电 子 科 技 大 学

UNIVERSITY OF ELECTRONIC SCIENCE AND TECHNOLOGY OF CHINA

专业学位硕士学位论文

MASTER THESIS FOR PROFESSIONAL DEGREE

论文题目 **基于大型语言模型的汽车客服自动问答系统研究与设计**

专业学位类别	机械
学　　号	202152040807
作者姓名	覃毅哲
指导教师	廖伟智 研究员
学　　院	机械与电气工程学院

分类号 TP3 密级 公开

UDC 注1 681.5

学 位 论 文

基于大型语言模型的汽车客服自动问答系统研究与设计

（题名和副题名）

覃毅哲

（作者姓名）

指导教师 廖伟智 研究员

电子科技大学 成 都

申请学位级别 硕士 专业学位类别 机械

提交论文日期 2024 年 4 月 11 日 论文答辩日期 2024 年 5 月 27 日

学位授予单位和日期 电子科技大学 2024 年 6 月

答辩委员会主席

评阅人

注 1：注明《国际十进分类法 UDC》的类号。

Research and Design of an Automated Question-Answering System for Automotive Customer Service Based on Large Language Models

A Master Thesis Submitted to

University of Electronic Science and Technology of China

Discipline **Mechanics**

Student ID **202152040807**

Author **Yizhe Qin**

Supervisor **Prof. Weizhi Liao**

School **School of Mechanical and Electrical Engineering**

摘 要

在汽车后市场业务中，客服是其中最基础、重要的环节之一，是建立企业与客户沟通桥梁的关键，既有助于提升客户满意度和忠诚度，也有助于客服部门获取客户需求、零部件故障等市场和售后数据，使企业能不断改进产品和服务，提升企业竞争力。现有的自动问答系统在实际应用中面临知识更新不及时、利用率不高、难以处理复杂问题以及知识库检索能力不足等问题，这些问题导致了问答系统的准确率下降，用户对线上服务的信任度也随之降低。本文针对 XX 汽车制造企业当前面临的客户服务挑战，围绕该企业的“智能客服问答系统建设”课题，根据企业客服业务现状，提出了基于大型语言模型的智能客服问答系统设计方案。该方案旨在通过整合和优化企业现有的知识库资源，以提升客户在车辆使用、维护和置换等环节的咨询和投诉处理体验。主要研究内容如下：

（1）针对现有系统难以有效利用多种结构化知识的问题，本文提出了一种多源异构客服知识融合方法，该方法通过预设规则，将结构化的表格、知识图谱、文本对统一为自由文本的形式，然后通过切分、标题增强及向量化等步骤进行处理，使这些信息能够在基于自由文本的自动问答系统中被高效利用。本文通过公开数据集评估知识融合效果：通过 NLPCC 2018 KBQA 数据集评估基于结构化知识的自动问答任务，本文方法答案匹配达到 74.0%；通过“公益 AI 之星”相似句对判定大赛数据集评估文本相似度任务，本文方法准确率达到 91.4%。实验验证结果证明了本文提出的知识融合方法的有效性。

（2）为提升系统对多轮交互问题的处理能力，提出一种基于大型语言模型的提示框架，并基于此框架设计了一种多轮对话问题的解决方案。该方案将提示框架应用于多轮对话的意图识别、槽位填充和自然语言生成等多个子模块。并通过公开数据集 ATIS 上验证本文提出的方法，实验结果表明引入基于大型语言模型可提升各模块性能， 其中意图识别任务上准确率达到 97.4%，槽位填充任务上 F1 达到 94.8，验证了该方法的有效性。

（3）针对现有系统检索能力不足的问题，本文提出了一种利用大型语言模型的检索增强生成技术来提升企业自动问答系统的知识检索能力的方法。该方法通过预训练的密集嵌入模型将问题文本与知识库转换为密集向量，这些向量包含了更丰富的语义信息。然后，利用大型语言模型作为阅读器来生成最终答案，以提高答案的准确性和可读性。本文还通过实验研究了系统模型选型和关键参数设计，以实现问答系统的准确性和响应速度之间的平衡。

（4）基于以上方法，本文设计并开发了面向汽车客服的自动问答 Web 系统，基于该系统完成了系统的应用验证。

关键词：汽车售后，客服，知识融合，自动问答系统，大型语言模型

ABSTRACT

In the automotive aftermarket industry, customer service plays a pivotal role in establishing effective communication between enterprises and customers. This not only enhances customer satisfaction and loyalty but also enables the customer service department to gather valuable market and aftermarket data, including customer demand and parts failure. Such information can be utilized to continuously enhance the competitiveness of products and services offered by the enterprise. However, automated Q&A systems often encounter challenges such as outdated knowledge, low utilization rates, difficulty in handling complex problems, and inadequate retrieval capabilities of knowledge bases in practical applications. These issues can lead to decreased accuracy of Q&A systems and a loss of user trust in online services. To address these challenges faced by XX automobile manufacturing enterprise's customer service department, this thesis proposes a design scheme for an intelligent customer service Q&A system based on a large-scale language model that is tailored to the current situation of the enterprise's customer service business. The solution aims to integrate and optimize existing knowledge base resources within the enterprise to enhance customers' experience when dealing with inquiries and complaints related to vehicle use, maintenance, and replacement. The main research contents are as follows:

(1)This thesis proposes a method for fusing multiple sources of heterogeneous customer service knowledge to address the challenge of existing systems being difficult to effectively utilize knowledge. The method unifies structured tables, knowledge graphs, and text pairs into free text using preset rules. Subsequently, the information is processed through slicing, caption augmentation, and vectorization to enable efficient utilization in a free-text-based automated Q&A system. The method achieved a 74.0% answer matching rate on the NLPCC 2018 KBQA dataset, which evaluates structured knowledge-based automatic Q&A tasks. Additionally, it attained an accuracy of 91.4% on the "Public Welfare AI Star" Similar Sentence Pair Judgment Competition dataset, which assesses text similarity tasks. The results of the validation demonstrate the effectiveness of the knowledge fusion method proposed in this thesis.

(2) In order to enhance the system's capability to handle complex problems, this thesis proposes a prompting framework based on a large-scale language model.

Subsequently, a method for addressing the multi-round dialog problem was developed within this framework. The method applies the prompting framework to multiple sub-modules, such as intent recognition, slot filling, and natural language generation for multi-round dialog. The method proposed in this thesis was validated on the publicly available dataset ATIS. Experimental results demonstrate that the incorporation of techniques based on large language models enhances the performance of various modules. Specifically the accuracy rate for the intention recognition task reached 97.4%, and the F1 score for the slot filling task reached 94.8%, thereby verifying the effectiveness of the proposed method.

(3) Aiming at the problem of inadequate retrieval ability of the existing system, this thesis proposes a technique to enhance the retrieval capability of automated enterprise Q&A systems by utilizing a large-scale language model. This method transforms question text and knowledge base into dense vectors using a pre-trained dense embedding model, which encapsulates richer semantic information. Subsequently, a large-scale language model serves as a reader to generate final answers, thereby improving the accuracy and readability of the responses. Additionally, this thesis explores model selection and key parameter design to strike a balance between the accuracy and response speed of the Q&A system.

(4) Based on the above methods, this thesis designs and develops an automatic Q&A Web system for automotive customer service, and completes the application verification of the system based on this system.

Keywords: Automotive Aftermarket, Customer Service, Knowledge Fusion, Automated Q&A Systems, Large Language Model

目　录

第一章 绪论

1.1 选题来源

近年来，随着我国经济的快速发展和人民生活水平的不断提高，汽车已逐渐成为家庭出行的主要交通工具，市场保有量持续攀升。据统计，2023 年我国汽车保有量已达到 4.35 亿[1]，巨大汽车保有量促进了汽车后市场产业的发展，我国汽车后市场消费规模已超万亿元[2]。汽车后市场巨大的商业价值已经成为汽车产业重要的发展方向。

在汽车后市场业务中，客服是其中最基础、重要的环节之一，是建立企业与客户沟通桥梁的关键，既有助于提升客户满意度和忠诚度，也有助于客服部门获取客户需求、零部件故障等市场和售后数据，使企业能不断改进产品和服务，提升企业竞争力。

目前，汽车行业的客户服务主要通过两种方式实现：人工客服和自动问答系统。人工客服通常通过 400 电话热线提供服务，而自动问答系统则通过网络应用、微信小程序和车主专用 APP 等平台，利用预设的问答逻辑自动回复客户问题。尽管人工客服模式在专业性、知识储备和新产品知识更新方面有其优势，但也带来了高昂的人力成本。因此，自动问答系统以其成本效益和效率，成为了行业发展的重点。然而，现有的自动问答系统在实际应用中还面临知识更新不及时、利用率不高、难以处理复杂问题以及知识检索能力不足等问题，这些问题导致了问答系统的准确率下降，用户对线上服务的信任度也随之降低。以 XX 汽车制造企业为例，尽管已经开发出自动问答系统，但绝大多数用户仍然倾向于使用传统的 400 电话服务，这不仅未能充分利用线上资源，也未能有效降低企业的人力成本。

据此，本文针对 XX 汽车制造企业当前面临的客户服务挑战，围绕该企业的“智能客服问答系统建设”课题，根据企业客服业务现状，提出了基于大型语言模型的汽车客服自动问答系统设计方案。该方案旨在通过整合和优化企业现有的知识库资源，以提升客户在车辆使用、维护和置换等环节的咨询和投诉处理体验。通过采用先进的语言处理技术，该系统将能够提供更加精确、及时和人性化的服务，从而提高客户满意度，降低企业运营成本，并提升企业的市场竞争力。

为解决现有线上客服系统在知识更新和利用率方面的不足，本文提出了一种多源异构知识融合的方法。该方法能够自动整合来自不同来源和结构的知识，为构建高效的问答系统打下坚实基础。此外，针对系统在处理复杂问题时的局限性，本

文提出了一种基于大型语言模型提示框架的多轮对话框架。该框架利用语言模型的泛化能力，显著提高了多轮对话中的意图识别和槽位填充等关键任务的性能。

为了进一步提升系统的知识检索能力，本文设计了一个基于大型语言模型的检索增强生成技术的知识问答模块。通过整合上述方法，本文开发了面向企业实际应用需求的汽车客服自动问答系统。经过一系列实验验证，本文提出的方法能够有效利用多种形式的知识资源，提供更为精确的检索结果，并能够高效、准确地响应用户的问题。这不仅提升了用户体验，也为企业的客服效率和成本控制带来了实质性的改进。

1.2 国内外研究发展现状

自动问答系统基本思想是通过对使用者发起的提问进行后端逻辑处理，然后返回一个准确答案[3]。根据其所利用的知识结构差异，可将自动问答系统分为基于结构化知识的问答系统、基于自由文本的问答系统，以及基于文本对即问题-答案对的问答系统[4,5]三类。因此，本节将首先根据知识结构的差异，分别论述基于结构化知识的问答系统、基于自由文本的问答系统，以及基于文本对的问答系统三类系统的研究发展现状。然后，研究解决复杂问题所需的多轮对话方法的发展现状。最后，结合大语言模型的研究进展，分析大型模型在问答系统中的研究现状和挑战。

1.2.1 自动问答系统研究现状

1.2.1.1 基于结构化知识的自动问答方法研究发展现状

结构化知识是指以一种有组织、可管理和易于理解的形式存储和表示的知识。它通常以清晰的层次结构、关系或模式来描述，以便计算机系统或人类能够有效地访问、理解和利用这些知识。结构化知识具有多种形式，但在问答系统中，学界通常以知识库（Knowledge Base）形式的知识为主要研究对象。知识库是一组结构化的数据库，它以实体、关系、实体三元组的形式存储实体与关系信息[1]。基于结构化知识的自动问答系统的核心思想是通过分析用户提出的问题，将问题转化为一个结构化的查询，之后在预定义的结构化数据中执行该查询，最终将查询结果作为问题的答案返回给用户。知识库问答的研究方法主要包括信息检索方法（Information Retrieval，IR）和基于语义解析方法（Semantic Parsing-based Methods，SP）两类。

信息检索方法的基本思想是识别问句中的实体并链接到数据库，获取以该实体为中心的知识库子图，将子图中的每个节点都视作候选答案，再学习问句和候选

答案的向量表示，然后对候选答案进行打分排序、筛选，最后得到问题的答案[6]。Bordes 等人[7,8]通过学习单词和知识库三元组的低维向量嵌入，将问题转化为向量表示，同时将答案映射到同一空间，并采用知识库三元组自动生成问题作为训练，实现了弱监督下问题解析，但该方法仍需要人工制定特征或规则，但是由于其忽略了词序信息，使其无法处理复杂问题。Chen 等人[9]提出双向注意力记忆神经网络（BAMnet），通过注意力机制来学习问题与数据库之间的相互作用，提高了模型的可解释性。Wang 等人[10] 则采用检索与重排框架来访问知识库，并使用预训练的 BERT 模型对检索到的候选项进行重排，提高了候选答案的相关性。

基于语义解析方法的思想是首先将自然语言的语句转变为结构化的逻辑表征形式，然后将逻辑表征改写成数据库查询语句，从而获取数据库中的实例并进行排序，最终获得问题的答案。早期的语义解析模型通常采用基于模板或基于查询图等方法。近几年来，通常采用深度学习的方法，这类方法核心思想是将自然语言转化为向量。Dong 等人[11]提出基于卷积神经（CNN）网络的端到端语义模型，使用三个 CNN 分别分析问题和答案的相似度，从而获得答案。Hao 等人[12] 则认为，将问题表示为固定向量的方法难以准确表达核心信息，因此提出了一种基于交叉注意力的端到端模型，通过动态的表示方法提高了模型的精确性和灵活性。

信息检索方法基于关键词匹配和统计模型，计算成本低且适用于简单问题，但对复杂问题和语义理解能力有限。相比之下，语义解析方法能够更好地理解问句的语义和上下文，提供更精确的答案，但计算复杂度高，需要大量标注的语义数据进行模型训练。随着深度学习和自然语言处理技术的不断发展，语义解析方法展现了更优越的性能，并且现代的语义解析方法通常将自然语言转化为向量，与其他类型的知识问答系统具有相似性，使之具备更强的迁移性能。

1.2.1.2 基于自由文本的自动问答方法研究发展现状

基于自由文本的问答系统是一种用于从自由形式的文本中提取信息并回答用户提出问题的系统。现有的自由文本问答系统主流模式主要采用“检索-阅读理解”架构，其主要由检索器、阅读器构成。

检索器可以分为稀疏检索、密集检索、基于交互的检索等。稀疏检索常用方法有 TF-IDF 和 BM25[13]。其中，TF-IDF 通过计算词语在问题中的频率与它在整个语料库中的逆文档频率的乘积来衡量其重要性。BM25 则计算文档中的词频、文档长度和查询项的信息增益来评估文档的相关性。密集检索比较有代表性的有 Karpukhin 等提出的密集向量检索器（DPR，Dense Passage Retrieval）[14]，它是一个双塔模型，使用预训练的语言模型来生成问题和文档的密集表示，该密集表示相

较于传统的 TF-IDF 等稀疏表示方法，能够更好地捕捉语义信息。基于交互的检索器代表性工作是由 Seo 等[15]提出的 BiDAF，该方法引入了双向注意力流和上下文编码，通过将问题到文档的注意力和文档到问题的注意力进行结合，使模型能够在两个方向上传递信息，该交互式的结构允许模型更好地捕获问题和文档之间的复杂关系。

在自然语言处理领域，阅读器主要分为抽取式和生成式两种类型。抽取式阅读器专注于从文本中提取关键信息，通过选择文本中的关键句子或段落，并将这些片段组合成一个摘要，而不是创造新的文本内容。相比之下，生成式阅读器在抽取信息的基础上，进一步理解文本的深层语义，并通过重新组织语言来生成答案。传统的生成式阅读器通常采用循环神经网络（RNN）等方法[16]。近年来，预训练语言模型如 UniLM[17]，Bert[18]，ERNIE-GEN[19]等在阅读理解任务中得到了广泛应用，这些模型经过少量的微调后，即可在多个任务中展现出卓越的性能[20]。

尽管抽取式阅读器在准确性和可解释性方面表现更优，但抽取式阅读器在语言创造性方面存在局限，对于需要综合信息或深入解释的问题，抽取式阅读器可能无法提供满意的答案。生成式阅读器则能够通过理解文本的深层语义，生成包含创造性语言的答案，从而在处理复杂问题时表现出更强的能力[4]。

综上，检索-阅读理解架构结合了信息检索的广泛覆盖能力和机器阅读理解的深度分析能力，使得该架构能够有效处理更为复杂的问题。随着算法的不断进步和计算资源的日益丰富，这种架构能够充分利用非结构化的自由文本知识，无需将其预先结构化，为问答系统的设计，特别是在知识整合方面，提供了新的视角和方法。

1.2.1.3 基于文本对的自动问答方法研究发展现状

基于文本对的自动问答方法与基于自由文本的自动问答方法在研究框架上相似，包括问题解析、信息检索和答案抽取三个阶段。主要差异在信息检索阶段。基于自由文本的自动问答方法在检索时目标是与用户问题对应的答案片段，而基于文本对的自动问答方法只需要找到与用户问题相似的问题。因此可定义为问题搜索任务。近年来的研究主要采用深度学习的方法，如 Peng 等人[21]提出了一种增强的循环卷积神经网络（Enhanced-RCNN）模型来学习句子相似性。Mueller 和 Thyagarajan [22]提出采用改进的孪生长短时记忆（Siamese-LSTM）网络的用于评估句子之间的语义相似性，通过在词嵌入向量中添加同义词信息，结合使用曼哈顿度量，形成高度结构化的空间，表现优于手工设计特征和更复杂的神经网络系统。潘理虎等人[23]建立 Text-CNN 问句分类模型对问句进行分类，然后通过 Word2vec 词向量模型将问句中的词与词的空间向量相似度转化为语义相似度，并结合句法规

则进行分析，结果优于 TF-IDF 方法。丁邱等人[24]通过注意力矩阵和句子矩阵相互生成彼此注意力加权后的新的句子表示矩阵，将获取的新矩阵同原始矩阵拼接融合，丰富句子特征信息，经过 Transformer 深层语义编码后，计算最终句子相似度。

基于文本对的自动问答的研究方法与基于自由文本的自动问答的研究方法相似。而文本对结构化程度高于自由文本，并且搜寻目标与原始问句形式上更接近，因而基于文本对的自动问答任务挑战小于基于自由文本的自动问答任务。因此，在考虑知识融合任务时，可以采用基于自由文本的问答方法，并通过适当的调整来处理文本对形式的知识。

1.2.1.4 融合多源异构知识的自动问答方法研究发展现状

综上，现有主要的三类自动问答系统即基于结构化知识的自动问答系统、基于自由文本的自动问答系统和基于文本对的自动问答系统在研究方法和关键任务上各有不同和优势。其中基于结构化知识的自动问答系统检索效率更高，结果更准确；而基于自由文本知识的自动问答系统可以直接利用海量的非结构化文档；基于文本对的自动问答系统知识相对于结构化知识更易于整理，相较于自由文本处理难度更小。为了同时利用这三类结构优势，融合多源异构知识的自动问答系统成为该领域的一个重要研究方向。

早期具有代表性的融合方法如 Ferriucci 等人[25]设计的多专家系统，通过在不同来源知识上，使用不同的算法检索，提出多个候选答案，通过训练不同专家系统下的答案权重，按照权重分数进行排序。这种通过不同模块处理不同来源知识的方法被称为后期融合方法，由于各模块算法各异，后融合的算法增加了系统的复杂度。Sun 等[26]认为早期融合可以更灵活地结合来自多个来源的信息，并提出一种早期融合策略，使用一个统一的基于图卷积神经网络的模型同时学习文本和知识库表示，而不是分别训练后再融合，提高了模型对不同来源信息融合的效率。Agarwal 等[27]提出整合结构化的知识图谱和自然语言语料库，将知识图谱转化为自然文本，使其能够无缝集成到语言模型中。Meta AI 在 Agarwal 等人的基础提出了一种新的统一知识表示（UniK-QA）方法[28]，将结构化数据如列表和表格，以及完全结构化的知识库展平为非结构化数据。Khashabi 等人[29]提出了与知识形式无关的预训练问答系统（UNIFIEDQA），实验结果表明，该系统能够有效地处理多种不同的问答任务，并且具有显著的泛化能力，并且在单一问答系统中引入多种形式的问答任务能够提升模型的推理能力。

综上，前融合策略在处理不同结构的数据时展现出了高度的灵活性，使其成为当前研究的主流方法。同时，现有研究也揭示了结构化知识转化为非结构化知识并将其应用于问答系统的潜力。

1.2.2 多轮对话方法研究发展现状

对话系统主要包括闲聊型对话（Chit-Chat Dialogue，Chit-Chat）和任务型对话（Task-Oriented Dialogue, TOD）[30,31]两类。本文研究的汽车领域客服自动问答系统属于任务型对话，常采用多轮问答交互模式以帮助用户完成一些相对复杂的任务。这类系统广泛应用于虚拟助手、客服机器人、知识库问答等场景。任务型对话的研究方法主要包括流水线（Pipeline）型方法和端到端（End to End）型方法[32]。其中，流水线型方法是一种分步骤的处理方法，它将一个复杂的问题分解为多个较小的任务或步骤，每个步骤独立处理数据并产生中间结果，这些结果随后被用作下一个步骤的输入。此类方法的每个步骤均有明确的输入和输出，且通常需要人工设计和干预以提取所需的特征或进行特定的处理。相比之下，端到端方法更为直接，它将整个流程视为一个整体，直接从原始数据中获得最终结果，而不涉及独立的中间步骤。

1.2.2.1 流水线型多轮对话方法

流水线型多轮对话方法主要包括自然语言理解、对话管理、自然语言生成三个模块。

（1）自然语言理解模块

自然语言理解模块包括意图识别和语义槽填充两个子任务。其中，意图识别用于获取当前对话的目标，语义槽则是完成该目标所需要的信息。语义槽填充过程即从用户输入中抽取所需信息的过程。传统的意图识别方法包括基于卷积神经网络的方法与基于循环神经网络的方法[33]。近年来，意图识别任务与槽位填充任务联合方法成为一种新的研究趋势。Hakkani-Tür 等[34]使用长短期记忆（Long Short-Term Memory，LSTM），提出双向 RNN-LSTM 架构，用于联合语义槽填充、意图识别和领域分类，并构建了一个联合多领域模型，在微软 Cortana 真实用户数据上展现了其优势。Liu 等[35]提出了一种基于注意力的循环神经网络方法，该方法用于同时进行意图识别和语义槽填充，通过在基于对齐的循环神经网络模型中引入注意力机制，为意图分类和语义槽填充提供了额外的信息。实验结果表明，联合建模的方法在性能上优于单独进行意图识别和语义槽填充。

