行动者、框架与论证：利用大语言模型对财经新闻中的气候话语进行跨数十年计算分析

Ruiran Su[1] **Janet B. Pierrehumbert**[1] **Markus Leippold**[2]

[1]牛津大学工程科学系，英国牛津[2]苏黎世大学金融系，瑞士苏黎世

{ruiran.su, janet.pierrehumbert}@eng.ox.ac.uk, markus.leippold@bf.uzh.ch

arXiv:2601.10142v1 [cs.CE] 15 Jan 2026

摘要

财经新闻媒体影响着价值数万亿美元的气候投资决策，但这一精英领域的话语仍未得到充分研究。我们利用行动者—框架—论证（AFA）流程，分析了道琼斯通讯社2000–2023年的气候相关文章，提取发言者、议题的框定方式以及所采用的论证。我们使用分解式验证框架，对照2,000篇人工标注文章验证提取结果，并评估其完整性、忠实度、连贯性和相关性。纵向分析揭示了结构性转变：2015年前的报道强调风险和监管负担；《巴黎协定》签署后，话语转向经济机遇与创新，金融机构则成为主导声音。在方法层面，我们为利用大语言模型开展纵向媒体分析提供了可复现的范式；在实质层面，我们揭示了金融精英如何在二十年间内化并重新诠释气候危机。

1 引言

财经新闻媒体是全球经济的神经系统，不仅为投资者提供信息，也塑造着价值数万亿美元的企业战略和资本配置决策（Shiller，2017）。因此，这一精英领域如何呈现气候变化影响重大，会左右气候风险究竟被框定为应尽量降低的成本，还是应把握的机遇。尽管一般新闻中的气候传播已得到广泛研究（Schmid-Petri等，2017a），但尽管财经新闻在塑造市场预期和影响政策辩论方面发挥着重要作用，这一高风险领域仍未得到充分研究。

本文通过对道琼斯气候新闻语料库进行纵向计算分析，弥补了这一空白。该语料库包含980,061篇气候相关文章（2000–2023年）。我们重点从一个经过分层且纳入不确定性因素的4,143篇文章样本中进行提取，该样本保留了完整数据集的时间多样性和主题多样性。我们超越主题模型和基于词典的方法，引入**施事者–框架–论元（AFA）**方法，利用大语言模型（大语言模型）识别谁在发言、气候问题如何被框定，以及提出了哪些论证。为确保该流程在跨越二十年的话语中保持可靠性，我们使用一个包含2,000篇文章的人工标注金标准对其进行验证，并引入**分解式验证框架（DVF）**，通过多评审者验证，从完整性、忠实性、连贯性和气候相关性等维度评估提取结果。

本研究围绕三个指导性研究问题展开：

- **RQ1（行动者）：** 在精英财经新闻中，哪些类型的金融和政策行动者主导着气候话语？他们的影响力随时间发生了怎样的变化？

- **研究问题2（框架）：** 2000–2023年间，气候变化框架发生了怎样的演变？哪些外部事件与框架的重大转变同时出现？

- **RQ3（论证）：** 采用了哪些论证策略和论证依据来支持气候立场？这些策略和依据在不同的行动者群体和时代之间有何差异？

我们的贡献主要有三点。第一，我们开发并验证了一套基于大语言模型的论证框架分析流程，用于纵向提取行动者、框架和论证，并以 DVF 为基础确保其稳健性和可审计性。第二，我们迄今为止最全面地描绘了精英财经新闻中的气候框架化，记录了不同时期、行动者群体和论证谱系之间的系统性差异。第三，我们揭示了财经新闻气候叙事的结构性转变：从 2010 年前对

成本与合规的强调，转向《巴黎协定》之后以机遇与创新为核心的框架化，这一转变与《巴黎协定》及后续政策转变等外部冲击相关。

通过将方法论创新与实质性洞见相结合，本研究确立了一种可复现的大规模话语分析范式。

2 相关研究

2.1 论证挖掘与话语分析

论证挖掘旨在从文本中提取论证结构——包括主张、证据和论证依据（Lawrence 和 Reed，2019年）。早期研究建立了用于组件检测和关系解析的端到端流水线，通常采用全局推理（Stab 和 Gurevych，2017a）。后续神经网络方法通过对非树结构进行建模，并对组件和链接进行联合解码，提高了稳健性（Liu 等人，2021）。话语解析领域的相关研究则提供了互补的连贯性表征，并已应用于论证质量评估和关系建模（Liu 等人，2020年；Prasad 等人，2019年）。

其他研究则拓展了任务范围（立场、证据检索、关系类型判定和质量评估）及应用领域，并采用了规模更大、更多样化的数据集（Hua 等人，2022年；Chen 等人，2021；Gleize 等人，2020年）。Transformer基线模型在片段／链接预测和跨领域迁移方面仍表现出色，而基于图的架构和话语信息增强模型有助于捕捉长程结构和论证连贯性（Ruiz-Dolz 等人，2021；Toledo 等人，2021）。

与我们的研究目标最为相关的，是那些大规模、跨时间考察论据结构的研究。然而，现有研究通常基于短时段语料库（论文、论坛、辩论）开展，且仅使用内在指标进行评估。新近证据表明，经过恰当提示的大语言模型在论证挖掘任务中能够达到或超越特定任务模型的表现，但人们仍担忧其忠实性、偏差与可复现性（Gorur 等人，2025年；Li 等人，2025年）。

2.2 媒体框架建构与气候传播

框架理论强调，议题的呈现方式会强烈影响公众解读和政策应对（Entman，1993）。早期研究依赖对小型语料库进行人工内容分析，研究对象通常来自美国或欧洲，并着重探讨政治极化以及气候变化报道中的媒体偏见（Nisbet，2009；Schmid-Petri 等，2017b；Boykoff 和 Boykoff，2004年）。近期研究将这一方向拓展至多国比较，显示美国、欧洲和中国媒体在气候辩论的框架呈现方式上仍存在持续分歧（Schmid-Petri 和 Arlt，2016a；Duan 和 Miller，2021）。

计算方法一直致力于扩大框架分析的规模。基于词典的方法（Card 等人，2015）和主题模型（Blei 等人，2003；Roberts 等人，2014年）已被应用于气候语料库，以识别风险、适应和责任等高层次主题（Schirmag 等人，2025年）。然而，这类方法往往缺乏细致度，难以深入揭示行动者如何随时间推移策略性地运用框架。计算框架分析的最新进展凸显了模型不仅要捕捉调用了哪些框架，还要捕捉支撑这些框架的论证策略（Demszky 等人，2019）。近期研究开始对新闻媒体中的叙事框架建构进行建模，捕捉气候话语中的冲突、解决方案以及实体角色（英雄、受害者、反派）（Frermann 等人，2023年；Otmakhova 和 Frermann，2025年），不过这些方法关注的是一般新闻，而非专业化的金融领域。

2.3 用于气候与金融话语的大语言模型

近期研究已开始将大语言模型应用于气候与金融文本，提出的方法和发现与我们的行动者—框架—论点（AFA）流程直接相关。例如，(Leippold，2023年) 探讨了 GPT-3 针对气候金融叙事开展的结构化访谈；(Jain 等人，2024年) 则展示了基于大语言模型的气候感知型投资决策支持工具。ClimateBERT (Webersinke 等人，2022年) 和 SusGen-GPT (Wu 等人，2024年) 等领域专用模型，为气候与可持续性文本提供了实用的嵌入和生成先验。其他研究则利用大语言模型估计围绕气候主题的公众舆论和定向影响 (Lee 等人，2024年；Islam 和 Goldwasser，2024年)，并分析大语言模型在包括气候议题在内的社会政治议题上的输出偏差和表征差异 (Chen 等人，2023年)。

2.4 我们的贡献

我们的研究以互补的方式推进了这三个研究方向。相较于论证挖掘

（§2.1），我们面向跨越数十年的财经新闻，开展纵向分析；将抽取任务分解为一个整合的行动者—框架—论证框架，联合分析谁在发声、议题如何被框定，以及采用了哪些论证；并将大语言模型抽取与分解式验证框架相结合，以审查完整性、忠实性、连贯性和相关性。相较于框架研究（§2.2），我们聚焦于财经新闻这一高风险领域，在此领域中，框架会影响资本配置；并分析不同行动者随时间推移采用的修辞策略。叙事框架模型能够捕捉故事层面的冲突与解决方案，而我们的方法则强调财经话语所特有的经济和监管维度。相较于大语言模型应用（§2.3），我们通过多评审验证，将抽取结果与人工标注进行核验，强调可复现性和忠实性，而非仅关注预测性能。

3 道琼斯气候新闻语料库

3.1 语料库构建

我们分两步构建道琼斯气候新闻语料库。首先，我们使用道琼斯智能标识符（DJID）筛选道琼斯通讯社（2000–2023年）的文章。DJID是一套专有主题分类体系，能够对财经新闻进行一致分类（道琼斯，2024年；ProQuest，2024年）。我们保留标注了气候相关DJID、涵盖三个领域的文章：（1）核心气候议题（例如，N/CO2 [二氧化碳] ）；（2）能源转型（例如，N/BFL [生物燃料]）；以及（3）受气候影响的行业（例如，N/AGR [农业]）。DJID及其描述的完整清单见附录A.1。

其次，为了量化依赖DJID代码可能引入的选择偏差，我们结合基于关键词的重新筛查和监督分类开展验证研究。结果摘要显示，DJID筛选具有较高的精确率（∼95%）和召回率（∼92%），这表明我们的气候语料库既准确，又具有广泛的代表性。完整的方法细节见附录A.2。

预处理包括标准化处理和近重复项去除。我们使用局部敏感哈希，移除了 ∼8.7% 的高度重叠文章，最终得到**980,061 篇具有完整元数据和时间戳的独特气候相关文章**，用于历时分析（见图4，附录A.5展示了道琼斯气候新闻语料库的时间分布，并标注了主要气候政策事件）。完整的预处理详情见附录A.4。

3.2 抽样方法

为了从气候语料库中构建一个规模可控且具有代表性的子集，我们实施了一个四阶段分层抽样程序。该设计确保4,143篇文章样本保留了整个98万篇文章语料库在时间和主题上的多样性，同时也增加了复杂论证文本的比例。

