CCFinance：从网络档案重构财经新闻样本

Marcus Gawronsky

arXiv:2611.30737v1 [q-fin.CP] 23 Nov 2026

摘要

专有新闻档案会将报道归入各家公司，因此，要复现一项财经研究，通常需要同时获取文章和供应商的归类结果。CCFinance 通过记录所抓取的报道版本、其中提及了哪些发行人以及每个发行人出现的频率，使研究人员能够依据不同的筛选规则重构企业报道覆盖情况。在 Fine News 的 1,480 万条英语记录中，这些记录源自公开的 Common Crawl News，我们识别出 522 万篇不同且日期一致、包含发行人关联的文章。精确重构需要受限快照及衍生制品。我们在 250 篇由模型标注的文章导语上比较 GLiNER 和 Stanza，并根据盲化模型判断评估发行人关联和报道重点。在 475 条已接受链接中，估计精确率为 96.8%。在包含检测到的公司的留出文章中，要求出现三次提及会使主体精确率从 31.9% 提高到 66.1%。在 2017–2025 面板中，同一规则将至少有 50 篇合格报道的发行人-年份数量从 13,918 个减少至 8,641 个。三项重新实施分别考察波动性跟踪、银行网络和隔夜收益。在网络研究中，1月至2020年4月期间，跨出版商共同出现的银行配对数量从 177 对增加到 242 对，但在路透社内部则从 50 对降至 35 对。

1 引言

由于实证研究使用的大部分财经新闻来自专有档案，研究人员若想复现已发表的估计值，往往需要同时获取原始报道，以及将这些报道与公司关联起来的供应商标注。例如，Heston 和 Sinha [13,Section 2], 使用时间戳和报道标识符，将专有 NewsScope 情绪分数及公司标签与路透社新闻档案中的全文匹配起来。即使拥有同一篇报道的副本，如果一位研究人员统计财报中提到的每家公司，而另一位只统计财报所报告业绩的公司，两人构建的公司样本也可能不同。以往向研究人员开放的资料集通常涵盖的出版商较少或时间跨度较短，而重建样本所需的字段可能分散在不同产品中。例如，公开发布的 SEntFiN 提供了 10,753 个标题的实体情绪标注 [30], ，而研究投资者能够观察到的内容时，可能还需要文章正文、发布时间和采集时间，

以及该日期的上市发行人身份。

在商业订阅之外开展工作的研究人员，通常以两种方式之一获取新闻：要么通过出版商或供应商的限定发布，要么直接收集文章。美国国家标准与技术研究院根据研究协议分发路透社语料库第一卷和汤森路透文本研究集（TRC2）。前者涵盖一年的英语新闻专线，后者涵盖2008年1月至 2009 [23]年2月这一时期。从出版商或聚合商收集文章，如Dong 等人 [11]利用纳斯达克的分代码页面所做的那样，使研究人员能够控制其所保留的材料，尽管采集时可用的页面以及约束其使用的条款决定了该样本中有多少可供其他人重新

获取。例如，R´emy和Ding [26], 记录了其原始路透社和彭博下载因版权原因被撤回的情况。

由于直接采集需要重新访问正在运行的网站，研究人员必须遵守出版商条款和爬取偏好，同时尽量减少对服务器造成的负载。即使网站允许已验证的搜索爬虫访问 [5], ，而这些爬虫带来的引荐流量可以为出版商带来读者和广告收入 [10]，用于防范滥用流量的网站保护措施也可能给研究爬虫带来挑战。共享档案库让多个团队能够在一次抓取后检查同一份响应。这对复现研究很重要，因为两次访问同一个 URL，返回的文本可能不同。如果出版商在发布后为报道添加了对另一家公司的回应，不同日期采集数据的研究人员甚至可能对报道中提及了哪些公司产生分歧。如果两位研究人员都能找回同一份采集记录，他们便可以转而考察各自的规则如何处理同一段文字。例如，Common Crawl 会遵循出版商爬取指令，并在遇到速率限制或服务器错误后放慢请求速度 [8]。

这种档案可重复利用的规模，在语言模型预训练中显而易见：按采样权重计算，经过筛选的 Common Crawl 数据占 GPT-3 训练数据混合的 60% [4, 表 2。2]非营利组织 Common Crawl 基金会向研究人员、政策制定者和公众开放其 URL 索引及底层网页响应 [7, 9]。商业搜索服务会为检索和排序结果而索引网页 [12], ，而公共档案还允许研究人员在网站内容发生变化后，检查网页的特定版本。2016年，该基金会推出了 CC-News，这是一个面向全球新闻网站的爬取项目，其每日 Web ARChive（WARC）文件记录了抓取的响应及其采集时间 [22]。访问这些副本仍须遵守来源条款及相关内容的权利要求 [6]。

在一份存档响应能够作为文章文本使用之前，研究人员必须识别其正文，并决定保留哪些副本。RefinedWeb 展示了筛选和去重在为语言模型预训练准备网络数据时所发挥的作用 [25]。对于新闻，Infini-News 从 CC-News 中抽取文章正文和结构化元数据，并使用其抓取溯源信息[16], 对这些内容建立索引；而 Fine News 则对这些记录应用语言和质量筛选器，并移除重复文本 [31]。这些转换使得从一个经过清理的正文开始的研究人员能够通过抽取过程追溯至一份存档响应。对于语言模型而言，该正文已经可以作为训练样本。对于金融面板数据而言，还需要做出进一步决策，因为价格和收益是针对特定证券在特定日期观测得到的。

即使经过这些处理，一篇文章也可能提及多家公司，报道只涉及其中一家公司的事件，或对每家公司表达不同的情绪 [29, 30]。命名实体识别能够定位这些提及，而发行人关联则为其分配带有日期的身份，使研究人员可以将它们与金融记录连接起来 [28]。文章形式标签可以区分新闻报道与新闻稿等形式，而事件标签则识别某段文字所描述的发展类型。相关内容对投资者的影响可能取决于他们持有哪家公司的股份，因此，发行人特定情绪所对应的对象不同于句子的整体语气。稠密嵌入还使研究人员能够比较共享词汇很少的段落，并将这些表征用于检索或收益模型 [17, 27]。因此，评估的相关单位取决于研究使用的是整篇文章、其中的一处提及，还是文章与发行人之间的关联。

在 CCFinance 中，我们对一个源自 CC-News 的冻结版 English Fine News 发布版本应用出版商筛选器，得到 14,807,287 条暂存记录。为描述这一筛选结果，我们估计同一主机名规则在四个固定源版本中的产出。在冻结样本中，我们核对发布日期与抓取日期，选取规范副本，并使用由架构驱动的 GLiNER2.5 Small 提取公司跨度和属性 [24]。我们将其跨度与 Stanza 进行比较，依据留出集上的模型判断评估带日期的发行人关联和主题分配，并审计新增标签和正文嵌入。通过固定年度文章要求，我们可以衡量当研究要求每篇文章都与该企业相关时，有多少发行人-年份不再符合条件。

上游 Fine News 修订版本可公开获取，但要重建完全一致的 CCFinance 样本，还需要受限的暂存快照和衍生制品。

我们通过三项同行评审研究中的若干设计，追踪这些选择所带来的影响。在出版商—月份层面，我们将 Baker 等人 [3] 的股票市场波动率跟踪器与基于同一日历中存档文章构建的跟踪器进行比较。当分析单位是一对银行（如 Kazinnik 等人 [15], 的共同提及设计）时，我们发现，相连配对的数量在不同出版商之间有所增加，而在路透社内部则有所下降。对于 Lopez-Lira 和 Tang [18], 的隔夜设计，报道还必须在交易截止时间之前已可获取。恢复发布时间后，我们便能在更多交易时段评估投资组合的多空腿，而重新拟合后平均漂移仍为负值。

2 文献综述

金融新闻语料库中的观测值如何界定，部分取决于新闻的采集方式。在构建 FNSPID 时，Dong 等人 [11] 从纳斯达克按股票代码划分的页面上的标题和网址入手，获取已关联页面的正文，并补充较早的新闻数据。他们的建模实验以 1570 万条时间对齐的新闻记录为起点，并分别考察文本和股票代码的可用性。按股票代码采集，为研究人员提供了从证券追溯到报道的路径。从网络档案入手则颠倒了这一顺序：研究人员看到的是文本副本，必须确定它涉及哪些证券（如果确实涉及证券的话）。一篇报道若以两个股票代码编入索引，就能形成两个公司观测值；而同一页面的两次抓取，则可以记录一次修订版本。因此，时间对齐记录的数量与存档副本的数量衡量的是新闻报道覆盖的不同方面，也提示了不同的完整性检查方法。从证券名单入手，可以查明其中哪些证券完全没有可检索的文本。从存档文章入手，则可以查明被识别出的公司有多少能对应到名单上的发行人。然而，若要了解哪些公司很少或完全没有报道，两种方法都需要一份独立于已发现文章而编制的名单。否则，没有报道的发行人就会连同其缺席的证据一起从分母中消失。

对于从档案网页着手的研究人员来说，语料库的选择也决定了他们会继承哪些编辑和处理决策。CC-News 的每日文件保留了从新闻网站抓取的响应内容 [22], ，而 Infini-News 提取文章文本和元数据，并可将其与抓取溯源信息一并检索 [16]。Fine News 则进一步按语言和质量筛选提取的记录，并为重复文本选出代表项 [31]。对于预训练而言，筛选和去重有助于决定模型会接触哪些段落，以及接触这些段落的频率 [25]。对于金融面板数据，同样的运营操作可能改变所度量的报道覆盖量。两家出版商刊载同一篇通讯社新闻稿时，在研究媒体关注度时，两者都可能各自为研究贡献一条发布记录；但在研究共同提及时，将两者都计入则会对同一银行配对额外增加一条观测。如何处理第二份副本，取决于研究所估计的具体指标。

在文本中找到一个组织，与识别出研究人员可以长期跟踪的证券，是两回事。GDELT 的全球知识图谱记录了受监测新闻中的组织、主题和语气，因此研究人员可以追踪事件的报道方式 [33]。FiNER-ORD 则使识别本身能够接受评估：Shah 等人 [29]为 201 篇财经新闻文章中的人物、组织和地点提供了人工标注。这些标签使研究人员能够判断抽取器是否找到了文段中的正确词语。在金融应用中，正确抽取出的组织可能是交易所、监管机构，或拥有上市母公司的子公司；这些组织对相关提及能否纳入发行人面板，各有不同影响。Schwenkler和Zheng [28] 明确指出了这一额外步骤，他们

