人工智能

基于大语言模型RAG搭建智能电话小结流程

徐奕婷，崔珊珊

（中国电信股份有限公司上海分公司客服服务中心，上海 200120）

摘　要：面向电话客服场景的智能标签生成需求，基于检索增强生成（RAG）技术构建了智能电话小结流程。通过BGE-M3-Embedding模型结合Milvus库的向量检索，实现对用户诉求的精准匹配，并借助提示词工程优化模型生成最符合的标签。在检索增强生成方法的对比实验中，RAG流程的测试集准确率相比传统训练方法提升了20%，召回率达到了0.95。结果表明，基于RAG的智能电话小结流程具备高度的可靠性和灵活性，能够为复杂通话场景下的客服系统提供有力支持。

关键词：大语言模型；RAG检索增强生成；电话小结

中图分类号：TP393

文献标志码：A

doi: 10.11959/j.issn.1000−0801.2025058

0　引言

随着大语言模型（large language model，LLM）的不断发展，其在客服场景中的应用愈发广泛，尤其在智能质检、工单撰写等标准化任务中，大幅优化了工作流程并提升了效率。然而，大语言模型存在数据时效性差、缺乏特定领域知识等问题，以电话小结场景为例，客服代表需要熟悉标签集及其描述，并根据通话内容选择匹配的标签，而这类知识储备并未包含在大模型的预训练数据集中，使得模型在处理垂直领域的场景时会出现大量幻觉问题，难以满足应用需求。

针对这一问题，本文提出了一种新的解决方案：结合大语言模型、文本嵌入技术与向量数据库，构建基于LLM和检索增强生成（retrieval augmented generation，RAG）的智能电话小结流程。该流程首先利用文本嵌入技术，将历史通话数据存储于本地向量数据库中，并通过检索逻辑筛选出与当前通话内容相关的知识。随后，召回的相关知识被交由大模型处理，以弥补其在话务领域知识上的不足。大模型据此理解用户的具体诉求，并将其与话务标签进行精准匹配。该方案不仅显著提升了企业在电话通话处理中的效率，还为未来客服工作的智能化发展提供了新的思路。

1　相关技术

1.1　大语言模型

LLM是基于深度学习的自然语言处理模型，通常由数十亿甚至上千亿的参数组成。这类模型经过大规模文本数据的训练，具备了强大的语言理解和生成能力。然而，LLM在特定领域的适用性存在明显局限。由于其训练数据主要来源于通用文本库，模型在面对专业领域或实时数据时，往往缺乏足够的背景知识，难以生成准确或相关的内容。此外，LLM还可能带来隐私问题，如无意间泄露训练数据中的敏感信息。

收稿日期：2024−11−03；**修回日期：**2025−02−12

1.2 RAG架构

传统的大语言模型主要依赖其预训练数据进行推理和生成，但在面对复杂或领域专属的任务时，容易因缺乏足够的背景知识而产生幻觉。由此，将大语言模型与外部知识源结合的技术方案诞生了。

RAG是一种将LLM与检索相结合的架构。如图1所示，当响应用户输入时，RAG首先从外部知识库或数据库中获取与输入相关的上下文，然后将这些上下文交给大模型，生成更加准确且具有现实依据的回答。通过先检索后生成的方式，RAG架构弥补了大语言模型在特定领域的知识空缺，显著提高了生成结果的可靠性。

1.3 Milvus向量库

Milvus是当前少数支持GPU加速的向量数据库之一，尤其在高吞吐量、低时延和高召回率的应用场景中表现优异。借助GPU的并行计算能力，Milvus能够显著提升向量搜索的性能与效率，满足生产环境中高并发请求的需求。相比于传统的CPU计算，GPU加速不仅加快了大规模向量数据的处理速度，还能缩短检索响应时间，并提高检索精度，在需要快速、高效查询的任务中具有明显优势。

此外，Milvus提供了多样化且灵活的搜索方式，包括Top-K搜索、范围近似最近邻（ANN）搜索、稀疏与稠密向量搜索、多向量搜索及分组搜索等，满足不同场景下的复杂查询需求。

2 电话小结的应用实践

2.1 向量数据库的构建

本阶段属于离线数据准备阶段，旨在将历史通话数据向量化后构建索引并存入向量数据库。此过程可将文本信息转化为高维语义向量，方便后续的检索和匹配操作。整个处理流程包括以下步骤。

（1）获取质检数据。首先，将通话文本交由大语言模型处理，通过提示词工程嵌入质检要求。模型依据提示生成通话的概述，提取用户诉求、诉求分类、客服代表的解决方案等多维度信息。

