【电子与信息科学／Electronics and Information Science】

融合大语言模型与向量知识库的应用文生成框架

秦斌[1]，陆平[2]，徐琰[2]，邓芳伟[2]，王旖洋[3]，曾渭钰[3]，李欣莹[3]，李灿亮[3]

1）深圳大学信息中心，广东深圳 518060；2）中兴通讯股份有限公司，广东深圳 518057；3）深圳大学电子与信息工程学院，广东深圳 518060

摘　要：为提高应用文编写效率，提出一种融合大语言模型(large language model，LLM)与向量知识库(vector knowledge base)的应用文自动生成框架．根据目标应用场景，以人工编写的标准应用文为范本，构建结构化辅助生成文件，并建立相应类型应用文的向量知识库．利用目标类型应用文的章节标题和用户输入的关键信息在知识库中进行检索，匹配相关文段；设置提示词引导LLM，以召回的参考文段及用户输入的提示信息为参考，使用末级标题作为分割标志，分章节生成应用文文本；最终按规定格式整合全文并输出完整的目标应用文．以应急预案为例，在同一评价标准下使用ChatGPT-4Turbo进行评测，自动生成的应急预案高度趋近于人工编写的质量，二者的文档质量相似度达95.87%．所提方法能够在算力资源有限的情况下突破字数限制，生成符合基本标准的长篇幅应用文，可供人工参考或直接使用，极大提高了编写人员的工作效率．

关键词：人工智能；应用文生成；大语言模型；向量知识库；提示词工程；模型评测；ChatGPT-4Turbo；DeepSeek-R1

中图分类号：TP391.1　　**文献标志码：**A　　**DOI：**10.3724/SP.J.1249.2025.05597

A framework integrating large language models and vector knowledge bases for practical document generation

QIN Bin[1], LU Ping[2], XU Yan[2], DENG Fangwei[2], WANG Yiyang[3], ZENG Weiyu[3], LI Xinying[3], and LI Canliang[3]

1) Information Center, Shenzhen University, Shenzhen 518060, Guangdong Province, P. R. China

2) Zhongxing Telecommunication Equipment Corporation, Shenzhen 518057, Guangdong Province, P. R. China

3) College of Electronics and Information Engineering, Shenzhen University, Shenzhen 518060, Guangdong Province, P. R. China

Abstract: To improve the efficiency of practical document writing, this study proposes an automated generation method that integrates large language models (LLMs) and vector knowledge bases. In diverse application scenarios, manually prepared standard documents are used as templates to construct structured auxiliary files to support document generation and vector knowledge bases tailored to specific document types. By leveraging target-type document chapter headings and user-provided key information, the system queries the knowledge base to retrieve relevant text segments. Prompt engineering is then applied to guide the LLM, which synthesizes coherent text by

Received: 2025-04-24; **Accepted:** 2025-07-11; **Online** (CNKI)**:** 2025-08-04

Foundation: University-industry Cooperation Foundation of ZTE Government Industry Large Language Model Technology Research (IA20231030016); National Key Research and Development Program of China (2020YFB1806405); Shenzhen Science and Technology Major Project (KJZD20230923114906013)

Corresponding author: Professor of engineering QIN Bin (qinbin@szu.edu.cn)

Citation: QIN Bin, LU Ping, XU Yan, et al. A framework integrating large language models and vector knowledge bases for practical document generation [J]. Journal of Shenzhen University Science and Engineering, 2025, 42(5): 597-605. (in Chinese)

retrieved reference segments with user input, generating content chapter by chapter in alignment the lowest-level heading structure. The generated texts are subsequently formatted into complete documents in compliance with predefined standards. Evaluation results demonstrate that, using emergency plans as a benchmark, when assessed under identical criteria with ChatGPT-4Turbo, the automatically generated emergency plans achieve 95.87% similarity to manually prepared counterparts, demonstrating comparable quality. The proposed method can overcome token limitations even under constrained computational resources, generating lengthy practical documents that meet baseline requirements. These outputs serve as reliable references or be directly adopted, significantly enhancing efficiency of document preparation.

Key words: artificial intelligence; practical document generation; large language model; vector knowledge base; prompt engineering; model evaluation; ChatGPT-4Turbo; DeepSeek-R1

应用文作为一种重要的信息载体，在各领域中扮演着举足轻重的角色．然而，人工编辑应用文本费时费力，过程繁琐且对编辑人员的专业知识要求很高．因此，研究一种高效且通用的应用文自动生成方法具有重要的应用价值．

近年来，自然语言处理(natural language processing，NLP)在文本生成方面取得了显著进展，大语言模型(large language model，LLM)生成的高质量文本已能够应用于部分实际场景．例如，王进强等[1]提出基于注意力机制的结构化文本生成方法，使用标书相关的特定语料训练模型，基于Transformer的双向编码器表示(bidirectional encoder representations from transformers，BERT)模型和TextRank算法生成结构化文本．然而，此类方法通用性不足，语料库的构建及模型训练耗时较多，模型的背景知识缺乏时效性．闫盈盈等[2]提出利用生成式人工智能构建患者药品说明书，通过设计药品说明书框架，调用“通义千问”2.0大模型分段生成文本，结构化地生成患者药品说明书，但该方法局限性在于参考文段直接由输入获得，未融合更广泛的医学背景知识，生成的文本内容较单一．智能合同起草平台通过训练Agent模型，或结合人工智能搜索与大数据知识图谱，提升了合同编写的便利性[3]．但此类智能合同起草平台的应用场景有限，且存在一定的信息泄露风险．

