arXiv:2305.02823v2 [econ.GN] 5 May 2023

调查生成式人工智能的经济预期 *

J. Leland Bybee

耶鲁大学

初稿：2023年2月16日 本稿：2023年4月26日

初稿，欢迎提出意见

摘要

我介绍了一种经济预期调查：通过向大语言模型（LLM）询问其对各类金融和宏观经济变量的预期，这些预期基于1984年至2021年间《华尔街日报》新闻文章构成的样本。我发现，由此得出的预期与现有调查十分吻合，包括专业预测者调查（SPF）、美国个人投资者协会调查以及杜克大学首席财务官调查。重要的是，我记录到，基于LLM的预期与这些现有调查序列所呈现的许多偏离完全信息理性预期的现象相吻合。LLM的宏观经济预期表现出共识SPF预测中常见的反应不足。此外，其收益预期具有外推性，与预期收益的客观指标脱节，并且与未来实现收益呈负相关。最后，通过使用LLM训练期之外的文章样本，我发现其与现有调查指标的相关性依然存在，这表明这些结果并非源于记忆，而是体现了LLM的泛化能力。我的研究结果为LLM有望帮助我们更好地理解人类信念、探索非理性预期的可能模型提供了证据。

*leland.bybee@yale.edu。我从与 Nick Barberis、Paul Fontanier、Will Goetzmann、Paul Goldsmith-Pinkham、Bryan Kelly、Alp Simsek、Kaushik Vasudevan 的讨论和他们提出的意见中获益良多，也受益于耶鲁大学管理学院午餐研讨会参与者的意见。

预期至关重要。消费或储蓄多少、定什么价格，以及是否雇用或解雇员工，只是影响宏观经济动态的若干基本决策，而这些决策取决于主体对未来的预期。然而，这些预期是如何形成的，以及如何最好地对这一过程建模，仍是一个悬而未决的问题。

Coibion和Gorodnichenko (2015)

信念是资产定价的核心。资产价格具有前瞻性，几乎所有资产定价模型都意味着，投资者根据自己对某个随机贴现因子（*SDF*） M_{t+1}与收益 X_{t+1}的联合分布的信念为资产定价。资产定价领域之外的观察者或许会因此猜想，资产定价研究工作的很大一部分都致力于理解投资者如何形成信念。但至少到目前为止，事实并非如此。

Brunnermeier等人(2021年)

任何着手分析行为成因的人，（若没有独立的神经生理学证据）都会关注仅有的可用数据，即输入有机体的记录以及有机体当前的反应。

乔姆斯基 (1959)[1]

[1]这段引文摘自乔姆斯基对Skinner(1957)的评论，概述了乔姆斯基对Skinner方法的总结。尽管这篇评论本身持批评态度，但如此概括的行为主义方法仍与本研究对大语言模型的探讨相关。

1 引言

Muth提出的理性预期假说 (1961)仍是宏观经济学和金融学中占主导地位的信念模型。其主导地位不难理解：理性预期为经济行为建模提供了一种易于处理的方法（Lucas (1972)），同时限制计量经济学家的自由度（Lucas(1976）。然而，理性预期并非没有证据，也并非没有质疑其主导地位的替代理论。这些模型让我们置身于Sims（Sims (1980）所说的替代性预期“荒野”之中。[2] 成功探索这片荒野的尝试，主要依赖调查数据来作为信念的代理指标（Coibion和Gorodnichenko (2015)， Bordalo等人(2020年)， Angeletos等人 (2021年)，Nagel 和 Xu(2022a)， Lochstoer 和 Muir (2022 ）。

本文提出一种利用大语言模型（LLM）构建信念代理指标的新方法。大语言模型是一类旨在学习人类语言结构的统计模型。这些模型通过估算给定文档中此前所有已观察到的词元后，下一个词元 s_i 出现的概率来实现这一目标，

$$p(s_i|s_1, \ldots, s_{i-1}), \tag{1}$$

采用一种称为 Transformer 的特定神经网络架构（Vaswani 等人（2017））。虽然这一目标看似简单，但借助海量训练数据语料库和极大的参数空间，这些模型展现出一种涌现能力，能够模仿“类人”行为（Brown 等人（2020）、Wei 等人（2022）、Bubeck 等人（2023）。[3]

这种模仿类人文本互动的能力，使大语言模型有可能模拟人类行为和信念。丰富的因果信念可以蕴含在文本中——这一假设构成了整个学术出版流程的基础。作为一种旨在生成与训练语料库相似文本的统计方法，大语言模型应当反映其训练数据中蕴含的信念。[4] 我利用这一洞见，通过向 OpenAI 的 GPT-3.5 实例提供《华尔街日报》(《华尔街日报》（《华尔街日报》）的一组历史新闻文章，并要求其预测各种金融和宏观经济指标，来形成对大语言模型预期的估计。随后，我将这些文章层面的预期汇总为月度和季度时间序列，并将所得的信念代理变量与现有调查进行比较。

[2]Sims本人提到“均衡失调经济学的荒野”，但此后，这片荒野被重新表述为Sargent （2001）所说的“有界理性的荒野”，以及Angeletos等人（2021年）所说的“预期形成与均衡的替代模型之荒野”。[3]此处的“涌现”指的是这些模型的规模，相对于较小的语言模型而言。例如，OpenAI的GPT-3实例使用了约45太字节的文本作为训练数据，参数量约为1750亿。[4]大语言模型会反映其训练语料库中的偏见与信念，这一点是当前许多人工智能对齐研究的核心（Bender等人（2021年）、Schramowski等人（2022）

首先，我检验 GPT 对股票市场的预期与美国个人投资者协会（AAII）调查和杜克大学首席财务官调查中的预期吻合得如何。我发现，GPT 的预期与这些现有调查显著相关，相关程度与其他现有调查指标相当。随后，我比较 GPT 对标准普尔的预期与各种其他预期收益指标之间的相关性，以及现有调查中的相应相关性。我发现，GPT 表现出同样的外推预期，并且与股票基金资金流显著相关。此外，我发现，GPT 的预期与预期收益客观指标之间也存在同样的脱节——其与对数股息价格比、Lettau 和 Ludvigson （2001）的 CAY 指标以及若干预期收益预测代理变量的相关性符号，与现有调查指标相符。最后，我发现，GPT 的预期与未来实现收益负相关，这与现有调查指标一致，但不同于预期收益客观指标。

其次，我将 GPT 对一系列宏观经济变量的预期与Coibion和Gorodnichenko (2015)研究中的专业预测者调查（SPF）进行比较。我发现，除两个 SPF 序列外，GPT 的预期均与修订显著相关。随后，我评估了与 GPT 预期相关的 SPF 修订变动是否是 SPF 共识预测中普遍存在的反应不足的驱动因素。通过同时使用 SPF 修订和 GPT 的预期运行 Coibion-Gorodnichenko（CG）回归，我发现 GPT 能够匹配观察到的反应不足。

这些结果为探索各种预期的广阔天地提供了新的证据，因为 GPT 的预期与现有调查序列中表现出的许多偏离完全信息理性预期（FIRE）的情况相符。GPT 的预期偏离 FIRE 可能有两个来源，F_t^{gpt}，关于第 k 个序列，X_{t+h}^k，如下所示：

$$F_t^{gpt}(X_{t+h}^k) = g_k(\theta_t). \tag{2}$$

此处 g_k 对应于针对给定序列提示的大语言模型权重 k 以及 θ_t时期 t所提供的新闻文本 g_k 在整个样本中保持不变。其次请注意，提供给 GPT 的新闻文本彼此独立且缺乏上下文——GPT 在 t 时的预期仅基于截至 t时可获得的新闻文章集合。[5] 这限制了能够解释 GPT 偏离现象的模型范围。

如果 GPT 的预期偏差源自估算得到的大语言模型权重，g_k，那么解释这些偏差的信念模型就必须体现对新闻的静态反应函数。依赖于过度强调具有代表性或显著新闻的模型（Bordalo等人 (2018),Bordalo等人 (2022）或许能够解释这种偏差，前提是代表性并不取决于

[5]事实上，情况比这更严重：GPT 的预期是基于单篇新闻文章形成的，甚至缺乏时间点 t 其他新闻文章的背景。

近期语境，而是给定新闻文章的某些更普遍属性。大语言模型捕捉这些偏差的潜力，凸显了新兴的跨学科文献的重要性；这些文献通过实验研究此类模型的行为（Aher等人 (2022), Horton(2023),Argyle等人 (2023), Brand等人 (2023)).

