大语言模型辅助决策系统：情景化应急演练与应急管理理论现代化的双重引擎

杨以恒[1]，熊伟男[1,*]
1. 海军大连舰艇学院，辽宁 大连，116001

摘要：突发事件的耦合性、不确定性与时间压力，对应急决策人才培养和应急管理理论创新提出了更高要求。本文以大语言模型（large language model，LLM）辅助决策系统为对象，分析其在情景化应急演练中的应用价值与在应急管理理论现代化中的支撑作用。研究认为：基于 LLM 的辅助决策系统实现了从“规则匹配”到“认知协同”的跨越，形成“感知—推理—行动—演化”的闭环机理；借助多 LLM 协同与知识增强，系统能够在演练中生成自适应、可解释、可追溯的高压力情境，并通过陪练、教练与镜像三种角色重构人机关系；在理论层面，LLM 推动应急管理研究从经验归纳走向数据驱动、从静态文本走向动态演化、从碎片化走向体系化。本文进一步给出“领域语料—指令微调—偏好对齐—知识增强—工具链”的技术路径、兼顾过程与结果的多维评估体系，以及数据、系统与伦理三个层面的保障机制，并讨论了事实性错误、评估效度与责任边界等局限。研究可为应急管理的数字化转型与相关学科建设提供参考。

关键词：大语言模型；辅助决策；情景化演练；应急管理；检索增强生成；人机协同

LLM-based decision support systems: a dual engine for scenario-based emergency drills and the modernization of emergency management theory

Yiheng Yang[1] , Weinan Xiong[1,*]
1. Dalian Naval Academy, Dalian, Liaoning, China, 116001

Abstract: The coupling, uncertainty and time pressure of emergencies place higher demands on decision-making competence and theoretical innovation in emergency management. Focusing on large language model (LLM)-based decision support systems, this paper examines their value in scenario-based drills and their role in modernizing emergency management theory. Such systems move decision support from rule matching to cognitive collaboration through a perception–reasoning–action–evolution loop. With multi-LLM collaboration and knowledge augmentation, they generate adaptive, explainable and traceable high-pressure scenarios, and reshape the human–LLM relationship through the roles of sparring partner, coach and mirror. Theoretically, LLMs shift emergency management research from experience-based induction to data-driven inquiry, from static documents to dynamic evolution, and from fragmented studies to systemic analysis. A technical path of domain corpora, instruction tuning, preference alignment, knowledge augmentation and tool chains is proposed, together with a process-and-outcome evaluation system and a safeguard framework at the data, system and ethical levels. Limitations regarding factual errors, evaluation validity and responsibility are discussed. The findings inform the digital transformation of emergency management.

Keywords: Large language model; Decision support; Scenario-based drill; Emergency management; Retrieval-augmented generation; Human–LLM collaboration

1 引言

我国是世界上自然灾害最为严重的国家之一，事故灾难、公共卫生事件与社会安全事件也时有发生。随着城镇化、工业化与信息化的深度交织，突发事件的致灾机理日益复杂，跨区域、跨行业、跨主体的耦合特征更加突出，事件演化往往在数小时内出现多次转折。应急决策者必须在信息不完备、资源受限、时间紧迫的条件下，同时完成态势研判、方案比选、资源调度与公众沟通等多重任务，决策质量直接关系到人民生命财产安全与社会秩序稳定。《“十四五”国家应急体系规划》明确提出提升应急管理能力与水平的要求[1]，而能力建设的落脚点，正是决策者能否在复杂情境中作出合理判断。

与日益复杂的决策需求相比，应急管理的能力建设仍存在三方面短板。其一是演练情境的“脚本化”。多数演练按照预先编写的脚本推进，参演单位按部就班完成规定动作，难以再现真实事件中的信息缺失、指令冲突与局势反复，演练压力不足。其二是理论验证手段的匮乏。突发事件样本稀少且不可重复，研究者多依赖事后复盘与案例比较，难以开展大样本、可控制的因果分析。其三是知识更新的滞后。预案与规范文本一经发布便趋于稳定，而风险形态、技术条件与社会预期持续变化，知识供给与决策需求之间的时滞不断扩大。这三重困境相互关联：情境不真实，训练就难以产生迁移；验证手段不足，理论就难以积累；知识更新滞后，训练与理论又会同时落后于实践。

大语言模型（large language model，LLM）的快速发展为破解上述困境提供了新的技术条件。以 Transformer 为基本架构的预训练模型[2]在海量文本上习得了丰富的语言知识与世界知识，并通过规模化展现出上下文学习[3]与多步推理[4]能力；检索增强生成（retrieval-augmented generation，RAG）[5]将外部知识引入生成过程，显著缓解了事实性错误；基于人类反馈的强化学习[6]与反思机制[7]则使模型能够遵循任务意图，并从反馈中持续改进。这些进展使 LLM 不再只是文本生成工具，而逐步成为能够理解任务、整合证据、提出方案并解释理由的“认知协作者”。

然而，现有研究多集中于模型能力评测或单一场景应用，对 LLM 辅助决策系统如何在“演练”与“理论”两个层面同时发挥作用，仍缺乏系统分析；对领域适配、效果评估与风险治理的讨论也相对分散。本文尝试弥补这一不足，主要工作包括：（1）剖析 LLM 辅助决策系统从规则匹配走向认知协同的机理，提出“感知—推理—行动—演化”闭环；（2）归纳情景化演练中的三类应用模式；（3）论证应急管理理论研究的三大范式转变；（4）给出系统建设的技术路径、评估体系与保障机制，并讨论其局限。研究方法以文献分析、框架构建与公开案例剖析为主，不涉及具体系统的实测数据，所得结论旨在为后续的实证研究提供分析框架。

