权威媒体语调与金融条件预测：基于《人民日报》的弱监督学习测度

0xBlank

摘要：本文基于《人民日报》全量文本语料，构建月度政策沟通指数并检验其对金融条件变化的预测增量。为兼顾经济可解释性与算法泛化能力，本文创新性地引入弱监督学习（Weakly Supervised Learning）框架，利用词典先验生成弱标签以训练分类模型，有效克服了传统词典法的覆盖不足与无监督学习的语义含混。区别于常规的样本内回归，本文采用严格的滚动样本外预测（Rolling OOS）与 Clark-West 检验，在防止前视偏差的前提下评估预测精度。实证结果显示，弱监督指数在疫情前（Pre-COVID）子样本中对短端利率变化具有显著的边际预测能力，表现优于词典及 PCA 指数；且在“四万亿”刺激等重大政策窗口期呈现出高度的方向一致性。本文展示了利用权威文本数据挖掘宏观信号的可行路径，为政策沟通研究提供了严谨的量化实证框架。

关键词：政策沟通；文本分析；弱监督学习；滚动样本外预测；金融条件

一、引言

在过去二十年的宏观经济学实践中，一个显著的范式转变是：政策的影响力已不再局限于利率升降或准备金率调整等传统的“量价工具”。在金融市场微观结构日益复杂、资产价格对预期高度敏感的当下，“央行沟通”（Central Bank Communication）本身已演化为核心政策工具。理论上，有效的政策沟通能够通过校准市场对未来路径的预期、修正风险溢价，从而在实质性政策落地前即改变金融条件（Financial Conditions）。在中国独特的制度语境下，这种沟通往往体现为“预期管理”：权威媒体不仅是政策发布的窗口，更是政策意图传递与市场情绪引导的关键节点。因此，一个极具实证价值的问题浮出水面：我们能否从海量的非结构化权威文本中提取出高频、可度量的“政策沟通信号”，并证明其蕴含了传统宏观变量历史信息之外的“增量预测能力”？

尽管已有文献尝试利用文本挖掘技术构建相关指数，但在实证应用中仍面临“解

释性”与“灵活性”的两难：传统的词典法（Dictionary Approach）虽然含义明确，但难以捕捉语境变化与新词汇；而纯粹的非监督学习（如 LDA、PCA）虽能降维，却往往生成难以赋予经济学含义的统计因子。

针对上述痛点，本研究基于《人民日报》的全量文本语料，试图构建一套兼具经济直觉与算法精度的月度政策沟通指数，并将其置于严格的宏观预测框架下进行检验。与简单的词频统计不同，本文在方法论上引入了“弱监督学习”（Weakly Supervised Learning）的思路：我们不仅保留了基于 TF-IDF 和主成分分析（PCA）的无监督特征，更创新性地利用词典情感信号生成“弱标签”，训练 Logistic 回归模型以构建政策语气指数（记为 tone_logit）。这一设计旨在结合两者的优势——既利用词典作为具有经济学先验的方向锚点，又通过机器学习模型的泛化能力缓解词表覆盖不足与语义多义性问题。

为确保实证结论并非源于数据挖掘（Data Mining）或过度拟合，本文建立了一套严谨的数据生成与核验流程。我们严格剔除任何形式的“未来数据”，确保所有预测模型仅依赖于 （t−1）期及更早的信息集；同时，我们对文本规模（如文章篇幅、字数）进行了正交化处理，以证明指数的波动源于政策态度的变化，而非媒体版面的机械性调整。

在实证策略上，本文并未止步于解释变量的样本内显著性（In-sample Significance），而是将样本外预测能力（Out-of-sample Predictability）作为核心检验标准。具体而言，我们选取短端利率作为金融条件的代理变量，构建了包含政策沟通特征的 ARX(12) 增强模型，并将其与基准 AR(12) 模型在滚动窗口下进行对比。通过计算均方预测误差（MSPE）比率并采用 Clark–West 检验，我们以此识别文本指标是否真正提供了统计显著的增量信息。

此外，为了验证指数的经济学逻辑自洽性，本文还引入了“事件研究”视角作为外部有效性（External Validity）的补充证据。我们考察了指数在 2008 年“四万亿”刺激、2014 年降息、2017 年金融工作会议以及 2019 年 LPR 改革等关键历史节点的表现。这不仅是为了检验指数能否在重大政策转向时给出正确的方向性信号，更是为了观察短端利率在这些时间窗口内是否存在符合预期管理逻辑的响应。这种基于典型事实（Stylized Facts）的定性核验，与定量的滚动预测互为印证，增强了结论的稳健性。

本文的边际贡献主要体现在三个维度：

第一，数据资产与方法融合。本文利用权威语料构建了包含词典法、PCA 降维与弱监督学习在内的多维度政策沟通指数体系，并公开了构建细节与核验结果，为后续研究提供了可复现的数据基础。

第二，严格的预测增量检验。区别于仅关注相关性的研究，本文在严格的时间对齐与滚动样本外框架下，证实了 AI 文本指标对金融条件具有实质性的预测增量，这对于理解高维数据在宏观预测中的价值具有实证意义。

第三，多层次的有效性验证。通过结合统计检验（Clark-West）与历史叙事（政策事件冲击），本文系统地展示了政策沟通信号的有效边界——既肯定了其在信息不对称时期的引导作用，也审慎地指出了复杂非线性模型在有限样本下的适用局限。

本文后续安排如下：第 2 节详细说明数据来源、预处理流程与变量构造；第 3 节阐述文本指标的构建算法及审计框架；第 4 节报告核心预测结果、稳健性检验及事件窗口分析；第 5 节深入讨论实证结果的经济含义与方法论启示；第 6 节总结全文并展望未来的研究方向。

本文检验：权威媒体语调是否在宏观变量历史信息之外，为短端利率变化提供统计显著的样本外增量预测能力；并比较词典 / PCA / 弱监督三类测度的优劣。

二、数据来源与变量构建

本研究致力于在严格的实证框架下检验政策沟通的有效性。为此，我们构建了一个包含高维非结构化文本与低维宏观经济变量的月度数据集。样本区间涵盖 2005 年 1 月至 2023 年 12 月，这一时段完整覆盖了中国利率市场化改革的关键进程及货币政策框架的转型期。考虑到 2020 年初爆发的新冠疫情（COVID-19）可能引发宏观经济的结构性断裂（Structural Break），我们在全样本分析之外，还专门设置了截至 2019 年 12 月的“疫情前子样本（Pre-COVID）”，以排除极端外生冲击对预测稳健性的潜在干扰。

2.1 文本数据工程：从非结构化语料到月度信号

本研究的文本语料基础来自于《人民日报》的全量历史归档（覆盖 1946-2023）。作为中国最具权威性的官方媒体，该报不仅是政策发布的载体，更是外界解读政策风向的核心窗口。为了从海量非结构化文本中提取可用于计量分析的结构化数据，