此外，当前中文语境下面向公众开放的商业化LLM，普遍存在着一定的技术瓶颈，单次输出篇幅受限，导致生成的内容出现细节缺失或结构不完整的问题．同时，随着生成文本长度(l)的增加，LLM的模型幻觉，即输出与输入信息无关或违背事实逻辑的现象会加剧．对该现象内在机制的分析可以从两个方面进行．在注意力机制层面，模型对早期关键信息的捕捉能力随l的增加而下降．这一注意力分散现象可用如式(1)的注意力权重熵H_t来量化．

$$H_t = -\sum_{s=1}^{t-1} P(s|t)\,\mathrm{lb}\,P(s|t) \tag{1}$$

其中，$P(s|t)$为生成第t个词时对第s个词的注意力权重．随着t增大，H_t随之上升，模型对关键信息的聚焦能力下降，幻觉加剧．在解码策略层面，LLM通过自回归方式逐词生成文本，每一步的局部生成误差(ϵ_t)会随l的增大累积为全局误差

$$E(l) = \sum_{t=1}^{l} \epsilon_t + \lambda \sum_{t=1}^{l-1} \epsilon_t \epsilon_{t+1} \tag{2}$$

其中，误差耦合系数λ反映了因上下文依赖产生的非线性放大效应．当l较大时，$E(l)$将被放大，导致生成内容偏离预期．模型幻觉会显著降低生成内容的质量，因此，适当缩短l是一种保证生成文本质量的可行方法．

针对上述问题，本研究基于LangChain框架[4]，提出一种融合LLM与向量知识库的应用文生成框架．使用向量数据库存储具有时效性的目标类型应用文背景知识．采用检索增强生成(retrieval-augmented generation，RAG)[5]方法，以向量数据库高效的检索能力召回目标知识，提升了LLM生成答案的准确性和上下文适配性，较传统主流方法的灵活性和通用性更高．为突破模型单次输出的字数限制，避免生成文本过长导致模型幻觉，基于结构化拆分处理思想，采用一种分步式生成策略完成长篇幅应用文的生成任务．

在LLM性能评测领域，特定指标评测[6-8]、基准评测[9-10]、模型评测[11-12]与人工评测[13]为目前主流的评测方法．鉴于生成式应用文无通用的特定指标，也无适配的基准评测数据集，本研究选择模型评测方法．首先介绍融合LLM与向量知识库的应用文生成框架的数学模型和算法流程，随后提出一

套针对应用文本的评价标准和评测方法，最后以生成应急预案为目标任务进行实验，评估所提应用文生成框架的有效性.

1　应用文生成方法

图 1 为本研究提出的融合 LLM 与向量知识库的应用文生成框架. 该框架采用结构化拆分策略处理应用文本，通过分步生成策略生成各章节内容，再使用文本语义对齐算法整合长文本的结构. 这种“分而治之”的策略既能充分利用 LLM 强大的语言生成能力，又能有效抑制生成长文本时的模型幻觉，保证长篇幅应用文输出的稳定性和高质量.

该框架依赖于内置的向量知识库，它存储了多个目标类型应用文的高质量章节文段. 通过结构化辅助生成文件统一管理应用文的章节框架、参考文段、生成提示词及提示信息. 以生成包含两级标题的应用文为例，其结构化辅助生成文件的结构如图 2. 其中，{一级标题}与{二级标题}为目标应用文的实际章节标题，每个{二级标题}下维护着 3 个与该章节内容生成有关的数据项：{content}记录了从向量知识库中检索并匹配得到的与当前章节相关度最高的参考文段；{prompt}存储了生成该章节内容时使用的提示词；{user_input}记录了从用户输入中提取的关键提示信息. {content}、{prompt}和{user_input}共同构成生成提示词组. 章节框架和生成提示词将依据用户所选应用文类型的格式规范进行配置：章节框架遵循目标类型应用文的文体结构，而生成提示词参照目标类型应用文中特定章节的内容特征和表述习惯来确定.

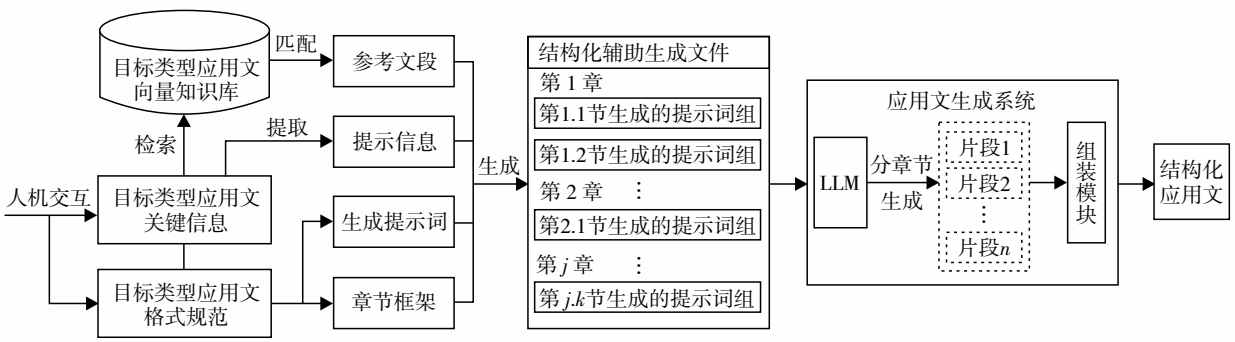

图 1　融合大语言模型与向量知识库的应用文生成方法框架

Fig. 1　Framework for practical document generation method based on large language model and vector knowledge base.

```
{
  “{一级标题}”:{
    “{二级标题}”:{
      “content”: “{content}”,
      “prompt”: “{prompt}”,
      “user_input”: “{user_input}”,
      “generate_data”: “{generate_data}”,
    },
  },
}
```

图 2　结构化辅助生成文件的结构

Fig. 2　Composition of structured file for assisting document generation.