2 LLM 辅助决策系统：从规则匹配到认知协同

2.1 规则型系统的局限与 LLM 的突破

传统的应急辅助决策系统本质上是一套“规则引擎”：将预案条款、处置流程与专家经验编码为“条件—行动”规则，由系统在特定条件下触发相应建议。这类系统在结构化、确定性的问题上表现稳定，例如依据预案生成响应等级、匹配处置流程、提示联络对象。但面对复杂多变的突发事件，其局限十分明显。一是规则难以穷尽。预案只能覆盖可预见的典型情形，对于次生衍生、跨域联动等“意料之外”的组合，规则库往往无以为继。二是系统缺乏学习能力。规则由人工维护，更新周期长，难以从每一次处置中自动积累经验。三是交互方式僵化。系统只能接受结构化输入，无法理解决策者的自然语言表述、隐含意图与价值偏好，人机之间难以形成真正的协作。

LLM 的引入改变了这一格局。预训练阶段积累的通用知识使模型具备了较强的语义理解与常识

推理能力[8]；指令微调与偏好对齐使模型能够按照任务要求组织回答，并在多个候选中选择更符合人类偏好的方案[6]；上下文学习使模型无需重新训练即可适应新的任务描述与示例[3]；思维链提示把复杂问题分解为可检验的中间步骤，提升了多步推理的稳定性[4]。更重要的是，模型能力的提升呈现出“涌现”特征：当参数规模与训练数据达到一定水平后，模型在类比、归纳与跨领域迁移等任务上表现出非线性的能力跃升[9]。这使得同一套模型可以同时承担信息抽取、方案草拟、文本生成与结果解释等异质任务，而不必为每一类任务单独开发系统。

这种能力使 LLM 的角色从“工具”变为“协作者”。它不再被动等待指令，而是可以主动澄清任务边界、指出信息缺口、提出多个备选方案并说明各自的适用条件与风险。人机关系也随之从“人操作机器”转向“人机共同判断”：机器承担信息整合、方案草拟与一致性检查，人保留价值判断、责任承担与最终决定权。需要强调的是，这种转变并不意味着把决策权让渡给模型。LLM 的输出仍具有概率性，可能生成看似合理却与事实不符的内容[10]，因此必须嵌入核验、溯源与人工确认机制。换言之，模型扩展的是决策者的信息处理能力，而不是替代其判断责任。

2.2 核心机理：感知—推理—行动—演化

LLM 辅助决策系统的运行可以概括为“感知—推理—行动—演化”的闭环（图 1）。四个环节并非线性串接，而是通过统一的证据链与反馈通道相互衔接，形成持续迭代的决策过程。证据链为每一条信息、每一步推理与每一项建议分配可追溯的标识，使后续环节能够核对“结论从何而来”；反馈通道则把行动结果与复盘结论送回感知与推理环节，使系统在下一轮运行中有所调整。

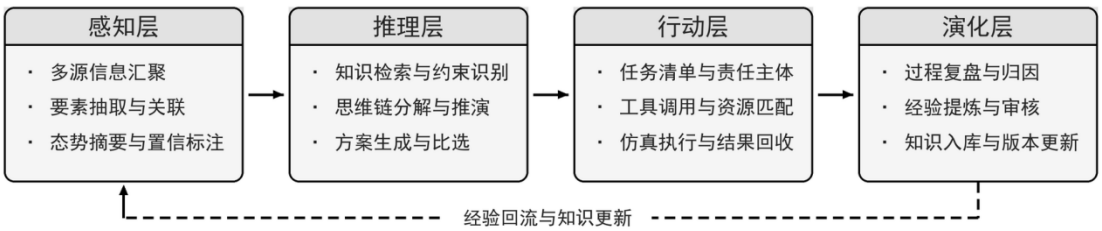


图 1 LLM 辅助决策的“感知—推理—行动—演化”闭环

Fig. 1 The perception–reasoning–action–evolution loop of LLM-based decision support

感知环节的任务是把分散、异构、碎片化的信息转化为可用的态势认知。突发事件的信息来源包括监测预警数据、接报信息、地理与气象数据、资源与队伍台账，以及社交媒体上的公众反馈。传统系统依赖结构化字段录入，大量非结构化信息被排除在外。LLM 可以借助语义理解能力，从文本、表格与对话记录中抽取事件要素（时间、地点、类型、规模、影响范围），识别关键实体及其关系，并将多源信息压缩为分层的态势摘要。同时，模型可以标注信息的时间、来源与置信度，为后续推理保留证据基础。感知环节的难点不在于“读得多”，而在于“分得清”：哪些信息已经确认，哪些仍是传言，哪些相互矛盾，必须在摘要中显式区分，否则错误会沿着证据链向下游传导。

推理环节是决策的核心。系统首先依据当前态势从知识库中检索相关的法规条款、预案要求、历史案例与专家经验，检索结果作为生成方案的约束条件与参考依据[5]。随后，模型通过思维链将复杂问题分解为“目标设定—约束识别—方案生成—后果推演—方案比选”等步骤[4]；对于存在多条处置路径的情形，可以借助树状搜索展开不同分支并比较其后果[11]。与纯规则系统相比，这种推理不是简单的条件匹配，而是在不完全信息下形成可修正的判断，并把判断依据显性化，使决策者能够

追问“为什么”。当检索不到可靠依据时，系统应明确指出知识缺口，而不是用流畅的表述掩盖不确定性。

行动环节把方案转化为可执行的任务。模型将处置方案拆解为任务清单，明确责任主体、时限要求与资源需求，并通过接口写入演练系统或业务系统；在仿真环境中，方案可以被执行并回收结果，形成“提出—执行—观察”的快速循环。工具调用能力使模型能够连接地理信息、路径规划、资源优化等专用模块，把语言层面的建议转化为可计算的行动方案[12]。在此过程中，关键节点应设置人工确认，避免模型在缺乏授权的情况下直接触发实际处置动作。行动环节的设计原则是“建议可以自动，执行必须受控”。

演化环节决定了系统能否持续改进。每次演练或处置结束后，系统依据结果与过程记录进行复盘：哪些判断被后续信息证实，哪些假设被证伪，哪些建议被采纳或被修改。反思机制使模型能够把成功经验提炼为可复用的策略，把失败教训转化为检查项[7]；这些经验以案例、提示模板与知识条目的形式沉淀下来，反哺下一轮感知与推理。需要指出的是，演化不等于让模型自行改写规则。错误经验一旦被固化，可能在后续场景中被反复放大，因此知识入库应经过专家审核与版本管理，重要条目的修改应保留理由与责任人。