我们实施了以下数据工程步骤：

首先，进行噪声剔除与清洗。原始语料包含大量非政策相关的广告、版面填充信息及格式乱码。我们构建了自动化清洗管线，剔除了无意义字符与非文本噪声，确保后续分析聚焦于具有实质内容的文本信息。

其次，实施日度到月度的频率聚合（Temporal Aggregation）。虽然原始数据为日度频率，但考虑到宏观经济数据多为月度发布，且单日文本波动可能包含较大的随机噪声，我们将经过清洗的日度语料按月进行聚合。这一处理不仅实现了与宏观变量的频率对齐（Frequency Alignment），更重要的是平滑了短期由于版面安排导致的异动，使指数更能反映中长期的政策基调变化。

2.2 政策沟通指数：多维度的测度体系

为了克服单一文本指标的片面性，本文构建三类月度政策沟通指数用于实证分析（详细构建算法见第 3.2 节）：基于词表净语气强度的 tone_dict、基于 TF–IDF 表示并由 PCA 提取的无监督文本因子 tone_pca，以及以词典信号生成弱标签并训练 Logistic 回归得到的弱监督指数 tone_logit。三者均在月度频率上与宏观变量对齐，后文所有回归与预测均使用其滞后一期值进入方程以避免信息泄漏。

2.3 宏观变量：金融条件与实体机制

在被解释变量的选择上，本文区分了“金融条件”与“实体经济”，旨在不仅检验“预测性”，还能以此推断政策沟通的传导机制。

(A) 金融条件变量（核心预测目标）

我们将短端市场利率的变化作为主要的被解释变量。理论上，相比于具有较强惯性的实体经济指标，金融资产价格——特别是短端资金价格——对央行的口径变化反应最为迅速。

dshibor3m：定义为 3 个月上海银行间同业拆放利率（Shibor 3M）的月度一阶差分，即 $dshibor3m_t = shibor3m_t - shibor3m_{t-1}$。选用差分形式是为了确保时间序列的平稳性，满足计量模型的基本假设。

(B) 价格与实体变量（对照与机制分析）

为了验证政策沟通影响的广泛性及其局限性，我们引入了价格和产出变量作为对照：

dcpi_yoy（价格对照）：消费者价格指数（CPI）同比增速的月度一阶差分。$dcpi\_yoy_t = cpi\_yoy_t - cpi\_yoy_{t-1}$，之所以使用一阶差分，是为了削弱通胀同比的强自相关，从而更适合检验文本信号的“增量预测信息”。

ip_yoy（中间量，用于构造机制变量）：工业生产序列通常以同比指数形式发布，记为 $IPIndex_t$（=100 表示同比持平），据此定义工业同比增速为

$$ip\_yoy_t = IPIndex_t - 100$$

dip_yoy（实体机制）：工业增加值（Industrial Production）同比增速的月度变化。$dip\_yoy_t = ip\_yoy_t - ip\_yoy_{t-1}$ 该变量用于捕捉政策沟通向实体经济的传导。需要说明的是，鉴于中国宏观数据的发布特征（如春节效应导致的 1-2 月数据合并），我们在处理该序列时保持了审慎态度：在主回归中严格遵循日历时间，而在涉及缺失值的长窗口滚动预测中，作为稳健性检验，我们仅在附录中报告经过线性插值处理的结果。

2.4 变量汇总

Table 1：总结了本文核心变量的定义、构建方法及对应的样本区间。所有涉及计算与清洗的代码脚本均已归档，确保研究过程的可复现性。

类别	变量代码	变量名称与定义	数据频率	样本处理说明
政策沟通	tone_logit	弱监督 Logit 指数：基于词典信号训练的分类概率得分	月度	核心指标：使用 $t-1$ 期值，方向已核验
	tone_dict	词典净语气指数：基于正负情感词频差构建	月度	对照组：基于规则的基准
	tone_pca	PCA 文本因子：基于 TF-IDF 矩阵的第一主成分	月度	对照组：纯数据驱动的无监督指标
金融条件	dshibor3m	短端利率变化：Shibor 3M 的月度一阶差分	月度	主预测目标：对政策信号敏感，反应迅速
价格水平	dcpi_yoy	通胀变化：CPI 同比增速的一阶差分	月度	对照目标：用于检验强惯性下的预测增量
实体经济	dip_yoy	产出变化：工业增加值同比增速的一阶差分	月度	机制变量：用于分析政策传导效应

Table 1：变量定义与样本设定

注：样本主区间为 2005:01-2023:12。考虑到疫情冲击，部分稳健性检验基于 Pre-COVID 子样本（截止 2019:12）。所有宏观变量均经过平稳性检验。

三、方法：测度构建、审计与预测评估框架

本研究的核心实证逻辑在于：首先，基于非结构化语料构建可信的政策沟通测度；其次，通过严格的审计程序确保指标不仅是统计上的数字，更具有经济学意义上的方向性；最后，在防止“前视偏差（Look-ahead Bias）”的前提下，检验该指标对宏观金融条件的增量预测能力。

3.1 文本预处理与时频聚合

原始的《人民日报》语料以日度为单位组织。考虑到宏观经济分析的惯例及变量频率，我们实施了由日向月的数据降维。具体而言，我们在清洗阶段移除了广告、版面填充符及非语义噪声，随后将当月所有版面的清洗文本合并为一个“月度文档”。

这一处理基于两点考量：第一，实现与 Shibor、CPI 等月度宏观序列的频率对齐；第二，平滑高频噪声——单日的版面变动可能受突发性非经济事件干扰，而月度聚合文本更能反映决策层在中期内稳定的政策口径与注意力分配。

3.2 政策沟通指数的构建策略

为避免单一测度方法的偏误，本文构建了三类指数，分别代表了“专家规则”、“数据驱动”与“人机结合”三种方法论视角。

3.2.1 词典指数（tone_dict）：基于规则的基准

作为基准，我们利用包含“扩张/支持”与“收紧/监管”等语义的金融情感词典，计算月度文本的净语气强度。该方法的优势在于经济含义直观（Explicit Semantics），但劣势在于词表的静态性难以捕捉不断演化的政策语境（Contextual Shift）。因此，在本文中，它主要作为确定指数方向的“锚点”以及生成弱标签的来源。

3.2.2 无监督因子（tone_pca）：纯数据驱动

我们采用 TF-IDF 向量空间模型表示月度文本，并利用主成分分析（PCA）提取第一主成分作为文本变动的共同因子。该指标完全由数据方差驱动，不依赖任何人为主观判断，但其最大的缺陷在于缺乏内生的经济学方向感——统计上的“正向冲

