

《基于SparkDesk大语言模型的Prompt引导碎片式虚假新闻检测研究——以“胖猫事件”为例》

摘要：【研究目的】针对社交媒体中碎片化短评的虚假信息检测难题，本文以“胖猫事件”为典型案例，旨在探索大语言模型（SparkDesk）结合Prompt提示工程在微博评论三分类——“真实”“虚假”“信息不足”——任务中的应用效果，以构建一个高度可迁移、低成本且具备强解释性的轻量级检测方案。【研究方法】项目提出一种融合权威知识库与知识检索增强（RAG）的多轮Prompt迭代框架，并在199条人工标注数据与9246条大规模真实评论上开展采用包括准确率、Cohen's Kappa、“蓝V语境符合度”等指标的评估实验。【研究结论】动态RAG检索+多轮Prompt迭代能显著提升模型对碎片化评论的语义对齐与事实核对能力，具备即时响应实时舆情且能适应少样本学习环境的双重优势。

关键词：大语言模型；碎片式虚假新闻；Prompt提示工程；RAG检索；舆情自动化分析；网络舆情

一、引言

在信息技术与移动终端普及的背景下，社交媒体平台如微博、抖音、知乎等，已逐渐演变为公众表达情绪、传播信息与参与公共议题的重要舆论场。这一传播生态的去中心化、自发性和情绪驱动特征，使得网络舆情在时效性和互动性方面取得前所未有的扩展。然而，这也加剧了信息环境的复杂性：评论内容呈现出高度碎片化、情绪化，真假信息交织混杂，事实与立场往往难以分辨。

以“胖猫事件”为代表的舆情突发案例，便突显了社交平台信息传播的裂变特征。事件引发大量围绕法律事实、性别议题、情感责任等方面的讨论，形成了典型的“话题集群扩散”现象。部分评论对事件进行了事实复述与理性探讨，但同时也伴随着大量未经证实的传言、煽动性语言和“带节奏”标签，构成了网络舆情中的“碎片式虚假新闻”——这类信息并非完整报道，而是以评论、转发语、片段式指控等形式存在，极易误导公众判断，干扰政策响应，甚至引发次生舆情危机。

面对海量、非结构化、实时更新的社交媒体文本，传统的人工研判方式在效率和准确性上均难以胜任，亟需引入具备理解语义、推理逻辑与任务泛化能力的智能技术手段。大语言模型（Large Language Model, LLM）以其强大的语义建模和Few-shot推理能力，展现出在舆情分析、信息筛选与虚假检测中的广泛应用潜力。特别是在无需大规模训练的前提下，LLM 可通过**提示工程（Prompt Engineering）**直接接受特定任务指令，在碎片信息识别与倾向性语言判别中具有独特优势。

本文立足于**大语言模型**在虚假新闻识别场景中的新型能力，聚焦“**碎片式虚假信息**”检测任务，构建了一套以 Prompt 为核心驱动机制的自动化识别流程。通过引入**权威知识库**、**人工标注样本**、**混合知识检索与评论分类标准**，利用 **Sparkdesk** 模型对“胖猫事件”相关微博评论进行三分类判断（真

实、虚假、信息不足），旨在探索LLM在现实复杂舆情场景中识别误导性言论、提升舆情治理质量的可行路径。

1.1 案例引入

“胖猫事件”是2024年中国网络舆论场的典型热点之一。该事件因其涉及情感关系、金钱交易、女性权益等敏感议题，迅速在微博平台引爆大量讨论。在事件演化过程中，不仅有媒体报道、警方通报等官方信息流传，也出现了大量民间账号的评论、引申与改写。这些文本往往表现为“断章取义”“推测定性”“情绪放大”等形式，形成大量碎片化信息，其真假难辨却影响极广。由于这类言论通常非正式发布、未形成完整逻辑，但极具感染力与传播力，因此具有典型的“碎片式虚假新闻”特征。

1.2 研究动机与目的

近年来，大语言模型在自然语言理解、文本分类、推理与情感分析等任务上展现出前所未有的性能优势，特别是在小样本学习和语境推断方面，其Few-shot推理能力使其在高复杂度、弱标注条件下表现优异。借助Prompt工程，研究者可无需模型训练，直接引导大模型完成特定任务判断。

基于此，本文尝试构建一个结合**权威知识库+人工标注样本+任务引导Prompt**的检测系统，对“胖猫事件”相关微博评论进行三类识别：①与官方信息一致、表达理性的“真实”；②与官方结论冲突、具有误导倾向的“虚假”；③事实不明、语义模糊或缺乏判断依据的“信息不足”。本研究旨在为碎片化社交媒体文本的虚假识别构建一套可复用、可部署、可评价的智能化识别流程，并通过实证验证其在提升舆情治理效率与信息筛选质量方面的可行性。

项目目标不仅是识别“虚假”与“信息不足”的评论，更是从中筛选出**可供政府舆情治理决策参考的“真实”与“建设性意见”**。拟解决“信息过载冗余”、“情绪化与非理性表达”、“伪装账号带节奏”等网络评论中的“虚假”成分问题，辅助政府在应对舆情时做出**快速、准确、可信的反应**，推动舆情治理向智能化、精准化转型。

1.3 研究问题

结合“胖猫事件”语境与社交媒体舆情特征，本文主要围绕以下问题展开探讨：

Q1：如何基于权威知识库与Prompt提示工程，构建SparkDesk对微博评论“真实—虚假—信息不足”三分类的结构化识别体系？

Q2：大语言模型在处理带有情绪色彩、事实模糊或刻意伪装的微博评论时，能否稳定实现内容归类与风险识别？

Q3：Prompt结构是否可通过版本控制与示例迭代优化实现性能增强？该方法在无标注大规模数据上是否具有实际部署价值？

二、研究背景

2.1 文献综述

随着社交媒体平台的普及与信息传播速度的成倍增长，“碎片化”舆情内容——尤其是评论、转发语、短句讨论——已成为虚假信息传播的主要载体。面对“胖猫事件”引发的舆论风暴，传统方法愈暴露出在实时性、细粒度与情绪化表达识别方面的局限。基于此，本文在承接已有虚假新闻检测和舆情治理研究的基础上，重点引入 **Prompt-based Few-shot Inference** 与 **Prompt 工程（Prompt Engineering）** 理论，实现对微博评论的自动化三分类判断。

2.2.1 虚假新闻检测方法研究

传统的虚假新闻检测方法主要依托于机器学习和深度学习等技术。机器学习方法包括支持向量机、逻辑回归、随机森林等，其核心在于特征工程，即提取文本的词汇特征、句法特征、语义特征及传播特征，并利用有监督分类算法进行真假判断。然而，这些方法在文本异构性增强与数据规模急剧扩大的背景下，难以适应不同领域虚假内容的泛化需求。

随着深度学习的发展，CNN、BiLSTM、HAN 等模型引入文本上下文建模机制，在一定程度上提升了语义捕捉能力。然而它们通常依赖大量标注数据，面对社交媒体中表达模糊、上下文缺失的短文本内容效果仍有限。

近年，研究重心开始向融合外部知识的建模策略倾斜，如引入知识图谱、图神经网络（GNN）以及结构化因果建模等，进一步增强模型对语义偏离与信息冲突的识别能力（Zhou et al., 2020）。但这些方法普遍存在部署成本高、迁移性弱的问题，特别是在快速爆发的新事件场景下难以落地。

2.2.2 大语言模型在虚假信息检测中的角色转变与 Prompt 工程

随着 ChatGPT、GLM、LLaMA 等大语言模型的兴起，虚假信息检测出现范式性变革：由“训练+分类”的封闭模式转向“Prompt+推理”的开放交互模式。基于提示工程（Prompt Engineering）的方法，不再依赖模型再训练，而是通过结构性任务引导，大幅降低了对标注数据的依赖，具备高度迁移性与任务适配能力。

王润周等提出将大模型与混合检索增强生成（RAG）架构结合，通过构建外部知识上下文，提高大模型判断虚假信息时的事实对齐能力。潘磊等通过大语言模型构建因果事件图，实现了对事件链条的溯源分析与误导信息定位，验证了 LLM 在复杂信息网络中建构语义关系的能力。

特别是在碎片化舆情场景中，Prompt + Few-shot 推理成为一种可行的**轻量级任务适配机制**，对动态、细粒度的虚假信息识别具有独特优势。

2.2.3 评论类与碎片化虚假新闻识别研究

与传统整篇新闻识别不同，社交平台中大量“虚假信息”表现为评论、转发语、断句指控、标签操控等非结构化表达，即所谓“碎片式虚假新闻”（fragmented fake news）。这种形式的谣言在认知心理学上往往更具误导性，在信息传播路径中具备“二次加工、立场注入”的特点，研究难度更高。

Shu et al. (2019) 在其综述中指出：**社交评论和用户行为文本是虚假新闻传播的关键载体**，对这些“次级内容”的识别能力直接决定了平台的抗谣言能力。Kaliyar et al. (2021) 提出 FakeBERT 模型，用于检测社交平台中带有主观色彩的短文本，强调上下文缺失环境下的微语义建模。Chen et al. (2022) 明确指出，针对“评论型虚假信息”的识别，传统分类器的表现大幅低于基于语境理解的大模型模型。

目前，一些研究开始探索结合大语言模型和社交语料的方式，对评论类信息进行“真实/虚假/不确定”三分类（García-Durán et al., 2023）。特别是在多情绪、多立场共存的表达下，模型需能准确识别语义漂移、话语操控与事实背离等特征。

因此，在真实舆情事件中构建评论级别的三分类机制（真实/虚假/信息不足），不仅可实现信息可信度重构，也为碎片式虚假新闻检测提供了任务框架支持。

2.2.4 舆情治理与大模型驱动的信息筛选机制

舆情治理从“信息监测”向“语义甄别”转型，是近年来智能治理体系的重要方向。传统的舆情研判依赖人工经验与关键词规则，面对大规模社交平台文本、情绪扩散机制与异构传播结构时显得力不从心。

梁炜等指出，大语言模型可嵌入舆情生命周期，实现情绪突变识别、风险传播路径预测与回应建议自动生成。陈健等则明确将“碎片化误导信息”作为舆情治理中最具风险的干扰源，强调模型对引流性、攻击性、误导性评论的精准甄别能力。