（2）数据拼接与清洗。在获得通话文本的多维度信息后，提取其中的通话概述和诉求分类字段，将其与对应的工单数据和历史标签数据进行拼接整合。由于工单和标签数据可能存在标注错误，经过人工校验后的数据才会用于后续步骤。

（3）构建向量库。借助BGE模型，将原始文本数据向量化，生成语义向量。然后根据工单数据和标签数据的类型差异，将生成的向量分别存储到Milvus不同的collection中，构建向量索引库。通过这种方式，可以为后续的检索操作提

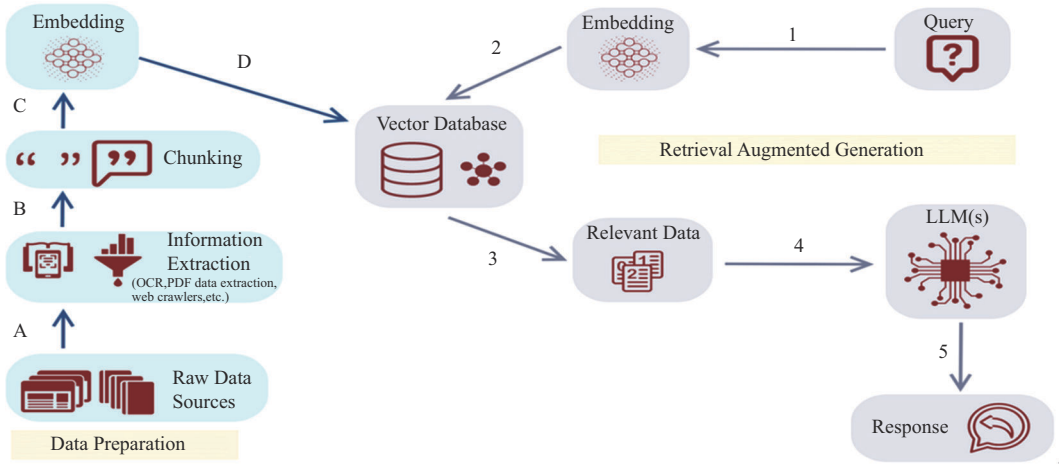

图1 RAG架构

供高效的数据结构支持。

通过上述流程，对历史通话数据进行了规范化处理，并通过文本嵌入技术构建了便于检索的语义索引库，为智能电话小结的实现提供了数据基础。

2.2　系统流程

在通话数据输入系统后，首先需要从质检数据中提取通话概述和诉求分类字段，并以此为基础进行后续的处理。根据工单数据的存在与否，流程分为两个分支。

（1）有工单处理流程（如图2所示）。当存在工单数据时，系统首先判断该通话的工单类型，并选择对应的向量数据库 collection 进行召回。例如，若该通话涉及投诉单，则系统只会从 fault_collection 中召回与投诉类相关的语义向量。然而，在一些情况下，用户的实际诉求可能涉及投诉工单进度查询，通话内容与投诉类语义高度重合，但并不属于投诉。因此，系统在召回后会加入一些工单咨询类标签，增强召回结果的容错性，确保标签匹配的精准度。