击”不一定对应经济上的“政策宽松”。

3.2.3 弱监督学习指数（tone_logit）：核心指标

针对上述两种方法的局限，本文采用弱监督学习（Weakly Supervised Learning）框架。具体做法是：利用词典生成的信号作为“弱标签”（例如，若词典显示当月显著偏宽松，则标记为 1），在文本特征空间上训练 Logistic 回归模型，并将模型输出的概率值转化为连续指数 tone_logit 。

$$p_t = \Pr(\text{expansion}|\mathbf{x}_t) = \sigma(\boldsymbol{\beta}^\top \mathbf{x}_t)$$

这一设计的精妙之处在于：它利用词典提供了方向性的先验（Prior），同时利用机器学习模型的泛化能力去捕捉那些不在词表中、但与“宽松”高度共现的隐含特征。

3.3 指数审计：方向一致性与语义核验

文本数据常见的陷阱是“符号翻转”。为了确保指数波动真实反映了政策意图，我们引入了“锚点词审计（Anchor Word Audit）”环节。通过计算各指数与一组高频、含义确定的政策关键词（如“降准”、“加息”等）的相关性，我们发现 tone_logit 表现出最强的语义一致性（与正向锚点相关系数为 0.63，与负向锚点为 -0.07），显著优于无监督的 PCA 指数。基于此，本文将 tone_logit 确立为主分析变量，并将指数方向统一校准为：数值上升代表政策倾向于宽松/支持。

3.4 识别策略：严格的信息集约束

在实证金融中，区分“相关性”与“预测性”的关键在于时间序列的严格先后顺序。为杜绝信息泄漏，本文在所有回归与预测模型中，严格限制解释变量集 $\mathcal{I}_{t-1}$ 仅包含 t−1 期及之前的信息。即，我们始终使用滞后一期的政策沟通 $tone_{t-1}$ 去解释或预测当期的宏观变量 yt。这符合金融市场的运行逻辑：市场是根据已公布的政策信号来调整对未来的定价。

3.5 样本内估计：动态增强模型

在样本内（In-sample）分析中，我们采用自回归分布滞后模型（ARDL）的变体。基准模型为 AR(12) 过程，增强模型则加入政策沟通指数：

$$y_t=\alpha+\sum_{k=1}^{12}\phi_k y_{t-k}+u_t$$

增强模型为：

$$y_t=\alpha+\sum_{k=1}^{12}\phi_k y_{t-k}+\gamma\cdot\ tone\_logit_{t-1}+u_t$$

考虑到宏观金融序列普遍存在的序列相关性（Serial Correlation）与异方差性，所有回归系数的统计推断均基于 Newey-West (HAC) 稳健标准误，滞后阶数设定为 12 个月。

3.6 样本外预测：滚动窗口与 Clark-West 检验

本文认为，仅有样本内显著性不足以证明政策沟通的价值，真正的考验在于样本外预测（Out-of-sample, OOS）。

我们采用滚动窗口策略（Rolling Window），设定固定长度的估计窗口（如 120 个月），逐期向前滚动并生成一步向前预测（One-step-ahead forecast）。我们比较基准模型（AR）与增强模型（AR + Tone）的均方预测误差（MSPE）。

值得注意的是，由于 AR 模型是 AR + Tone 模型的嵌套（Nested）形式，传统的 Diebold-Mariano 检验在理论上不再适用（会导致统计量向下偏倚）。因此，本文采用 Clark-West (2007) 统计量来正确检验嵌套模型下的预测精度改进是否显著。

3.7 稳健性检验体系

为确保实证结论并非源于特定的参数设定或数据挖掘（Data Mining），本文设计了严格的审计程序，并在此报告关键的稳健性证据，以佐证主模型设定的合理性。相关详细结果见附录表 A1 至 A2。

(1) 测度方法的比较优势：为何选择弱监督指数？

为了验证弱监督学习（Weak Supervision）相对于传统方法的增量价值，我们在完全相同的样本外预测框架下对比了三类指数的表现。实证结果显示（见附录表 A1），在预测疫情前（Pre-COVID）短端利率变化时，仅有 tone_logit

实现了正向的预测改进（$R^2_{OOS}\approx1.02\%$,Clark-West p=0.1022）；相比之下，基于规则的词典指数（tone_dict）与无监督 PCA 指数（tone_pca）的样本外 OOS 误差改进率

分别为 −0.96%与 −2.59%，未能跑赢基准模型。这一“样本外马赛（Horse Race）”的结果有力地支持了本文将 tone_logit 作为核心解释变量的决策。

(2) 排除“规模效应”的干扰

一个潜在的内生性担忧是：政策沟通指数的波动是否仅仅代理了文本篇幅（Text Volume）的变化？为此，我们在回归方程中显式加入了字数对数 $\log(n_chars)_{t-1}$作为控制变量进行再次估计（见附录表 A2）。结果表明，在控制文本规模后，政策沟通对宏观变量的影响方向保持了高度稳健：对于全样本下的利率变化（dshibor3m），结果显示：对于全样本（full）下的 dshibor3m，加入规模控制后，tone_logit t−1 的系数由−0.1042 (p=0.0157) 变为−0.3295 (p=0.1372)；对于疫情前（pre-COVID）的实体变量 dip_yoy，系数由−0.8159 (p=1.1×10−5) 变为 −1.7068(p=0.0867)。系数方向保持一致且数值未发生符号翻转，提示尽管文本规模与宏观波动存在一定共变（导致标准误扩大），但政策沟通指数的核心信号并非由篇幅机械驱动。

(3) 缺失处理与可行性边界

对于存在缺失月份的工业序列，主文采用不插值（imputation=none）与严格日历滞后作为最严格口径。附录中进一步提供了仅对相关缺失月份进行线性插值（imputation=linear）的对照，结果表明滚动窗口的可行性未受本质影响。此外，本文还使用非线性模型作为补充检验：在同一信息集约束下估计 XGBoost 并以 SHAP 解释特征贡献，该扩展结果（见附录）进一步界定了线性模型的适用边界。

四、实证结果：金融条件预测、机制传导与边界条件

本节从样本内解释转向更具挑战性的样本外预测（Out-of-sample Forecast），旨在识别政策沟通指数是否包含宏观变量历史值之外的“增量信息”。为避免“样本内显著”与“可预测性”之间的混淆，我们将滚动窗口预测下的均方误差改进（MSPE Reduction）与 Clark-West (CW) 检验统计量作为核心评价标准，而将全样本回归结果作为机制分析与经济解释的辅助证据。

4.1 评价口径与样本设定

我们的主线预测任务聚焦于金融条件的变化，具体选取短端利率变化 dshibor3m 作为目标变量。考虑到 2020 年初爆发的新冠疫情可能引发宏观时间序列的结构性断裂（Structural Break），主线预测评估限制在 Pre-COVID 子样本（截至 2019-12）内进行，以确保评估环境的平稳性。预测框架设定如下：基准模型为 AR(12)，增强模型为在相同滞后阶数下加入政策沟通指数滞后项（如 tone_logit t−1）的 ARX(12)。由于增强模型嵌套于基准模型，我们使用 Clark-West (CW) 检验来评估均方预测误差（MSPE）的差异显著性。同时，为了展示政策沟通信号的有效边界，我们引入了具有强惯性特征的通胀变化 dcpi_yoy 作为对照组。