结合前述评论型虚假信息特征，在舆情治理场景中实现评论级别的“可用/误导/模糊”筛选机制，是当前虚假信息识别的重要现实落点。Prompt 驱动的 LLM 能力为构建这一机制提供了技术基础。

2.2.5 文献述评与研究定位

当前学界研究已逐渐转向认识到，虚假新闻不再局限于完整文章形式，而是以碎片化语言广泛嵌入评论、转发与互动语料中；评论类虚假信息更隐蔽、更具有情绪操控性，成为平台治理与风险识别的重要挑战；大语言模型具备推理能力和上下文建构能力，在碎片语境中可借助 Prompt 实现轻量化的判断机制；当前仍缺乏专门面向“评论型碎片虚假信息”的三分类任务机制、系统的 Prompt-engineering 驱动方案与可迁移实现路径。

因此，本文基于“胖猫事件”这一典型舆情案例，提出了一种面向微博评论的自动化真实性识别框架：该框架融合了权威知识库、少量人工标注、多轮 Prompt 迭代（包括 Chain-of-Thought 与 RAG 检索技术）、SparkDesk 大语言模型推理以及多维度评估指标（Accuracy、Cohen's Kappa、混淆矩阵和蓝 V 语境符合度），从而构建出一个高度可迁移、低成本且具备强解释性的轻量级检测体系。在已有成果基础上，进一步结合舆情治理中的“建设性评论识别”与“情绪噪声剔除”需求，探索大语言模型在具体复杂网络舆情中的实际应用边界。

2.2 概念界定

2.3.1 碎片式新闻与社交语境下“新闻”概念的扩展

在传统语境中，“新闻”一般指由新闻机构或专业记者通过正式渠道发布的、经过编辑审查的事实性报道。然而，在网络舆情治理场景下，特别是在社交媒体平台（如微博、抖音、小红书）成为信息传播主阵地的背景下，“新闻”的边界正在发生扩展。公众不再仅仅通过传统媒体获取信息，越来越多的信息来源于网络平台上经由自媒体、意见领袖、公众账号转发和评论的信息碎片化集合。

有研究指出，在社交平台中，网络用户对信息“是否来自媒体”与“是否具有公共影响力”的判断标准正在融合（梁炜，2025；陈健，2024）。也就是说，具有高关注度、高互动量、并引发广泛讨论的社交媒体内容，也在事实上承担了“新闻”的社会功能。尤其在舆情高峰期，微博等平台中非传

统媒体账号所发布的信息，往往比传统新闻更早、更广传播，成为公众认知事实、判断态度的重要依据。

在此基础上，本文引入“**碎片式新闻**”的概念加以区分与界定：**碎片式新闻**指的是在社交媒体环境中，以评论、转发语、标题控评、图文组合等“非完整报道”形式出现的信息单元。其常见形式包括断章取义式引用、模糊表述性指控、极端立场标签化表达等。虽然缺乏系统结构，但往往具有高度传播力、情绪动员力和舆论引导效应，是当下网络舆情传播的关键形态之一。

基于以上界定，本文在研究“胖猫事件”中虚假新闻检测时，对“新闻”概念进行**扩展性界定**：不仅包括由认证媒体（蓝V账号）发布的博文，也包括由具有一定粉丝规模、舆论影响力的认识个人账号（如红V、黄V、金V）发布的**观点强烈、表达明确、互动活跃的舆情内容**。这类内容在事件演化过程中具有显著的舆论引导或激化作用，已成为网络舆情生态的重要组成部分。

该定义兼顾传播力、表达性和账号权重，有助于在数据上明确“可检测”对象，为后续模型判断建立清晰样本边界。

2.3.2 虚假新闻的类型化界定

虚假新闻（Fake News）在本研究中并不局限于传统意义上的整篇虚构报道，而是聚焦于**碎片化传播链条中的误导性表达**。结合社交评论语料的表现特征与权威知识支撑体系，本文将“虚假新闻”界定为：在语义上与权威知识库中已有结论存在明显冲突，或在内容表达上严重偏离事件核心，以夸张、误导、断章取义、操控情绪等方式构造和传播的网络文本信息。

此类内容通常缺乏事实依据，或通过扭曲叙述方式引导公众误读事件本质，易在网络空间中形成舆论偏差、激化情绪对立，甚至误导政策响应方向。

在本研究中，虚假内容主要包括三种主要表现形式，其一为**伪信息**，捏造或歪曲事件关键事实（如时间、地点、人物关系、金额数据等），与权威通报信息发生直接冲突；其二为**情绪放大内容**，以煽动性语言、情绪化修辞强化指责、同情、对立等立场表达，但缺乏事实支撑或理性论据；其三是**引流类评论**，借助热点事件博取流量，发布与事件无实质关联的信息，通过蹭热度、制造争议等手段干扰正常舆情秩序。

值得注意的是，真实不等于主流立场，虚假不等于负面情绪；判断标准依据是否具备**可核实的事实支撑**、是否偏离事件核心信息与是否只有情绪攻击性等。

2.3.3 权威知识库的构建逻辑与功能定义

社交平台的开放性导致信息源极度异质，真假难辨、情绪失控、立场冲突成为网络舆情治理的常态。为提高大语言模型在识别碎片式虚假信息过程中的判断准确性与一致性，本文构建以**事实核查为核心的“权威知识库”**，作为Prompt输入的事实背景支撑，是提升虚假信息检测准确性的重要手段。

在“胖猫事件”中，官方机构发布的定性通报、警方权威调查结论、主流媒体（如央视、新华社）发布的深度报道，构成了事实层面的统一叙述，是当前事件中少数**可以信赖的权威源头**。这些内容包括但不限于：事件发生时间与地点；当事人之间的经济往来明细；官方对事件定性的通报内容；对网络传言的回应或澄清；政策层面的法律适用或舆情引导措辞等。

本文将上述内容进行提取、梳理与整理，构建为“权威知识库”，作为大语言模型执行真假判断时的核心参照标准。该知识库的作用主要体现在三个方面：**事实对齐**：用于判断网络新闻中是否存在与权威结论相矛盾的信息，如错误时间描述、虚构人物关系、故意误导结论等；**相关度评价**：用于评估评论或博文是否在事件主题范围内，是否存在明显的“内容漂移”或“蹭热度引流”等行为；**情绪偏差识别辅助**：当内容高度情绪化但缺乏事实支撑时，可结合权威信息辅助识别其虚假倾向。

三、研究设计

为应对“胖猫事件”引发的社交媒体舆情信息爆炸，传统舆情分析方法在处理实时性、复杂性与情绪化表达方面存在明显瓶颈。近年来，大语言模型（Large Language Models, LLM）在文本理解、推理、事实核对与观点分类等任务中展现出卓越表现，因此，本文选择以**Prompt-based Few-shot Inference**（基于提示的少样本推理）作为核心分析范式，构建针对微博评论真实性的自动化判断框架。

本研究提出一种以SparkDesk大语言模型为核心、Prompt提示工程为引导机制的自动化三分类判断框架，主要包含数据准备、多轮Prompt迭代优化、大语言模型推理调用与多维度评估验证四个核心模块。相较于传统机器学习方法，本研究所设计的框架具有如下显著优势：

高适应性：无需大规模数据微调，即可快速响应实时舆情；

低成本：依托Few-shot提示与API调用，减少了人工标注和训练的开销；

强解释力：模型输出包含“分类”与“推理理由”，提高了分类结果的透明性；

可拓展性：框架具备高度的迁移性，可迅速推广至其他热点舆情分析场景。

本研究围绕“碎片式虚假新闻”的识别问题，构建了由“事实卡片 + Few-shot示例 + 用户输入评论”组成的组合式Prompt，实现在无需微调模型的情况下，自动对网络评论进行真实性识别。为进一步提升分类的精度和实际应用性能，本研究设计了三轮Prompt结构优化机制（Prompt V1-V3）。在此过程中，通过结合Embedding RAG（知识增强检索）与动态Prompt构造，对错误类型和任务结构进行持续分析与优化，有效增强了模型的边界识别能力与表达一致性。

在完成模型结构与Prompt优化后，为验证其在真实舆情场景下的部署可行性，本研究将最优版本Prompt V3应用于包含9246条微博评论的大规模数据集。同时，本研究建立了如“蓝V语境符合度”等语义一致性评价指标，从权威内容识别与误导性评论过滤两个维度，验证模型的稳定表现，最终形成如下闭环流程：