3.2.1 时间分层

我们根据气候金融话语的主要阶段，将语料库划分为四个时期：危机前（2000–2007年）、金融危机（2008–2012年）、危机后复苏期（2013–2018年）和气候金融激增期（2019–2023年）。在每个分层中 t，样本规模按照比例分配确定：

$$n_t = \left\lfloor n_{\text{total}} \times \frac{|C_t|}{|C|} \right\rfloor,$$

其中，$|C_t|$ 是分层 t 的规模，$|C|$ 是整个语料库的规模。

3.2.2 主题聚类

为保留主题覆盖，每篇文章均以标题和首段嵌入按 2:1 加权拼接的方式进行表示，嵌入由 SBERT `all-MiniLM-L6-v2` 模型计算得出（Reimers 和 Gurevych，2019年）。我们采用 Ward连接法进行凝聚式聚类（Ward，1963年），并在以下范围内通过最大化轮廓系数来选择最佳聚类数量 k_t（Rousseeuw，1987） $[5, \min(15,\ \lfloor n_t/20 \rfloor)]$。试点研究表明，$k < 5$ 会将不同主题混为一谈，而 $k > 15$ 则会产生冗余的微型聚类。

3.2.3 代表性文章选择

我们从每个聚类中应用最大边际相关性（MMR）（Carbonell 和 Goldstein，1998）选择代表性文章，以平衡中心性与多样性：

$$\text{MMR}(a_i) = \lambda \cdot \text{sim}_c(a_i) - (1-\lambda) \cdot \max_{a_j \in S} \text{sim}_s(a_i, a_j).$$

我们设定 $\lambda = 0.7$，并略微偏重中心性。针对 $\lambda \in [0.5, 0.9]$ 进行的敏感性分析显示，下游性能保持稳定（附录 B）。

时间层	原始	样本	%
危机前（2000–2007年）	5,441	155	0.8
金融危机（2008–2012年）	320,157	1,355	0.8
危机后（2013–2018年）	215,351	911	0.8
气候激增（2019–2023年）	439,112	1,722	0.8
总计	**980,061**	**4,143**	**0.8**

表1：原始语料库与策略性样本中各时间分层的文章分布。

3.2.4 主动学习扩充

为丰富样本中的高难度论证案例，我们引入了主动学习循环。我们将初步的 RoBERTa-base 论点检测器（在种子集上的 F1 分数= 79.5%）应用于更大的语料库（附录C）。优先处理预测熵高于第 90 百分位的文章。根据熵分布的詹森–香农散度（<0.05，迭代间）衡量，经过四轮后，迭代抽样趋于收敛。

3.2.5 最终样本

最终的4,143篇文章样本与原始语料库高度一致：跨时间和主题分布的詹森–香农散度为 <0.1，样本质心与总体质心之间的余弦相似度为 >0.85；与2万篇随机样本相比，下游论证提取性能仅有极小幅度的下降（−3.3 个百分点）。

4 行动者–框架–论证（AFA）提取流程

我们设计了一个模块化流程，用于从财经新闻中提取行动者–框架–论证（AFA）结构。其目标是以可扩展且可复现的形式，呈现谁在发言、气候变化如何被框架化，以及采用了哪些论证策略。该流程按顺序执行；每个阶段都以前一阶段的结果为基础，从而确保行动者、框架和主张之间的一致性。图1展示了流程架构。

行动者—立场识别。 第一阶段识别行动者及其对气候议题表达的立场。每条提取结果都会注明行动者名称、类型（公司、金融机构、政府、非政府组织／倡议团体、个人）和表达的立场。此外，还会标明立场（支持、反对、中立或混合）。为确保归因准确，每条提取结果都关联到文章中的直接证据引文。

框架分类。 第二阶段会从我们的八框架类型学中指定一个主要框架，并可选择指定一个次要框架。

八框架类型学借鉴了气候传播研究中的既有框架（Nisbet，2009；Schmid-Petri 和 Arlt，2016b），并将其调整用于金融话语。主要调整包括：（1）将笼统的“经济发展”拆分为经济风险与经济机遇，以捕捉金融语境中的关键区别；（2）增加市场动态，以捕捉财经新闻中特有的竞争定位；（3）将“道德/伦理”细化为社会责任，以反映企业框架惯例。该类型学源于对500篇文章进行的试点编码。与源框架的完整映射见附录E。

论证提取。最后，模型将文章中的论证分解为各个组成部分：核心主张、支持性证据，以及论证依据，用于连接主张和证据。

4.1 模型与提示

尽管 Climate-BERT 等领域专用编码器（Webersinke 等，2022年）或 SusGen-GPT 等数据中心型模型（Wu 等，2024年为气候文本提供了颇具吸引力的预训练先验，但我们的重点是提取稳健性，以及在异质的金融新闻风格中实现纵向泛化。

因此，我们采用闭源与开源配对的提取策略。主要提取器为 Gemini-2.5-flash（Comanici 等，2025年），它具备出色的指令遵循可靠性和效率。为确保研究发现并非来自单一专有模型，我们使用同一组提示，在多种开放权重模型之一 LLaMA-4 Maverick-17B（MetaAI，2025年）上进行基准测试。这种配对兼顾了性能（闭源模型）与可复现性和透明度（开源模型），遵循近期关于 LLM 作为标注者流程的研究最佳实践（Tan 等，2024年；Törnberg，2024年；Pavlovic 和 Poesio，2024年）。

尽管存在多个开放权重模型，但为确保清晰度，我们在主要分析中聚焦于一个具有代表性的基线模型。所有提示词均采用结构化的思维链设计，并在附录F 中发布，以确保可复现性。

4.2 人工标注的黄金标准

为建立可靠的评估基准，我们构建了一个由以下文章组成的 2,000 篇文章黄金标准集

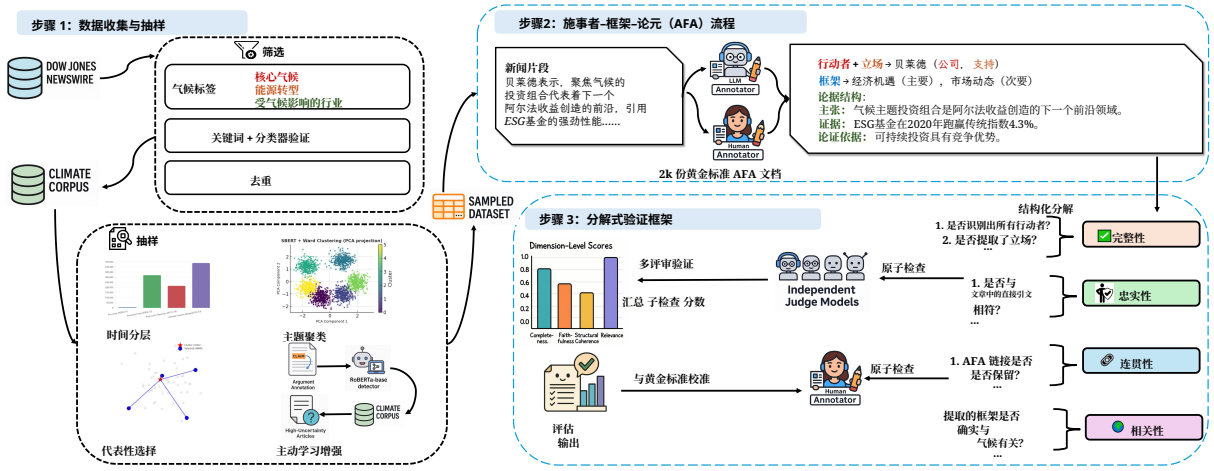


图1：本研究的端到端流程。步骤1：从道琼斯气候语料库进行数据收集和抽样，包括分层抽样、聚类和主动学习增强。步骤2：施事者–框架–论元（AFA）抽取流程，用于识别行动者、框架和论元结构。步骤3：分解式验证框架，将多评审评估与人工标注金标准验证相结合。

来自我们的语料库。每个标注单元都是一篇完整文章（平均=1,087 个词元，中位数为=982 个词元）。“新闻片段”（见图1 ）仅用于示例；实际处理使用完整文章。文章在 Zooniverse 公民科学平台上进行标注（见附录 G），每个项目至少获得五份独立标注。对于分类任务（立场、框架），最终标签通过多数投票确定。对于基于跨度的任务（行动者、论证），我们采用两步共识流程：首先计算标注员之间的词元级一致性分数，保留获得多数支持的词元，然后将连续的多数同意词元合并为统一片段（见附录H.6）。这种聚合既保留了短提及的边界精确率，又避免了较长论元片段中过度碎片化。

我们采用适用于多标注者场景的指标计算标注者间一致性（IAA）：对行动者跨度进行成对 F_1（Stab 和 Gurevych，2017b），立场采用 Krippendorff α（Krippendorff，2004年），框架采用 Krippendorff α，论点提取采用跨度级 F_1（Stab 和 Gurevych，2014年；Habernal 等，2018年）。表2 汇总了观察到的标注者间一致性分数，均处于“高度一致”范围（>0.7）（Landis 和 Koch，1977年），证实了公民科学标注方案的可靠性。

4.3 分解式验证框架

为确保评估可靠且可审计，我们引入一种**分解式验证框架（DVF）**，突破单一整体

组件	指标	IAA
施事者识别	成对 F1 分数	0.81
立场分类	Krippendorff 的 α	0.76
框架分配	Krippendorff 的 α	0.72
主张抽取	片段级 F1（宏平均）	0.74

表 2：Zooniverse 公民科学协议下各 AFA 组件之间的标注者间一致性。

分数的局限。DVF 不再要求评审对抽取结果进行整体评分，而是将评估拆解为细粒度的子项检查，在多个模型系列中对其进行验证，并以人工金标准标注作为基准。

结构化分解。 DVF 从四个维度评估抽取结果：（1）完整性（捕获所有组件），（2）忠实性（与源文本保持一致），（3）连贯性（模式和行动者—框架—论元链接），以及（4）相关性（领域特异性）。完整的子检查项见附录 I.2。

多评审验证。 为减轻自我评分偏差，DVF采用了多种不同的评审：GPT-4o（OpenAI，2024年）、Claude-Sonnet-4（An-thropic，2024年）以及两个开放权重评估器（Qwen3-30B A3B（Qwen，2025年）、Mixtral-8×22B（Mistral，2024年））。评审对各项子检查进行评分，再将评分汇总为维度级分数。为校准这些自动化分数，我们创建了一组人工评估数据，其中包含随机抽取的500条大语言模型输出。这些输出由不同于AFA金标准

标注员的编码员依据DVF评估标准进行标注。这组独立的评估数据可用于验证自动化DVF分数，确保评审的可靠性以人工评估为依据，同时避免循环论证。

5 结果与讨论

我们分四个部分呈现研究结果。首先，我们验证 AFA 抽取流程的可靠性（§5.1）。其次，我们使用五类别架构分析行动者显著性的长期趋势（§5.2，RQ1）。第三，我们记录过去二十年间框架策略的结构性转变（§5.3，研究问题2）。最后，我们考察行动者如何将框架与论证策略相结合，从而揭示出不同的修辞特征（§ 5.4，RQ3）。附录J.4提供了对抽样选取的 150 个错误预测所进行的定性错误分析。