生成应用文时，首先通过人机交互从前端获取用户输入的关键信息和格式规范，用于从向量知识库中匹配到与目标类型应用文各章节相关度最高的参考文段；随后，基于匹配到的参考文段和用户提示信息，使用 LLM 分步生成各章节内容；最终，使用文本语义对齐算法，将结构化辅助生成文件中的各级标题与生成的对应章节内容进行整合，并输出完整的长篇幅应用文. 所提框架主要由人机交互、知识库构建、知识库检索和文本生成 4 个核心部分组成.

1.1　人机交互

人机交互通过表单填写实现. 用户在前端界面中填写生成目标类型应用文所需的关键信息，所填数据构成集合 $X = \{x_i \mid i = 1, 2, \cdots, n\}$. 其中，$x_i$ 为用户填写的第 i 条数据；n 为用户填写的总数据条数.

X 中的部分数据将同步存储至结构化辅助生成文件中的 1 个或多个{user_input}. 对于需要同步的数据项 x_i，其处理过程可表示为如式(3)的映射关系.

$$x_i \rightarrow \text{user\_input}_{jk}, \quad \forall (m_j, n_k) \in C_i \tag{3}$$

其中，user_input_{jk} 为结构化辅助生成文件中一级标题为 m_j，二级标题为 n_k 时对应的{user_input}；C_i 为需要同步数据项 x_i 的所有一级标题与二级标题的

集合.

1.2 知识库构建

为不同应用场景构建对应的向量知识库时，采用以下步骤：首先，采用递归文本分割方法，将人工编写的标准范文(T)按句子和段落等自然边界初步分割为多个文本块. 若某个文本块长度超出预设字符数量限制(L)则对其进行递归分割，直至所有文本块均满足长度要求. 然后，利用预训练的嵌入模型将每个满足要求的文本块T_u转换为对应的高维向量表示$\boldsymbol{d}_u$，以保留其语义信息. 该转换过程表示为

$$\boldsymbol{d}_u = \mathrm{Embed}(T_u), \quad \forall T_u \in \mathrm{Split}(T, L) \tag{4}$$

其中，函数Embed为嵌入模型实现的向量映射；$\mathrm{Split}(T, L)$为对T应用递归分割方法后得到的所有文本块的集合.

最后，将所有生成的文本块的高维向量存储在向量知识库中，从而完成知识库构建.

1.3 知识库检索

获取对应用文生成有参考价值的相关文本块，关键在于有效的知识库检索. 具体而言，依据结构化辅助生成文件中的章节顺序，依次将每个章节的{一级标题}、{二级标题}和{user_input}组合并向量化为查询向量$\boldsymbol{q}$，然后采用如式(5)的欧式距离作为相似度度量在知识库中进行检索.

$$L(\boldsymbol{d}_u, \boldsymbol{q}) = \sqrt{\sum_{v=1}^{\mu_u} \left(\boldsymbol{d}_u[v] - \boldsymbol{q}[v]\right)^2} \tag{5}$$

其中，$d_u[v]$为目标类型应用文知识库内文段向量$\boldsymbol{d}_u$的第v个元素；μ_u为文段向量$\boldsymbol{d}_u$的长度.

最后，选择令$L(\boldsymbol{d}_u, \boldsymbol{q})$最小的$\boldsymbol{d}_u$，即与该章节标题和提示信息相关度最高的文本块作为检索结果，并存入该章节的{content}.

1.4 文本生成

对于一级标题为m_j、二级标题为n_k的章节，基于检索到的知识库文段$\mathrm{content}_{jk}$、关联的用户输入数据$\mathrm{user_input}_{jk}$和内置的提示词prompt_{jk}，调用大模型生成文本内容$\mathrm{generate_data}_{jk}$的过程可表示为

$$\mathrm{generate\_data}_{jk} = \mathrm{LLM}\left(P\left(\mathrm{content}_{jk}, \mathrm{prompt}_{jk}, \mathrm{user\_input}_{jk}\right)\right) \tag{6}$$

其中，P为套用提示词模板构建生成提示词组的函数，LLM为大语言模型的文本生成函数.

通过上述分步式生成策略生成各章节内容后，采用文本语义对齐算法，即在保持上下文语义一致性的条件下，按照结构化辅助生成文件中的章节顺序，依次整合各级标题与各章节内容，进而生成完整的长篇幅应用文R. 该过程可表示为

$$R = \sum_{j,k}^{(m_j, n_k) \in C} \left(m_j + n_k + \mathrm{generate\_data}_{jk}\right) \tag{7}$$

其中，C为目标应用文所有一级和二级标题的集合.

2 评测方法

为评估所提应用文生成框架的效果，建立对生成式应用文的评价标准并提出对应的评测方法，最后以生成式应急预案的评测为例进行分析和验证.