数据收集 → 权威知识库构建 → 样本标注 → 动态Prompt设计与迭代 → SparkDesk推理调用 → 分类输出 → 多维度评估验证 → 大规模部署应用

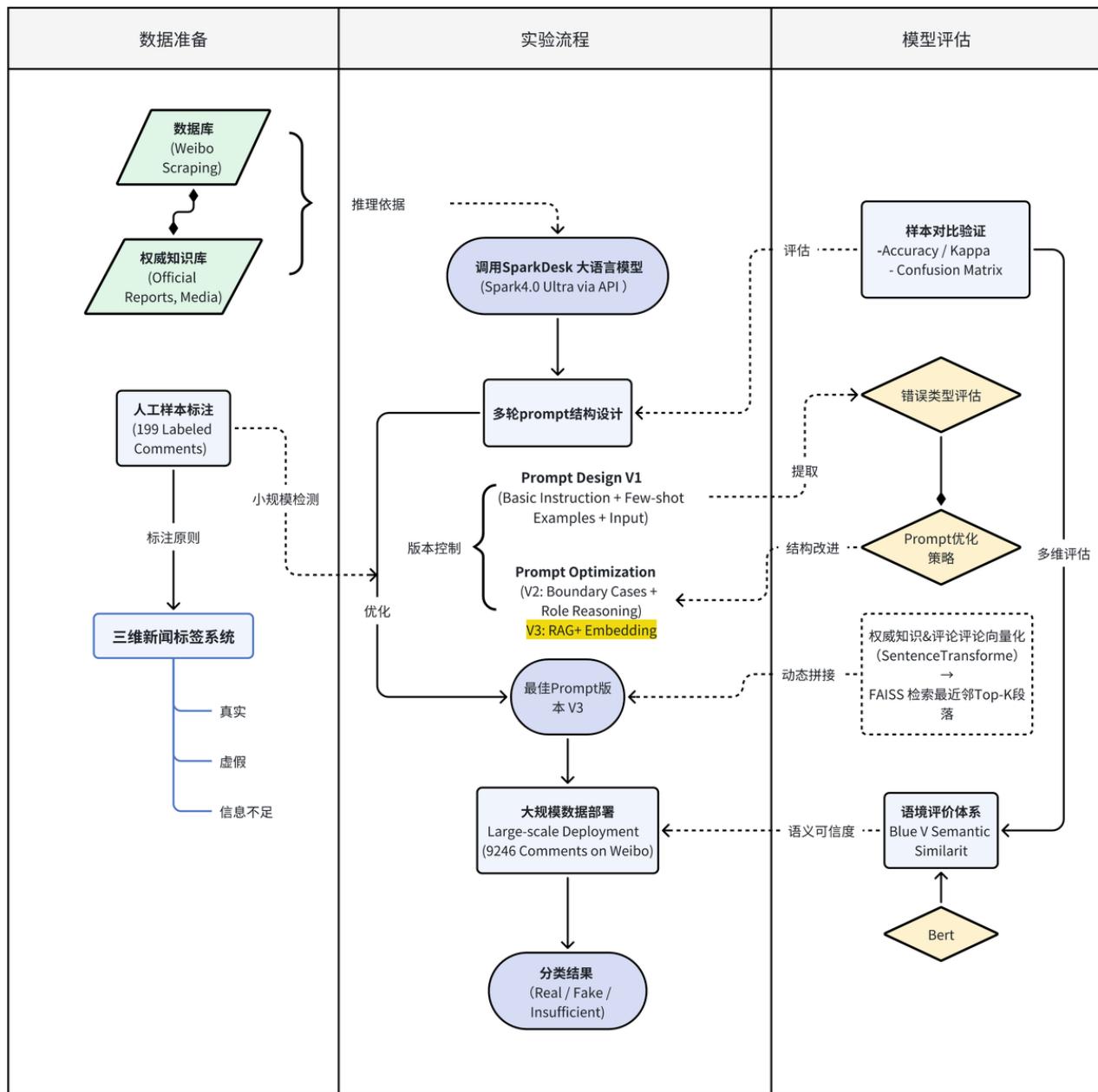


图1 流程设计图

3.1 数据库与权威知识库

3.1.1 数据采集与预处理

数据采集与分类

本研究所使用的数据来源于微博社交媒体平台，选取“胖猫事件”作为特定舆情案例。为高效、系统地获取相关博文与评论信息，本文调用GitHub上开源的微博话题爬虫项目

(<https://github.com/dataabc/weibo-search>)，结合Scrapy框架进行关键词式内容抓取。通过设定关键词（“胖猫”、“胖猫事件”、“警方通报胖猫事件”等）、时间范围、账号认证类型、转发数和点赞数等条件，共获得9246余条微博评论。采集涵盖广泛舆论反应的微博数据，构建具有代表性的碎片式新闻代表集。

爬取微博/话题	爬取阶段	爬取时间	有效数据
#胖猫事件# #21岁重庆男子跳江#	“胖猫事件”前期--官方介入前	2024.5.2-2024.5.11	3042
#警方通报胖猫事件##胖猫通报里的三个关键点##胖猫案细节公布#	“胖猫事件”后期--5.19 官方介入后	2024.5.19-2024.5.28	3050
#胖猫事件90多吨食物被浪费##记者回访胖猫事件#	“胖猫事件”新回应--2025.3.15报道后	2025.3.15-2025.3.20	3154

表1” 胖猫事件 “微博数据爬取来源结果表

所爬取的字段信息包括：微博ID、评论ID、文本内容、发布时间、点赞数、转发数、评论数、评论者关注数、粉丝数、账号认证类型（`user_authentication`）等。按照微博认证类型结合热度影响力等权重，将数据划分为媒体类（蓝V、金V账号-17.2%）、民间意见领袖类（红V、黄V账号-22.4%）与民众类（普通用户-60.4%）。

Distribution of Comments by User Authentication

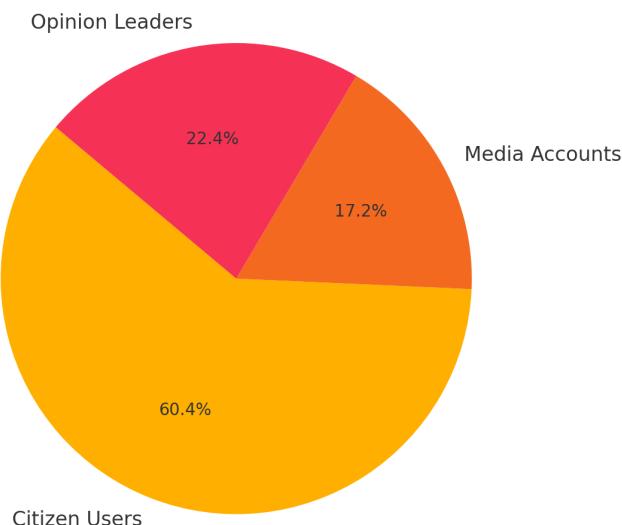


图2 数据用户类型分类占比图

数据清洗

由于社交媒体文本具有强烈的个体表达特征，存在大量重复、情绪化、结构混乱甚至无关噪音信息，故在样本筛选后对数据进行了系统的文本清洗处理。清洗流程包括：删除文本中的URL链接、HTML标签、表情符号、特殊字符与广告语；统一语料格式，去除冗余空格、制表符、换行符等格式噪音；筛除长度过短、内容明显无关的异常评论或机器生成文本。

3.1.2 权威知识库构建

语料采集

为了提高大语言模型在“胖猫事件”舆情识别中的事实判断准确性与上下文理解能力，本文基于权威信源整理构建了一个面向特定事件的**舆情知识库**，用于支撑后续大语言模型对网络评论的“真实 / 虚假 / 信息不足”三分类判断。

借助第三方舆情监测平台“知微数据”（Zhiwei Data）对“胖猫事件”进行事件发展全过程回溯与传播节点梳理，识别出舆情爆发、情绪拐点、官方回应、媒体反转等关键时间节点。通过该平台所提供的事件热度曲线、媒体分布图、传播路径网络图等辅助工具，本研究明确了事件的核心传播阶段与话题演化路径，从而为知识库构建提供时序线索与语境边界。



图3”知微数据“-重庆警方通报“胖猫”跳江事件趋势图

在此基础上，本文从众多公开报道中精挑细选出三份高可信度的**权威语料**，构成该事件的核心知识源，具体包括：警情通报（重庆市公安局南岸区分局，2024年5月19日）；中央广播电视台总台专访报道（2024年5月24日）；央视专题回访舆情调查报道（2025年3月15日）。

prompt事实卡片

上述三类文本在语义表达、数据来源与时效性上均具备高度权威性，具备构建事件“信息基准线”的能力。经过文本清洗与段落级整理后，本文将这三份语料中涉及的**事件时间线、当事人关系、经济数据、官方态度、事实争议点与结论性语言**提取为结构化“事实卡片”，作为大语言模型进行判断时的Prompt知识补充。

类别	内容摘要
【事件基本信息】	“胖猫”（刘某，男，21岁，湖南人）于2024年4月11日凌晨在重庆长江大桥跳江身亡，排除刑事案件。
【关系背景】	自2021年底起，“胖猫”与谭某以真实身份恋爱两年多，互见家长并共同生活，2023年底共同创业花店。
【经济往来】	两人互转账共计百万元，属正常恋爱期间经济往来，公安机关明确“不构成诈骗”。
【警方定性】	谭某不存在虚构感情或诈骗行为，刘某发布隐私内容并组织网暴，警方已介入并依法查处。
【网络谣言】	网上广泛传播的“谭某骗钱”“胖猫被PUA”“警方包庇”等均系失实谣言；部分网友恶意造谣、假冒视频、煽动转账辱骂行为被依法处罚。
【社会后果】	事件引发网络暴力，造成谭某工作生活严重影响，重庆全市被迫出动大量行政资源清理外卖、加强巡逻；部分外卖祭奠浪费食物近百吨。
【公安立场】	重庆警方坚持“公开透明、以证据说话”，通报中每句表述均有证据支撑，接受历史与公众监督。
【媒体观察】	舆论场被“标签化叙事”主导，部分营销号推动“人设塑造+引流牟利”产业链，造成重大信息偏差。

表2”事实卡片“示例

3.2 人工标注机制与验证流程

为确保模型分类判断具备现实语义标准与事实依据，本文构建了一套系统化的人工标注机制与验证流程，结合“胖猫事件”微博评论样本，建立高质量、可复用的标注数据集，并以此作为后续Prompt优化与模型评价的标准参照。

3.2.1 标注原则与标签体系

本研究以“可用于政府舆情治理的真实与建设性意见甄别”为目标，制定人工标注体系，构建了199条高质量标注样本。标注工作严格以权威知识库为依据，围绕“事实一致性”“情绪强度”和“表达完整性”三个核心维度，将微博评论划分为三类标签：

真实 (Real)：评论内容与权威知识库中核心事实一致，表达理性客观，具备政策参考价值和公共建设性。

典型特征：转述警方结论、事实补充、理性建议；

简要说明示例：事实转述 / 理性建议 / 高度一致。

虚假 (Fake)：评论存在明显事实错误、虚构情节或引导性强的情绪化倾向，与知识库中的事件基线冲突，具有误导性。

典型特征：歪曲时间、金额、人物关系，煽动性修辞；

简要说明示例：虚假数字 / 情绪宣泄 / 引流信息 / 伪装叙述 /

信息不足 (Insufficient)：评论虽部分信息属实，但表达模糊、判断不明或缺乏关键信息支撑，难以准确归类。

典型特征：语义模棱两可、无明确立场、语义混杂；

简要说明示例：观点表达 / 混合信息 / 暂不可判。

该三类标签体系不仅具备现实场景解释性，也为Prompt设计提供了可结构化示例支持。

3.2.2 标注执行流程

人工标注过程采用分步骤执行方式，确保标注结果的客观性与可复用性。

步骤一：样本筛选与抽样策略

利用“user_authentication”字段筛选认证账号（蓝V、红V、黄V、金V）评论，辅助评论量转发量等热度参考，作为“碎片式新闻”语料来源。样本涵盖媒体类与民间意见领袖类账号，确保多样性与代表性。从中随机抽取199条评论，构成小规模测试数据集。