5.1 流程验证

5.1.1 基于大语言模型抽取的理由

一个关键问题是：当微调预训练语言模型（PLM）在特定任务上也能达到类似性能时，大语言模型的计算成本是否合理。我们检验两个假设：（H1）大语言模型在复杂的多类别任务上性能更优；（H2）通过上下文学习利用序列依赖，将提升性能。

实验设置 我们将包含 2,000 篇文章的金标准划分为训练集（1,400 篇）、验证集（300 篇）和测试集（300 篇）。我们将三个 RoBERTa-Large 基线模型（Zhuang 等人，2021）与我们的零样本大语言模型流程进行了比较：

独立 RoBERTa： 每个 AFA 组件都作为独立模型进行训练，仅接收原始文章文本：(1) 用于识别行动者的词元级命名实体识别，(2) 用于识别立场的序列分类器，(3) 用于识别框架的多标签分类器，以及 (4) 用于识别论点的跨度提取器。

顺序 RoBERTa（金标准）： 下游组件将上游组件的金标准标注作为额外特征（例如，框架分类器接收与文章文本拼接的金标准行动者信息）。

所有 RoBERTa 模型均使用相同的超参数。训练详情见附录 K。

结果 表 3显示，大语言模型在所有组件上的表现均优于经过微调的 RoBERTa-Large，其中框架分类（+15.2 F1 分数）和论证提取（+17.2 F1 分数）也有显著差距，这支持 H1。

模型	行动者	立场	框架	Arg.
独立 RoBERTa	.762	.718	.504	.563
+金标准上游	.762	.742	.631	.595
Gemini-2.5	**.819**	**.791**	**.783**	**.767**
Δ (大语言模型 – 最佳 RoBERTa)	+5.7	+4.9	+15.2	+17.2

表3：在包含300篇文章的测试集上的性能比较（宏平均 F1）。

例如，RoBERTa将“可再生能源基金的表现优于传统投资组合4.3%”分类为技术解决方案（由“可再生能源”触发），而正确的框架应为经济机遇（聚焦于财务回报）。大语言模型通过更长范围的语境推理，能够正确识别论证焦点。

RoBERTa在复杂论证结构中的长距离依赖方面也表现不佳。我们发现，它要么在分句边界处过早截断主张，要么过度延伸至支持性证据，这表明其训练数据不足以学习细微的论证边界。

序列依赖与错误传播表3 揭示了支持H2的关键发现：序列条件化有益：金标准上游特征显著提升了RoBERTa的性能（+12.7 用于框架，+3.2 用于论证），证实AFA组件并非彼此独立：框架依赖于行动者，而论证依赖于二者。但独立RoBERTa分类器无法自然地实现这一点。

启示 总而言之，对于我们的应用场景（跨越数十年、抽取多个组件且人工标注有限），大语言模型能够提供更优的成本效益权衡。它们在以下方面表现出色：（1）需要语义区分的复杂多类别任务；（2）通过上下文学习进行序列推理，而非依赖脆弱的特征工程。

5.1.2 基于大语言模型的性能

为验证基于大语言模型的抽取流程的可靠性，我们使用通过 Zooniverse 标注创建的 2,000 篇文章金标准来评估模型输出。

表4报告了 Gemini-2.5-flash（主要提取器）和 LLaMA-4-Maverick-17B 在全部三个 AFA 组件上的系统性能

组件	Gemini-2.5	LLaMA-4
行动者–立场识别		
行动者类型	**0.847**	0.823
立场	**0.791**	0.768
总体 F1	**0.819**	0.796
框架分类		
主要框架	**0.783**	0.761
次要框架	**0.692**	0.671
论证提取		
主张	**0.806**	0.782
证据	**0.774**	0.753
论证依据	**0.721**	0.698
总体 F1	**0.767**	0.744
*DVF*汇总分数		
完整性	**0.831**	0.809
忠实性	**0.887**	0.863
连贯性	**0.792**	0.771
相关性	**0.856**	0.834

表4：相对于2,000篇文章黄金标准的流水线性能（宏平均F1）。Gemini-2.5-flash作为主要提取器；LLaMA-4-Maverick-17B提供开放权重基线性能。加粗值表示每个组件的最佳性能。

（开放权重基线）。在各个组件上，Gemini 的表现始终比 LLaMA 高约两到三个百分点。行动者类型识别的性能最高（$F_1 = 0.847$），次要框架分类的性能最低（$F_1 = 0.692$），这反映了次要框架标签固有的模糊性和稀疏性。与通常对应段落主导叙事焦点的主要框架不同，次要框架往往捕捉边缘或重叠的解读线索（例如，监管框架与技术框架同时出现），因此更难进行自动预测。

除了内在F1之外，我们还使用DVF汇总分数对提取结果进行验证，这些分数从四个维度确认质量。忠实性尤其始终保持较高水平，表明提取出的组件均有充分依据，扎根于源文本之中。为确保透明度，我们使用500个样本的人工评估集对DVF分数进行验证。各评审的详细分解结果和人工验证统计数据见附录 I.2；为保持清晰，正文仅呈现汇总结果。

5.2 行动者随时间的显著性

为回答**RQ1（行动者）**，我们考察不同类别的行动者如何参与财经新闻中的气候话语。我们采用五类别架构（公司、金融机构、政府、非政府组织/倡导组织、个人），追踪各类行动者在不同时间分层中的长期变化，

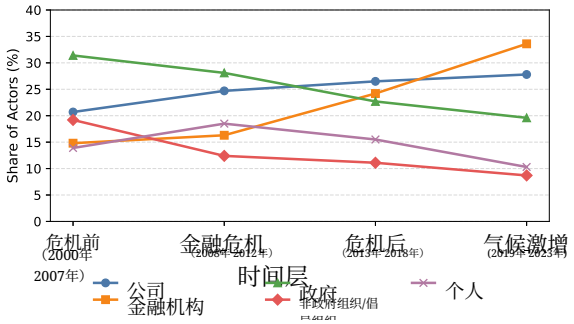


图2：行动者随时间的显著性，展示四个时间分层中五类行动者的比例分布。

包括四个时间分层（2000–2007年、2008–2012年、2013–2018年、2019–2023年）中的行动者显著性变化。图2 展示了行动者随时间变化的分布。

我们观察到话语权出现了明显的结构性转变（$\chi^2 = 217.3$，$p < 0.001$）。在21世纪初，政府（≈ 31%）和公司（≈ 21%）主导了与气候相关的金融报道，这与以监管和倡议为驱动的气候变化框架相一致。例如，一篇2006年的文章引用了世界自然基金会的警告：“监管不作为可能导致高碳基础设施被锁定。”在金融危机期间，个人所占提及比例上升至 ≈ 18 %，而政府在相对存在度方面开始下降（≈ 28 %），反映出报道转向市场导向的声音。

在危机后时期，随着《巴黎协定》的签署，金融机构的占比从危机前的≈ 15%飙升至2015年后的25%以上，并在近些年增长至 ≈ 34%，成为占比最大的单一行动者群体。资产管理公司和银行越来越多地将自己定位为气候领域的领导者。例如，一篇2021年的道琼斯文章引用贝莱德的话称：“可持续金融如今已成为阿尔法收益创造的核心。”

与此同时，在气候议题热潮期间，非政府组织的占比降至 10% 以下，而政府则稳定在 20% 左右，往往提供背景监管语境，而非引领叙事。

这些发现表明，话语权力正在重新平衡：从由政府主导的体制转向金融市场主导体制，在这一体制中，金融机构和公司成为精英气候话语中的主导声音。具体比例见附录 J.1。

5.3 框架中的叙事转变 s

为回答 **研究问题2（框架）**，我们分析精英金融新闻中气候问题框架随时间发生的变化。我们追踪此前提出的八框架类型学（经济风险、法规遵从、经济机遇、技术解决方案、市场动态、环境紧迫性、社会责任、不确定性/怀疑）在四个时间分层中的表现，并检验其分布是否随时间发生显著变化。

从风险/合规转向机遇/创新。图 3直观展示了这一明显趋势：以风险和合规为导向的报道逐渐减少，而以机遇和创新为导向的报道则不断增加。变点分析（PELT）识别出 **2015年第四季度至2016年第二季度**出现了具有统计显著性的结构突变，这与《巴黎协定》及相关政策公告的发布时间相吻合。此后，经济机遇和技术解决方案变得日益突出。

统计证据。我们观察到，在 PELT 分析识别出的 2015—2016 年变点前后（见 §5.3），经济风险与经济机遇框架的相对权重发生了逆转。后期机遇框架显著增加，而风险框架则有所下降；技术解决方案也有所增加，且常与机遇框架同时出现（呈正相关），这与前瞻性投资叙事相符。各时期的具体百分比和显著性检验结果见附录J.2。

对比示例。早期时期的文章往往强调合规成本或风险敞口（例如，强制性调整、负债），而2015年后的报道则日益突出增长和竞争定位（例如，绿色金融管线、创新领导力）。以下列出简短示例（为简洁起见，已作释义）：

- **风险（2015年前）：**“新的碳规则将推高运营成本并挤压利润率。”
- **机遇（2015年后）：**“气候主题投资组合和清洁技术投资将开辟新的阿尔法收益与市场领导力途径。”

财经新闻中对气候的框架出现了明显且有统计依据的转变：叙事从风险和监管负担转向机遇和技术创新，而且

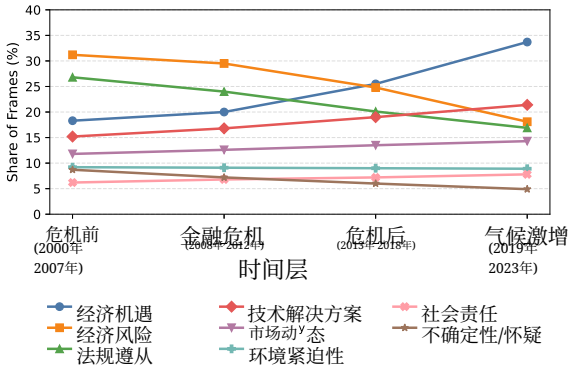


图3：四个层面中气候框架的时间动态。

这一转折点集中在2015—2016年。这种重新框定与该行业转向绿色金融和以投资为主导的理由相一致。

5.4 行动者—框架—论证对齐（RQ3）

为回答**RQ3（论证）**，我们分析行动者群体如何将框架与论证结合起来构建气候叙事。除了总体变化之外，这项分析还揭示了不同的修辞 repertories，展现了不同的行动者在精英财经新闻中采用的表达方式。