（2）对话策略管理模块

对话策略管理模块包含对话状态追踪和对话策略学习两部分。其中，对话状态追踪主要用于在对话系统中跟踪用户和系统之间对话状态的过程，其任务是不断更新语义槽值，以便系统能够在后续对话轮中更好理解用户的需求。在不同的对话系统架构中，对话状态追踪常使用规则驱动的方法、基于统计的方法或基于机器学习的方法来实现。由于深度学习在自然语言领域具有良好表现，结合注意力机制等的深度学习方法也逐步成为其主流方法之一。如 Zhou 等[36]提出通过多层自注意力逐渐构建了用户输入和系统回应的多粒度语义表示，并使用分段-分段相似性矩阵进行匹配。通过共享组件的 Attentive Module 实现自注意力和交叉注意力，提取并融合了跨多轮对话上下文的重要匹配信息。Wu 等[37]提出一种适用于多领域对话状态追踪的可转移生成器 TRADE，该方法无需预定义领域本体即可学习状态追踪，并通过领域共享实现了零样本对话状态追踪和少样本快速适应。

对话策略学习负责决定系统在与用户的对话中采取何种行动。通过确定系统在特定对话状态下应该采取的响应或动作实现系统的目标。早期通常采用手动定义规则的方法。近年来对话系统中常选择强化学习来学习对话策略，在强化学习的上下文中，系统通过与用户的交互来学习对话策略，以最大化预定义的奖励信号[38]。

（3）自然语言生成模块

自然语言生成模块负责生成系统对用户的自然语言响应。这个模块需要根据对话的上下文和当前对话状态来产生连贯、合理且与用户期望相符的文本。自然语言生成的方法包括基于模板的方法、统计机器翻译方法（SMT）以及神经网络方法，其中，神经网络方法包括循环神经网络、LSTM[39]等。近年来，随着大模型技术的不断发展，以 BERT 为代表的生成式大模型已成为自然语言生成领域的主流方法[18,40,41]。

综上，流水线型多轮对话方法将多轮对话系统划分为多个独立的子任务，并按顺序处理各子任务，每个子任务负责特定的功能，其优点在于结构清晰、易于理解和实现，每个子任务可以单独优化，能够灵活地集成各种模型和技术。然而，多个子模块的结构也可导致误差累积，并且系统难以实现端到端的全局优化。

1.2.2.2 端到端型多轮对话方法

端到端型多轮对话通常依赖于深度学习技术，例如循环神经网络（RNN）、长短期记忆网络（LSTM）和变换器（Transformer）等。这些模型能够捕捉对话中的上下文信息，并处理长距离依赖关系。端到端的方法将对话系统中的多个任务，如

意图理解、对话流程管理、系统响应生成等，整合到一个统一的模型中，实现从输入到输出的直接映射。Wen 等人[42] 出了一种基于神经网络的端到端训练模型，该模型能够与知识库结合，解决了序列到序列模型在处理任务型对话时的局限性，但该模型存在需要大量训练数据的问题。Eric 与 Manning[43]提出了基于注意力机制的序列到序列模型，该模型无需显式地建模用户的意图和槽位状态，但同样需要大量的训练数据。Lei 等人[44]设计了一种名为“信念跨度”机制，用于追踪对话中的信念状态，使得任务导向型对话系统可以采用序列到序列的方式进行建模，并通过强化学习进行优化，其性能优于传统的监督学习方法，尽管它在处理用户意图变化或不精确表达时存在挑战。Yang 等人[45] 通过微调大型语言模型 GPT-2，在对话会话级别上实现了对任务型对话的建模，并在 MultiWOZ 数据集上展示了良好的性能。

综上，尽管端到端对话系统在简化系统设计方面取得了进展，但这些方法通常需要大量的训练数据。相比之下，流水线型方法每个模块都有明确的输出，系统的可解释性更好。因此，尽管端到端方法是当前的研究热点，流水线型方法在工业界仍然是主流[31]。

1.2.3 大型语言模型在问答系统中的应用发展现状

语言模型的发展经历了统计学语言模型（Statistical Language Model, SLM）、神经语言模型（Neural Language Model, NLM）、预训练语言模型（Pre-trained Language Model, PLM）和大型语言模型（Large Language Model, LLM）四个阶段[46]。其中，大型语言模型（下文简称大模型）由于其基于巨大的语料库训练，在多个自然语言处理任务中均不断刷新最优性能，被广泛应用于机器翻译、情感分析、信息抽取及问答系统等自然语言的多个任务中[47]。

当前最具有代表性的基于大模型的问答系统为 ChatGPT[48]。ChatGPT 以广泛的语料作为训练数据，覆盖了丰富的知识，在应用中展现了通过利用模型自身知识实现强大问答的能力。Tan 等[49]广泛测试了 ChatGPT 及其他多个大模型，在 CQA 数据集上使用大模型自身的知识回答问题的能力，发现 ChatGPT 性能可以接近传统的专用问答模型，但当前的大模型在推理能力上仍有很大的改进空间，可采用思维链（COT）启发的提示以改善原始模型在某些特定类型问题上的性能。尽管大模型在问答任务上取得了重大的进展，但仍存在幻觉问题[50]，即产生看似合理但不符合事实的输出，且受到原始训练语料库的限制，大模型也难以整合最新知识。此外，由于采用公共数据集训练，大模型更适合于回答开放域知识，无法直接用于回答非公开的领域，尤其是与企业内部知识相关的问题。

针对上述问题，主流的解决方法包括大模型微调（Fine-Tune）和检索增强生成（Retrieval Augmented Generation，RAG）等技术。其中，微调是指使用专门的数据集在特定领域进行额外的训练[19]。对于超大型的大模型，更新所有参数是一项艰巨的任务，因此，在实际应用中，研究者与工程师通常采用冻结大部分参数，只训练部分参数的训练方法[51]。大模型微调方法包括参数高效微调（Parameter-Efficient Fine-Tuning, PEFT）[52]、大模型低秩适应（Low-Rank Adaptation of Large Language，LoRA）[53]和提示微调[54,55]（Prefix-Tuning 或 Prompt-Tuning，P-Tuning）等，微调过程重新训练少量参数效果即可接近全量参数训练。然而，微调方法仍需要一定的计算成本，并且当存在知识更新时，就需要再次进行微调，难以应对知识频繁更新的情况。检索增强生成技术是一种结合了检索和生成技术的方法，利用检索技术从大规模语料库检索与用户查询相关的信息，然后向语言模型传递检索到的信息[56]。Martino 等[57] 用一种称为知识注入的技术，将实体上下文数据从知识图谱映射到文本空间，以纳入大模型提示中；实验表明经过知识注入的 bloom-560m 模型优于未经知识注入的 text-davinci-003 模型，实验中后者模型参数量是前者的 300 倍。Jiang 等人[58]提出前瞻式主动检索增强生成模型（FLARE），通过迭代地使用对即将生成的句子的预测来预测未来的内容，然后将其用于查询来检索相关文档，再重新生成句子，该方法在四个长篇知识密集型生成据集上进行了全面测试，取得了优良的性能，但多次调用大模型检索产生了更多的性能开销。

综上，尽管微调的方法可达到与全量训练相媲美的效果，但该方法将新增知识存储于模型参数中，因此每当知识更新则需重新微调模型，灵活性和可维护性不佳。检索增强生成技术则将知识储存在模型外部，该技术架构解决了大模型中的信息获取问题，新增知识无需重新训练或微调模型。这一特性不仅可缓解公共领域大模型的幻觉问题，也适用于企业构建基于内部知识的自动问答系统。对于本文研究中汽车客服自动问答系统所面临的知识更新频繁和私有化程度高的挑战，检索增强生成技术为客服问答系统提供了一个新的解决方案，但目前检索增强生成技术正处于起步阶段，在实际工程应用中面临着诸多挑战。例如其庞大的参数量导致的推理延迟问题，直接影响了问答系统的响应速度。因此，如何设计出既能够保证快速推理又能维持整体性能的模型，仍需要进一步的研究和探索。

1.3 论文结构安排

1.3.1 主要工作内容

本文围绕 XX 汽车制造企业“智能客服问答系统建设”课题，针对现有客服自动问答系统中面临的知识利用率低、无法应对需要多轮交互的复杂问题以及问答系统检索性能弱等挑战，开展基于大型语言模型的汽车客服自动问答系统研究与设计。具体工作内容包括：

（1）针对现有自动问答系统无法利用多种结构知识的问题，本文提出一种多源异构客服知识的融合方法，通过预设规则，将结构化的表格、知识图谱、文本对统一为自由文本的形式，然后通过切分、标题增强及向量化等步骤进行处理，以便在基于自由文本的自动问答系统中高效利用。在此基础上，通过公开数据集进行实验验证，以评估知识融合的效果。

（2）针对系统无法处理多轮交互的复杂问题，本文基于提示工程的理念，提出一种基于大型语言模型的提示框架，并在此框架下设计了一种多轮对话问题解决方案。该方案将提示框架应用于多轮对话的意图识别、槽位填充和自然语言生成等多个子模块，在此基础上，利用大型语言模型的泛化能力提升各子模块的性能。

（3）针对现有系统检索能力不足的问题，本文提出了一种将大型语言模型增强的检索生成技术应用于企业自动问答系统的方法。首先，使用预训练的密集嵌入模型将问题文本与知识库转换为高维向量，这些向量比传统的稀疏向量包含更丰富的语义信息。然后，利用大型语言模型作为阅读器，生成最终答案，从而提升答案的抽取和总结能力。最后，通过实验研究模型选型和关键参数设计，以平衡问答系统的准确性和响应速度。

（4）针对项目要求，本文设计并开发面向汽车客服的自动问答 Web 系统，并完成系统的应用验证。

1.3.2 论文组织结构

本文结构安排如图 1-1 所示，主要包括六个章节，具体结构如下。

第一章：绪论。主要介绍课题研究背景及意义，并分析自动问答系统、多轮对话方法及大型语言模型在问答系统中的研究发展现状，以及阐述本文的主要研究内容与论文组织结构。

第二章：需求分析及总体设计。本章首先分析了 XX 汽车制造企业建设汽车客服自动问答系统的需求，然后指出现有自动问答系统存在的问题，并分析了现有知

识结构的特点。在此基础上，提出了整个问答系统的设计方案，并详细介绍了知识统一表示模块、对话管理模块和知识问答模块三大核心模块的技术路线。

第三章：多源异构客服知识融合方法研究。针对现有自动问答系统无法有效利用多种结构知识的问题，本章提出了一种融合多源异构客服知识的方法。通过预设规则，将结构化的表格、知识图谱、半结构化的文本对统一为非结构化的自由文本形式，并通过切分、标题增强和向量化等步骤进行处理，以便在基于自由文本的自动问答系统中使用。最后，在 NLPCC 2018 KBQA 和阿里天池公益之星问题相似度挑战大赛数据集上验证了本方法的有效性。

第四章：基于大模型的提示工程设计及多轮问答实现。针对现有自动问答系统无法处理多轮交互的复杂问题，本章提出了一种基于大模型的提示框架，并阐述了该框架的原理和设计。然后，基于该框架设计了一种多轮对话方法，将提示框架应用于多轮对话的意图识别、槽位填充和自然语言生成等多个子模块。通过大模型的泛化性能提升各子模块的性能。最后，在 ATIS 数据集上验证了本方法的有效性。

第五章：基于大模型检索增强生成的知识问答模块设计。针对现有自动问答系统检索能力不足的问题，本章提出了一种将大模型增强的检索生成技术应用于企业自动问答系统的方法。首先使用预训练的密集嵌入模型，将问题文本与知识库转化为高维向量，然后利用大模型作为阅读器生成最终答案。

第六章：汽车自动客服问答系统设计及验证。在前几章研究的基础上，本章整合了知识统一表示、多轮对话和知识问答等核心模块，设计了完整的汽车客服自动问答系统，并通过系统阐述了本文方法的应用验证情况。

第七章：总结与展望。对本文的工作进行分析和总结，提出未来的改进和研究方向。

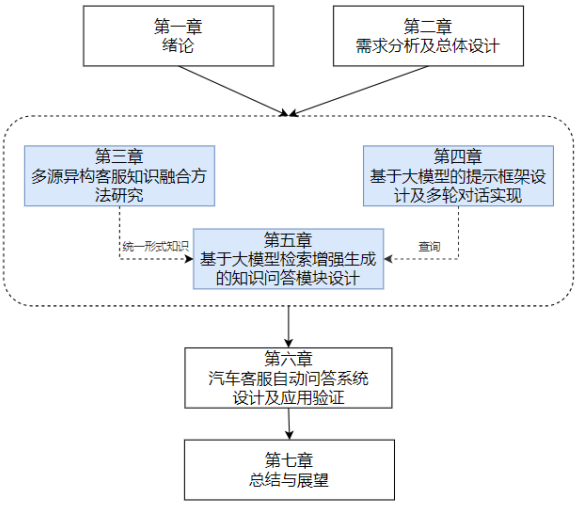

图 1-1 论文组织架构

第二章 需求分析及总体设计

2.1 汽车客服自动问答系统需求

客服业务的核心目标是及时、有效解决用户问题、减少用户抱怨并最终提高用户满意度。为达成这一目的，XX 汽车制造企业建立了三线客服体系，如图 2-1 所示。在该流程中，一线客服包括普通坐席专员和线上自动问答系统，可直接处理客户的常见问题，常见问题可通过查阅整理好的知识库解决；当普通坐席专员遇到无法解决的问题时，则将问题流转至二线客服，二线客服由专业坐席组成，二线客服不仅专业性更强，可利用的知识也更为丰富，包括维修手册、服务专案、营销活动通知等未及时整合到知识库中的文档；当二线客服仍无法有效处理用户问题，则将问题流转至三线客服，三线客服由技术支持工程师组成，三线客服具备更全面、详细的维修手册、技术文档等。

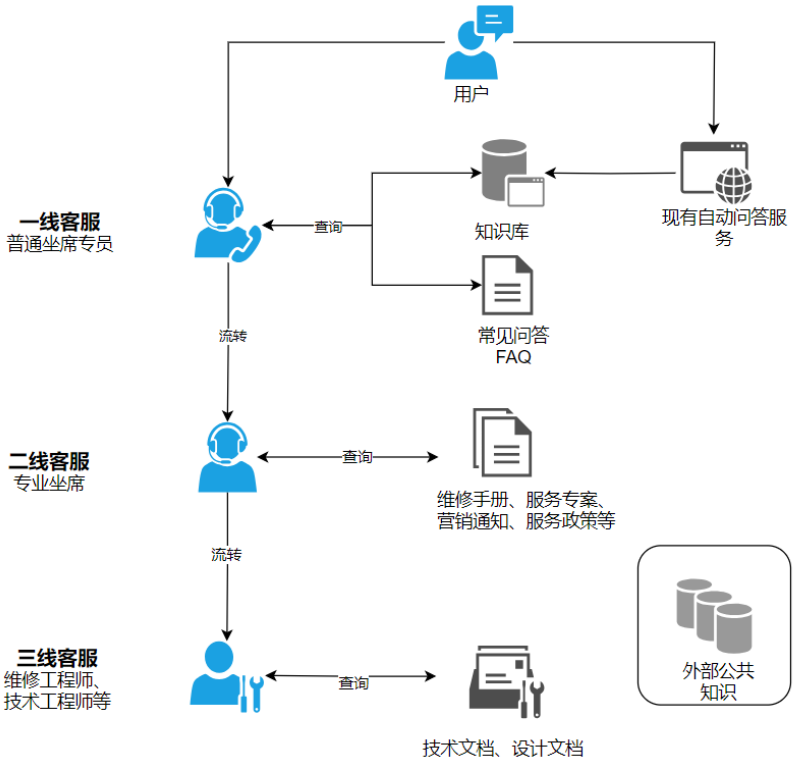

图 2-1 XX 汽车制造企业三线客服流程

可见，客服坐席和自动问答系统是响应用户最直接快速的渠道。然而，现有的自动问答系统存在的问题限制了系统在实际应用中解决用户疑问的能力，进而影响了用户对系统的信任，未能实现减轻人工客服工作负担的预期目标。因此，企业

不得不投入至少 60 名客服坐席，以确保服务的连续性，这导致了人力投入成本大，效果不佳。因此，XX 汽车制造企业提出了“线上智能客服问答系统建设”的研究课题，研究现有自动问答系统的改进方法，提升系统知识利用率，扩大问题处理范围；并增强处理复杂问题的能力和知识检索能力，以提高答案的准确性。本文即是以该课题为研究背景，围绕目前 XX 汽车制造企业的汽车在客服自动问答系统的问题和需求，开展改进方案设计和关键技术研究。

2.2 企业现有汽车客服自动问答系统现状及知识特点

2.2.1 企业现有汽车客服自动问答系统现状

根据调研，XX 汽车制造企业现有的自动问答系统采用基于文本对问答的方法。首先需人工将结构化、非结构化和文本对等多种结构的原始知识文档转换为特定的文本对格式，将其录入系统作为知识库。系统检索答案时采用关键词检索技术，检索出与原始问题近似的标准问题，然后获取标准问题对应的答案。该系统存在问题具体表现如下：

（1）知识更新慢、利用率低

在知识库构建过程中，将多源异构知识转换为标准文本对格式的工作难以实现自动化，需要人工处理。面对庞大的知识库，人工录入方法效率低，导致大量知识未能及时整合利用，进而造成知识更新滞后和缺失。此外，随着新知识持续产生，而旧知识的录入速度远远滞后于新知识的生成速度，进一步加剧了知识利用率低的问题。

（2）无法解决复杂问题

售后问题的专业性要求解决此类问题必须获取特定的、必要的前置信息，但往往需要多轮问答才能获取到所有必要信息。而现有系统仅能处理单轮对话输入，缺乏上下文理解及多轮交互的能力，因此无法有效解决这类复杂问题。

（3）知识库检索能力弱，影响回答准确性与效率

现有的关键词检索方式常导致大量无关结果的出现，也常出现遗漏形式不同而意义相同的信息，导致回答不准确。

综上，根据现有自动问答系统存在的问题，需要对知识库进行重构，并且改进问题理解及答案检索的流程和方法。

2.2.2 企业现有知识特点

企业可用于构建自动问答系统的知识库的知识来源丰富、结构各异。主要包括结构化、文本对、自由文本三类知识，具体如下：

（1）结构化知识

企业涉及的结构化知识主要包括企业信息系统中的数据表和知识图谱两大类。其中，数据表包括车型配置表、配件电子目录等。此外，虽然人工维护的电子表格和文档中抽取的表格包含非结构化的知识，但经过简单的清洗和整理，可以被视为结构化的数据表，从而便于在单一系统中被应用。知识图谱主要来源于企业内部整合的知识资源。尽管其规模相对较小，但经过人工的精心整理，能确保数据的高质量和准确性，其格式遵循资源描述框架（RDF）标准。

尽管结构化知识因其格式的规范性使其在单一系统中具有较强的应用优势，但由于不同信息系统之间在数据格式化方式上存在差异，跨系统使用时仍然面临挑战。现有的问答系统往往无法直接利用这些结构化知识。

（2）文本对知识

文本对知识来源于现有自动问答系统中知识库的导出数据，以及 400 客服中心基于历史通话收集整理的常见问答，数据规模小，仅约 10000 条，但由于其主要来源于历史客户的真实问题，因此能够覆盖用户的常见问题，蕴含较高质量的问答数据知识。其数据形式如图 2-2。

标准问题	检索关键词	答案
车机腾讯我的车-集结的方式有几种？	腾讯我的车	您好，腾讯我的车-发起集结的方式有两种，一种是在手机端腾讯我的车小程序发起集结，另一种是在车机导航端我的车发起集结。
车机如何连接蓝牙播放蓝牙音乐？	连接蓝牙	您好，在车机本地应用找到【蓝牙音乐】，在下端点击打开蓝牙开关，连接手机蓝牙后即可播放蓝牙音乐。
车机可以拨打蓝牙电话吗？是否同步通讯录？	通讯录	您好，车机是可以连接手机蓝牙拨打蓝牙电话的，连接蓝牙后在车机更多应用找到【蓝牙电话】就可以拨打蓝牙电话，蓝牙电话界面也可以找到通讯录和最近通话。
车机能否播放U盘歌曲、视频？	U盘视频	您好，车机支持播放U盘音乐和视频，插入U盘后在车机本地应用界面找到USB视频和USB音乐播放。
车机如何使用多媒体功能播放视频；	播放视频	您好，车辆不支持在线网络播放视频，可以观看本地视频。播放方法：（1）将视频拷贝到U盘；（2）插入U盘；（3）打开本地媒体，找到USB视频，点击进去播放您想观看的视频！
车机支持哪些USB视频格式？	播放格式	您好，支持的播放视频格式有：ASF/AVI/FLV/WMV/MKV/MOV/VOB/MP4/M4V/3GP/TS。
车机收听本地电台需要消耗数据流量吗？	消耗流量	您好，收听本地电台不需要消耗数据流量，车机无剩余流量也可以正常收听本地电台。

图 2-2 文本对数据形式（部分）

（3）非结构化知识

企业中可用于构建问答知识库的非结构化知识主要以自由文本的形式存在，其来源广泛，包括用户手册、保养手册、维修手册、服务方案、政策法规、保险知识以及用车常识等。作为知识库中最为丰富和庞大的部分，非结构化知识具有形式

多样、更新频繁的特点，有助于显著扩大知识库的覆盖范围。然而，这些特性也使得将非结构化知识转换为结构化知识或文本对极具挑战性。结构化知识和文本对由于其规范性，更易于被系统处理和利用。传统的问答系统设计多依赖于这两类知识结构，本课题中 XX 汽车制造企业现有的问答系统也主要采用这一设计模式。然而，在该模式下，大量的非结构化知识难以被有效利用，大大限制了问答系统的性能和准确率。

2.3 汽车客服自动问答系统总体设计

根据企业现有汽车客服自动问答系统需求、现有自动问答系统现状、现有知识特点，本文提出了自动问答系统设计方案，该方案基于自由文本问答方法，采用“检索-阅读理解”架构。为提高系统知识利用率和解决复杂问题能力，设计了知识融合模块、对话管理模块和知识问答模块三大核心模块。同时，方案还包括必要的用户前端与管理前端，以实现人机交互。其总体设计方案如图 2-3 所示，组成及原理如下：

