子计算和人工智能量化策略代码。

多模态大语言模型：勾勒图片中蕴含的语义信息

传统的大语言模型仅能处理单一模态的数据，例如文本，而多模态大模型拥有多模态感知和生成能力，例如图像、语音和视频，在交互性上更贴近通用人工智能的愿景。近年来，多模态大模型已经得到了长足的发展，目前多模态大模型已经具备同时感知文本、图像、音频的能力，并且逐渐发展出生成多模态内容的能力。在量化研究领域中，研究报告同样蕴含较为丰富的多模态数据，包括文本和图像数据，在多模态大语言模型图像感知能力的加持下，代码复现系统可能会更具准确性与实用性。

检索增强生成：为大语言模型披上知识之铠

RAG（Retrieval-Augmented Generation，检索增强生成）是一种通过使用外部来源事实构建向量知识库，用于提升生成式人工智能准确性和可靠性的技术。具体而言，在大模型生成内容时，RAG 将首先在外部知识库中检索相关信息，随后依据相关信息大模型将生成更为准确的回答。基于可随时更新的外部知识库，RAG 在有效提升大模型生成内容的时效性与准确性的同时，降低了大模型产生“幻觉”现象的可能性。对于代码复现系统，RAG 能支撑海量文档并提供高效信息检索，为代码生成提供信息依据。

“GPT 如海”可复现因子计算和人工智能量化策略代码

我们针对不同应用场景设计了两套代码复现模板，分别是因子代码复现和人工智能量化策略代码复现。对于因子代码复现场景，测试结果显示“GPT 如海”能够较准确地提取出图片或 PDF 文档中的因子表达式以及因子构建步骤，并能进一步以此为基础构建因子计算代码；对于人工智能量化策略复现场景，测试结果显示，“GPT 如海”能够较准确地提取出人工智能量化策略的构建步骤和细节参数，并依次输出数据集构建、模型架构和模型训练的.py 代码，在代码细节上，我们观察到“GPT 如海”能够敏锐捕捉到策略构建的细节，例如损失函数设计等，尽管这些细节信息散布在文章各处。

多模型复现效果对比：小模型也可“纳须弥于芥子”

在大模型如火如荼的发展进程中，专为响应大模型落地需求的小型大语言模型也渐为兴盛。我们简易对比了 GPT-4、通义千问（Qwen 1.5，7B）和 Llama3（8B）的代码生成效果，在不同评价维度上，三个模型展现出差异性特征，例如 GPT-4 并不严格遵守代码模板，而通义千问和 Llama3 较为遵守代码模板。客观而言，通义千问和 Llama3 以小参数体量博得不错的代码生成效果，体现出小型模型“纳须弥于芥子”的潜力。

风险提示：大模型存在幻觉现象，模型生成结果可能不符合事实。大模型生成的代码可能存在错误，使用需谨慎。大模型提取信息可能存在遗漏。大模型训练集广泛，可能存在过拟合风险。

正文目录

导言 **3**

多模态大语言模型 **4**

RAG：检索增强生成 **6**

RAG vs Fine-tuning 7

RAG 实现：检索增强生成开源框架 Embedchain 8

方法 **9**

索引（Indexing） 9

检索（Retrieval） 9

生成（Generation） 9

部署（Deployment） 9

结果 **10**

WebUI 主要功能 10

因子计算代码复现 10

人工智能量化策略代码复现 11

多模型测试：代码复现效果对比 13

总结 **15**

参考文献 16

风险提示 16

导言

“善学者，其如海乎” ——《随园诗话·补遗 卷四》

自 OpenAI 推出 ChatGPT 以来，大语言模型（Large Language Model， LLM）强大的指令遵循、语义理解、内容生成等能力屡屡超出预期，在办公、编程、游戏等多个领域带来革命性进展。与此同时，多模态大语言模型（Multimodal Large Language Model， MLLM）将视觉（图像）、听觉（语音）感知和文本理解相结合，展现出更加广阔的应用前景。在华泰金工前期研究《GPT 因子工厂：多智能体与因子挖掘》（2024 年 2 月 20 日）中，我们利用基于大模型的多智能体系统实现端到端的量价因子挖掘系统，在因子构建、逻辑解释、代码生成等方面表现出色，展现出大语言模型赋能量化研究的潜力。

本文将继续探索多模态大模型和 RAG（Retrieval-Augmented Generation，检索增强生成）技术在量化研究中的应用。量化研究中多模态数据丰富，包括且不限于文本和图像等。以往的深度学习通常只能处理单一模态的数据，应用范围较为有限；多模态大模型则可以同时从多模态多类型的数据中抽取信息，显著提高模型的感知范围，在量化研究领域可能具有广阔的应用潜力。但另一方面，量化研究领域对内容的时效性、准确性有很高要求，而目前大模型的参数难以实时更新，并且生成的内容可能含有事实性错误，极大地限制了大模型在量化领域的应用。

RAG 技术可能成为大模型应用于量化研究领域的破局关键。RAG 技术可以从外部知识库中检索相关知识，辅助大模型生成内容，不仅极大地缓解了模型的幻觉现象，而且省去了微调大模型的昂贵成本。结合金融领域沉淀的丰富数据，检索增强生成技术适用于多种生成式任务，例如智能问答、综述总结、股票预测等，具有广阔的应用前景。

本文主要探究多模态大模型和 RAG 技术在代码复现中的应用。在量化研究领域，研究者通常需要广泛地阅读和学习海量研究报告，从中收集信息、汲取灵感，提高自身研究能力。然而，复现量化研究报告相关的代码往往较为复杂，不仅包含文本信息、图像信息的理解和处理，还包括处理数据、撰写代码等多个步骤。我们希望借助多模态大模型在图像感知、语义理解、代码生成等能力，结合 RAG 技术的准确性和可靠性，辅助研究员复现研究报告，提高复现效率。本研究基于 OpenAI 的 GPT-4 系列模型，通过检索增强生成的方式，将量化研究报告中的因子或人工智能量化策略转化为.py 代码文件，有效降低代码复现工作的复杂度，大幅提升工作效率。

图表1： 华泰金工-GPT 复现因子研报 WebUI 界面

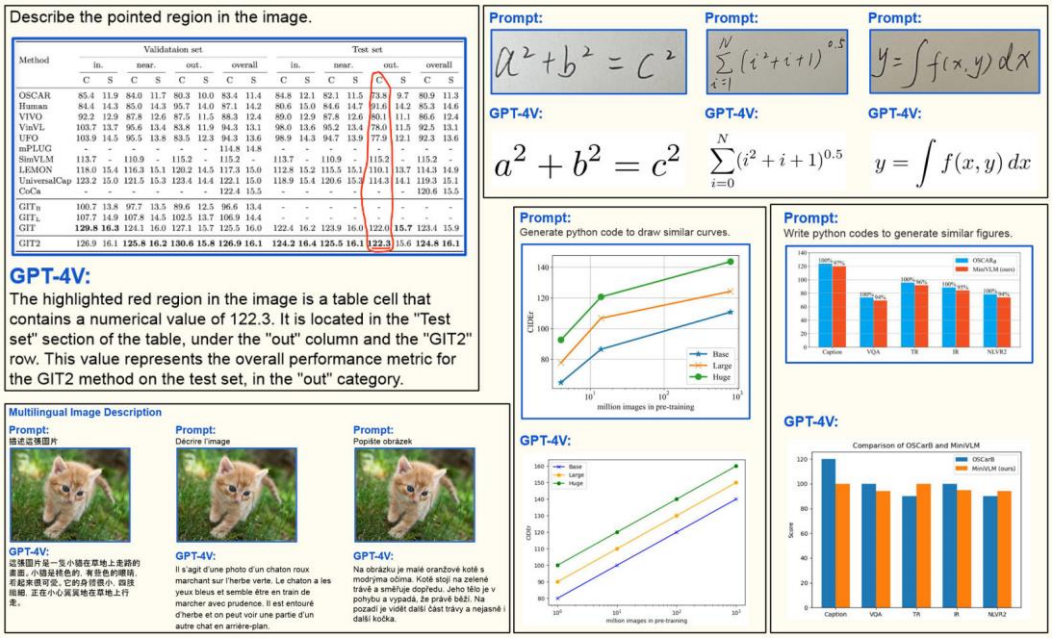
资料来源：Embedchain，Streamlit，OpenAI，华泰研究

多模态大语言模型

传统的神经网络仅能处理单一模态的数据，如视觉图像、文本、音频等。然而，人类在和外界环境交互时通常需要同时处理多种模态的信息，进而能更准确地感知、理解、决策。以研究报告为例，其中的图表能够直观准确地展现时序趋势或者条件比较等信息，这是文字和数字难以表述的。因此，为人工智能模型赋予多模态感知和生成能力，可能是实现通用人工智能（Artificial General Intelligence， AGI）的必由之路。