在新闻中识别组织并关联公司，从而构建网络。漏掉一个名称可能会导致一对真实关系被遗漏，而错误的归属则可能将这对关系归到另一家企业名下。

即使指派正确，文本究竟表达了什么仍是一个未解的问题。SEntFiN 中由人工标注的 10,753 条标题里，有 2,847 条包含多个实体，而针对这些实体表达的情绪可能各不相同 [30]。例如，一条标题宣布一家公司收购另一家公司。买方和目标公司都应能按名称检索到，但投资者在解读交易条款时，可能会对每只证券得出不同结论。Heston 和 Sinha [13]可以通过将新闻文本与供应商公司标签及情感分数相匹配，研究这类公司层面问题。对于存档文本，可以分别考察相应的判断：名称是否关联到了正确的发行人、报道是否与该发行人有关，以及赋予该发行人的情绪是否符合结合上下文作出的判断。

金融应用决定了哪些判断会纳入估计值。在 Baker 等人 [3], 的波动率追踪器中，符合词典规则的文章会计入出版商的月度份额，因此该出版商当月刊载的其他文章也会通过分母进入计算。在 Kazinnik 等人 [15], 的银行网络中，建立关联需要有证据表明同一篇报道涉及两家机构。Lopez-Lira 和 Tang [18] 的隔夜收益设计还增加了时间要求，因为只有在建仓之前就已发布的标题，才能为交易决策提供依据。对于抓取的文章正文，这项要求涉及包含模型所用信息的版本。因此，我们从存档副本开始构建 CCFinance，并在指派发行人、评估其在报道中的角色以及确定哪些观测值可以纳入各项设计时，始终追踪该副本的身份。

3 方法论

为了确定可据此构建财经观测值的存档副本，我们从 Fine News 的英语 eng Latn 数据分片开始。若记录的主机名与我们的 115 家出版商列表中的某个条目或该条目的点分隔子域名相匹配，我们就保留该记录。该列表包括提供财经报道覆盖的一般兴趣类出版商。对于每个匹配项，我们保留文章 ID、正文、源URL，以及可用的抓取和出版商元数据。上游修订版本和一个经过独立校验和验证的已分阶段修订版本，在 168 个已分阶段的 Parquet 数据分片中标识出 14,807,287 个不同的标识符（来源比较见附录表A1）。

为了了解来源选择如何影响这一筛选条件的产出率，我们将相同的主机名规则回溯应用于四个固定版本，其中包括 FineWeb-HQ [21]。在每个版本内，我们进行四次独立的有放回抽样：先均匀抽取一个文件，再在该文件中均匀抽取一个行组。我们使用逆概率权重估计与出版商匹配的总量和英语总量，二者之比即为估计的出版商产出率。两个新闻集合的点估计值较高（附录表A1）。由于这些版本在英语报道覆盖范围以及对重复项的处理方式上存在差异，这些产出率描述的是每个版本内的物理行；而若要估计某个来源新增了多少独立报道，则需要在不同版本之间匹配记录。

在选定的版本中，若一篇文章是规范副本，且其报告发布日期与抓取时间一致，则该文章可用。尽管 Fine News 保留了上游重复集群的代表项，暂存记录中仍存在完全相同文本和规范化 URL 的重复项。我们分别在每个文本组和每个规范化 URL 组中选取最早抓取的副本；若时间相同，则按 ID 排序，并在两种比较都判定为重复项时记录文本匹配结果。一个副本必须在两种比较中都胜出，才有资格入选，即使优先级更高的较早副本随后未能通过日期检查也是如此。由于出版商可能会修改同一 URL 下的文本，规范 ID 标识的是所保留的特定正文。

日期检查允许抓取时间晚于报告发布日期最多七天；考虑到时区差异，也允许早于发布日期最多一个日历日。因此，在日级语料库中，标注为星期一的页面即使到星期四才被抓取，仍可计入星期一的新闻；不过，对于交易信号而言，星期四才是我们首次观察到该版本的日期。在冻结语料库中，11,260,839 个可用副本里有 237,193 个是在报告发布日期后的两至七天内抓取的。另有直接证据表明，之后再次访问可能会返回不同的正文：在观测到文本哈希的 26,897 个重复规范化网址中，有 23,779 个具有多个哈希值。哈希值变化本身并不能说明出版商是否添加了具有财务相关性的信息，但这确实解释了为什么我们会保留计算信号所依据的特定文本的抓取时间。

为确定该版本所命名的公司，我们使用 GLiNER2.5 Small [24] 提取候选公司跨度，保留其文本和偏移量，并将其与一张包含带日期的上市记录和别名的表进行匹配。该表整合了 SEC注册人标识符和当前及曾用名称，以及来自 Wikidata 的上市、别名和子公司证据。在已知有效日期的情况下，我们根据文章日期评估别名证据。交易所股票代码或路透代码可以锚定一个身份，并帮助解析文章中随后出现的提及；对于含义不明确的候选项，则予以保留，链接器也可以选择不作判断。例如，在锚点审计中，一篇较早的 Sprint 新闻中与某股票代码关联的发行人，与该股票代码后来被分配给的发行人并不相同。

Schwenkler和Zheng [28] 同样先使用 Stanford CoreNLP 识别组织，然后在新闻衍生网络中链接公司。为考察链接决策之前的公司识别情况，我们在出版商筛选后的样本的导语中，比较 GLiNER2.5 Small 和 Stanza（另一种斯坦福工具包）提取的精确公司跨度（附录表 A2）。FiNER-ORD 还记录了将命名实体识别迁移到金融文本中的困难 [29]。

即使发行人的身份很明确，其在报道中的角色也可能有所不同。一场涉及 Gap 和 Tapestry 品牌的争议可能与两家公司都有关，而一篇关于 Tesla 的文章可能只是顺带提及 Nvidia。我们保留每个获接受的文章–发行人对及其关联证据；当一篇文章至少包含三次对某发行人的提取出的提及时，就将其计为重点报道该发行人，不再设置额外的相关性阈值。因此，一篇提及某家已关联公司三次、另一家一次的可用报道，在语料库中计数一次，同时计入两家公司的可用报道覆盖，并计入第一家公司的重点报道覆盖。

为了描述每个层级的文本，GLiNER 在包含 64 个词重叠的 384 词时间窗口上使用联合模式，记录提取出的提及的文章形式、局部相关性和面向投资者的情绪，以及各个时间窗口的事件得分。对于隔夜探测，EmbeddingGemma 会将每篇报道中最多 2,048 个词元编码为固定的 768 维正文向量 [27]。我们检查向量范数及其与冻结的文章标识符之间的对齐情况，并在单独的探索性检索任务中考察公司特定表示（附录表 A9）。

正文与发行人的区别同样适用于其表征。如果一篇报道同时点名交易双方，那么无论研究人员查询哪一方，存储的正文向量都相同。以提及其中一方的句子为条件生成嵌入，或许有助于区分这两个账户，前提是这些句子包含不同的信息。探索性检索任务通过要求每家公司的表征在名称被遮蔽后检索出其自身账户，来检验这一条件。只有16篇经主模型判定为符合条件的留出报道适用于每个编码器，低于原计划的最低数量60篇，而且两家公司仍可能共享同一个选定句子。因此，我们在隔夜实验中使用正文向量，并将文章—发行方链接保留为单独记录，用于说明该报道涉及或提及了哪些公司。

A面板。文章要求	文章标识符
已分阶段	14,807,287
日期一致	11,287,476
可用的规范副本	11,260,839
带有已接受的发行方链接的可用内容	5,220,290
带有聚焦的发行方链接的可用内容	2,547,597

面板 B。年度最低值	Raw	可用	重点
1篇文章	33,109	33,024	31,745
10篇文章	26,930	26,434	22,243
50篇文章	14,539	13,918	8,641

表1：通过文章级规则筛选后保留的不同暂存标识符（A面板），以及在2017–2025年间达到各年度文章最低数量要求的发行人-年份（面板 B 中共有34,781个符合条件的发行人-年份）。一篇文章可以计入多个发行人；对于每个被提及至少三次的关联发行人，该文章均被视为重点文章。

4 数据与实证设计

我们先按文章、再按发行人统计报道覆盖，因为同一篇规范副本可能提供有关多家公司的新闻。表1的 A面板按行对每个不同的文章 ID 计数一次，而面板 B 则追踪2017年至2025年这九个完整日历年中的发行人。对于某一年，如果发行人拥有有日期的上市记录，或拥有无日期的当前上市记录，且该记录在当年或前一年得到公众流通股证据的支持，则其符合条件。即使没有任何文章能够与符合条件的发行人建立关联，这些发行人仍保留在分母中。

由于我们按新闻的发表年份归类，但在抓取时观察到新闻，因此某一年存档网页数量多，并不意味着该年有大量关于企业的报道符合条件。2017年抓取队列体现了这一区别。当年抓取的3,659,035条暂存记录中，有2,886,747条是在报告发布日期七天之后才获取的，其中英国和印度的路透社媒体平台各有超过一百万条延迟抓取记录（附录表A7）。若将所有这些网页都计入2017年，就会混淆爬虫访问出版商的日期与出版商报道新闻的日期。按照发表年份规则，2017年有3,240个符合条件的发行人-年份，其中854个达到50篇可用文章的最低要求；相比之下，2020年有3,607个符合条件的发行人-年份，其中1,844个达到该要求。因此，某企业在一年中符合条件，并不能证明其在整个研究期间都持续获得报道覆盖。

在34,781个符合条件的发行人-年份中，33,024个至少有一篇可用文章。若年度最低要求为50篇文章，则13,918个发行人-年份凭借可用链接符合条件；若要求每篇报道至少三次提及该发行人，则有8,641个发行人-年份达到同一最低要求。这一限制带来的影响也因公司规模而异。在测量的公众流通股年度五分位组中最低的一组，6,533个发行人-年份中有329个（5.0%）达到可用文章的最低要求，105个（1.6%）达到重点要求。在最高五分位组中，相应数量分别为6,526个中的5,790个（88.7%）和4,561个（69.9%）。附录表A8列出了中间五分位组的数据以及按年份划分的细目。对于规模较小的企业，即便是初始要求也筛选出一个范围狭窄的群体：最低五分位组中，约二十分之一的发行人-年份达到50篇文章的最低要求。进一步要求主题聚焦后，首轮筛选通过的329个中仅保留105个；而最高五分位组中，首轮筛选通过的5,790个有4,561个得以保留。因此，对于一项对符合条件的企业估计单一系数的分析，相关报道的定义一旦改变，估计该系数所依据的样本构成也会随之改变。