行动者—框架关联。 表17 报告了独立性检验的标准化残差 χ^2，反映各行动者类别对不同框架的代表程度是高于还是低于预期。公司和金融机构明显偏好经济机遇框架，而非政府组织则格外强调环境紧迫性。政府和监管机构对风险和法规遵从框架的使用更为均衡，体现了其制度角色。

行动者的论证策略。 论证依据分析揭示了不同的论证逻辑：

- **公司**常采用竞争优势和创新机遇论证依据，将气候行动视为业务增长战略。

- **金融机构**强调市场领导地位和风险调整后收益，将气候投资与受托责任联系起来。

- **政府/监管机构**高度依赖监管必要性和合规论证依据。
- **非政府组织／倡导团体**强调环境紧迫性和伦理责任。

•**个人（研究人员/专家）**强调科学证据并进行风险沟通，通常强化非政府组织的框架，但以技术依据为支撑。

论证复杂度。我们以每项主张所含前提数量的中位数来衡量论述的详尽程度。机会导向型论证更为复杂（前提数量中位数为 2.8–3.2），而风险论证则更为简洁（前提数量中位数为 1.4–1.7），这表明前瞻性框架需要更多论证投入。

示例说明。

• **非政府组织（2010年）**："政府必须立即采取行动；若不加遏制，排放将加速生态崩溃。"（框架：环境紧迫性）。

• **公司（2018年）**："投资清洁技术能够确保公司在全球市场的长期竞争力。"（框架：经济机遇）。

•**金融机构（2021）**："聚焦气候的投资组合代表着阿尔法收益创造的下一个前沿领域。"（框架：经济机遇）。

行动者群体不仅偏好不同的框架，还采用各具特色的论证策略。行动者、框架与论证之间的这种对应关系，展现了叙事如何发生转变：金融机构和公司日益将自身定位为气候转型中的市场领导者，而非政府组织强调紧迫性，政府则聚焦合规。总体而言，这些模式揭示了金融气候话语中一个既碎片化又结构化的修辞场域。

这些金融新闻研究发现与金融领域之外关于气候话语的基于大语言模型的研究相一致。例如，Lee 等人（Lee 等人，2024年）表明，模型推导出的框架信号与全球变暖相关的公众舆论趋势相关，这说明自动化框架提取能够反映更广泛的社会叙事。这种跨领域一致性增强了我们关于金融媒体中气候议题框架重塑的推论的外部效度。

6 结论

我们提出了一种用于纵向话语分析的行动者—框架—论证（AFA）流程，该流程将基于大语言模型的提取与分解式验证框架相结合。将这一方法应用于跨越二十年的气候相关金融新闻语料，既展现了方法上的稳健性，也带来了实质性的洞见。

我们记录了精英金融报道中的结构性转变：从早期强调经济风险和法规遵从，转向2015年后以机遇和技术解决方案为核心的话语。我们还表明，不同行动者采用了各异的修辞策略：企业和金融机构强调竞争优势与风险调整后收益，政府着重于合规和政策工具，非政府组织则突出环境紧迫性，呈现出一个模式鲜明但又多元异质的气候金融叙事领域。

在方法层面，我们的流程提供了一套可复现的方案，将行动者识别、框架分析和论证挖掘整合起来，并辅以可审计的评估。研究结果凸显了金融媒体如何塑造人们对气候转型速度和方向的预期。

本研究将自然语言处理领域的进展与气候传播和金融领域的核心问题相结合，展示了计算社会科学如何揭示引导资本和政策的叙事。

局限性

我们的研究基于单一的英文新闻来源（道琼斯通讯社）。未来研究应拓展新闻来源，以评估研究发现的普遍性。尽管我们验证了基于大语言模型的标注，但大语言模型本身存在偏差。我们的验证能够减轻这种偏差，但无法将其彻底消除。近期研究指出，大语言模型可能会再现或放大气候及社会政治传播中的偏差（Chen 等人，2023年；Islam 和 Goldwasser，2024年）。我们的分解式验证框架通过多评审验证和忠实性检查降低了部分风险，但全面的公平性审计和子群体稳健性测试仍是重要的未来研究方向，尤其是对于可能影响投资或政策决策的应用。最后，我们的研究绘制了话语图景，但没有直接衡量其对资本流动的现实影响；这是未来研究的一项关键方向。未来研究应将覆盖范围扩展到单一通讯社和单一语言之外，将话语变化与可衡量的市场结果和政策周期联系起来，并探索用于监测新兴

叙事转向的实时监测
工具。

伦理声明

我们的分析不涉及私人信息或个人身份信息。我们承认大模型训练和推理所带来的环境成本。近期分析表明，硬件效率和软件优化的提升已开始降低许多工作负载的边际碳强度（Patterson 等，2022年），但这些成本仍不可忽视。为减少我们的环境足迹，我们采用了分层抽样、高效提示和多阶段提取，而不是对大模型进行端到端重新训练。我们报告这些操作选择，以促进可复现性，并鼓励开展具有能源意识的复现研究。未来的工作还应进一步探索轻量级领域适配器和检索增强方法，以进一步减少计算资源消耗和排放。我们还承认，大语言模型可能反映其训练数据中的偏差；我们针对人工标注员开展的验证是一项关键检查。所介绍的方法可用于监测“漂绿”，但也可能被用于编造更具说服力的虚假信息。我们本着开放学术探究的精神发布研究结果，相信透明地理解媒体叙事所带来的益处超过了被滥用的风险。

致谢

Ruiran Su 获得怡和基金奖学金的支持。我们衷心感谢 Zooniverse 平台志愿者的贡献，他们的投入使金标准数据集的构建成为可能。我们还感谢道琼斯为本研究提供新闻语料库的访问权限。此外，我们感谢使用牛津大学高级研究计算（ARC）设施为本研究提供计算资源。最后，我们感谢匿名审稿人的建设性反馈，这些反馈帮助改进了本文。

参考文献

Anthropic。2024年。Claude Sonnet 4。模型文档／系统卡。Claude Sonnet 4，专有模型。

David M. Blei、Andrew Y. Ng 和 Michael I. Jordan。2003年。潜在狄利克雷分配。机器学习研究期刊，3：993–1022。

Maxwell T. Boykoff 和 Jules M. Boykoff。2004年。作为偏差的平衡：全球变暖与美国主流报刊。全球环境变化，14(2)：125–136。

Jaime Carbonell 和 Jade Goldstein。1998年。使用 MMR（一种基于多样性的重排序方法）对文档重新排序并生成摘要。载于第 *21* 届年度国际 *ACM SIGIR* 信息检索研究与开发会议论文集，第 335–336 页，澳大利亚墨尔本。

Dallas Card、Amber E. Boydstun、Justin H. Gross、Philip Resnik 和 Noah A. Smith。2015年。媒体框架语料库：跨议题框架标注。载于第*53*届计算语言学协会年会（*ACL*）暨第*7*届自然语言处理国际联合会议（*IJCNLP*）论文集，第438–444页。计算语言学协会。

Kaiping Chen、Ruitong Duan、Ying Peng 和 Jing-wen Zhang。2023年。GPT-3 如何回应不同公众对气候变化和黑人的命也是命的看法：对对话式人工智能公平性的批判性评估。*arXiv* 预印本 *arXiv:2209.13627*。预印本，最后修订于2023年3月。

Zheng Chen、Yulan He 和 Yu Zhang。2021。Hitting Your Marq：多模态论证质量评估。载于*2021*年自然语言处理实证方法会议（*EMNLP*）论文集，第6958–6970页。计算语言学协会。

Gheorghe Comanici、Eric Bieber、Mike Schaekermann、Ice Pasupat、Noveen Sachdeva、Inderjit Dhillon、Marcel Blistein 等。2025年。Gemini 2.5：以先进推理、多模态、长上下文和新一代智能体能力拓展前沿。技术报告，谷歌 / DeepMind。预印本 / 技术报告，包含 Gemini 2.5 Flash 模型。

Dorottya Demszky、Nikhil Garg、Rob Voigt、James Zou、Jesse Shapiro、Matthew Gentzkow 和 Dan Jurafsky。2019年。分析社交媒体中的极化：方法及其在 21 起大规模枪击事件推文中的应用。载于*2019*年北美计算语言学协会分会会议论文集：人类语言技术，第 *1* 卷（长篇与短篇论文），第 2970–3005 页，明尼阿波利斯，明尼苏达州。计算语言学协会。

道琼斯。2024年。道琼斯智能标识符（DJID）分类体系 API。财经新闻内容的专有主题分类系统。

Ran Duan 和 Serena Miller。2021。中国的气候变化：党派支持型报纸与市场化报纸新闻多样性研究。新闻学，22(10)：2493–2510。

Robert M. Entman。1993。框架：澄清一个支离破碎的范式。《传播学杂志》，43(4):51–58。

Lea Frermann、Jiatong Li、Shima Khanehzar 和 Gosia Mikolajczak。2023年。冲突、反派与解决方案：迈向叙事媒体框架模型。载于 第*61*届计算语言学协会年会论文集（第*1*卷：长篇论文），第8712–8732页，多伦多，加拿大。计算语言学协会。Martin Gleize、Eyal Shnarch、Ran Levy、Ben Bogin、Ranit Aharonov 和 Noam Slonim。2020年。高效的论证质量成对标注。载于 第*58*届计算语言学协会年会论文集（*ACL*），第5772–5781页。计算语言学协会。

Deniz Gorur、Antonio Rago 和 Francesca Toni。2025年。大语言模型能否执行基于关系的论证挖掘？载于第*31*届计算语言学国际会议论文集，第8518–8534页，阿布扎比，阿联酋。计算语言学协会。

Ivan Habernal、Henning Wachsmuth、Iryna Gurevych 和 Benno Stein。2018年。论证推理理解任务：隐含论据的识别与重构。载于 *NAACL-HLT 2018*年论文集，第1930–1940页。

Xinyu Hua、Zichao Yang 和 Yiming Yang。2022年。IAM：一个用于集成论证挖掘的综合性大规模数据集。载于计算语言学协会第*60*届年会论文集（第*1*卷：长篇论文），第2052–2064页。计算语言学协会。

Tunazzina Islam 和 Dan Goldwasser。2024年。基于大语言模型对社交媒体广告中的气候微定向进行事后研究：主题洞见与公平性评估。*arXiv*预印本*arXiv:2410.05401*。预印本。