过去三年中，多模态大模型得到了长足的发展。OpenAI 于 2021 年提出 CLIP 模型，实现了文本数据和图像数据之间的映射。CLIP 模型不仅可以识别图像类型，而且可以生成图像、搜索与文本对应的图像，成为了目前多模态大语言模型的基石。DeepMind 在 2022 年推出 Flamingo 模型，将预训练视觉模型和语言模型相连接，使得大语言模型能够同时理解视觉和文本信息，并生成文本输出。2023 年 10 月，OpenAI 推出多模态大语言模型 GPT-4-Vision，其展现出强大的多模态感知能力和多任务能力。GPT-4-Vision 可以从图表中找到用户需要的数据，也可以回答图片相关的问题。除了将视觉信息融入大语言模型以外，研究者们还将语音信息和大语言模型相结合。目前，多模态大模型已经具备同时感知文本、图像、音频的能力，并且逐渐发展出生成多模态内容的能力。

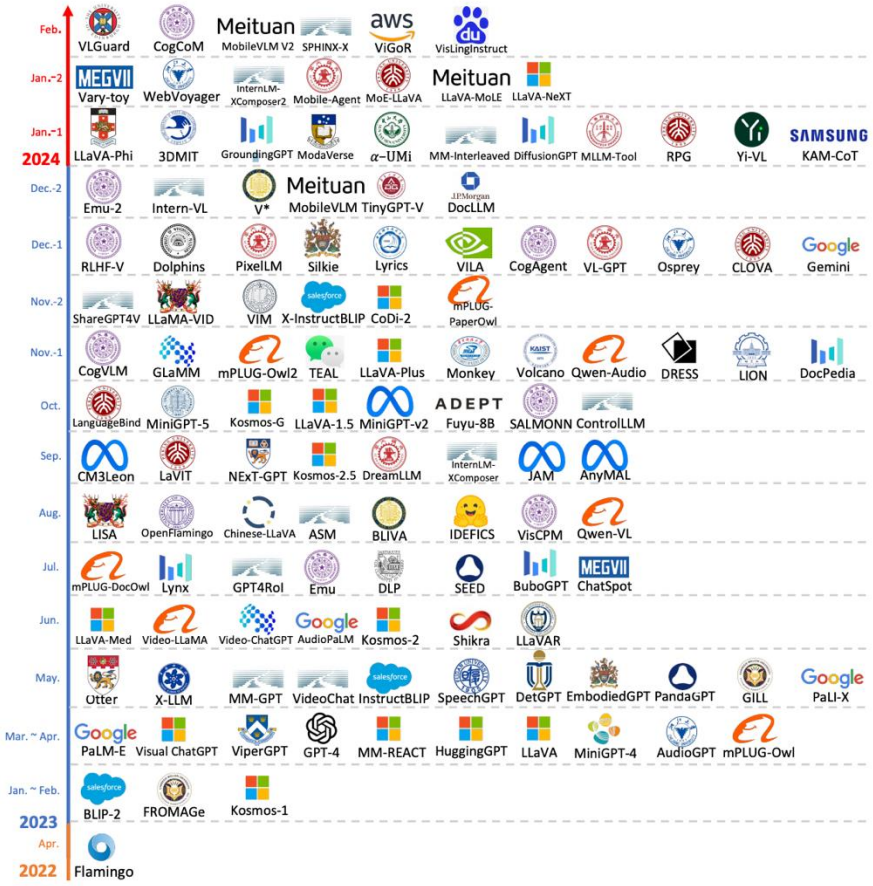
图表2： 多模态大语言模型 GPT-4-vision 能力测试示例



资料来源：*The Dawn of LMMs: Preliminary Explorations with GPT-4V(ision)*，华泰研究

多模态大模型已经成为人工智能发展的趋势，具有广阔的应用前景。多模态大模型在多种模态的感知和理解能力的加持下，可以从多模态数据中获取更为全面综合的信息，显著提高模型性能。此外，多模态大模型的多模态输入和输出能力使得人与模型的交互更加灵活多样，落地应用场景也更丰富多元，有望赋能娱乐、医疗、电商、游戏、金融等诸多领域。

图表3：多模态大语言模型发展时间线



资料来源：*MM-LLMs: Recent Advances in MultiModal Large Language Models*，华泰研究

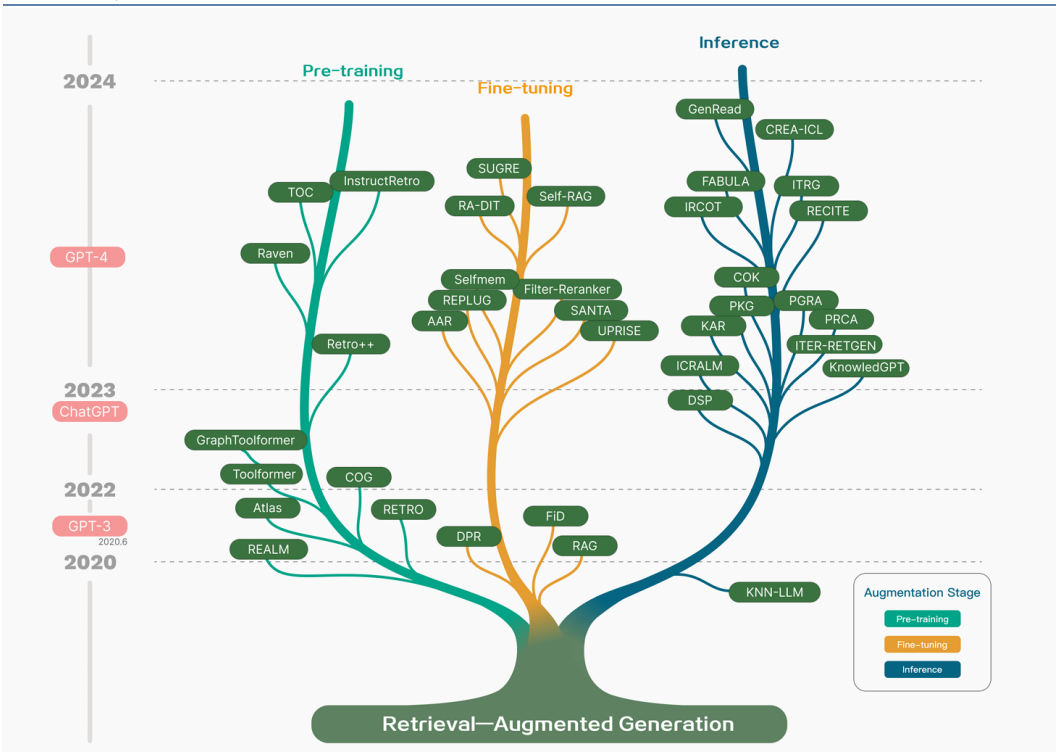
在量化研究领域中，研究报告同样蕴含较为丰富的多模态数据，包括文本和图像数据。通常意义上的大语言模型只能处理文本数据，却无法提取报告中以图表形式呈现的信息，因此在研究报告复现上可能具有较大的局限性。因此，本研究尝试借助多模态大模型的图像感知能力，助力研究报告相关代码的复现工程。