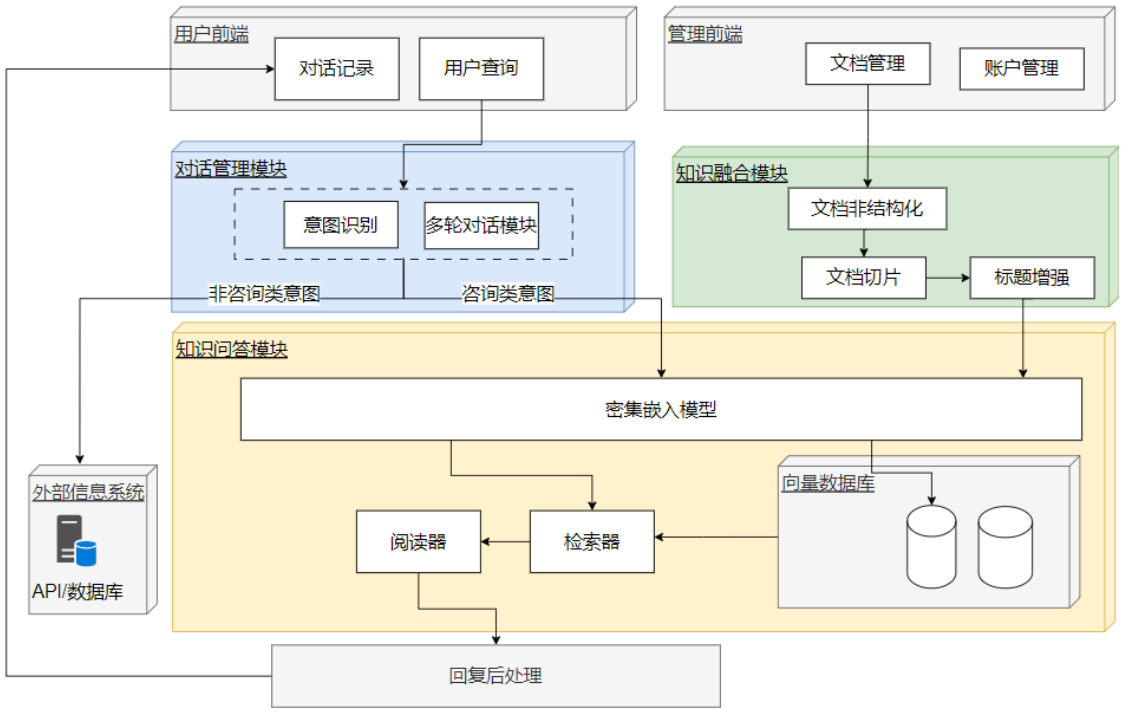

图 2-3 问答系统整体结构

（1）知识融合模块：本文采取了一种与传统问答系统不同的构建知识库的方法，即逆向处理策略。该策略将结构化知识和文本对逆向转换为非结构化的文本形式。该模块将管理员上传的知识原始文件作为输入，通过识别文件扩展名来识别其知

识结构类型，并根据结构类型应用预定义的规则进行自动化处理，将其转换为文本并切分为特定长度的文档切片。这些文档切片随后通过密集嵌入模型编码为高维向量，并存储在向量数据库中。

（2）对话管理模块：该模块包含意图识别和多轮对话两个子模块。其中，多轮对话模块采用流水线型架构，包括自然语言理解、对话策略管理、自然语言生成等子模块。这些子模块采用本文提出的提示框架方法进行设计，该方法利用大型语言模型的泛化能力，无需修改模型的原始参数，即可灵活适应不同的下游任务。模块的输入来源于用户输入的内容，通过大模型对用户的意图进行分类，并根据意图调用相应的下游任务，初始化所需的语义槽位。在多轮对话模块中，大模型用于抽取槽值以填充语义槽，并根据槽位的填充情况决定回复策略。一旦槽位填充完毕，大模型首先修改原始问句，然后将解决问题所需的槽值信息融入其中，并将修改后的问句输出至知识问答模块以检索答案。

（3）知识问答模块：该模块基于检索增强生成的思想，结合当前非结构化知识常用的检索器-阅读器的架构进行设计。其中，检索器通过密集向量检索机制工作，将问题文本和知识文本映射到同一密集向量空间中，并计算它们之间的向量相似度。检索器召回与问题向量相似度最高的若干篇相关文档，并将其输出至阅读器。阅读器采用大型语言模型作为生成式阅读器，理解问题的语义并生成符合自然语言规范的答案。

2.4 本章小结

本章分析了企业客服业务运作流程与建设汽车客服自动问答系统与需求，指出重构现有自动问答系统的重要性。然后分析了现有的自动问答系统存在的问题，以及现有可利用的知识结构特点，在此基础上，提出自动问答系统的设计方案，该方案主要由知识统一表示模块、对话管理模块及知识问答模块三大核心模块的组成。

第三章 多源异构客服知识融合方法研究

本章提出了一种多源异构客服知识的融合方法，该方法无需大量人工转化和清洗，能够将结构化知识、文本对及自由文本等多源异构的客服知识融为统一的自由文本形式，以便在基于自由文本的自动问答系统中被使用，提高知识的利用率。

3.1 多源异构客服知识融合方法

为提高客服知识利用率，需要将多种结构的知识进行融合。根据本文 2.2.2 对企业现有知识结构分析，现有知识结构主要由结构化、文本对、自由文本三类知识组成。尽管结构化的知识更易于利用，但将知识结构化需要持续、繁杂的处理，且企业现有的结构化客服知识在企业中占比小，对其重点处理和利用的价值不高。因此，根据企业实际情况出发，以及现有自然语言处理技术对自由文本处理的良好表现，将结构化知识、文本对、自由文本知识统一融合为自由文本的方法更为高效。其中，结构化文档中又包括知识图谱和数据表，两者结构存在显著差异，故需分别处理。综上，本文提出如图 3-1 所示的多源异构知识融合模块整体流程。

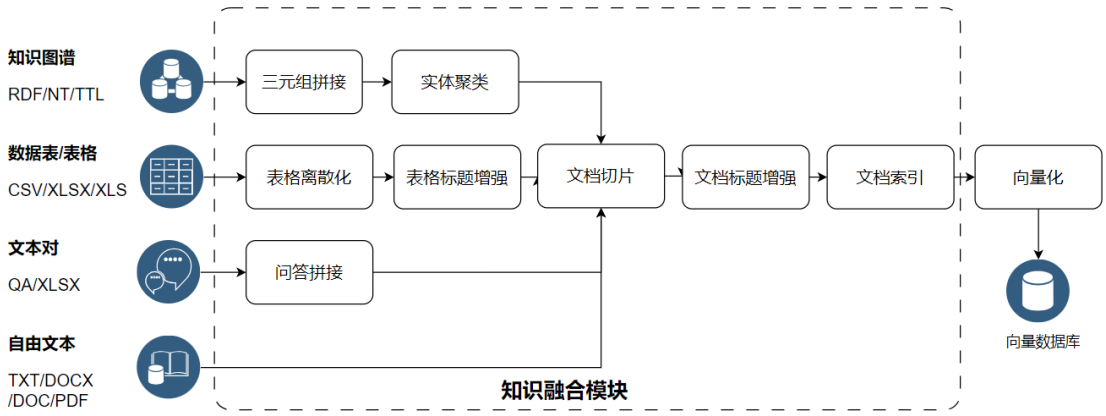

图 3-1 多源异构知识融合模块流程

如图所示，主要包括知识图谱、数据表及表格、文本对、自由文本的处理，以及通过切片文档进行融合等。其中，不同结构的知识以不同文件后缀区分，根据其结构特点采用预设流程进行处理和融合。

3.1.1 知识图谱处理

如图 3-1 所示， 知识图谱的处理逻辑包括三元组拼接、实体聚类及文本切片三个步骤。

（1）三元组拼接

三元组拼接的目的是将知识图谱中的三元组转化成自然语言语句。知识图谱采用形如$\mathrm{T}=(e_1,\mathrm{r},e_2)$的三元体结构，其中$\mathrm{e}_1$、$\mathrm{e}_2$分别表示实体 1 与实体 2，$r$ 表示实体关系。实体 1 为名词或名词短语，用于描述事物的具体实例或类别。实体关系为动词或形容词短语，用于描述两个实体之间的某种关系或属性。实体 2 为主体相关联的值或对象，可以是一个具体的数值、实体的名称、另一个三元组等。因此实体 1、关系、实体 2 分别可以作为语句中的主体、谓词和客体，共同构成中文句法的完整主谓宾语句，可直接按照 Sentence=concat（e_1，r，e_2），拼接成自然语言语句，其中 concat 表示拼接函数，用于将字符串直接首尾相连进行拼接。在知识图谱给定的子图中，边的数量反映了关系实例的总数，同时三元组拼接过程将生成相应数量的句子以描述这些关系。如图 3-2 所示，三元体非结构化处理流程，图中左侧为本文的知识图谱部分子图，其中，存在三对三元组，实体 1 均为“车身稳定系统”，三条边代表三组关系，分别连接三个实体 2，因此可拼接成三个自然语言句子的文本，如图 3-2 中右侧所示。

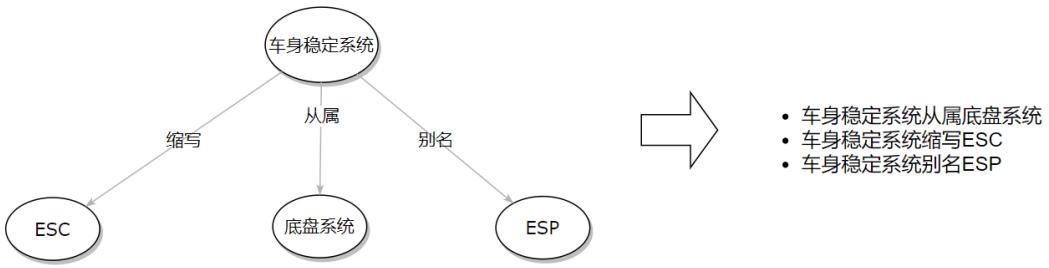

图 3-2 三元体非结构化处理

（2）实体聚类

由于本文研究的自动问答系统主要面向保有用户，在回答中增加适当的拓展信息可增强对用户的支持，故采用实体聚类处理有助于在召回文档时返回较多同一实体的其他信息，提高答案丰富性。实体聚类具体过程是将所有相同实体 1，即e_1相同的句子归到同一中间文档中，如图 3-2 中的以“车身稳定系统”为例，将相关的三元体关系归类到一个中间文档。

（3）文本切片

文本切片通过将长文档分割成更短、更易于管理的文本片段，从而加快处理和检索速度，并提高答案的相关性。具体方法是设定超参数l_{max}，用于控制单个文本片段的最大字符长度。在文本切片过程中，选择最接近第l_{max}个字符左侧的第一个句末标识符（如句号）作为切分断点。根据现有主流大模型能够支持的最大上下文长度的限制[59,60]，本文中l_{max}取 250 个字符。

3.1.2 数据表及表格处理

数据表及表格的处理过程包括表格离散化与拼接、表格标题增强与文本切片三个步骤。具体流程如图 3-3 所示。

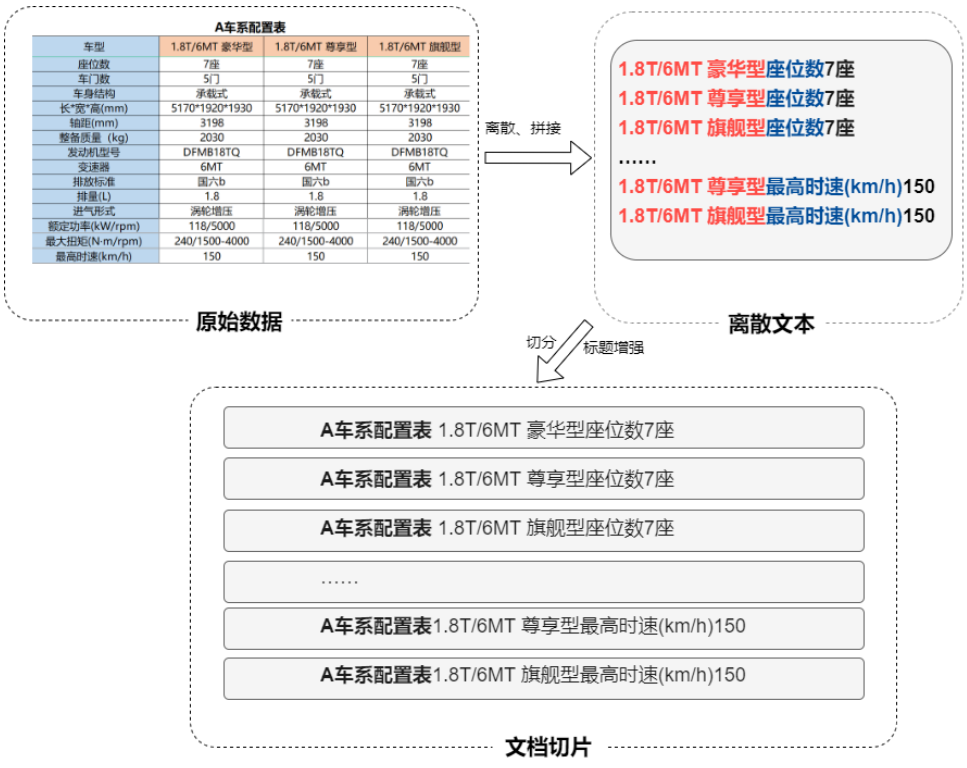

图 3-3 表格类文档的处理

（1）表格离散化

表格离散化是将表格离散为多个自然语言语句描述，离散化前需要进行预处理。人工编制的表格通常包含明确含义的显性字段名称。而信息系统中的数据表通常使用缩写或编码作为字段名称，其可读性较差。因此，需将数据表字段名称解析

为具有可读性的字段名称，该过程根据字段的抽象程度选择人工或机器自动翻译。经过预处理的数据表与表格采用相同的方法进行后续处理。

经过预处理后的表格参照知识图谱的三元组拼接过程，将表格的行标签视为实体标识，列标签被视为属性，而对应的行列值则代表另一实体，然后以类似于处理知识图谱的方法进行拼接，将有组织的表格离散为多条“实体标识-属性-值”文本的结构，本文将此过程称为表格离散化。由于在后续的自动问答系统中采用“检索-阅读理解”架构，原始知识经过检索后需经过生成式大模型抽取、总结后生成答案。因此原始知识不需要严格符合语法规范，即使行标签与列标签的位置发生转置，大模型仍能正确识别并将其转化为符合句法规则的语句。故无论处理实际数据表中行列标签的顺序，均采用相同的方法进行转换。

经过离散化后生成的各文本之间内容仍然高度相关，因此无需类似知识图谱的聚类过程。

（3）表格标题增强

在 XX 汽车制造企业的客服知识中，数据表中通常为车辆参数、配件信息等数据。此类数据在结构上相同、内容高度相似，但因车型或配置的差异而有所区分。例如：A 车型和 B 车型的配置表在形式上完全一致，但参数值不同。而表格标题中通常包含可区分上述内容的信息，如车型名称。表格标题增强具体过程是将表格标题直接拼接到离散的文本之前，表格的标题通过文件名或正则式从文档中抽取。对于多维表格，需要解析成二维表格后按上述方式进行处理。

（3）文档切片

对数据表和表格的文档切片的过程与处理知识图谱的方法一致，因此不再赘述。

3.1.3 文本对处理

文本对的处理过程包括问答拼接和切分文档两步。

（1）问答拼接

由于文本本质上是由自由文本组成，因此仅需将问题与答案进行拼接。为了提升拼接后的文本的可读性，拼接时额外引入固定形式连接文本，形如：“以下是$\{Q_i\}$的答案：$\{A_i\}$”。其中$\{Q_i\}$与$\{A_i\}$分别表示第 i 组文本对的问题和答案，如表 3-1 所示。

表 3-1 文本对处理示意

原始数据	问题：轮胎的更换周期是多久？
	答案：建议三年或六万公里，具体与您的轮胎型号、使用路况有关
拼合文档	以下是轮胎的更换周期是多久的答案：建议三年或六万公里，具体与您的轮胎型号、使用路况有关。

（2）文本切片

由于相邻文本对之间关联较差，因此无需聚类，可按照每一组文本对单独形成一个切片的方式进行切分。

3.1.4 自由文本处理

对于自由文本，可直接按照字符数量进行切分，切分过程中采用重叠切分的方法，以避免遗漏重要上下文信息。具体方式是切分的每一文本片段均包含上一文本片段末尾至少n_p个字符，其中n_p是一个预设参数，代表重叠字符的数量。切分断点为上一片段从末尾往前数第n_p个字符位置之前的句末标识符。本文中n_p设为 50。如图 3-4 所示，将原始文档自由文本切分的前后对比，加粗部分字体为重叠切分内容。受篇幅所限，该例中$l_{max} = 100$，$n_p = 15$。

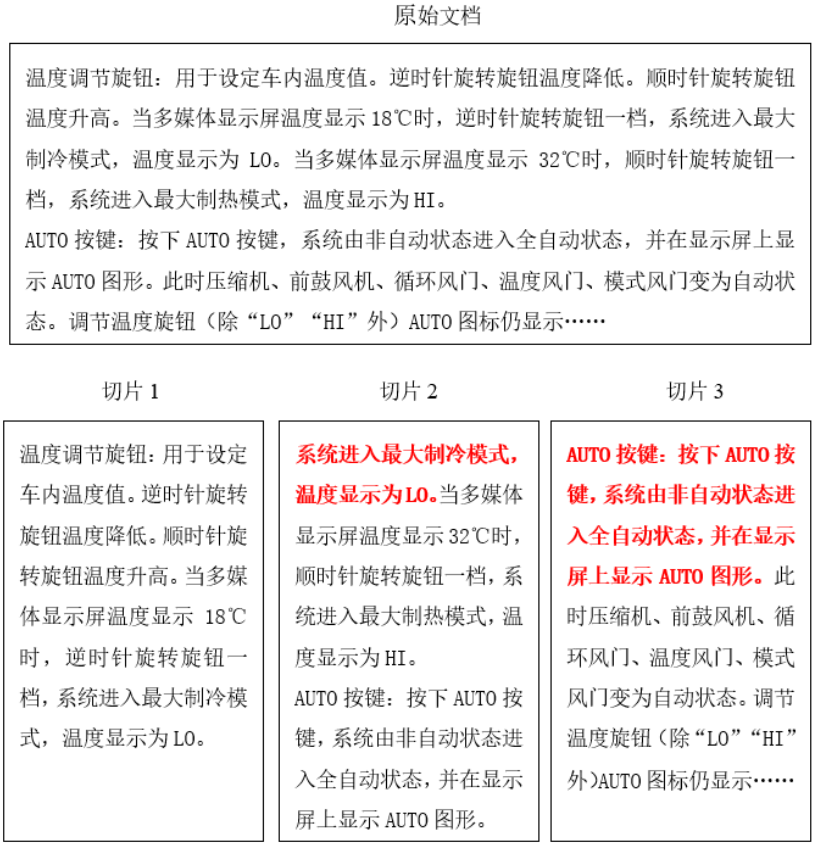

图 3-4 自由文本处理前后对比（部分）

3.1.5 切片文档的处理

（1）文档标题增强

经过 3.1.1~ 3.1.4 步骤完成各类文档的处理后形成的形式相近的非结构化切片文档。对切片文档进一步进行文档标题处理，其过程与处理表格形式知识中采用的表格名称增强的方法类似，即将文件名与文档切片拼接，不在赘述具体过程。

（2）索引数据构建

索引数据构建是将文档切片转化为向量数据库输入所需的标准格式，该格式取决于选用的向量搜索框架。本文选用 FAISS[61] 作为向量搜索框架，并使用 Langchain 封装。因此需将文本切片转换为 Langchain 所要求的 JSON 格式。基于此格式，还需将文档名称等必要信息作为元数据写入，以便后处理时使用，统一表示模块最终输出示意如图 3-5 所示。

```
{"docment":
    {"pagecontent":"以下是《A车型使用手册》内容：\nA车型外观参数\n豪华型车长4600mm……",
     "metadata":
         {"source":"XX车型使用手册",
          "title":"A车型外观参数"
          }
     }
}
```

图 3-5 知识融合模块最终输出示意

3.2 实验验证及分析

3.2.1 数据集及评价指标

本文提出的多源异构客服知识融合方法，是将结构化知识、文本对及自由文本均转化为自由文本形式，并通过一个统一的基于自由知识自动问答系统框架（详见本文第五章）进行处理。为评估这一转化过程对知识利用的影响，本文通过实验，将所提出的方法分别与现有的基于结构化知识的问答方法、基于文本对的问答方法进行对比。其中，结构化知识方面，本文评估了知识图谱问答（KBQA）任务；文本对方面，文本对问答系统主要任务是寻找相似问句，即文本相似度任务。

（1）KBQA 任务数据集及评价指标

KBQA 的实验验证采用 NLPCC（CCF 国际自然语言处理与中文计算会议）2018 KBQA 评测任务[62]提供的数据集。该评测任务提供了一个大型中文知识库，其中包含 872 万个实体和 4794 万对三元组[63]。训练集包含 24479 对问答，测试集共有 618 个问题。

评价指标采用答案精确匹配（EM，Extract Match）：

$$EM = \frac{\text{抽取正确的答案数}}{\text{总样本数}} \quad (3\text{-}1)$$

由于本文提出的方法采用生成式模型作为阅读器，生成答案文本与正确答案文本可能不完全一致，答案需人工进行评价。以下列问答为例，对样本评价标准进行说明。答案评价标准如表 3-2 所示。

问题：马丁·泰勒青少年时期在哪个球队踢球

答案：克拉姆灵顿少年队（正确答案）

表 3-2 KBQA 评测任务答案评价标准

情形描述	生成答案举例	判定
生成的答案与真实答案完全一致	克拉姆灵顿少年队	Y
生成的答案包含真实答案，同时也包含为了生成自然语言产生的虚词、问候语、引述题干的内容	你好，根据相关文档显示，马丁泰勒青少年时期在克拉姆灵顿少年队踢球。	Y
除了包含答案外，还包含与问题和答案密切相关的补充信息	马丁泰勒青少年时期在克拉姆灵顿少年队踢球，司职后卫	Y
生成答案错误（事实、实体错误）	1、马丁·凯利青年队克拉姆灵顿少年队（实体混淆） 2、马丁·泰勒青年队利物浦（事实错误）	N
虽然包含答案，但包含过多无关信息	马丁·泰勒运动项目足球，马丁·泰勒青年队克拉姆灵顿少年队，马丁·泰勒所属运动队沃特福德足球俱乐部，马丁·泰勒国籍英格兰	N
正确回答问题，但在检索器中并未召回包含答案的文档	/	N