2.1 应用文评价

2.1.1 应用文评价标准

参考文献[14]，本研究采用完整性(E_1)、一致性(E_2)、相关性(E_3)、清晰性(E_4)和规范性(E_5) 5个指标评价所生成文件的质量. 其中，完整性指生成的应用文需包含预设文章结构的所有必要部分，如生成的应急预案需涵盖总则、应急组织体系、应急预警、应急响应、应急保障、应急后期处置及附则的所有相关信息；一致性指应用文各章节对同一命名实体或关键概念的描述应前后一致，如在生成的应急预案中，对事故类型的定义和预案编写目的的阐述等，均需保持上下文一致；相关性指文本内容应与特定场景高度相关，如应急预案中提及的应急组织机构和应急保障措施等，均需适配相应的事故类型和预案适用范围；清晰性指生成的应用文需用词准确、逻辑合理且可读性强，能为特定应用场景的行动提供明确指导；规范性指生成文本的语言风格应符合目标应用文的表述习惯，语体规范且用词专业，确保文件正式且具权威性.

2.1.2 评价指标的使用规则

对生成式应用文进行质量评测时，需综合所有指标，并在统一标准下给出单一评价结果. 设定评价指标分数集合$G = \{g_\tau \mid 0 \leqslant g_\tau \leqslant 100, \tau = 1, 2, 3, 4, 5\}$，对应的权重值集合$W = \{w_\tau \mid 0 \leqslant g_\tau \leqslant 1, \tau = 1, 2, 3, 4, 5\}$，且满足$\sum_{\tau=1}^{5} w_\tau = 1$. 其中，$g_1$—$g_5$和$w_1$—$w_5$分别为$E_1$—$E_5$对应的实际分数和权重系数. 因此，待测文本单次评测的结果为

$$S = \sum_{\tau=1}^{5} w_\tau g_\tau, \quad 0 \leqslant S \leqslant 100 \tag{8}$$

2.1.3　利用层次分析法确定评价指标权重

为降低主观因素对评测结果造成的影响，采用层次分析法(analytic hierarchy process，AHP)确定各指标的权重[15].

1）构建指标判断矩阵. 为衡量判断指标间的相对重要性，基于专家经验，使用1～9标度方法(标度含义请扫描论文末页右下角二维码查看补充材料表S1)，构建判断矩阵为

$$\boldsymbol{A}=\left(a_{pq}\right)_{5\times5}=\begin{bmatrix}1 & 3 & 2 & 3 & 5\\ 1/3 & 1 & 1/2 & 1 & 3\\ 1/2 & 2 & 1 & 2 & 4\\ 1/3 & 1 & 1/2 & 1 & 3\\ 1/5 & 1/3 & 1/4 & 1/3 & 1\end{bmatrix} \tag{9}$$

其中，矩阵元素 a_{pq} 为指标 p 与指标 q(p，q = E_1，E_2，…，E_5)的重要性比值. 得到5个评价指标 E_1、E_2、E_3、E_4和 E_5对应的归一化权重依次为0.405 2、0.143 4、0.249 7、0.143 4和0.058 3.

2）一致性检验. 计算得到 $\boldsymbol{A}$ 的最大特征根 λ_{max} = 5.056 7. 随机一致性指标(random consistency index，RI)记为 I_{RC}，查表可得 I_{RC} = 1.12（请扫描论文末页右下角二维码看补充材料表S2)，一致性指标(consistency index，CI)记为 I_C，且 $I_C=\left(\lambda_{max}-\tau\right)/(\tau-1)$ = 0.014 2(本研究指标数 τ = 5)，一致性比率(consistency ratio，CR)记为 R_C，且 $R_C=I_C/I_{RC}$ = 0.012 7 < 0.1. 由此可证，构建的判断矩阵符合一致性检验，所得权重是可信的.

2.2　应用文评测方法

在采用模型评测方法时，参考文献[14]，本研究的主裁判模型选择目前公认评测能力较强的ChatGPT-4Turbo(以下简称GPT-4)，其语义理解能力和长文本输入能力均较出色，且与人类裁判偏好的一致性可达80%[15]，具有较高的可信度和可解释性. 为进一步增强实验结果的可靠性，同时引入Deepseek-R1大模型作为补充评测模型. Deepseek-R1的训练数据中包含大量中文语料，在中文理解和文化适应性方面较GPT-4更具优势.

评测采用回答评分方法，即人工制定评分标准并使用恰当的提示词，引导裁判模型依据该固定标准对生成的待测文档打分，从而完成自动评测.

3　实验与评估

为全面评估所提融合LLM与向量知识库的应用文生成框架，采用如图3的流程，设计了一个生成应急预案的系统性实验.

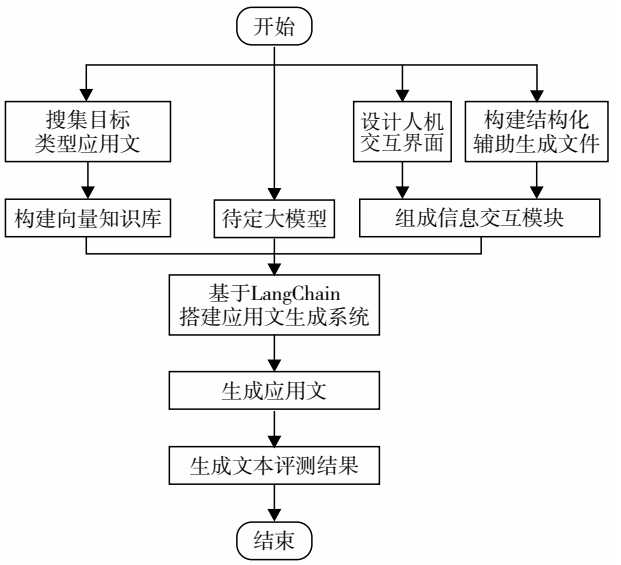

图3　应急预案生成实验流程图

Fig. 3　Flowchart of the emergency plan generation experiment.