综合来看，闭环的有效性并不取决于模型规模的大小，而取决于四个环节之间的衔接质量：感知是否为推理提供了足够且可靠的证据，推理是否给出了可检验的依据，行动是否保留了人的控制权，演化是否把经验转化为可复用的知识。离开这些条件，所谓闭环很容易退化为“模型自言自语”——输出看似完整，却既无法验证，也无法沉淀。

2.3　多 LLM 协同：从单点能力到体系能力

单一模型难以独立承担完整的决策支持任务。一方面，模型的上下文窗口有限，难以同时容纳海量态势信息、法规文本与历史案例；另一方面，不同任务对模型能力的要求差异很大，规划、生成、调度与评估混杂在一个模型中，容易相互干扰。为此，系统通常采用多 LLM 协同架构，通过角色分工与协议约束，把单点能力组织为体系能力[13-15]。这一思路与多代理系统的研究传统一脉相承，但其重点不在于增加模型数量，而在于让不同单元在明确的职责边界内交换结构化信息。

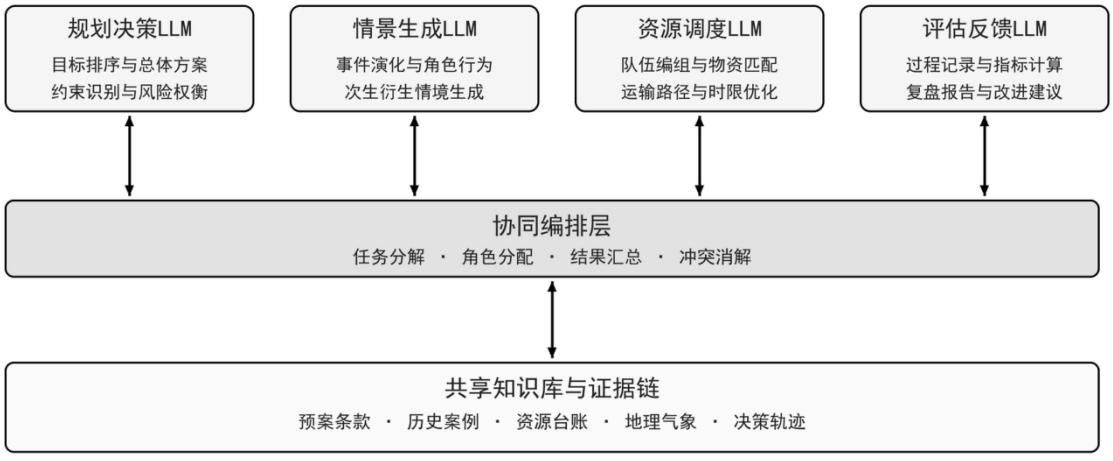


图 2　多 LLM 协同的辅助决策架构

Fig. 2　Architecture of multi-LLM collaborative decision support

一种可行的划分是设置四类模型单元（图 2）。规划决策单元负责全局态势评估、目标排序与总体方案设计；情景生成单元负责推演事件演化、生成次生衍生情境与参演角色的行为；资源调度单

元负责队伍编组、物资匹配与运输路径优化；评估反馈单元负责过程记录、指标计算与复盘报告生成。四类单元共享统一的知识库与证据链，并通过协同编排层完成任务分解、角色分配、结果汇总与冲突消解。这种划分的依据是任务的认知性质，而不是组织机构的设置，因此可以根据实际需要增减单元，例如在公共卫生事件中增设舆情研判单元。

协同机制包括三个方面。其一是信息共享。各单元使用统一的态势摘要与证据编号，避免各自维护一套事实版本，减少“同一事件、多种说法”的情况。其二是任务协同。编排层根据任务依赖关系把子任务分配给相应单元，并规定输入输出的格式与时限；当某个单元的输出不满足约束时，编排层可以要求其重做或转交其他单元。其三是冲突消解。当不同单元给出相互矛盾的判断时，系统通过证据比对、交叉质询与置信度加权形成初步结论；对于影响重大的分歧，则提交人类决策者裁决，并保留完整的争论文本供事后复盘。

多 LLM 协同的优势在于分工带来的专业性、交叉校验带来的可靠性，以及并行处理带来的时效性。但协同也会引入新的问题：通信开销随单元数量增加而上升，错误可能在单元之间传播并被放大，责任归属也变得更加模糊。因此，协同架构的设计应遵循“必要分工”原则，单元数量与任务复杂度相匹配，并保证每个关键结论都能追溯到具体的证据与责任人。在工程实现上，可以通过共享上下文、结构化消息格式与统一的工具接口降低协同成本，而不是让各单元以自然语言自由对话。

3 情景化演练：三类应用模式

3.1 自适应博弈：构建高压力情境

演练的价值取决于情境的逼真程度与压力水平。脚本化演练之所以难以达到训练目的，根本原因在于它把“不确定性”排除在外：参演者知道事件会如何发展，也知道自己的动作会得到什么回应，演练退化为流程演示。LLM 辅助决策系统可以通过“挑战方”角色，把不确定性重新引入演练。这里的挑战方不是固定的题库，而是能够根据参演者行动调整策略的生成机制。

自适应博弈是其中的核心机制。系统不是重复固定的“剧情”，而是根据参演方的行动动态调整事件演化。当参演方加强了某一环节的处置，挑战方会寻找其薄弱环节施加压力；当参演方调整策略，挑战方会识别变化并作出回应。这种“你来我往”的博弈依托两类技术：一是基于规则的约束生成，确保事件演化符合物理规律、法规要求与社会常识；二是基于反馈的策略优化，利用演练数据调整挑战方的行为倾向，使博弈强度始终处于“略高于当前能力”的区间。二者缺一不可：只有规则约束，情境会重新僵化；只有策略优化，情境可能滑向不合常理的极端。

多层次难度设置可以满足不同对象的训练需求。基础级情境以单一灾种、局部影响为主，信息相对完整，重点训练流程规范与基本判断；进阶级情境引入多灾种叠加、跨区域联动与信息冲突，重点训练方案比选与资源统筹；挑战级情境则设置复合链式事件、极紧时间窗与多目标竞争，重点训练临机决断与全局协调。难度的调整不仅体现在事件规模上，更体现在信息完备度、时间压力与资源约束的组合方式上（表 1）。