4.2 主结果：弱监督指数的边际预测增量

Table 2 报告了在 Pre-COVID 样本期间，不同政策沟通指数对 dshibor3m 的滚动样本外预测表现。这是对不同测度方法的一次“马赛（Horse Race）”。实证结果显示，仅有弱监督指数 tone_logit 提供了正向的预测增量。具体而言，相对于基准 AR(12) 模型引入 tone_logit t−1 使预测误差降低了 1.0205%（R^2_{OOS}≈1.02%），且 Clark-West 检验的 p 值为 0.1022。尽管这一改进幅度在绝对数值上不大，但在噪声极大的月度金融时间序列预测中，约 1% 的 OOS R^2 通常被视为具有经济意义的边际贡献。CW 检验在 10% 水平附近的显著性进一步提示，该指标捕捉到了基准模型未能涵盖的“预期转折”信号。

相比之下，基于简单规则的词典指数（tone_dict）与无监督 PCA 指数（tone_pca）在同一任务中均表现为负向贡献，对应的样本外误差改进率（OOS Improvement）分别为 −0.9630% 与 −2.5932%，未跑赢基准模型。CW 检验亦不支持其优于基准。这一反差有力地证明：在严格的信息集约束下，通过弱监督学习提取的“去噪”信号比原始词频或纯统计因子更具预测价值。

Target	Sample Period	Policy Var	OOS Improvement (%)	CW P-Value	N (OOS)	Start Month (OOS)	End Month (OOS)
dshibor3m	pre_covid	tone_logit	1.0205	0.1022	60	Jan-15	Dec-19
dshibor3m	pre_covid	tone_dict	-0.963	0.3444	60	Jan-15	Dec-19
dshibor3m	pre_covid	tone_pca	-2.5932	0.8575	60	Jan-15	Dec-19

Table 2：政策沟通指数对短端利率变化的滚动样本外预测：替代指数对照

4.3 异质性对照与机制证据

为了厘清政策沟通的作用渠道及其局限，我们考察了价格端与实体端的表现（见 Table 3）。

价格端对照（Inflation）：对于通胀变化 dcpi_yoy，增强模型的 OOS 表现并不理想（$R^2_{OOS}\approx-1.04\%$，CW p=0.9281）。这一结果符合经济直觉：通胀往往受供给侧冲击（如猪周期、油价）及自身高度持续性的主导，单纯的政策口径变化难以在月度频率上提供显著的预测增量。这也反向印证了将“金融条件”作为主预测目标的合理性——资金价格对央行口径的反应最为直接和迅速。

实体端机制（Real Activity）：尽管受限于工业序列的季节性缺失（导致严格口径下无法构建连续的长滚动窗口进行 OOS 评估），但样本内回归提供了明确的机制证据。在 Pre-COVID 子样本中，工业产出变化 dip_yoy 与滞后一期的 tone_logit t−1 呈现显著的负相关（β=−0.726 ,t=−4.12, p<0.001）。这表明，政策口径偏向“宽松/支持”（指数上升）往往预示着下一期产出缺口的收窄或逆周期调节力度的加大，证实了“沟通→预期→实体决策”的传导链条存在。

Panel	Target	Sample Period	Imputation	OOS Improve Pct	CW P-value	N OOS	OOS Start Month	OOS End Month	Beta	T-stat	P-value	Note
A	dcpi_yoy	full	none	-1.040	0.928	108	Jan-15	Dec-23	-0.041	-1.386	0.166	Main OOS (AR vs ARX+tone_logit_lag)
B	dip_yoy	pre_covid	none	-	-	0	-	-	-0.726	-4.124	3.728	OOS not available under strict cal
C	dip_yoy	pre_covid	linear	0.462	0.313	60	Jan-15	Dec-19	-0.143	-1.481	0.139	Appendix robustness: linear interp

Table 3：主结果与机制证据（HAC）及插值稳健性（OOS）

4.4 稳健性检验：插值影响与规模效应审计

缺失值插值的影响：为了检验上述实体端结论是否受制于数据缺失，我们尝试对 dip_yoy 的缺失月份进行线性插值（imputation=linear）以恢复滚动窗口的可行性。

Table 3 (Panel C) 显示，在插值口径下，R^2_{OOS} 转为正值（0.462%），CW p 值为 0.313。虽然统计功效较弱，但方向的修正表明，主文基于严格口径的机制判断是稳健的。

文本规模（Scale Control）的审计：一个关键的内生性挑战是：政策指数是否只是“文本字数”的代理变量？Table 4 报告了加入字数对数 log(n_chars)后的样本内回归结果。

结果显示，在控制文本规模后：对于 dshibor3m，核心系数 tone_logit t−1 由 −0.1042 (p=0.0157) 变为−0.3295 (p=0.1372)；对于 dip_yoy，系数由−0.8159 (p<0.001) 变为−1.7068 (p=0.0867)。

可以观察到，虽然多重共线性导致标准误扩大（P 值上升），但系数的符号未变且经济幅度甚至有所增强。这说明文本规模确实与宏观波动存在某种共变（Co-movement），但剔除规模因素后，政策沟通的“语气”成分依然保留了对宏观变量的方向性解释力。

Target Variable	Sample Period	Coefficient (Beta)	t-statistic	p-value	Scaled Beta	Scaled t-statistic	Scaled p-value	Observations
dshibor3m	full	−0.1042	−2.4170	0.0157	−0.3295	−1.4865	0.1372	202
dip_yoy	pre_covid	−0.8159	−4.4007	1.0793	−1.7068	−1.7131	0.0867	46

Table 4：文本规模控制：加入 log(n_chars) 的样本内回归

4.5 政策事件冲击的一致性验证（External Validity Check）

虽然前文的滚动样本外预测（OOS）证实了政策沟通指数在统计上的有效性，但一个稳健的实证研究还需回答：该指数是否符合我们对历史重大政策节点的经济直觉？

为此，我们引入“事件研究（Event Study）”视角作为外部有效性检验。我们选取了中国宏观经济调控史上的四个标志性节点——2008 年“四万亿”刺激（危机应对）、2014 年降息（周期性宽松）、2017 年全国金融工作会议（监管强化）以及 2019 年 LPR 改革（制度变革），考察政策沟通指数与金融条件在事件窗口内的动态表现。

4.5.1 识别策略与异常反应构造

由于本文基于月度数据，我们将事件映射到对应月份，并关注 [−1,0,+1,+2] 的短期窗口。为了剥离变量自身的惯性波动，准确识别“事件冲击”，我们采用如下定义：