本实验旨在评估知识统一表示方法的有效性，大模型中使用了广泛的语料进行训练，模型参数中包含大量的知识，需剔除大模型利用自身知识回答问题的情况。故本文将表 3-2 中最后一种情形“正确回答问题，但在检索器中并未召回包含答案的文档”视为错误，因为在此情形中，统一表示后的知识并未被检索器成功召回。

本文的基线模型采用 NLPCC 2018 KBQA 评测任务赛事前三名的方法：SEU-WDSKBQA、Yiwise-KBQA、NDers、XJBot。

（2）文本相似度任务数据集及评价指标

1）实验数据集

本文中文本对任务数据集及评价方法来源于阿里天池“公益 AI 之星”挑战赛-新冠疫情相似句对判定大赛。该数据集中整理了近万条真实语境下的提问问句对，要求选手通过自然语言处理技术识别相似的患者问题。该数据集源自真实环境，并经过专业人工清洗，数据的质量优良。此外，数据集中包含了丰富的专业名词，因此适合作为专业领域问句相似度算法评价数据集。

数据格式如图 3-6 所示。每组数据提供了 query1 和 query2 两个句子、问题分类及相似度标签。

	id	cate...	query1	query2	label
1	0	咳血	请问呕血与咯血有什么区别?	请问呕血与咯血这两者之间有什...	1
2	1	咳血	请问呕血与咯血有什么区别?	请问呕血与咯血异同?	1
3	2	咳血	请问呕血与咯血有什么区别?	请问呕血与咯血怎么治疗?	0
4	3	咳血	请问呕血与咯血有什么区别?	请问呕血与咯血是什么原因导致...	0
5	4	咳血	请问呕血与咯血有什么区别?	请问呕血与咯血与其他疾病有关...	0
6	5	咳血	老年人吃百令胶囊能喝鸽子汤吗?	老年人可以百令胶囊和鸽子汤一...	1
7	6	咳血	老年人吃百令胶囊能喝鸽子汤吗?	老年人服用百令胶囊后能否喝鸽...	1
8	7	咳血	老年人吃百令胶囊能喝鸽子汤吗?	老年人吃百令胶囊和鸽子汤对病...	0
9	8	咳血	老年人吃百令胶囊能喝鸽子汤吗?	老年人吃百令胶囊能起作用吗?	0

图 3-6 阿里天池“公益 AI 之星”挑战赛-新冠疫情相似句对判定大赛数据集格式（部分）

2）评价指标

为便于与基线方法对比，采用数据集提供方提供的评价指标，即准确率：需要预测 query1 与 query2 是否相似，是则为 1，否则为 0。计算公式为：

$$p = \frac{\text{正确预测数目}}{\text{总问题数目}} \tag{3-2}$$

由于该评价任务仅评估问句相似率，不考虑答案匹配阶段，本文参照该任务的要求也仅评估检索阶段性能，而非最终答案匹配指标。此外本文检索阶段采用归一化向量评估相似度，相似度范围取值范围在 0 至 1 之间的连续数值，且该数值非均匀分布，集中于 0.6 至 1 之间，因此需按照赛事评估方式，设定阈值将连续的相似度数值进行二值化处理。通过在验证集上的性能评估，本研究选取 0.92 作为区分正负例的阈值。

3）基线模型及实验环境

与本文进行对比的基线模型为：

Siamese-CNN[64]：该模型是一种卷积神经网络架构，该架构最初设计用于人脸验证和签名验证等任务，但也可以用于其他文本相似性的任务。

Siamese-LSTM：结合了 Siamese 网络和长短时记忆网络（LSTM）的神经网络架构。类似于 Siamese-CNN 的思想，但在处理序列数据时采用了 LSTM 的结构。

Siamese-BiLSTM[65]： Siamese-LSTM 的改进，将其中的 LSTM 替换为双向长短时记忆网络。

QSTransformer：基于 Transformer 的问句相似度计算方法。在获取句子语义特征前引入交互注意力机制比较句子间词粒度的相似性。

epidemic-sentence-pair：该数据集赛事第一名的方案，该方案采用了多个模型 BERT-wwm-ext、Ernie-1.0 和 RoBERTa-large-pair 融合，并且在训练过程中引入外部数据。

实验环境如表 3-3。

表 3-3 多源异构知识融合实验环境

配置	参数
CPU	48 vCPU AMD EPYC 9654 96-Core Processor
GPU	RTX 4090(24GB) ×4
内存	240GB
系统	Ubuntu 20.04.4
语言环境	python 3.8.10
CUDA 版本	11.8

3.2.2 实验结果与分析

（1） KBQA 组实验结果与分析

KBQA 组结果如表 3-4 所示。如表 3-4 所示，本文提出的方法在不同阅读器配置性能存在差异，但均展现了优于基线的性能。

表 3-4 实验结果-KBQA 组

方法/系统名称	EM
SEU- WDSKBQA	69.3%
Yiwise-KBQA	63.6%
NDers	62.9%
XJBot	62.9%
本文方法（基于 ChatGLM2-6B[59]阅读器）	71.2%
本文方法（基于 GPT-4[48]阅读器）	74.0%

在无需额外训练的情况下，该方法在精确匹配得分（EM）上可以达到 71.2%（使用 ChatGLM2-6B 作为阅读器）和 74.0%（使用 GPT4 API 作为阅读器）。分别优于竞赛最优方法 1.9%至 4.7%。实验结果表明，在将结构化的知识图谱转化为非结构化文档之后，问答系统仍能有效地利用这些知识，且其性能显著优于基线方法，验证了本文方案技术可行性与先进性。

（2）文本对组实验结果与分析

文本对组结果如表 3-5 所示。

表 3-5 实验结果-文本对组

方法/系统名称	准确率
Siamese-CNN	75.6%
Siamese-LSTM	83.9%
Siamese-BiLSTM	84.7%
QSTransformer	90.2%
epidemic-sentence-pair	96.4%
本文方法	91.4%

在文本相似度判断任务中，本文方法与赛事第一的方案 epidemic-sentence-pair 方法存在差距，但相比该方案，本文方法无需重新训练，具有更强的泛化性，而 epidemic-sentence-pair 模型不仅需要使用原有数据集训练，还需引入外部数据集 CHIP2019 数据集联合训练，同时还根据训练集数据特征作了数据增强。另一方面，本文方法仅使用单个嵌入模型，模型更简单，总参数量更小。

此外，如表 3-5 所示，本文方法优于其他常用于文本相似度判定的基线方法，实验证明了本文方法在处理文本对问答任务上的实用性。

3.3 本章小结

针对企业客服知识多源异构的特点，本章提出了一种多源异构知识融合方法。通过自动化将结构化知识、文本对和自由文本知识统一融合为的自由文本形式，进而通过基于自由文本知识的自动问答的方法进行处理，实现了多种知识结构的有效利用。本文方法的设计简洁高效，无需进行额外的训练。本章分别在 NLPCC 2018 KBQA 与阿里天池公益之星问题相似度挑战大赛数据集上验证本文方法的有效性，实验结果表明，本文方法不仅可有效处理多源异构知识，且性能优于传统基于结构化知识的自动问答方法。

第四章 基于大模型的提示框架设计及多轮对话实现

4.1 基于大模型的提示框架设计

4.1.1 提示工程基本原理

提示工程是当前 NLP 领域热点研究内容之一，由于其研究尚处于起步阶段，目前缺乏统一标准的定义。文献[66]提出提示工程的形式化定义，对于给定数据集 D 和用于特定任务大模型M_{task}，找到提示词p^*，使得模型输出达到最佳的表现，如下式所示：

$$p^* = \arg\max_{p} \sum_{(x,y)\in D_{dev}} f\left((x;\ p), y\right) \quad (4\text{-}1)$$

本文将提示工程过程总体描述为如下步骤。

（1）初始提示词添加

初始提示词添加方式取决于后续第二步模板搜索的方法，采用人工搜索的方式根据任务需求，通过自然语言描述任务的方式撰写初始的提示词，同时可添加一定数量的样例以提示模型如何输出。根据样例的数量提示词可分为零样例提示（zero-shot prompt）、单样例提示（one-shot prompt）和少样例提示（few-shot prompt）。选择提示样例的数量取决于任务的复杂性和模型的能力[67] [68]。然后，添加若干个待填充的槽位[X_i]和[Z_j]。其中，用户输入内容使用[X_i]表示，系统定义、环境变量由[Z_j]表示。例如，对于文本分类任务，定义[X]为用户输入的一段文本，[Z]为分类标签，则初始的提示词可以写成：“请根据标签[Z]对输入内容[X]进行分类。”

而若采用机器学习类的方法自动搜索提示词模板，则无需用自然语言表述需求，只需使用间隔符将输入按特定顺序分隔后拼接，形如：“[X]## [Z]”。

（2）模板搜索

提示工程的搜索目标是寻找最优的提示函数。函数搜索方式分为两类，一类为人工不断修改自然语言的方式，人工的方法较直观，但依赖于人工经验和技巧，即使是经验丰富的提示工程师也很难搜索到最佳的模板。另一类方式是采用自动搜索的方法，自动搜索的方法学习到的并非直接的提示词，而是中间的向量或模型前缀参数表示，更为抽象。该类方法包括 prefix-tuning[54]、prompt-tuning[69]和 P-tuning[55]。其中，prefix-tuning 通过优化特定任务的小型连续前缀向量，在保持语言模型参数冻结的情况下，将显性提示词转换为虚拟标记。prompt-tuning 是在 prefix-tuning 的基础上进行了简化，其只允许每个下游任务在输入文本前添加额外

的 k 个可调标记。P-tuning 的思想与 prompt-tuning 相似，而 P-tuning v2 则在 P-tuning 上进行优化，其最显著的改进是对预训练模型的每一层均加入可调参数。并通过多种设置弥补了与全量训练之间的差距，特别是对于小型模型和困难任务，P-tuning v2 可以在仅微调 0.1%~3%的参数的情况下，实现与全量微调相当的效果[55]。这类基于提示微调的方法并不引入人类可读的显性提示，称之为软提示。软提示中没有显性的自然语言，导致解释性不足，且可能降低模型的泛化性。提示微调原理的主要方法思想如图 4-1 所示。

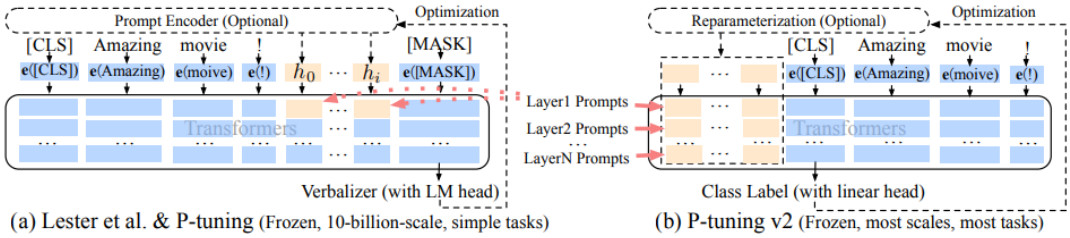

图 4-1 提示微调的原理[55]：P-tuning 与 Prompt Tuning 方法与之类似，加了可微的虚拟标记，但是仅限于输入层；P-tuning v2 在 transformer 块的每层输入前加入可微调的参数

Liu[70]等人观察到了利用基于提示的工具可以更好地利用大模型的能力进行下游任务的趋势，总结出“提示框架”（PF）的概念，即提示框架是使大模型能够与外部世界进行交互的上层。提示框架管理、简化和促进这种交互，帮助大模型克服数据滞后等挑战。提示框架自下而上可以分成数据层、基础层、运行层和服务层。数据层是基础层，作为与外部环境直接的接口。数据层主要处理诸如数据传输和预处理等任务，同时负责管理与外部数据源的交互。基础层位于数据层和执行层之间的计算中枢，负责大模型的管理，作为计算和控制中心，涉及接收和理解指令、执行命令以及进行各种计算，支持知识管理和决策过程。执行层构成业务逻辑的核心，负责与各 LLM 交互以完成特定的实际任务。

4.1.2 提示框架设计

基于提示工程的思想，在 Liu 等人[70]提出的框架基础上，本文针对客服自动问答系统中涵盖语言理解、语言生成等多个自然语言处理相关的子任务的特点，提出基于大模型的提示框架设计，以便在多个不同的子任务中使用同一框架，灵活地调用大模型，简化模型设计。区别于 Liu 等人将提示框架定义为一种抽象化的方法

论，本文提出的提示框架是一个具体的多任务处理模型。本文提出的提示框架如图4-2所示，其基本结构包括输入层、决策层、模型层和输出层。

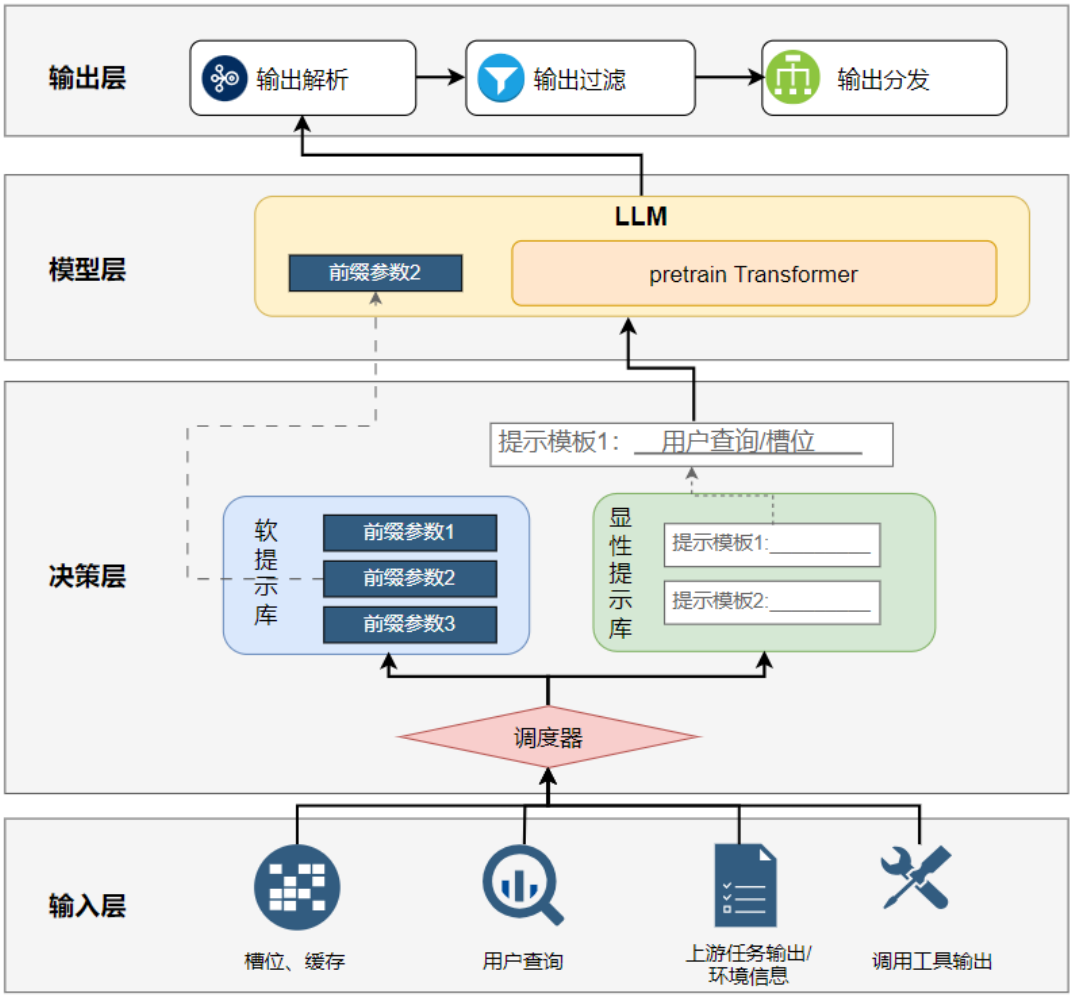

图 4-2 本文的提示框架结构

（1）输入层：用于定义初始环境输入，包括槽位与缓存、用户输入、上游任务输出结果和外接工具的输出等。

（2）决策层：决策层是提示框架的核心，根据输入层的输入选择合适的下游任务提示，以引导大模型根据输入与任务提示输出相应的内容。决策层由调度器与提示库两部分组成。其中，调度器根据当前任务从提示库中选择相应的提示模板或微调参数作为任务适配器。调度器可以设计成预设的程序、智能代理，或者嵌套另一个提示框架模型实例，本文中采用预设的程序形式。提示库包含预设提示词模板的显性提示库和包含微调参数的软提示库。根据不同的训练参数和下游任务需求，决

策层可以选择分别或同时从显性提示库及软提示库加载模板或参数。该设计使得决策层能够灵活地适应不同的任务需求，提高框架的整体性能和效率。

（3）模型层：模型层接受经过决策层处理后的显性提示输入或软提示参数，负责实际内容生成。模型层由原始的预训练大模型组成，模型大部分参数被冻结，仅在决策层输入中包含软提示参数时，才会增加少量的前缀参数，该设计保证了模型在处理各种任务时的鲁棒性和高效性。模型层根据任务要求生成原始输出，由输出层进一步处理。

（4）输出层：负责对模型层的输出进行必要的后处理，确保输出结果的质量，生成可被后续任务利用的形式。输出层包含输出解析、过滤、分发等子模块。

通过本文的提示框架设计，原始输入、提示库以及大模型本体被划分为模块化单元，同时提示库的形式被标准化。这种模块化设计意味着只需修改提示库的内容，就能适配多种下游任务，主要具有以下优势：

（1）减少开发工作量。高度模块化设计，使得多种不同任务在一个算法框架内集成运行，提高了代码的复用率。

（2）优化项目周期。本文提示框架采用并行策略，具有既能利用显性提示无需训练、快速实现的特点，又能利用软提示优良的性能。在传统数据驱动的系统建设中，数据准备和清洗占据了大量时间，延长项目上线周期。通过提示框架，企业可以在数据质量较低、标注数据不足的情况下，利用显性提示方法快速构建系统。随着系统上线后数据积累，可逐步切换到性能更高的软提示库，进一步提升系统性能。

4.2 基于提示框架的多轮对话设计

本节基于 4.1 节的提示框架模型，提出多轮对话的实现方案。该多轮对话采用流水线型架构，包含自然语言理解、对话管理及自然语言生成三个模块，如图 4-3 所示。其中自然语言理解中的意图识别、槽位填充子模块及自然语言生成模块均为提示框架的实例。

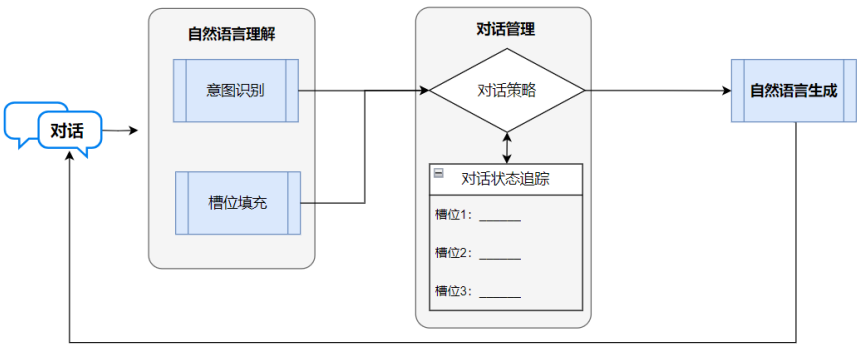

图 4-3 基于提示框架的多轮对话框架

4.2.1 自然语言理解模块设计

自然语言理解模块的核心任务是理解用户的意图，并从用户输入中提取实现该意图所需的信息，这些信息通常被称为“槽位”。该模块主要包括两个子模块：意图识别和槽位填充。

4.2.1.1 意图识别模块设计

意图识别包含输入意图识别和话题识别两个任务。用户输入新的内容后，首先进行输入意图识别，评估用户意图是开启一个新话题或是延续此前对话，以应对用户在自动问答系统中临时改变意图的情形。如果识别为新话题，则进一步识别具体的话题类型。

（1）输入意图识别方法

输入意图采用提示框架实现。在提示框架中，不同任务中软性提示的微调、加载方法基本一致，因此本节涉及提示框架实例均仅介绍其基于显性提示的实现。输入意图识别的提示模板如图 4-4 所示。模板中的[X]为格式化的历史对话记录。当用户输入新的内容时，系统按如图 4-4 的形式将新输入内容与l_h条历史对话记录拼接，其中l_h是人工设置的超参数。l_h越长，历史对话记录附带的上下文信息越丰富，判别更准确，但也导致推理成本的增加、响应延迟，并且大模型对总体输入长度也在限制，因此将l_h设置为 3 至 5。

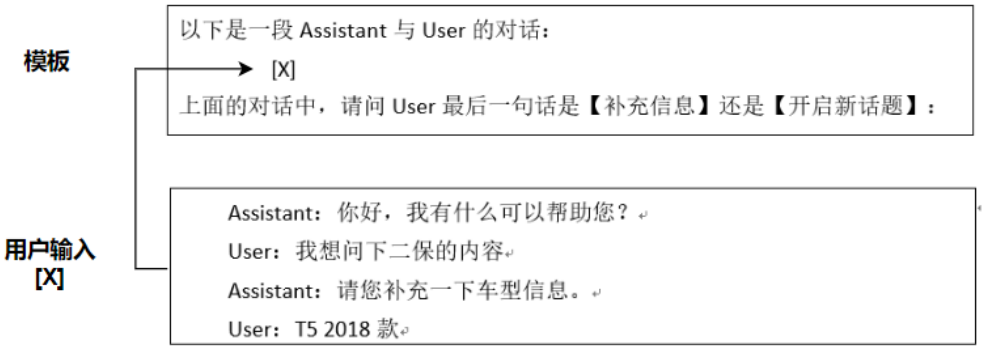

图 4-4 输入意图识别提示词模板

（2）话题识别

话题识别也采用提示框架方法构建，话题识别采用的提示词模板如图 4-5。图中[X]、[Z]为两个待填充内容，[X]表示用户键入的查询或拼接好的历史对话，[Z]为一个当前话题识别任务中的分类标签的值域。