Ayush Jain、Manikandan Padmanaban、Jagabondhu Hazra、Shantanu Godbole 和 Hendrik Hamann。2024年。赋能可持续金融：利用大语言模型进行气候意识型投资。载于*ICLR 2024*机器学习应对气候变化研讨会。Climate Change AI。

Klaus Krippendorff。2004年。内容分析：方法论导论。Sage Publications。J. Richard Landis 和 Gary G. Koch。1977年。分类数据观察者一致性的测量。*Biometrics*，33(1):159–174。John Lawrence 和 Chris Reed。2019年。论证挖掘：综述。《计算语言学》，45(4):765–818。Sanguk Lee、Tai-Quan Peng、Matthew H. Goldberg、Seth A. Rosenthal、John E. Kotcher、Edward W. Maibach 和 Anthony Leiserowitz。2024年。大语言模型能否估计公众对全球变暖的看法？对算法忠实度与偏差的实证评估。*PLOS Climate*，3(8):e0000429。

Markus Leippold。2023年。GPT-3如是说：采访一个气候金融大语言模型。金融研究快报，53:103617。Hao Li、Viktor Schlegel、Yizheng Sun、Riza Batista-Navarro 和 Goran Nenadi´c。2025年。大语言模型与论证挖掘：综述。Yang Liu、Christian Stab 和 Iryna Gurevych。2021。用于论证挖掘的神经转移模型。载于计算语言学协会第*59*届年会论文集（*ACL*），第2135–2149页。计算语言学协会。

Yang Liu、Sheng Zhang、Jingbo Shang 和 Jiawei Han。2020年。面向话语关系文本级解析的自顶向下神经架构。载于 第*58*届计算语言学协会年会（*ACL*）论文集，第6386–6395页。计算语言学协会。

MetaAI。2025年。LLaMA 4：Maverick（17b）——一种多模态混合专家模型。LLaMA-4 Maverick-17B（128个专家）的模型卡片/文档，可通过ModelScope/模型库访问。

Mistral。2024年。Mixtral8×22b。https://huggingface.co/mistralai/Mixtral-8x22B-v0.1。开放权重的混合专家语言模型。

Matthew C. Nisbet。2009年。气候变化传播：为何框架对公众参与至关重要。环境：科学与可持续发展政策，51(2)：12–23。

OpenAI。2024年。GPT-4o 系统卡。arXiv 预印本 arXiv:2410.21276。OpenAI 推出的多模态“全能”模型。

Yulia Otmakhova 和 Lea Frermann。2025年。政治话语中的叙事媒体框架。载于计算语言学协会论文集：*ACL 2025*，第 9167–9196 页，奥地利维也纳。计算语言学协会。

David Patterson、Joseph Gonzalez、Urs Hölzle、Quoc Le、Chen Liang、Lluis-Miquel Munguia、Daniel Rothchild、David So、Maud Texier 和 Jeff Dean。2022年。机器学习训练的碳足迹将达到平台期，随后缩减。《计算机》，55(7):18– 28。Maja Pavlovic 和 Massimo Poesio。2024年。大语言模型作为标注员的有效性：直接表征的比较综述与实证分析。载于*NLPerspectives 2024*。计算语言学协会。Rashmi Prasad、Bonnie Webber、Alan Lee 和 Aravind Joshi。2019 年。宾夕法尼亚话语树库 3.0。语言数据联盟。

ProQuest。2024年。主页 - 道琼斯 Factiva - LibGuides。https://proquest.libguides.com/factiva。该页面称 DJID 包含约 350,000 个分类代码。Qwen。2025年。Qwen3 30b a3b。https://huggingface.co/unsloth/Qwen3-30B-A3B-GGUF。开放权重评估模型。

Nils Reimers 和 Iryna Gurevych。2019年。Sentence- BERT：使用孪生 BERT- 网络的句子嵌入。载于*2019*年自然语言处理实证方法会议暨第九届国际联合自然语言处理会议（*EMNLP-IJCNLP*）论文集，第 3982–3992 页，中国香港。计算语言学协会。

Margaret E. Roberts、Brandon M. Stewart、Dustin Tin- gley、Christopher Lucas、Jetson Leder-Luis、Sharon Gadarian、Bethany Albertson 和 DavidRand。2014年。面向开放式调查回答的结构主题模型。《美国政治科学杂志》，58(4)：1064–1082。

Peter J. Rousseeuw。1987年。轮廓系数：用于解释和验证聚类分析的图形辅助工具。《计算与应用数学杂志》，20：53–65。

Maria Ruiz-Dolz、Santiago Cortes、Javier Garcia 和 Ana Garcia-Serrano。2021。基于 Transformer 的模型：自动识别论证关系。《*IEEE* 智能系统》，36(4)：54–62。

Tatjana Schirmag、Jakob H. Wedemeyer、Annika Stechemesser 和 Leonie Wenz。2025年。神经主题建模揭示德国电视的气候变化报道。《通讯：地球与环境》，6(1)：441。

Hannah Schmid-Petri、Silke Adam、Ivo Schmucki 和 Thomas Häussler。2017a。怀疑氛围的变化：塑造美国新闻界气候变化报道的因素。《科学公众理解》，26(4)：498–513。

Hannah Schmid-Petri、Silke Adam、Ivo Schmucki 和 Thomas Häussler。2017b。不断变化的怀疑氛围：塑造美国媒体气候变化报道的因素。《公众理解科学》，26(4):498–513。Hannah Schmid-Petri 和 Dorothee Arlt。2016a。构建科学不确定性的假象？德国和英国印刷媒体中的气候变化框架。《传播学》，41(3):265–289。Hannah Schmid-Petri 和 Dorothee Arlt。2016b。构建科学不确定性的假象？德国和英国印刷媒体中的气候变化框架。《传播学》，41(3):265–289。Robert J. Shiller。2017。叙事经济学。《美国经济评论》，107(4):967–1004。

Christian Stab 和 Iryna Gurevych。2014。说服性文章中论证成分及其关系的标注。载于*COLING 2014* 会议论文集，第1501–1510页。Christian Stab 和 Iryna Gurevych。2017a。说服性文章中论证结构的解析。载于*2017*年北美计算语言学协会分会会议论文集：人类语言技术，第985–995页，加利福尼亚州洛杉矶。计算语言学协会。

Christian Stab 和 Iryna Gurevych。2017b。说服性文章中的论证结构解析。载于《计算语言学》，第43卷，页码619–659。

Zhen Tan、Dawei Li、Song Wang、Alimohammad Beigi、Bohan Jiang、Amrita Bhattacharjee、Mansooreh Karami、Jundong Li、Lu Cheng 和 Huan Liu。2024年。用于数据标注与合成的大语言模型：综述。载于《*2024*年自然语言处理实证方法会议论文集》，第 930–957 页，美国佛罗里达州迈阿密。计算语言学协会。

Assaf Toledo、Yoav Kantor、Gabriel Stanovsky 和 Ido Dagan。2021。基于图的论证质量评估。载于《自然语言处理最新进展国际会议（*RANLP 2021*）论文集》，第 1427–1437 页。

Petter Törnberg。2024年。使用大语言模型进行文本标注的最佳实践。*arXiv* 预印本 *arXiv:2402.05129*。

Joe H. Ward. 1963年。用于优化目标函数的层次分组。《美国统计学会杂志》，58(301):236–244。

Nicolas Webersinke、Mathias Kraus、Julia Anna Bin- gler 和 Markus Leippold。2022年。Climatebert：面向气候相关文本的预训练语言模型。载于*AAAI 2022*年秋季研讨会论文集：人工智能在应对气候挑战中的作用。arXiv：2110.12010。Qilong Wu、Thibault Masson、Aritra Mukherjee、Yi- hao Zhao、Chaima Driouich、Xingjian Xing、Patrick Paroubek 和 Mickael Coustaty。2024年。Susgen-gpt：面向金融自然语言处理和可持续发展报告生成的以数据为中心的大语言模型。*arXiv* 预印本 *arXiv:2412.10906*。预印本。Liu Zhuang、Lin Wayne、Shi Ya 和 Zhao Jun。2021。一种经过稳健优化的 BERT 预训练方法及后训练。载于第二十届中国计算语言学大会论文集，第1218–1227 页，中国呼和浩特。中国中文信息学会。

A 语料库整理与验证

A.1 DJID代码参考

表5 列出了我们在语料库构建中使用的所有道琼斯智能标识符（DJID）代码，并按第 3.1 节介绍的三个领域分组：核心气候议题、能源转型和受气候影响的行业。

领域	Code	描述
核心气候议题	N/ENV	环境
	N/二氧化碳	二氧化碳 / 排放量
	N/RNW	可再生能源
能源转型	N/BFL	生物燃料
	N/COA	Coal
	N/NUK	核能
	N/NGS	天然气
受气候影响的行业	N/AGR	农业
	N/FST	林业

表5：构建气候新闻语料库时使用的道琼斯智能标识符（DJID）代码。

A.2 DJID筛选验证

为评估基于DJID的筛选的可靠性，我们开展了一项两部分验证研究。

假阴性分析。 我们抽取了1,000篇未标记气候相关DJID的文章，并使用以下方法重新筛查：

- **关键词词库：** 一组经过整理的152个气候相关术语（见附录 A.3），用于匹配文章文本。

- **监督分类器：** 一个在 5,000 篇气候与非气候新闻文章上进行微调的 RoBERTa-base 模型（宏平均 F1 = 0.91）。由任一方法标记为“与气候相关”的文章均经过人工核验。

在 1,000 篇文章中，有 73 篇被标记；人工检查确认其中 60 篇确实与气候相关。这意味着假阴性率为 ∼8%。

假阳性分析。 我们随机抽取了 500 篇带有气候 DJID 的文章。两位标注员分别独立核验每篇文章是否实质性地讨论了气候问题（ Krippendorff's $\alpha = 0.79$）。仅有 21 篇（4.2 %）被判定为假阳性，因此估算精确率为 >95%。

摘要。 总体而言，DJID筛选的召回率为 ∼92 %，精确率为 >95%。尽管准确率很高，未来的工作仍可采用补充方法（基于词典或分类器的扩充）进一步提高覆盖率。

类别	示例关键词
总体气候	气候变化；全球变暖；温室效应
碳	碳；二氧化碳；碳税；碳上限-ture
能源	可再生能源；太阳能；风能；化石燃料；生物燃料
金融	环境、社会和治理；绿色债券；碳市场
政策	《巴黎协定》；《京都议定书》；净 zero