或者，如果 GPT 的预期偏差源于新闻文章本身的属性，θ_t那么，解释这些偏差的模型就需要考虑这些属性。具体而言，若要用学习、外推以及近期状态的代表性来解释这些偏差（Fuster等人(2010）、Afrouzi等人(2020年）、Farmer等人(2021年），那么这些动态就必然嵌入新闻文本生成过程本身。这一潜在的偏差来源凸显了研究叙事的起源及其动态如何驱动信念的重要性（Shiller (2019）、Bybee等人(2021年）、Andre等人(2021年）、Michalopoulos和Xue(2021年）、Bybee等人(2022）、Flynn和Sastry(2022）。

更进一步，值得注意的是，GPT 能够匹配多个现有调查序列中的不同矩。鉴于 GPT 的预期基于同一算法和同一新闻来源，这表明这些调查中观察到的不同偏差源自同一来源，而试图解释这些现象的模型应当能够共同解释这些结果。

另一种可能影响我研究结果的解释是，训练语料库中可能存在数据泄漏。对我的论述而言，重要的是，我假设 GPT 学到的是其训练样本中所包含的信念因果结构的某种泛化，而不是在回应时复述记忆中的文本。为检验这一问题是否会影响研究结果，我进行了两项测试。首先，我使用 GPT 训练期之外（即 2021年9月之后）的《华尔街日报》文章样本，检验 GPT 的预期是否仍与现有调查指标相关；结果确实如此。其次，如果 GPT 依赖关于未来的信息，那么相较于预测修正，相应序列的实现值理应更能解释 GPT 的预期。但我发现事实并非如此。

一个尚未解决的问题是：GPT 的信念究竟属于谁？我的研究结果表明，GPT 的信念与多个不同群体的信念高度吻合，包括专业预测人员、个人投资者和首席财务官。GPT 的训练语料库基于广泛的来源和主题，因此它或许可以代表某种“代表性主体”。然而，未来这一领域很可能会发展出旨在更直接反映不同群体信念和利益的大语言模型。具体而言，可以通过对现有的开源大语言模型进行微调（例如Touvron 等人（2023）中的模型），或者像Wu 等人（2023）那样，基于不同语料库从头构建模型来实现。我的研究结果提供了概念验证，表明这些方法是进一步深入理解人类行为的一条有价值的路径。

图 1：提示格式

```
Here is a piece of news:

"%s"

Do you think this news will increase or decrease %s?

Write your answer as:

{increase/decrease/uncertain}:
{confidence (0-1)}:
{magnitude of increase/decrease (0-1)}:
{explanation (less than 25 words)}
```

注。 报告向 GPT 发出的查询所用的提示格式。“%s”表示提示中插入标题和目标文本的位置。

2 衡量生成式人工智能的预期

我的主要数据集由 《华尔街日报》 1984年到2021年间刊载的所有文章构成，数据购自道琼斯历史新闻档案。数据清理流程的完整摘要详见附录 A.1。在完成这项初步清理后，我从每个月的数据中随机抽取 300 篇文章。这样做是因为每次向 OpenAI 的 API 发出请求都需要支付一笔边际费用。

查询使用 OpenAI 的 GPT-3.5 模型实例。图 1 展示了用于查询 GPT 的提示格式。每次向 GPT 发出的查询都会提供一个标题和对所需序列的描述——例如，对于消费者价格指数，我会询问“美国消费者价格指数”的上升或下降。作为回应，GPT 会生成一段符合提示中所要求格式的文本。

表1 报告了调查序列以及回答的汇总统计。附录7 报告了B 中各调查序列可能出现的上升/下降组合之间的共现情况。[6] 表2 列出了若干示例新闻标题，以及针对标准普尔500指数、消费者价格指数和失业的相应回答。此外，我还要求 GPT 解释其回答，相关内容列于表2中。这些解释为分析 GPT 回答背后的推理提供了有用方法。例如，当给出一篇关于向

[6]与经典的监督式机器学习不同，GPT 并非针对某项特定任务（即预测经济指标）进行训练，而是针对通用语言建模任务进行训练。因此，在机器学习文献中，这类分类被称为零样本学习。

表 1：GPT 调查的汇总统计

序列	提示	日期范围	计数	增长率	下降率	不确定性百分比
SNP	标准普尔500指数	1984—2021年	136345	15.13	26.84	58.02
CPI	美国的消费者价格指数	1984—2021年	132736	7.86	6.45	85.69
HS	美国住房开始于the联合州	1984—2021年	132212	2.50	5.68	91.82
IP	工业生产in the美国	1984—2021年	132892	10.11	11.72	78.17
DEFL	国内生产总值其中的价格平减指数美国	1984—2021年	132760	9.63	12.50	77.88
AAA	美国的AAA级债券利率美国	1984—2021年	133467	11.09	14.88	74.03
C	美国的实际消费州	1984—2021年	131574	11.53	17.67	70.80
GF	联邦政府消费在美国	1984—2021年	132839	9.86	10.88	79.26
GY	实际国内生产总值的美国	1984—2021年	132148	20.54	20.91	58.56
NRI	实际非住宅投资美国	1984—2021年	132961	17.46	22.94	59.59
RI	美国实际住宅投资美国	1984—2021年	133157	8.66	16.49	74.85
GS	州和地方政府消-费（美国）	1984—2021年	131428	13.44	17.30	69.26
3TB	3个月期美国国库券利率	1984—2021年	134609	15.45	11.21	73.34
UE	美国的就业	1984—2021年	120102	9.81	11.24	78.95

注。 报告调查标签、传递给 GPT 的文本、日期范围、非缺失回答的数量、回答为“增加”的比例、回答为“减少”的比例，以及回答为“不确定”的比例。

美国企业开放日本电话市场的新闻报道时，GPT 预计这将为美国电话公司带来增长机会，从而推高标准普尔指数；通过在美国电信业创造新的就业机会来降低失业率；并通过加剧竞争来降低消费者价格指数。类似地，当给出一篇关于布什执政时期减税和收紧政府预算的报道时，GPT 预计减税将刺激美国经济，从而推高标准普尔指数；由于不确定哪种影响会占主导地位，预计不会对失业产生影响；并预计减税带来的需求增加将推高消费者价格指数。这些解释凸显了 GPT 为其回答提供清晰推理链条的能力，也为理解信念的因果结构开辟了新的可能途径。

表2：示例回答

序列	方向	标题/回应
		(1989年-06-29) 美国达成扩大使用权限的协议，涉及 Mo- 日本的移动电话业务
S&P	1	这一消息预示着美国移动 电话公司将迎来潜在的增长机会，这可能会对标准普尔500指数产生积极影响—— dex.
UE	-1	扩大进入日本手机市场的渠道可能会 为美国电信业创造新的就业机会—— 业。
CPI	-1	竞争加剧以及进入日本手机市场的渠道扩大 市场可能会导致美国消费者面临更低的价格。
		（2001-02-28）布什提出减税和紧缩预算以提供援助 “疲软”的经济
S&P	1	减税刺激经济，这可能导致增加 企业利润和更高的股价。
UE	0	减税可能会刺激投资，但紧缩预算可能会重新- 减少政府支出并减缓就业增长。
CPI	1	减税通常会刺激消费，从而增加需求 并推高价格，但财政紧缩可能会抵消这一影响。

注。 列出了示例文章标题样本及 GPT 对应的回答。第一列列出所查询的序列。第二列列出查询的方向。第三列列出标题及 GPT 对应的回答。

最后，由于 GPT 生成的是文章层面的二元预期，我将这些预期汇总，使其频率与调查数据具有可比性。为进行汇总，我计算一个“平衡统计量”：GPT 回答预期增加的文章比例，减去 GPT 回答预期减少的文章比例：

$$F_t^{gpt}(X_{t+h}^k) = \frac{\sum_{i\in A_t} \mathbb{I}(\text{Increase})_i^k - \mathbb{I}(\text{Decrease})_i^k}{\sum_{i\in A_t} \mathbb{I}(\text{Increase})_i^k + \sum_{i\in A_t} \mathbb{I}(\text{Decrease})_i^k}. \tag{3}$$

这一做法常用于多种调查指标，也用于其他调查，例如盖洛普调查、密歇根大学消费者调查和美国个人投资者协会调查。

3 收益预期

我首先将 GPT 的收益预期与 Greenwood 和 Shleifer (2014年）中使用的两个公开基准收益序列进行比较，以验证这些预期。第一个是美国个人投资者协会（AAII）投资者情绪调查。AAII 调查自 1987 年起每周对 AAII 成员进行调查，至今仍在开展；调查衡量参与者对未来六个月股票市场走势持看涨、看跌或中性看法的比例。按照 Greenwood 和 Shleifer(2014），在大部分分析中，我将每周的回答汇总为月度数据。第二项调查是杜克大学或 Graham 和 Harvey 对首席财务官的调查，由 John Graham 和 Campbell Harvey 于 1998 年发起。该调查征询首席财务官对各种宏观和企业特定指标的看法，包括他们对未来十二个月美国股票市场收益的预期。该调查自 2000 年起持续至今。

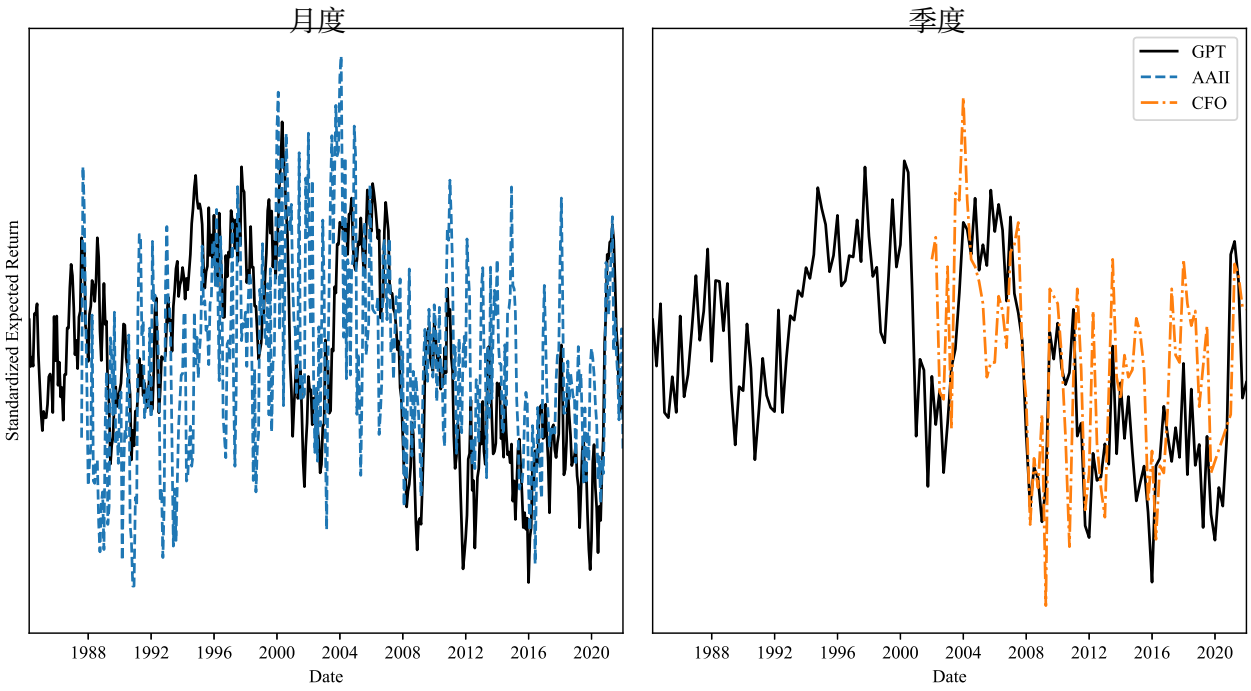
图 8 报告了 GPT 预期的时间序列，采用三个月的窗口进行汇总，并同时展示相应的调查序列。此外，还报告了每个调查序列与 GPT 预期的各种汇总指标之间的相关性。我考察了一个月、两个月和三个月窗口的汇总指标，以及对每日汇总指标进行指数加权平均的结果。对于指数加权平均，我报告了在最优平滑参数下的相关性 λ 该参数可使其与调查序列的相关性最大化。