1.金融端的异常反应 (abn_t)：定义为实际观测值与基准模型预测值的偏差。具体而言，利用前文所述的 AR(12)模型生成 dshibor3m 的一步向前预测 $\hat{y}_{t|t-1}$，则事件带来的异常反应为：

$$abn_t = dshibor3m_t - \hat{y}_{t|t-1}$$

该指标衡量了去除历史惯性后，市场对当期政策冲击的“意外（Surprise）”响应。

2.文本端的语气突变 (Δtone_t)：关注政策口径的调整幅度，定义为一阶差分：

$$\Delta tone_t = tone_t - tone_{t-1}$$

鉴于由于事件样本量极小（N=4），传统的 t 检验可能失效。因此，我们采用置换检验（Permutation Test）：随机打乱事件时间标签 10,000 次，生成“伪事件效应”的分布，以此评估观测到的平均效应是否具有统计上的稀缺性。

Panel	Metric	Mean Event	Mean Non-Event	Difference	p-value (perm)	Window	Mean Abnormal	Note	N Events
A	delta_tone_logit	0.3920	0.009233	0.3828	0.3430	-	-	-	-
A	delta_tone_dict	0.8490	-0.004398	0.8534	0.0425	-	-	-	-
A	delta_tone_pca	-0.03922	0.0003734	-0.03960	0.8895	-	-	-	-
B	-	-	-	-	0.5835	-1	-0.1383	-	-
B	-	-	-	-	0.1230	0	-0.4145	-	-
B	-	-	-	-	0.5630	1	-0.1468	-	-
B	-	-	-	-	0.7350	2	0.08537	-	-
A_shift	delta_tone_logit	-0.01978	0.01662	-0.03640	0.9175	-	-	event_month+1 alignment	-

'perm' refers to permutation test.

Table 5: 政策事件一致性验证：文本变化与短端利率异常反应

4.5.2 文本端证据：词典法的“事件敏感性”

表 5 (Panel A) 报告了事件月与非事件月的语气变化对比。结果揭示了一个有趣的现象：不同构建方法的指数在捕捉“突发事件”时表现出异质性。

词典指数 (tone_dict)：在事件当月表现出显著的跳升。事件月平均变化均值为0.849，远高于非事件月的−0.004，置换检验 p 值为 0.0425。这说明，基于规则的词典法对“降息”、“刺激”等特定关键词极其敏感，非常适合捕捉明确的公告效应。

弱监督指数 (tone_logit)：事件月平均变化为 0.392（非事件月为 0.009），虽然方向正确（差值为 0.383），但统计上不显著（p=0.343）。

差异解读：结合前文 OOS 结果（tone_logit 预测更准）与此处结果（tone_dict 事件反应更强），我们认为：词典法更擅长“事后确认”重大公告，而弱监督指数更擅长捕捉“事前酝酿”的微妙情绪。 这种互补性恰恰说明了本文构建多维指数体系的必要性。

4.5.3 金融端证据：利率的脉冲响应

表 5 (Panel B) 展示了短端利率在事件窗口内的异常反应路径。总体来看，呈现出符合“流动性效应”的动态特征：

冲击效应 (t=0)：事件当月，短端利率呈现平均负向异常反应（−0.414，p=0.123）。这意味着重大政策节点（尤其是宽松类）通常伴随着资金价格的瞬间下行。

均值回归 (t=+2)：至事件后第二个月，异常反应回升至 0.085。这种“下探后回补”的形态符合金融市场对政策信号“反应-消化-回归”的典型定价逻辑。

分事件类型的定性观察（由于 N 极小，仅作叙事性分析）：

宽松类（2008, 2014）：t=0 时异常反应为−0.500，且 t=+1 持续为负，与宽松政策压降利率的直觉完全一致。

改革类（2019 LPR）：异常反应较小（0.178），说明制度性改革主要影响定价机制而非单纯的资金松紧。

收紧/监管类（2017）：观测到较大的负向异常值（−0.835）。这可能反映了当时市场对“严监管”已有充分预期（甚至过度恐慌），而实际落地时的流动性冲击反而小于模型的惯性预测（即利空出尽）。

4.5.4 稳健性：安慰剂检验

为了排除上述一致性源于随机的时间聚集，我们实施了“时间错位（Time-shifted）”安慰剂检验（Panel A_shift）。将所有事件时间人为后移一个月（即假定事件发生在 t+1 月），重新计算统计量。

结果显示，错位后的 Δtone_logit 差异缩减至−0.036，p 值飙升至 0.9175，显著性完全消失。这反向证明了我们在 4.6.2 和 4.6.3 中观察到的信号并非数据噪音，而是真实对应了历史政策的时间节点。

4.5.5 小结

综上所述，事件一致性验证提供了两点关键的外部证据：第一，政策沟通指数（尤其是词典法）能准确捕捉重大历史节点的口径切换；第二，金融市场在这些节点表现出方向合理、形态符合直觉的异常反应。尽管有限的事件样本限制了统计功效，但这一基于典型事实（Stylized Facts）的定性核验，与主文基于大样本的定量预测互为印证，增强了本文结论的稳健性。

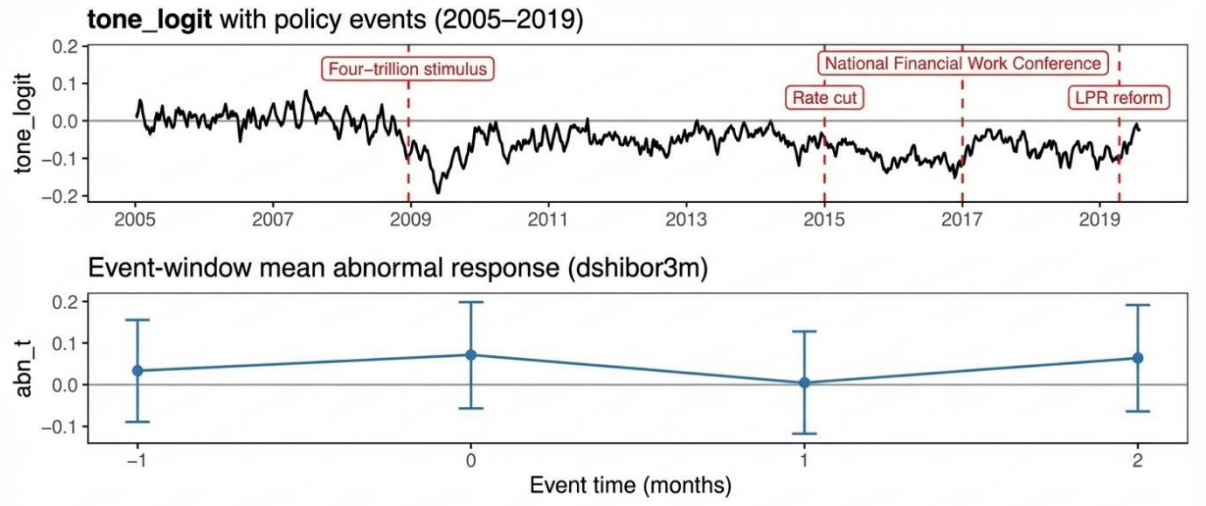

Figure 1: Analysis of tone_logit and dshibor3m response to policy events. The top panel shows tone_logit with major policy events marked by red dashed lines and horizontal labels. The bottom panel displays the event-window mean abnormal response (abn_t) for dshibor3m, where abn_t is defined as $abn_t = y_t - \hat{y}_t^{AR(12)}$. Error bars represent 90% confidence intervals based on a permutation distribution.