表 6：气候关键词词典的代表性子集。完整词典（152 个术语）见补充发布材料。

A.3 关键词词典构建

气候关键词词典旨在捕捉可能未被标注相关 DJID 的文章。该词典基于既有气候传播研究、政府间气候变化专门委员会词汇表以及金融报道中的领域专用术语构建而成。最终词典涵盖五个主题类别中的 152 个术语。表 6列出了具有代表性的示例；完整词典随我们的补充材料一并发布，以确保可复现性。

A.4 预处理详情

规范化。 所有文章均经过统一处理，包括删除模板内容（例如重复标题、版权声明）、空白规范化和 Unicode 规范化。为保留领域专用术语，我们避免进行词元级清理。

近重复检测。 由于通讯社经常发布同一报道的略微修改版本，我们采用基于 MinHash 的局部敏感哈希（LSH）来识别重复项。Jaccard 相似度 ≥ 0.9 的文本对被标记为近重复项。每个聚类仅保留一个规范版本，移除了 ∼8.7% 的文章。这样可以防止冗余，并避免对通讯社供稿报道进行过度加权。

最终语料库。 预处理后，语料库包含 980,061 篇独特的气候相关文章，并保留了元数据（时间戳、DJIDs、文章来源标识符等），确保历时分析的完整性。

A.5 语料库时间分布

B 抽样超参数分析

B.1 MMR λ 敏感性分析

最大边际相关性（MMR）参数 λ 的选择决定了以下两者之间的权衡

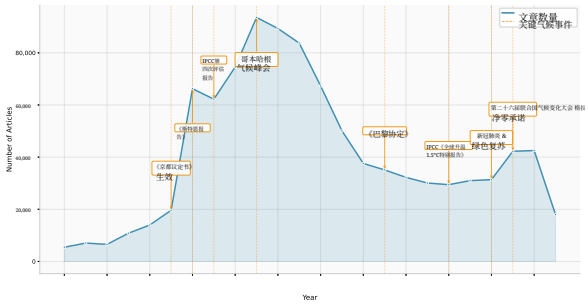


图4：道琼斯气候相关文章的时间分布（2000–2023年），并标注关键气候事件。

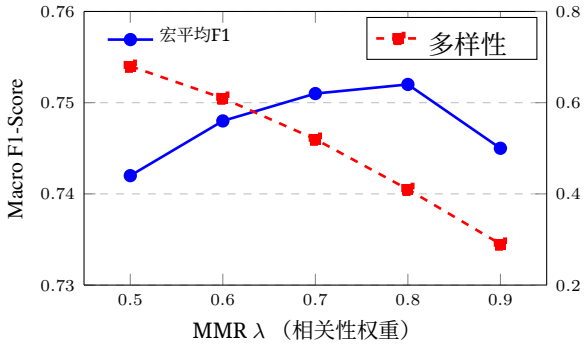


图 5：不同 MMR λ 取值下，模型性能（宏平均F1）与样本多样性之间的权衡。

筛选与主题相关的文章（相关性）以及具有新颖性的文章（多样性）。我们通过使用不同的 λ 值生成五个不同的子样本，并评估下游论点提取模型在每个子样本上的性能，进行了敏感性分析。如表7 和图5所示， $\lambda = 0.7$ 的取值实现了近乎最优的平衡：在保持较高 F1分数的同时，也保留了较高的样本多样性。样本多样性以同一聚类内所选文章两两之间的平均余弦距离衡量。虽然 $\lambda = 0.8$ 带来了小幅的 F1分数提升，但代价是多样性显著下降，因此不太适合我们兼顾主流论点和离群论点的目标。

表 7：下游论点提取模型的性能及样本多样性随 MMR 参数 λ变化的情况。最高的 F1分数以粗体显示。

指标	$\lambda = 0.5$	$\lambda = 0.6$	$\lambda = 0.7$	$\lambda = 0.8$	$\lambda = 0.9$
宏平均F1	0.742	0.748	0.751	**0.752**	0.745
精确率	0.739	0.745	0.750	0.758	0.761
召回率	0.745	0.751	0.752	0.746	0.730
多样性	0.68	0.61	0.52	0.41	0.29

表8： k 取不同值时，2019–2023年时间段中涌现聚类主题的示例。

k**的值**	**聚类主题示例**
$k = 4$	- 可再生能源与政策（混合） - 企业ESG与金融（混合）
$k = 8$	- 可再生能源技术 - 气候金融政策 - 企业可持续发展报告 - 碳市场
$k = 16$	- 公用事业规模太阳能项目（冗余） - 陆上风电项目（冗余） - 自愿碳信用额（冗余）

B.2 主题聚类数量的依据（k）

主题聚类数量的范围 k_t根据定性分析限定为 $[5, 15]$ 。表8列出了2019—2023年分层中的示例，说明 $k < 5$会将不同主题混为一谈，而 $k > 15$则会产生虚假的、粒度过细的微型聚类。

C 主动学习详情

C.1 初步模型架构

主动学习循环由一个基于*RoBERTa-base*的初步论证组件检测器驱动。主要训练超参数详见表9。该模型在留出的 1,000 篇文章种子集上取得了 79.5% 的宏平均 F1 分数。

表 9：主动学习循环中使用的初步论证组件检测器的超参数。

超参数	**值**
基础模型	‘RoBERTa-base‘
学习率	2e-5
优化器	AdamW
批次大小	16
最大序列长度	256
训练轮数	4
预热步数	100

为阐明我们各个标注集之间的关系：

1. **主动学习种子集**（1,000篇文章）：从完整的98万篇语料库中抽取，并在最终抽样前由内部人员进行标注，以引导训练初步的论点检测器。

2. **最终样本**（4,143篇文章）：使用主动学习，并基于种子集训练的检测器构建。

3. **金标准**（2,000篇文章）：4,143个样本的子集，通过 Zooniverse 标注，用于流水线评估。

4. **20k 验证语料库**（附录 D）：从 980k 语料库中另行抽取的分层随机样本，由内部人员标注，仅用于验证我们的 4,143 个样本是否保留了下游任务性能。不用于训练或主要分析。

C.2 主动学习收敛

主动学习过程在 4 次迭代后收敛。我们将收敛定义为：连续两轮迭代中选出的文章的熵分布之间的詹森-香农散度（JSD）低于 0.05 的阈值。图6 展示了这一趋势，表明模型识别出的“最难”文章集合很快趋于稳定。

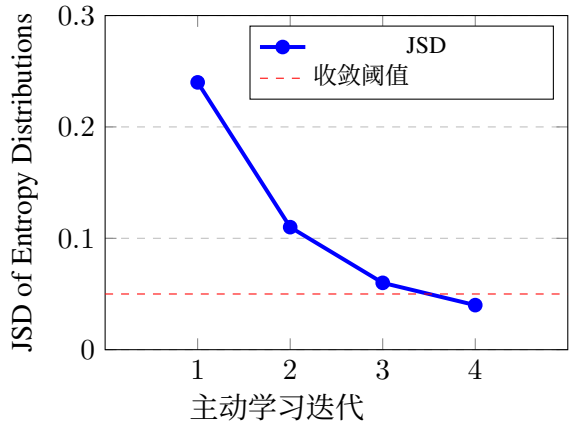


图 6：主动学习循环的收敛情况。在第 4 次迭代时，所选文章集合之间的詹森-香农散度降至 0.05 阈值以下。

D 扩展验证结果

为验证我们包含4,143篇文章的样本完整性，我们实施了严格的验证方案，重点关注两个关键标准：(1) 保留下游任务性能；(2) 与母语料库相比，元数据分布的保真度。

D.1 下游任务性能保持

对一个子语料库而言，最关键的检验在于其对预期下游任务的实用性。我们通过在样本上训练论证提取模型，并将其性能与在规模大得多的20,000篇文章分层随机样本上训练的相同模型架构进行比较来评估这一点。

模型架构。 论证抽取模型是一个词元分类器，基于`RoBERTa-large`构建。我们在 RoBERTa 模型的最终隐藏状态之上添加了一个线性层，将每个词元分类为适用于目标组件（“主张”“前提”）的 IOB格式（内部、外部、开头）。模型采用端到端微调。通过在开发集上进行有限搜索选定的关键超参数详见表 10。

表10：验证实验中使用的 RoBERTa-large 论证抽取模型的超参数。

超参数	数值
基础模型	roberta-large
学习率	1e-5
优化器	采用线性衰减的 AdamW
权重衰减	0.01
批次大小	8
最大序列长度	512 个词元
训练轮次	5

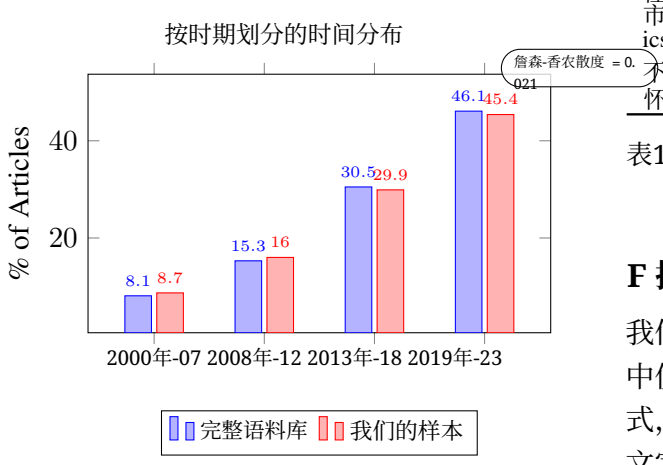
结果。如表11所示，在我们的4,143篇文章样本上训练的模型取得了75.1%的宏平均F1分数。与在2万篇文章基线模型上训练的模型所取得的78.4% F1分数相比，性能仅下降了3.3个百分点。这一结果令人鼓舞，表明我们的抽样策略在使用不到21%的训练数据的情况下，保留了95%以上的性能，从而证实了其出色的数据效率。

表11：论证提取任务的详细性能比较。两个模型均在同一留出测试集上进行了评估，该测试集包含 1,500 篇文章。每项指标的最佳分数以粗体显示。

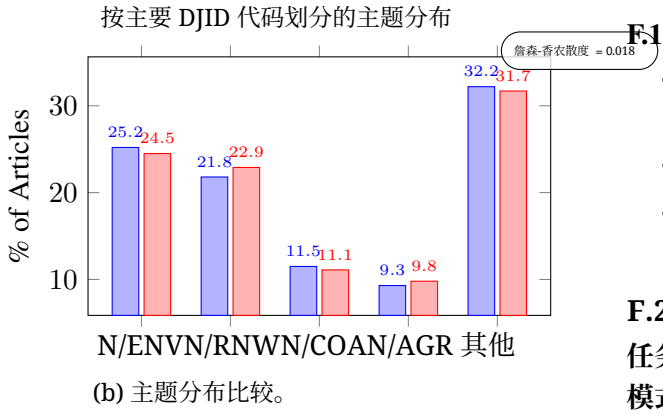
训练样本	主张			前提			宏平均
	P	R	F1	P	R	F1	F1
2万分层随机抽样	**0.801**	**0.753**	**0.776**	**0.769**	**0.824**	**0.795**	**0.784**
我们的样本（4千）	0.782	0.725	0.752	0.723	0.781	0.751	0.751