尽管如此，仅靠多模态大模型自身仍然难以完成这一任务。一方面，目前的主流多模态大模型的上下文长度有限，难以将研究报告文件直接作为模型输入；另一方面，GPT-4 等支持长文本输入的模型难以从量化研究报告中捕捉重点信息和复现细节，生成内容的质量可能不尽人意。因此，我们在多模态大模型的基础上，尝试叠加 RAG（检索增强生成）技术，从而显著提高模型的复现质量。

RAG：检索增强生成

RAG（Retrieval-Augmented Generation，检索增强生成）是一种通过使用外部来源获得的事实，用于提升生成式人工智能准确性和可靠性的技术。虽然目前大模型展现出很强的生成能力，但由于其训练数据和模型结构的局限性，生成的内容往往并不准确甚至脱离现实，也即幻觉（Hallucination）现象的主要来源之一。为了提高大模型生成内容的时效性、准确性，研究者们提出了 RAG 方法，已有研究者如 Peng et al. (2023)的研究显示出检索增强能有效降低幻觉现象。

图表4：检索增强生成技术进化树

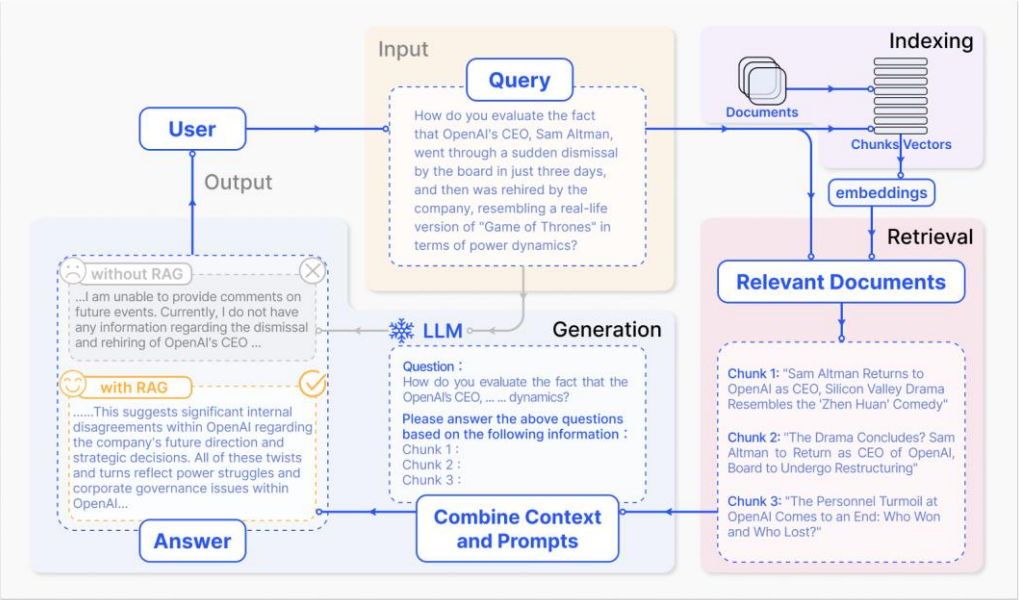


资料来源：*Retrieval-Augmented Generation for Large Language Models: A Survey*，华泰研究

具体而言，RAG 是指在大模型生成内容时，首先在外部知识库中检索相关信息，然后再根据相关信息生成回答的技术。经典的检索增强生成技术包括三个步骤，分别是索引（Indexing）、检索（Retrieval）和生成（Generation）：

1. **索引（Indexing）**：构建外部知识库的过程。首先需要收集任务相关的文本数据；然后将这些文本数据切分为更小的块（Chunk），以便适配大模型的输入长度；最后进行嵌入和索引。具体而言，嵌入模型（Embedding Model）将文本段落编码映射为固定长度的向量，并利用嵌入向量和文本段落建立起键值对，储存在向量数据库中，形成索引。
2. **检索（Retrieval）**：根据用户输入的问题检索相关的外部知识。首先利用嵌入模型将用户的问题映射为嵌入向量；然后和外部数据库中的嵌入向量计算相似性，将相似性最高的 K 个嵌入向量对应的文本块作为当前问题的补充信息。
3. **生成（Generation）**：大模型根据用户指令和补充信息生成最终结果。系统将检索得到的 K 条文本段落和用户的问题合并为新的提示信息，将其作为大模型的输入并得到最终的结果。

图表5：RAG 应用架构示意图



资料来源：*Retrieval-Augmented Generation for Large Language Models: A Survey*，华泰研究

RAG 使得模型不仅可以利用其内部知识，还可以利用外部知识辅助生成。其优点在于：

1. 不需要微调大模型，大大降低了使用成本；
2. 不同的知识库可以适配不同的任务，同时知识库可以不断更新，更加动态灵活；
3. 大模型生成的答案可以追溯到具体的数据来源，具有更强的可追溯性和可靠性。

因此，检索增强生成是大模型落地应用的关键技术之一。为了准确复现报告中涉及的代码，我们采用 RAG 技术，先从图片或文档中检索因子或人工智能量化策略相关信息，再由大模型进行复现。

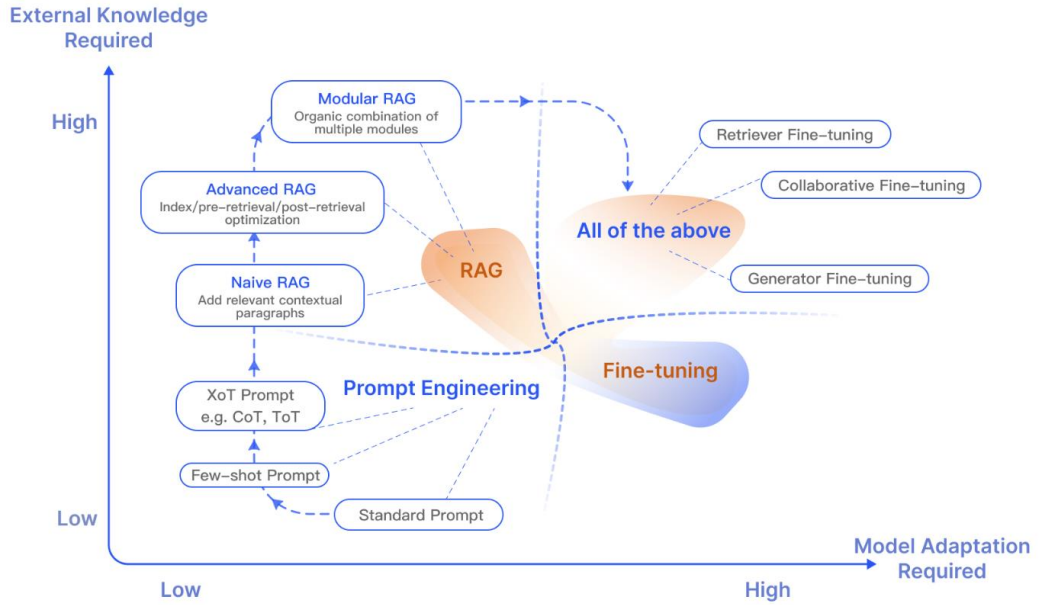
RAG vs Fine-tuning

在构建大语言模型应用程序时，我们通常可使用两种方式整合特定领域的知识：RAG 和微调（Fine-tuning）。RAG 与微调各有其优势，我们可以从外部知识需求和模型适应性需求两个维度进行对比：

1. **外部知识需求方面**：RAG 通常高于微调技术，RAG 需要大量丰富的外部知识构建向量知识库，以便于后续的检索和生成任务；而微调技术所需的往往是一个优质训练集，训练集知识纳入模型内部之后，我们往往寄希望于模型通过顿悟（Grokking）或涌现（Emergence）为我们提供惊喜；
2. **模型适应性需求方面**：微调通常高于 RAG 机制，通过特定垂直领域知识的学习，微调后的模型善于捕捉风格的微妙变化、特定领域的词汇等等，而 RAG 机制依赖于静态模型，风格切换或特定能力维度可能弱于微调技术。