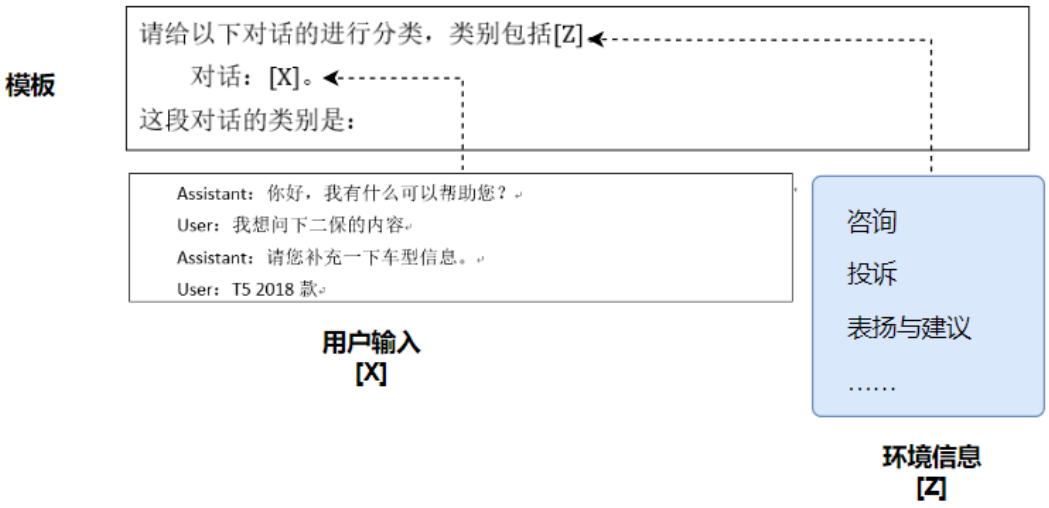

图 4-5 话题识别提示词模板

4.2.1.2 槽位填充模块设计

槽位填充模块的任务是从用户输入的信息中抽取出与预设语义槽相关的内容，填充到语义槽中。槽位填充模块也采用提示框架的方法。本文在研究过程中发现，直接使用显性提示方法进行槽位填充时，难以一次性成功抽取所有相关内容，并容易出现假阳性结果，即模型错误地从输入中抽取了不存在于预设槽位的内容。为了减少这一现象，本文设计了槽位检测环节，当特定槽位检测为阳性时，再对特定槽位内容进行抽取。

槽位检测用于识别用户输入的内容是否包含需要填充的槽位的内容，如图 4-6 所示。其中[X]为用户输入的句子，$[\hat{Z}]$ 为当前需要填充的槽位名称，如“车型”“年款”“零部件”等，[Z]为该槽位的取值空间，该取值空间用于指示模型的输出范围。当该取值空间巨大而不便于列举时，则直接依赖大模型的通用语言理解能力抽取，不在模板中约束取值空间。例如，在抽取“零部件名称”语义槽时，穷举零部件名称的取值空间不具有可操作性，故依托于大模型原始能力直接抽取出零部件名称。

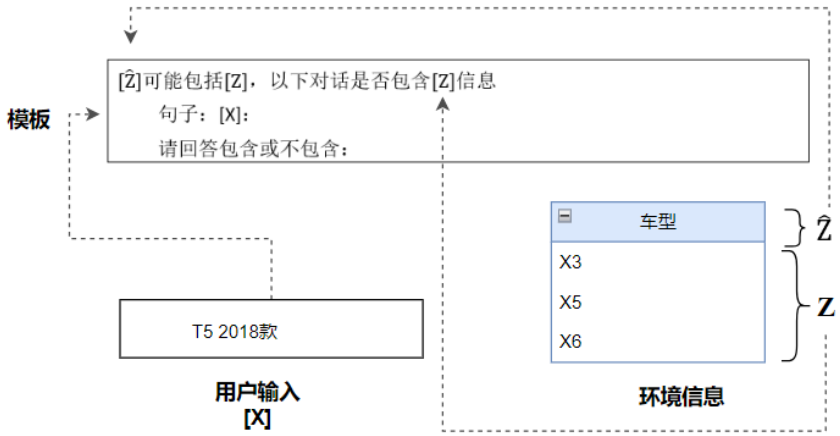

图 4-6 槽位检测

槽位检测结果通过后再进行槽位抽取和填充，抽取提示词模板如图 4-7 所示，其中[X]、[Z]与$[\hat{Z}]$与槽位检测中定义相同：

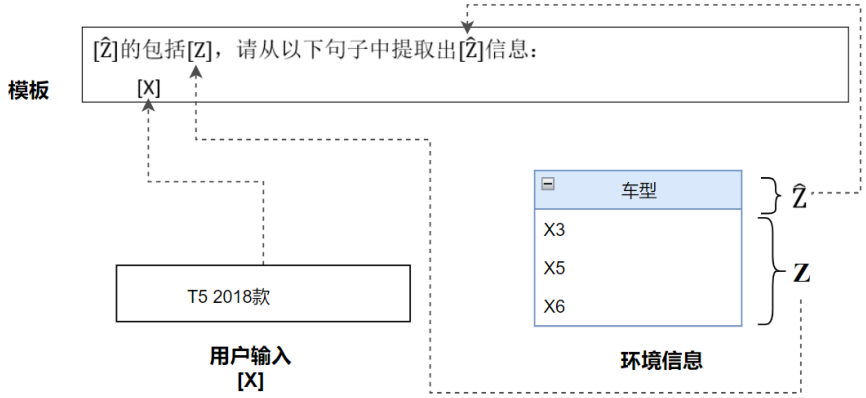

图 4-7 槽位填充

通过上述槽位填充的提示构造过程可见，使用显性提示方法存在两个主要缺陷：

（1）需要反复调用大模型。在填充 N 个槽位的任务中，需要调用多达 2N 次大模型，算法复杂度高，不适合需要填充多个槽位的情况；

（2）不适用于要求语义槽值具有精确格式或取值范围要求的任务。因为该情形下需要在提示词中枚举槽位的取值空间，以指示模型输出。导致提示词冗长，从而增加模型的计算负担。

综上，显性提示方法更适合于涉及少量槽位且对填充内容无特殊格式要求的槽位填充任务。而软提示方法在单次大模型调用中即可完成所有槽位的抽取，更适

合槽位填充任务，但其效果依赖于训练的数据量。本文研究的 XX 汽车制造企业的客服自动问答系统中，需要的语义槽主要包括车型、零部件、行驶里程、购车年份、经销商名称、用户所在城市等信息，信息主要用于检索知识库，因此不需要精确的格式和内容，显性提示的方式可满足需求。

4.2.2 对话管理模块设计

本文提出的对话策略为人工定义的程序。以咨询类意图为例进行阐述，整体对话策略流程如图 4-8 所示，主要包括：输入意图判断、槽位管理、构造提问、问句改写和构造答案流程。其他意图流程基本相似，仅在需填充的槽位与最终输出目标不一致。

（1）输入意图判断

首先拼接用户输入和对话历史，通过输入意图识别模块判断用户意图。若判断为开启新话题，则调用分类模块对话题意图进一步分类，并按照分类结果继续对话。若判断为补充信息，则更新对话状态，并追踪轮询槽位内容。

（2）槽位管理

当首次进入新话题时，首先初始化语义槽，该语义槽包含该话题任务下必要的前置信息。通过查询历史槽位，获取相同的槽位名，取同名槽位的槽值继承作为初始值，如历史槽位信息不存在，则初始值为空。然后检测并抽取用户当前输入的内容是否包含相关语义槽信息。本轮抽取完毕后，判断语义槽是否填充完毕。

（3）构造提问

若槽位未填充完毕，系统将利用模板或大模型构造问句，对缺失的槽位进行追问。该过程会持续进行至所有槽位都被填充，或者对任何一个槽位的连续追问达到预设的上限。如果达到追问上限用户仍未提供或系统无法提取，那么未能获取信息的槽位将被标记为“未提供”。引入追问上限的目的是避免系统与用户之间陷入无休止的追问循环，从而提高对话的流畅性和用户体验。

（4）问句改写

问句改写的目的是生成一个更加精确和完整的查询，以便于后续的模块处理。系统将用户输入的原始问题以及所有槽位的键和值信息输入到大模型，大模型槽位信息融入原始问句中，形成包含完整信息的问句。

（5）构造答案

返回大模型生成的答案与参考文档，形成答案。通过设定参数 R，可设定快速回答、准确回答两种回复策略。当设定 R=1 时，采用快速回答策略，即用户槽位未填充完毕时也根据当前已满足槽位信息检索并返回答案，以实现简单问题快速

响应；让 R=0 时，采用准确回答策略，该策略下仅当槽位全部满足时检索并返回答案。

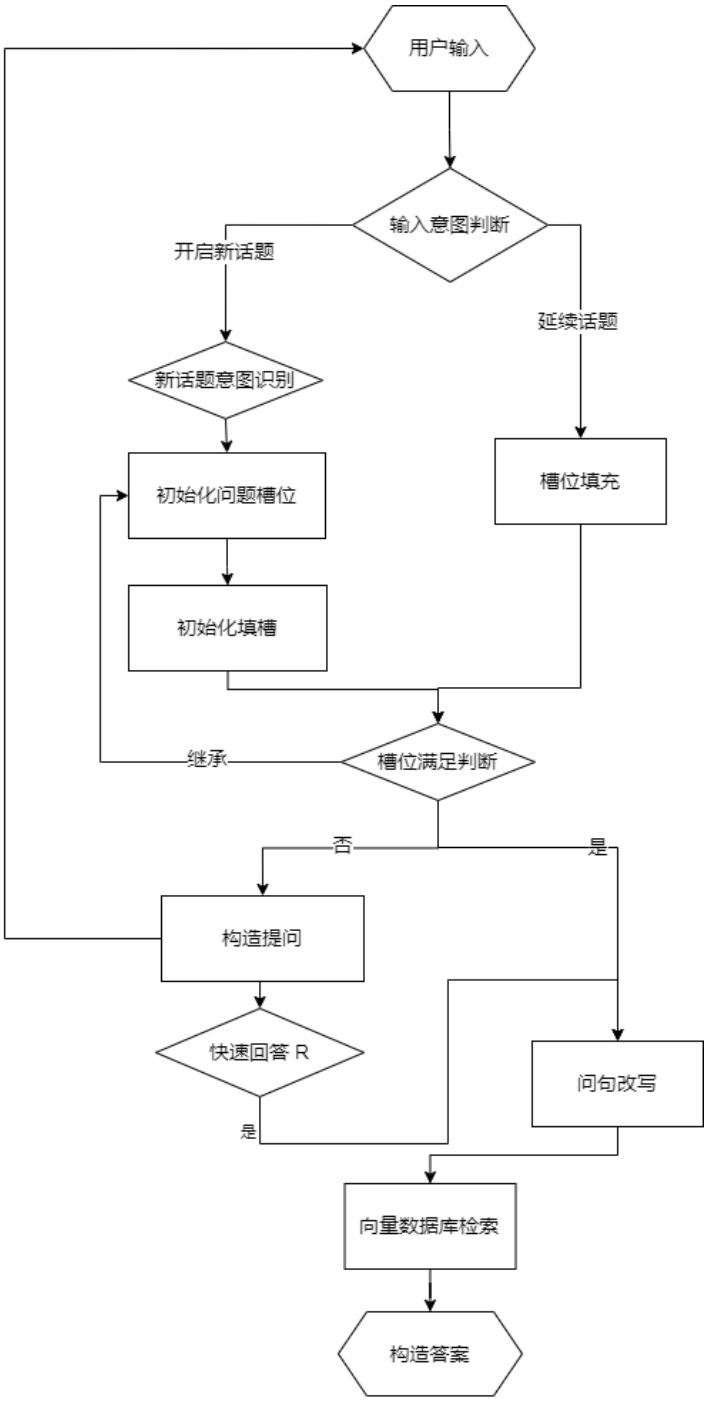

图 4-8 对话策略流程

4.2.3 自然语言生成模块设计

自然语言生成（NLG）在多轮对话模块中的主要作用是构建针对用户的回复。生成任务包括闲聊、构造提问和构造答案三类，其作用及流程如下。

（1）闲聊：用于回应用户的无特定目的性的输入，如闲聊、问候或致谢等。该部分无需引入额外的提示或引导，由大模型生成相应的闲聊回复。

（2）构造提问：该部分用于在当前话题所需的语义槽不能满足时，模块将针对未满足的槽位构造问题。

（3）构造答案：若当前话题所需的语义槽满足，模块将根据构造答案来回应用户的问题。

4.3 实验验证及分析

为验证基于提示框架的多轮对话方法的可用性，本节将通过实验进行验证。

4.3.1 数据集及评价指标

（1）评价范围

对任务型对话系统进行整体性评价常采用基于人工的评价方法、用户模拟的方法和混合方法，受多种因素影响，评价结果可能无法完全匹配用户体验感受[71]。即使人工制定的评价指标也可能引入偏差，使得现有的评价过程往往难以准确满足用户的要求[72]。因此学者们通常对任务型多轮对话的各个模块进行单独评价。即单独对自然语言理解、对话策略管理和自然语言生成进行单独评价。其中，对自然语言理解的评价又进一步细分对意图识别与槽位填充两个子任务进行评估；而对话策略管理通常不单独评价，而是体现在整个任务型对话性能的指标中；自然语言生成的评价则通常使用 BLEU[73]评分来评估生成文本与参考文本之间的相似性，BLEU 是一种常用的自动评估指标，用于衡量生成文本的质量。

综上，多轮对话性能的评价可通过意图识别、槽位填充、自然语言生成三个子任务进行评估。由于本章使用的自然语言生成模块是基于大模型的提示框架，BLEU 为模型的通用性能，与下游任务关联性不强，其性能指标在其技术报告[59,74]中已有详细论述，因此在本研究中将不再重复评估自然语言生成指标，仅评估意图识别和槽位填充两个子任务的性能。

（2）数据集

为使结果便于与现有方法比较，本文采用公开数据集 ATIS（航空公司旅行信息系统）进行实验。ATIS 是一个被广泛应用于意图识别与槽位填充任务的数据集，该数据集包括用户在自动化航空公司旅行查询系统上请求航班信息的音频记录和相应的手动转录。数据包含 17 个唯一的意图类别。原始数据分为训练集、开发集和测试集，分别包含 4478、500 和 893 个带有意图标签的语句。其意图相对本文中的汽车客服系统数量更多、意图混淆度更大，如标签中有 Flight、Flight + Airline、

Flight Number、Flight Number + Airline 等存在包含、交叉、组合关系的意图。槽位填充方面，ATIS 共有 129 个槽位（含无槽位信息的零标识槽位）。意图复杂性与槽位数量均大于本文研究的面向汽车客服的自动问答系统，挑战性更大。

（3）评价指标

针对意图识别任务，采用意图识别准确率（accuracy）作为指标，对于槽位填充任务，采用 F1 分数作为指标。

4.3.2 实验环境及参数设置

4.3.2.1 实验环境

实验环境如表 4-1 所示。

表 4-1 意图识别及槽位填充实验环境

配置	参数
CPU	12 vCPU Intel(R) Xeon(R) Platinum 8352V CPU @ 2.10GHz
GPU	RTX 4090(24GB) ×1
内存	90GB
系统	Ubuntu 20.04.4
语言环境	python 3.8.10
CUDA 版本	11.8
LLM 型号	ChatGLM3-6B
模型参数量	60 亿

4.3.2.2 提示词设置

（1）意图识别

意图识别实验构造了零样例提示、少样例提示、不包含显性任务信息的软提示和包含显性任务信息的软提示等四种提示词构造方法，软提示均采用 P-Tuning v2 方式训练。各提示词根据数据集为英文的特点，将提示词适应性修改为英文，零样例提示如图 4-9 所示，模型输出的温度参数为 0.1，旨在输出相对确定的结果；少样例提示是在零样例提示的基础上增加了三条从训练集抽取的样本作为样例，如图 4-10 所示，模型输出温度参数与零样例提示设置一致；不包含任务信息的软提示构造如图 4-11 所示，训练和推理提示词中均不包含显性描述任务的内容，仅包含用户输入和标签；包含显性任务信息的软提示则在训练和测试样本中均加入了显性任务的说明，如图 4-12 所示。

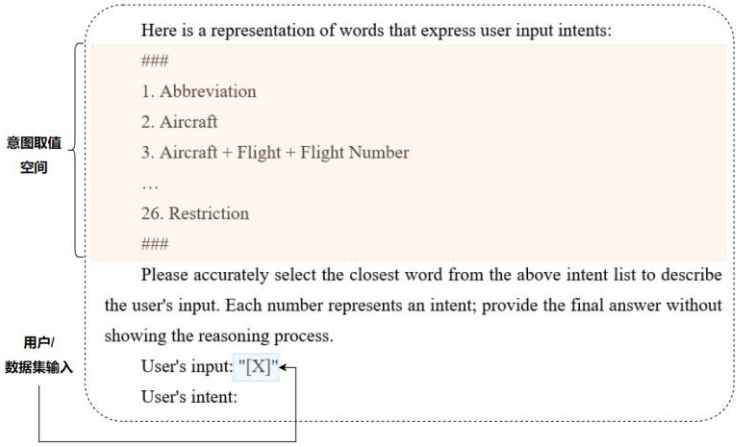

图 4-9 零样例提示构造情况

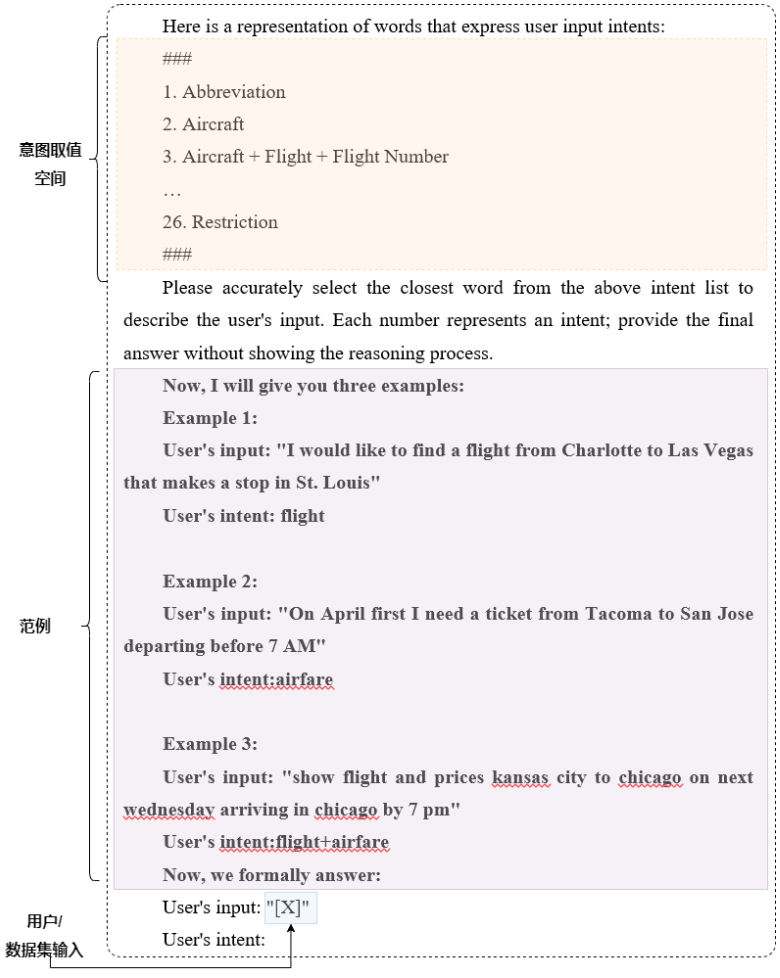

图 4-10 少样例提示构造情况

图 4-11 不含任务信息的软提示构造情况

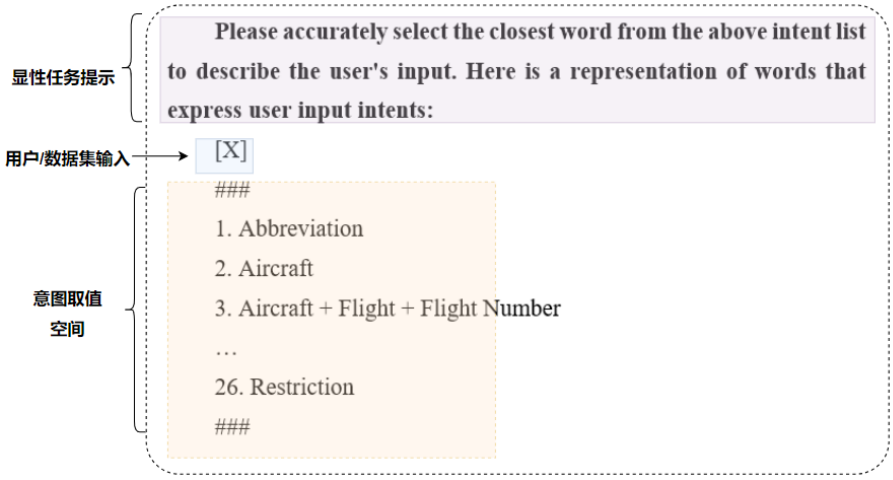

图 4-12 包含任务信息的软提示构造情况

软提示的训练参数如表 4-2 所示。

表 4-2 意图识别任务的软提示训练参数

训练参数	参数值
前缀长度 PRE_SEQ_LEN	128
学习率 LR	0.02
使用 GPU 数量 NUM_GPUS	1
最大输入长度 MAX_SOURCE_LEN	256
最大目标输出长度 MAX_TARGET_LEN	128
梯度累积步数 GRAD_ACCUMULARION_STEPS	16
最大学习步数 MAX_STEP	2000

（2）槽位填充

由于 ATIS 数据集槽位标记为 129 个，共计 2318 个字符，如采用显性的提示，仅罗列槽位值就将占用 1108 个标记（token），这对模型的最大输入能力提出了要求，同时标记的数量直接影响推理速度，继而影响系统响应。因此在该数据集上仅采用无显性提示词软提示，其训练参数见表 4-3 所示。