我发现，无论采用哪种汇总方法，GPT 的预期都与现有调查指标显著相关。本节余下部分将重点关注三个月的汇总窗口。在此窗口下，GPT 预期与现有调查指标之间的平均相关性为 0.47。作为比较，Greenwood 和 Shleifer(2014) 发现，在他们评估的全部收益预期调查中，平均相关性为 0.43。

这些结果表明，GPT 的预期与现有调查指标的变动密切吻合。因此，我接下来将评估 GPT 的预期在多大程度上符合现有收益预期的关键特征。为此，我将比较 GPT 预期与文献中此前研究的替代指标之间的相关性，并将其与现有调查指标进行比较。

资产定价领域的大量文献聚焦于外推预期的重要性（Cutler 等人（1990）、Barsky 和 De Long（1993）、Lakonishok 等人（1994）、Barberis 等人（2015）、Jin 和 Sui（2022））。因此，我参照 Greenwood 和 Shleifer（2014）以及 Nagel 和 Xu（2022b），评估不同调查预期与美国股票市场过去十二个月收益率（R_{t-12}）之间的相关性。此外，近期有研究利用日益丰富的已实现交易行为数据，将其作为投资者信念的代理变量（Giglio 等人（2019）、Gabaix 和 Koijen（2021）、Alekseev 等人（2022））。

图2：收益预期时间序列



	一个月平均值	两个月平均值	三个月平均值	最优指数加权移动平均
首席财务官（样本量=76）	0.49	0.56	0.59	0.59
AAII（Q，样本量=138）	[4.84] 0.32	[5.88] 0.40	[6.21] 0.39	[6.24] 0.43
AAII（M，样本量=414）	[3.93] 0.32	[5.15] 0.35	[5.01] 0.35	[5.63] 0.36
	[6.83]	[7.65]	[7.65]	[7.82]

注。 分别展示 GPT 月度/季度标准化预期的时间序列，并叠加 AAII 和首席财务官调查结果。底部表格报告相关系数，并在方括号中列出对应的 t统计量。 t统计量采用滞后 12 个月的 Newey-West标准误。一个月/两个月/三个月平均值分别对应 GPT 预期在一个月/两个月/三个月窗口内的汇总值。最优指数加权移动平均对应采用最优调节参数计算的指数加权平均。

遵循这类文献，我评估不同调查预期与流入股票的共同基金资金流量之间的相关性。为计算这一指标，我将 Thomson Reuters Mutual Fund Holdings S12 数据库与 CRSP Survivor-Bias-Free US Mutual Funds 数据库相结合，并遵循 Alekseev 等人(2022年).

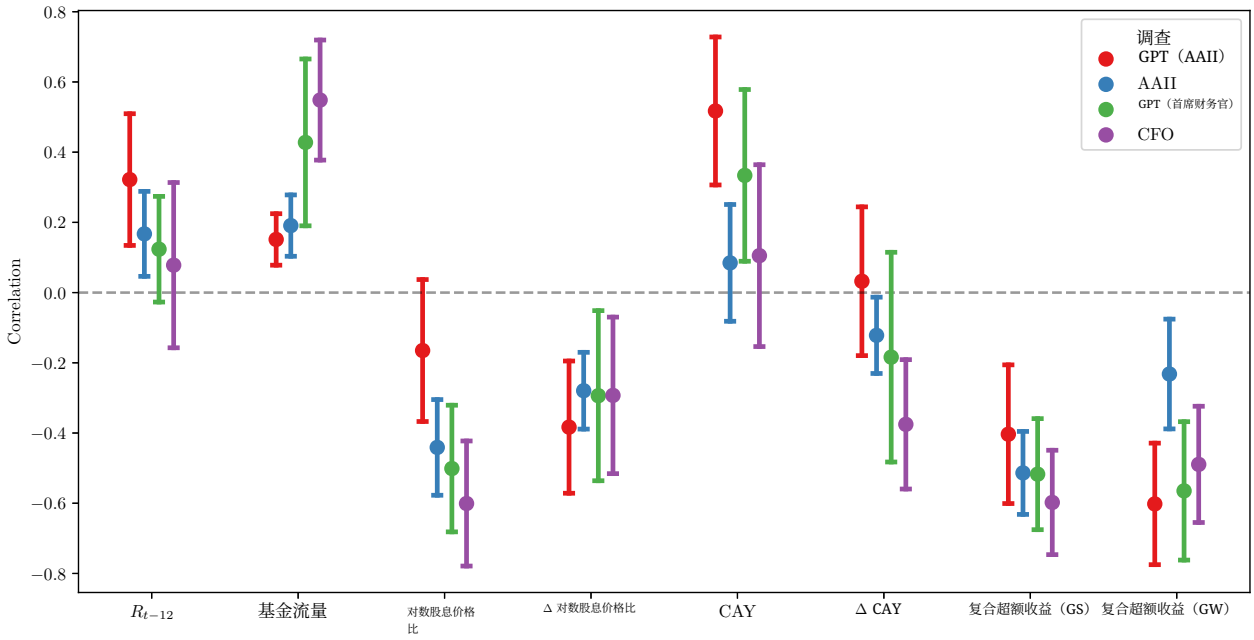
资产定价领域的大量文献也研究了预期收益的客观指标。重要的是，众所周知，这些客观指标与主观调查指标之间存在脱节（Greenwood 和 Shleifer (2014)，Nagel 和 Xu

(2022年b））。因此，接下来我会考察若干客观预期收益代理指标。首先，我按照大量实证文献的做法，考察对数股息价格比及其12个月变化。其次，我考察Lettau 和 Ludvigson(2001）提出的消费财富比（CAY）。最后，我考察两个预期收益指数，这些指数通过将未来12个月收益率对不同组的预测变量进行回归而构建。具体而言，我考察 Greenwood 和 Shleifer (2014）（沿用Fama 和 French (1989年））使用的同一预期收益指数，该指数采用对数股息价格比、国库券收益率、违约利差（BAA级债券收益率减去AAA级债券收益率）和期限利差（十年期政府债券收益率减去三个月期国库券收益率）。此外，我还使用Welch 和 Goyal(2008年）的“厨房水槽式”预测变量构建一个预期收益指数；这些变量是实证资产定价文献中的常见基准。

图 3 报告了每个调查序列与上述相应指标之间的相关性。对于每个调查序列，我都纳入了一个基于 GPT 的预期序列，该序列按相同频率构建，覆盖相同的时间样本。除 CAY 的变化和 AAII 的 GPT预期代理指标外，GPT 的相关性均与现有调查指标的符号相同。这些结果表明，GPT 不仅能匹配现有调查指标的变动，更重要的是还能匹配文献中指出的偏离理性预期的情况，这说明这些指标有望成为研究非理性预期的工具。

最后，根据理性预期模型，预期收益的指标应当能够预测未来收益率。然而，现有调查指标与未来收益率的相关性反而为负。因此，接下来我将 GPT 的预期与未来收益率之间的相关性，与若干其他主观和客观预期收益指标进行比较。图 4 报告了以上讨论的预期指标对未来收益率进行的一系列预测回归的相关系数。GPT 与未来收益率呈现出与现有调查指标相同的负相关性，这与客观指标呈现的正相关性形成对比。