D.2 分布相似性与保真度

我们进一步验证了样本的元数据分布忠实地反映了其所属的98万篇文章语料库的分布。我们针对关键分布计算了詹森-香农散度（JSD），其中JSD分数越低，表示相似度越高。图7 直观比较了时间分布和主题分布，证实了二者具有极高的保真度（两者的JSD < 0.03）。



(a) 时间分布比较。



(b) 主题分布比较。

图7：完整的98万篇文章语料库与我们的4000篇文章样本之间关键元数据分布的比较，证实了分布上的高度忠实度。

E 框架类型学

我们的分析采用了预先定义的财经新闻气候话语八框架类型学。该类型学综合了传播学研究中的类别（Nisbet，2009；Schmid-Petri andArlt，2016b），以及在初步语料探索中观察到的金融特定框架（例如经济风险、市场动态）。表 12 给出了每个框架的定义。

框架	定义
经济机遇-遇	将气候变化框定为增长、创新-创新，以及投资潜力。
经济风险	强调财务损失、搁浅资产——以及对企业或市场的风险。
监管合——规	聚焦法律、政策及监管负担或激励措施。
技术解决方案	强调创新、研发及技术解决气候挑战的方法。
环境紧迫性	强调生态严重性以及采取迅速行动的必要性。
社会责任性	涉及伦理、企业责任，以及社会预期。
市场动-ics	将气候置于竞争的框架中，供需和市场定位。
不确定性怀疑	对气候科学、政策或影响表示质疑或怀疑。

表12：财经新闻气候话语的八框架类型学。

F 提示模板与模式

我们公开了行动者—框架—论证（AFA）流程中使用的所有提示模板和模式。为适应双栏格式，我们采用紧凑的 JSON 模式，并将说明性文字保持简短。完整的 JSON 模式文件和脚本收录于我们的补充材料中。

F.1 通用协议

- 输出必须是有效的 JSON 对象（不得包含说明文字）。
- 缺少字段：返回 null 或空列表。
- 框架必须与预定义的类型学完全一致。

F.2 阶段1：行动者—立场

任务： 提取行动者、类型、立场及支持性引文。
模式：

{ "行动者": [{ "名称": "字符串", "行动者_类型": "公司|政府|个人|组织", "立场": "支持|反对|中立|混合", "引文_文本": "字符串", "气候_相关性": "字符串" }]}

F.3 阶段2：框架分类

任务： 指定一个主要框架，并可选择指定一个次要框架。**模式：**

{ "primary_框架": "框架_枚举",

"secondary_框架": "框架_枚举_或_空值", " 理由 ": " 字符串 ", " 气候 _ 关联 ": " 字符串 " }

F.4 阶段 3：论证提取

任务： 提取主张、证据、论证依据、影响以及可选的支持性论证。**模式：**

{ "claim": "字符串", "evidence": ["字符串", "..."],"warrant": "字符串", "impact": "字符串", "supporting_arguments": [{ "claim": "字符串", "evidence": ["字符串", "..."]," warrant": "字符串" }]}

F.5 DVFJudge 评审提示

维度： 完整性、忠实性、连贯性、气候相关性。每项原子检查均在 $[0, 1]$, 连续量表上评分，其中 0 表示完全失败，1 表示完全成功，中间值表示部分正确。

模式（示例）：

{ "完整性": {"行动者":0.0-1.0, "立场":0.0-1.0, "框架":0.0-1.0, "论证":0.0-1.0}, "忠实性": {"引用_对齐":0.0-1.0

"段落_等价性":0.0-1.0}, "连贯性": {"链接_保留":0.0-1.0, "模式_格式正确":0.0-1.0}, "气候_相关性": {"切_合主题":0.0-1.0, "排除外围内容_": 0.0-1.0
}

F.6 推理设置

温度 = 0.2，Top-p = 0.9，最大词元数 = 512（阶段 1–2）、768（阶段 3）、256（DVF）。停止序列：\{"\{\n ","“\}。根据模式验证输出；失败时重试。

G 人工标注金标准

为我们的 AFA 流程建立可靠的评估基准，我们从语料库中抽取了 2,000篇文章，构建了一个金标准数据集。本节详细介绍标注协议、指南、裁定流程和一致性指标。

G.1 标注员培训与背景

我们通过*Zooniverse*收集标注数据。Zooniverse 是一个广泛使用的公民科学平台，用于分布式人工标注。我们没有采用一支小型内部编码员团队，而是利用了大量志愿者。为确保标注质量，我们采取了以下措施：

- **入门教程：**所有参与者均完成了交互式教程，学习行动者、框架和论元标注的示例，并完成了带有反馈的练习任务。
- **冗余标注：**每篇文章均由至少 $k = 5$ 名志愿者独立标注，以减少个人错误。
- **聚合：**最终标签通过多数投票（适用于立场和框架等分类任务）或共识启发式方法（适用于论元片段，依据词元级一致性）确定。

尽管标注员没有接受过语言学或气候传播方面的正式培训，但冗余标注与聚合流程仍取得了较高的标注员间一致性（见表 2）。这种方法d证明了针对复杂话语任务开展可扩展的公民科学标注e的可行性。

G.2 标注指南

该模式涵盖行动者—框架—论元（AFA）流程中的所’有组件：

- **行动者—立场**：识别所有行动者（公司、政府、非政府组织、个人），并将立场归类为支持、反对、中立或混合。每项行动者标注都需要附上支持性引文。
- **框架分类**：指定一个主要框架，并（如有）指定一个次要框架，从八框架类型学中选择（见附录E）。标注员提供简短的理由说明。
- **论证提取**：将论据结构分解为：(i) 核心主张，(ii) 证据片段，(iii) 连接主张与证据的论证依据，以及 (iv) 影响（明示或暗示）。如有其他支持性论据，也会予以标注。

带注释文档示例及指南摘录随补充材料一同发布。

G.3 标注流程

每篇文章均由至少五名独立贡献者在 Zooniverse 上进行标注。对于分类任务（立场、框架），最终标签由多数投票决定。对于基于跨度的任务（行动者、证据、论证依据），我们通过词元级一致性汇总标注结果，并保留至少 60% 的贡献者标记的跨度。分歧通过共识汇总而非个别裁决解决。这种冗余协议在利用公民科学规模优势的同时，为专家标注质量提供了稳健的近似评估。

H 标注指南（摘录）

本节摘录了用于培训人工编码员的完整标注指南。这些示例展示了标注员如何在行动者—框架—论元（AFA）模式中区分标签并处理常见的边缘情况。完整指南和编码手册随我们的补充材料一并发布。

H.1 行动者—立场标注

任务： 识别行动者（组织、公司、政府、非政府组织或个人）并判定其立场。

规则：

- 行动者必须在文本中被明确提及。诸如“分析师”或“评论家”等泛称不予纳入，除非与某个具名实体相关联。
- 如果提到多个子公司或关联机构，只有在它们表达不同立场时，才将其视为不同的行动者。

- 立场定义：

  – **支持：** 明确支持气候行动、投资或遵守监管规定。– **反对：** 拒绝或抵制气候措施，或强调其负面影响。– **中立：** 提及气候问题，但不作评价性判断。– **混合：** 在同一语境中表达支持和反对立场。

- 每项立场标注都必须有逐字引文作为依据。

示例： *“埃克森美孚表示，新的碳排放规则将增加能源生产商的成本。”* → 行动者：埃克森美孚；立场：反对；引文：原文照录。

H.2 框架分类

任务： 从八种框架类型中指定一个主要框架，并在存在次要框架时指定一个。

规则：

- 主要框架 = 议题呈现的主导方式。

- 次要框架 = 可选，仅当文本明确涉及次要维度时使用。
- 如果出现多个框架，请选择与主张或论点最相关的框架作为主要框架。

示例： *“投资者认为，可再生能源是下一个重大的增长机遇。”* → 主要框架：经济机遇；次要框架：技术解决方案。

H.3 论证提取

任务： 分解为主张、证据、论证依据和影响。

规则：

- **主张：** 行动者提出的主要命题。

- **证据：** 文本中引用的事实性或统计性支持。使用逐字摘录。
- **论证依据：** 连接主张与证据的推理。这部分可以是隐含的，但应进行释义。

- **影响：** 接受该主张后明示或暗示的结果。

示例： 句子：“贝莱德认为，绿色投资组合的表现优于传统基金，ESG基金在2020年增长了4.3%。”

- 主张：“绿色投资组合的表现优于传统基金。”

- 证据：“ESG基金在2020年增长了4.3%。”

- 论证依据：“财务回报证明可持续投资是合理的。”

H.4 歧义消解

- 如果一段文字符合多个合理标签，请选择证据支持最充分的那个。
- 只有当单个行动者明确表达了支持和反对两种立场时，才使用“混合”立场。
- 标注员会将不明确的案例标记出来，交由裁定，而不是猜测。

H.5 标注者间一致性

我们针对每种组成部分分别计算一致性，并采用适用于多标注员场景的指标：

- **施事者识别**：对标注员两两计算词元级精确率、召回率和 F1 分数，再以宏平均F1进行汇总。
- **立场分类**：Fleiss' κ 和 Krippendorff 的 α （名义尺度），针对四种立场类别。
- **框架分配**：针对八框架集合中主要框架的 Krippendorff α ；次要框架则计算标注者两两之间的 Jaccard 相似度平均值。
- **论证成分**：论证成分（主张、证据、论证依据、影响）的片段级 F1，通过多数投票进行聚合。报告各成分的宏平均 F1。

H.6 基于跨度的任务的共识推导

对于基于跨度的任务（行动者、论证），我们采用两步式词元聚合和跨度重构程序，根据多个人工标签推导出单一的黄金标准标注。这种方法整合细粒度的词元一致性，同时避免较长论证成分出现过度碎片化。