步骤二：标注规范与执行机制

标注人员使用结构化Excel模板逐条判断样本，填写两项内容：①“标注结果”（真实 / 虚假 / 信息不足）；②“简要说明”（关键词或理由说明）。每条评论均需明确标注依据，并标注对应关键词以辅助分析。

步骤三：知识对齐与一致性校验

标注前，执行者需熟读事件权威知识卡片内容（如重庆南岸警方通报、央视报道等），统一判断标准。标注过程中严格参照知识库事实框架，结合评论的内容重心、事实支撑度、语气强度与建设性进行归类。部分样本由两人独立标注并校对一致性。

3.2.3 验证方法与评估指标

为系统评估大语言模型分类性能，本文将该199条人工标注样本分别输入三轮Prompt版本（Prompt V1/V2/V3），记录**SparkDesk**输出结果并与人工标签进行对比，评估模型判断的准确性与一致性。采用如下指标：

准确率（Accuracy）：模型预测与人工标注完全一致的比例；

Cohen's Kappa值：用于衡量模型预测结果与人工标签之间在排除随机因素后的契合度；

混淆矩阵：统计各类标签判断混淆情况，辅助识别误判模式。

该验证体系兼顾定量统计与误差诊断，有助于模型行为调优与Prompt设计改进。

3.3 Prompt引导结构设计

为了应对碎片化、情绪化高度交织的网络舆情环境，本研究提出了一种多轮Prompt优化策略，旨在提高大语言模型在对“胖猫事件”相关评论进行虚假新闻分类时的准确率和一致性。该策略通过逐步强化Prompt中所包含的信息量、角色引导和背景知识输入，从而引导模型从单纯的浅层文本匹配，转向深入语义边界的识别与逻辑自洽的推理。通过多轮Prompt优化策略，本文从最初基础模板

（Prompt V1）逐步引入角色设定和自省引导（Prompt V2），再到结合知识增强检索的动态Prompt构造（Prompt V3），形成了一个完整的优化流程。各个版本的设计差异如下：

Prompt V1 测试模型的基础判断能力；

Prompt V2 强化推理链条与细致自省，改善模型对情绪性和表达模糊信息的辨识；

Prompt V3 利用RAG策略实现定制化动态背景补充，显著提高模型对复杂评论的事实关联识别能力和整体分类准确性。

3.3.1 Prompt基本构成

每轮Prompt均由以下三个主要部分构成：

权威知识背景（Knowledge Card）

此部分主要来源于警方通报、央视报道以及新华社文章，通过结构化的方法提炼出核心事实卡片。事实卡片内容通常包括事件的时间线、关键人物关系以及定性结论等信息，同时整合版的权威话语料也被加入其中，以保证模型获得高可信度的背景知识。

Few-shot示例

从人工标注样本中筛选出3-5条代表性评论，这些示例涵盖“真实”、“虚假”以及“信息不足”三种类型，并为每条示例附带标注结果。Few-shot示例为模型提供标准参考案例，有助于模型建立任务内部的判断标准和分类依据。

待判定评论输入

将实时抓取的微博评论以原始文本的形式拼接在Prompt末尾，要求模型基于以上提供的信息输出相应的判断结果。严格规范不同的prompt版本下模型输出格式。

3.3.2 多轮Prompt结构优化（静态）

在传统的单轮Prompt中，面对复杂、碎片化且情绪化的评论，简单指令往往难以捕捉潜在的表达模糊性和多样化的伪装方式。因此，本研究采用了多轮Prompt优化策略（Multi-round Prompt Refinement Mechanism），其优化过程分为以下三个阶段：

Prompt V1：基础任务引导模板（Baseline）

V1版本作为初始基线模板，仅提供任务说明和少量Few-shot示例。该版本不针对特定角色进行设定，也没有针对情绪宣泄或表达模糊做额外引导。其主要目标是测试大语言模型在面对基本事实判断任务时的原生能力，为后续优化提供性能参考。

Prompt V2：多轮推理角色设定版（任务背景 + 自省引导）

在V1的基础上，V2版本增加了“公安舆情审查专员”角色设定，并明确要求模型不仅给出分类结果，还需输出详细的推理理由和判断依据。进一步嵌入了自省提示（如“请再次确认此评论是否存在误导倾向”），以促使模型对边界性表达进行二次审视。这一增强措施旨在使模型在识别复杂语境下伪装性信息时，能形成更稳定的判断偏好和更充分的可解释性。

3.3.3 Prompt V3：RAG动态Prompt

Prompt V3在V2版本的基础上进一步引入了动态知识增强策略，即RAG（Retrieval-Augmented Generation）。其主要设计思想在于利用实时检索获取与待判定评论最相关的权威知识背景，动态构造Prompt，从而使模型判断更具针对性和事实依据。

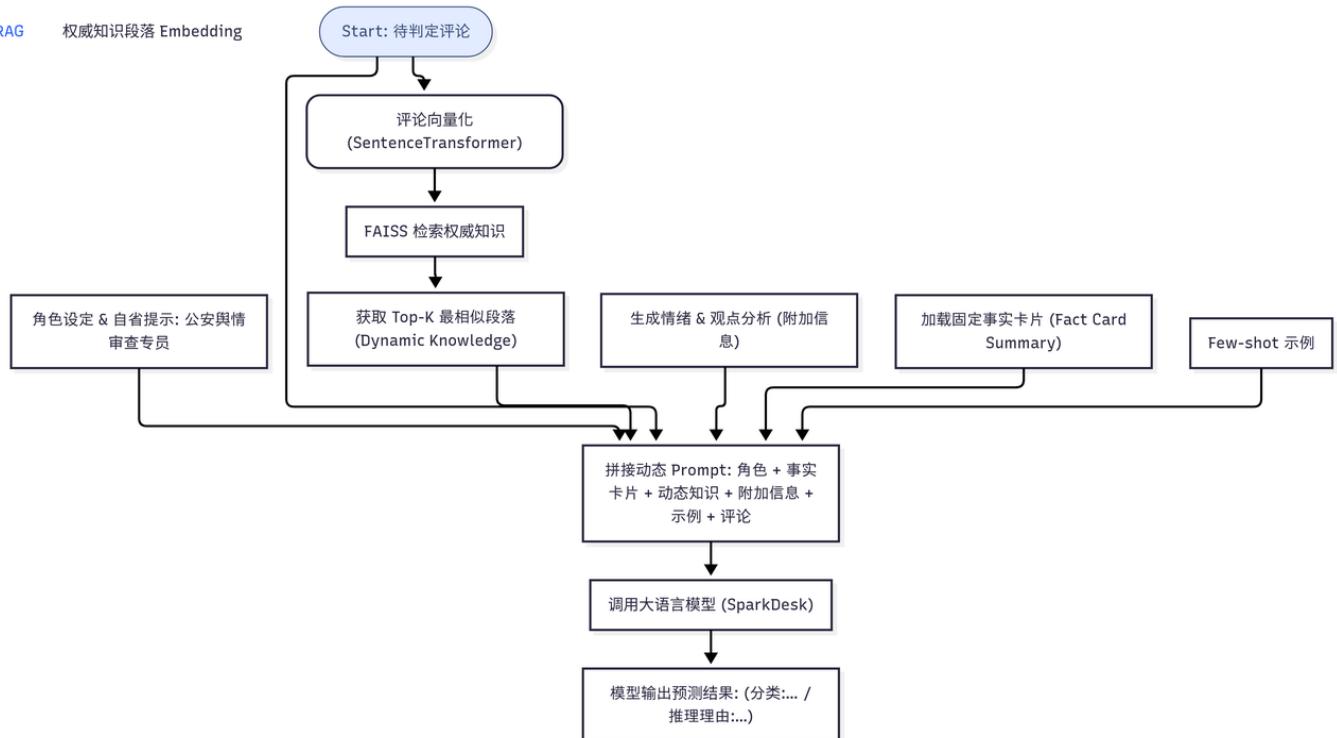


图4 动态prompt构造流程图

权威知识段落Embedding

首先，通过预训练的语义Embedding模型（*SentenceTransformer*模型 *paraphrase-MiniLM-L6-v2*）对权威知识卡片中的文本段落进行编码，并将所有向量存储至FAISS向量检索引擎中。这一步确保背景信息具备高质量的语义表示。其主要计算方法如下：

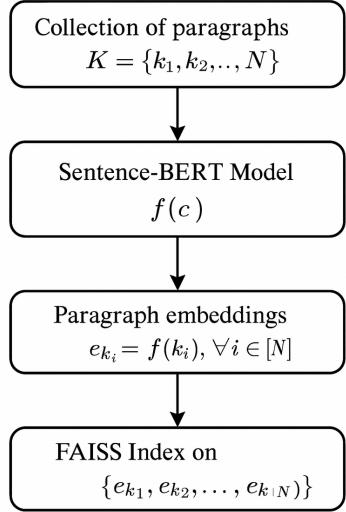


图5 权威知识语义编码与索引构建图

评论向量化与知识检索

对于每条待判定的评论，同样生成其Embedding表示。利用FAISS索引从权威知识库中检索与评论最相似的Top-K段落，这些段落构成了动态的权威知识背景，反映了针对当前评论的最新、最相关的事事实依据。其主要计算过程如下：