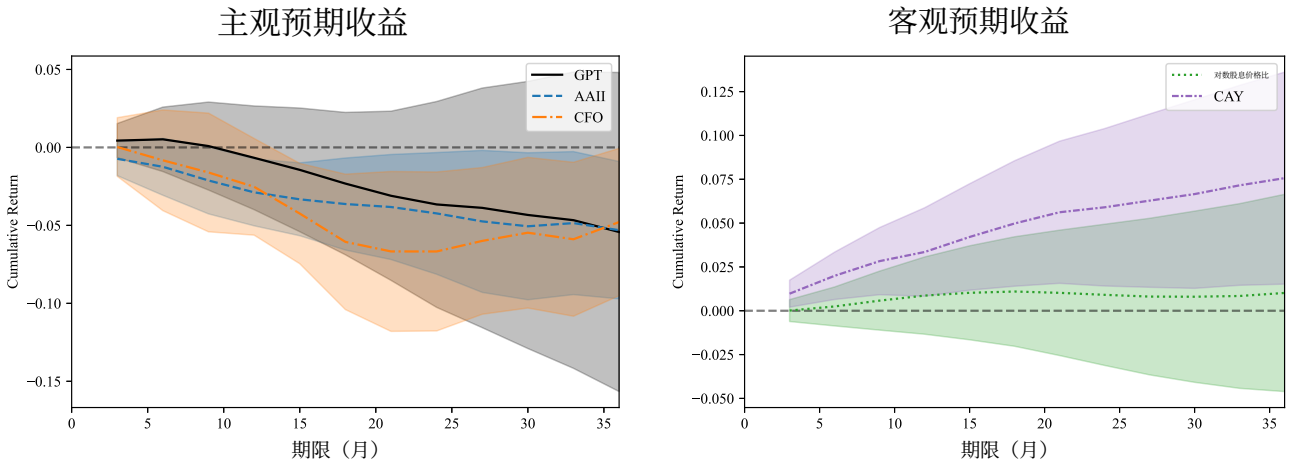
图3：调查结果与现有矩的相关性



	GPT（AAII）		AAII		GPT（首席财务官）		CFO	
	相关系数	*t*-统计量	相关系数	*t*-Stat	相关系数	*t*-统计值	相关系数	*t*-统计量
R_{t-12}	0.32	2.82	0.17	2.27	0.12	1.35	0.08	0.54
基金流量	0.15	3.39	0.19	3.58	0.43	2.95	0.55	5.26
对数股息价格比	-0.17	-1.34	-0.44	-5.31	-0.50	-4.56	-0.60	-5.53
Δ 对数股息价格比	-0.38	-3.34	-0.28	-4.19	-0.29	-1.99	-0.29	-2.15
CAY	0.52	4.03	0.08	0.83	0.33	2.24	0.11	0.67
Δ CAY	0.03	0.25	-0.12	-1.84	-0.18	-1.01	-0.38	-3.34
复合预期收益率（政府支出）	-0.40	-3.35	-0.51	-7.13	-0.52	-5.36	-0.60	-6.59
复合预期收益率（GW）	-0.60	-5.70	-0.23	-2.43	-0.56	-4.70	-0.49	-4.85

注。 报告了各预期序列与相应序列之间的相关性及90%置信区间，标准误采用具有12个月滞后的Newey-West标准误。下方表格报告了相应的相关系数和 *t*统计量。 *t*统计量采用具有12个月滞后的Newey-West标准误。GPT（AAII）表示在AAII 调查数据可用的日期，GPT的月度预期与之对应的相关性。GPT（CFO）表示在首席财务官调查数据可用的日期，GPT的季度预期与之对应的相关性。 R_{t-12} 表示滞后12个月的收益率。基金流量表示流入股票的共同基金总流量。Log(D/P)为对数股息价格比，CAY为Lettau 和 Ludvigson (2001 年）。 Δ Log(D/P) 和 ΔCAY 分别表示相应的 12 个月变化。Comp. ER （GS）表示以对数股息价格比、国库券收益率、违约利差和期限利差为自变量，对未来一年收益率进行回归所得的拟合值——即Greenwood 和 Shleifer(2014年）所使用的综合预期收益率指标。Comp. ER（GW）表示以Welch 和 Goyal(2008年）中的“厨房水槽式”预测变量为自变量，对未来一年收益率进行回归所得的拟合值。

图4：收益预测回归



注。 报告了一系列预测回归的系数：以主观和客观预期收益代理指标预测给定期限内未来的累计收益。阴影区域表示使用Newey-West标准误计算的90%置信区间，滞后阶数对应于相应期限。

4 宏观经济预期

接下来，我通过将GPT的预期与一组标准的宏观经济预期——专业预测者调查（SPF）进行比较，来验证GPT的预期。SPF是由费城联邦储备银行开展的季度宏观经济预测调查，也是研究宏观经济预期的黄金标准（Coibion和Gorodnichenko (2012年),Coibion和Gorodnichenko (2015), Bordalo等人(2020年), Angeletos等人 (2021年), Farmer等人 (2021年)）。作为专业预测人员，受访者是经济中信息最充分的一些主体，因此他们预测中的误差值得关注。

在我的样本期内，SPF涵盖了广泛的宏观经济变量，包括消费者价格指数、实际国内生产总值、失业率和联邦基金利率。我将GPT的预期与Coibion和Gorodnichenko (2015)考察的13个序列的SPF预期进行比较。对于GPT的预期，我采用调查期前一季度内文章层面预期的季度汇总值。SPF会发布多个期限的预测；然而，由于GPT预期的期限未知，我将GPT预期与1至4季度期限预测的平均值进行比较。我的结果对这一选择具有稳健性，附录D中展示了1至4季度各期限预测对应的结果。

我首先评估 GPT 的预期与 SPF预测的相关程度。对于水平值 $F_t(X^k_{t+h})$ 和修订 $F_t(X^k_{t+h}) - F_{t-1}(X^k_{t+h})$，我都报告了结果。修订反映了预测中包含的新信息；鉴于 GPT 的预期是根据近期新闻的细粒度内容形成的，因此它们可能与修订而非水平值的相关性更强。图5 报告了这些相关性——附录中提供了 GPT 预期与相应 SPF 修订的完整时间序列图集。

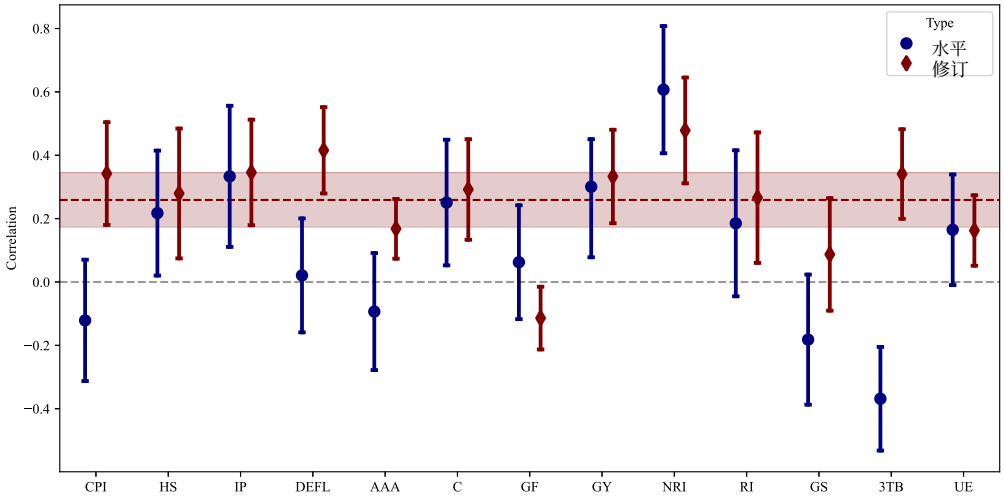
相应 SPF 修订的图集见附录C。在所有抽样序列中，除联邦政府消费以及州和地方政府消费这两项外，GPT 的预期与相应修订之间的相关性均达到 90% 的置信水平，且具有统计显著性。

大量研究偏离完全信息理性预期（FIRE）的文献，都聚焦于预测误差的可预测性。具体而言，研究者特别关注信念对新信息的迟滞反应，并使用 Coibion和Gorodnichenko (2015)（CG）回归进行衡量。CG 回归以预测误差为因变量，以相应的预期修订为自变量：正系数表示反应不足，负系数则表示反应过度。

鉴于 GPT 与 SPF 修订之间存在显著相关性，我接下来评估：宏观经济预期中广为人知的反应不足，是否可以用与 GPT 相关的修订成分来解释。为此，对于 GPT 与 SPF 修订之间呈现显著相关性的序列（即不包括政府消费），我将 SPF 预测误差对 GPT 预期的相应版本进行回归。图6 报告了这些回归的相关性，并将其与使用原始 SPF 修订所得的相关性进行比较。除两个序列外，GPT 预期与 SPF 预测误差之间的相关性均为正，合并后的总体结果为正，且在 90% 的置信水平上显著。[7]

[7]失业在原始 SPF 修订中表现出显著的过度反应，这与相关文献的结论相悖。我的样本包括 2020年 新冠肺炎疫情，这可能会影响这一结果。为检验这一点，我将样本期限定在 2019 年之前，并重新估计了相同的程序。相关结果报告于附录 E中。GPT 的预期与 SPF 预测误差之间显著的正相关性依然存在，且原始 CG 结果与相关文献的结论一致。

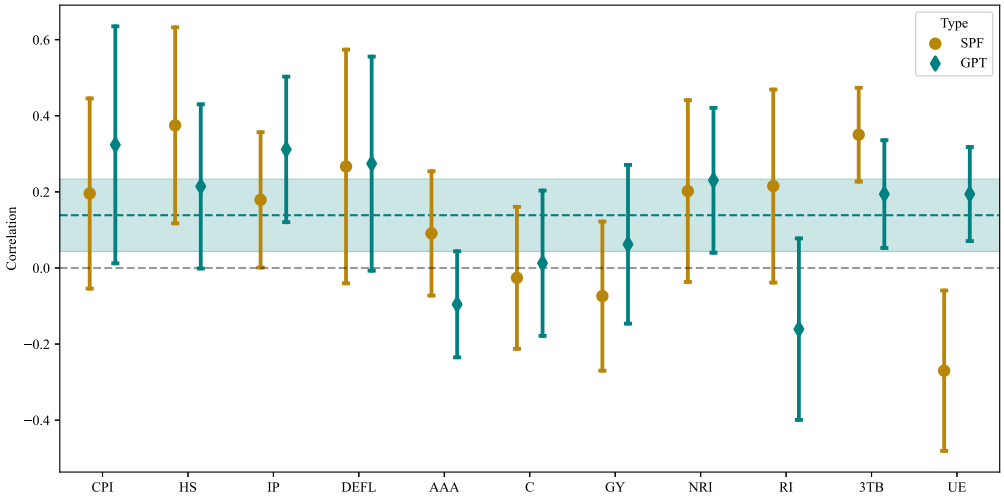
图 5：GPT/SPF相关性



	水平		修订	
	相关系数	t 统计量	相关系数	t 统计量
面板	0.07	1.49	0.26	4.93
CPI	-0.12	-1.04	0.34	3.46
新屋开工	0.22	1.81	0.28	2.24
工业生产	0.33	2.46	0.35	3.40
国内生产总值平减指数	0.02	0.19	0.42	5.01
AAA级公司债券收益率	-0.09	-0.83	0.17	2.92
消费增长	0.25	2.07	0.29	3.02
联邦政府支出	0.06	0.57	-0.11	-1.89
GDP增长	0.30	2.22	0.33	3.70
非住宅投资	0.61	4.96	0.48	4.70
住宅投资	0.19	1.32	0.27	2.12
州政府支出	-0.18	-1.45	0.09	0.80
3个月期美国国库券	-0.37	-3.70	0.34	3.96
失业率	0.16	1.55	0.16	2.40