为了解哪些报道会纳入某发行人的报道覆盖范围，我们首先考察抽取器是否识别出公司名称，然后考察链接器是否将该名称分配给正确的发行人。其中

在250篇抽样导语中的712个模型参考公司跨度里，按精确匹配计算，GLiNER找回了412个，Stanza找回了421个；而允许边界重叠后，两者的公司跨度F1值相差不到半个百分点（附录表A2）。精确评分要求预测名称的首字符和末字符都与参考标注一致。例如，在一篇同时提及瑞银集团的导语中，Stanza将所有格附加到了瑞士信贷集团名称上。读者或许能够识别出这两家银行，但这些额外字符导致Stanza在精确评分中失去了这对匹配。相反，如果某个组织在Stanza更宽泛的标签下被正确识别，但不属于仅限商业实体的参考标注，那么它会被计为假阳性。因此，该比较描述了每个抽取器为公司链接任务提供的内容；该任务的下一步是判断该名称是否指向符合条件的发行人。

对于身份比对，我们独立于链接器的判定抽取提取片段，并要求模型评审员根据标记文本和中立搜索结果识别文章日期时的发行人，同时不向其提供预测结果。在留出抽取片段中，475 个已接受链接的估计精确率为 96.8%（95% 区间为 94.0–99.6%），在可与参考链接关联的提及中，召回率为 70.8%（64.6–77.1%）。在相同的 886 条留出提及中，使用完整链接器时，加权发行人链接 F1 为 81.81%；若规范化名称必须唯一对应一个带日期的发行人，则为 59.46%。将学习得到的别名及其频次限制在先前日历年内，结果为 78.57%（附录表 A5）。这两种时间限制的目的不同。上市日期决定候选发行人当时是否可能是该文章的发行人，而先前时间窗口决定哪些较早的名称用例可用于消歧。在 475 个生产分配中，前一年回放保留了 464 个，将其中一个的身份更改，并对十个弃权。因此，即使两个版本的链接器查询的是同一条带日期的上市记录，历史用户仍可能失去一个链接。

确定一个名称所指代的发行人后，我们进一步考察文章是否与该发行人相关。这一比较采用文章级判断，并以 GLiNER 在文章中至少检测到一家公司的结果为条件。在对 217 篇开发集文章进行比较的相关性、位置和提及次数规则中，三次提及规则取得了最高的主体 F1。在另外 183 篇测试文章上，该规则保留了 65 个经判定的主体配对中的 37 个。相较于统计所有已接受链接，要求至少提及三次使主体精确率从 31.9% 提高到 66.1%（54.8–77.3%），但使召回率从 92.3% 降至 56.9%（44.3–69.5%）。在 188 个已接受的测试配对中，判定员确认其中 60 个属于主体。要求至少提及三次后，保留了 56 个配对，其中 37 个是主体。当研究需要文章主要围绕该企业展开时，精确率的提升很有价值；但与此同时，这一规则也排除了更宽泛规则所保留的 23 个主体配对。若研究人员愿意纳入更多附带提及，则可以基于同一批记录作出不同选择：在留出文章中，至少提及两次可找回经判定主体的 76.9%，精确率为 55.6%（附录表 A4）。

由于生产面板统计的是具名提及，即使文章继续用代词或描述来讨论企业，提及次数也可能不足三次。当我们也统计同一批测试文章中 Stanford CoreNLP 共指链里锚定到单一关联发行人的名词性和代词性提及时，另有十二个文章—发行人配对符合条件，其中七个被判定为主体。召回率从 56.9% 上升至 67.7%（配对增幅为 10.77 个百分点，95% 区间为 $[3.92, 19.61]$），而精确率从 66.1% 降至 64.7%。另一次由 GPT-6.1 Terra 进行的审查认定，这些配对中新增的 42 条引用有 28 条指向锚定的发行人。十二个配对中有七个仅凭这些已确认的引用仍符合条件，其中六个被判定为主体。

即使一则报道已被归属给某家企业，其语气对该企业投资者而言，仍可能与新闻所暗示的不同。在主要评审员标注为积极、负面或中性的 400 条标注提及中，有 358 条符合这一条件；GLiNER 的公司特定情感加权宏平均F1为 46.79%，而将每家公司对应为其所在句子的 Stanford情感时，该指标为 40.84%。两者相差 5.95 个百分点，其配对95%区间涵盖零值 $[-1.33, +13.98]$。在一项盲评中，GPT-6.1 Terra 将 40 个抽样目标中的 23 个标记为无效公司跨度。因此，我们将发行人

情绪视为探索性分析。

当研究人员按形式或事件筛选文章时，也会遇到类似的问题。为了评估这些选择如何影响样本构成，盲法模型评审员在涵盖输入窗口中的文章形式、局部提及属性和事件族的三个确认框架中，各自对 400 个抽样单位进行标注。在可用文章中，分类器判定的加权新闻稿占比为 42.1%，主要评审员判定的占比为 4.2%；参考标注为正的抽样文章有 14 篇。若按预测交易设置筛选条件，将会漏掉许多被评审员标记为交易型的时间窗口，因为相对于这些标注，召回率仅为 23.6%。附录表 A3 按类别报告了估计值。这些误差会以不同方式影响两种常见的设计。剔除预测为新闻稿的文章，会舍弃参考评审员认为属于其他新闻形式的文章；而仅保留预测为交易型的文章，则会漏掉许多根据该评审员的判断包含交易的时间窗口。这里的单位同样重要：漏掉一个时间窗口意味着局部遗漏，而排除整篇文章则取决于其其他时间窗口中检测到了什么。

在这三种金融应用中，相关决策所作用的单位各不相同。对于股票市场波动率跟踪器，我们用包含所需词组的文章数量，除以该出版商同月所有符合条件的文章数量。对于银行网络，我们还需要知道哪些机构共同出现，因为每篇符合条件的文章都会针对其在指定周提及的映射银行中的每一组无序对贡献一条观测。因此，一篇涉及三家银行的报道会产生三组无序对，而每组无序对的权重则取决于提及该组合的报道数量。

在隔夜实验中，识别证券也就确定了要使用哪个交易所交易时段。我们将来自两家印度出版商的及时新闻与美国上市证券配对；在固定面板中，纳斯达克交易的证券对应 740 条文章—证券记录，纽约证券交易所交易的证券对应 364 条。对于其中 270 条记录，登记册缺少上市起始日期，因此，纳入资格还要求存在与发行人已关联的、2021 年之前可用的新闻。为确定开盘前可能观测到的内容，我们要求首次发布晚于前一交易所收盘时间且早于纽约时间 09:00；抓取时间则须在发布时间与截止时间之间，且包含截止时间。如果爬虫在 09:00 之后抓取隔夜报道，我们只能在截止时间之后看到正文。我们度量开盘至收盘的漂移，因为从收盘至开盘的反应在文章出现之前就已开始。滚动探测器使用每次重新拟合前已实现的收益进行训练，而编码器是在研究期间结束后发布的，其预训练与档案中报道的重叠情况尚未测量。每个启用的多头腿或空头腿至少需要两种证券，因此，当两个腿均不符合条件时，即使某个交易时段有预测，该时段对投资组合均值的贡献也为零；而没有预测的交易日则不纳入评估。

5 结果

表2比较了这三个估计值对决定其样本的各种决策的敏感性。我们改变纳入波动率跟踪器的月份和出版商、银行可以共同出现的报道，以及证明某家企业的新闻在交易截止时间之前已经发布的证据。

5.1 发布方—月份波动率

波动率基准衡量每个档案中的追踪器能够解释多少相同的月度 VIX 波动。我们首先使用作者的归档数据 [1]，按其显示精度重现 Baker 等人 [2, 3] 的六个月度数据列。CCFinance跟踪指标采用来自 Benzinga、Forbes、MarketWatch 和《华尔街日报》的 675,234 篇日期一致的规范文章；这四家媒体中，只有《华尔街日报》属于作者的十一家报纸样本。要计为命中，标题和正文中必须包含来自以下三个

练习	评估单位和支持度	基线和参考	敏感性分析
波动性	49 个共同月份，675,234 篇符合条件的文章，四家出版商	VIX R^2 0.515 对比 0.647，gap -0.132 $[-0.215, +0.110]$	Gap -0.041 剔除后来自2020年3月至4月两者 fits
银行网络	不同的无序每组中的银行配对	177 至 242 对，而此前为 210 至 2023年7月手稿中为 172	仅路透社，从 50 降至 35。匹配后的平均增幅 $+56.3$ $[5, 83]$ 在六周内及 $+19.1$ $[-25, 59]$ 在十四周内
隔夜漂移	财报周，34个中有33个参考节点映射的277 个评估会话，42 个双腿均有的情况，以及 123 个两者皆无的情况	手稿 -10.15基点 $[-25.2, +4.2]$ 相比之下， $+11$ 基点的参考 MiniLM 标题样本	恢复的时钟给出 340 个评估会话，其中 110 个带有双腿，以及在重新拟合后 -7.73 基点重新拟合

表 2：对于波动性，配对六个月区块区间度量的是同期拟合差异中的不确定性。对于银行，范围涵盖匹配报道抽样的中间95%；六周分析中每周有682篇报道，十四周分析中每周有158篇。漂移以每个评估交易时段的总基点数度量，并采用以已保存的预测为条件的21个交易时段区块区间。

词典组的词语，分别涉及经济、股票市场和波动性。我们将每家出版商的命中占比除以其在截至2022年12月的 40 个校准月份中的离散度，计算标准化占比的平均值，再将该平均值缩放至校准期 VIX 的均值。从 2019年9月到 2023年9月，每个追踪器分别进入一次回归，回归中包含截距，并使用来自 FRED 的相同月度 VIX 结果。CCFinance跟踪指标的R^2 为 0.515，作者的追踪指标为 0.647，差值为 -0.132 （配对 95% 区间 $[-0.215, 0.110]$）。