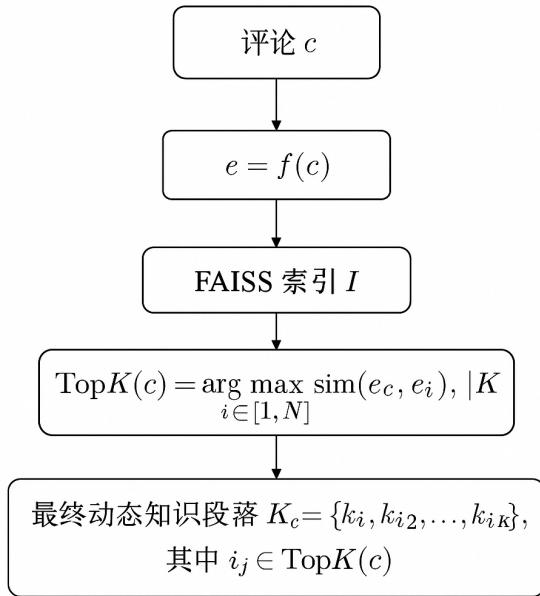


图6 评论向量化与相关知识检索图

动态Prompt构造

将从知识库中检索到的动态权威知识与预先设定的Few-shot示例、附加信息（例如情绪分析和观点完整性）以及固定的“事实卡片总结”拼接在一起。最终的Prompt包含以下几部分：

事实卡片总结：提供固定的事事实摘要，例如“胖猫”事件的关键事实及权威结论；

动态权威知识：利用RAG策略检索到的与评论最相关的权威段落；

附加信息及Few-shot示例：提供情绪、观点分析与示例案例，辅助模型理解任务要求；

待判定评论文本：作为最终输入，要求模型依据上述背景信息输出分类结果和推理理由。

模型输出格式要求严格，即第一行以“分类：”开头，第二行以“推理理由：”开头，以保证结果的可解析性和一致性。

3.4 SparkDesk调用与大规模部署

为提升对网络舆情中“真实”“虚假”与“信息不足”类信息的识别能力，本文引入基于SparkDesk的大语言模型（Large Language Model, LLM）对“胖猫事件”中的微博评论进行自动化真假分类。

3.4.1 Prompt版本控制与SparkDesk调用

使用Python自定义了query_spark函数（封装在spark_utils.py中）来构造和发送API请求。系统首先通过build_prompt(comment)加载预设的Prompt模板，该函数自动整合权威知识卡片和Few-shot示例，生成完整任务指令。接着，根据SparkDesk API要求配置请求头和鉴权信息（如APP_ID、API_PASSWORD），并设置诸如model（"4.0Ultra"）、temperature、max_tokens等参数，确保推理过程稳定。

调用过程中，通过requests.post将JSON格式的请求体发送至API_URL，SparkDesk接收Prompt后返回包含候选答案列表的JSON响应。为降低噪音，重点提取返回中choices字段的第一个答案，进一步用字符串处理获取类似“分类：真实”、“分类：虚假”或“分类：信息不足”的最终预测结果。

针对返回格式的多样性，设计了提取函数，通过正则匹配或字符串分割提取准确分类标签，并使用标准化函数统一“信息不足”、“信息不定”为同一标签，确保评估数据的一致性。

为防止请求超时或频繁失败，系统采用try/except捕获异常，对解析失败的响应默认标记为“信息不足”，同时设置适当延时（例如0.3秒）控制请求频率，保证数据完整性。

```
if 'true_label' in df.columns:
    df['true_label'] = df['true_label'].apply(normalize_label)
    df['prediction'] = df['prediction'].apply(normalize_label)
else:
    print("⚠️ 输入文件中没有 'true_label' 列，因此跳过评估指标的计算。")

output_file = 'predicted_weibo_comments.csv'
df.to_csv(output_file, index=False, encoding='utf-8-sig')
print(f"⚠️ 已保存预测结果到 {output_file}")

if 'true_label' in df.columns:
    acc = accuracy_score(df['true_label'], df['prediction'])
    kappa = cohen_kappa_score(df['true_label'], df['prediction'])
    print(f"⚠️ 指标准确率: ({acc:.4f})")
    print(f"⚠️ Cohen's Kappa 值: ({kappa:.4f})")
else:
    print("⚠️ 未检测到 'true_label' 列，无法计算评估指标。")

# 调用 predict_spark.py
python predict_spark.py
```

图7 调用api代码运行界面示例

Prompt可视化与前端集

本项目前端交互基于Streamlit构建，支持单条评论预测与批量CSV预测两种模式。用户可输入评论与权威知识，情绪与观点完整性信息（可选），系统会动态生成Prompt，并将其展示在页面中并

通过大语言模型（SparkDesk）进行推理，返回分类结果（真实/虚假/信息不足）与推理理由。同时批量模式支持上传 CSV 文件，自动完成逐条处理并生成可下载文件。



图8 prompt可视化交互界面示例

3.4.2 大规模数据部署评估

在前述基础实验（199条人工标注数据与三轮Prompt优化）确立最佳Prompt版本后，本研究进一步扩展了Prompt评估框架，设计并实施了面向9000余条评论的大规模真实数据分类部署。为更贴合实际舆情治理情境，本文创造性地引入了“蓝V语境符合度”作为评价指标，即模型判断为“真实”的评论中与实际蓝V用户评论在内容特征和语义表达上的匹配程度。这一指标旨在真实、客观地评价模型对权威媒体与官方口径信息的识别和拟合能力。

大规模数据部署方案

在大规模评估阶段，本文利用前文提到的 `weibo-search` 爬虫工具获取的9000多条涵盖不同用户类型（蓝V、红V、黄V、普通用户）的“胖猫事件”相关评论作为部署评估数据集。

随后，将此前确定的最优Prompt（Prompt V3：RAG动态匹配）应用到全部9246条数据上，通过调用SparkDesk（Spark4.0 Ultra）API，自动完成大规模评论真实性的分类任务，获得每条评论的“真实”、“虚假”、“信息不足”的分类结果。

蓝V语境评价指标设计

在突发舆情治理任务中，如何评价大语言模型所判定“真实”评论的内容质量与实际参考价值，是衡量模型实用性与可信度的关键。本文基于用户认证信息的弱监督特征，引入以“蓝V官方账号语境”为参考的外部评价体系，构建了更加贴近现实应用场景的语义相似度评价框架。

实证研究发现，在微博等社交平台中，**蓝V账号（官方媒体、政府部门、企业机构等）发布内容的真实性与权威性显著高于其他类型用户**。其内容通常与官方通报、调查结论具有高度一致性，体现出良好的事实对齐能力。因此，蓝V账号内容可作为模型“真实”分类判断是否接近权威表达逻辑的弱监督外部参照。

基于此，本文提出两个衍生指标用于评估模型输出的语义可信度：

指标一：蓝V账号匹配准确率

该指标关注模型对蓝V账号评论的判断能力，即：

蓝V账号匹配准确率 = 模型将蓝V账号发布内容判定为“真实”的比例。

通过将模型应用于大规模数据样本，观察蓝V用户数据被模型判定为“真实”的比例，以此检验其对典型权威内容的识别能力。该指标用于衡量模型的事实立场对齐能力和弱监督语义一致性。

指标二：蓝V语境符合度（BERT语义相似度）

为进一步量化模型输出评论与官方表达风格的一致性，本文引入BERT语义编码器，通过向量空间语义相似度计算方式，构建蓝V语境符合度指标。

采用预训练语义匹配模型（BERT）进行向量化表示，并计算模型判断为“真实”的评论与蓝V账号抽样语料之间的平均余弦相似度。根据阈值（设置为0.7以上）统计高匹配评论占“真实”评论总数的比例。

蓝V语境符合度 = 达到语义相似度阈值的“真实”评论数 ÷ 总“真实”评论数

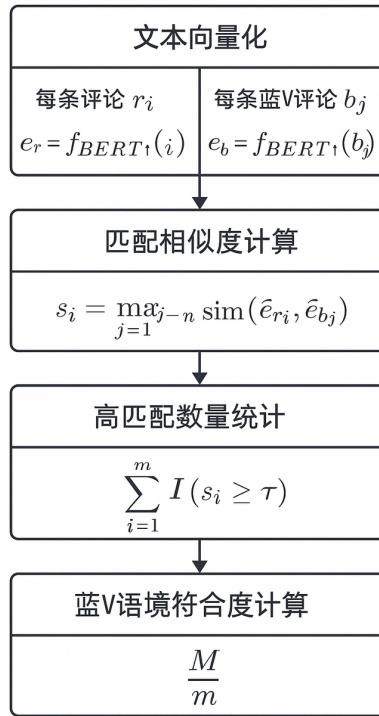


图9 BERT语义相似度计算流程图

四、实验结果与分析

4.1 小规模人工标注样本检验

为评估不同 Prompt 版本在三分类任务中的原生表现与优化效果，本实验选取了 199 条 人工标注评论样本，依次测试 Prompt V1、V2、V3 三种策略。所有实验均基于 Spark 4.0 Ultra 模型，以相同的测试集、Few-shot 示例和评价指标对比性能。

4.1.1 Prompt V1：基础任务模板验证

首先采用最简化的 **Prompt V1**（仅含基础任务说明 + Few-shot 示例）对 199 条样本进行分类。模型被要求在“真实”、“虚假”与“信息不足”三类中做出单一选择，不输出推理理由。结果显示，准确率（Accuracy）为 49.75%；Cohen's Kappa 为 0.0956。