注。 报告SPF预期和GPT 预期的水平值与修订之间的相关性。此外，还报告相关系数的90%置信区间。深红色虚线报告面板相关系数，深红色阴影带表示相应的90%置信区间。单变量回归的标准误采用Newey-West方法，面板标准误采用Driscoll-Kraay方法。底部表格报告相应的相关系数和 t-统计量。

图 6：Coibion-Gorodnichenko 回归



	SPF		GPT	
	相关系数	t 统计量	相关系数	t 统计量
面板	0.12	1.31	0.14	2.39
CPI	0.20	1.28	0.32	1.70
新屋开工	0.37	2.39	0.21	1.63
工业生产	0.18	1.65	0.31	2.67
国内生产总值平减指数	0.27	1.42	0.27	1.60
AAA级公司债券收益率	0.09	0.91	-0.10	-1.12
消费增长	-0.03	-0.23	0.01	0.11
GDP增长	-0.07	-0.62	0.06	0.49
非住宅投资	0.20	1.39	0.23	1.99
住宅投资	0.22	1.39	-0.16	-1.11
3个月期美国国库券	0.35	4.67	0.19	2.25
失业率	-0.27	-2.10	0.19	2.58

注。 报告 SPF 修订和 GPT 预期的 CG 系数。此外，还报告系数的 90% 置信区间。青绿色虚线表示面板CG 系数，青绿色阴影带表示相应的 90% 置信区间。单变量回归的标准误采用 Newey-West 方法，面板标准误采用 Driscoll-Kraay 方法。底部表格报告相应的相关系数和t-统计量。

5 记忆还是泛化?

鉴于 GPT 的训练期与我的样本期有所重叠，接下来我将评估 GPT 的预期是否源于数据泄漏或对训练样本的记忆。对记忆的担忧不同于对过拟合的常见担忧。大语言模型的训练目标是预测序列中的下一个词元，因此，若这里存在直接过拟合，就意味着训练样本中包含我在此指定格式的提示，这似乎不太可能。然而，GPT 的训练语料库可能包含有关我所关注变量未来取值的信息和讨论，而它会在收到提示时复述这些内容。为了解决这一担忧，我进行了两项检验。首先，我扩展了《华尔街日报》 文章样本，纳入 GPT 训练期之后的文章，即 2021年9月 之后的文章。其次，我检验 GPT 的预期究竟是由 SPF预测 的修订所驱动，还是由变量的实现值所驱动。

为评估 GPT 的样本外预期，我抓取了《华尔街日报》档案中从 2021年9月 到 《华尔街日报》2023年3月所有可获取的文章。随后，我按照附录A.2 中详述的清理流程进行处理，并像对主样本那样，询问 GPT 对每篇文章的预期。AAII回报预期每周发布，因此，这使我获得了 GPT 训练期之外的 79 个观测值。对于 SPF，只有五个季度的数据可用，因此我针对上述研究的全部序列，对 SPF预期修订进行面板回归，并以 GPT 的预期作为解释变量。表 3 报告了这些检验的结果，表明我在主样本中观察到的显著相关性并非记忆所致。

表3：与GPT样本外的相关性

	AAII（每周）	SPF平均值	SPF 1	SPF2	SPF 3	SPF 4
相关性	0.46	0.61	0.49	0.64	0.57	0.53
t统计量	[4.54]	[6.11]	[4.50]	[6.61]	[5.45]	[4.93]
N	79	65	65	65	65	65

注： 报告 GPT 的样本外预期与各项基准之间的相关性。每个 SPF 列对应不同的期限，标准误采用 Driscoll-Kraay 方法计算。

接下来，我评估 GPT 的预期究竟是由 SPF预测的修订所驱动，还是由变量的实现值所驱动。为此，我针对上述研究的全部序列，将 SPF预期的修订对 GPT 的预期进行面板回归。表 4报告了这些检验的结果，表明我在主样本中观察到的显著相关性，并非源于 GPT 掌握了超出 SPF预期范围的未来实现值信息。

表4：GPT 与 SPF 修订和实现值之间的相关性

期限	Avg.	1	2	3	4
SPF 修订	0.25	0.20	0.17	0.16	0.14
	[11.65]	[8.92]	[7.92]	[7.02]	[6.19]
实现值	0.02	0.01	0.02	0.02	0.01
	[1.80]	[0.47]	[1.44]	[1.36]	[0.97]
R^2	6.71	4.07	3.27	2.53	1.95

注。 报告 GPT 的预期与 SPF 修订和实现值之间的相关性。标准误采用 Driscoll-Kraay 方法。

6 结论

本文介绍了一种利用大语言模型构建预期代理指标并研究信念的新方法。我发现，由此得到的预期代理指标与现有调查对应指标一样，表现出许多偏离完全信息理性预期的特征。这些结果为指导预期形成模型的发展提供了新的证据。

这些事实根据偏差的来源，对信念模型提出了新的约束。如果这些偏差源自所估计的大语言模型本身，就需要进一步研究这些方法的静态行为。另一方面，如果偏差源自《华尔街日报》 文章本身，就需要进一步研究这些文章的叙事动态。总的来说，关于作为人类模拟体的大语言模型，我们仍有许多需要了解之处，但此处的结果表明，这些方法有望开启一个新的研究方向。

参考文献

Afrouzi、Hassan、Spencer Yongwook Kwon、Augustin Landier、Yueran Ma 和 David Thes-mar，2020年，《预期过度反应：证据与理论》，可在 *SSRN 3709548* 获取。

Aher、Gati、Rosa I Arriaga 和 Adam Tauman Kalai，2022，使用大语言模型模拟多个人类，*arXiv* 预印本 *arXiv:2208.10264* 。

Alekseev、Georgij、Stefano Giglio、Quinn Maingi、Julia Selgrad 和 Johannes Stroebel，2022，《基于数量的方法：构建气候风险对冲投资组合》，美国国家经济研究局 。

Andre、Peter、Ingar Haaland、Christopher Roth 和 Johannes Wohlfart，2021，关于宏观经济的叙事。

Angeletos、George-Marios、Zhen Huo 和 Karthik A. Sastry，2021，《不完善的宏观经济预期：证据与理论》，《美国国家经济研究局宏观经济学年刊》 。

Argyle、Lisa P.、Ethan C. Busby、Nancy Fulda、Joshua R. Gubler、Christopher Rytting 和 David Wingate，2023，《由一而众：使用语言模型模拟人类样本》，《政治分析》 1–15。

Barberis, Nicholas、Robin Greenwood、Lawrence Jin 和 Andrei Shleifer，2015，X-CAPM：一种外推型资本资产定价模型，《金融经济学杂志》115，1–24。

Barsky, RobertB 和 J Bradford DeLong，1993，股票市场为何波动？，《经济学季刊》108，291–311。

Bender, Emily M.、Timnit Gebru、Angelina McMillan-Major 和 Shmargaret Shmitchell，2021，论随机鹦鹉的危险：语言模型会不会过于庞大？xn–cc6c，载于 *FAccT '21*：*2021*年*ACM*公平、问责与透明度会议论文集，610–623（美国计算机协会，纽约，纽约州，美国）。

Bordalo, Pedro、Nicola Gennaioli、Yueran Ma 和 Andrei Shleifer，2020，宏观经济预期中的过度反应，《美国经济评论》110，2748–82。

Bordalo、Pedro、Nicola Gennaioli 和 Andrei Shleifer，2018，诊断性预期与信贷周期，《金融学杂志》73，199–227。

Bordalo、Pedro、Nicola Gennaioli 和 Andrei Shleifer，2022，显著性，《年度经济学评论》14，521–544。

Brand、James、Ayelet Israeli 和 Donald Ngwe，2023，Usinggptformarketresearch，可在 *SSRN 4395751* 查阅。

Brown, Tom B.、BenjaminMann、Nick Ryder、Melanie Subbiah、Jared Kaplan、Prafulla Dhari- wal、Arvind Neelakantan、Pranav Shyam、Girish Sastry、Amanda Askell、Sandhini Agar- wal、Ariel Herbert-Voss、Gretchen Krueger、Tom Henighan、Rewon Child、Aditya Ramesh、Daniel M. Ziegler、Jeffrey Wu、Clemens Winter、Christopher Hesse、Mark Chen、Eric Sigler、Mateusz Litwin、Scott Gray、Benjamin Chess、Jack Clark、Christopher Berner、Sam Mc- Candlish、Alec Radford、Ilya Sutskever 和 Dario Amodei，2020年，《少样本学习者》$arXiv$ 。

Brunnermeier、Markus、Emmanuel Farhi、Ralph S. J. Koijen、Arvind Krishnamurthy、Syd- ney C. Ludvigson、Hanno Lustig、Stefan Nagel 和 Monika Piazzesi，2021年，评论文章：资产定价的未来展望，金融研究评论 34，2126–2160。