3.1　实验环境

实验硬件配置为Intel Core i9-14900K处理器，128 Gbyte内存，使用NVIDIA GeForce RTX4090 24 Gbyte图形处理器(graphics processing unit，GPU)进行加速，基于Windows10操作系统，采用Python 3.10、LangChain 0.0.330和PyTorch2.1框架.

使用本地部署的开源双语大模型ChatGLM3-6B[17]生成文本；基于Faiss[18]构建向量知识库；采用基于千万级中文句对数据集训练的嵌入模型m3e-base[19]执行向量化操作；裁判模型使用GPT-4和Deepseek-R1大模型.

3.2　实验过程及结果分析

3.2.1　应急预案生成及评分

基于30份国家应急管理部门发布的标准应急预案构建向量知识库，并使用所提方法生成应急预案. 构建包含10个不同应急场景的样本集，覆盖自然灾害、事故灾难、医疗卫生事件和社会安全事件等关键领域，每个场景生成10份应急预案，共计100个样本.

分别使用GPT-4和Deepseek-R1作为裁判模型，对生成的100份应急预案及10份对应场景的国家或行业发布的标准应急预案进行评分. 评测过程为：向裁判模型提供待评分的应急预案文档，并详细描述评价标准及赋分规则；裁判模型阅读文档后给出各指标的分数和简要的评分依据；记录该评测结

果，并依据2.1节中的指标权重，计算各应急预案文档的得分.

图4为使用GPT-4和Deepseek-R1评测时，不同应急场景的应急预案的平均得分(具体分值请扫描论文末页右下角二维码看补充材料表S3). 由图4可见，当GPT-4担任裁判时，标准应急预案得分稳定在95～97分，而本研究方法生成的应急预案得分集中在90～94分，存在小幅波动. 当Deepseek-R1作为裁判模型时，标准应急预案得分区间为82～93分，而本研究方法生成的应急预案得分区间为77～82分. Deepseek-R1的评分整体较GPT-4低且波动更大，但两类预案得分的相对高低关系并未改变. 图4显示，国家或行业发布的标准应急预案在GPT-4和Deepseek-R1评测中的平均得分分别约为96.20和88.66，而本研究方法生成的应急预案在GPT-4和Deepseek-R1评测中的平均得分分别约为92.23和79.99. 二者的文档质量相似度在GPT-4和Deepseek-R1的视角下，分别为92.23/96.20 ≈ 95.87%和79.99/88.66 ≈ 90.22%，达到较高水平，表明本研究所提框架生成的应急预案在质量上具有较高的可靠性.

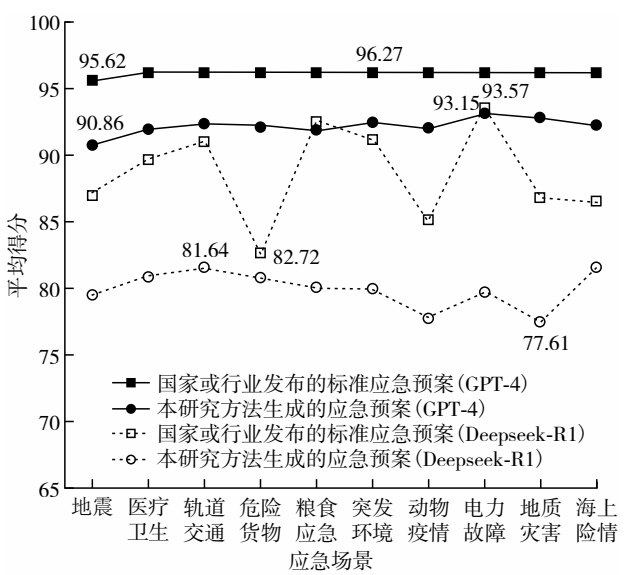

图4 各应急场景下两类应急预案的平均得分

Fig. 4 Average scores of two types of emergency plans across various scenarios evaluated by GPT-4 (solid line) and Deepseek-R1 (dashed line), generated by the proposed method of this paper (point) and issued by government agencies (square).

图5为本研究框架生成的100份应急预案在5个评价标准上的平均得分. 由图5可见，生成式应急预案在完整性和相关性方面表现良好，且Deepseek-R1与GPT-4的完整性均是最优指标. 所提方法结构化拆分处理强化了文本的结构完整性，向量知识库的融合则提升了文本内容的相关性. 在一致性、清晰性和规范性评估中，GPT-4与Deepseek-R1的评测结果存在一定差异. 分析认为，这源于Deepseek-R1对中文表达规范具有更严格的要求. 该评估意见也为生成文本在以上方面的优化提供了方向.

表1为使用所提框架生成50份和100份应急预案时，不同应急场景下的评测结果(GPT-4评分). 由表1可见，当各场景样本数量由5增至10时，各应急场景生成应急预案的平均得分并未发生显著变化，验证了此框架具有良好的稳定性. 同时，各应急场景下生成预案的平均得分差异较小，表明此框架在多种场景中具有较好的通用性，具备实际应用的可行性.