（2）无工单处理流程（如图3所示）。如果没

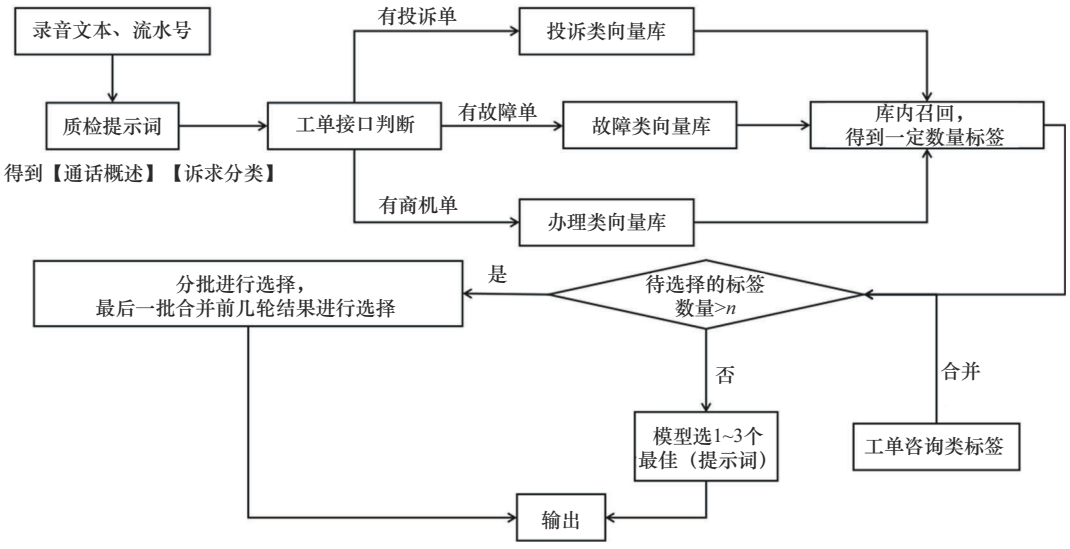

图2　有工单分支

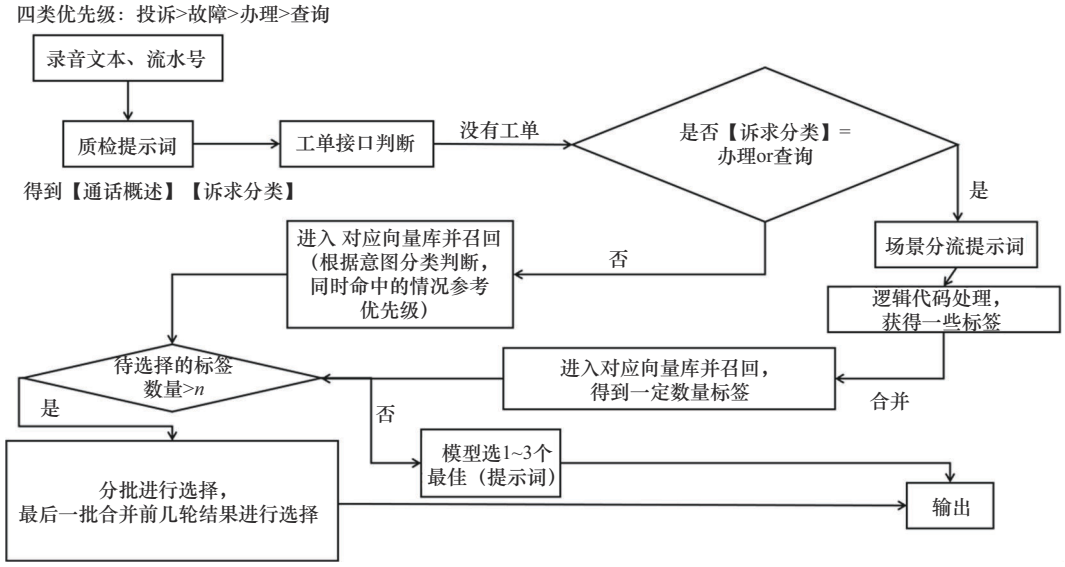

图3　无工单分支

有关联的工单数据，系统将依据大模型生成的质检数据中诉求分类字段进行进一步判断。对于诉求类型为办理或查询的情况，由于二者在流程上存在较高的重合度，系统会首先进入场景分流，通过预设的提示词引导模型进一步分析业务场景。根据分析结果，系统判断该通话需要匹配哪些标签，然后进入相应的向量数据库 collection 进行语义向量召回，最后将两个标签集进行合并。

结果整合与输出。在标签召回完成后，系统将通话文本、召回的标签及其对应的标签解释一并交付给大模型进行最终处理。由于大语言模型的理解力会随着输入文本长度的增加而逐渐下降，经过多次测试，本文确定了临界值N。当标签集的数量超过N时，系统会对标签进行分批处理，每批标签分别交给大模型生成小结，最后合并生成的结果。

智能电话小结的生成不仅考虑了工单数据、用户诉求等多方面因素，还通过容错机制和分批处理优化了模型在复杂场景中的表现，提升了系统整体的召回精准度和处理效率。

2.3 RAG优化策略

为了提高智能电话小结流程的生成效率，本系统在RAG架构下实施了两项优化策略：意图分流和重排序。这两项策略从减少召回数据量入手，缓解了大模型侧的处理压力问题，同时保证了结果的可靠性。

（1）意图分流

意图分流是本系统在召回阶段的第一层优化措施，依据不同的工单类型和用户诉求类别，将输入的通话数据分流至不同的向量数据集中进行检索。这一策略有效减少了系统需要处理的向量数据量，避免了冗余数据的召回，大幅降低了大模型在生成标签阶段的处理压力。

（2）重排序

在某些情况下，召回的结果集仍可能过大，尤其是当通话内容较为复杂或用户诉求模糊时。为进一步优化召回的结果，本系统引入了重排序机制。当召回量超出预定范围时，系统将根据通话概述与召回内容之间的关联性进行排序，优先选择与当前通话语义最相关的结果。