Bubeck, Sébastien, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro 和 Yi Zhang，2023，《通用人工智能的火花：GPT-4 的早期实验》，$arXiv$ 。

Bybee, Leland, Bryan T Kelly, Asaf Manela 和 Dacheng Xiu，2021，商业新闻与商业周期。

Bybee, Leland, Bryan T Kelly 和 Yinan Su，2022，叙事资产定价：从新闻文本中提取可解释的系统性风险因子，约翰斯·霍普金斯大学凯瑞商学院研究论文 。

乔姆斯基, Noam，1959，《对 Skinner《言语行为》的评论》，《语言》35，26–58。

Coibion, Olivier 和 Yuriy Gorodnichenko，2012，调查预测能告诉我们有关信息僵化的什么？《政治经济学杂志》。

Coibion, Olivier 和 Yuriy Gorodnichenko，2015，《信息僵化与预期形成过程：一个简单框架与新事实》，《美国经济评论》105，2644–78。

Cutler, David M、James M Poterba、LawrenceH Summers 等，1990，《投机动态与反馈交易者的作用》，《美国经济评论》80，63–68。

Fama, Eugene F., 和 Kenneth R. French，1989年，股票和债券的商业状况与预期收益，《金融经济学杂志》25，23–49。

Farmer、Leland、Emi Nakamura 和 Jón Steinsson，2021，《关于长期的学习》，技术报告，美国国家经济研究局。Flynn、Joel P 和 Karthik Sastry，2022，《叙事的宏观经济学》，可通过 *SSRN 4140751* 获取 。Fuster、Andreas、David Laibson 和 Brock Mendel，2010，《自然预期与宏观经济波动》，《经济展望杂志》24，67–84。Gabaix、Xavier 和 Ralph S. J. Koijen，2021，《探寻金融波动的起源：市场缺乏弹性假说》，美国国家经济研究局 。Giglio、Stefano、Matteo Maggiori、Johannes Stroebel 和 Stephen Utkus，2019，《关于信念与投资组合的五个事实》，美国国家经济研究局 。Greenwood、Robin 和 Andrei Shleifer，2014，《收益预期与预期收益》，《金融研究评论》27。Horton、John J.，2023，《作为模拟经济主体的大语言模型：我们能从 Homo Silicus 中学到什么？》，美国国家经济研究局 。Jin、Lawrence J 和 Pengfei S ui，2022，《收益外推与资产定价》，《金融经济学杂志》145，273–295。Lakonishok、Josef、Andrei Shleifer 和 Robert W Vishny，1994，《逆向投资、外推与风险》，《金融学杂志》49，1541–1578。Lettau、Martin 和 Sydney Ludvigson，2001，《消费、总财富与预期股票收益》，《金融学杂志》56，815–849。Lochstoer、Lars A. 和 Tyler Muir，2022，《波动率预期与收益》，《金融学杂志》77，1055–1096。Lucas、Robert E.，1972，《预期与货币中性》，《经济理论杂志》4，103–124。Lucas、Robert E，1976，《计量经济学政策评估：一种批判》，载于卡内基—罗切斯特公共政策会议系列，第1卷，19–46，北荷兰出版社。Michalopoulos、Stelios 和 Melanie Meng Xue，2021，《民间传说》，《经济学季刊》136，1993–2046。Muth、John F，1961，《理性预期与价格变动理论》，《计量经济学：计量经济学会杂志》315–335。

Nagel, Stefan 和 Zhengyang Xu，2022a，《具有衰退记忆的资产定价》，金融研究评论 35，2190–2245。

Nagel, Stefan 和 Zhengyang Xu，2022b，《主观风险溢价的动态》，美国国家经济研究局 。

Sargent, Thomas J，2001，征服美国通货膨胀 （普林斯顿大学出版社）。

Schramowski, Patrick、Cigdem Turan、NicoAndersen、Constantin A. Rothkopf 和 Kristian Kersting，2022，大型预训练语言模型包含关于何为正确与错误行为的类人偏见，《自然·机器智能》 4，258–268。

Shiller, Robert J.，2019，叙事经济学 （普林斯顿大学出版社，普林斯顿，新泽西州，美国）。

Sims, Christopher A，1980，宏观经济学与现实，《计量经济学》：计量经济学会期刊 1–48。

Skinner, BF, 1957，《言语行为》（科普利出版集团）。

Touvron, Hugo, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-AnneLachaux, Tim- othée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Ro- driguez, Armand Joulin, Edouard Grave, 和 Guillaume Lample，2023，《LLaMA：开放且高效的基础语言模型》，$arXiv$ 。

Vaswani, Ashish, NoamShazeer, Niki Parmar, JakobUszkoreit, LlionJones, AidanN. Gomez, Lukasz Kaiser, 和 Illia Polosukhin，2017，《注意力就是一切》，$arXiv$ 。

Wei, Jason, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, DennyZhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, 和 William Fedus，2022，《大型语言模型的涌现能力》，$arXiv$ 。

Welch、Ivo 和 Amit Goyal，2008年，《股票溢价预测的实证表现综合考察》，金融研究评论 21，1455–1508。

Wu、Shijie、Ozan Irsoy、Steven Lu、Vadim Dabravolski、Mark Dredze、Sebastian Gehrmann、Prabhanjan Kambadur、David Rosenberg 和 Gideon Mann，2023，《BloombergGPT：面向金融的大型语言模型》，$arXiv$ 。

A 构建《华尔街日报》样本

A.1 构建《华尔街日报》 1984—2021年样本

我按以下顺序执行数据处理步骤:

1. 删除1984年1月之前和2021年12月之后的所有文章。2. 排除页码引文标签对应于 A、B、消费以外的任何版块或缺失页码引文标签的文章。3. 排除周末刊登的文章。4. 排除主题标签明显属于非经济内容（如体育）的文章。排除清单可向作者索取。5. 排除标题符合特定模式的文章（例如与数据表相关的文章，或属于常设体育、休闲或书籍专栏的文章）。排除清单可向作者索取。6. 排除字数少于100的文章。7. 排除标题少于10个词的文章。8. 每个月份的数据抽取300篇文章作为样本。

A.2 构建《华尔街日报》2021年10月至2023年3月样本量 e

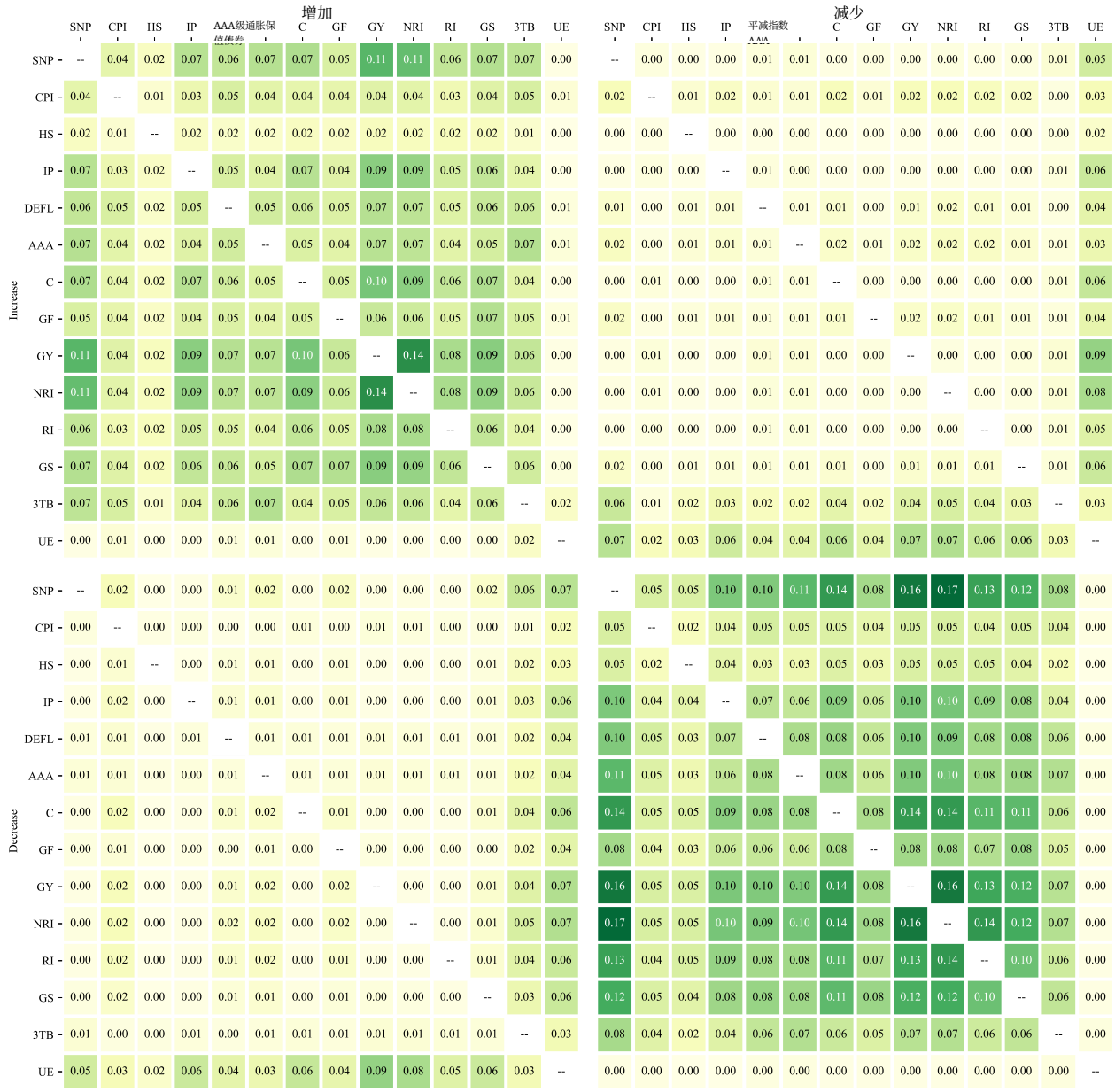
我按以下顺序执行以下数据处理步骤:

1. 排除标题少于 8 个单词的文章。2. 排除不属于以下版块的文章：“美国”“商业”“世界”“政-治”“《华尔街日报》独家新闻”“街头观察”“金融”“财报”“市场”“美国市场”“美国经济”“经济”。

B GPT 回复的共现

本节报告所研究的 GPT 预期序列中，文章层面的上升/下降回答之间的共现情况。

图7：共现矩阵

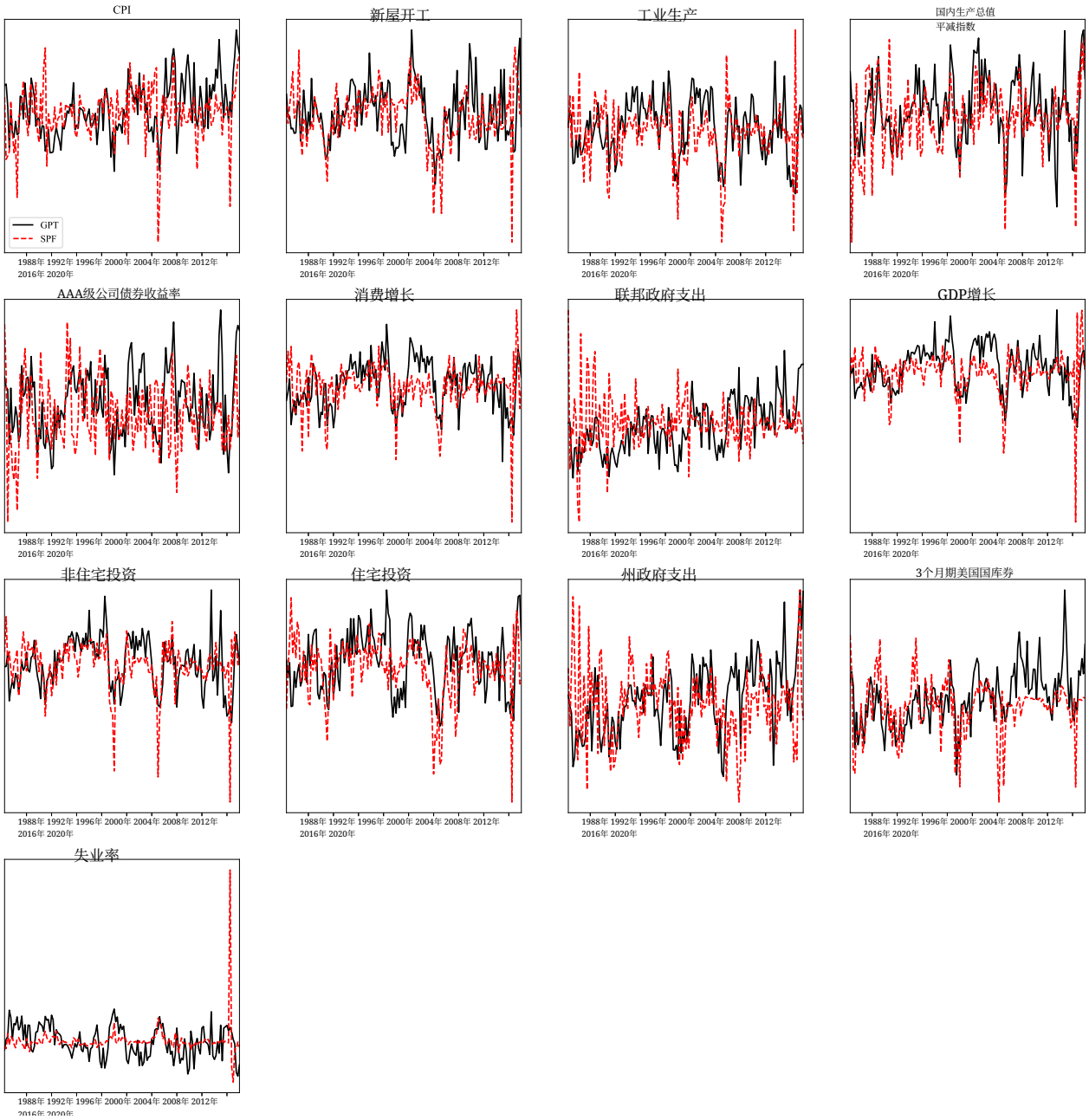


*No*等。 报告不同调查序列中各项指标的共现比例。

消费 GPT/SPF预期的时间序列

本节以时间序列图展示第4节所研究的 SPF 序列的 GPT 预期和平均修订。

图 8：GPT 预期与 SPF 修订的时间序列

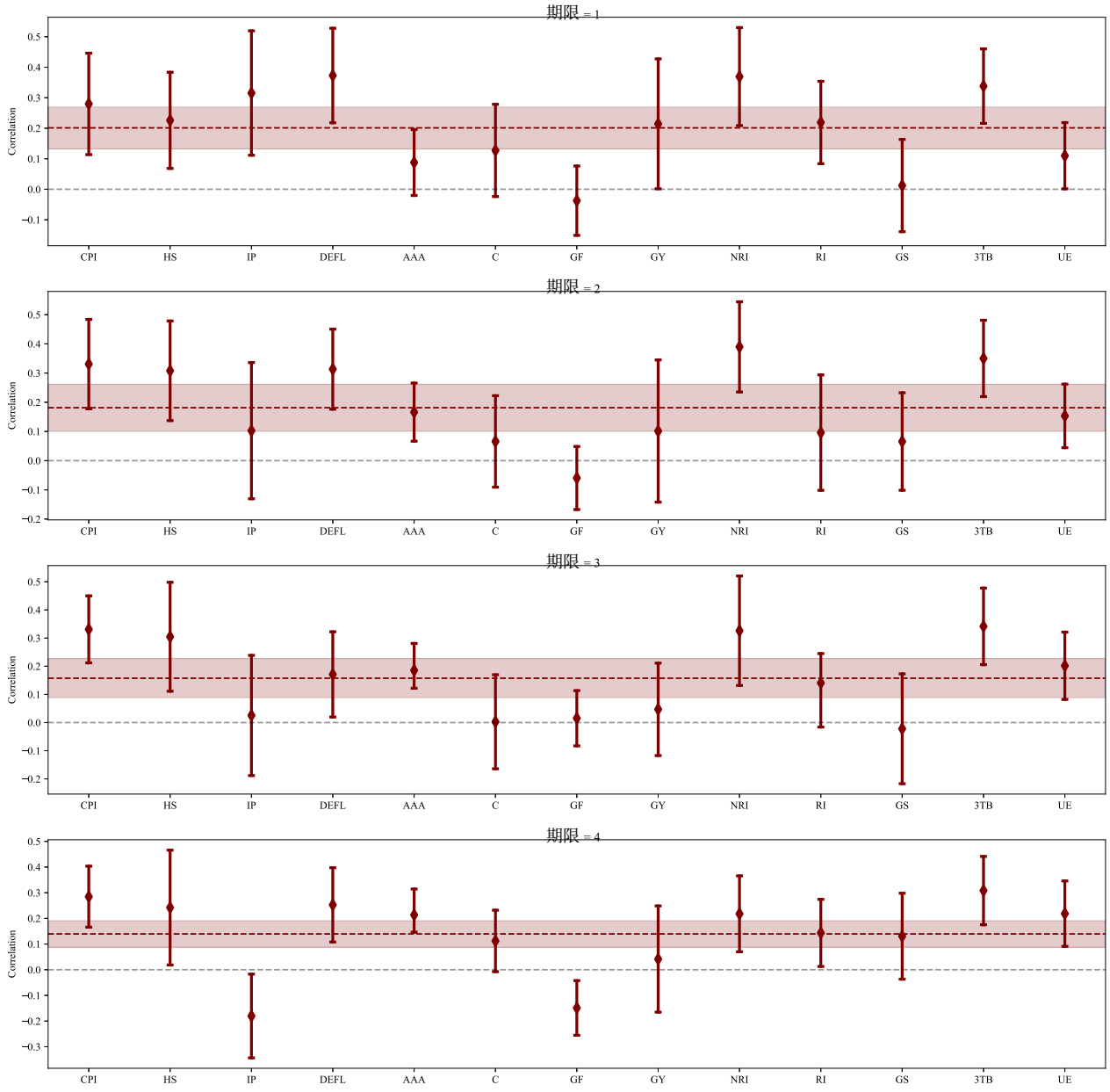


注释。 报告了 GPT 预期（黑色）和相应 SPF 修订（红色）的时间序列。

D　G不同期限下 PT/SPF 的相关性　S

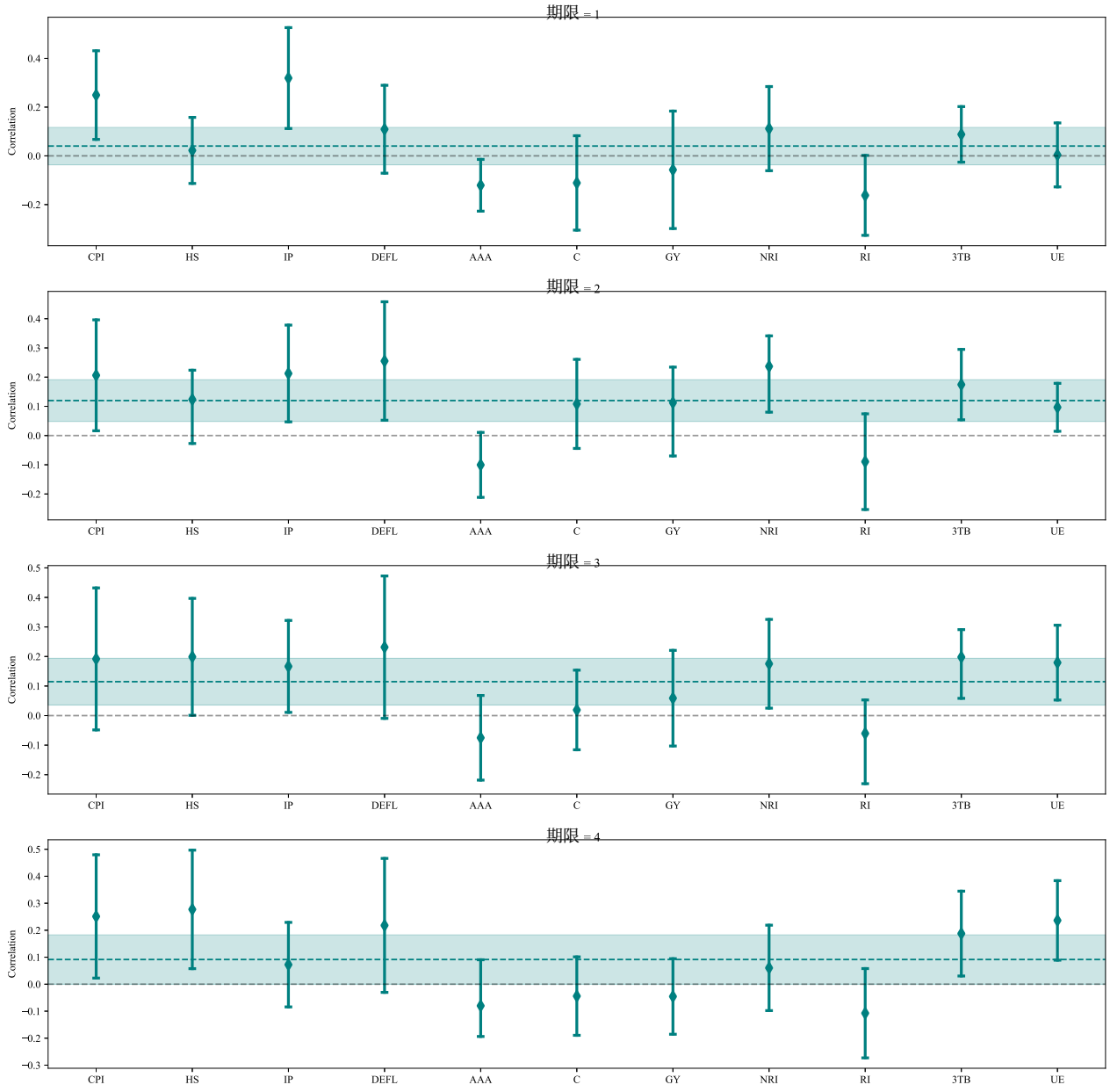
本节报告了与第 4 节相似的结果，但按照不同的期限进行了划分。

图 9：不同期限下的 SPF 预期修订相关性



注。 报告了不同期限下 SPF预期与 GPT 预期修订之间的相关性，并报告了相关系数的 90% 置信区间。酒红色虚线表示面板相关系数，酒红色阴影带表示相应的 90% 置信区间。单变量回归的标准误采用 Newey-West 方法，面板标准误采用 Driscoll-Kraay 方法。

图 10：不同期限的 Coibion-Gorodnichenko 系数

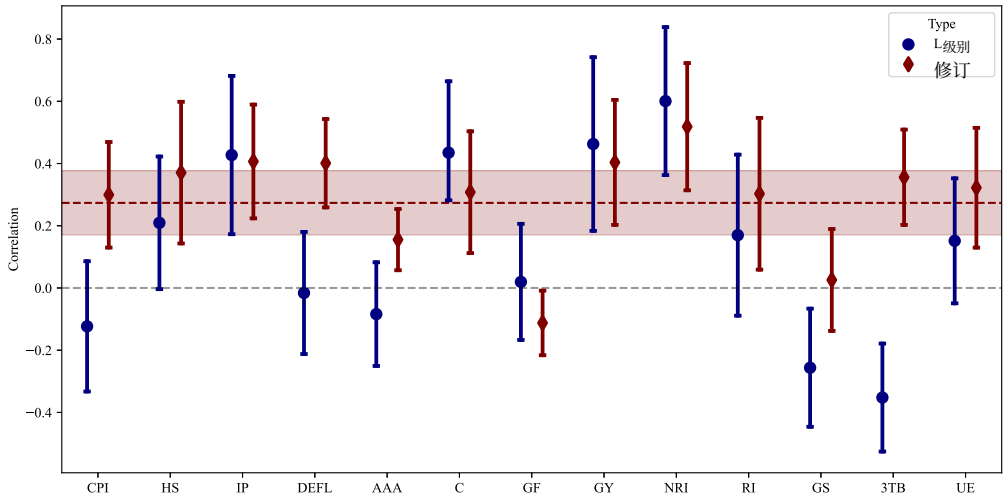


注。报告不同期限下SPF 修订和 GPT 预期的 CG系数。此外还报告系数的 90% 置信区间。青绿色虚线表示面板CG系数，青绿色阴影带表示相应的 90% 置信区间。单变量回归的标准误采用 Newey-West 方法，面板标准误采用 Driscoll-Kraay 方法。

E 新冠肺炎

本节报告了与4 相似的结果，但排除了 2019 年之后的所有数据，以避免新冠肺炎疫情的影响。

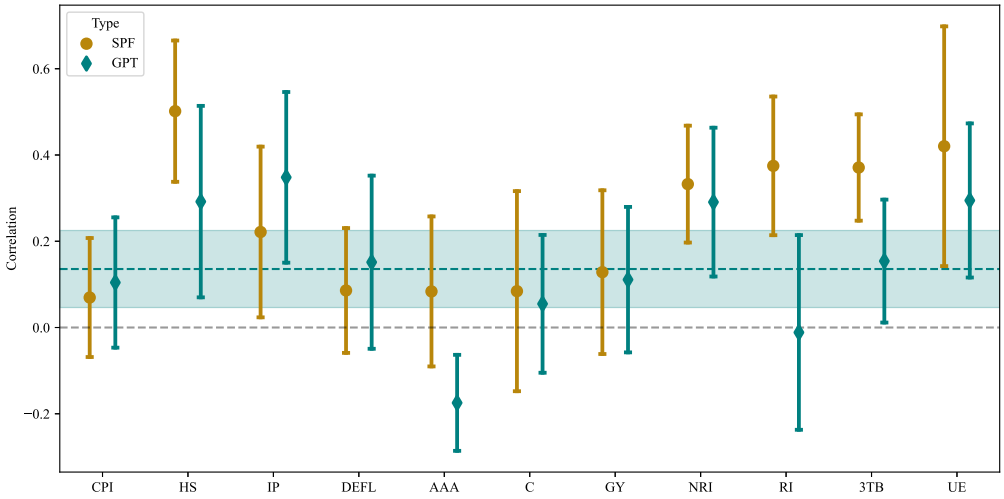
图11：GPT/SPF相关性（2019年以前）



	级别		修订	
	相关系数	t 统计量	相关系数	t 统计量
面板	0.07	1.23	0.27	4.35
