文章类型：论文|刊号（ISSN）：2630-4740 /（中图刊号）：715GL012

生成式 AI 在气象预警信息制作中的可控应用

应爽[1]　吕珊珊[2]　解彦维[3]
1 长春市气象局　2 农安县气象局　3 九台区气象局
DOI:10.32629/eep.v8i9.2882

[摘　要] 气象预警信息制作须短时完成文案撰写、渠道适配与CAP封装，但存在口径不一致、重复劳动高、质检低效、审计弱等问题。本文提出面向气象局内部制作环节的“可控人机协同”框架：以“预警要素单”为唯一事实来源,使用生成式AI完成多版本草拟与风格迁移；以红线约束与自动质检确保要素一致、规范用词与渠道限长；以CAP实现跨平台一致封装与版本关联；以权限分离、日志审计与KPI看板构建治理闭环。框架基于面向影响理念,建立“分级—措辞—行动”的一致性映射,并以城市内涝、高温热浪情景作为落地示例。方案与WMO指南、CAP标准及风险治理规范一致,可快速集成,提升一致性、时效性与可审计性。
[关键词] 生成式AI；气象预报预警；面向影响预警；人机协同
中图分类号：S152.4+2　**文献标识码**：A

Controllable Application of Generative AI in Meteorological Warning Information Production

Shuang Ying[1]　Shanshan Lv[2]　Yanwei Xie[3]
1 Changchun Meteorological Bureau
2 Nong'an County Meteorological Bureau
3 Jiutai District Meteorological Bureau

[Abstract] Meteorological alert production faces challenges of inconsistent wording, repetitive labor, low efficiency, and weak auditing. This paper proposes a controllable framework integrating human–machine collaboration, CAP encapsulation, and quality governance. The framework uses alert element sheets as the source of truth, employs AI for multi–version drafting, enforces constraints to ensure consistency, uses CAP for cross–platform encapsulation, and constructs governance via auditing and KPI tracking. The impact–based framework establishes severity–phrasing–action mapping, illustrated through flooding and heat–wave cases. It aligns with WMO guidelines and CAP standards, integrates quickly, and improves consistency, timeliness, and auditability.
[Key words] Generative AI; Meteorological forecasting and warning; Impact–based warning; Human–machine collaboration.

引言

气象预警服务重在促进防灾减灾,信息的清晰、一致和可达性至关重要。当前业务面临三类主要问题：跨班次与跨渠道口径一致性；跨终端与多格式文本的传播适配；对结构化标准的稳定合规。生成式AI有利于解决上述问题,但预警与公共安全紧密相连,不可“自动发布不经审”。因此,本文旨在给出明确责任、严格约束、全链路可审计前提下,AI生成气象信息的可控应用路径。

国际上,WMO的面向影响预警强调“从不确定性到行动”的转译[1],OASIS发布的共同警报协议(CAP)标准提供了跨平台一致的机读框架[2],Sendai框架将早期预警纳入减灾治理[3]。本文据此设计“要素单驱动的人机协同框架”,聚焦文案草拟、质检与CAP封装等非决策环节的提质增效。

1 方法与架构

1.1设计原则与总体架构

本文方法的核心是把预报决策与文本生成剥离：预报员会商形成“预警要素单”,明确灾种、等级、范围与重点人群等；AI在硬约束不变前提下生成公众版、行业简报与各渠道摘要等。架构分四层：

在规范层,建立分级与术语规范[4]、行动建议库、禁用词与风格守则,及标准地名与区域库。旨在把“面向影响”的理念固化为可执行的词汇与句式,并提供等级到行动强度的一致性映射。

在模型层,采用可控的文本生成模型(优先本地化部署或合规云端)与检索增强组件,结合规则引擎完成输出前置约束与产出后校验。该层不触碰预报要素决策,仅围绕表达、结构与适配工作。

在业务层,流程设计为：要素单表单化、AI草拟与多版本生成、自动质检、人工复核与编辑、CAP封装与版本关联、渠道化适配与分发、留痕归档。各环节以接口衔接,避免人工复制粘贴导致的口径漂移。

在治理层,实施身份与权限管理(风险分级与权限分离)、日志与审计(提示词、输入、输出与修改轨迹)、看板与KPI,及模板与词汇库的版本管理。治理层确保“越高风险,越多人工在环与审批层级”[5]。

1.2人机协同流程与控制点

流程由“预警要素单”触发,要素单是唯一事实来源,含灾种、等级、时间窗、区域几何或行政清单、要素强度、预期影响与重点人群,并记录“下一次计划更新时间”(图1)。AI基于要素单与本地化模板生成公众版与行业简报草案,并产出渠道摘要(短信、App弹窗等)。

自动质检覆盖四类：要素一致性(灾种、等级、时间、区域与数值单位不可改动)、术语与禁用词(措辞与分级匹配,避免夸张与越权承诺)、渠道限长与结构(不同渠道以最小冗余承载关键信息)、CAP预校验(时间格式、区域几何、消息ID唯一性与引用链)。通过后进入人工审核,预报员通读并补充本地化要点,中高等级实施双人复核。最终CAP封装并按通道冗余策略分发;提示词、输入输出、修改记录、审批人、时间戳与回执全程留痕,便于审计与改进。