表 1　演练情景参数配置示例

Table 1　Configuration examples of drill scenario parameters

情景要素	基础级	进阶级	挑战级
事件类型与规模	单一灾种，局部影响	多灾种叠加，跨区域影响	复合链式、跨区域联动
信息完备度	关键信息基本齐全	部分信息缺失或相互冲突	信息过载与虚假信息并存
时间压力	小时级，节奏可控	分钟级，阶段压缩	分钟级且持续变化
资源约束	资源基本充足	部分资源短缺	多目标竞争，资源硬约束
演化不确定性	路径基本确定	存在 2～3 条演化分支	多分支并发、概率变化
训练重点	流程规范与基本判断	方案比选与资源统筹	统筹协调与临机决断

注：表中参数为示例性配置，实际演练应结合灾种特征、区域条件与训练目标调整。

复合与非对称情境拓展了演练的维度。真实事件中，参演方往往需要同时应对次生灾害、谣言传播、公众质疑与部门间分歧，而这些情形在传统演练中难以低成本复现。LLM 可以依据历史案例与领域知识生成合理的事件链，扮演媒体、公众、专家等不同角色[16]，提出带有情绪色彩或利益诉求的质询，迫使参演者在信息甄别、沟通表达与价值权衡中作出选择。应当明确，情境生成必须接受专家设定的约束集审核，防止为了追求难度而制造违背常识的“神剧情”，否则会损害演练的信度，甚至让参演者形成错误的应对习惯。

3.2　人机协同：重构演练角色关系

LLM 的介入改变了演练中的角色结构。传统演练由组织者编写脚本、参演者执行动作、评估者事后打分，三者相互分离。引入 LLM 后，模型可以同时承担陪练、教练与镜像三种角色（图 3），使演练从“一次性表演”转变为“持续学习”的过程。三种角色对应三种不同的交互节奏：陪练发生在演练进行中，教练贯穿演练全程，镜像则集中在复盘阶段。

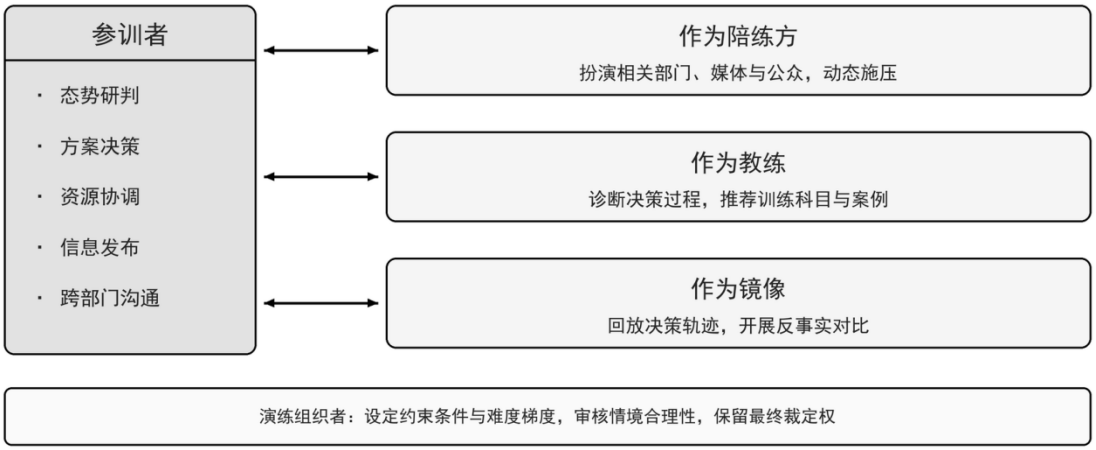


图 3　人机协同的演练模式

Fig. 3　Human–LLM collaborative drill mode

作为陪练方，LLM 承担部分角色的即时响应。在跨部门联动的演练中，参演者往往难以同时与多个部门、多类主体互动，模型可以扮演尚未到场的部门、上级机构、周边群众或媒体，给出符合其立场与信息条件的回应。参演者因此可以把注意力集中在统筹研判与关键决策上，而不必疲于应付“角色互动”。这种分工降低了演练的组织成本，也使小规模演练能够模拟大规模联动。当然，模型

扮演的角色应保持立场的一致性，不能在不同轮次中随意改变态度，否则会削弱情境的可信度。

作为教练，LLM 提供过程性指导。系统记录参演者的信息获取路径、判断时点与决策依据，识别其在哪些环节出现迟疑、遗漏或反复。例如，若参演者在态势研判阶段反复追加同类信息而迟迟不作判断，可能反映出对信息优先级缺乏把握；若其在资源调度中频繁调整方案，可能反映出对约束条件理解不清。针对这些问题，模型可以推荐相应的训练科目、提供相似案例的处置要点，并在演练结束后生成个性化复盘报告。指导的关键是“指出现象、给出证据、留出选择”，而不是直接给出标准答案。

作为镜像，LLM 促进反思性学习。系统可以回放参演者的决策过程，并把模型自身在相同情境下的推理路径并列呈现，使参演者看到自己的思维盲区与可选路径。更进一步，系统可以构造反事实场景——“如果当时优先疏散而不是先控制源头，结果会怎样”——让参演者在对比中理解决策的因果结构，而不是简单记住“标准答案”。这种反事实分析的价值不在于证明某种做法绝对正确，而在于帮助参演者建立条件化的判断能力：在什么条件下某种选择更合适，在什么条件下需要调整。

角色重构也带来边界问题。模型扮演的角色越多，参演者越容易把模型视为“全能角色”，从而产生两种偏差：一是过度依赖模型的提示，把演练变成“跟着提示走”；二是把模型的判断当作标准答案，忽视真实情境中的价值冲突。为避免这些偏差，应明确规定模型的角色权限与信息边界，模型不得提前泄露“剧情”走向，不得代替参演者作处置决定，其建议应标注依据与不确定性。演练组织者则应保留情境调整与结果裁定的最终权力。