CPI	-0.12	-0.97	0.30	2.90
新屋开工	0.21	1.61	0.37	2.67
工业生产	0.43	2.75	0.41	3.65
国内生产总值平减指数	-0.02	-0.13	0.40	4.63
AAA级公司债券收益率	-0.08	-0.83	0.16	2.59
消费增长	0.43	3.11	0.31	2.58
联邦政府支出	0.02	0.17	-0.11	-1.78
GDP增长	0.46	2.72	0.40	3.30
非住宅投资	0.60	4.14	0.52	4.16
住宅投资	0.17	1.08	0.30	2.04
州政府支出	-0.26	-2.22	0.03	0.26
3个月期美国国库券	-0.35	-3.33	0.36	3.81
失业率	0.15	1.24	0.32	2.75

注。 报告 SPF预期与 GPT 预期的水平值和修订值之间的相关性。此外，还报告相关系数的 90% 置信区间。酒红色虚线表示面板相关系数，酒红色阴影带表示相应的 90% 置信区间。单变量回归的标准误采用 Newey-West 方法，面板标准误采用 Driscoll-Kraay 方法。底部表格报告相应的相关系数和 t-统计量。

图 12：Coibion-Gorodnichenko 回归（2019年以前）



	SPF		GPT	
	相关系数	t 统计量	相关系数	t 统计量
面板	0.21	4.19	0.14	2.50
CPI	0.07	0.83	0.10	1.13
新屋开工	0.50	5.02	0.29	2.16
工业生产	0.22	1.84	0.35	2.89
国内生产总值平减指数	0.09	0.97	0.15	1.24
AAA级公司债券收益率	0.08	0.79	-0.17	-2.57
消费增长	0.08	0.60	0.05	0.56
GDP增长	0.13	1.11	0.11	1.08
非住宅投资	0.33	4.02	0.29	2.76
住宅投资	0.37	3.82	-0.01	-0.08
3个月期美国国库券	0.37	4.94	0.15	1.77
失业率	0.42	2.48	0.29	2.70

注。 报告 SPF 修订和 GPT 预期的 CG 系数。此外，还报告系数的 90% 置信区间。青绿色虚线表示面板CG系数，青绿色阴影带表示相应的 90% 置信区间。单变量回归的标准误采用 Newey-West 方法，面板标准误采用 Driscoll-Kraay 方法。底部表格报告相应的相关系数和*t*-统计量。