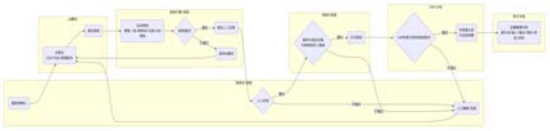

图1 信息制作与分发的端到端流程图

1.3可控生成技术与模型部署

模型选型应以“可控、稳态、低泄露风险”为优先目标。相比追求通用大模型能力,更合理的是采用本地部署的小中模型,并通过检索增强接入术语库、行动建议库与标准地名库,降低幻觉与语义漂移的风险。部署侧遵循最小必要原则：不输入敏感信息,推理日志存放受控域；配置模板回退,确保关键产出可离线生成。模型生命周期管理需以稳定基线评测集为依托,覆盖口径一致性、长度合规与禁用词漏检等维度,配合小流量灰度上线与版本回滚策略,形成“先评测、后放量”的变更管控。运行监测建议纳入推理时延、失败率与质检不通过率等指标的突变告警,一旦超过阈值立即熔断并切换人工或回退链路,从工程层面保证连续可用与风险可控[6][7]。

为确保生成内容严格对齐“要素单”与规范库,应引入结构化与受限解码。先以受约束结构(如JSON模式或函数调用)产出,仅含灾种、等级、时间窗、区域与行动建议等关键要素,再以模板渲染自然语言,避免自由生成对硬约束的越权修改。后引入独立于模型的规则引擎进行二次校验,对灾种、等级、时空范围、数值与单位进行逐项比对,不通过即拦截,形成“生成—校验”的双重保险。术语与行动建议使用“字典锁定与等价词白名单”控制表达边界,限制可能导致强度感知变化的随意改写。渠道适配坚持“模板优先、生成补充”,先满足长度、无障碍与读秒要求,再让模型填充要素化语块,以减少口径漂移。为提升可解释性与审计能力,内部留存生成所引用的知识片段标识,使每条行动建议都能回溯至规范条目或历史模板,有助于值班员快速定位问题并完善库表。

2 产品与互联

2.1面向影响的表达与渠道适配

预警信息的表达遵循“分级—措辞—行动”的一致性映射：黄色为“准备级”,侧重提醒与建议；橙色为“行动级”,强调立即执行的规避与调整；红色为“紧急级”,突出停止非必要活动并就地避险[1][4]。表达不追求概率数字的精确呈现,而采用“可能、局地、短时、显著增大”等范围词,将不确定性翻译为具体、可执行的行为。渠道适配遵循“短而全”的原则：短信以最小要素与一条行动建议构成,App弹窗与政务号兼顾可读性与无障碍,广播口播稿采用短句与三段式结构,网站通告提供完整长文与CAP下载。通过AI的风格迁移能力,可显著降低在多个渠道重复改写导致的口径漂移风险。

2.2 CAP封装与版本关联

CAP是跨平台一致传播的“最小可行消息体”。封装时填写identifier、sender、sent、status、msgType、scope等字段,并在info块中统一event、urgency、severity、certainty、effective、expires与area等关键元素。升级或更正通过references字段建立版本关联[2],形成可追溯链。区域优先采用多边形表达,必要时配合行政区清单与标准地名编码。时间格式、时区一致与几何合法性在自动质检中前置校验,减少分发失败与语义歧义。遵循CAP使得本地系统可以无缝对接上游或平行系统,符合多部门协同与Sendai框架倡导的互操作性目标[2]。

3 治理与评估

3.1质量治理与KPI

质量治理以“自动质检+人工复核+事后复盘”三位一体实现闭环[5]。自动质检负责刚性一致性,人工复核负责语义正确与本地语境,事后复盘通过演练与运行日志定位模板不足、术语库缺口与质检规则盲点。KPI建议以定义优先、统计其次：时效类包括从要素单到初审通过的时长、升级重发时长；质量类包括自动质检一次通过率、人审修改率、CAP校验通过率；一致性类关注跨渠道口径一致率与跨分区风格一致性；可达性以分发日志统计投递成功率与时延；安全类以红线违规拦截次数与审计留痕完整率衡量。KPI用于内部改进与资源配置,不作为对个人的简单考核,以避免“为了指标而服务”的逆向激励[6][7]。

3.2评估设计与红队测试

可信应用需可重复的评估与对抗性检验。评估建议采用“运

文章类型：论文|刊号（ISSN）：2630-4740 /（中图刊号）：715GL012

行日志+对照实验”的组合设计：在同灾种、相近复杂度中,分别采集纯人工、人机协同流程样本,比较从“要素单确认到初审通过”的时长、自动质检一次通过率、人审修改率与CAP校验通过率,并跟踪跨渠道关键要素一致性与分发成功率,以运行日志为数据源开展滚动监测。为验证安全边界与鲁棒性,应组织红队测试,构造能够诱发越权承诺、等级措辞不匹配、单位与范围表达错误和禁用词渗透的挑战用例,记录自动质检拦截与人工兜底的成功率,并据此归纳失败模式,形成“问题—规则补丁—再验证”的闭环机制。定期开展事后复盘,结合演练与真实运行样本更新模板、术语库与禁用词清单,同时审视看板指标的漂移与异常,确保系统在持续迭代中维持既定的风险门槛与服务水平[8]。上述评估与红队流程不依赖外部敏感数据,能够在不增加一线负担的前提下,提供足够的证据支持与改进指引。