通过计算召回结果与输入文本的余弦相似度，系统选取最相关的 top-k 个结果用于后续生成处理。虽然这种机制在一定程度上会损失精确度，但能够显著提升小结生成的效率，确保系统在高并发场景下仍然能够快速响应。

3 实验与结果

3.1 数据集

实验数据来源于客服系统中的历史通话数据，包括已打上标签的通话记录。为确保模型的通用性，数据涵盖了不同种类的用户需求。

3.2 实验设计

本实验将采用真实通话数据，使用RAG技术进行标签化实验，评价标准为准确率和召回率。实验将分别对单独的检索步骤以及结合RAG的完整流程进行评估。

3.3 实验结果

如表1所示，系统整体召回率达到了0.952，能够较为全面地检索出与通话相关的候选标签集。在不同需求类别下，准确率表现有所差异，其中查询类需求的准确率最高，达到0.855。

表1 电话小结流程测评数据

流程	召回率	准确率
查询类需求	—	0.855
办理类需求	—	0.757
故障类需求	—	0.676
投诉类需求	—	0.787
综合	0.952	0.810

4 基于RAG的电话小结应用的探讨

在电话客服场景中，基于RAG技术的小结应用具有显著的优势，但RAG架构存在其局限

性。本节将从RAG技术的应用价值及其不足之处进行讨论，并探讨未来可能的发展方向。

4.1　技术优势

基于RAG的系统相较于传统模型训练方法具有显著的优势，尤其在业务快速上线的需求下表现尤其突出。由于RAG技术不需要大量的数据训练，只需要依赖现有的知识库和提示词设计，便能够迅速适配特定业务场景。因此，对于一些需要快速部署和频繁更新的业务场景而言，RAG提供了更为灵活的解决方案。

此外，RAG的高度定制能力使其能够通过简单的配置快速适应不同领域的需求。通过调整知识库的内容以及设计合理的提示词，系统可以迅速获得特定领域的专业知识。

同时，RAG具有明确的调优机制。当系统在某个场景下出现错误时，可以根据具体问题进行针对性优化。若是因知识点遗漏导致检索不完整，只需将该错误案例加入知识库即可确保下次检索时不会再次遗漏；若错误来源于模型生成回答不准确，则通过调整提示词改善生成效果。这样逐步优化的方式，使系统在经过充分测试后，能够稳定提升整体的精度，确保在复杂业务场景中的表现更加可靠。

4.2　局限性与未来展望

尽管RAG技术在电话小结场景中展现出诸多优势，但其也存在一定的局限性。首先，RAG无法完全保证提供给模型的信息是绝对准确的。在电话小结场景中，数据源相对单一，数据格式也较为统一，但在更复杂的场景中，数据可能来源于不同渠道，且格式多样。例如，企业知识库可能同时包含PDF、CSV、TXT等不同格式的文档，在这种情况下，要准确地理解用户的问题并从多个数据源中找到关键信息，这不仅要求系统能够处理异构数据，还要求其具备更强的上下文理解和推理能力。

此外，由于RAG依赖于知识库的内容质量，如果知识库中存在错误或信息不全，将直接影响模型的生成结果。这种情况下，即便模型能够准确检索并生成答案，所提供的回答仍可能存在误导性。

4.3　未来展望

针对上述RAG系统的局限性，未来笔者将尝试引入知识图谱技术，进一步提升系统的知识整合能力。通过图谱将不同数据源的知识结构化、关联化，使系统能够更好地理解和处理复杂的多源数据。

5　结束语

本文展示了RAG技术在电话小结标签化场景中的创新应用，并验证了其在标签生成中的有效性。通过结合检索和生成，RAG不仅解决了大语言模型在特定领域知识不足的问题，还显著提高了电话通话处理的效率和准确性。

参考文献：

[1] RADFORD A, NARASIMHAN K, SALIMANS T, et al. Improving language understanding by generative pre-training[J]. OpenAI preprint, 2018: 1-12.

[2] DU Z X, QIAN Y J, LIU X, et al. GLM: general language model pretraining with autoregressive blank infilling[J]. arXiv preprint, arXiv: 2103.10360, 2021.

[3] GAO Y, XIONG Y, GAO X, et al. Retrievalaugmented generation for large language models: a survey[J]. arXiv preprint, arXiv: 2312.10997, 2023.

[4] ZHAO W X, ZHOU K, LI J, et al. A survey of large language models[J]. arXiv preprint, arXiv: 2303.18223, 2023.

[作者简介]

徐奕婷（1999– ），女，现就职于中国电信股份有限公司上海分公司客服服务中心，主要研究方向为大模型应用开发等。

崔珊珊（1995– ），女，现就职于中国电信股份有限公司上海分公司客服服务中心，主要研究方向为大模型应用开发等。