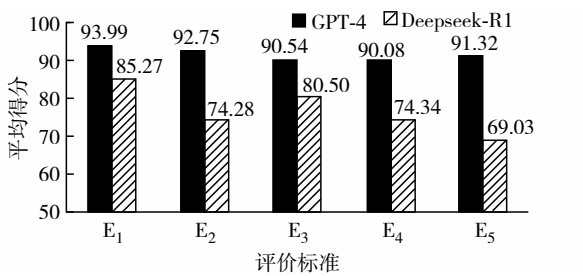

图5 各评价指标在GPT-4和Deepseek-R1下的平均得分

Fig. 5 Average scores of each evaluation criterion assessed by GPT-4 (solid black) and Deepseek-R1 (striped diagonal).

表1 GPT-4评测下不同大小样本集在各应急场景下的平均得分及方差

Table 1 Average scores and variances evaluated by GPT-4 for sample sets of different sizes in each emergency scenario

应急场景	50份应急预案		100份应急预案	
	平均得分	方差	平均得分	方差
地震	91.50	15.29	90.86	9.31
医疗卫生	92.32	2.68	91.98	2.72
轨道交通	93.13	4.51	92.38	3.73
危险货物	92.32	2.68	92.23	2.97
粮食应急	92.32	2.68	91.93	3.77
突发环境	92.32	2.68	92.51	1.97
动物疫情	92.32	2.68	92.13	2.39
电力故障	93.53	3.46	93.15	3.08
地质灾害	93.33	2.77	92.83	3.24
海上险情	92.32	2.68	92.34	1.94

为验证裁判模型GPT-4和Deepseek-R1的可靠性，对评测结果进行可重复性实验. 从实验样本集

的每个应急场景抽取1份生成式应急预案，共10份组成测试样本集．使用同一账号及相同评测提示词，在不同时间进行4次无历史记录的独立评测，得分见表2．由表2可见，使用GPT-4和Deepseek-R1大模型对同一生成文档进行多轮评测时，分值波动保持在7分以内(方差较小)，表明二者对评测时间点的变化有良好的鲁棒性，评测结果较可靠．

表2　评测结果的可重复性实验结果

Table 2　Experimental results on the reproducibility of evaluations

文档名称	GPT-4得分						Deepseek-R1得分					
	2024-08-12	2024-11-17	2024-11-18	2024-11-19	2024-11-21	方差	2025-03-10	2025-03-11	2025-03-12	2025-03-13	2025-03-14	方差
地震(全国)	95.29	92.18	92.39	93.81	91.87	1.63	81.98	78.71	79.50	83.42	82.15	3.10
医疗卫生(浙江省)	92.32	91.34	88.51	94.81	90.18	4.46	81.89	79.91	83.22	82.98	84.49	2.36
城市轨道(全国)	94.07	92.32	91.71	89.81	89.44	2.88	79.00	80.17	79.29	78.71	81.95	1.37
危险货物(浙江省)	92.32	90.41	92.26	93.27	90.78	1.12	84.00	79.93	82.37	79.29	79.91	3.22
粮食应急(山东省)	94.22	91.47	91.30	93.39	93.21	1.30	69.95	71.38	73.71	72.16	68.58	3.13
突发环境(山东省)	94.32	92.94	90.24	89.81	88.79	4.28	79.29	80.01	80.73	79.58	79.00	0.37
动物疫情(广州市)	94.07	93.60	91.78	94.81	90.96	2.09	80.19	79.29	83.22	82.98	84.00	3.41
电力故障(广西省)	94.22	92.17	90.70	91.84	88.70	3.29	80.01	79.00	79.29	79.00	77.75	0.53
地质灾害(广西省)	93.82	89.60	90.75	90.50	90.28	2.15	79.00	77.04	78.71	77.00	80.01	1.37
海上险情(广州市)	90.23	93.53	88.19	90.53	91.94	3.19	84.32	84.81	87.48	86.95	85.17	1.54

3.2.2　消融实验

为验证所提应用文生成框架中关键模块的有效性，基于3.2.1节实验，对人机交互获取章节框架与知识库检索获取参考文段两模块进行消融实验．

消去章节框架模块，即不再采用分步式生成策略，仅使用提示信息检索参考文段，将参考文段、提示信息和生成提示词构成的提示词组作为输入，引导LLM一次性生成完整应急预案．受模型输出长度限制，生成的应急预案长度显著缩减．

消去参考文段模块，即不进行知识库检索，仅将提示信息和生成提示词构成的提示词组作为输入，引导LLM生成无参考文段辅助的应急预案片段，再根据预设的章节框架进行整合．此时，生成文本的表述能力将完全依赖于生成模型自身的能力．

从3.2.1节实验样本集的每个应急场景抽取5份，得到共计50份文档作为基准集．同时，维持每个应急场景生成5份样本的方式，生成无章节框架模块及无参考文段模块的应急预案各50份．使用GPT-4作为裁判模型，对基准集和消融集生成的应急预案进行评分，结果见表3．由表3可见，消去任一关键模块都会降低应急预案的生成质量．

表3　知识库检索模块和预设章节框架模块消融实验结果

Table 3　Results of ablation experiment on the knowledge base retrieval module and the preset chapter framework module