但 RAG 和微调技术并非针锋相对或水火不容，同时结合 RAG 和微调技术可能会创造出既具有广泛知识又拥有特定领域专长的大模型应用。举例而言，Zhang et al. (2024)提出 RAFT（Retrieval-Augmented Fine-Tuning，检索增强型微调）的技术，通过构造有价值文档和干扰文档数据集对大模型进行微调，从而使得大模型能够更好地执行 RAG，提升其通过检索和推理回答问题的能力。

图表6：RAG 与微调技术的对比



资料来源：*Retrieval-Augmented Generation for Large Language Models: A Survey*，华泰研究

RAG 实现：检索增强生成开源框架 Embedchain

Embedchain 是一个用于实现 RAG 架构的轻量级开源框架，专注于快速搭建基于 RAG 的大模型问答系统。该框架集成了加载数据、创建嵌入向量、向量数据库存储和相似度检索等多项功能，不仅可以处理各种非结构化数据，包括文本、PDF 文件、网页、图像、视频等，还可以无缝衔接主流的大模型，包括 OpenAI 的 GPT 系列、Meta 的 Llama 系列和 Anthropic 的 Claude 系列模型等。除此之外，基于 Embedchain 构建的应用可以轻松通过 Streamlit 和 Gradio 等框架进行部署。本研究依赖于 Embedchain 框架，搭建 RAG 系统并使用 Streamlit 进行部署。

图表7：Embedchain 框架的文件及模型支持

资料来源：Embedchain 官网，华泰研究

方法

本研究基于多模态大模型和 RAG 技术进行代码复现，可实现从图片和 PDF 文件等多模态数据中获得因子构建信息或人工智能量化策略，并直接生成对应的.py 代码文件，实现端到端的代码复现流程。本系统遵循基本的 RAG 架构，分为四个步骤，分别是索引、检索、生成以及最终的部署。

索引（Indexing）

索引阶段负责根据量化研究报告构建外部知识库。首先，系统会根据用户输入的文件格式进行数据预处理，将多模态的数据输入统一转换为文本数据。如 PDF 格式的文件会被解析为文本部分和图片部分，而图片部分会通过 GPT-4 vision 模型转换为文本描述；若单独输入图片格式的文件，同样会经此处理。随后，预处理所得的文本数据会被分割为相同大小的块，并通过 OpenAI 的嵌入模型 text-embedding-3-large 转化为相同长度的向量，存储至向量数据库中。

检索（Retrieval）

检索阶段负责从外部知识库中检索相关知识。为了精确复现代码，我们将复现步骤进一步细化，例如复现人工智能量化策略被细分为“输入和输出数据”“数据预处理方法”“网络结构”“训练参数”四个步骤，每个步骤都有对应的查询语句。在检索时，查询语句先由嵌入模型映射为向量，然后与向量数据库中的向量计算相似度。相似度最高的 K 个向量对应的文本块会作为该复现步骤的相关内容，辅助模型生成。

生成（Generation）

生成阶段分为三个阶段：（1）首先，检索阶段获得的查询语句和相关知识会输入大模型，由模型总结生成对应的研究报告细节，如神经网络类因子的数据、结构、参数等；（2）其次，我们设计了“自我反思”环节。我们将第一步获得的报告细节输入大模型，让模型对目前的复现内容进行反思，生成进一步的优化建议，从而辅助研究员阅读研究报告、改进复现代码；（3）最后，系统将自然语言格式的复现细节输入大模型，由模型生成各个复现步骤对应的代码文件。为了提高代码质量，我们设计了代码模版，辅助模型生成。由此，我们实现了基于多模态大模型和RAG的量化代码复现系统，我们为它起了一个有趣的名字：GPT 如海，寓意其善学如海。

图表8：基于多模态大模型和 RAG 的代码复现系统具体参数

步骤	参数	参数值
索引	块大小/块间重叠大小	500/200
	多模态大模型	gpt-4-vision-preview
	嵌入模型	text-embedding-3-large
	向量数据库	Chroma
	嵌入向量长度	1536
检索	相似度度量	余弦相似度
	K	15
生成	大模型	如无特殊说明，默认为 gpt-4-0125-preview

资料来源：华泰研究

部署（Deployment）

“GPT 如海”构建完毕后，我们基于 Streamlit 框架将整个系统部署于本地 HTTP 端口，通过本地 localhost（默认 8501 端口）即可轻松访问，既方便文件上传，也便于生成代码后直接下载。

结果

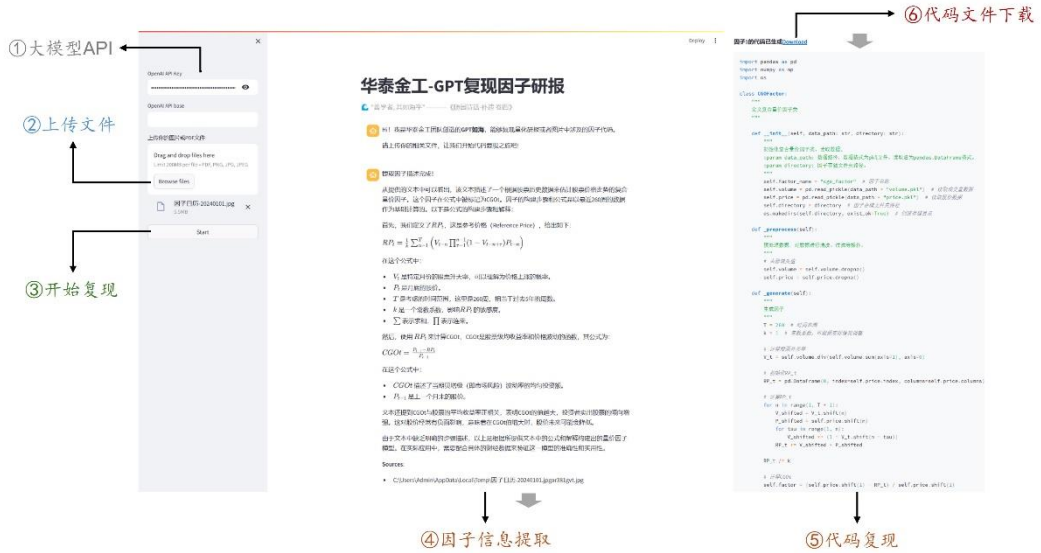
对于“GPT 如海”系统，我们一共设计了两个版本，其一专门用于复现图片或 PDF 文档中涉及的量化因子，其二则用于复现图片或 PDF 文档中的人工智能量化策略，二者在界面和使用上几乎没有差异。以下我们将通过界面展示和代码功能展示对“GPT 如海”进行更为详尽的展示。

WebUI 主要功能

在部署完毕的“GPT 如海”系统中，关于代码复现的各项功能已趋于完善。系统运行严格遵循以下工作流：

1. **填入大模型 API**：填入 OpenAI 或其他可调用模型的 API，以供代码复现系统生成代码；
2. **上传文件**：支持图片或 PDF 格式文件上传，可上传多个文件；
3. **开始复现**：点击 Start 按钮代码复现系统即开始运转，无需自行输入或设定提示（Prompt）；
4. **因子信息提取**：系统中大模型将从 RAG 架构中的向量数据库提取复现代码所需信息，例如因子表达式和因子构建步骤，并将其输出，便于人工监督检查；
5. **代码复现**：大模型输出的代码将以代码风格展示（仅支持 Python 代码），便于阅读；
6. **代码文件下载**：点击代码旁的 Download 可直接将.py 文件下载至本地。