表 4-3 槽位填充任务的软提示训练参数

训练参数	参数值
前缀长度 PRE_SEQ_LEN	192
学习率 LR	0.03
使用 GPU 数量 NUM_GPUS	1
最大输入长度 MAX_SOURCE_LEN	256
最大目标输出长度 MAX_TARGET_LEN	256
梯度累积步数 GRAD_ACCUMULARION_STEPS	16
最大学习步数 MAX_STEP	7500

基线选用代表性的方法如下：

RNN+LSTM[34]：使用双向循环网络与长短期记忆的方法；

Attention BiRNN [35]：循环神经网络与注意力结合的方法；

Slot-Gated [75]：在注意力中增加槽门机制的方法；

Joint BERT[76]：基于预训练语言模型，采用意图识别与槽位填充联合建模的方法。

4.3.3 实验结果及分析

意图识别微调 38 分钟，槽位填充任务微调约 1 小时 30 分，最终训练损失函数 loss<0.02，结果如表 4-4 所示。

表 4-4 意图识别及槽位填充实验结果

方法		意图识别（acc）	槽位填充（F1）
双向 RNN+LSTM		92.6%	94.3
Attention BiRNN		94.2%	91.1
Slot-Gated		94.1%	95.2
Joint BERT		97.4%	96.1
本文方法	零样例显性提示	85.9%	
	少样例显性提示	93.0%	
	软提示	97.4%	94.8
	零样例显性提示+软提示	97.4%	

（1）意图识别结果分析

本文所提出的方法中，通过提示微调的方式构造软提示，意图识别的准确率优于多数基线水准，与 Joint BERT 方法相当。尽管人工构建的显性提示方法性能未达最优，但在合适的构建方法下，仍可达到 93.0%的较好水平。因此，在考虑模型迁移性、灵活性和标注数据缺乏的情况下，人工构建提示方法具有可行性。

（2）人工提示构造方式对结果影响分析

少样例提示在复杂意图分类任务中优于零样例提示，体现在意图分类准确性和生成答案格式上。如表 4-5 中 ID264 的样本所示，零样例提示在明确回答要求下仍倾向完整表述，需增加后处理环节，而少样例提示方法则直接给出答案，造成上述现象的原因可能为基座模型经 RLHF 对齐，倾向于完整友好回答，给出样例示意后能够缓解该情形；两类人工构造提示方法共同的问题为枚举过多意图，如表 4-5 中 ID 为 121 的样例，此问题可通过基于样例特征工程二次清洗，但增加后处理难度。经实验观察发现枚举的额外意图仍与原用户输入有一定相关性，表明模型陷入局部最优。

表 4-5 zero-shot 与 few-shot 典型负例分析

样本 ID	Query 及正确答案	零样例提示输出	少样例提示输出
264	on april first i need a ticket from tacoma to san jose departing before 7 am'（Airfare）	Thought: The closest word to the user\'s intent is "Airfare".\n\nAnswer: Airfare	Flight + Airfare
121	show me all the flights from charlotte to minneapolis on tuesday morning（Flight）	Flight + Airline + Flight Number + Time + Location	Ground + Fare

（3）软提示构造形式对结果影响分析

根据实验结果，在构造用于软提示的提示词时，添加显性的任务信息与否对模型最终收敛结果没有影响，但添加任务信息可使训练过程更快收敛。从理论上分析，通过提供与微调目标相关的信息，减少了训练过程中需要搜索的解空间，从而加速收敛过程。

（4）槽位填充任务结果分析

从实验可知基于软提示的槽位填充 F1 值达到 94.8，尽管低于最优基线 Joint-Bert，但本文方法仍优于基线中各类基于 RNN 的方法。实验验证了本章提出方法在槽位抽取任务上的可行性。

4.4 本章小结

本章提出了一种基于大型语言模型的提示框架，该框架根据下游任务需求动态加载显性提示词和软提示参数，以实现将大模型的通用能力迁移至特定领域的任务。针对自动问答系统中的多轮对话需求，本章在该框架基础上提出了多轮对话的实现策略。并在公开数据集上对意图识别和槽位填充任务进行验证。实验结果表明，通过合理构造提示词，显性提示可以降低系统或模型领域迁移的难度，同时达到较好的性能；引入软提示可进一步提升模型性能，使其接近最优基线。实验结果验证了本章提出的多轮对话方法的有效性和实用性。

第五章 基于大模型检索增强生成的知识问答模块设计

基于自由文本的自动知识问答系统常采用“检索器-阅读器”的架构。其中，检索器负责接收用户的问题并检索相关文档，阅读器则从这些文档中抽取或生成答案。检索增强生成（RAG）技术同样包含检索和生成两个阶段，其架构与非结构化知识问答系统类似。本章将对基于检索增强生成技术的知识问答模块进行研究设计，该模块将传统的阅读器替换为生成式大模型。并研究影响该模块准确性与回答效率的关键因素，提高整个问答系统的准确性和响应速度。

5.1 密集向量检索器设计

密集向量检索器的基本原理是将文本或图像映射到向量空间后，在该空间中测量用户查询与文档相似性。密集向量检索器的工作流程主要包括密集嵌入阶段和向量检索阶段。

5.1.1 密集嵌入

密集嵌入（dense embeddings）是一种将单词、句子等高维的离散数据映射到低维连续向量空间的技术，相比于传统的独热（one-hot）编码，其表示更为紧凑，因为每个元素都映射为一个连续的实数向量，维数远小于整个词表的数量。如图 5-1 所示，以“车机”一词为例说明独热编码与密集嵌入的编码形式的区别。

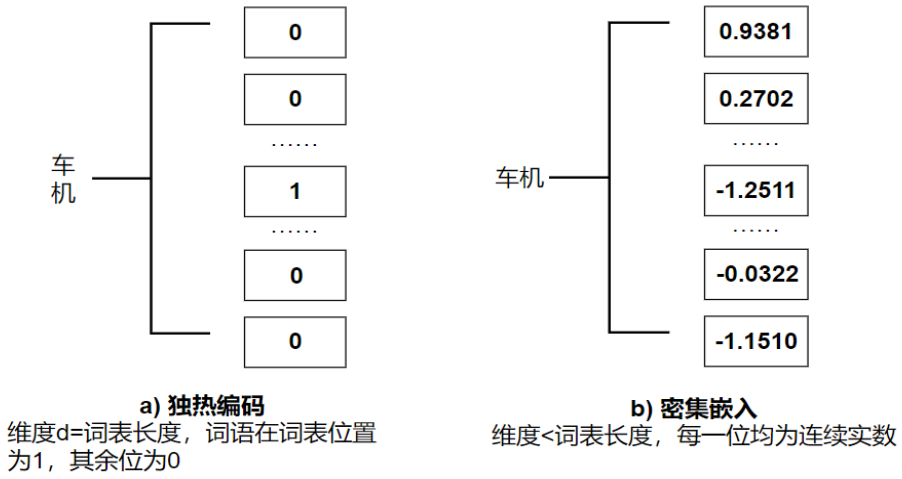

图 5-1 密集嵌入与传统独热编码对比

如图中所示，独热编码将词汇表中的每个词均表示为一个向量，其中只有一个元素为 1，其余为 0，而密集嵌入则将每个词映射为一个连续的实数向量。因此密集嵌入的向量不仅包含了词汇的表示，还隐含了更丰富的语义信息。语义上相似的单词或句子在密集嵌入后通常在向量相似度上也是更为接近的，因此使用密集嵌入能够更深层次地理解语义，有助于从词义、句意层面理解用户查询，提高问答系统检索准确率。

5.1.2 向量检索

向量检索的原理是根据问题向量与知识库向量计算相似度，并返回向量相似度最高的若干条文档。常用的向量相似度计算方法有欧氏距离、曼哈顿距离、向量点积、余弦相似度等。本文采用问题向量与文档向量内积的方式计算相似度，如式 5-1：

$$\mathrm{sim}(x, \mathrm{y}_i) = \mathrm{x} \cdot \mathrm{y}_i \tag{5-1}$$

其中 x 表示查询向量，y_i表示向量数据库中第i个向量，通过对比向量间点积大小计算不同y_i与 x 的相似度。

为提高知识问答模块的检索效率，本文采用 FAISS（Facebook AI Similarity Search）库作为密集向量检索工具[61]。该工具提供了平坦索引（Flat Indexing）、IVF（Inverted File）和 HNSW（Hierarchical Navigable Small World）几种检索方法。其中平坦索引是 FAISS 中最简单的索引方法。它将所有向量放在同一层级，通过计算向量之间的相似性来进行检索，其计算向量相似度采用点积的方法，适用于小规模数据集。IVF 是一种基于倒排索引的方法[77]。它将向量空间划分为多个子空间，每个子空间构建一个倒排索引。当进行检索时，首先在子空间中搜索，然后合并结果。 HNSW 是基于图的检索方法，适用于大规模高维数据，其构建了一个具有导航能力的图结构，使得在图上的局部搜索可以快速找到相似的向量。由于本文研究的汽车客服自动问答系统知识库的数据规模相对大规模开放域知识库较小，故采用平坦索引的方式进行检索，平坦索引的实现简单，无需复杂的层次结构，即可快速有效地实现向量检索。

5.2 密集向量检索器性能验证及分析

本节研究将通过实验验证检索器的检索性能，并研究问答系统设计时，召回文档数量 top-k 的设计依据。

5.2.1 实验数据集及评价指标

（1）数据集

本文采用课题依托单位 XX 汽车制造企业私有的汽车客服问答测试集作为本部分的数据集，该数据集包含一个文本知识库和一个测试集。测试集由 512 个问题组成，其中约 25%的问题来自真实客户咨询，经过 GPT-4 进行改写和人工审核修改，以避免与知识库中已有的文档内容重叠。其余 75%问题则是基于原始知识文档，借助 GPT-4 生成的新问题。该举措旨在增加问题的文本多样性，从而验证模型处理各类问题的能力。为了便于引用和讨论，本文将该数据集简称为 AUTO 数据集。

由于 AUTO 数据集涉及企业内部不便公开的私有数据，为使结果可复现，本章实验还引入 NLPCC-2018-KBQA-TEXT 数据集（为方便叙述，简称 KBText），该数据集的知识库源自于 NLPCC 2018 KBQA 评测任务提供的原始知识库，经过本文第三章中所描述的非结构化处理方法处理后，生成的非结构化文本库。包含 4794 万个句子组成的文本片段。测试集借用 NLPCC 2018 KBQA 测试集的 618 个问题。

（2）评价指标

本节采用平均倒数排名 MRR（Mean Reciprocal Rank）衡量检索器性能，该指标常用于评估信息检索系统的性能：

$$\mathrm{MRR} = \frac{1}{Q}\sum_{i=1}^{Q}\frac{1}{Rank_i} \qquad (5\text{-}2)$$

其中 Q 表示查询的问题总数，$Rank_i$表示的第i个问题在召回的文档结果中找到的第一个正确答案的排名。MRR 的值范围在 0 到 1 之间，数值越接近 1 说明检索系统在平均排名上的表现越好，即系统对查询问题给出的正确答案排名越靠前。MRR 关注检索系统返回的结果中第一个正确答案的排名，而不会考虑排名之后的文档情况。

5.2.2 实验环境及参数

实验参数如表 5-1 所示。

表 5-1 实验环境及参数设定

配置	参数
CPU	12 vCPU Intel(R) Xeon(R) Platinum 8352V CPU @ 2.10GHz
GPU	RTX 4090(24GB) ×1
内存	90GB
系统	Ubuntu 20.04.4
嵌入模型	BGE-base-zh
嵌入维度	768
检索方式	Flat Indexing

5.2.3 实验结果及分析

（1）整体检索性能

实验结果如表 5-2 所示。尽管 KBText 数据集检索空间庞大且测试问题复杂度高，检索器仍展现出了优良的性能，正确答案平均出现在召回文档的前两位（1/0.511≈2）。在 AUTO 数据集上，检索模块实现了更高的 MRR 值，达到了 0.939。表明在多数情况下，检索器召回的第一条文档即包含了问题的正确答案。实验结果表明，密集检索器不仅可应对大规模开放域知识，还能够有效地处理专业术语密集且私有化的汽车客服知识。

表 5-2 本文方法在 KBText 及 AUTO 数据集检索性能表现

数据集	MRR
KBText	0.511
AUTO	0.939

（2）召回文档数量（top-k）设计分析

如图 5-所示，随着 top-k 值的增加，KBText 和 AUTO 数据集实验中召回文档中包含正确答案的概率变化情况。

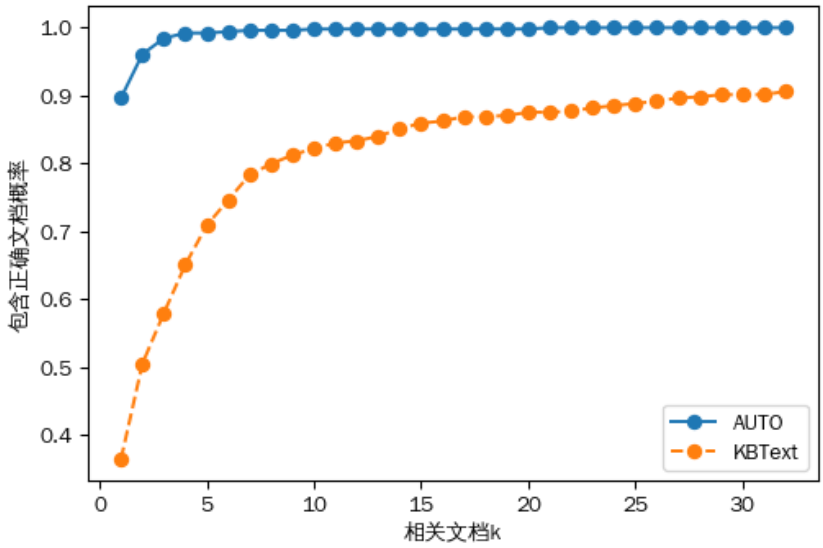

图 5-2 召回文档中包含正确答案的概率随 top-k 值变化

图中显示，在 KBText 数据集上，当 top-k 小于 8 时，召回文档包含正确答案概率随着 top-k 的增加而迅速上升，在 top-k=8 时达到 0.799。而当 top-k 超过 8 后，增长逐渐平稳。对于 AUTO 数据集，在 top-k=4 时，召回文档中包含正确文档的概率已超过 0.99，且在 top-k=21 时达到 1。继续增加 top-k 值对提升模型性能的收益极小，反而可能引入过多无关文档，并且文档数量增加直接导致输入下游模型的标记数量增加，继而增加大模型的推理开销。因此，在企业小规模数据集上，将 top-k 值设为 4 即可以较好地平衡性能与效率。

5.3 基于大模型的阅读器设计

阅读器主要任务是从检索器返回的文档集中，理解问题并提供最终答案。这一过程与检索增强生成的过程一致：检索增强生成的核心思想是利用检索结果来提升生成模型的性能，因此生成式阅读器成为近年来研究的热点。本文基于该思想，将阅读器按照本文第四章中的提示框架方法进行模块化设计，以简化开发流程。具体流程如图 5-3 所示。

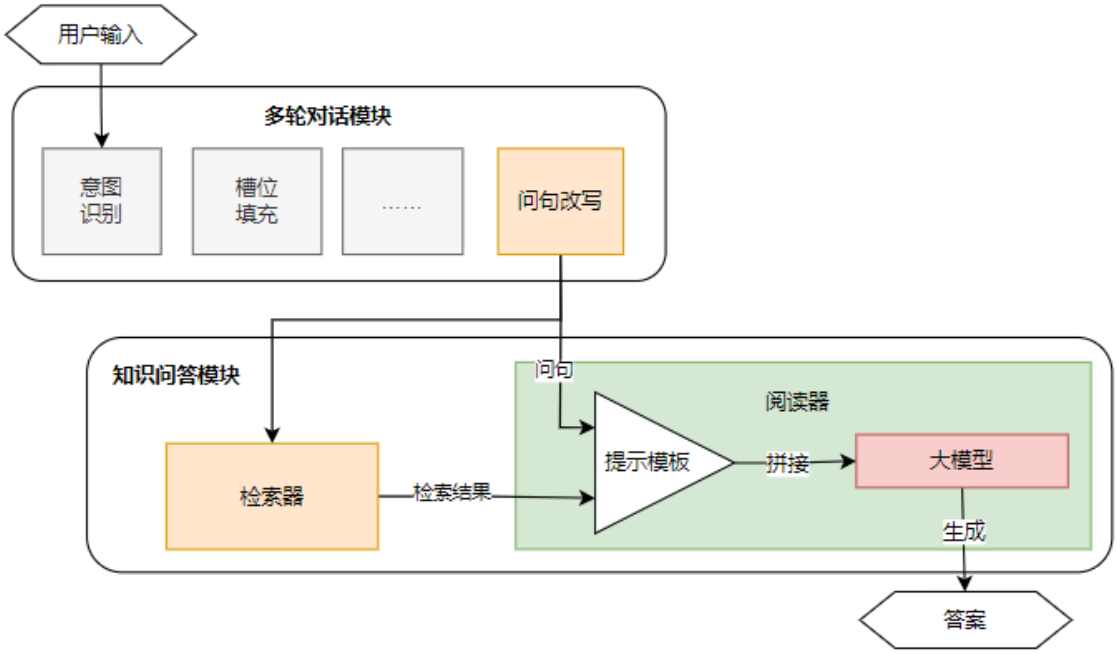

图 5-3 阅读器工作流程

首先，阅读器同时接收经过多轮对话模块改写后的问句和检索器检索结果作为输入，然后通过显性提示模板拼接成提示词，其中提示模板如图 5-4 所示，其中[X]为问句，[Z]为调用密集检索器对问题查询后返回的 top-k 篇文档内容。拼接后的提示输出至大模型，由大模型生成最终答案。

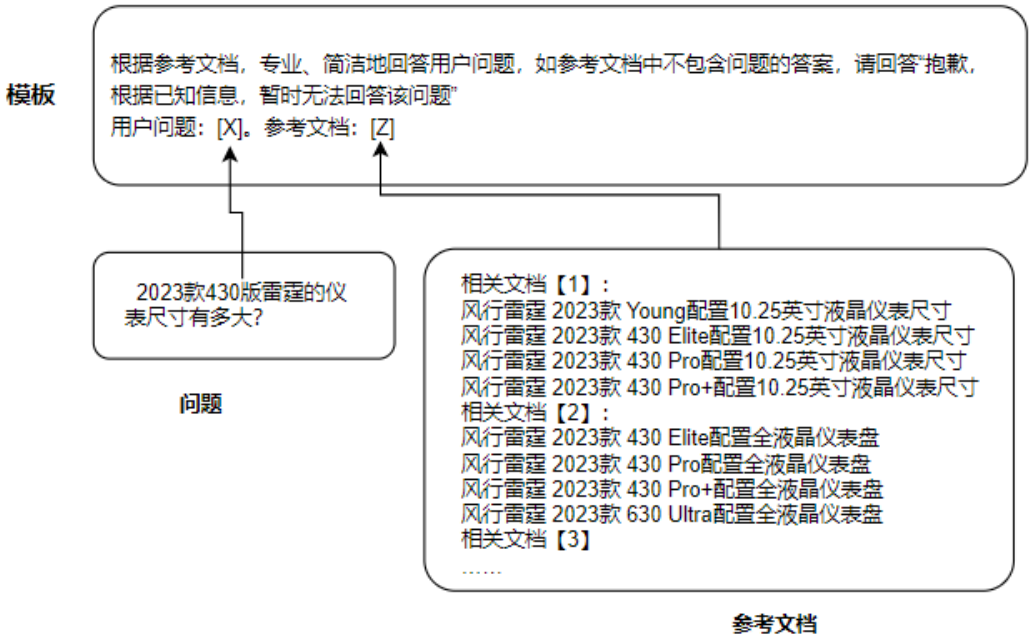

图 5-4 阅读理解任务提示模板

5.4 大模型阅读器性能验证及分析

本节将通过实验分析大模型生成式阅读器的性能，并与当前具有代表性的中文大模型任务上的表现进行对比，分析大模型的模型选型、参数量设定依据。

5.4.1 实验数据集及评价

（1）数据集

对 KBText 测试集按照本章 5.2 节实验的方法进行检索，对每个样本问题召回的前 8 条相关文档，并剔除了召回文档中未包含正确答案的样本。最终形成 491 组有效样本，作为本实验测试集。

（2）基线模型及评价指标

本节采用 FiD（Fusion in decoder）[78]作为基线模型，FiD 使用预训练的生成式模型 T5 进行训练，原始 T5 模型不包含中文。本实验中将基线模型的 T5 模型替换为结构完全相同、但在中文语料集上预训练的中文模型 MengziT5。同时将训练数据由英文 Squard 替换为中文翻译后 Squard 数据集。参照原 FiD 方法训练 50000 步，训练步数约为基线方法的 3 倍，以尽可能获得更优性能。将验证集上获得最佳性能的模型参数固定，用作对比实验的模型。

衡量阅读器的性能主要包括精确度匹配 EM、F1 分数等指标。在阅读理解任务中，F1 在字符级别对比标准回答与预测回答的交集。而本文所采用的生成式阅读器，在生成正确的答案的同时，可能改变答案句子的表述方式或添加答案相关的补充信息，这将导致生成的答案文本与原始答案文本有所差异，从而降低 F1 分数。因此 F1 分数更适用于评价抽取式阅读器的性能，而不适合用于评估生成式阅读器。因此，本章节仅采用精确度匹配 EM 来评估阅读器的性能。

5.4.2 实验设计

实验将检验大模型选型、参数量及提示词构造方式对阅读理解任务的影响，具体如下。

（1）对比不同的模型与基线方法，验证所提方法的有效性并选择合适的模型作为后续模块的基座。

本实验中评估多个大模型效果，具体如下。

ChatGLM[59]系列：GLM 由清华大学与智谱提出，在使用 Transformer 的基础上重新排列了层规范化和残差连接的顺序，使用单个线性层来对输出进行预测，并将激活函数 ReLU 用 GeLUs 替换。基于 GLM 结构，研究者相继开源了 ChatGLM-

6B、ChatGLM2 和 ChatGLM3，其中 ChatGLM2、ChatGLM3 改进了训练方式和训练数据，同时在推理速度和上下文长度上优化模型。

QWEN[60]：QWEN 系列由阿里巴巴开发，基于 LLaMA 的基础上改进并在万亿级别的词条的数据集上训练的大模型，使用监督微调（SFT）和人类反馈强化学习（RLHF），提高了语言模型参与自然对话的能力。