从两项回归中剔除2020年3月和4月后，各追踪器所能解释的变异都会减少，VIX拟合差距缩小至 -0.041。在校准保持不变的情况下，剔除Forbes会使两个日历下的差距都扩大；在全样本中，差距从 -0.132 扩大到 -0.360 。当我们仅保留两组报纸共同包含的出版商《华尔街日报》时，其追踪器对月度 VIX 波动的解释占比较小（$R^2 = 0.110$），且与作者构建的十一份报纸综合指标的相关系数为0.397。这些比较有助于理解拟合差距：共同日历使波动性结果保持一致，但每个追踪器的新闻分母仍然不同。如果某出版商在波动较大的月份扩大市场覆盖范围，那么词典命中占比既取决于新增报道中有多少讨论波动性，也取决于该出版商刊载了多少其他新闻。对每家出版商的占比进行标准化，可将其变异置于共同尺度上，同时保留其对报道构成的依赖。单一报纸比较采用基准综合指标，因为提供的月度基准汇总了全部十一份报纸（附录表A10）。

5.2 财报周的银行配对

出版商的选择也决定了哪些银行会共同出现，因为只有当保留的报道同时提及两家机构时，这一对机构才会进入网络。在 Kazinnik 等人 2023年7月设计中针对 1月13日至19日和2020年4月13日至19日这两周的分析里， $[14, 15]$, 我们将 34 个参考节点中的 33 个的 Factiva标签替换为 CCFinance链接，在某些情况下通过外国集团的上市母公司进行映射。重构网络中最大的六家机构构成一个完全核心，相连配对的数量从 177 增至 242，而 2023年7月手稿报告的数量则从

210 降至 172。将 CCFinance 限定为路透社会改变出版商构成和可用报道数量，在该样本中，数量从 50 降至 35。仅排除纳斯达克时（162 增至 230），以及保留所有出版商但仅使用直接银行映射时（77 增至 133），这一增长仍然存在。

为检验在固定出版商配额下这一增幅是否仍然存在，我们合并联合发行副本，并匹配从各出版商抽取的报道数量。在每周匹配682篇报道的六周期间，2020年1月至4月的平均增幅为56.3个相连配对，其中间95%的重抽样范围为 $[5,\ 83]$。在同样的六周期间，移除Truist，以获得一个在其2019年合并前即可使用的名单后，增幅降至45.4个配对。随后要求十四周期间采用相同的出版商配额，每周便只剩158篇报道，并排除纳斯达克；纳斯达克在原先的682篇报道中提供了271篇。平均增幅降至19.1个配对 $[-25,\ 59]$，接近在相同配额下2020年7月至10月增加的19.0个配对。这些范围反映了从所选周次的报道抽取中产生的变异，而六个安慰剂时期则展示了网络在其他日期的变化情况（表A11）。因此，延长周数进行比较，就要求档案在更长时间内持续提供一组规模更小的出版商。在这一共同样本上，2020年1月至4月的增幅与随后7月至10月的变化大致相当，尽管未匹配的1月和4月网络相差65个配对。

对于网络而言，只要某家银行在同一周与另一家银行一同出现，该银行就可以提供证据。隔夜设计在两个方面对时间要求更严格：每篇文章都必须在其交易截止时间之前发布，而且必须有多只证券在同一日期符合条件，才能将其收益合并到投资组合中。

5.3 发行人-交易时段漂移

对于隔夜实验，报道覆盖方面的问题是：同一交易时段内是否有足够多的证券及时获得新闻，从而构建投资组合多空腿。我们将 Lopez-Lira 和 Tang[17, 18] 选定的设计应用于 Business Standard 和 Mint Lounge 抓取的 996 篇正文，这些正文已关联至 171 家发行人；我们使用 EmbeddingGemma 向量和雅虎价格历史数据，替代作者使用的与 RavenPack 匹配的标题和 CRSP 价格。按照他们的档案，我们在六个月时间窗口内，基于此前的收益符号分别拟合滚动逻辑回归 [19]。在每个交易时段内，最早符合条件的预测文章为其对应证券提供信号。我们在每个符合条件的多空腿中对证券等权配置，并按每个活跃腿一单位名义金额汇总带符号的收益；不活跃的腿则赋值为零。虽然这些探针基于更早的收益进行训练，但发行人链接器会利用整个档案估计别名频率，因此历史关联可能会借助后续证据。

在32个月中，有12个月的训练样本行少于100行；在拟合的20个月中，有16个月的基准求解器在达到10次迭代上限时发出了警告。固定样本检验评估了2021年10月至2024年5月期间670个可能交易时段中的277个。评估的交易时段中，只有42个同时有多头和空头投资组合，另有123个两者皆无。在不计成本的情况下，每个评估交易时段的平均开盘至收盘漂移为-10.15 基点；以已保存的预测为条件，21个交易时段的区块区间为 $[-25.2,\ +4.2]$ 基点。分母包含每个评估交易时段，因为既未建立活跃腿的投资组合不会部署资本，当天的收益率为零。在至少有一个活跃腿的条件下，相同的毛收益在154个日期上的平均值为 -18.26 基点。这两个均值分别描述每个评估交易时段的收益率，以及规则实际建立了头寸的日期上的收益率（附录表 A12）。

从历史版 Business Standard 格式中恢复首次发布时间，新增了 667 个不同的公司—会话。有了这些额外观测值，评估交易时段的数量从 277 增至 340，其中有 110 个包含双腿（此前为 42 个），另有 74 个两者皆无。由于恢复的观测值也纳入了训练窗口，我们在扩展面板上重新拟合探针，平均

漂移仍为负值，每个评估交易时段为 -7.73基点。在两个模型都评估的 277 个日期中，恢复使平均毛漂移变化了 $+2.45$ 基点（配对95%区间 $[-6.80, +13.74]$）。这一增益能否在交易成本下持续，取决于这些日期中实际交易了多少条腿。若每条活跃腿在入场和出场时均单向承担五个基点的成本，则共同日期上的配对变化为 -1.27 基点 $[-11.19, +9.25]$。即使日历保持不变，额外头寸也会产生额外成本；这些区间涵盖毛值和净值变化的正负两种可能（附录表 A12）。

Lopez-Lira 和 Tang 的表 8 [17]报告了在范围更广的 RavenPack 和 CRSP 样本中，MiniLM 标题的平均日漂移基点数，其中涵盖了不同的文本、发行人报道覆盖、表征和价格。对于固定的 CCFinance 样本，标题与正文的配对比较仍无法确定哪种表征带来的平均漂移更高（附录 E）。

拟合探针的稳定性在这里很重要，因为某只证券的信号发生变化，就可能决定其双腿是否满足至少包含两个成员的要求。在原有正则化强度下，允许基准求解器最多运行 1,000 次迭代，可消除 16 条警告，并使 738 个证券–时段信号中的 9 个发生变化。评估的仍是相同的 277 个日期，但如今只有 39 个日期支持双腿，平均漂移为 -8.73 基点。恢复时间戳能增加构建投资组合的机会，而分配给合格证券的正负号及其随后获得的收益，则决定了增加的报道覆盖有何价值。附录表A13列出了每种表征的成本情景。

6 结论

一篇存档文章只要确定了来源、日期以及其作为独立副本的状态，就可以计入出版商的月度新闻计数。对于银行网络，只有当同一篇文章提及两家已识别银行时，它才构成一项关系；而如何处理联合发行副本，则决定了这项关系获得多大权重。构建隔夜投资组合还需要进一步判断截止时间之前哪些信息可观测，以及同一交易时段内是否有足够的合格证券来构成一个投资组合部分。将这些决策与估计值一并记录，研究人员便能了解，更严格的规则会导致报道覆盖减少多少，以及哪些公司或日期造成了这种差异。

CCFinance 为冻结后的选样保留了这份记录，并提供发行人身份、主题焦点和支持性元数据的审计信息。尽管上游档案及其文档均可根据各自条款公开访问，但访问暂存快照的匿名请求需要经过身份验证，而在审查期间，项目代码库和冻结输出的访问受到限制。恢复所需输入后，这些流程即可重新生成报告中的分析结果（附录表 A14）。我们计划在审查结束后发布代码和符合条件的衍生记录，并为每件制品分别规定使用条款。对于构建新面板的研究人员而言，台账能够追溯公司报道覆盖范围或投资组合支持度的变化，并将其关联到符合条件的单篇报道。

引用

[1] Scott R. Baker、Nicholas Bloom、Steven J. Davis 和 Kyle Kost。《政策新闻与股市波动性：复现数据》。Mendeley Data，第1版，2025年8月。URL https: // data.mendeley.com/datasets/rmhtcwnvrk/1。

[2] Scott R. Baker、Nicholas Bloom、Steven J. Davis 和 Kyle Kost。政策新闻与股市波动性。手稿，2025年7月25日，2025年7月。URL https://static1.squarespace.