图表9：WebUI 界面功能解析（因子代码复现）



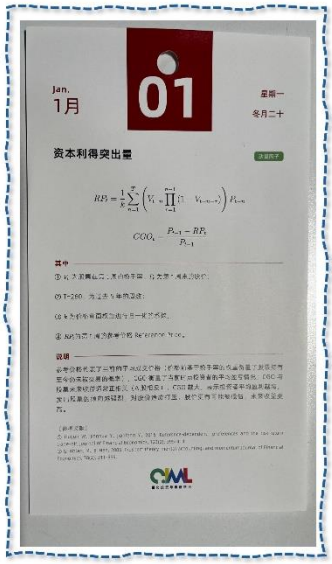
资料来源：Embedchain，Streamlit，OpenAI，华泰研究

因子计算代码复现

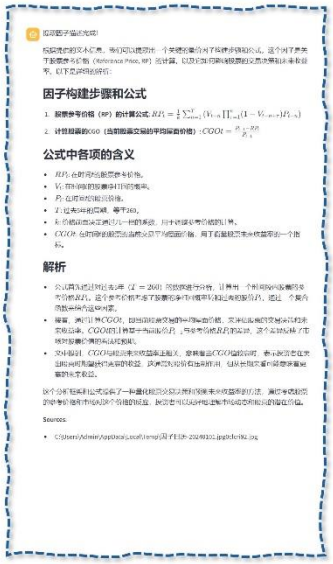
首先我们测试“GPT 如海”系统的因子代码复现能力。本文尝试向系统输入一张用手机拍摄的.jpg 格式照片（像素大小为 2268×4032），图片中包含一个较复杂量价选股因子的表达式和解析。从大模型输出中可以看到，“GPT 如海”可以从具备噪音（如阴影、褶皱等）的图片中准确无误地提取出因子数学表达式，并使用 Latex 输出，显示出其依赖的多模态大模型（即 gpt-4-vision-preview）的多模态信息提取能力。随后，“GPT 如海”输出该因子的计算代码，从代码上看，复现代码基本符合因子构建步骤。

图表10：单因子代码复现（图像文件输入）

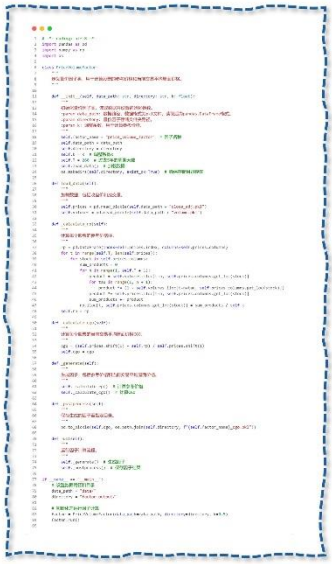
①输入文件为包含因子详细构建信息的图片



②“GPT如海”从图片中准确提取因子公式和构建步骤



③“GPT如海”直接输出因子构建代码

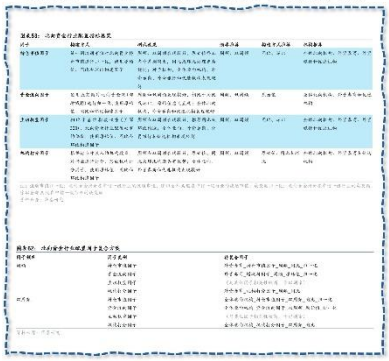


资料来源：Embedchain，Streamlit，OpenAI，量化投资与机器学习，华泰研究

对于因子复现版本的“GPT 如海”，其同样支持多个因子的信息提取与代码复现。在华泰金工前期报告《析精剖微：机构拆解看北向资金》（2022 年 10 月 27 日）中，我们定义了 4 类可用于行业配置的北向因子：持仓市值因子、资金流向因子、主动权重因子和机构打分因子，但最终构建行业配置策略时由于主动权重因子与其他因子相关性较高，只纳入了其余 3 个因子。当我们将完整的 38 页数据输入复现系统后，“GPT 如海”成功提取用于构建策略的 3 个因子相关信息，并依次生成 3 个因子的.py 计算代码，显著表明多模态大模型对表格图片的理解，以及 RAG 系统加持下大模型的信息提取能力和代码生成能力。

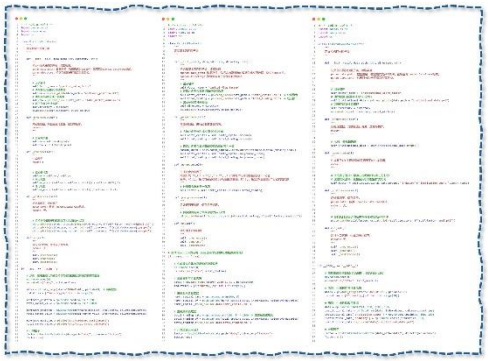
图表11：多因子代码复现（PDF 文件输入）

①输入文件为原始PDF文件，因子构建信息位于第25页图表53，共4个因子，在构建策略时，仅选择3个因子



②模型成功提取3个构建策略因子的信息

③模型依次生成三个因子的.py代码



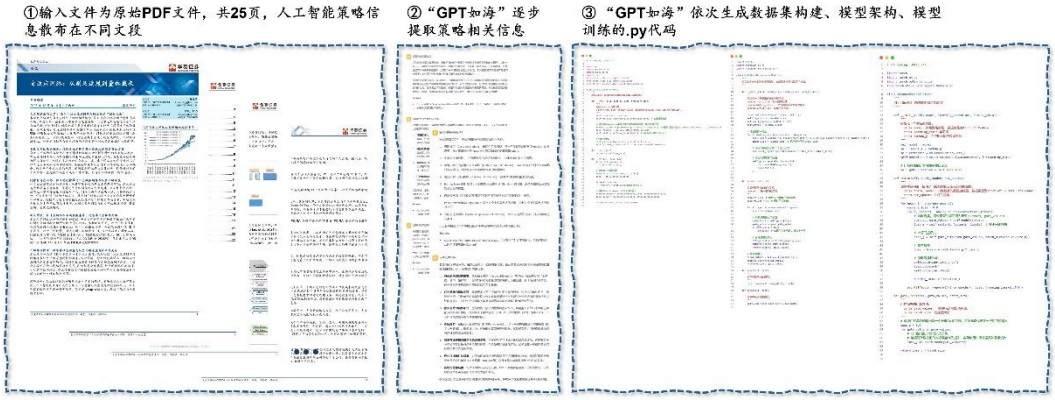
资料来源：Embedchain，Streamlit，OpenAI，华泰研究

人工智能量化策略代码复现

除因子复现外，“GPT 如海”还可进行人工智能量化策略代码的复现。随量化研究领域的日益发展，量化模型渐趋复杂，人工智能技术的应用也愈发广泛，在学习公开论文或研究报告时难免需要复现相关代码。对于一个较完善的人工智能策略而言，可能包含诸多组成部分，包括数据集构建、模型架构搭建和模型训练评估等等，在良好代码能力的基础上想要复现也是一件需花费颇多精力的事情。

华泰金工团队已公开发布 70 余篇人工智能相关研报，我们选取其中两篇让“GPT 如海”尝试复现。第一篇报告为华泰金工人工智能系列第 71 篇：《自适应网络：从削足适履到量体裁衣》（2023 年 12 月 1 日），此篇报告主要探究动态自适应网络架构能否在传统网络的基础上带来效果提升，其中动态自适应网络概念繁杂、变体众多，相关代码的实现可能具备一定难度。我们直接将 25 页原始 PDF 文档输入复现系统，“GPT 如海”可逐步提取出因子数据集、数据预处理、训练参数、神经网络架构等策略相关信息，同时进行自我反思，用于判断复现依赖信息的缺失与否。信息提取完毕后，“GPT 如海”依次生成数据集构建、模型架构和模型训练的 3 个.py 代码。