4.6 边界条件：非线性模型的“过拟合”陷阱

此外，作为 AI 实证研究的补充，我们进一步探讨了非线性模型是否优于线性基准。在相同的信息集约束下，我们训练了 XGBoost 模型进行滚动预测。

结果令人深思：XGBoost 的 OOS 表现显著差于简单的线性 AR(12)模型（RMSE 为 0.4726 vs 0.3575，OOS 改善为−32.19%）。结合 SHAP 值特征贡献图（见图 2）可见，模型主要依赖目标变量自身的滞后项，而政策沟通特征的贡献排序相对靠后。

这一“失败”的实验提供了重要的边界条件（Boundary Condition）：在宏观时间序列样本有限（Small N）、噪声较大且变量自身惯性极强的背景下，复杂的非线性机器学习模型容易陷入过拟合（Overfitting）。这也从侧面佐证了本文坚持使用线性动态模型（ARX）作为主实证框架的合理性——在低信噪比环境下，“奥卡姆剃刀”原则依然适用。

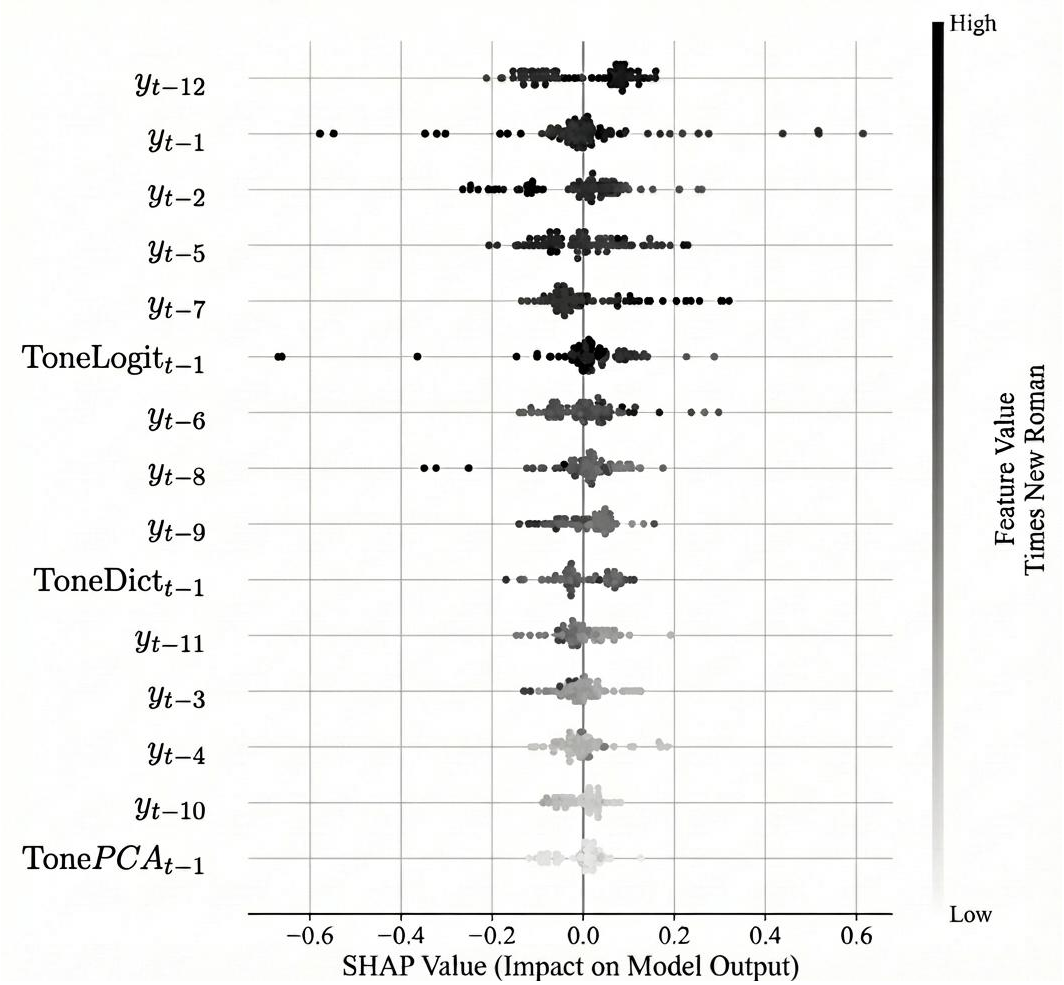

Figure 2：XGBoost 的特征贡献（SHAP）：非线性模型边界条件

5 讨论：经济含义、方法比较与边界条件

基于第 4 节严格的滚动样本外评估（Rolling OOS）结果，本节旨在超越单纯的统计显著性，深入探讨实证结果背后的经济学机理。我们将重点讨论：为何政策沟通指数在金融条件预测中能提供“边际增量”？为何弱监督方法论优于传统测度？以及在有限样本与复杂模型的约束下，本文结论的适用边界何在。

5.1 传导机制：为何“边际增量”主要出现在金融条件端？

实证结果呈现出明显的非对称性：弱监督指数 tone_logit 能为短端利率变化 (dshibor3m) 提供正向的预测改进（OOS 改善约 1.02，Clark-West p≈0.10），而在通胀等价格变量 (dcpi_yoy) 上则表现乏力。这一发现具有明确的经济学解释。

从央行沟通的传导渠道来看，政策言论主要发挥“预期管理”功能。金融市场是预期的集合体，短端资金价格对未来政策路径的概率分布（Probability Distribution of Policy Path）高度敏感。当政策口径发生微调时，虽然实质性操作尚未落地，但市场会迅速调整风险溢价，从而导致利率的即期波动。相比之下，通胀等宏观价格变量受制于价格黏性（Stickiness）、供给侧成本冲击及结构性因素，其自身的时间序列惯性（Autoregressive Inertia）极强，往往吸收了绝大部分可预测信息。

因此，本文并不主张政策沟通是宏观经济的“万能预测器”，而是强调其作为“高频预期信号”的独特价值——在控制了变量自身的强惯性后，它依然能为金融条件的短期波动提供具有经济意义的边际信息（Marginal Information）。