com/static/5e2ea3a8097ed30c779bd707/t/688847f536b2b31218aeefab/1753761782006/Policy%2BNews%2Band%2BStock%2BMarket%2BVolatility%2BJuly%2B2025.pdf。第2.1节和表1。[3] Scott R. Baker、Nicholas Bloom、Steven J. Davis 和 Kyle Kost。政策新闻与股市波动性。《金融经济学杂志》，175:104187，2026年。doi: 10.1016/j.jfineco. 2025.104187。[4] Tom B. Brown、Benjamin Mann、Nick Ryder、Melanie Subbiah、Jared D. Kaplan、Prafulla Dhariwal 等。语言模型是少样本学习者。载于神经信息处理系统进展，第33卷，第1877–1901页，2020年。URL https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html。[5]Cloudflare。挑战恶意机器人。Web 应用防火墙文档，2026年4月。URL https://developers.cloudflare.com/waf/custom-rules/use-cases/ challenge-bad-bots/。[6] Common Crawl 基金会。Common Crawl 使用条款，2024年3月。URL https://www.commoncrawl.org/terms-of-use。[7] Common Crawl 基金会。关于 Common Crawl，2026年。URL https://commoncrawl.org/ about。访问于2026年11月。[8] Common Crawl 基金会。常见问题，2026年。URL https://commoncrawl. org/faq。有关 CCBot 的技术问题，访问于2026年11月。[9] Common Crawl 基金会。Common Crawl 语料库概览，2026年。URL https: //commoncrawl.org/overview。访问于2026年11月。[10]竞争与市场管理局。在线平台与数字广告：附录 S，大型数字平台与出版商之间的关系。技术报告，英国竞争与市场管理局，2020年7月。URL https: //assets.publishing.service.gov.uk/media/5efb22fbd3bf7f768fdcdfae/Appendix_S_-_the_relationship_between_large_digital_platforms_and_publishers.pdf。[11] Zihan Dong、Xinyu Fan 和 Zhiyuan Peng。FNSPID：一个全面的时间序列财经新闻数据集。arXiv:2402.06698v1，2024年。URL https://arxiv.org/abs/2402.06698v1。[12]谷歌。深入了解谷歌搜索的运作方式。谷歌搜索中心文档，2026年。URL https://developers.google.com/search/docs/fundamentals/how-search-works。访问于2026年11月。[13] Steven L. Heston 和 Nitish Ranjan Sinha。新闻与情绪：从新闻报道中预测股票收益。《金融分析师杂志》，73(3):67–83，2017年。doi: 10.2469/faj.v73.n3.3。[14] Sophia Kazinnik、Cooper Killen、Daniela Scid´a 和 John Y. Wu。新闻与网络：利用金融新闻报道度量银行关联性。CEBRA 手稿，2023年7月6日，2023年7月。URL https://congress-files.s3.amazonaws.com/2023-07/Scida_News_and_Networks_CEBRA2023.pdf。[15] Sophia Kazinnik、Cooper Killen、Daniela Scid´a 和 John Y. Wu。新闻与网络：利用金融新闻报道度量银行关联性。《银行与金融杂志》，187:107694，2026年。doi: 10.1016/j.jbankfin.2026.107694。

[16] Ruggero Marino Lazzaroni、Jana Lasser 和 Kirill Solovev。Infini-news：高效检索 13 亿篇经处理的 Common Crawl News 文章。arXiv:2605.18337，2026年。URL https://arxiv.org/abs/2605.18337。[17] Alejandro Lopez-Lira 和 Yuehua Tang。ChatGPT 能否预测股价变动？收益可预测性与大语言模型。arXiv:2304.07619v6，2025年10月。URL https://arxiv.org/abs/2304.07619v6。[18] Alejandro Lopez-Lira 和 Yuehua Tang。Chat GPT 能否预测股价变动？收益可预测性与大语言模型。《金融经济学杂志》，184:104335，2026年。doi：10.1016/j.jfineco.2026.104335。[19] Alejandro Lopez-Lira 和 Yuehua Tang。ChatGPT 能否预测股价变动？收益可预测性与大语言模型：复现包。Mendeley Data，版本 2，2026年6月。URL https://data.mendeley.com/datasets/f39x226htv/2。[20] Christopher Manning、Mihai Surdeanu、John Bauer、Jenny Finkel、Steven Bethard 和 David McClosky。Stanford CoreNLP 自然语言处理工具包。载于计算语言学协会第*52*届年会论文集：系统演示，第55–60页，2014年。URL https://aclanthology.org/P14-5010/。[21] Bettina Messmer、Vinko Sabol˘cec 和 Martin Jaggi。通过基于模型的数据选择增强多语言大语言模型预训练。arXiv:2502.10361，2025年。URL https://arxiv.org/abs/ 2502.10361。[22] Sebastian Nagel。新闻数据集现已开放。Common Crawl 基金会，2016年10月。URL https://commoncrawl.org/blog/news-dataset-available。[23] 美国国家标准与技术研究院。路透社语料库。文本检索会议数据文档，2009年。URL https://trec.nist.gov/data/reuters/reuters.html。[24] Mary Newhauser 和 Urchade Zaratiana。推出 GLiNER2.5：采用模式驱动接口的高效无跨度信息抽取。Fastino 研究成果发布，2026年8月。URL https://fastino.ai/blog/gliner2-5-span-free-information-extraction。[25] Guilherme Penedo、Quentin Malartic、Daniel Hesslow、Ruxandra Cojocaru、Alessandro Cap- pelli、Hamza Alobeidli、Baptiste Pannier、Ebtesam Almazrouei 和 Julien Launay。用于 Falcon 大语言模型的 RefinedWeb 数据集：仅凭网络数据，超越精选语料库。arXiv:2306.01116，2023年。URL https://arxiv.org/abs/2306.01116。[26] Philippe R´ emy 和 Xiao Ding。彭博和路透社财经新闻数据集。数据集发布记录，2015年。URL https://github.com/philipperemy/financial-news-dataset。[27] Henrique Schechter Vera、Sahil Dua、Biao Zhang、Daniel Salz、Ryan Mullins 等。EmbeddingGemma：强大而轻量的文本表示。arXiv:2509.20354v3，2025年。URL https://arxiv.org/abs/2509.20354v3。[28] Gustavo Schwenkler 和 Hannan Zheng。新闻所揭示的企业网络。欧洲系统性风险委员会工作论文第108号，2020年。URL https://www.econstor.eu/handle/ 10419/244260。

[29]Agam Shah、Abhinav Gullapalli、Ruchit Vithani、Michael Galarnyk 和 Sudheer Chava。FiNER-ORD：金融命名实体识别开放研究数据集。arXiv:2302.11157v2，2024年。URL https://arxiv.org/abs/2302.11157v2。[30] Ankur Sinha、Satishwar Kedas、Rishu Kumar 和 Pekka Malo。SEntFiN 1.0：财经新闻的实体感知情感分析。《美国信息科学与技术学会会刊》，73(9):1314–1335，2022年。doi：10.1002/asi.24634。URL https://doi.org/10.1002/ asi.24634。[31]Kirill Solovev。Fine news。数据集文档，版本 2，2026年。URL https:// huggingface.co/datasets/ksolovev/fine-news/tree/v2。[32]斯坦福自然语言处理组。Stanza 命名实体识别模型，2026年。URL https:// stanfordnlp.github.io/stanza/ner_models.html。模型文档，访问于2026年11月。[33]GDELT 项目。GDELT 数据与文档，2026年。URL https://www. gdeltproject.org/data.html。

语料库沿革与来源试点研究

要将历史出版商列表应用于英语 Fine News 记录，我们从元数据中提取主机名，必要时则改用 URL，并纳入所列媒体平台的点分隔子域名。复现包注明了上游修订版本，以及从 168 个数据分片中冻结选取的 14,807,287 个不同标识符。若要验证从全新的上游数据中进行筛选能否重现该快照，仍需逐行比较标识符和正文。

在四个固定版本的抽样框中（表 A1），回溯性试点研究先抽取一个文件，再从中抽取一个行组。我们将每个计数乘以其两项选择概率乘积的倒数进行扩展，并计算所得汉森–赫维茨总量之比，以估计出版商产出率；同时，在三个自由度下计算探索性的 95% 线性化 t 区间。清单文件保留了版本、抽取概率和观测计数。由于各版本的筛选和去重方式不同，这些总量统计的是物理行。若要估计登记册之外的独立报道、重叠情况或财经报道覆盖范围，则需要另行抽样。

仅抽取四次，Fine News 匹配总量的区间仍然足够宽（215 万至 1,867 万条物理记录），涵盖了 1,481 万个暂存标识符。这些标识符占固定版本中 219,192,987 条英文记录的 6.76%，不过具体选择尚未经过独立重放验证。Fine News、Infini-News、RefinedWeb 和 FineWeb-HQ 框架分别包含 168、49,748、5,534 和 9,246 个文件。抽取的四个 Fine News 文件分别包含 437、519、71 和 319 个行组。每个抽样的 Infini-News 文件包含一个行组，而每个 RefinedWeb 文件包含 175 个行组。FineWeb-HQ 文件分别包含 352、336、164 和 331 个行组。

在 11,260,839 个可用副本中，有 237,193 个是在发布后两至七天内抓取的。在 26,897 个具有观测文本哈希的重复规范化网址中，有 23,779 个包含多个正文哈希。仅凭较早的发布日期，无法确定某个修订版本何时出现。

历史选择。 原始筛选条件包括以下115个条目，但即使是同一出版商的子域名，其抓取历史也各不相同。在五个最大的观测主机中，nasdaq.com 的已分阶段占相应总数的9.0%，可用占11.6%；uk.reuters.com 分别为8.9%和

表 A1：四个源版本中的估计出版商产出率。在每个源版本中，我们进行四次独立的有放回文件—行组抽样，并将匹配英语物理记录的估计数量除以抽样框中英语物理记录的估计数量。

来源	英语定义	抽样英语	匹配数	估计值匹配数	产出率 [95%区间]
Fine News	英语分片路径	16,384	799	10,408,482	4.50% [2.36, 6.63]
Infini-News	原生英语拉丁文	37,632	2,575	32,025,275	6.84% [3.37, 10.32]
RefinedWeb	发布范围	4,000	20	4,842,250	0.50% [0.16, 0.84]
FineWeb-HQ	原生英语	4,000	25	18,887,267	0.69% [0.13, 1.25]

0.2%；in.reuters.com 分别为7.7%和0.9%；reuters.com 分别为6.5%和5.9%；business-standard.com 分别为5.7%和7.5%。
accountingtoday.com，afr.com，agriculture.com，agweb.com，barrons.com，bdlive.co.za，benzinga.com，bitcoinmagazine.com，bloomberg.com，bloombergquint.com，business-standard.com，businessdailyafrica.com，businessday.co.za，businessinsider.com，businessinsider.com.au，businessinsider.de，businessinsider.pk，businesslive.co.za，businesstimes.com.sg，businesstimesafrica.net，businessweek.com，businesswire.com，cfo.com，channelnewsasia.com，cityam.com，cnbc.com，coindesk.com，cointelegraph.com，dailyfx.com，dawn.com，economictimes.indiatimes.com，economist.com，efinancialnews.com，etf.com，etfdb.com，etftrends.com，etfworld.co.uk，fin24.com，finance.yahoo.com，financial-planning.com，financialexpress.com，financialpost.com，fnlondon.com，fool.co.uk，fool.com，fool.com.au，forbes.com，forexcrunch.com，fortune.cnn.com，fortune.com，freightwaves.com，ft.com，globenewswire.com，gurufocus.com，handelsblatt.com，ibtimes.co.in，ibtimes.co.uk，ibtimes.com，ibtimes.sg，institutionalinvestor.com，investing.com，investingnews.com，investmentweek.co.uk，japantimes.co.jp，kitco.com，koreabizwire.com，koreaherald.com，koreajoongangdaily.joins.com，law360.com，livemint.com，malaymail.com，marketwatch.com，mining.com，miningmonthly.com，moneycontrol.com，moneyweb.co.za，morningstar.com，morningstar.com.au，morningstar.in，morningstaronline.co.uk，nasdaq.com，nikkei.com，nikkeibp.co.jp，nikkeibp.com，offshore-energy.biz，oilprice.com，pionline.com，platts.com，prnewswire.co.uk，prnewswire.com，proactiveinvestors.co.uk，proactiveinvestors.com，proactiveinvestors.com.au，professionalpensions.com，qz.com，reuters.com，rigzone.com，rttnews.com，scmp.com，seekingalpha.com，spglobal.com，straitstimes.com，taipeitimes.com，theedgemalaysia.com，theedgemarkets.com，t heedgesingapore.com，thefinancialexpress.com.bd，themalaymailonline.com，thestreet.com，thinkadvisor.com，thisismoney.co.uk，upstreamonline.com，worldoil.com，wsj.com，zacks.com。