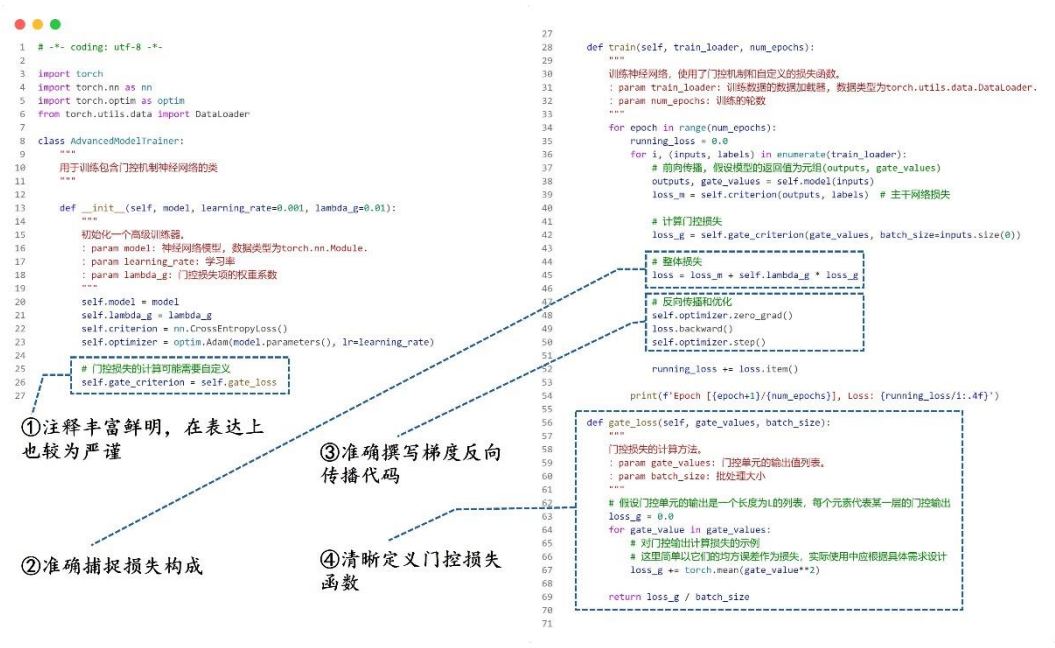
图表12：人工智能量化策略代码复现示例一（PDF 文件输入）



资料来源：Embedchain，Streamlit，OpenAI，华泰研究

我们可详细观察“GPT 如海”代码复现的特点。对于代码注释而言，注释丰富鲜明，表达上也较为谨慎，有效提升了代码的可读性和可参考性；对于原文核心信息捕捉上，“GPT 如海”复现系统也较为准确，例如原文强调神经网络架构的整体损失为主干网络和门控网络的合成，复现代码中也的确书写正确，但“GPT 如海”也并非没有错误，例如主干网络的损失实际是收益率衰减加权的 mse 损失，但它写为交叉熵损失（代码第 22 行）；在代码准确性上，我们发现“GPT 如海”依赖的 gpt-4-0125-preview 模型能力较强，无论是神经网络的反向传播、类的撰写以及函数的调用等等都较为准确，表明该系统的代码复现水平较高。

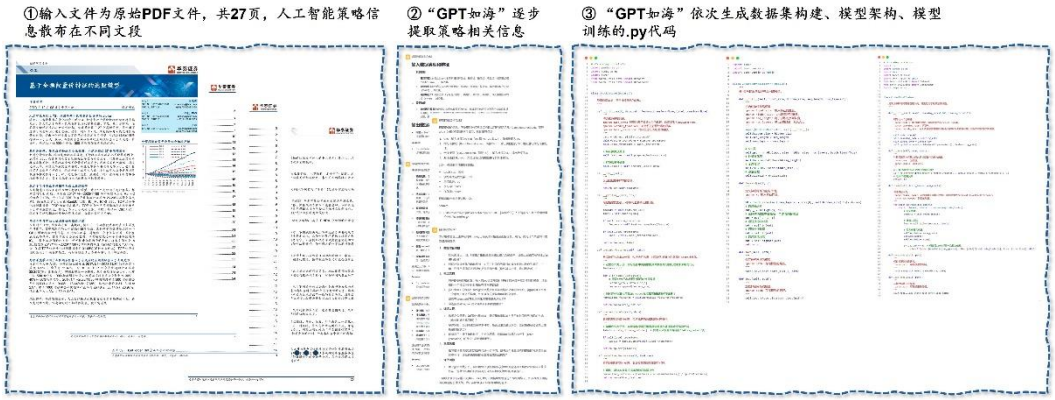
图表13：人工智能量化策略代码复现示例一中的代码亮点（以训练代码为例）



资料来源：Embedchain，Streamlit，OpenAI，华泰研究

第二篇测试人工智能量化策略代码的文章为华泰金工人工智能系列第 72 篇：《基于全频段量价特征的选股模型》（2023 年 12 月 8 日），此篇报告聚焦于高频和低频量价数据的信息挖掘，基于多频段量价数据构建了一个端到端神经网络选股模型。我们将 27 页原始 PDF 文档输入“GPT 如海”复现系统，策略相关信息同样被稳定提取出来，“GPT 如海”根据这些信息依次生成数据集和神经网络相关代码。

图表14：人工智能量化策略代码复现示例二（PDF 文件输入）



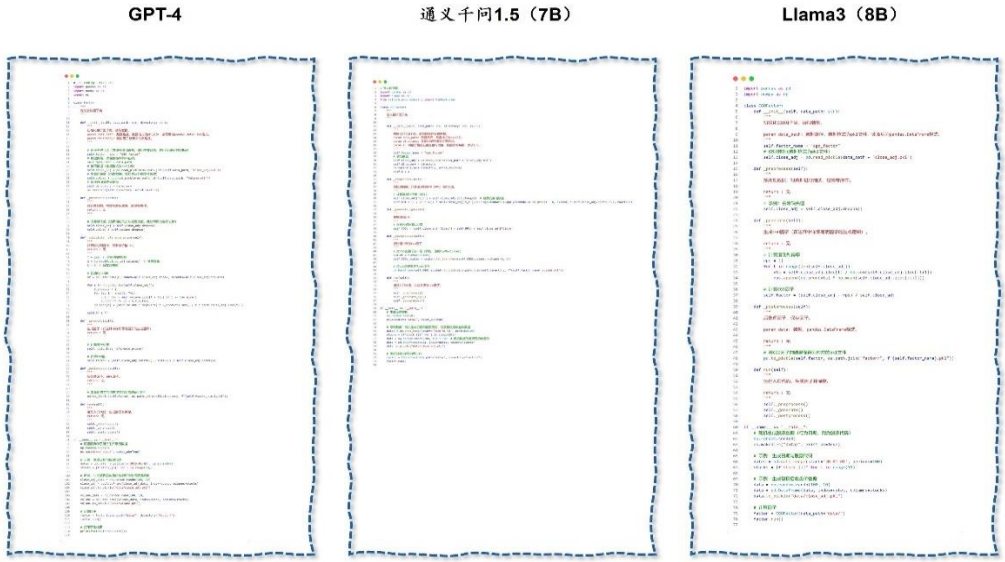
资料来源：Embedchain，Streamlit，OpenAI，华泰研究

多模型测试：代码复现效果对比

随着大语言模型的发展，大模型的终端落地需求逐步推动小型模型的开发迭代进程，迄今已有一批较出色的小型大语言模型脱颖而出，例如阿里巴巴公司推出的开源大模型通义千问、Meta 公司推出的 Llama 系列开源大模型。小型模型部署便利、输出速度较快的优势使得个人能够以较低廉的成本部署使用大语言模型。为检验小型模型在代码复现上的效果，我们选择两个开源模型与 GPT-4 进行对比，模型分别为 7B 参数量的 Qwen 1.5 和 8B 参数量的 Llama3。在对比测试中，“GPT 如海”除生成代码的模型改为 Qwen 1.5 或 Llama3，其余设置均不变。

我们测试了如图表 10 中的单因子图片输入，三个模型输出的代码复现结果如下图所示。简单从输出代码行数上看，GPT-4 输出的代码行数最多，Llama3 次之，通义千问输出的代码长度最短。