3.3 数据驱动：从经验组织到科学设计

演练全过程的数据化，使训练从“凭经验组织”走向“依数据设计”。系统记录的不只是最终评分，还包括信息何时到达、参演者何时作出判断、调用了哪些资源、与哪些角色发生了交互。这些细粒度的过程数据，构成了分析训练效果的基础，也为评估从“结果导向”转向“过程导向”提供了条件。

能力画像刻画个体与团队的能力结构。依据过程数据，可以从态势研判、方案构造、资源统筹、沟通协调、风险控制等维度构建能力画像，并以能力剖面图、趋势图等形式呈现。与依赖主观印象的评分相比，过程数据支持的评估更为客观，也更容易定位具体短板。对于团队演练，还可以分析信息在成员之间的流转效率、决策权的实际分布与协同模式，从而判断团队是“分工明确、衔接顺畅”，还是“各自处置、反复确认”。

规律挖掘揭示训练中的结构性问题。通过对多轮演练数据的统计分析，可以发现一些仅凭观察难以察觉的规律：哪些环节的决策时延最长，哪些信息最容易被忽略，哪些角色在跨部门协同中成为瓶颈，哪类情境下的方案修改次数最多。这些规律既可以用于改进演练设计，也可以反馈到预案修订与岗位培训。需要注意的是，统计数据只能说明“哪里有问题”，不能自动说明“为什么”，因果解释仍需结合个案复盘。

效果预测优化训练资源配置。基于历史演练数据，系统可以估计不同训练方案的效果，例如增加某一科目的训练时长、调整难度梯度、改变团队编成，对能力提升的可能影响。需要提醒的是，这类预测依赖样本质量与模型假设，在数据有限时容易产生过度拟合。因此，预测结果应作为方案设计的参考，而不是资源分配的唯一依据；同时应通过后续演练对预测进行检验，形成“预测—实施—验证”的循环。

4 应急管理理论现代化：三大范式转变

4.1 从经验归纳到数据驱动

传统的应急管理研究主要遵循“案例复盘—经验总结—理论提炼”的路径[17,18]。这一路径的价值不容否认，其局限也十分明显：重大突发事件的样本数量有限，难以支撑大样本统计；每次事件都嵌在特定的制度、地理与社会条件中，结论的普适性存疑；复盘多在事后进行，容易受到结果导向的归因偏差影响。应急管理学科的发展历程也表明，理论概括往往滞后于实践创新[19,20]，而理论的滞后又会反过来削弱实践的解释力。

LLM 辅助决策系统为理论研究提供了新的“实验场”。系统可以在设定参数后批量生成虚拟案例，覆盖不同的致灾强度、响应时序、资源禀赋与信息条件，从而把“不可重复的历史”转化为“可重复的实验”。研究者可以设计对照实验：固定其他条件，只改变某一变量（如预警提前量、信息共享程度、跨部门协同机制），观察处置结果的差异，进而识别变量之间的因果关系。这种研究方式并不能取代真实案例，但可以弥补真实样本的不足，为理论命题提供更大范围的检验。

以“信息共享程度与处置效率的关系”为例。传统研究多通过少数案例的比较得出“共享越充分、效率越高”的结论，但这一结论在资源极度紧张或信息质量低下的条件下是否成立，缺乏系统检验。借助仿真实验，可以设置不同的共享范围、共享时延与信息噪声水平，观察其对研判准确率、响应时延与资源浪费的影响，从而识别共享的边际收益递减区间与适用边界。类似地，“先期处置窗口”的长度、“统一调度”与“属地负责”的配比等问题，也可以在可控条件下得到更细致的回答。

也应看到，虚拟案例的可信度取决于模型假设与参数设定。如果仿真规则本身偏离现实，大规模实验只会放大偏差。因此，虚拟实验必须与真实案例相互校准：用真实案例检验仿真结果，再用仿真结果解释真实案例中未被观察到的机制。这种“仿真—实证”互证的研究策略，比单纯依赖其中一种方法更为稳健，也更符合应急管理研究重视情境的传统。

4.2 从静态文本到动态演化

应急管理知识长期以预案、规范、手册等文本形式固化。文本化带来了权威性与可操作性，但也带来了刚性：预案一经发布便趋于稳定，修订周期往往以年计，而风险形态、技术条件与社会预期却在持续变化。结果是，预案文本与实际处置之间出现越来越大的缝隙，基层执行者不得不用“经验补丁”填补空白，而这些补丁又难以被系统地记录和传播。

LLM 辅助决策系统提供了一种动态的知识生产方式。系统的知识并非只存在于模型参数中，而是由模型能力、外部知识库与案例经验共同构成。每一次演练或处置产生的决策轨迹，都可以被抽取为结构化的“知识条目”——在什么条件下、依据什么证据、采取了什么行动、产生了什么结果。这些条目经过审核后进入知识库，成为下一轮推理的检索对象。知识因此不再是静态文本，而是随实践不断更新的“活文档”。

隐性的判断经验也可以被部分外化。资深决策者的许多能力难以用文字表述，例如对局势“不对劲”的直觉、对不同部门配合默契程度的把握。通过分析大量决策轨迹，可以提取出其中的共性模式：在哪些信号出现时会提高响应等级，在哪些条件下会暂缓大规模调动，在哪些环节会主动向上级报告。这些模式并非精确的规则，而是带有概率色彩的行为倾向，可以被转化为提示模板、检查清单或评估维度，供经验不足者参考。

动态演化还体现在策略的博弈更新上。在演练中，参演方与挑战方的策略相互作用，形成“策略—反策略—再策略”的循环。系统记录这一过程，可以识别哪些策略是稳定的，哪些会被迅速破解，哪些组合能够产生协同效应。这种分析把理论研究从“总结成功经验”推进到“刻画策略演化”，为预案的动态修订与处置策略选择提供依据。当然，演化必须有边界：知识更新需要版本管理、变更审计