Baichuan[79]：由百川智能开发，属于仅解码器的结构，采用与 LLaMA 相同的模型设计，主要优化了分词器、位置编码和激活函数等。并使用万亿级 token 规模的高质量语料上训练。

ChatGPT：ChatGPT 是由 OpenAI 于 2022 年 11 月推出的基于 GPT-3.5 的人工智能聊天机器人程序。GPT-4 是相对于 ChatGPT 更先进的语言模型，拥有更多的参数和更复杂的架构，能够处理更长的文本，在语言理解、推理和生成文本方面表现出更高的能力。

本节实验中 ChatGLM、QWEN 及 Baichuan 等模型均采用本地配置的环境部署，旨在验证真实工程场景下的实际应用能力；ChatGPT 及 GPT-4 两个模型则将通过 API 方式调用，以研究当前基于大模型为基础的阅读器的所能达到的性能边界。由于人工提示词无法达到最优设计、GPT-4 也无法完全代表大模型的最佳性能，该性能边界非最优状态，仅为近似参考值。

（2）评估模型参数量对阅读器性能影响

通过对比 QWEN 模型的 18 亿、70 亿、140 亿和 720 亿参数的版本，评估模型参数量对阅读器性能的影响。并据此选择最合适的模型参数。

（3）评估不同提示词构造方式对模型性能的影响。

在 Baichuan-7B 与 ChatGLM-6B 模型上，采用将指令（I）、问题（Q）和文档（C）的按照不同排列顺序形成六种不同的提示方法构造提示词，如指令-问题-文档、指令-文档-问题、文档-问题-指令等，评估提示词构造对模型性能的影响。

5.4.3 实验结果及分析

（1）总体分析

实验结果总体结果如表 5-3 所示。实验表明，采用大模型作为阅读器显著提升了阅读器性能。其中使用 GPT-4 API 作为阅读器 EM 达到 93.27%，实现了最佳的性能；本地部署模型中表现最佳的模型为 QWEN -72B（91.24%）。从开源协议友好性、推理成本、数据安全角度与性能差距综合度量，本地的 QWEN-7B 和 ChatGLM2-6B 更具工程应用价值。

表 5-3 阅读理解总体结果

模型		参数量	EM
基线	FiD	1.1 亿	8.96%
本文方法	ChatGLM-6B	60 亿	82.38%
	ChatGLM2-6B	60 亿	89.61%
	ChatGLM3-6B	60 亿	86.35%
	Baichuan-7B	70 亿	76.64%
	QWEN-1.8B	18 亿	72.70%
	QWEN-7B	70 亿	87.98%
	QWEN-14B	140 亿	87.57%
	QWEN-72B	720 亿	91.24%
	GPT-3.5-Turbo API	未公布	89.66%
	GPT-4 API	未公布	93.27%

与基线比较及分析：尽管使用了更多的训练步数，FiD 性能（8.96%）仍然远低于本文使用的基于大模型的方法。这一结果也远低于其英文 NQ，SQuAD Open 数据集上 48.2%及 53.6%的性能表现。从参数量与模型结构上分析，首先，T5 模型仅 2.2 亿，已有研究表明，相同的模型对不同语言难度具有显著差异[80]，并且同一模型处理同样内容的文本，中文比英文也需要更多的标记[81]，导致更高的计算成本，进而影响了 FiD 在中文任务上的性能。同时，T5 模型 2.2 亿的参数也显著低于本实验中的各大模型。其次，结构上看，T5 为包含编码器和解码器的完整 Transformer 结构，对比的大模型则均采用仅解码器的结构，已有研究[82]表明在未针对下游任务进行微调的情况下，同等参数下仅解码器结构的结构更有优势，因为编码器的双向注意力会存在低秩问题，这可能会削弱模型表达能力，就生成任务而言，在编码器中引入双向注意力并无实质优势。

（2）模型参数量对阅读器性能影响分析

如图 5-5 所示，QWEN 模型在参数量从 18 亿、70 亿、140 亿增加到 700 亿的过程中，虽然阅读理解性能提升，然而，参数量的增长并未带来持续的性能改进。实验结果表明，在参数量达到 70 亿左右时，模型性能已达到相对理想的状态，将参数量增加至 720 亿，模型性能的提升幅度仅为 3.7%。同时，根据表 5-4 数据显示，本实验中 60 亿至 70 亿参数量的模型，如 ChatGLM2-6B、ChatGLM3-6B 等，与 QWEN-7B 性能表现相似。因此，70 亿左右的参数量对于自动问答系统中的阅读理解任务已经足够。在现有的大模型推理技术下，对于该参数量级的模型，通过 INT4 量化技术进行量化处理后，模型在推理过程中仅需要 6 至 7GB 的显存空间，实现了在计算资源利用和模型性能之间的平衡。

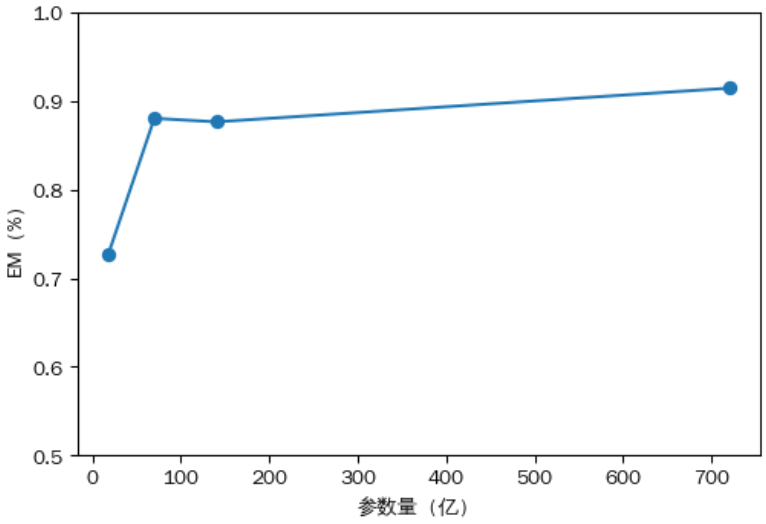

图 5-5 QWEN 模型参数量与阅读理解性能关系

表 5-4 部分 60~70 亿参数量中文大模型的阅读理解性能对比

模型	参数量	EM
ChatGLM-6B	60 亿	82.38%
ChatGLM2-6B	60 亿	89.61%
ChatGLM3-6B	60 亿	86.35%
Baichuan-7B	70 亿	76.64%
QWEN-7B	70 亿	87.98%

（3）不同提示词构造方式对阅读理解性能影响

表 5-5 所示为使用不同的提示构造方式，在 ChatGLM 6B 与 Baichuan 7B 两个模型上评估阅读理解性能的结果。

表 5-5 不同提示词构造方式对阅读理解性能影响

构造方式	ChatGLM 6B	Baichuan 7B
Q-C-I	68.85%	76.43%
Q-I-C	68.85%	71.93%
C-Q-I	82.38%	70.29%
C-I-Q	81.97%	67.42%
I-Q-C	64.55%	76.64%
I-C-Q	77.66%	73.57%

实验结果表明，提示词的构造方式对模型性能有显著影响，且这种影响在不同模型之间存在差异。例如，在 ChatGLM 模型中，将篇幅冗长的相关文档（Q）置于输入的前半部分能够获得更好的效果，而将其置于输入的末尾则会明显降低性

能。该结论在 Baichuan 模型中并不适用。因此，为优化阅读器的性能，需要针对特定模型调整提示词构造策略。

5.5 本章小结

本章介绍了基于大模型检索增强生成技术的知识问答模块设计，该模块包含检索器和阅读器两大关键部件。实验结果表明，本章所采用检索器能实现较好的性能，可有效解决当前自动问答系统在语义理解、知识检索方面所面临的挑战。而在阅读器中引入生成式大模型，不仅可提高答案的文本丰富性，而且显著提升了阅读理解的性能。

在企业生产实际应用中，不仅要考虑系统的准确率等性能指标，还需考虑推理效率与响应速度，对此，本章提出了 top-k 的设置应与知识库的规模相关联，并结合企业现有的客服知识的规模，确定了该参数的设置范围。在阅读理解任务中，本文对比了当前中文开源大型语言模型在阅读理解任务中的表现，并分析了不同参数量对性能的影响。本章最后提出了适用于企业实际生成环境部署的参数量。这些研究对企业实际部署应用提供了参考依据。

第六章 汽车客服自动问答系统设计及应用验证

本章在以上章节的研究基础上，整合知识融合、多轮对话、知识问答等核心模块，完成完整的汽车客服自动问答系统设计。通过该系统，进行实际应用验证。

6.1 系统设计

6.1.1 功能设计

系统功能设计如图 6-1 所示，包含用户前端、知识管理前端、文档处理、知识问答、多轮对话和日志 6 个模块 32 个子模块。其中文档处理、知识问答和多轮对话为三大核心模块，包含 20 个子模块，由于其已在本文第三至五章中详细介绍，故不再赘述，本章将对其应用进行验证。其中，用户前端、知识管理前端实现系统交互的主要界面与基本功能，日志系统记录各环节操作日志。系统包括设计用户、知识管理员、运维管理员三种角色。用户通过用户前端模块与系统进行交互，知识管理员使用知识管理前端相关功能，其余功能由运维管理员使用和维护。

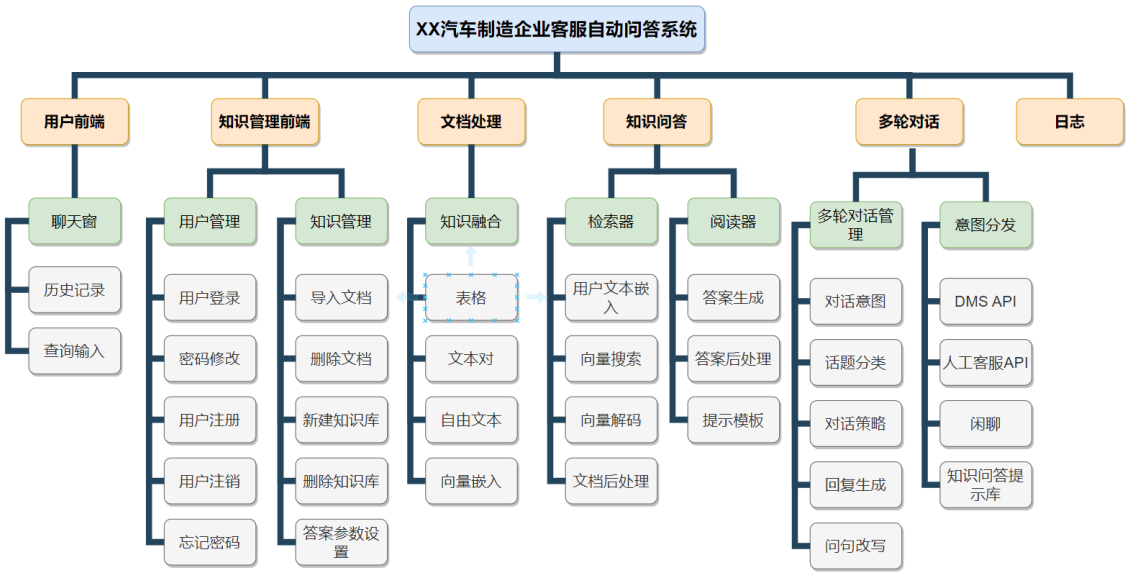

图 6-1 系统功能设计

6.1.2 架构设计

系统架构如图 6-2 所示。分为数据层、模型层、框架层、算法层、服务层和应用层六个层次。数据层是架构的基础，由储存知识的向量数据库、系统账户数据库和第三方 API 构成，其中 API 主要用于与经销商管理系统、外部数据库等信息进行集成。模型层由大模型和提示库组成。该层由深度学习框架和大模型管理框架共同驱动，其中深度学习框架采用 PyTorch 和 Transformers。PyTorch 提供了构建深度学习模型的灵活工具和库，Transformers 库由 Hugging Face 基于 PyTorch 开发，专用于自然语言处理任务，封装了大模型的微调、调用等基础功能。LangChain 作大模型管理框架，负责管理模型、向量数据库和向量检索工具，通过封装常用算法进一步减少算法层的开发工作。算法层由本文第三章至第五章提出的各模块构成。服务层分为两部分：下层采用 FastAPI 实现模块的交互，便于未来实现多样化应用；上层为 Streamlit，是一个包含前后端的 web 框架，其作用是基于 FastAPI 的封装构建 web 应用，实现应用层的管理员端与用户端交互界面。

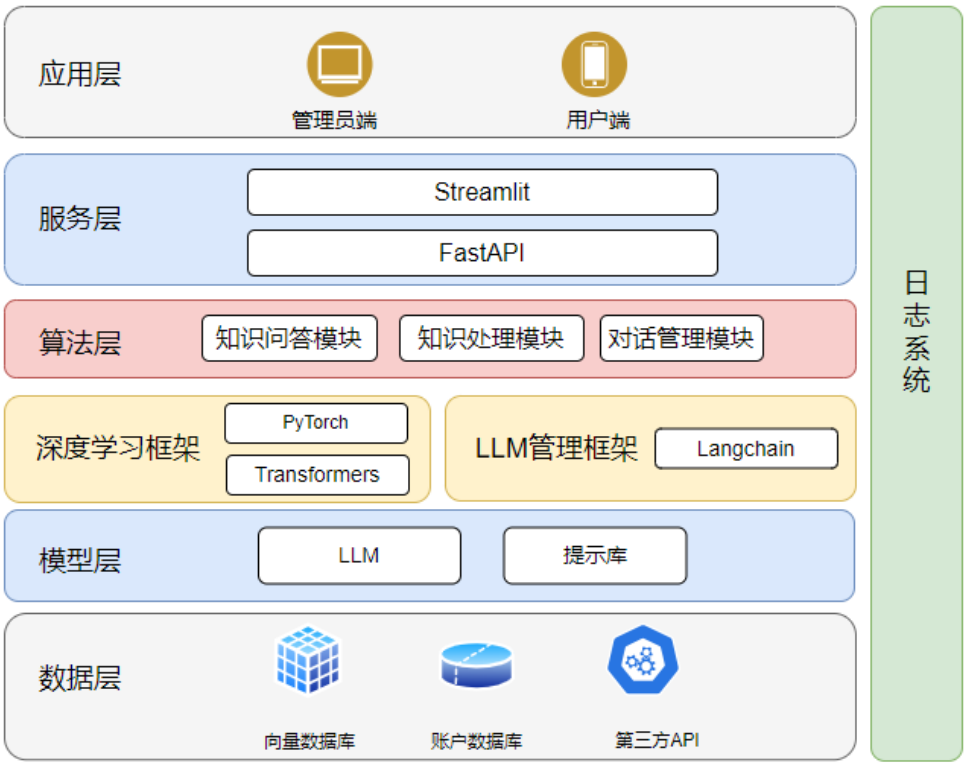

图 6-2 系统架构

6.2 应用验证

6.2.1 前端应用验证

前端应用包含用户前端与管理员前端两部分，如图 6-3、图 6-4 所示。其中，用户前端实现了问答系统与用户交互的基本功能，包括了对话框、历史记录等功能。知识前端则实现了知识上传、删除和添加到向量数据库等功能，知识库的管理人员通过前端直接上传原始知识文件，系统根据文件扩展名自动识别文件内容形式并调用相应结构的文档处理器，按照本文融合表示的方法处理后添加知识到向量数据库中。

图 6-3 XX 汽车客服自动问答系统用户前端应用

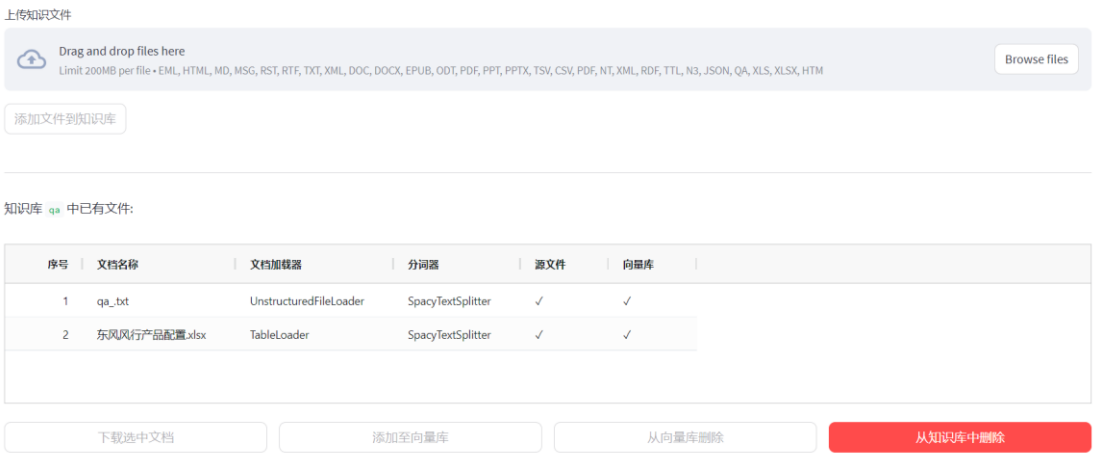

图 6-4 XX 汽车客服自动问答系统知识管理前端应用

6.2.2 多源异构知识利用验证

（1）结构化知识利用验证

如图 6-5 所示，为结构化知识利用的验证，其中，图 6-5（a）为原始的表格知识形式，在经过统一转化后，形成了非结构化的文本片段，并被问答系统正确检索和生成答案，如图 6-5（b）所示，结构化知识经过融合后，能够被自动问答系统利用。

车型名称	风行T5 马赫版 1.5TD/7DCT 激擎款五座	风行T5 马赫版 1.5TD/7DCT 燃擎款五座	风行T5 马赫版 1.5TD/7DCT 劲擎款五座	风行T5 马赫版 1.5TD/7DCT 激擎款七座	风行T5 马赫版 1.5TD/7DCT 燃擎款七座	风行T5 马赫版 1.5TD/7DCT 劲擎款七座
指导价(元)	99900	109900	119900	102900	112900	122900
最大功率(kW)	140	140	140	140	140	140
最大功率转速（rpm）	5200	5200	5200	5200	5200	5200
最大扭矩(N·m)	300	300	300	300	300	300
最大扭矩转速（rpm）	2000-4000	2000-4000	2000-4000	2000-4000	2000-4000	2000-4000
发动机特有技术	缸内直喷	缸内直喷	缸内直喷	缸内直喷	缸内直喷	缸内直喷
燃料形式	汽油	汽油	汽油	汽油	汽油	汽油
燃油标号	92#	92#	92#	92#	92#	92#
供油方式	直喷	直喷	直喷	直喷	直喷	直喷
缸盖材料	铝合金	铝合金	铝合金	铝合金	铝合金	铝合金
缸体材料	铝合金	铝合金	铝合金	铝合金	铝合金	铝合金
变速箱						
挡位个数	7	7	7	7	7	7
变速箱类型	湿式双离合变速箱(DCT)	湿式双离合变速箱(DCT)	湿式双离合变速箱(DCT)	湿式双离合变速箱(DCT)	湿式双离合变速箱(DCT)	湿式双离合变速箱(DCT)
简称	7DCT	7DCT	7DCT	7DCT	7DCT	7DCT
底盘转向						
驱动方式	前置前驱	前置前驱	前置前驱	前置前驱	前置前驱	前置前驱
前悬架类型	麦弗逊式独立悬挂+横向稳定杆	麦弗逊式独立悬挂+横向稳定杆	麦弗逊式独立悬挂+横向稳定杆	麦弗逊式独立悬挂+横向稳定杆	麦弗逊式独立悬挂+横向稳定杆	麦弗逊式独立悬挂+横向稳定杆
后悬架类型	多连杆式独立后悬	多连杆式独立后悬	多连杆式独立后悬	多连杆式独立后悬	多连杆式独立后悬	多连杆式独立后悬

T5 EVO | T5马赫版 | 风行SX6 | 风行游艇 | 菱智PLUS | 菱智M5 | PLUS(CNG) | M7领爵款 | 风行M7 | M5EV | S50EV

（a）

（b）

图 6-5 结构化知识利用验证。（a）知识的原始形式；（b）结构化知识经过知识融合后应用于问答系统中

（2）文本对知识验证

文本对知识形式如图 2-2，验证情况如图 6-6，在该案例中，“相关文档[1]”与“相关文档[2]”均源于文本对，验证结果显示文本对知识经过融合后，能够被自动问答系统利用。

图 6-6 文本对知识利用验证

本文将多源异构知识统一为非结构化的自由文本形式，故不再赘述非结构化知识的利用验证。

6.2.3 多轮对话验证

如图 6-3、图 6-7 和图 6-8 所示。图 6-3 验证模型的上下文理解能力，在此案例中，模拟用户在与系统第二轮交互提出的问题“那 V3 呢？”，并未具体指明询问内容，但系统能根据历史交流内容，准确推断出用户所询问的是关于座椅数量的问题。图 6-7 验证了模型在缺乏必要信息时的行为模式，系统通过向用户提问以获取所需信息，待获取补充信息后生成相应的回答。图 6-8 则验证了模型意图识别的能力，模型正确识别了用户的“保养预约”和“表扬与建议”等不同意图。

图 6-7 多轮对话验证

图 6-8 意图识别验证

6.2.4 知识问答模块验证

如图 6-3~图 6-7 均已从不同角度验证了知识问答模块的语义理解能力与检索能力。其中如图 6-3 验证了模型的上下文理解能力；图 6-5 验证了常规问句的理解能力，如图 6-6 中模型能够正确将“机头”理解为发动机舱，并检索出相应的文档，而原始的知识库中并不包含“机头”一词，验证了模型对口语的理解能力。

6.3 本章小结

本章在前序章节研究的基础上，完成了汽车客服自动问答系统的功能设计与架构设计，研发了 XX 汽车制造企业客服自动问答系统。通过该问答系统，重点阐述文档处理模块、多轮对话模态、知识问答模块的应用验证，验证了本文所提出的方法的可用性。

第七章 总结与展望

7.1 总结

本文以 XX 汽车制造企业“智能客服问答系统建设”为背景，针对现有自动问答系统面临的问题，提出一种基于大模型的汽车客服自动问答系统，将大模型应用于整个问答系统的多个模块中。具体研究成果如下：

（1）针对现有自动问答系统无法利用多种结构知识的问题，本文提出了一种多源异构客服知识的融合方法，通过预设规则，将结构化的表格、知识图谱与文本对统一为自由文本的形式，然后通过切分、标题增强及向量化等步骤进行处理，以便在基于自由文本的自动问答系统中高效利用，使用该方法的有效性已在公开数据集上得到验证。

（2）为提升系统对多轮交互问题的处理能力，本文基于提示工程的理念，提出了一种基于大型语言模型的提示框架，并基于此框架设计了一种多轮对话问题的解决方案。该方案将提示框架应用于多轮对话的意图识别、槽位填充和自然语言生成等多个子模块，利用大型语言模型的泛化能力提升各子模块的性能。

（3）针对现有系统检索能力不足的问题，本文提出了一种将大型语言模型检索增强生成技术应用于企业自动问答系统的方法。首先，使用预训练的密集嵌入模型将问题文本与知识库转换为密集向量。然后，利用大型语言模型作为阅读器，生成最终答案，从而提升答案的抽取和总结能力。最后，通过实验探讨了模型选型和关键参数设计，以平衡问答系统的准确性和响应速度。

7.2 展望

本文面向汽车客服自动问答系统面临的主要挑战提出了解决方案，但限于时间、研究角度和研究能力，提出的方法仍有待进一步优化，且部分理论也需进一步深入讨论。具体可以从以下几个方面开展：

（1）本文将大模型用于问答系统的多个模块中，一次查询中需要多次调用大模型推理，增加了算力资源消耗，影响响应速度；如何优化推理速度，尤其是高并发时的推理速度，不仅是需要解决的工程问题，也是当前学术界研究的热点方向。

（2）本文提出多轮对话的方法仍然一定程度依赖人工制定规则。近年来，已有一些研究者在问答系统中尝试引入智能体（AI Agent），由 AI 自主决策下一步回

复策略。但其性能尤其是鲁棒性尚未达到可部署到生产环境的要求，如何更有效地利用 AI Agent 技术，减少手工设计规则，值得进一步研究和探讨。

未来，随着国家、各行业在算力上的投入提升，算力将不再是制约大模型广泛应用的因素；同时，模型结构和机理的深入研究也将进一步提高同等算力下的模型性能。因此，在以汽车客服自动问答系统为代表的专业领域中，大模型的应用也将更为广泛和深入。将为用户提供更智能、高效的信息交互体验。

参考文献

[1] 中国汽车工业协会. 数据统计[EB/OL]. [2024–04–03]. http://www.caam.org.cn/tjsj.