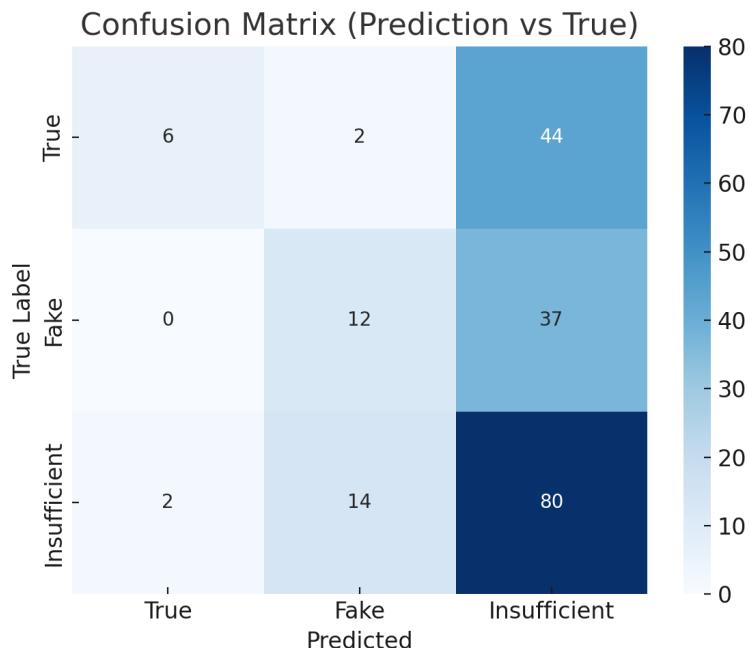


图10 Prompt V1：预测结果与标注结果混淆矩阵图

Prompt V1 虽能完成基本分类，但整体判定准确性较低，且与人工标签存在显著偏差。**Kappa 值远低于 0.1**，表明模型尚未形成稳定的判断偏好，同时边缘样本误判问题较为严重，对“真实”与“虚假”样本的区分能力不足，对语义模糊或信息缺失的评论过度保守，频繁归为“信息不足”。从大模型返回结果来看，模型完成文本分类检测任务时缺乏推理支持，无上下文或角色引导，难以对复杂语句给出可靠判断。

4.1.2 多轮prompt优化结果

在基础版的基础上，依次引入“角色自省+推理输出格式”（Prompt V2）和“动态背景知识+RAG 检索”（Prompt V3），进行多轮优化实验，其准确率和一致性均显著提升。

Prompt 版本	结构增强策略	特征说明	任务定位	准确率	Kappa 值
Prompt V1	基础任务说明 + Few-shot 示例	使用通用分类任务描述，普通身份角色，训练以基本样本为主	基准性能测试 (Baseline)	49.75%	0.0956
Prompt V2	增强任务引导语 + 角色自省 + 推理输出格式	设置“公安网评专员”角色，要求输出判断理由，强化分类推理链条与边界识别自反性能力	增强边界识别与模糊语义适应能力	57.29%	0.2752
Prompt V3	多轮引导 + Embedding RAG+动态 prompt 构建	增加权威知识段落 Embedding 匹配动态背景知识，辅助百度api情感分析，细化区分规则	模拟真实情境下的批量推理与事实核查	72.86%	0.5514

表3 不同 Prompt 版本结构策略与实验结果

Prompt V2 新增元素设定“公安舆情审查专员”角色，要求输出“分类+推理理由”，核心改进方向为情绪提示与自省引导，帮助模型更好捕捉谣言与带攻击性的表达。Prompt V3 结合 RAG 技术，动态检索权威事实段落，并接入情感与观点完整性信息，实现实时背景补充与二次审视机制，旨在使模型在复杂评论上具备更强的事实关联识别能力和分类稳定性。

实验结果说明，随着 Prompt 结构复杂度的递进，特别是在引入“RAG动态背景知识”和“角色身份模拟”后，模型在准确率和 Cohen’s Kappa 一致性指标上的表现呈现显著提升趋势。如图 11 所示，从 Prompt V1 到 V3，模型准确率由 49.75% 提升至 72.86%，一致性系数 Kappa 由 0.0956 增至 0.5514，说明模型判断结果与人工标注的一致性显著增强。

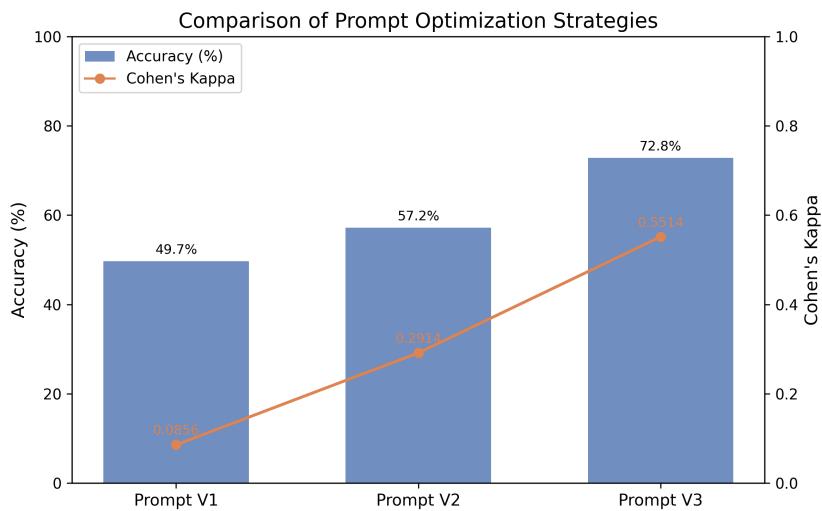


图11 Prompt多轮优化策略下模型准确率与一致性对比图

为了更直观地观察 Prompt V3 对各类别的区分效果，我们绘制了其混淆矩阵（图12），结果显示：

真实 (Real)：正确识别 $32 / (32 + 3 + 18) \approx 57.1\%$

虚假 (Fake)：正确识别 $30 / (0 + 30 + 20) \approx 60.0\%$

信息不足 (Insufficient)：正确识别 $83 / (11 + 2 + 83) \approx 83.8\%$

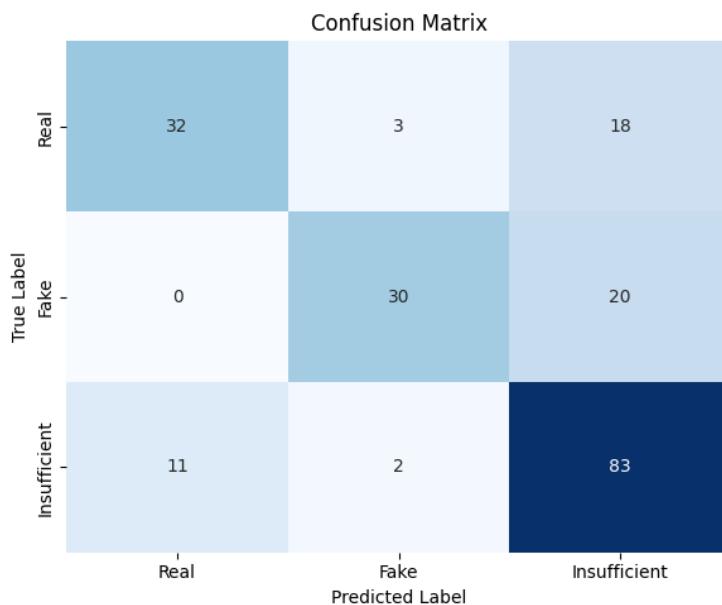


图12 Prompt V3：模型混淆矩阵图

由此可见，V3 在“信息不足”类别上表现最为稳健，同时“真实”与“虚假”评论的识别率也实现了显著跃升——再一次验证了动态背景知识与自省提示的强大增益效果。

从 V1 到 V3，随着 Prompt 结构不断丰富——从基础示例→角色自省→动态检索，模型的准确率和一致性持续改善。Prompt V3 不仅全面提升了总体性能，还在高歧义场景下展现出更强的鲁棒性与可解释性。

4.2 Prompt 版本控制实验分析

在上述对比实验的基础上，本节进一步剖析各版本设计背后的优势机理与结构优化启示，并提出未来的改进方向。

4.2.1 动态 Prompt 的核心优势

动态 Prompt 在构造提示词时，不仅依赖固定的模板和 Few-shot 示例，而是借助外部知识库、实时检索信息及情绪与观点补充等动态因素，对任务背景进行定制化扩展。结合知识增强检索策略（RAG 模型），Prompt V3 通过动态补充权威事实信息与背景数据，使模型在判断复杂评论时能充分考虑上下文和边缘信息。

在前述小样本对比实验中，Prompt V3 相较于基础版本（V1）与自省引导版本（V2）展现了显著的性能提升，细化了上下文理解能力，具备更强的稳定性与泛化能力，增强模式分类任务的可解释性与透明度，在实际领域（如网络舆情监控、内容审核等方面）展现出较强的应用潜力。

究其原因，核心在于动态 Prompt 将固定模板与 Few-shot 示例的“静态”提示，升级为基于检索的“动态”背景补充。当模型面对碎片化、情绪化或断裂语境的评论时，RAG（检索增强生成）模块能够从权威知识库中实时提取最相关的事段落，并融合情绪与观点信息，为推理过程注入强有力的上下文支持。混淆矩阵结果显示，Prompt V3 在“真实”“虚假”“信息不足”三类上的正确识别率均大幅提升，Cohen's Kappa 值由 0.0956 飙升至 0.5514，表明模型输出与人工标注在去除偶然一致后的匹配度显著增强。

与此同时，通过在提示词末尾强制要求“分类+推理理由”双重输出，Prompt V3 不仅提升了可解释性，还为审查人员提供了可追溯的决策链条。这种透明度的增强，使得模型在网络舆情监控、内容审核甚至智能客服等多种应用场景中，都能以“机器判断+人机协同”的方式，支持快速且可信的决策。

4.2.2 Prompt 结构优化启示

错误案例类型的深入分析

Prompt V3 在整体分类准确性和事实对齐能力方面相较前代版本已有显著提升，但其实验结果中仍暴露出一部分具有代表性的错误案例，反映出当前模型在复杂舆论语境下仍面临挑战。从误判分布来看，主要可归纳为两类典型情形：其一是“模糊边界型”误判，主要体现在部分语义模糊或上下文依赖性较强的评论中，尤其是“真实”与“信息不足”标签之间出现界定不清的问题。在此类样本中，尽管评论本身具备部分事实信息，但由于其表达形式松散、缺乏明确的引用依据，模型往往出于保守性策略将其判定为“信息不足”，导致“真实”样本被低估。其二是“情绪量化型”误判，即模型在处理情绪化语言时存在过度谨慎的判断倾向，部分不具备攻击性的主观表达被误判为“虚假”，反映出当前情绪判断机制在强度区分与语义解析上仍显粗糙，未能有效识别批判性表达与虚假信息之间的本质差异。