图表15：多模型代码复现效果对比



资料来源：Embedchain，Streamlit，OpenAI，HuggingFace，LM-Studio，华泰研究

我们详细解读三份代码，可以发现三个模型在不同维度上显示出不同特点。在代码能否直接运行方面，仅有 GPT-4 输出的代码可以直接运行，而通义千问和 Llama3 输出的代码均会报错；在代码正确情况上，我们详细阅读了三份代码，发现或多或少都存在一些错误，但 GPT-4 的错误是最少的，几乎接近正确答案；在是否遵循代码模板方面，通义千问和 Llama3 均在遵守模板的情况下生成了代码，而 GPT-4 与模板略有不同，增加了一个因子项的计算函数；在注释丰富程度上，GPT-4 和通义千问均较丰富，Llama3 的注释则简单不少。值得强调的是，以上对比仅为简易探索，且仅进行了一次代码生成测试，可能不具备结论性，评价维度与内容也较为主观，作为个人观点仅供参考。

图表16：多模型代码复现效果对比详细信息

	GPT-4	通义千问	Llama3
模型全称	gpt-4-0125-preview	Qwen1.5-7B-Chat	Llama3-8B-Chinese-Chat
参数量	-	7B	8B
来源	OpenAI	Hugging Face: Qwen/Qwen1.5-7B-Chat-GGUF	Hugging Face: shenzhi-wang/Llama3-8B-Chinese-Chat-GGUF-8bit
代码能否直接运行	**是**	否	否
代码正确情况	**少量错误**	部分正确	部分正确
是否遵循代码模板	否	**是**	**是**
注释丰富程度	**丰富**	**丰富**	一般
模型运行环境	联网 api 调用	本地运行	本地运行
模型来源	OpenAI	Hugging Face: Qwen/Qwen1.5-7B-Chat-GGUF	Hugging Face: shenzhi-wang/Llama3-8B-Chinese-Chat-GGUF-8bit

资料来源：OpenAI，HuggingFace，LM-Studio，华泰研究

总结

本文在华泰金工前期研究《GPT 因子工厂：多智能体与因子挖掘》（2024 年 2 月 20 日）的基础上，继续探索大语言模型在量化研究领域的潜在应用。以代码复现需求为切入点，我们基于多模态大模型和 RAG 技术构建了一个完善易用的代码复现框架，“GPT 如海”。在框架内部，“GPT 如海”既依赖于多模态大模型对于不同模态信息的提取能力，也离不开 RAG 技术对海量文档的检索支持，最终使得基本的 GPT-4 模型在应用框架的加持下发挥出了极为出色的代码复现效果。

我们针对不同应用场景设计了两套代码复现模板，分别是因子代码复现和人工智能量化策略代码复现。对于因子代码复现场景，我们尝试向系统输入包含因子构建信息的图片或 PDF 文档，测试结果显示“GPT 如海”能够较准确地提取出图片或 PDF 文档中的因子表达式以及因子构建步骤，并能进一步以此为基础构建因子计算代码；除此之外，如果 PDF 文档中包含多个因子，即使因子信息分散在不同文段，“GPT 如海”也能有效检索，并逐个输出因子代码。

对于人工智能量化策略复现场景，我们向系统输入两篇华泰金工人工智能深度报告：《自适应网络：从削足适履到量体裁衣》（2023 年 12 月 1 日）和《基于全频率量价特征的选股模型》（2023 年 12 月 8 日）。测试结果显示，“GPT 如海”能够较准确地提取人工智能量化策略的构建步骤和细节参数，并依次输出数据集构建、模型架构和模型训练的.py 代码。在代码细节上，我们观察到“GPT 如海”敏锐捕捉到策略构建的细节，例如损失函数设计等，尽管这些细节信息散布在文章各处。

在大模型如火如荼的发展进程中，专为响应大模型落地需求的小型大语言模型也渐为兴盛。我们简易对比了 GPT-4、通义千问（Qwen 1.5，7B）和 Llama3（8B）的代码生成效果，在不同评价维度上，三个模型展现出差异性的特征，例如 GPT-4 并不严格遵守代码模板，而通义千问和 Llama3 较为遵守代码模板。不过对比结果仅依赖于单次代码生成，不具备结论性，仅供参考。

本文作为大语言模型在量化研究领域代码复现场景下的应用探索，搭建的“GPT 如海”工具已趋于完善，但仍有诸多不足之处。例如，“GPT 如海”系统效果不稳定，有时无法提取出因子信息或提取信息有误；作为大语言模型为基础的应用系统，“GPT 如海”内部的提示工程仍存在可提升之处；此外，WebUI 部署设计较为简单，值得进一步优化。

参考文献

Yang, Z., Li, L., Lin, K., Wang, J., Lin, C. C., Liu, Z., & Wang, L. (2023). The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421, 9(1), 1*.

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., ... & Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Peng, B., Galley, M., He, P., Cheng, H., Xie, Y., Hu, Y., ... & Gao, J. (2023). Check your facts and try again: Improving large language models with external knowledge and automated feedback. *arXiv preprint arXiv:2302.12813*.

Zhang, D., Yu, Y., Li, C., Dong, J., Su, D., Chu, C., & Yu, D. (2024). Mm-llms: Recent advances in multimodal large language models. *arXiv preprint arXiv:2401.13601*.

风险提示

大模型存在幻觉现象，模型生成结果可能不符合事实。大模型生成的代码可能存在错误，使用需谨慎。大模型提取信息可能存在遗漏。大模型训练集广泛，可能存在过拟合风险。

免责声明

分析师声明

本人，林晓明、何康，兹证明本报告所表达的观点准确地反映了分析师对标的证券或发行人的个人意见；彼以往、现在或未来并无就其研究报告所提供的具体建议或所表达的意见直接或间接收取任何报酬。

一般声明及披露

本报告由华泰证券股份有限公司（已具备中国证监会批准的证券投资咨询业务资格，以下简称“本公司”）制作。本报告所载资料是仅供接收人的严格保密资料。本报告仅供本公司及其客户和其关联机构使用。本公司不因接收人收到本报告而视其为客户。

本报告基于本公司认为可靠的、已公开的信息编制，但本公司及其关联机构(以下统称为“华泰”)对该等信息的准确性及完整性不作任何保证。

本报告所载的意见、评估及预测仅反映报告发布当日的观点和判断。在不同时期，华泰可能会发出与本报告所载意见、评估及预测不一致的研究报告。同时，本报告所指的证券或投资标的的价格、价值及投资收入可能会波动。以往表现并不能指引未来，未来回报并不能得到保证，并存在损失本金的可能。华泰不保证本报告所含信息保持在最新状态。华泰对本报告所含信息可在不发出通知的情形下做出修改，投资者应当自行关注相应的更新或修改。

本公司不是 FINRA 的注册会员，其研究分析师亦没有注册为 FINRA 的研究分析师/不具有 FINRA 分析师的注册资格。

华泰力求报告内容客观、公正，但本报告所载的观点、结论和建议仅供参考，不构成购买或出售所述证券的要约或招揽。该等观点、建议并未考虑到个别投资者的具体投资目的、财务状况以及特定需求，在任何时候均不构成对客户私人投资建议。投资者应当充分考虑自身特定状况，并完整理解和使用本报告内容，不应视本报告为做出投资决策的唯一因素。对依据或者使用本报告所造成的一切后果，华泰及作者均不承担任何法律责任。任何形式的分享证券投资收益或者分担证券投资损失的书面或口头承诺均为无效。

除非另行说明，本报告中所引用的关于业绩的数据代表过往表现，过往的业绩表现不应作为日后回报的预示。华泰不承诺也不保证任何预示的回报会得以实现，分析中所做的预测可能是基于相应的假设，任何假设的变化可能会显著影响所预测的回报。

华泰及作者在自身所知情的范围内，与本报告所指的证券或投资标的不存在法律禁止的利害关系。在法律许可的情况下，华泰可能会持有报告中提到的公司所发行的证券头寸并进行交易，为该公司提供投资银行、财务顾问或者金融产品等相关服务或向该公司招揽业务。