[2] 中华人民共和国商务部. 2020 年我国汽车后市场消费规模超万亿元[EB/OL].[2024–04–03]. http://www.mofcom.gov.cn/article/news/202103/20210303044933.shtml.

[3] 郑实福刘挺. 自动问答综述[J]. 中文信息学报, 2002(06): 46–52.

[4] 闫悦, 郭晓然, 王铁君, 等. 问答系统研究综述[Z].(2023).

[5] 毛先领, 李晓明. 问答系统研究综述[J]. 计算机科学与探索, 2012, 6(3): 193-207.

[6] 仇韫琦, 王元卓, 白龙, 等. 面向知识库问答的问句语义解析研究综述[J]. 电子学报, 2022, 50(9): 2242–2264.

[7] Bordes A, Weston J, Usunier N. Open question answering with weakly supervised embedding models[C]. Calders T, Esposito F, Hüllermeier E, et al., eds.//Machine Learning and Knowledge Discovery in Databases. Berlin, Heidelberg: Springer, 2014: 165–180. DOI:10.1007/978-3-662-44848-9_11.

[8] Bordes A, Chopra S, Weston J. Question answering with subgraph embeddings[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar: Association for Computational Linguistics, 2014: 615–620[2022–12–04]. http://aclweb.org/anthology/D14-1067. DOI:10.3115/v1/D14-1067.

[9] Chen Y, Wu L, Zaki M J. Bidirectional attentive memory networks for question answering over knowledge bases[C]//Proceedings of the 2019 Conference of the North. Minneapolis, Minnesota: Association for Computational Linguistics, 2019: 2913–2923[2022–12–04]. http://aclweb.org/anthology/N19-1299. DOI:10.18653/v1/N19-1299.

[10] Wang Z, Ng P, Nallapati R, et al. Retrieval, re-ranking and multi-task learning for knowledge-base question answering[C]. Merlo P, Tiedemann J, Tsarfaty R, eds.//Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. Online: Association for Computational Linguistics, 2021: 347–357[2024–04–04]. https://aclanthology.org/2021.eacl-main.26. DOI:10.18653/v1/2021.eacl-main.26.

[11] Dong L, Wei F, Zhou M, et al. Question answering over freebase with multi-column convolutional neural networks[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Beijing, China: Association for Computational Linguistics, 2015: 260–269[2022–12–04]. https://aclanthology.org/P15-1026. DOI:10.3115/v1/P15-1026.

[12] Hao Y, Zhang Y, Liu K, et al. An end-to-end model for question answering over knowledge base with cross-attention combining global knowledge[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vancouver, Canada: Association for Computational Linguistics, 2017: 221–231[2022–12–04]. https://aclanthology.org/P17-1021. DOI:10.18653/v1/P17-1021.

[13] Robertson S E, Walker S, Jones S, et al. Okapi at trec-3[J]. Nist Special Publication Sp, 1995, 109: 109.

[14] Karpukhin V, Oguz B, Min S, et al. Dense passage retrieval for open-domain question answering[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online: Association for Computational Linguistics, 2020: 6769–6781[2023–03–14].https://aclanthology.org/2020.emnlp-main.550.

[15] Seo M, Kembhavi A, Farhadi A, et al. Bidirectional attention flow for machine comprehension[C]. [2023–12–30]. https://openreview.net/forum?id=HJ0UKP9ge.

[16] Chen D, Fisch A, Weston J, et al. Reading wikipedia to answer open-domain questions: arXiv:1704.00051[Z].arXiv, 2017(2017–04–27)[2023–12–30]. http://arxiv.org/abs/1704.00051. DOI:10.48550/arXiv.1704.00051.

[17] Dong L, Yang N, Wang W, et al. Unified language model pre-training for natural language understanding and generation[C]//Advances in Neural Information Processing Systems. Curran Associates,Inc., 2019[2024–04–05]. https://proceedings.neurips.cc/paper_files/paper/2019/hash/c20bb2d9a50d5ac1f713f8b34d9aac5a-Abstract.html.

[18] Devlin J, Chang M-W, Lee K, et al. Bert: pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

[19] Xiao D, Zhang H, Li Y, et al. ERNIE-gen: an enhanced multi-flow pre-training and fine-tuning framework for natural language generation: arXiv:2001.11314[Z]. arXiv, 2020(2020–06–08)[2024–04–05]. http://arxiv.org/abs/2001.11314. DOI:10.48550/arXiv.2001.11314.

[20] 钱锦，黄荣涛，邹博伟，等. 基于多任务学习的生成式阅读理解[J]. 中文信息学报，2021,35(12): 103-111+121.

[21] Peng S, Cui H, Xie N, et al. Enhanced-rcnn: an efficient method for learning sentence similarity[C]//Proceedings of The Web Conference 2020. New York, NY, USA: Association for Computing Machinery,2020:2500–2506[2024–01–28]. https://doi.org/10.1145/3366423.3379998. DOI:10.1145/3366423.3379998.

[22] Mueller J, Thyagarajan A. Siamese recurrent architectures for learning sentence similarity[C]//Proceedings of the AAAI conference on artificial intelligence. 2016, 30(1).

[23] 潘理虎, 刘杰, 白尚旺, 等. 基于 Word2vec 和句法规则的自动问答系统问句相似度研究[J]. 计算机应用与软件, 2021, 38(03): 169-174+201.

[24] 丁邱, 迟海洋, 严馨, 等. 基于 Transformer 模型的问句语义相似度计算[J]. 计算机工程与设计, 2023,44(03): 887–893. DOI:10.16208/j.issn1000-7024.2023.03.034.

[25] Ferrucci D A. Introduction to" This is Watson"[J]. IBM Journal of Research and Development, 2012, 56(3): 235-249.

[26] Sun H, Dhingra B, Zaheer M, et al. Open domain question answering using early fusion of knowledge bases and text[C]. Riloff E, Chiang D, Hockenmaier J, et al., eds.//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium: Association for Computational Linguistics, 2018: 4231–4242[2024–01–31]. https://aclanthology.org/D18-1455. DOI:10.18653/v1/D18-1455.

[27] Agarwal O, Ge H, Shakeri S, et al. Knowledge graph based synthetic corpus generation for knowledge-enhanced language model pre-training[C]. Toutanova K, Rumshisky A, Zettlemoyer L, et al., eds.//Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Online: Association for Computational Linguistics, 2021: 3554–3565[2024–01–31]. https://aclanthology.org/2021.naacl-main.278. DOI:10.18653/v1/2021.naacl-main.278.

[28] Oguz B, Chen X, Karpukhin V, et al. UniK-qa: unified representations of structured and unstructured knowledge for open-domain question answering: arXiv:2012.14610[Z]. arXiv, 2022(2022–05–03)[2022–12–04].http://arxiv.org/abs/2012.14610.

[29] Khashabi D, Min S, Khot T, et al. UnifiedQA: crossing format boundaries with a single qa system[R]//arXiv E-Prints. [2022–12–09]. https://ui.adsabs.harvard.edu/abs/2020arXiv200500700K.

[30] Yan Z, Duan N, Chen P, et al. Building task-oriented dialogue systems for online shopping[C]. .

[31] 于长宏, 詹志强. 基于深度学习的复合任务多轮对话系统研究与实现[D]. 北京邮电大学, 2023.

[32] Chen H, Liu X, Yin D, et al. A survey on dialogue systems: recent advances and new frontiers[J]. ACM SIGKDD Explorations Newsletter, 2017, 19(2): 25–35. DOI:10.1145/3166054.3166058.

[33] Mesnil G, He X, Deng L, et al. Investigation of recurrent-neural-network architectures and learning methods for spoken language understanding[C]//Interspeech. 2013: 3771-3775.

[34] Hakkani-Tür D, Tur G, Celikyilmaz A, et al. Multi-domain joint semantic frame parsing using bi-directional rnn-lstm[C]//Interspeech 2016. ISCA, 2016: 715–719[2024–02–13].

https://www.isca-archive.org/interspeech_2016/hakkanitur16_interspeech.html. DOI:10.21437/Interspeech.2016-402.

[35] Liu B, Lane I. Attention-based recurrent neural network models for joint intent detection and slot filling: arXiv:1609.01454[Z]. arXiv, 2016(2016–09–06)[2024–02–13]. http://arxiv.org/abs/1609.01454. DOI:10.48550/arXiv.1609.01454.

[36] Zhou X, Li L, Dong D, et al. Multi-turn response selection for chatbots with deep attention matching network[C]. Gurevych I, Miyao Y, eds.//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Melbourne, Australia: Association for Computational Linguistics, 2018: 1118–1127[2024–01–03]. https://aclanthology.org/P18-1103. DOI:10.18653/v1/P18-1103.

[37] Wu C-S, Madotto A, Hosseini-Asl E, et al. Transferable multi-domain state generator for task-oriented dialogue systems[C]. Korhonen A, Traum D, Màrquez L, eds.//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics, 2019: 808–819[2024–01–04]. https://aclanthology.org/P19-1078. DOI:10.18653/v1/P19-1078.

[38] Kwan W-C, Wang H-R, Wang H-M, et al. A survey on recent advances and challenges in reinforcement learning methods for task-oriented dialogue policy learning[J]. Machine Intelligence Research, 2023, 20(3): 318–334.

[39] Wen T-H, Gasic M, Mrksic N, et al. Semantically conditioned lstm-based natural language generation for spoken dialogue systems[EB/OL].(2015–08–07)[2024–01–04]. https://arxiv.org/abs/1508.01745v2.

[40] Topal M O, Bas A, van Heerden I. Exploring transformers in natural language generation: gpt, bert, and xlnet: arXiv:2102.08036[Z]. arXiv, 2021(2021–02–16)[2024–01–04]. http://arxiv.org/abs/2102.08036. DOI:10.48550/arXiv.2102.08036.

[41] Zhang H, Song H, Li S, et al. A survey of controllable text generation using transformer-based pre-trained language models[J]. ACM Computing Surveys, 2023, 56(3): 64:1-64:37. DOI:10.1145/3617680.

[42] Wen T-H, Vandyke D, Mrkšić N, et al. A network-based end-to-end trainable task-oriented dialogue system[C]. Lapata M, Blunsom P, Koller A, 编//Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers. Valencia, Spain: Association for Computational Linguistics, 2017: 438–449[2024–04–05]. https://aclanthology.org/E17-1042.

[43] Eric M, Manning C D. A Copy-Augmented Sequence-to-Sequence Architecture Gives Good Performance on Task-Oriented Dialogue[C]//Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers. 2017: 468-473.

[44] Lei W, Jin X, Kan M-Y, et al. Sequicity: simplifying task-oriented dialogue systems with single sequence-to-sequence architectures[C]. Gurevych I, Miyao Y, 编//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Melbourne, Australia: Association for Computational Linguistics, 2018: 1437–1447[2024–04–06]. https://aclanthology.org/P18-1133. DOI:10.18653/v1/P18-1133.

[45] Yang Y, Li Y, Quan X. UBAR: towards fully end-to-end task-oriented dialog system with gpt-2: 16[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(16): 14230–14238. DOI:10.1609/aaai.v35i16.17674.

[46] Zhao W X, Zhou K, Li J, et al. A survey of large language models: arXiv:2303.18223[Z]. arXiv, 2023(2023–11–24)[2024–02–25]. http://arxiv.org/abs/2303.18223.

[47] Liu P, Yuan W, Fu J, et al. Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing: arXiv:2107.13586[Z]. arXiv, 2021(2021–07–28)[2022–12–07]. http://arxiv.org/abs/2107.13586. DOI:10.48550/arXiv.2107.13586.

[48] openAI. GPT-4[EB/OL].[2024–02–01]. https://openai.com/gpt-4.

[49] Tan Y, Min D, Li Y, et al. Can chatgpt replace traditional kbqa models? an in-depth analysis of the question answering performance of the gpt llm family[C]. Payne T R, Presutti V, Qi G, et al., eds.//The Semantic Web – ISWC 2023. Cham: Springer Nature Switzerland, 2023: 348–367. DOI:10.1007/978-3-031-47240-4_19.

[50] Welleck S, Kulikov I, Roller S, et al. Neural text generation with unlikelihood training: arXiv:1908.04319[Z].arXiv, 2019(2019–09–26)[2024–03–12]. http://arxiv.org/abs/1908.04319. DOI:10.48550/arXiv.1908.04319.

[51] Jeong C. A study on the implementation of generative ai services using an enterprise data-based llm application architecture[J]. Advances in Artificial Intelligence and Machine Learning, 2023, 03(04): 1588–1618. DOI:10.54364/AAIML.2023.1191.

[52] Han Z, Gao C, Liu J, et al. Parameter-efficient fine-tuning for large models: a comprehensive survey:arXiv:2403.14608[Z].arXiv,2024(2024–04–01)[2024–04–06]. http://arxiv.org/abs/2403.14608. DOI:10.48550/arXiv.2403.14608.

[53] Hu E J, Shen Y, Wallis P, et al. LoRA: low-rank adaptation of large language models[C]. [2024–04–06]. https://openreview.net/forum?id=nZeVKeeFYf9.

[54] Li X L, Liang P. Prefix-tuning: optimizing continuous prompts for generation: arXiv:2101.00190[Z].arXiv, 2021(2021–01–01)[2024–01–13]. http://arxiv.org/abs/2101.00190. DOI:10.48550/arXiv.2101.00190.

[55] Liu X, Ji K, Fu Y, et al. P-tuning v2: prompt tuning can be comparable to fine-tuning universally across scales and tasks: arXiv:2110.07602[Z]. arXiv, 2022(2022–03–20)[2024–01–13]. http://arxiv.org/abs/2110.07602. DOI:10.48550/arXiv.2110.07602.

[56] Gao Y, Xiong Y, Gao X, et al. Retrieval-augmented generation for large language models: a survey: arXiv:2312.10997[Z]. arXiv, 2024(2024–03–27)[2024–04–06]. http://arxiv.org/abs/2312.10997. DOI:10.48550/arXiv.2312.10997.

[57] Martino A, Iannelli M, Truong C. Knowledge injection to counter large language model (llm) hallucination[C]. Pesquita C, Skaf-Molli H, Efthymiou V, et al., eds.//The Semantic Web: ESWC 2023 Satellite Events. Cham: Springer Nature Switzerland, 2023: 182–185. DOI:10.1007/978-3-031-43458-7_34.

[58] Jiang Z, Xu F, Gao L, et al. Active retrieval augmented generation[C]. Bouamor H, Pino J, Bali K, eds.//Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Singapore: Association for Computational Linguistics, 2023: 7969–7992[2024–04–06]. https://aclanthology.org/2023.emnlp-main.495. DOI:10.18653/v1/2023.emnlp-main.495.

[59] Du Z, Qian Y, Liu X, et al. GLM: general language model pretraining with autoregressive blank infilling[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). .

[60] Bai J, Bai S, Chu Y, et al. Qwen technical report: arXiv:2309.16609[Z]. arXiv, 2023(2023–09–28)[2024–02–27]. http://arxiv.org/abs/2309.16609. DOI:10.48550/arXiv.2309.16609.

[61] MetaAI. Facebookresearch/faiss[Z]. Meta Research, 2024(2024–02–04)[2024–02–04]. https://github.com/facebookresearch/faiss.

[62] 中国计算机学会自然语言处理专业委员会. NLPCC 2018 home[EB/OL]. [2024–01–01]. http://tcci.ccf.org.cn/conference/2018/index.php.

[63] Duan N. Overview of the nlpcc 2018 shared task: open domain qa[C]. Zhang M, Ng V, Zhao D, et al., eds.//Natural Language Processing and Chinese Computing. Cham: Springer International Publishing, 2018: 452–456.

[64] Koch G, Zemel R, Salakhutdinov R. Siamese neural networks for one-shot image recognition[C]//ICML deep learning workshop. Lille, 2015.

[65] Neculoiu P, Versteegh M, Rotaru M. Learning text similarity with siamese recurrent networks[C]. Blunsom P, Cho K, Cohen S, et al., eds.//Proceedings of the 1st Workshop on

Representation Learning for NLP. Berlin, Germany: Association for Computational Linguistics, 2016: 148–157[2024–02–01]. https://aclanthology.org/W16-1617. DOI:10.18653/v1/W16-1617.

[66] Ye Q, Axmed M, Pryzant R, et al. Prompt engineering a prompt engineer: arXiv:2311.05661[Z]. arXiv, 2023(2023–11–09)[2024–01–07]. http://arxiv.org/abs/2311.05661.

[67] Chen B, Zhang Z, Langrené N, et al. Unleashing the potential of prompt engineering in large language models: a comprehensive review: arXiv:2310.14735[Z]. arXiv, 2023(2023–10–27)[2024–01–07]. http://arxiv.org/abs/2310.14735.

[68] Reynolds L, McDonell K. Prompt programming for large language models: beyond the few-shot paradigm: arXiv:2102.07350[Z]. arXiv, 2021(2021–02–15)[2024–01–08]. http://arxiv.org/abs/2102.07350. DOI:10.48550/arXiv.2102.07350.

[69] Lester B, Al-Rfou R, Constant N. The power of scale for parameter-efficient prompt tuning: arXiv:2104.08691[Z].arXiv, 2021(2021–09–02)[2024–03–01]. http://arxiv.org/abs/2104.08691.

[70] Liu X, Wang J, Sun J, et al. Prompting frameworks for large language models: a survey[EB/OL].(2023–11–21)[2024–01–07]. https://arxiv.org/abs/2311.12785v1.

[71] 赵阳洋, 王振宇, 王佩, 等. 任务型对话系统研究综述[J]. 计算机学报, 2020, 43(10): 1862–1896.

[72] 张伟男, 张杨子, 刘挺. 对话系统评价方法综述[J]. 中国科学: 信息科学, 2017, 47(8): 953–966.

[73] Papineni K, Roukos S, Ward T, et al. Bleu: a method for automatic evaluation of machine translation[C]. Isabelle P, Charniak E, Lin D, eds.//Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, 2002: 311–318[2024–01–12]. https://aclanthology.org/P02-1040. DOI:10.3115/1073083.1073135.

[74] Zeng A, Liu X, Du Z, et al. Glm-130b: an open bilingual pre-trained model[J]. arXiv preprint arXiv:2210.02414, 2022.

[75] Goo C-W, Gao G, Hsu Y-K, et al. Slot-gated modeling for joint slot filling and intent prediction[C]. Walker M, Ji H, Stent A, eds.//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). New Orleans, Louisiana: Association for Computational Linguistics, 2018: 753–757[2024–02–14]. https://aclanthology.org/N18-2118. DOI:10.18653/v1/N18-2118.

[76] Chen Q, Zhuo Z, Wang W. BERT for joint intent classification and slot filling: arXiv:1902.10909[Z].arXiv, 2019(2019–02–28)[2024–02–14]. http://arxiv.org/abs/1902.10909. DOI:10.48550/arXiv.1902.10909.

[77] Babenko A, Lempitsky V. The inverted multi-index[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(6): 1247–1260. DOI:10.1109/TPAMI.2014.2361319.

[78] Kapanipathi P, Abdelaziz I, Ravishankar S, et al. Leveraging abstract meaning representation for knowledge base question answering: arXiv:2012.01707[Z]. arXiv, 2021(2021–06–02)[2022–12–04]. http://arxiv.org/abs/2012.01707. DOI:10.48550/arXiv.2012.01707.

[79] Yang A, Xiao B, Wang B, et al. Baichuan 2: open large-scale language models[EB/OL].(2023–09–19)[2024–02–27]. https://arxiv.org/abs/2309.10305v2.

[80] Cotterell R, Mielke S J, Eisner J, et al. Are all languages equally hard to language-model?[EB/OL].(2018–06–10)[2024–02–27]. https://arxiv.org/abs/1806.03743v2.

[81] Petrov A, La Malfa E, Torr P, et al. Language model tokenizers introduce unfairness between languages[J]. Advances in Neural Information Processing Systems, 2023, 36.

[82] Wang T, Roberts A, Hesslow D, et al. What language model architecture and pretraining objective work best for zero-shot generalization? arXiv:2204.05832[Z]. arXiv, 2022(2022–04–12)[2024–02–27]. http://arxiv.org/abs/2204.05832. DOI:10.48550/arXiv.2204.05832.