B 抽取器和元数据审计

精确的公司跨度。这项微基准测试从暂存快照中日期一致的规范英语文章里，按出版年份（2019 至 2023 年）选取种子化 ID 哈希值最低的 50 篇。对于每篇文章的导语，我们取前 150 个以空格分隔的词；若正文更短，则取全文。GPT-6.2-sol medium 主流程覆盖了全部 250 篇导语，包括 40 篇试点样本；另一个独立的 GPT-6.2-sol low 流程则覆盖了五个年份中排名前 12 的样本。两轮标注在重复的 60 篇导语中有 52 篇完全一致（跨度 F1 为 95.68%）。重复标注员可以查看已发布的试点示例，评分前依据指南解决了分歧。最终的模型参考集包含 250 篇导语中的 712 个跨度，其中 73 篇不含任何跨度。指南将具名企业纳入标注，并排除产品、裸股票代码和非商业组织。

Stanza 1.11.0 使用英语 OntoNotes CharLM 模型，并结合分词和多词扩展。GLiNER2.5 Small 使用已部署的公司模式，截止时间为 0.5 [24, 32]。评分按文章 ID 和精确的左闭右开 Uni-代码偏移量，匹配 Stanza ORG 或 GLiNER 公司跨度。因此，即使正确识别出的非公司 ORG，也会在

表A2：公司跨度精确微平均分数（%）。平衡数据集包含250条导语和712个模型参考跨度。

系统	TP	FP	FN	P	R	F1
Stanza ORG	421	835	291	33.52	59.13	42.78
GLiNER 公司	412	427	300	49.11	57.87	53.13

此任务中计为假阳性。表A2显示，在这项公司专属任务中，GLiNER优势为 10.34 个微平均F1百分点；按发表年份进行 2,000 次段落自助抽样后，配对95%区间为 $[5.84, 14.56]$ 。差异基于未四舍五入的分数计算。如果预测和参考标注只需共享一个字符，采用最大一对一匹配后，Stanza 的 F1值为 58.13%，GLiNER 为 58.54%。在按年份分层的 2,000 次重抽样中，配对差异为 0.41 个百分点（-3.36 至 $+4.40$）。两种评估均保留仅限商业实体的参考标注。斯坦福发布的 OntoNotes 88.8% 和 CoNLL03 92.1% 分数，均为其基准数据集划分上的全类别微平均F1 [32]。在本地任务中，附着于 Credit Suisse Group AG 公司跨度的所有格使 Stanza 未能与 UBS Group AG 配成一对；GLiNER 则在一条提及五只股票的导语中漏掉了美国银行，因此在 15 个可能的提及对中损失了五对。这些都是基于日期的发行人消歧之前的提及对。

文章、提及和窗口元数据。GLiNER2.5 Small 处理长度为384个词、重叠64个词的窗口。每个窗口都会获得一个针对七种文章形式的 softmax 分布，并针对六个事件族中的每一个获得独立分数。文章采用其各窗口中置信度最高的预测形式。事件族按窗口评估，截止值为0.5，因此一个窗口可能属于多个事件族，也可能不属于任何事件族。抽取器返回公司、人物、公共机构和地点跨度在正文中的偏移量。每个公司跨度都会分别获得局部相关性和面向投资者的情绪 softmax 分数。在文章–发行人对内，链接器取最高的实质相关性分数，并以提及链接分数为权重，对情绪向量求平均。下方的提及级审计评估的是发行人级聚合之前的局部标签。

seed-46确认排除了seed-45开发文章，并从每个可用语料框架中抽取400个样本单元。对于形式和情绪，半数样本按预测类别规模分配，另半数则在可用样本范围内平衡各类别。事件样本在所有分数均低于0.5的窗口与按最高类别分层的正向窗口之间均分。GPT-6 Luna低投入判定器会查看文章，以及局部上下文或事件窗口中标记的提及，但看不到预测结果。另一次Luna调用和一次GPT-6.1 Terra低投入调用分别对每个框架中相同的100个随机样本单元独立标注。估计值采用纳入概率倒数权重，并使用在层总量固定条件下的文章聚类乘数区间（表A3）。

使用可用文章—发行人链接时，新闻稿差异依然存在（-37.3 个百分点，95% 置信区间 $[-47.0, -27.6]$）。若将三篇上下文不足的文章计为新闻稿，加权基准流行率将从 4.18% 升至 4.73%。在加权提取提及中，7.4% 的情感标签呈现相反的积极和负面极性。Luna 重复判定的加权情感一致率为 82.3%，Luna 与 Terra 之间的加权情感一致率为 57.2%。在所有预测均低于阈值的抽样窗口中，主要评审员在 200 个窗口中的 160 个里至少发现了一种事件类别。另行进行的 10,000 篇文章抽取基准测试速度为每秒 106.1 篇文章，CUDA 显存峰值预留量为 7.22 GiB。合成示例为开发集提供指导，且对确认估计的权重贡献为零。

句子与公司情感。主要评审员为 400 条标注提及中的 358 条赋予积极、负面或中性情感。我们将 GLiNER 的公司标签与 Stanford CoreNLP 4.5.10 得出的所在句子的根节点情感进行比较，并合并每个方向上的两种强度

表A3：根据盲化模型判断进行的加权元数据验证。每个框架均包含400个抽样单位（全文文章、提取片段和输入窗口）。n统计未加权参考阳性项。精确率和召回率的文章聚类区间为95%。罕见参考类别的样本支持有限，包括3篇上下文不足的文章、9篇综述文章，以及7个与相关性有关的无效片段。破折号表示分母为空。

Task	类别	n	精确率 [95% 区间]	召回率 [95% 区间]
文章形式	分析	44	.40 [.27,.53]	.53 [.39,.67]
	上下文不足	3	–	.00 [.00,.00]
	访谈	15	.18 [.08,.32]	.31 [.14,.55]
	意见	42	.18 [.10,.28]	.34 [.20,.49]
	其他	13	.14 [.04,.27]	.02 [.01, .07]
	新闻稿	14	.08 [.04, .14]	.80 [.58, .94]
	报告	260	.86 [.77, .93]	.26 [.23,.28]
	汇总	9	.12 [.04,.26]	.16 [.05,.41]
局部相关性	附带的	103	.41 [.20,.65]	.08 [.04,.15]
	无效跨度	7	–	.00 [.00,.00]
	实质性的	283	.71 [.65,.77]	.67 [.61,.73]
	不确定	7	.00 [.00,.00]	.00 [.00,.00]
情绪	无效跨度	6	–	.00 [.00,.00]
	混合	21	.02 [.00,.09]	.01 [.00,.02]
	负面	92	.50 [.40,.60]	.56 [.47,.65]
	中性	157	.46 [.35,.58]	.19 [.15,.23]
	无法评估	15	.04 [.00,.11]	.06 [.01,.19]
	积极	109	.41 [.33,.49]	.73 [.66,.79]
事件类别	表现	150	.83 [.70,.93]	.21 [.17,.24]
	融资	55	.60 [.43,.77]	.17 [.11,.25]
	交易	60	.65 [.49,.79]	.24 [.16,.33]
	运营	96	.66 [.49,.80]	.09 [.06,.12]
	法律与声誉	93	.81 [.66,.92]	.02 [.02,.03]
	市场与宏观经济学	126	.87 [.75,.95]	.10 [.08,.12]

。混合及无法评估的预测均未能命中判定类别。有一个斯坦福句子超过了 200 个词元的解析上限。加权三分类宏平均 F1 为：GLiNER 46.79%，斯坦福 40.84%，配对差异为 5.95 个百分点 $[-1.33, +13.98]$。在按方向判定的单元中，极性反转的比例分别为 8.38% 和 10.73%。新一轮 GPT-6.1 Terra 评估与主要评审员在 40 个盲测案例中的 10 个上意见一致。它将 23 个目标标记为无效公司跨度，其中包括一个产品和一个政党。跨度是否有效以及周边文本的金融解读，都会影响这项比较。

C 发行人身份、重点与报道覆盖

seed-43 确认样本按别名频率和共享同一键的发行人数提取片段，并另行从至少包含一个提取出的公司提及片段的文章中抽取均匀样本。两种抽样均独立于链接器的预测。对于片段判定，模型会看到标记的文本和中性的发行人搜索结果，但候选链接会被隐藏；文章判定员则会收到最多 12,000 个正文字符。如果链接指向判定出的发行人，或指向文章日期当天同一上市证券，我们就将其计为正确；未链接精确率则按被判定为没有符合条件的发行人的弃答所占比例来度量。加权估计使用

表 A4：识别发行人为文章主题的精确率、召回率和 F1值（%）。217 篇开发集文章和 183 篇留出文章中，分别包含 80 和 65 组经判定的主体配对。每行展示应用于已接受的发行人关联的一条不同规则。位置和相关性两行分别显示单独使用这两种度量时，开发集上取得的最佳 F1值。

	开发集			留出集		
Rule	P	R	F1	P	R	F1
每个链接	27.9	83.8	41.9	31.9	92.3	47.4
至少两次提及	53.8	71.3	61.3	55.6	76.9	64.5
至少三次提及	78.8	65.0	71.2	66.1	56.9	61.2
至少三次提及	88.2	56.3	68.7	75.6	52.3	61.8
正文第一季度中首次提及	58.1	67.5	62.4	54.9	69.2	61.2
实质相关性 ≥ 0.40	48.1	63.8	54.8	41.7	53.8	47.0