评测项(文档数)		文档平均得分		
		基准实验	无参考文段	无章节框架
评价指标	完整性(50)	95.20	88.02	81.48
	一致性(50)	92.84	88.16	82.70
	相关性(50)	89.80	88.64	85.24
	清晰性(50)	89.84	87.70	82.58
	规范性(50)	91.74	88.18	81.06
应急场景	地震(5)	91.50	83.14	80.45
	医疗卫生(5)	92.32	89.04	81.77
	轨道交通(5)	93.13	87.65	82.39
	危险货物(5)	92.32	89.04	83.34
	粮食应急(5)	92.32	89.06	83.14
	突发环境(5)	92.32	89.37	83.59
	动物疫情(5)	92.32	88.89	82.32
	电力故障(5)	93.53	88.23	83.08
	地质灾害(5)	93.33	88.26	83.58
	海上险情(5)	92.32	88.91	83.62

3.2.3　通用性验证实验

为验证所提应用文生成方法的通用性，对框架中的生成模型和嵌入模型进行替换实验，以评估所提方法在不同模型组合上的有效性．

在验证生成模型的替换效果时，嵌入模型固定为m3e-base，选用与ChatGLM3-6B参数量相近的Baichuan2-7B和Qwen-7B作为生成模型，遵循与

3.2.1节相同的实验流程，使用GPT-4作为裁判模型，结果如表4．由表4可见，ChatGLM3-6B、Baichuan2-7B和Qwen-7B模型的平均得分分别为92.54、92.16和88.78，分数波动处于合理范围．

表4 GPT-4评测下不同生成模型在各应急场景下平均得分

Table 4 Average scores evaluated by GPT-4 for different generative models in each emergency scenario

应急场景	ChatGLM3-6B	Baichuan2-7B	Qwen-7B
地震	91.50	89.38	87.09
医疗卫生	92.32	94.27	89.48
轨道交通	93.13	89.38	87.47
危险货物	92.32	94.27	88.88
粮食应急	92.32	94.27	88.80
突发环境	92.32	94.27	88.72
动物疫情	92.32	94.27	88.60
电力故障	93.53	89.46	88.60
地质灾害	93.33	89.38	91.59
海上险情	92.32	92.65	88.58

在验证嵌入模型的替换效果时，选用与m3e-base（102×10^6个参数）参数量相近的bge-base-zh（102×10^6个参数）和Dmeta-zh（102×10^6个参数）作为嵌入模型．截取知识库检索得到的参考文段，即结构化辅助生成文件中的｛content｝内容，与人工定位的最佳参考文段比对，以准确召回数衡量嵌入模型的检索效果．依据设定，生成的应急预案共有25个二级标题，即检索总次数为25．实验结果如图6，不同嵌入模型在文段召回能力尽管存在差异，但在不同应急场景下的变化趋势基本一致；m3e-base模型在召回性能上表现最优．

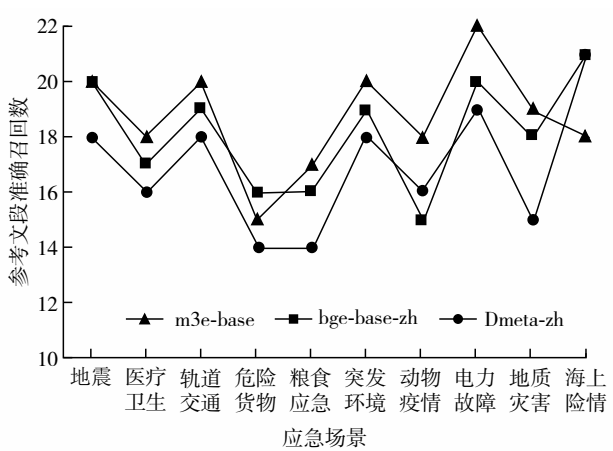

图6 不同嵌入模型在各应急场景下的参考文段准确召回数

Fig. 6 The number of accurately recalled reference paragraphs by different embedding models across scenarios (m3e-base (triangle), bge-base-zh (square) and Dmeta-zh (circle)).

所提生成框架在不同的生成模型与嵌入模型上均取得了良好效果，证明该框架具有较强的泛化能力．据此推测，在算力资源充足的条件下，采用性能更卓越的生成模型（如Deepseek-R1）和嵌入模型，有望进一步提升应用文的生成效果．

综上所述，本研究所提应用文生成框架能够高效生成结构完整、内容丰富且质量较好的应急预案．但是，上述由裁判模型所给出的评测结果，不能完全等同于人类偏好标准，由所提框架生成的应用文，仍需人工审核后方可投入实际应用．

结　语

提出一种融合LLM与向量知识库的应用文自动生成框架，通过分步式生成策略实现各章节的串联创作，采用文本语义对齐算法完成多段落语义关联与结构整合．实验结果表明，基于本研究所提出框架的应急预案生成任务在多个评价指标上表现良好，证明了该生成框架在实际应用中具有一定的潜力．

尽管本研究取得了一定成果，但仍存在进一步改进和扩展的空间．未来的工作将集中在扩展向量知识库的覆盖范围，优化结构化辅助生成文件的构造方式，以及优化LLM的文本生成提示词等．我们期待该方法能够在更广泛的应用场景中得到验证和应用．

基金项目： ZTE政务行业大模型技术研究产学研基金资助项目（1A20231030016）；国家重点研发计划资助项目（2020YFB1806405）；深圳市科技重大专项资助项目（KJZD20230923114906013）

作者简介： 秦斌(qinbin@szu.edu.cn)，深圳大学正高级工程师．研究方向：人工智能；
陆平(lu.ping@zte.com.cn)，中兴通讯股份有限公司正高级工程师．研究方向：产业数字化．
秦斌和陆平为共同第一作者．

引　　文： 秦斌，陆平，徐琰，等．融合大语言模型与向量知识库的应用文生成框架[J]．深圳大学学报理工版，2025，42(5)：597-605.