3.3风险与合规

AI制作预警信息的风险包括内容幻觉、措辞与等级不匹配、敏感承诺越权、日志与隐私暴露。风险控制的总原则是“规则先行、模型后置；人工在环、权限分离”。硬约束字段(灾种、等级、时间、区域、要素强度)禁止模型改写；对未知与未签发要素不生成;敏感承诺与行政决定性表述(如停课停业)一律禁用;质检不通过即熔断回退人工。模型与日志部署遵循信息安全与隐私保护规范,访问可审计。上述做法与国际上对AI风险管理的理念相一致：明确上下文、识别与控制风险、设置监测与改进机制[6][7]。此外,系统全程不启用“自动发布不经审”模式；在任何情况下,签发责任与最终口径均由人工岗位承担。

4 应用与实施

4.1情景化示例

以强降水引发的城市内涝(橙色)为例,会商形成要素单:主城区与南部片区,18:00至次日06:00,1-3小时雨强30 - 50毫米(以本地技术规定为准)、局地更大,关注下凹式立交与低洼点积涝,重点人群为晚高峰通勤与地下空间作业人员。AI生成公众版草案时,将不确定性转译为“短时强降雨和能见度下降风险增大”,行动建议聚焦“减少夜间出行、避开积水与地下空间、车辆远离易涝点”。短信摘要限定在120字内(以本地运营商与平台配置为准),广播口播稿采用短句,网站通告则提供更完整的影响场景描述与下一次计划更新时间。CAP封装中设置severity为Severe、urgency为Expected、certainty为Likely,多边形覆盖主城与南部,references串联后续升级或更正。

以高温热浪(红色)为例,要素单指出“全市大部未来三天≥40℃(以本地技术规定为准)、体感湿热高、重点人群为户外作业与老年慢病人群”。公众版文案强调“减少午后外出、补水降温、照护老幼慢病、错峰安全用电”等行动,App弹窗提供简化版本与无障碍打字支持。CAP设置expires与多天有效期一致,区域采用行政合并表示。

两例均体现“要素单驱动、AI迁移、质检兜底、CAP合规”的生产线特征。

4.2实施路线与讨论

实施过程中,建议“三步走”推进：完成要素单表单化、术语与行动建议库建设、AI草拟与自动质检在蓝/黄色级别的试点;上线CAP封装与全渠道适配、引入双人复核机制与日志看板;扩展至多灾种与多语言摘要,引入A/B模板对照与风格一致性评估,并与应急部门建立口径协同与联合演练。需要正视三点局限：生成式AI的幻觉风险无法被完全消除,只能通过规则、检索增强与人工在环降低；行动转化效果需要长期运营数据验证,本文仅提出KPI定义；行业深度定制(如电力负荷或城市排水能力)仍需跨部门数据协同。尽管如此,把AI用于“草拟—改写—封装—质检”的非决策环节,已能在不改变签发责任的前提下显著提升产能与一致性。

5 结论

本文提出一套制作气象预警信息的可控人机协同框架：以要素单为事实源,以生成式AI完成多版本草拟与渠道迁移,以红线约束与自动质检兜底一致性与规范性,以CAP封装实现跨平台互操作,并通过权限、日志与KPI构建治理闭环。该方法与WMO面向影响预警、OASIS CAP标准及国际风险治理规范一致,可在2-8周内分阶段集成到现有流程。后续工作建议围绕联合演练、运行日志评估与模板迭代展开,逐步形成可推广的标准化实践。

[参考文献]

[1]World Meteorological Organization.Guidelines on multi-hazard impact-based forecast and warning services:WMO-No.1150[R].Geneva:WMO,2015.

[2]OASIS.Common alerting protocol version 1.2[S].Burlington:OASIS,2010.

[3]United Nations Office for Disaster Risk Reduction. Sendai framework for disaster risk reduction 2015-2030[R].Geneva:UNDRR,2015.

[4]International Organization for Standardization.Societal security—Emergency management—Guidelines for colour-coded alerts:ISO 22324:2015.Geneva:ISO,2015.

[5]International Organization for Standardization. Security and resilience—Crisis management—Guidelines:ISO 22361:2022.Geneva:ISO,2022.

[6]International Organization for Standardization.Information technology—Artificial intelligence—Risk management:ISO/IEC 23894:2023.Geneva:ISO,2023.

[7]National Institute of Standards and Technology.Artificial intelligence risk management framework (AI RMF 1.0)[R].Gaithersburg:NIST,2023.

[8]United Nations International Strategy for Disaster Reduction. Developing early warning systems:A checklist[R].Bonn:UNISDR,2006.

作者简介：

应爽(1982--),女,汉族,吉林农安人,硕士,正高级工程师,研究方向：天气预报预警服务。