线性化区间（表 A6）。Cohen 的 κ 在对 164 条提及进行重复判定时为 0.935，在由独立审计判定员对 339 条提及进行判定时为 0.881。

选择主体规则。 我们比较了提及次数、发行人首次提及的位置占正文长度的比例，以及其最高实质相关性分数。在由六个次数截止值、四个位置截止值和 19 个相关性截止值构成的网格中，三次提及且不附加其他限制的规则取得了最高的开发集 F1值。在留出文章上，计算两次提及比计算三次提及找回了更多主体，且 F1值更高（表 A4）。开发集选定的 0.40 相关性阈值在留出文章上得到 41.7% 的主体精确率和 53.8% 的召回率。

由于判定员只读取前 12,000 个字符，完整正文规则可用的一些证据可能落在判定文本之外。开发集正文中有七篇、留出集正文中有五篇超过了这一限制。在一篇 Schlumberger 财报电话会议文字记录中，Goldman Sachs、花旗集团和美国银行的第三次已关联提及都出现在第 20,000 个字符之后。完整正文规则选中了这三家银行，而判定员则根据可见的前缀将 Schlumberger 判定为主体。

与生产环境相比，前一年回放保留了464个发行人身份，变更了1个，并对先前已分配的10条提及弃判。生产环境中的411条弃判均依然存在。注册记录有效性描述的是记录的生效日期，而先前时间窗口则决定频次证据何时可以纳入决策。明确锚点可以直接消解一条提及。否则，领先候选项需要获得至少75%的证据支持，且单独的运营评分须达到0.5。对于含义模糊的提及，保留其候选列表。频次分层分为少于10次、10至999次以及至少1,000次，并与匹配发行人数量为零、一个或多个的情况交叉划分。

锚点审计从种子集 43 篇文章之外抽取明确的上市锚点样本。修饰名称审计则从规范化股票代码和股票形式后新接受的文章–发行人对中抽样。在300个新接受的配对中，297个带有后置股票代码，另有3个使用了“stock”一词。另行检查了6个交易所前缀案例。规范化新增了49,852个接受的配对，其中47,808个来自可用文章。身份一致并不意味着主题角色也已确定，后者需要单独评估。含义并非公司名称的短键，以及存在相互竞争的发行人候选项的子公司，都可能导致错误或弃判。带日期记录限制了有效关联，而全语料别名先验可以使用之后观察到的锚点。

在 2017 年抓取记录中，有 2,886,747 条是在发布后超过七天才获取的，另有 15 条的发布日期超出了未来日期容差（表 A7）。仅英国和印度的路透社媒体平台就分别占据了这些延迟抓取记录中的 1,297,451 条和 1,042,605 条。发行人面板根据文章的发表年份为其归类，并针对每个关联发行人各计数一次。某个发行人—年份组合若具有有日期的上市记录，或具有起始时间未知的当前上市记录，且在该年或前一年观察到公众持股量，则进入样本范围

发行人分配	精确率	召回率	F1
唯一带日期名称	99.12	42.47	59.46
WikiDict 和交叉对照表	99.78	20.42	33.91
前一年 CCFinance	96.66	66.18	78.57
完整 CCFinance	96.81	70.83	81.81

表 A5：886 条共同提取出的提及的加权发行人链接得分（%）。其中，578 条有经模型裁定的合格发行人。名称规则要求发行人唯一且带有日期。斯坦福的 WikiDict 为 464 条提及返回了页面，其中 97 条与带有 CIK 的 Wikidata 条目的英文维基百科规范站点链接相匹配，49 条得出了唯一且带有日期的发行人。前一年回放会在日历年开始前固定学习得到的别名证据，并保留冻结注册表。完整版本与前一年版本的 F1值相差 3.24 个百分点 [1.22，6.23]，这是在抽样层内重新加权的 2,000 次配对文章级乘数抽样中所得。

表 A6：链接与焦点诊断。分母是所述度量的评估预测或参考正例的未加权计数。跨度估计值使用逆抽样权重。文章估计值使用包含检测到的公司的 183 篇留出文章的均匀样本。锚点行和修饰名称行给出了不同盲化模型审计的原始比例。

度量	分母	估计值 [95% 区间]
抽取片段精确率	475 个已接受链接	96.8% [94.0,99.6]
抽取片段召回率	578 个判定为可链接的片段	70.8% [64.6,77.1]
未链接精确率	411 个未链接片段	84.8% [80.6,88.9]
多个候选项时的精确率	225 个已接受片段	92.0% [86.4,97.6]
文章召回率，所有提及的发行方	202 个参考配对	68.8% [61.0,76.7]
文章召回率，主题发行方	65 个参考配对	92.3% [86.3,98.3]
焦点精确率，每个链接	188 个预测配对中的 60 个	31.9% [24.7,39.1]
焦点精确率，三次提及	56 个预测配对	66.1% [54.8,77.3]
焦点召回率，三次提及	65 个参考配对	56.9% [44.3,69.5]
显式股票代码锚点	300 个抽样锚点	294/300 [95.70%，99.26%]
修饰名称增补	300 个抽样新配对	300/300 [98.74%,100%]

在具有有日期的上市记录，或具有起始时间未知的当前上市记录，且在该年或前一年观察到公众持股量的情况下。满足任一条件的发行人，即使在没有关联文章的年份，也会继续留在面板中。

D 正文嵌入与发行人条件化

生产环境中的 EmbeddingGemma 制品 [27] 为每篇正文存储 768 个 FP16 坐标，令牌上限为 2,048 个。在普查中，我们核对了所有标识符、位置和文本标记，并检查了向量范数及有限性。全部 14,807,287 篇冻结正文都包含文本，因此我们使用空白正文样例来检查零向量约定。重点筛选统计抓取正文中的所有提及；而在隔夜面板中，添加文档前缀和特殊标记后，分词器截断了 996 篇正文中的 44 篇。在 1,104 条文章–发行人记录中，有 6 条的第三个已关联提及位于编码器可见文本范围之外。

为了评估嵌入能否区分同一篇报道中的两家公司，发行人条件化试验要求模型参考集为每家公司识别出一个不同的账户。当两个以公司为条件的视图都更偏好各自的名称屏蔽账户时，即视为配对成功。在

表 A7：按捕获队列划分的报道覆盖。每列统计不同标识符的数量。「可用」指日期一致且具有规范形式。「已关联」和「重点」分别将可用标识符限定为至少有一个已接受的关联和至少有一个重点关联的标识符。暂存标识符即为队列分母。

捕获年份	已分阶段	已注明日期	可用	已关联	重点
2016	61,324	54,579	54,422	18,085	6,381
2017	3,659,035	772,273	770,001	302,957	129,627
2018	1,055,267	1,030,279	1,028,757	418,127	171,662
2019	1,584,644	1,252,003	1,250,744	537,153	229,016
2020	1,578,892	1,565,227	1,563,051	793,549	366,484
2021	1,588,551	1,577,085	1,571,857	765,957	362,237
2022	1,719,346	1,557,848	1,557,355	762,816	400,894
2023	1,420,564	1,378,531	1,375,864	680,572	357,587
2024	1,033,012	1,016,670	1,009,969	477,955	274,446
2025	852,474	834,880	830,745	364,949	197,499
2026	254,178	248,101	248,074	98,170	51,764

150 篇留出故事中，Luna 识别出 18 个符合条件的案例，其中 16 个适用于每种编码器；而 Terra 在审计中发现了 8 个符合条件的报道。由于两个分母都低于预定的最低数量 60 和 20，因此，通过选择发行方句子而获得的观察收益仍需在更大的样本上加以确认（表 A9）。

收购检索试点在抽取的80篇报道中发现38篇具有方向性的报道，并在19组独立交易配对中为每种条件进行了76次比较。主要增幅为 -6.6 个百分点，配对95%区间为 $[-18.4, +5.3]$。Terra在20篇经审计的报道中有16篇判断一致。上下文句子池化重复使用了25篇符合条件的报道，Gemma的得分为12/25，OpenSearch为11/25，均比各自的提取句基线低一例。在受控的角色互换测试中，两家公司共用句子掩码，两者的得分均为0/24。OpenSearch的提及池化得分为4/24。生产向量支持固定编码器实验，而这些发行方特定实验仍属探索性分析。

E 研究详情与敏感性分析

股票市场波动率。 2025年7月的手稿和档案v1提供了参考追踪器及规格说明 [1, 2]。六次档案重跑均以显示精度复现了月度拟合结果。为根据196个出版商-月份单元构建CCFinance追踪指标，我们将每家出版商的词典命中占比除以其在截至2022年12月的40个校准月份内的离散度，再对各出版商取平均，并重新缩放，使其均值与校准期内的VIX均值一致。随后，我们以截距项和同期追踪指标作为解释变量，对每个共同结果进行回归。对于每个10,000个自助法样本，我们以有放回抽样的方式抽取六个月的循环区块，直到获得49个观测值，并截去最后一个区块多出的部分。每次抽样中，两种追踪指标和共同结果使用相同的月份。所得平方相关系数差值的第2.5和第97.5百分位数构成配对区间。实现波动率 R^2的CCFinance值为0.336，存档追踪器值为0.569，差距为 -0.233 $[-0.311, +0.041]$。

表A10先在两种拟合中剔除2020年3月和4月，然后也在出版商离散度和 VIX 尺度的校准中剔除这两个月。纳入全部四家出版商时，47个月的 VIX 差距在两种校准下均四舍五入为 -0.041 。在所有设定中，剔除 Forbes 导致的缺口都是四种出版商剔除方案中最大的。仅使用《华尔街日报》时，49个月的 R^2数值分别为：VIX 为 0.110，已实现波动率为 0.120。档案中的月度基准提供了十一份报纸综合指标，因此，0.397 的相关系数比较的是单一出版商追踪器与

表A8：按发表年份和SEC年度公众持股量五分位划分、文章数量达到50篇的发行人-年份。括号内为行内百分比。公众持股量采用当前年份或前一年中的最新测量值。年度排名分配使合并五分位组的数量略有不均。在所有符合条件的发行人-年份中，9,338个有带日期的上市证据，另有25,443个凭当前上市记录（起始时间未知）和已观测公众持股量符合条件。