后续优化方向

基于上述问题，未来的优化路径需从多个维度同步推进。一方面，应持续扩展并动态更新权威知识库，提升基于 RAG 策略的检索语料覆盖率与相关性，确保语义补充的实效性与针对性。另一方面，Prompt 模板中可引入更细粒度的情绪分级机制与语义模糊度量模型，以强化对模糊陈述、暗示性表达等隐性特征的判断能力，避免“语气即标签”的误判误导。此外，构建以误判样本为核心的人工反馈闭环机制，也将成为 Prompt 优化的重要方向。通过引入高频误判案例至 Prompt 示例库，并对模型推理路径进行显式引导，有助于实现分类标准的自我强化，从而提升系统的学习能力与泛化鲁棒性。

总体而言，本次小规模样本检验研究，依托 Prompt 版本控制实验，形成了基线测试（Prompt V1：验证模型基础判断能力）——引导推理（Prompt V2：强化推理链和自省提示，改善情绪与语义模糊样本的处理）——动态补充（Prompt V3：利用定制化背景知识与 RAG 策略，大幅提升模型对复杂评论事实关联的识别能力和整体分类准确性）的闭环优化流程，为大语言模型在碎片化文本分类任务中的准确性、稳定性和可解释性提供了系统性证据，并为后续在更复杂舆情场景中完善 Prompt 设计与模型部署奠定了坚实基础。

4.3 大规模数据部署评估结果分析

在完成基于小样本人工标注数据的 Prompt 结构优化与模型对比实验后，本文进一步在大规模舆情数据上部署了最优 Prompt 版本（Prompt V3），对 9246 条与“胖猫事件”相关的微博评论进行三分类自动判别。部署结果表明，该模型在实际数据环境下展现出稳定的判断能力和良好的分类效果。

4.3.1 整体分类结果

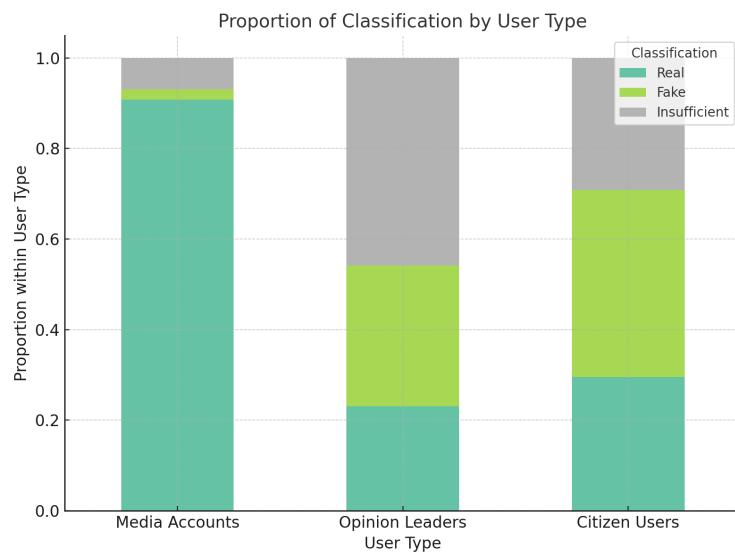


图13 用户类型 × 评论分类矩阵混合占比图

Sparkdesk 大语言模型在不同用户群体之间展现出显著的类型分布差异，进一步验证了用户角色差异对评论内容结构的显著影响。**媒体类账号（Media Accounts）** 中，真实评论占比高达 88.2%，虚假评论极低，仅占 2.3%。这验证了蓝 V/ 金 V 等官方账号的事实表达具有显著一致性，模型对其有良好的权威对齐能力。**以红 V、黄 V 为主的民间意见领袖类账号（Opinion Leaders）** 中，“信息不足”类评论占比最大（约 46.8%），显示其在表达立场时常使用模糊或调侃方式，模型较难进行明确归类。而**普通用户（Citizen Users）**，虚假占比最高（约 44.3%），反映出普通用户在情绪化表达、信息不完整与误导倾向方面的显著特征，模型识别能力得到有效验证。

在整体分布结构上，模型预测显示“信息不足”类别在三类用户占据主导（约64%），符合当前社交媒体舆论传播的“长尾效应”特征，即大量评论处于事实模糊、内容不完整、缺乏验证依据的中间态。

4.3.2 蓝V指标评估结果

指标	数据
模型判定为“真实”的评论总数	2347
与蓝V语料语义相似度 ≥ 0.7 评论数	1796
蓝V语境符合度	76.53%
蓝V账号匹配准确率	90.81%

表4 蓝V指标评估结果

进一步利用BERT模型对“真实”类别的评论与蓝V用户真实内容进行语义匹配度分析后，显示：模型判定为“真实”类别评论中，达到蓝V账号风格语义相似度阈值（ ≥ 0.7 ）的评论数量为1796条；“蓝V语境符合度”为 $1796/2347 = 76.5\%$ 。

该结果说明，在大规模真实数据部署情境下，**SparkDesk + Prompt V3策略**下分类模型不仅在小样本人工标注任务上表现出色（准确率达到72.86%，Kappa达到0.5514），更重要的是，在无标注的真实环境下也能取得76.5%的权威媒体（蓝V）语境拟合度。这表明本作品的Prompt版本控制与迭代优化策略显著提高了模型对权威信息的理解与识别能力，大规模数据环境中展现出优异的“权威语义风格拟合能力”，也为后续在其他舆情热点事件中的推广应用提供了经验支撑和可复用框架。

五、结论与拓展

5.1 研究总结

本研究以“胖猫事件”为典型个案，针对社交媒体舆情环境中“碎片式虚假新闻”的识别难题，提出并验证了一种基于SparkDesk大语言模型的Prompt引导三分类判断框架。与传统需要大规模微调或重训练的模型不同，本框架通过“权威知识卡片+few-shot示例+用户输入”的结构化Prompt输入体系，配合三轮迭代优化策略，有效实现了对微博评论“真实”“虚假”“信息不足”三类标签的自动判别，并兼具实时性、低标注成本与高可解释性。

研究方法聚焦于**Prompt版本控制与多轮优化**，通过设计阶梯度化的不同prompt版本结构，集合权威知识库和混合知识检索开展Sparkdesk大模型进行文本分类任务时的提示工程研究。其中，**Prompt V1（基础模板）**仅含任务说明与典型示例，用以测试模型原生分类能力；**Prompt V2（角色与自省引导）**增加“公安舆情审查专员”角色设定与自省提示，要求模型输出“分类+推理理由”，提升对情绪化和模糊表达的敏感度；**Prompt V3（RAG 动态检索增强）**则引入实时检索的权威事实段落（RAG），结合情感与观点完整性信息，动态补充背景知识，显著提升复杂评论的事实对齐与分类稳定性。

在实验结果上，在小样本对比实验阶段中，针对199条人工标注样本，Prompt V3的分类准确率达到72.86%，Cohen's Kappa值达到0.5514，相比基础版本（Accuracy≈49.8%，Kappa≈0.096）实现了跃升。该结果充分证明了**动态背景知识匹配、多轮引导、角色设定与自省机制**

在细粒度、碎片化文本分类任务中的有效性。在大规模部署与权威拟合评估阶段中，将 Prompt V3 部署于 9,246 条真实微博评论后，模型在“真实”类别的蓝V语境符合度达到 76.5%。即分类为“真实”的评论，其语义风格与官方蓝V账号的表达高度一致，验证了框架在真实舆情环境下对权威内容逻辑风格的识别与拟合能力。

本文提出的“**Prompt 版本控制 + 迭代优化 + 大规模数据部署**”策略，不仅在算法性能指标上取得显著进步，更在真实舆情治理场景中展现出高度的可部署性与可解释性。研究成果为大语言模型在政务舆情监测、内容审核与虚假信息过滤等任务中的落地应用，提供了可复制的理论框架与实践样本。

5.2 后续改进方向

尽管本研究已初步构建了“Prompt 引导 + SparkDesk 推理 + 权威知识核对”的虚假新闻检测闭环流程，为实时舆情治理提供了切实可行的技术方案，但在实际部署与跨场景应用中，仍有多重方向可进一步拓展。首先，鉴于当下社交平台内容日益呈现图文并茂、短视频流行的趋势，仅依赖文本信息难以全面捕捉多样化的误导手段，因此应当引入跨模态信息融合策略，通过同步分析评论中所附图片、视频帧或图文混排的视觉线索，实现对图文一致性与多模态伪信的深度识别，从而进一步提升虚假信息检测的覆盖面与准确率。

与此同时，为了降低对外部 API 调用的依赖，以及满足对延迟和安全性有更高要求的政务或行业级场景，需要探索轻量化部署与本地化适配路径。借助开源大模型（如 LLaMA、Qwen 等）结合知识蒸馏与量化剪枝等技术，可在保证推理性能的前提下，将模型私有化落地，从而显著降低运营成本并提升响应速度，为政府部门和企业提供可控、可审计的部署方案。

在标签体系方面，原有的“真实—虚假—信息不足”三分类虽然能够覆盖大多数评论类型，但对于引流、重组、煽动等更细粒度的虚假手段尚无专门区分。未来研究可基于情绪识别与立场分析，将虚假信息进一步拆分为“引流型”“重组型”“情绪操控型”等子类，并以多任务学习框架联合训练分类与情感、立场等模块，既丰富了事件结构建模维度，也提升了对不同误导策略的辨识能力。

此外，为了更好地对齐权威账号的语言风格，可以在“蓝V语境符合度”之外，进一步训练专门的“权威语境风格识别器”，将其作为生成风格控制或分类辅助模块，实时监测模型输出与官方口径的一致性，进而增强整体判断结果的可信度与专业度。

在可解释性层面，除了要求模型输出推理链与判断依据外，还可配套开发可视化人机协同界面，将分类结果、证据段落、情绪指标等关键要素以交互式图表形式呈现，辅助舆情分析人员快速锁定潜在风险，并实现“人机共治”模式下的高效决策流程。

最后，为检验以上改进策略的综合效果，应与政府舆情系统或内容审核平台进行对接，模拟突发事件中的“检测—对齐—干预”全流程，开展跨系统闭环实战演练，以评估模型在实时高峰期的稳定性、可扩展性及政策落地可行性，从而为未来面向更多热点场景的智慧舆情治理奠定坚实基础。