与回溯机制，防止错误经验在系统中扩散。

4.3 从碎片化到体系化

应急管理研究长期按灾种与环节分领域展开：自然灾害、事故灾难、公共卫生、社会安全各有各的理论积累，监测预警、应急响应、恢复重建各有各的研究传统。分领域研究有利于专业化，但也容易形成“隔断”：各领域的知识难以整合，跨领域的耦合效应难以刻画，体系层面的涌现现象更难以解释。公共安全科技领域提出的框架方法，虽然强调多学科交叉与多环节衔接[21]，但在具体研究中仍缺少能够同时容纳多要素、多尺度、多主体的分析工具。

LLM 辅助决策系统为体系化研究提供了工具。系统可以集成致灾因子、承灾体、应急资源、信息传播、社会响应等多维模型，在统一的时间与空间框架下模拟事件的演化与处置。与单领域模型相比，体系化仿真能够呈现要素之间的相互影响：一次交通中断如何影响医疗资源的可达性，一条谣言如何改变公众的避险行为，一个部门的延迟响应如何通过联动机制放大为系统性延误。这些影响在单领域研究中往往被抽象掉，在体系仿真中则可以被观察和度量。

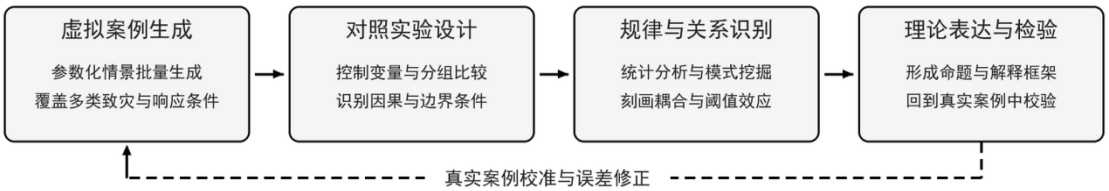


图 4 基于 LLM 辅助决策系统的应急管理理论研究框架

Fig. 4 Research framework for emergency management theory based on LLM-based decision support

体系化研究至少可以回答三类问题。其一是耦合关系。不同要素之间是替代、互补还是相互制约？在什么条件下耦合会由“缓冲”转为“放大”？其二是协同效能。不同协同模式（统一调度、分区负责、混合模式）在不同情境下的效能差异如何？协同收益是否存在阈值？其三是体系脆弱性。哪些节点一旦失效会引发连锁反应？哪些环节看似冗余实则是关键备份？这些问题的回答，有助于把“体系能力”从抽象概念转化为可测量的指标。

需要指出，体系化并不等于把所有要素都纳入模型。模型的价值在于有选择地简化，而不是无差别地复制现实。因此，体系研究的关键是确定“必要的抽象层级”：哪些要素必须显式建模，哪些可以作为参数或边界条件，哪些可以在结果层面进行校正。这一判断需要领域专家与建模人员共同完成，不能交给模型自行决定。过度建模不仅会推高计算成本，还可能因参数估计误差的累积而降低结论的可靠性。

5 系统建设的关键路径与保障

5.1 技术路径：从通用模型到领域模型

构建可用的 LLM 辅助决策系统，关键是把通用模型转化为理解领域语言、遵循领域规则、能够调用领域工具的专用系统。这一转化通常包括五个环节（图 5）：领域语料构建、增量预训练与指令微调、专家偏好对齐、知识库与检索增强、工具链与编排评测。五个环节并非严格串行，实践中往往需要根据评测结果回到前序环节迭代。

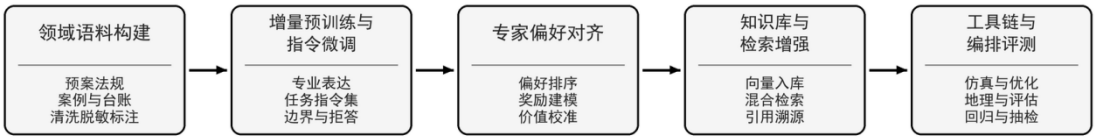


图 5　从通用模型到领域模型的技术路径

Fig. 5　Technical path from general-purpose models to domain-specific models

领域语料构建是基础。通用模型虽然掌握了丰富的语言知识，但对应急预案、处置规程、行业术语与地方实际的了解有限。需要系统收集法规预案、历史案例、专家问答、资源台账、地理与气象数据等资料，进行清洗、去重、脱敏与结构化标注，形成高质量的领域语料。语料的质量比数量更重要：错误或过时的文本一旦进入训练或检索，会在后续环节被反复引用，纠正成本远高于前期审核成本。

增量预训练与指令微调是核心。在领域语料上进行继续预训练，可以使模型熟悉专业表达与知识分布；在此基础上构造指令数据集，把“任务描述—输入信息—期望输出”组织成训练样本，通过监督微调提升模型遵循任务要求的能力。指令数据应覆盖研判、生成、比选、评估等典型任务，并包含边界情形与拒答示例，使模型学会在信息不足时明确表示不确定，而不是编造答案。

偏好对齐与知识增强是保障。通过专家对模型输出进行排序或评分，训练奖励模型并优化生成策略，可以使模型的回答更符合领域规范与价值取向[6]。同时，构建外部知识库，把预案条款、案例要点、资源信息向量化存储，通过混合检索在生成前引入相关证据，并在输出中标注引用来源[5,22]。这种“模型+知识库”的架构，既保留了模型的推理能力，又提高了事实的准确性与时效性。

工具链与编排评测是延伸。模型需要借助专用工具完成计算、查询与优化任务，例如路径规划、资源匹配、损失评估与地理分析[12]。通过工具调用接口，模型可以把语言建议转化为可执行的操作，并对执行结果进行解释。多单元协同则需要编排层统一管理任务流、上下文与权限，并通过回归测试、边界测试与人工抽检持续检验系统表现。任何一个环节的薄弱，都会限制系统的整体可用性。

5.2　评估体系：从结果评价到过程评价

评估是系统建设的“风向标”。如果只看处置结果，系统可能倾向于选择“看起来有效”的方案，而忽视过程中的判断质量；如果只看过程合规，又可能纵容“程序正确但效果不佳”的做法。合理的评估应兼顾结果与过程，并把人的因素纳入其中（表 2）。