参考文献 / References：

[1] 王进强，刘金硕．基于注意力机制的结构化文本自动生成[J]．武汉大学学报工学版，2022，55(2)：198-203.
WANG Jinqiang, LIU Jinshuo. Automatic generation of structured text based on attention mechanism [J]. Engi-

neering Journal of Wuhan University, 2022, 55(2): 198-203. (in Chinese)

[2] 闫盈盈，何娜，张志玲，等. 生成式人工智能构建患者药品说明书的方法研究[J]. 临床药物治疗杂志，2024，22(5)：1-6.
YAN Yingying, HE Na, ZHANG Zhiling, et al. Methods for constructing patient medication instructionsusing generative artificial intelligence [J]. Clinical Medication Journal, 2024, 22(5): 1-6. (in Chinese)

[3] 全民互联科技(天津)有限公司. 智合同[EB/OL]. [2024-07-16]. https://www. shenht. com.
National Internet Technology (Tianjin) Co., Ltd. Smart contract [EB/OL]. [2024-07-16]. https://www. shenht. com. (in Chinese)

[4] CHASE H. LangChain [EB/OL]. (2022-10-14)[2024-07-16]. https://github. com/langchain-ai/langchain.

[5] LEWIS P, PEREZ E, PIKTUS A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks [C]// Advances in Neural Information Processing Systems. Red Hook, USA: Curran Associates Inc., 2020: 9459-9474.

[6] SITENDER, BAWA S, KUMAR M, et al. A comprehensive survey on machine translation for English, Hindi and Sanskrit languages [J]. Journal of Ambient Intelligence and Humanized Computing, 2023, 14(4): 3441-3474.

[7] FAISAL M, MAWARIDI B H, AFRAH A S, et al. Enhancing indonesian text summarization with latent dirichlet allocation and maximum marginal relevance [J]. International Journal of Advanced Computer Science and Applications, 2024, 15(8): 519-528.

[8] LI Daiyi, TU Yaofeng, ZHOU Xiangsheng, et al. End-to-end Chinese entity recognition based on BERT-BiLSTM-ATT-CRF [J]. ZTE Communications, 2022, 20(Suppl. 1): 27-35.

[9] LI Haonan, ZHANG Yixuan, KOTO F, et al. CMMLU: measuring massive multitask language understanding in Chinese [C]// Findings of the Association for Computational Linguistics: ACL 2024. Stroudsburg, USA: ACL, 2024: 11260-11285.

[10] HUANG Yuzhen, BAI Yuzhuo, ZHU Zhihao, et al. C-EVAL: a multi-level multi-discipline Chinese evaluation suite for foundation models [C]// Advances in Neural Information Processing System. Red Hook, USA: Curran Associates Inc., 2023: 62991-63010.

[11] MUKHERJEE S, MITRA A, JAWAHAR G, et al. Orca: progressive learning from complex explanation traces of GPT-4 [EB/OL]. (2023-06-03) [2024-07-16]. https://arxiv. org/abs/2306. 02707.

[12] WANG Yidong, YU Zhuohao, ZENG Zhengran, et al. PandaLM: an automatic evaluation benchmark for LLM instruction tuning optimization [EB/OL]. (2024-05-24) [2024-07-16]. https://arxiv. org/abs/2306. 05087.

[13] CHIANG W L, ZHENG Lianmin, SHENG Ying, et al. Chatbot arena: an open platform for evaluating LLMs by human preference [EB/OL]. (2024-03-07) [2024-07-16]. https://arxiv. org/abs/2403. 04132.

[14] NI Weijian, SHEN Quanle, LIU Tong, et al. Generating textual emergency plans for unconventional emergencies: a natural language processing approach [J]. Safety Science, 2023, 160: 106047.

[15] 邓雪，李家铭，曾浩健，等. 层次分析法权重计算方法分析及其应用研究[J]. 数学的实践与认识，2012，42(7)：93-100.
DENG Xue, LI Jiaming, ZENG Haojian, et al. Research on computation methods of AHP wight vector and its applications [J]. Mathematics in Practice and Theory, 2012, 42(7): 93-100. (in Chinese)

[16] ZHENG Lianmin, CHIANG W L, SHENG Ying, et al. Judging LLM-as-a-Judge with MT-bench and chatbot arena [C]// Advances in Neural Information Processing Systems. Red Hook, USA: Curran Associates Inc., 2023: 46595-46623.

[17] Team GLM. ChatGLM: a family of large language models from GLM-130B to GLM-4 all tools [EB/OL]. (2024-07-30) [2024-10-10]. https://arxiv. org/abs/2406. 12793.

[18] DOUZE M, GUZHVA A, DENG Chengqi, et al. The Faiss library [EB/OL]. (2024-09-06) [2025-01-10]. https://arxiv. org/abs/2401. 08281v2.

[19] WANG Yuxin, SUN Qingxuan, HE Sicheng. M3E: moka massive mixed embedding model [EB/OL]. (2023-06-06) [2025-01-10]. https://github. com/liqiangus/m3e-base.

【中文责编：英子；英文责编：木柯】

补充材料