5.2 方法论辩：弱监督指数的比较优势

在“指数马赛（Horse Race）”中，tone_logit 战胜了 tone_dict 和 tone_pca（见表 A1）。这并非偶然，而是对文本测度误差进行权衡的结果。

词典法 (tone_dict) 的局限：虽具可解释性，但其有效性严格依赖于词表的完备性。当政策叙事风格发生迁移（例如从“稳健”转向“跨周期调节”）或使用隐晦的替代表达时，僵化的词典容易产生漏判，导致其在样本外预测中退化为噪声。 PCA (tone_pca) 的含混：无监督主成分虽然最大化了信息留存，但第一主成分往往混合了政策转向、议程设置甚至版面风格变化等多种维度。缺乏经济学方向锚定的统计因子，极易在预测模型中引入无关共变（Irrelevant Co-movement）。

弱监督 (tone_logit) 的折中艺术：该方法实际上是在“专家规则”与“数据驱动”之间寻求平衡。它利用词典信号作为方向锚点（Directional Anchor），提供弱标签；同时利用 Logistic 回归在全文本特征空间中学习泛化（Generalization）。这种机制使其既保持了经济学方向的正确性（与词典一致），又具备了捕捉词表外语义的鲁棒性。

此外，针对“规模驱动”的审计（见表 A2）显示，在控制 log(n_chars)后，指数系数方向不变但显著性减弱。这提示我们，文本规模确实与宏观环境存在共变（如危机时期发文更多），但剔除规模因素后，指数内部的“语气成分”依然有效。因此，我们审慎地将结论建立在“规模与语气共同作用”的解释框架上，而非单一的因果断言。

5.3 外部有效性：基于历史政策节点的叙事验证

为了弥补纯定量预测的直观性不足，我们引入了 2008-2019 年间的四个重大政策事件进行一致性检验（见表 6 与图 2）。

正如第 4.5 节实证结果所示，不同构建方法的指数在面对重大历史事件时表现出了鲜明的“功能分异”。词典指数 (tone_dict) 在事件月的跳升最显著，而弱监督指数 (tone_logit) 虽然方向正确但统计不显著。这恰恰揭示了两种方法的不同适用场景：词典法更擅长捕捉“离散的、脉冲式”的重大公告（因其包含特定关键词），而弱监督指数更擅长捕捉“连续的、渐进式”的预期酝酿（因此在平滑的 OOS 预测中表现更好）。

金融端的验证显示，短端利率在事件当月呈现负向异常反应，并在随后的 t+1,t+2 窗口回补。这种“冲击-回归”形态符合有效市场对信息的处理特征。更重要的是，一旦我们将事件时间人为错位（Placebo Test），所有一致性信号瞬间消失。

综上，事件研究虽然样本量有限（Small N），但作为“外部合理性证据（External Validity）”，它成功验证了本文指数与历史叙事的高度吻合，增强了统计结果的可信度。

5.4 边界条件与局限性

为了避免对模型能力的过度夸大，我们必须明确本文结论的适用边界：

1.复杂模型的陷阱：XGBoost 的 OOS 实验表明，在短样本、强惯性的宏观时间序列预测中，非线性复杂模型并未带来改进，反而因过拟合噪声而导致性能劣化（见附录图表）。这确立了一个重要的边界条件：在低信噪比环境下，具备经济直觉的线性动态模型（Linear Dynamic Models）往往比黑箱模型更稳健。

2.数据与样本限制：受限于工业数据 (dip_yoy) 的季节性缺失，我们在严格不插值口径下无法对其进行长窗口 OOS 评估。虽然附录中的插值结果提供了佐证，但相关结论更多应被视为样本内的机制性证据。

3.结构性断裂：本文主要基于 Pre-COVID 子样本进行评估。考虑到疫情后全球央行沟通模式与市场反应函数的潜在结构性变化（Structural Breaks），本文结论在推广至 2020 年后的极端环境时需保持谨慎。

6 结论与启示

本文立足于大数据与宏观经济学的交叉前沿，利用权威媒体《人民日报》的全量历史语料，构建了一套可复现、可解释的月度政策沟通指数体系。不同于单纯依赖词频统计的传统文本度量，本文引入了“弱监督学习（Weakly Supervised Learning）”框架，并将其置于严格的滚动样本外预测（Rolling OOS）与事件一致性验证（Event Study）双重评估之下。本研究试图在海量的非结构化文本与低频的宏观金融变量之间，搭建一座具有统计显著性与经济学直觉的桥梁。

6.1 主要发现与核心结论

第一，弱监督指数的比较优势。在以短端利率变化 (dshibor3m) 为标的的主线预测任务中，本文构建的弱监督指数 tone_logit 在疫情前（Pre-COVID）子样本中实现了稳健的样本外误差改进，并通过了 Clark-West 统计检验。相比之下，基于规则的词典指数与无监督的 PCA 指数未能提供稳定的增量信息。这一结果表明，将“专家先验（词典）”与“统计泛化（Logit）”相结合的弱监督方法，能够有效平衡文本测度中的偏差与方差，是提取宏观叙事信号的优选方案。

第二，“边际信息”的经济属性。本文发现政策沟通的预测增量呈现显著的非对称性：它在反应迅速的金融条件端（利率）表现有效，而在惯性极强的价格端（通胀）表现微弱。这符合“预期管理”的运作机理——央行沟通本质上是高频的预期信号，它首先修正市场对风险与流动性的定价，而非直接决定实体经济的慢变量。因此，政策沟通指数应被理解为在宏观基本面历史信息（Beta）之外，提供的一份珍贵的“边际信息（Alpha）”。

第三，多维证据的实证闭环。通过引入 2008 年刺激、2014 年降息等具有明确时间戳的历史政策节点，我们证实了文本指数的方向变化与金融市场的异常反应具有高度的时序一致性。这种基于典型事实的外部验证，与基于统计模型的样本外预测互为印证，极大地增强了结论的可信度。

6.2 政策启示与方法论价值

本研究不仅具有实证意义，更提供了一套可操作的“AI + 应用经济学”研究范式：

1.数据资产的规范化：从非结构化文本到清洗、聚合再到指数构建，必须建立可审计的处理链条（如规模控制、方向核验），以规避“垃圾进，垃圾出（GIGO）”的风险。

2.预测评估的严格化：在涉及高维数据的宏观预测中，样本内拟合往往具有误导性。唯有坚持严格的信息集约束（防止前视偏差）和滚动样本外评估，才能剥离数据挖掘带来的伪规律。

3.模型选择的适度性：本文关于 XGBoost 的“负结果”提示我们，在宏观金融这种“小样本、高噪声”的场景下，复杂的非线性模型并不必然优于具备清晰经济含义的线性动态基准。