表 2　LLM 辅助决策系统的评估指标体系

Table 2　Evaluation indicator system for LLM-based decision support systems

一级指标（权重）	二级指标	观测方式
过程质量（0.35）	态势要素识别完整性；研判结论与证据的一致性；方案对约束条件的满足程度；备选方案的充分性	轨迹日志分析、专家评分、规则校验
处置效果（0.35）	关键环节响应时延；损失控制程度；资源使用效率；秩序恢复速度	仿真指标、时间戳、资源台账
人机协同（0.20）	建议采纳率；建议修正率；人工接管次数及其效果；信任与依赖水平	交互日志、问卷与访谈
可解释与合规（0.10）	依据可追溯率；同类情境结论一致性；权限与流程合规性	证据链审计、复评与交叉验证

注：权重为示例值，实际应用应结合演练目标、灾种特征与专家判断动态调整。

结果指标关注处置的最终效果，包括关键环节的响应时延、损失控制程度、资源使用效率与秩序恢复速度。这些指标直观、易比较，但受情境难度影响较大，单独使用容易产生误判。过程指标关注决策质量，包括态势要素识别的完整性、研判结论与证据的一致性、方案对约束条件的满足程度、备选方案的充分性以及沟通协调的有效性。过程指标更能反映能力，但采集与评分成本较高，需要借助日志分析与专家评审。已有研究尝试以模型作为评分者对开放式回答进行批量评价[23]，但其与专家评分的一致性仍需检验，宜作为初筛与辅助，而不宜直接作为终评依据。

人机协同指标是新增维度。系统给出的建议是否被采纳、被修改了多少、在哪些环节需要人工接管、接管后效果如何，这些数据既反映了系统的可用性，也反映了使用者的信任程度。若采纳率过高而修正率极低，可能意味着使用者过度依赖系统；若修正率长期偏高，则说明系统建议与实际需求脱节。两类偏差都需要通过界面设计、培训与模型迭代加以纠正，而不能简单归因于“系统不好用”或“使用者不配合”。

评估实施应遵循三项原则。一是权重动态化。不同阶段、不同科目的评估重点不同：基础训练更关注流程规范，综合演练更关注统筹协调，复盘阶段更关注改进空间。权重应随训练目标调整，并在评估前明确告知。二是依据可追溯。每一项评分都应对应具体的行为记录与证据编号，评估者可以看到“为什么得这个分”，受评者也可以提出申辩。三是信效度检验。通过专家评分一致性分析、复评与交叉验证，检验评估工具是否稳定、是否真正测量了想要测量的能力。

5.3 安全保障：从技术安全到伦理规范

LLM 辅助决策系统涉及大量敏感信息与重要判断，安全保障必须贯穿系统全生命周期（图 6）。保障体系可以分为数据、系统与伦理三个层面，三者相互支撑，缺一不可。数据安全解决“信息不被滥用”，系统安全解决“服务不中断、输出不越界”，伦理规范解决“决定由谁作出、责任由谁承担”。



图 6 LLM 辅助决策系统的多层次安全保障体系

Fig. 6 Multi-level safeguard system for LLM-based decision support systems

数据安全是底线。应建立数据分级分类制度，按照敏感程度确定采集、存储、使用与共享的规则；对个人信息与敏感业务数据实施脱敏、去标识化与访问控制；在数据分析与模型训练中采用差分隐私等技术，降低个体信息被推断的风险；通过审计日志记录数据访问与使用行为，确保可追溯、可问责。跨部门数据共享应遵循“最小必要”原则，明确共享范围与使用目的，防止数据在流转中失控。

系统安全是保障。应采用模块化架构，把模型服务、知识库、工具链与业务系统解耦，降低单点故障的影响；建立降级运行机制，在模型服务不可用时能够回退到规则引擎或人工流程，保证业

务连续性；对模型输出设置过滤与校验环节，拦截明显错误、越权建议与不当内容；定期开展压力测试与安全评估，检验系统在异常输入、数据污染与高并发条件下的稳定性。

伦理规范是约束。LLM 在应急领域的应用涉及生命权、公平性与责任归属等敏感问题，必须坚持“人在回路”原则：涉及人员疏散、资源调配、信息发布等重大决定，应由具备权限的决策者作出，模型只提供建议与依据。应建立算法审查机制，对模型的训练数据、评估结果与典型输出进行定期审查，防止系统性偏见；明确责任边界，模型出错时由使用机构承担相应责任，而不能以“技术原因”推诿；做好告知与解释，让公众理解决策的依据与局限。

还需警惕技术依赖带来的能力退化。如果参演者长期依赖模型给出方案，其独立研判与临机决断能力可能下降；如果组织把评估完全交给系统，其对复杂情境的敏感度也可能钝化。因此，系统设计应保留“无提示”训练环节，定期开展不依赖模型的基础能力考核，并在制度上明确模型的使用边界。

6 讨论：局限与边界

本文提出的框架仍面临若干限制。首先是事实性问题。LLM 可能生成与事实不符的内容，在应急场景中，错误信息可能被迅速放大。检索增强与引用溯源可以缓解这一问题，但不能完全消除；关键结论仍需人工核验，系统应明确标注不确定性，并在证据不足时给出保守建议。

其次是评估效度问题。演练中的表现能否迁移到真实处置，取决于情境的逼真度、压力水平与参演者的心理状态，而这些都是仿真难以完全复现的。把演练评分直接等同于真实处置能力，可能导致“演练表现好、实际处置差”的偏差。因此，评估体系应把演练数据与真实处置的复盘结果结合起来，通过长期跟踪检验其预测效度。

第三是数据与组织壁垒。跨部门数据共享涉及权责、安全与利益，现实中往往难以打通；领域语料的积累需要长期投入，短期内难以形成规模。第四是成本与公平。高性能模型的部署、维护与更新需要持续投入，资源有限的基层单位可能难以负担，从而加剧能力差距。第五是责任与信任。当模型建议与人类判断冲突时，如何界定责任、如何建立合理的信任水平，仍缺乏成熟的制度安排。