华泰的销售人员、交易人员或其他专业人士可能会依据不同假设和标准、采用不同的分析方法而口头或书面发表与本报告意见及建议不一致的市场评论和/或交易观点。华泰没有将此意见及建议向报告所有接收者进行更新的义务。华泰的资产管理部门、自营部门以及其他投资业务部门可能独立做出与本报告中的意见或建议不一致的投资决策。投资者应当考虑到华泰及/或其相关人员可能存在影响本报告观点客观性的潜在利益冲突。投资者请勿将本报告视为投资或其他决定的唯一信赖依据。有关该方面的具体披露请参照本报告尾部。

本报告并非意图发送、发布给在当地法律或监管规则下不允许向其发送、发布的机构或人员，也并非意图发送、发布给因可得到、使用本报告的行为而使华泰违反或受制于当地法律或监管规则的机构或人员。

中国香港

本报告由华泰证券股份有限公司制作,在香港由华泰金融控股（香港）有限公司向符合《证券及期货条例》及其附属法律规定的机构投资者和专业投资者的客户进行分发。华泰金融控股（香港）有限公司受香港证券及期货事务监察委员会监管，是华泰国际金融控股有限公司的全资子公司，后者为华泰证券股份有限公司的全资子公司。在香港获得本报告的人员若有任何有关本报告的问题,请与华泰金融控股（香港）有限公司联系。

香港-重要监管披露

- 华泰金融控股（香港）有限公司的雇员或其关联人士没有担任本报告中提及的公司或发行人的高级人员。
- 有关重要的披露信息，请参华泰金融控股（香港）有限公司的网页 https://www.htsc.com.hk/stock_disclosure

其他信息请参见下方 **“美国-重要监管披露”**。

美国

在美国本报告由华泰证券（美国）有限公司向符合美国监管规定的机构投资者进行发表与分发。华泰证券（美国）有限公司是美国注册经纪商和美国金融业监管局（FINRA）的注册会员。对于其在美国分发的研究报告，华泰证券（美国）有限公司根据《1934 年证券交易法》（修订版）第 15a-6 条规定以及美国证券交易委员会人员解释，对本研究报告内容负责。华泰证券（美国）有限公司联营公司的分析师不具有美国金融监管（FINRA）分析师的注册资格，可能不属于华泰证券（美国）有限公司的关联人员，因此可能不受 FINRA 关于分析师与标的公司沟通、公开露面和所持交易证券的限制。华泰证券（美国）有限公司是华泰国际金融控股有限公司的全资子公司，后者为华泰证券股份有限公司的全资子公司。任何直接从华泰证券（美国）有限公司收到此报告并希望就本报告所述任何证券进行交易的人士，应通过华泰证券（美国）有限公司进行交易。

美国-重要监管披露

- 分析师林晓明、何康本人及相关人士并不担任本报告所提及的标的证券或发行人的高级人员、董事或顾问。分析师及相关人士与本报告所提及的标的证券或发行人并无任何相关财务利益。本披露中所提及的“相关人士”包括 FINRA 定义下分析师的家庭成员。分析师根据华泰证券的整体收入和盈利能力获得薪酬，包括源自公司投资银行业务的收入。
- 华泰证券股份有限公司、其子公司和/或其联营公司，及/或不时会以自身或代理形式向客户出售及购买华泰证券研究所覆盖公司的证券/衍生工具，包括股票及债券（包括衍生品）华泰证券研究所覆盖公司的证券/衍生工具，包括股票及债券（包括衍生品）。
- 华泰证券股份有限公司、其子公司和/或其联营公司，及/或其高级管理层、董事和雇员可能会持有本报告中所提到的任何证券（或任何相关投资）头寸，并可能不时进行增持或减持该证券（或投资）。因此，投资者应该意识到可能存在利益冲突。

新加坡

华泰证券（新加坡）有限公司持有新加坡金融管理局颁发的资本市场服务许可证，可从事资本市场产品交易，包括证券、集体投资计划中的单位、交易所交易的衍生品合约和场外衍生品合约，并且是《财务顾问法》规定的豁免财务顾问，就投资产品向他人提供建议，包括发布或公布研究分析或研究报告。华泰证券（新加坡）有限公司可能会根据《财务顾问条例》第 32C 条的规定分发其在华泰内的外国附属公司各自制作的信息/研究。本报告仅供认可投资者、专家投资者或机构投资者使用，华泰证券（新加坡）有限公司不对本报告内容承担法律责任。如果您是非预期接收者，请您立即通知并直接将本报告返回给华泰证券（新加坡）有限公司。本报告的新加坡接收者应联系您的华泰证券（新加坡）有限公司关系经理或客户主管，了解来自或与所分发的信息相关的事宜。

评级说明

投资评级基于分析师对报告发布日后 6 至 12 个月内行业或公司回报潜力（含此期间的股息回报）相对基准表现的预期（A 股市场基准为沪深 300 指数，香港市场基准为恒生指数，美国市场基准为标普 500 指数，台湾市场基准为台湾加权指数，日本市场基准为日经 225 指数，新加坡市场基准为海峡时报指数，韩国市场基准为韩国有价证券指数），具体如下：

行业评级

增持：预计行业股票指数超越基准
中性：预计行业股票指数基本与基准持平
减持：预计行业股票指数明显弱于基准

公司评级

买入：预计股价超越基准 15%以上
增持：预计股价超越基准 5%~15%
持有：预计股价相对基准波动在-15%~5%之间
卖出：预计股价弱于基准 15%以上
暂停评级：已暂停评级、目标价及预测，以遵守适用法规及/或公司政策
无评级：股票不在常规研究覆盖范围内。投资者不应期待华泰提供该等证券及/或公司相关的持续或补充信息

法律实体披露

中国：华泰证券股份有限公司具有中国证监会核准的“证券投资咨询”业务资格，经营许可证编号为：91320000704041011J

香港：华泰金融控股（香港）有限公司具有香港证监会核准的“就证券提供意见”业务资格，经营许可证编号为：AOK809

美国：华泰证券（美国）有限公司为美国金融业监管局（FINRA）成员，具有在美国开展经纪交易商业务的资格，经营业务许可编号为：CRD#:298809/SEC#:8-70231

新加坡：华泰证券（新加坡）有限公司具有新加坡金融管理局颁发的资本市场服务许可证，并且是豁免财务顾问。公司注册号：202233398E

华泰证券股份有限公司

南京
南京市建邺区江东中路 228 号华泰证券广场 1 号楼/邮政编码：210019

电话：86 25 83389999/传真：86 25 83387521
电子邮件：ht-rd@htsc.com

北京
北京市西城区太平桥大街丰盛胡同 28 号太平洋保险大厦 A 座 18 层/邮政编码：100032
电话：86 10 63211166/传真：86 10 63211275
电子邮件：ht-rd@htsc.com

深圳
深圳市福田区益田路 5999 号基金大厦 10 楼/邮政编码：518017
电话：86 755 82493932/传真：86 755 82492062
电子邮件：ht-rd@htsc.com

上海
上海市浦东新区东方路 18 号保利广场 E 栋 23 楼/邮政编码：200120
电话：86 21 28972098/传真：86 21 28972068
电子邮件：ht-rd@htsc.com

华泰金融控股（香港）有限公司
香港中环皇后大道中 99 号中环中心 53 楼
电话：+852-3658-6000/传真：+852-2567-6123
电子邮件：research@htsc.com
http://www.htsc.com.hk

华泰证券（美国）有限公司
美国纽约公园大道 280 号 21 楼东（纽约 10017）
电话：+212-763-8160/传真：+917-725-9702
电子邮件: Huatai@htsc-us.com
http://www.htsc-us.com

华泰证券（新加坡）有限公司
滨海湾金融中心 1 号大厦, #08-02, 新加坡 018981
电话：+65 68603600
传真：+65 65091183