六、项目应用价值

6.1 项目调研与需求背景

随着社交媒体平台日益成为公众获取信息和表达观点的主要渠道，网络空间中的内容呈指数级增长，其信息复杂度与传播速度也随之显著提升。传统舆情治理手段——例如关键词拦截、人工研判、黑白名单机制等——虽在早期具备一定效果，但在面对评论类碎片信息中真假难辨、情绪操控与立场引导等现象时，已逐渐暴露出应对乏力的问题。

首先，信息高度碎片化，评论内容多为非结构化的个体表达，其中往往包含模糊立场、隐晦指控或未标注来源的引用，导致传统文本分类模型难以适应其语义复杂性；其次，误导性言论的隐蔽性较强，虚假内容不再以直接伪造事实的方式出现，而是通过情绪包装、图文混排、引导性语言等形式巧妙掺杂于真实语料中，增加了识别难度；再次，随着网络评论体量剧增，人工审核的效率与主观性问题逐渐暴露，审核者往往在短时间内难以做出一致性判断；最后，当前已有的AI检测模型大多依赖具体训练语料，对新兴事件的适应性差，迁移能力弱，难以跨事件泛化使用。

相关数据显示，在85%以上的突发舆情事件中，评论的扩散速度远快于权威回应的速度，而其中约48%的高互动评论存在事实偏离或潜在引导倾向。这些现实问题表明，亟需引入一种能够理解语义深层结构、具备高泛化能力并可在弱监督甚至无监督条件下进行推理的大语言模型方法。因此，结合Prompt提示工程、权威知识卡片与少量判断示例的大语言模型检测框架，成为提升虚假评论识别效率与治理响应能力的切实可行路径。

6.2 应用场景说明

本Prompt引导的Sparkdesk大语言系统在检测碎片式虚假新闻上具备良好的任务泛化性与部署灵活性，因而可广泛应用于多个关键场景中，赋能各类涉舆主体实现更高效、更智能的内容识别与管理。

在政府部门应对突发事件的场景中，当微博、抖音、小红书等平台短时间内爆发与公共安全、社会热点相关的舆情浪潮时，本系统可实现对评论内容的实时三分类处理，从中快速筛选出具有潜在误导性或信息不足的言论，从而辅助宣传与网信部门开展精准信息对齐、风险预警及正向舆论引导。

在内容平台审核机制优化方面，系统可作为审核员的AI助手嵌入后台管理工具。通过模型的智能判断与标签打标功能，平台可构建基于分类结果的“评论预警池”，将疑似虚假或信息不足的评论优先呈现，提升整体审核效率，减少人工判断压力，并支持未来构建更为复杂的“内容可信度分级体系”。

对于研究机构与主流媒体单位而言，模型的结构化输出结果亦可用于深入的事件分析与信息溯源。例如，评论立场分布图、虚假评论情绪强度图等二次分析产出，可为“谣言传播链路建模”、“观点扩散路径重建”等研究提供支持，推动学术界与媒体行业对新型舆情模式的理解与应对能力。

随着智慧城市建设的推进，系统还可接入智慧政务平台的舆情子系统，作为实时事件语义分析模块运行，为政务管理部门提供事件识别、评论可信度量化评分、重点评论聚类分析等多项辅助功能，强化地方政府对民意动态的实时掌控与科学干预能力。

6.3 应用市场潜力与推广可行性

从市场角度看，本项目具备广泛推广价值：

目标机构/平台	潜在应用形式	实施优势
政务舆情中心	舆情评论筛查、重大事件信息对齐监控	实时响应能力强、部署成本低
主流社交平台（微博、抖音）	智能审核插件、AI打标模块	降低人工成本、避免漏审风险
传媒机构与智库	舆论趋势研究、话语空间分析	数据可用性高、结构化输出
高校研究平台	舆情结构分析、教学案例复用	可视化强、研究边界广
大模型平台合作方	Prompt调优服务、模型标准测试集	可迭代、多模型横向评估

综上所述，本文在理论框架构建、方法路径创新与实验效果验证三个层面均取得积极成果，提出的Prompt引导检测策略兼具科学性、实用性与创新性，为**SparkDesk**大语言模型在舆情治理与公共信息可信度识别等场景中开拓出一条可执行、可扩展的路径。

七、参考文献

- [1] 张坤丽,王影,付文慧,等.大语言模型驱动下知识图谱的构建及应用综述[J/OL].郑州大学学报(理学版),1-9[2025-04-17].<https://doi.org/10.13705/j.issn.1671-6841.2024165>.
- [2] 方全,张金龙,王冰倩,等.基于组合上下文提示的大型语言模型领域知识问答研究[J/OL].计算机科学,1-13[2025-04-17].<http://kns.cnki.net/kcms/detail/50.1075.TP.20250417.1135.022.html>.
- [3] 张强,高颖,任豆豆,等.融合DeepSeek-R1和RAG技术的先秦文化元典智能问答研究[J/OL].现代情报,1-20[2025-04-17].<http://kns.cnki.net/kcms/detail/22.1182.G3.20250414.0937.002.html>.
- [4] 王亮.检索增强生成(RAG)驱动的知识服务：原理、范式及评估[J/OL].科技与出版,1-10[2025-04-17].<https://doi.org/10.16510/j.cnki.kjycb.20250409.002>.
- [5] 潘磊,袁鸿霄,钟准,等.基于大模型构建图网络的事件因果关系识别[J/OL].西南交通大学学报,1-10[2025-04-17].<http://kns.cnki.net/kcms/detail/51.1277.u.20250408.1541.004.html>.
- [6] 张文杰.提示词治理:DeepSeek等国产大模型内容生成的人机协同模式[J/OL].苏州大学学报(哲学社会科学版),1-12[2025-04-17].<http://kns.cnki.net/kcms/detail/32.1033.C.20250407.1357.004.html>.
- [7] 王晓诗,景少玲,孙飞,等.社交媒体优化综述[J/OL].计算机学报,1-27[2025-04-17].<http://kns.cnki.net/kcms/detail/11.1826.tp.20250401.1521.002.html>.
- [8] 郭宇豪,陈庆奎.PGCA-RAG：面向大语言模型检索增强的并行图缓存架构[J/OL].小型微型计算机系统,1-8[2025-04-17].<http://kns.cnki.net/kcms/detail/21.1106.tp.20250328.1618.014.html>.
- [9] 陈文泰,张帆.提示词构成要素对大语言模型跨模态内容生成质量的影响研究——基于讯飞星火大模型文生图功能的探索性实验[J].郑州大学学报(哲学社会科学版),2025,58(02):99-104.
- [10] 秦董洪,李政韬,白凤波,等.大语言模型参数高效微调技术综述[J/OL].计算机工程与应用,1-30[2025-04-17].<http://kns.cnki.net/kcms/detail/11.2127.TP.20250326.1554.025.html>.

- [11]宋佳磊,左兴权,张修建,等.大语言模型评估方法综述[J/OL].宇航计测技术,1-30[2025-04-17].<http://kns.cnki.net/kcms/detail/11.2052.V.20250325.1347.002.html>.
- [12]陆小飞,金檀.大语言模型微调技术在语言分析与测试中的应用与展望[J/OL].现代外语,1-9[2025-04-17].<https://doi.org/10.20071/j.cnki.xdwy.20250314.009>.
- [13]潘辉.我国大语言模型的研究进展:基于知识图谱的研究主题识别[J].情报探索,2025,(03):123-134.
- [14]刘雪颖,云静,李博,等.基于大型语言模型的检索增强生成综述[J/OL].计算机工程与应用,1-31[2025-04-17].<http://kns.cnki.net/kcms/detail/11.2127.TP.20250312.1303.008.html>.
- [15]Ron S .计算认知架构、双过程理论和大语言模型（英文）[J].宁波大学学报(理工版),2025,38(02):17-28.DOI:10.20098/j.cnki.1001-5132.2024.1014.
- [16]梁炜,许振宇.大语言模型赋能舆情治理现代化：价值、风险与路径[J].中国应急管理科学,2025,(01):93-103.
- [17]祁凯,周燕生.基于大语言模型生成内容的负面舆情态势恶化牵引作用研究[J/OL].情报杂志,1-10[2025-04-17].<http://kns.cnki.net/kcms/detail/61.1167.G3.20250311.1110.002.html>.
- [18]王润周,张新生,王明虎,等.基于混合检索增强生成大语言模型的网络舆情多任务分析[J/OL].情报杂志,1-14[2025-04-17].<http://kns.cnki.net/kcms/detail/61.1167.G3.20241212.0931.008.html>.
- [19]魏晨光,江旺,范海亮,等.基于大模型的网络钓鱼攻击检测方法研究[J].网络安全技术与应用,2025,(03):26-29.
- [20]许德龙,林民,王玉荣,等.基于大语言模型的NLP数据增强方法综述[J/OL].计算机科学与探索,1-23[2025-04-17].<http://kns.cnki.net/kcms/detail/11.5602.TP.20250218.1110.002.html>.
- [21]刘洋.“大模型+RAG”技术在档案工作中的应用探析[J].中国档案,2025,(03):64-65.
- [22]赵静,汤文玉,霍钰,等.大模型检索增强生成（RAG）技术浅析[J].中国信息化,2024,(10):71-72+70.
- [23]魏宏程,杨建林.大语言模型+检索增强方法的关键技术及其在情报任务中的应用流程[J].情报理论与实践,2025,48(03):178-188+206.DOI:10.16353/j.cnki.1000-7490.2025.03.021.
- [24] Shu K, Wang S, Lee D, Liu H. Beyond News Contents: The Role of Social Context for Fake News Detection[C]// Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining (WSDM). 2019: 312–320.
- [25] Zhou X, Zafarani R. Fake News Detection: A Survey[J]. ACM Computing Surveys, 2020, 53(5): 1-40.
- [26] Kaliyar R K, Goswami A, Narang P. FakeBERT: Fake News Detection in Social Media with a BERT-based Deep Learning Approach[J]. Multimedia Tools and Applications, 2021, 80(8): 11765–11788.
- [27] García-Durán A, Maillard J, Srivastava M, et al. Few-shot Fact Checking with Large Language Models[J/OL]. arXiv preprint arXiv:2305.10601, 2023.