这些限制并不意味着应当放缓技术应用，而是提示我们需要同步推进三项工作：一是建立面向应急领域的模型评测基准，覆盖事实性、推理质量、鲁棒性与价值对齐；二是完善人机协同的操作规程，明确何时使用、如何使用、何时停用；三是加强跨学科研究，把应急管理、计算机科学、心理学与伦理学的视角结合起来，避免技术方案脱离组织实际。

7 结论

本文围绕 LLM 辅助决策系统在情景化演练与应急管理理论现代化中的作用进行了系统分析，主要结论如下。

（1）LLM 辅助决策系统实现了从规则匹配到认知协同的转变。其核心机理是“感知—推理—行动—演化”的闭环：感知把碎片信息转化为可核查的态势认知，推理把领域知识与多步分析结合为可解释的方案，行动把方案转化为可执行的任务并保留人的控制权，演化把经验沉淀为可复用的知识。闭环的有效性取决于环节之间的衔接质量，而非模型规模。

（2）在演练层面，系统通过自适应博弈、人机协同与数据驱动三种模式，提升了情境的逼真度、指导的针对性与训练设计的科学性。LLM 既是陪练方，也是教练与镜像，但其角色权限必须受到约束，处置决定权始终属于人。

（3）在理论层面，LLM 推动应急管理研究从经验归纳走向数据驱动、从静态文本走向动态演

化、从碎片化走向体系化。虚拟案例与对照实验扩展了理论检验的空间，知识条目与决策轨迹提供了动态更新的素材，体系仿真使耦合关系、协同效能与脆弱性分析成为可能。

（4）系统建设需要技术、评估与治理的协同推进。领域语料、指令微调、知识增强与工具链构成技术路径；结果与过程并重的评估体系提供质量保障；数据、系统与伦理三个层面的机制构成安全底线。

展望未来，多模态信息融合将使系统能够处理图像、视频与语音等更丰富的输入；因果推断与可解释方法将提升结论的可靠性；多 LLM 协同与工具生态的成熟将扩展系统的任务边界。与此同时，制度建设、人才培养与伦理规范需要同步跟进。技术只是手段，其价值最终取决于能否帮助人们在复杂情境中作出更负责任的判断。

参考文献

[1] 国务院. “十四五”国家应急体系规划[Z]. 2022-02-14.

[2] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Advances in Neural Information Processing Systems, 2017, 30: 5998-6008.

[3] Brown T B, Mann B, Ryder N, et al. Language models are few-shot learners[C]//Advances in Neural Information Processing Systems, 2020, 33: 1877-1901.

[4] Wei J, Wang X, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models[C]//Advances in Neural Information Processing Systems, 2022, 35: 24824-24837.

[5] Lewis P, Perez E, Piktus A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks[C]//Advances in Neural Information Processing Systems, 2020, 33: 9459-9474.

[6] Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback[C]//Advances in Neural Information Processing Systems, 2022, 35: 27730-27744.

[7] Shinn N, Cassano F, Gopinath A, et al. Reflexion: Language agents with verbal reinforcement learning[C]//Advances in Neural Information Processing Systems, 2023, 36: 8634-8652.

[8] Bommasani R, Hudson D A, Adeli E, et al. On the opportunities and risks of foundation models[EB/OL]. (2021-08-16)[2026-10-01]. https://arxiv.org/abs/2108.07258.

[9] Wei J, Tay Y, Bommasani R, et al. Emergent abilities of large language models[J]. Transactions on Machine Learning Research, 2022(8): 1-30.

[10] Ji Z, Lee N, Frieske R, et al. Survey of hallucination in natural language generation[J]. ACM Computing Surveys, 2023, 55(12): 1-38.

[11] Yao S, Yu D, Zhao J, et al. Tree of thoughts: Deliberate problem solving with large language models[C]//Advances in Neural Information Processing Systems, 2023, 36: 11809-11822.

[12] Yao S, Zhao J, Yu D, et al. ReAct: Synergizing reasoning and acting in language models[C]//Proceedings of the 11th International Conference on Learning Representations, 2023: 1-17.

[13] Hong S, Zhuge M, Chen J, et al. MetaGPT: Meta programming for a multi-agent collaborative framework[C]//Proceedings of the 12th International Conference on Learning Representations, 2024: 1-25.

[14] Wu Q, Bansal G, Zhang J, et al. AutoGen: Enabling next-gen LLM applications via multi-agent conversation[EB/OL]. (2023-08-16)[2026-10-01]. https://arxiv.org/abs/2308.08155.

[15] Wang L, Ma C, Feng X, et al. A survey on large language model based autonomous agents[J]. Frontiers of Computer Science, 2024, 18(6): 186345.

[16] Park J S, O'Brien J C, Cai C J, et al. Generative agents: Interactive simulacra of human behavior[C]//Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, 2023: 1-22.

[17] 张海波, 童星. 中国应急管理结构变化及其理论概化[J]. 中国社会科学, 2015(3): 58-84.

[18] Boin A, 't Hart P, Stern E, et al. The politics of crisis management: Public leadership under pressure[M]. 2nd ed. Cambridge: Cambridge University Press, 2017: 1-30.

[19] 闪淳昌, 薛澜. 应急管理概论：理论与实践[M]. 北京: 高等教育出版社, 2012: 45-78.

[20] 薛澜, 刘冰. 应急管理体系新挑战及其顶层设计[J]. 国家行政学院学报, 2013, (1): 10-14.

[21]范维澄, 刘奕, 翁文国. 公共安全科技的“三角形”框架与“4+1”方法学[J]. 科技导报, 2009, 27(6): 3.
[22]Gao Y, Xiong Y, Gao X, et al. Retrieval-augmented generation for large language models: A survey[EB/OL]. (2023-12-18)[2026-10-01]. https://arxiv.org/abs/2312.10997.
[23]Zheng L, Chiang W L, Sheng Y, et al. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena[C]//Advances in Neural Information Processing Systems, 2023, 36: 46595-46623.

[1] **第一作者简介：**杨以恒（2004—），男，汉族，山东青岛人。研究方向：人工智能与电子信息工程。E-mail：yangyiheng2004@foxmail.com

[*] **通讯作者简介：**熊伟男（1992—），女，汉族，黑龙江牡丹江人。研究方向，智能决策。E-mail：xiongweinan6616@163.com