步骤 1：词元级一致性。将每篇文章标记为词元序列 $T = [t_1, t_2, \ldots, t_m]$。每位标注者 a_i 提供一个二元标签序列$y_i(t_j) \in \{0, 1\}$，其中 1 表示词元 t_j属于某个感兴趣的跨度（例如行动者提及、主张、证据）。对于每个词元，我们计算一致性得分：

$$A(t_j) = \frac{1}{N} \sum_{i=1}^{N} y_i(t_j),$$

其中 N 是标注员的数量（通常为 $N = 5$）。满足 $A(t_j) \geq \tau$ （且 $\tau = 0.5$）的词元会被保留，从而生成共识掩码 $M = [m_1, m_2, \ldots, m_m]$，其中 $m_j = 1$ 当且仅当 $A(t_j) \geq \tau$。

步骤 2：跨度重构。连续保留的词元会合并为统一片段：

$$S_k = \{t_p, \ldots, t_q\} \text{ such that } m_p = \cdots = m_q = 1, (m_{p-1}, m_{q+1}) =$$

如果多个标注者跨度在语义上发生重叠，我们将计算成对的片段级 F1 分数：

$$\mathrm{F1}(H_i, S_k) = \frac{2 \times |\mathrm{Tokens}(H_i) \cap \mathrm{Tokens}(S_k)|}{|\mathrm{Tokens}(H_i)| + |\mathrm{Tokens}(S_k)|},$$

其中， H_i 表示人工跨度。具有 $\mathrm{F1} > 0.7$的跨度将按照最大覆盖原则（最小起点、最大终点）进行合并。

处理非跨度词元。未被任何标注者标记的词元（例如，功能词）将被分配外部标签 O：

$$y_i(t_j) = 0 \quad \forall i, \text{ if } t_j \text{ unmarked.}$$

这些词元不参与聚合，但会保留用于评估。词元级 F1 分数计算如下：

$$\mathrm{F1} = \frac{2 \times \mathrm{TP}}{2 \times \mathrm{TP} + \mathrm{FP} + \mathrm{FN}},$$

其中，始终获得一致标注的 O 词元既不计入分子，也不计入分母。

示例。句子：“贝莱德表示，气候主题投资组合将推动增长。”

三位标注员标注：

- A_1： [贝莱德]； [气候主题投资组合将推动增长]
- A_2： [BlackRock Inc.]； [气候主题投资组合将推动增长]
- A_3：[贝莱德]； [投资组合将推动增长]

经过词元对齐和多数投票后，只有贝莱德 和以气候为重点的投资组合将推动增长达成共识。所有其他词元仍标记为 O。由此得到干净且互不重叠的金标准跨度。

算法摘要（伪代码）。

输入：分词文本 T、标注矩阵 Y、每个词元的阈值 tau 对于 T 中的词元 t_j：A(t_j) =mean(Y[:, j]) M[j] = 1 if A(t_j) >=tau else 0 将 M[j] = 1 为 1 的连续词元合并为跨度 S_k 对于每个候选跨度 H_i：计算 F1(H_i, S_k) 若 F1 > 0.7，则合并 H_i 并采用最大覆盖 返回共识跨度 S

这种两步法既能为短提及保留边界保真度，又能确保对较长的分句级论证具有稳健性，从而生成可复现的、基于跨度的评估金标准。

H.7 在评估中的应用

经裁定形成的共识标签构成第 5 节所有内在评估的金标准。模型输出通过精确率、召回率和 F1 分数与该集合进行比较评估。此外，金标准还为分解式验证框架（DVF）提供验证基线，确保自动验证分数以可靠的人工标注为依据。

I 分解式验证框架

I.1 分解式验证子检查

DVF 将评估分解为原子级子检查，使模型更易于验证，也便于人工审查。

- **完整性：** (1) 是否识别出所有行动者? (2) 是否提取了立场? (3) 是否分配了框架? (4) 是否完整捕捉了论证结构?

- **忠实性：** (1) 每个提取的组件是否都与直接引文相符? (2) 释义在语义上是否等价?

- **结构连贯性：** (1) 是否保留了行动者–框架–论证之间的关联? (2) 模式是否格式良好?

- **气候相关性：** (1) 提取的框架是否确实与气候有关? (2) 是否排除了外围问题（例如一般市场新闻）?

维度	GPT-4o	Claude	Qwen3	Mixtral
完整性	0.842	0.837	0.814	0.808
忠实性	0.902	0.895	0.871	0.868
连贯性	0.801	0.794	0.772	0.769
相关性	0.869	0.861	0.842	0.836

表13：基于 2,000 篇文章金标准数据集，由单个评审评定并取平均值的 DVF 维度分数。

本附录完整列出了 DVF 评估的细分结果，是对第 5.1 节所报告汇总结果的补充。请注意，DVF 从四个维度评估提取结果： 完整性、忠实性、结构连贯性和气候相关性。分数基于包含 2,000 篇文章的黄金标准集取平均值，并通过独立的 500 个样本人工评估进行验证。

I.2 分解式验证框架结果

I.3 各评审的性能

DVF采用四种不同的评审模型来缓解自评分偏差: GPT-4o（主要模型）、Claude-Sonnet-4、Qwen3-30B A3B和Mixtral-8×22B. 表格13 报告了汇总前各评审的分数。GPT-4o和Claude的一致性最高，而开放权重模型提供了有竞争力但略显嘈杂的估计值，并通过跨系列三角验证确保稳健性。

I.4 人工验证研究

为校准自动化DVF分数，我们构建了一个独立的人工评估集，其中包含随机抽取的500个模型输出。每个输出均由三名受训编码员使用DVF评估标准进行标注，他们独立于黄金标准标注员。编码员之间的一致性较高： $\kappa = 0.78$ （忠实度）、$\alpha = 0.74$（完整性）和 $\alpha = 0.71$（连贯性），表明结果具有较高的可靠性。

表14 比较了模型评审的汇总分数与人工标注。自动化 DVF 评估与人工评分高度一致，各维度的偏差均低于 0.03。这一验证证实，DVF 能够忠实地替代专家评估，同时具备高效扩展能力。

这些结果表明，尽管不同评审之间存在差异，但总体 DVF 分数仍经过了充分验证，与人工判断相符，因此有理由将其用作论文正文中的主要评估指标。

维度	人工评分	DVF（平均值）
完整性	0.842	0.821
忠实性	0.895	0.857
连贯性	0.781	0.792
相关性	0.864	0.816

表14：在包含500个条目的验证集上，对比人工编码的DVF分数与汇总后的自动化DVF分数。

J 结果

J.1 行动者类别占比

表 15 报告了四个时间分层中各行动者类别的确切比例，补充了图2中的趋势可视化。

时期	比较	金融机构	Gov.	NGO	个人
2000–2007年	20.7	14.8	31.4	19.2	13.9
2008–2012年	24.7	16.3	28.1	12.4	18.5
2013–2018年	26.5	24.2	22.7	11.1	15.5
2019–2023年	27.8	33.6	19.6	8.7	10.3

表15：各时间分层中的行动者类别占比（%）。

J.2 随时间变化的框架分布

表16报告了早期（2000–2014年）与近期（2019–2023年）时期的精确框架比例（%），并对时间独立性进行了显著性检验。这些数值补充了图3以及第5.3节中所述的变化点分析。

框架	2000–2007年	2019–2023年	Δ
经济机遇	18.3	33.7	+15.4
经济风险	31.2	18.1	−13.1
法规遵从	26.8	16.9	−9.9
技术解决方案	15.2	21.4	+6.2
市场动态	11.8	14.3	+2.5
环境紧迫性	9.2	8.9	−0.3
社会责任	6.2	7.8	+1.6
不确定性/怀疑	8.7	4.9	−3.8

表16：早期（2000– 2007年）与近期（2019–2023年）时期的框架分布转变（百分比）。卡方检验表明，大多数框架都发生了显著的时间变化（*** $p<.001$，** $p<.01$，* $p<.05$；n.s. 表示不显著），与正文趋势叙述相符。

J.3 行动者–框架关联矩阵

表17 报告了以下检验的标准化残差：χ^2 用于检验行动者群体与框架使用之间独立性的检验。正值表示过度代表；负值表示代表不足。

行动者类型	经济机遇	经济风险	技术解决方案	环境紧迫性
公司	$+4.2^{***}$	-2.8^{**}	$+2.1^{*}$	-3.9^{***}
金融机构	$+5.7^{***}$	-1.9^{*}	$+1.6$	-4.2^{***}
政府/监管机构	$+0.8$	$+1.2$	-0.6	$+1.8^{*}$
非政府组织/倡导组织	-5.1^{***}	$+0.9$	-1.3	$+6.8^{***}$
研究人员/专家	-1.4	$+2.3^{*}$	$+3.4^{***}$	$+2.7^{**}$

表17：行动者—框架关联矩阵（ χ^2 检验的标准化残差）。正值表示过度代表。显著性：*** $p<.001$，** $p<.01$，* $p<.05$。

表示。显著性水平已使用Bonferroni校正进行调整。

J.4 定性误差分析

为更好地了解模型的局限性，我们从所有施事者–框架–论元（AFA）组件中随机抽取了150个错误预测，并对其进行了人工检查。两位作者按照主评估中使用的同一模式对每个案例进行了标注。最终归纳出三类主要错误，共解释了超过80%的观测失败案例。

行动者歧义（32%）。在引用多个实体或发言人的段落中，常会出现歧义。例如：

> “贝莱德和几位市场分析师表示，绿色债券的表现将优于传统固定收益产品。”

模型经常将“贝莱德”和“分析师”合并为一个行动者片段，或将其归为“金融机构”这一集体类型。要区分发言者角色，可能需要运用话语层面的共指关系或依存关系线索。

框架重叠（27%）。当机遇主题和技术主题同时出现时，系统难以确定主导框架。示例：

> “投资可再生能源创新，能够确保适应净零政策的企业保持竞争力。”

在此例中，人工标注员将经济机遇视为主要框架，将技术解决方案视为次要框架，而模型则将二者的主次顺序颠倒了。引入层级化或多标签框架建模，或许能够缓解这种混淆。

主张边界漂移（25%）。在包含从句的复杂句子中，主张边界往往会过度延伸。示例：

> “专家认为，更严格的信息披露规则虽然起初可能会加重企业负担，但最终将提升透明度和投资者信心。”

模型将整句话提取为一个主张，未将主张与证据区分开来。跨度级注意力或话语分段有助于约束提取。

较少见的错误。其他次要问题包括讽刺语气下的立场误分类（8%），以及对国际货币基金组织或世界银行等超国家组织的行动者类型误识别（6%）。

启示。总体而言，定性检查表明，大多数错误源于修辞或句法复杂性，而非词汇缺口。这些发现表明，未来的工作应将大语言模型标注与具备话语感知能力的解析器及细粒度验证相结合，以便更好地处理金融评论中的歧义。