组别	符合条件	可用（%）	重点（%）
2017	3,240	854（26.4）	326（10.1）
2018	3,341	903（27.0）	358（10.7）
2019	3,429	1,222 (35.6)	593 (17.3)
2020	3,607	1,844 (51.1)	1,184 (32.8)
2021	3,950	1,923 (48.7)	1,209 (30.6)
2022	4,088	2,048 (50.1)	1,414 (34.6)
2023	4,176	2,038 (48.8)	1,475 (35.3)
2024	4,334	1,778 (41.0)	1,228 (28.3)
2025	4,616	1,308（28.3）	854（18.5）
最低	6,533	329（5.0）	105（1.6）
第二	6,530	889 (13.6)	311 (4.8)
第三	6,529	2,084 (31.9)	914 (14.0)
第四	6,527	3,860 (59.1)	2,161 (33.1)
最高值	6,526	5,790 (88.7)	4,561 (69.9)
未测量	2,136	966 (45.2)	589 (27.6)

表 A9：配对发行方—账户检索。每个单元格统计在相同的 16 篇符合条件且适配良好的报道中，两种公司视角都更倾向于各自的掩码账户的报道数量。

编码器	公司标题	选定指令	发行方句子
EmbeddingGemma-300M	2	2	10
OpenSearch neural embed v1	2	2	9
Qwen3-Embedding-0.6B	2	3	10

该综合指标。要求至少有15个发布日期会剔除三个数据稀疏的单元格，并得到 -0.142的全日历 VIX 差距。将校准期截至2021年12月，得到的差距为 -0.167。

银行共同提及。 2023年7月的 CEBRA 手稿定义了参考节点和时间窗口[14]。交叉对照表将带日期的发行人映射到其中34家银行里的33家，其中包括外国母公司代理项。每篇符合条件的文章中，每个无序映射配对最多计数一次。在匹配设计中，联合刊发构成连通的报道组件。固定配额在比较组和两项安慰剂变化中，每周从31个媒体平台保留682篇报道。从2020年1月至4月，匹配网络增加了56.3个不同配对 $[5,\,83]$ 和307.6个配对观测值 $[178,\,421]$。从2020年7月至10月，不同配对减少了9.8个 $[-46,\,28]$ ，同时观测值增加了45.7个 $[-117,\,217]$。从2021年1月至4月，不同配对减少了21.8个 $[-43,\,-1]$ ，同时观测值增加了202.0个 $[70,\,341]$。方括号给出1,000次匹配报道抽样中变化值的中间95%。

扩展日历要求在14周内保持相同的名单和出版商配额，并在32个符合条件的节点上，每周保留来自14个媒体平台的158篇报道。纳斯达克不在共同出版商支持集之内。由于合并其前身机构会改变网络单位，因此排除Truist。尽管2020年1月至4月的两个抽样区间都跨越零点，但在七个时期中，相连配对的平均增幅按带符号变化排名第一，仅比次高者多0.1对；而配对权重的增幅排名第二。这些区间度量的是匹配报道之间的变异，

表A10：CCFinance减去档案同期 R^2。每个47个月的拟合均省略2020年3月和4月。校准使用原始40个月，或剔除同两个月后剩余的38个月。每个由三家出版商组成的追踪器在完整日历中保留147个单元格，在受限日历中保留141个单元格。

	VIX			已实现波动性		
拟合月份	49	47	47	49	47	47
校准月份	40	40	38	40	40	38
全部四家出版商	−.132	−.041	−.041	−.233	−.051	−.047
省略Benzinga	−.098	−.067	−.075	−.115	−.010	−.014
省略 Forbes	−.360	−.233	−.240	−.382	−.072	−.072
省略 MarketWatch	−.149	+.004	+.003	−.284	−.050	−.045
省略《华尔街日报》	−.202	−.017	+.001	−.320	−.036	−.026

表 A11：14周共同支持集上的银行网络。均值和 95% 区间概述了匹配报道抽样的结果，每周包含 158 篇报道和 32 家符合条件的银行。配对权重汇总文章层面的共同提及。区间概述了周内抽样结果。

日历比较	不同配对数量的变化	配对权重的变化
2018年1月至4月	+11.3 [−13,30]	+26.9 [−36,89]
2018年7月至10月	−5.2 [−24,16]	−100.1 [−161,−42]
2019年1月至4月	+4.6 [−18,22]	+18.7 [−77,120]
2019年7月至10月	−2.5 [−41,23]	+11.1 [−85,99]
2020年1月至4月	+19.1 [−25,59]	+56.2 [−40,155]
2020年7月至10月	+19.0 [−13,52]	+89.8 [−14,199]
2021年1月至4月	−49.5 [−71,−29]	−13.8 [−99,67]

所选的我们eks，而六个安慰剂时期提供了历史对照（表A11） .

隔夜投资组合。 固定面板采用了 arXiv v6 和参考档案 v2 [17, 19]，中选定的设计，使用抓取的正文和冻结的雅虎图表价格，替代专有标题和 CRSP。时钟解析器恢复了 11,196 个首次发表标记。其中，2,738 个落在隔夜区间之外，274 个是在截止时间之后抓取的。在剩余的 8,184 篇文章中，5,392 篇没有固定名册关联。聚焦规则筛除了 2,510 个名册内关联。在剩余的 995 个关联中，五个未通过时段内相似度筛选，一个重复了近期标题，还有一个未通过价格变动标题筛选。其余 988 个通过了有日期的上市记录、向量和价格检查，并增加了 943 篇唯一正文和 667 个公司—会话。20 个滚动窗口中仍有相同的 20 个完成拟合，不过恢复面板改变了训练和评估行。允许 L-BFGS 最多迭代 1,000 次后，全部 16 条基线漂移警告均已消除。在原始逆正则化强度 $C = 1$ 下，738 个选定的证券–时段信号中有九个发生变化，剩下 39 个双腿时段，平均漂移为 -8.73 基点。更强的正则化（$C = 0.1$）改变了 221 个信号，剩下六个双腿时段，平均漂移为 -13.67 基点。两次重新拟合都评估了原来的 277 天。

在空头腿有交易的天数中，原始面板平均亏损 27.72 基点，恢复面板平均亏损 14.64 基点（表 A12）。在所有评估日期中，5 基点情景的均值分别为 -17.23 和 -18.79 基点。将两个面板都限制在原始的 277 个日期内，配对净变化为 -1.27 基点 $[-11.19, +9.25]$。

代表性 探针使用相同的996篇基线故事、发行人、价格和时间限制 -

表 A12：按投资组合支持度划分的 Gemma 正文漂移均值（基点）。每个面板使用各自的评估日期。净值表示在每个活跃腿的入场和出场时单向收取 5 个基点。空头腿的收益率为其证券收益率的负值。方括号给出基于已保存预测、从第一个到最后一个评估日期之间的交易所日历中抽取 10,000 个循环21交易日区块样本所得的 95%区间。

	原始面板			恢复面板		
支持度	Days	毛值 [95% 区间]	Net	Days	毛值 [95% 区间]	Net
任一活跃腿	154	−18.26 [−52.15, 7.36]	−30.98	266	−9.88 [−27.70, 7.73]	−24.02
双腿	42	−18.18 [−93.80, 30.97]	−38.18	110	−6.64 [−40.12, 22.53]	−26.64
多头腿	118	−5.50 [−35.83, 16.66]	−15.50	194	0.18 [−17.88, 17.08]	−9.82
空头腿	78	−27.72 [−65.95, 3.03]	−37.72	182	−14.64 [−34.61, 4.70]	−24.64

表A13：277 个评估会话中，毛值及情景净值的开盘至收盘漂移，以分配给每条活跃腿的单位名义本金的基点表示。成本列给出假设的单边成本。零腿交易时段计入所有均值。

表征	零腿交易时段	毛值	1 bp	5 bp	10 基点	25 基点
Gemma 正文	123	−10.15	−11.57	−17.23	−24.30	−45.53
Gemma 标题	124	−4.43	−5.76	−11.07	−17.72	−37.64
MiniLM 标题	125	−4.69	−6.03	−11.40	−18.12	−38.26

。Gemma标题减去Gemma正文的平均日漂移为 $+5.72$ 基点，配对的21交易时段区块区间为 $[-10.67, +21.39]$。MiniLM标题减去Gemma标题为 -0.26 基点 $[-18.32, +17.47]$。成本情景按假设的单向费率，对每条活跃腿按其入场名义本金的两倍收取费用，且不计借券费用。平均漂移涵盖全部277个评估交易时段，零腿交易时段的收益率和成本均为零。毛值均值为负，这意味着在这些情景下，盈亏平衡需要返利（表A13）。

F 访问与再现

Common Crawl 提供其抓取内容，并保留第三方权利 [6]。Infini-News 为衍生元数据规定 CC BY 4.0，并为批量文本规定学术研究使用条件[16]。Fine News 将其代码标注为 AGPL-3.0-only，并未为底层新闻报道指定通用许可证 [31]。CCFinance 对其项目代码采用 Apache-2.0 许可证，其选择器通过 DuckDB 读取 Fine News 的 Parquet 输出。Baker 和 Lopez-Lira–Tang 档案均列为 CC BY 4.0 [1, 19]。Factiva、RavenPack 和 CRSP 等专有输入数据具有单独的访问要求。

要重现图表，需要访问冻结输入数据和代码。审查期间，项目代码和制品存储库均为私有；2026年11月22日进行的匿名检查访问到了上游发布版本，但在暂存快照处需要身份验证。我们计划在审查后发布代码和衍生记录，并附带特定制品条款。历史出版商注册表、捕获标识符、副本判定、提及位置偏移量和发行人映射均可在不依赖文章正文的情况下单独记录。Vector 的再分发需要另行进行条款审查。冻结标签允许重新评分，而重新裁定则需要访问相应的模型。Codex 协助完成了手稿准备。

表A14：重现图表所需的输入。

图表	所需证据
语料库	捕获ID、来源框架、出版商注册表
报道覆盖	带日期的上市记录、公众持股量、副本和发行人判定
抽取	冻结预测结果、偏移量和参考判定
检索	正文向量和匿名发行方账户
波动性	出版商月度份额、存档追踪器、市场结果
银行配对	银行对照表和带日期的报道配对
隔夜漂移	发布与采集时钟、向量和证券价格