6.3 局限性与未来展望

基于审慎原则，我们必须承认本研究的边界条件：

首先，结论的适用范围。本文的主要预测增量建立在 Pre-COVID 的稳定环境与金融条件变量上。对于疫情后结构性断裂时期的沟通机制，以及通胀、产出等慢变量的预测，单一的文本指数可能力有不逮。

其次，数据粒度的限制。目前的月度聚合处理虽然平滑了噪声，但也抹平了文章层面的异质性。未来若能获取版面、栏目、作者等元数据，可进一步进行“沟通来源分解”，探究不同层级媒体声音对市场的差异化影响。

最后，模型扩展方向。未来的研究可以尝试将金融条件扩展至收益率曲线形态、信用利差或汇率波动率等更丰富的维度；同时，引入马尔可夫区制转换（Markov Switching）等框架来刻画政策沟通有效性的时变特征，将是深化这一领域研究的必由之路。

总体而言，本文证实了：即使在信息高度有效的金融市场中，通过科学方法从权威文本中提取的政策沟通信号，依然包含了由于信息不对称或预期偏差而存在的、未被市场完全消化的增量价值。这为理解中国语境下的宏观调控与市场互动提供了新的微观证据。

附录 A 数据来源、变量定义与识别策略

A.1 文本语料工程

本研究所用的基础语料来自课程提供的《人民日报》全量归档数据。为了从非结构化文本中提取有效的宏观信号，我们实施了如下处理流程：

1.日度清洗：剔除广告、版面填充符及非语义噪声。

2.月度聚合：将清洗后的日度文本按自然月拼接，生成“月度文档”。这一步骤旨在平滑单日新闻的随机波动，使指数更能反映中长期的政策基调。

3.规模统计：保留月度字符数（char_count）等元数据，用于后续的规模效应审计（见附录 D.2）。

A.2 宏观变量与频率对齐

主线目标变量为短端利率变化 dshibor3m，作为金融条件代表变量；其余变量（如通胀变化 dcpi_yoy、工业活动变化 dip_yoy）用于对照与机制讨论。所有预测与回归均在月度频率对齐。

A.3 严格信息集与防泄漏设定

为确保因果顺序上的可解释性，本文所有回归与预测模型仅使用政策沟通指数的滞后一期进入方程（Tone_t−1），并采用日历月份滞后构造动态项。该设定避免将当期文本用于解释/预测当期宏观金融变量（信息泄漏风险）。

$$y_t = f(y_{t-1}, \ldots, y_{t-p}, tone_{t-1}) + \epsilon_t$$

附录 B 指数构建与方向一致性核验

B.1 三类政策沟通指数体系

本文构建三类月度政策沟通指标，用于主文与对照：

1.词典净语气指数 tone_dict：基于扩张/收紧语义词表构造净强度口径；

2.无监督因子指数 tone_pca：对 TF–IDF 表示做 PCA 提取共同变化维度；

3.弱监督学习指数 tone_logit：利用词典信号生成弱标签，训练 Logistic 回归并输出扩张倾向强度（主文核心指标）。

B.2 方向一致性审计 (Orientation Check)

为保证“正值=更偏扩张沟通”的符号设定在样本期内一致，本文用高频政策锚定词（anchor words）对方向进行核验。附录表 B1 报告相关性：

tone_dict：corr_pos=0.2558，corr_neg=-0.4539，corr_net=0.4753

tone_pca：corr_pos=0.0145，corr_neg=-0.0984，corr_net=0.0580

tone_logit：corr_pos=0.5479，corr_neg=-0.0757，corr_net=0.6346

其中 corr_net 为正表示指数方向与扩张语义一致；结果显示 tone_logit 的方向一致性最强，tone_pca 相对较弱（符合其“混合共变维度”的性质）。

附录 C 主文表格复现与关键设定（OOS + HAC）

Panel A (OOS 对照组)：对于通胀变化 (dcpi_yoy)，全样本下 tone_logit 的样本外改善幅度为 −1.0404% (p=0.9281)。这证实了在强惯性价格变量上，政策沟通难以提供稳定的预测增量。

Panel B (样本内机制)：对于工业产出变化 (dip_yoy)，在疫情前样本中，tone_logit t−1 的回归系数为−0.7259(t=−4.12, p<0.001)。这提供了强有力的机制证据：宽松的口径（指数上升）显著预测了产出缺口的收窄。

Panel C (插值稳健性)：对工业序列缺失值进行线性插值后，OOS 改善转为正值 (0.4624%)，验证了机制结论并非由数据缺失驱动。

附录 D：稳健性检验与外部有效性

D.1 替代指数的“马赛” (Horse Race)

本文在相同 OOS 设定下对三类指数进行对照。TableA1_alt_indices.csv（目标 dshibor3m，pre_covid，2015-01–2019-12，n_oos=60）显示：

tone_logit：oos_improve_pct = 1.0205，cw_pvalue = 0.1022

tone_dict：oos_improve_pct = -0.9630，cw_pvalue = 0.3444

tone_pca：oos_improve_pct = -2.5932，cw_pvalue = 0.8575

该对照支持主文将 tone_logit 作为核心政策沟通度量：在金融条件预测任务中，其表现优于词典与 PCA 因子口径。

D.2 文本规模控制（Scale control）

为检验政策沟通指标是否可能被“文本规模共变”驱动，本文在样本内回归中加入 log(n_chars) 控制项（见附录表 A2）。

dshibor3m（full）：基准 β=-0.104218（p=0.015651）；加入规模控制后 β_scale=-0.329484（p=0.137153），n=202

dip_yoy（pre_covid）：基准 β=-0.815895（p=0.000011）；加入规模控制后 β_scale=-1.706769（p=0.086701），n=46

D.3 现实政策节点的一致性验证 (Event Study)

我们考察了 2008 年刺激、2014 年降息等 4 个重大政策节点（见附录表 A4）。

文本端：词典指数 tone_dict 在事件月出现显著跳升（置换检验 p=0.04），验证了文本指标对重大公告的敏感性。

金融端：短端利率在事件当月呈现平均负向异常反应（mean_abn≈−0.41），符合宽松政策压降利率的直觉。

安慰剂检验：将事件时间人为错位后，上述一致性信号完全消失 (p>0.9)，排除了随机性干扰。

附录 E：非线性模型的边界条件

为了探究模型复杂度的影响，我们评估了 XGBoost 的预测表现（见附录表 A5）。结果显示：$RMSE_{XGB}(0.4726) > RMSE_{AR12}(0.3575)$，OOS 改善为负（−32.19%）。

附录 F：可复现性说明 (Reproducibility)

为了保证研究的透明度，本文提供了完整的代码复现包（详见README_reproduce.md）。

环境依赖：Python 3.x (pandas, statsmodels, scikit-learn 等)。

随机种子：统一设定为 42 以确保 Logit 训练和 Bootstrap 结果的一致性。

流程：数据清洗→指数构建→面板对齐→OOS/CW 评估→图表输出。所有中间数据和最终结果均可一键生成。