

面向钻井地质设计知识问答的检索增强与生成优化方法

魏小涛^{1*}, 段俊卿², 梁任腾³, 陈瑞军⁴, 张伟⁴, 王世明⁴, 刘永杰⁴

¹中国石油大学(北京)克拉玛依校区石油学院, 新疆 克拉玛依

²北京交通大学软件学院, 北京

³中兴通讯股份有限公司, 北京

⁴红有软件股份有限公司, 新疆 克拉玛依

收稿日期: 2026年7月16日; 录用日期: 2026年8月18日; 发布日期: 2026年8月25日

摘要

在企业知识问答场景下, 大模型支撑系统经常会存在知识缺失、泛化回复、实体关系提取能力弱以及可追溯性不足等问题, 因此, 本文提出一种面向油田钻井地质设计领域的检索增强与生成优化协同方法。该方法将Ict-Retriever作为检索前端, 通过企业文档切片与向量索引实现外部知识召回。同时构建钻井地质设计专用指令微调数据集ICT-DrillGeoDesign, 该数据集涵盖地层岩性、压力预测、钻井液设计、井身结构等核心领域知识。除此之外, 对ChatGLM3-6B模型进行QLoRA微调, 进而得到面向钻井地质设计业务的大模型IctLM-Drill。最后, 后续研究将围绕钻井地质设计规范、油气层特征、钻井工程参数以及安全预警标准等核心文档构建领域知识图谱, 让系统能够完成地质设计问答、实体关系抽取以及设计参数推荐等任务。实验结果表明, IctLM-Drill在原始ChatGLM3-6B基础上, BLEU、PPL、Distinct指标上分别达到0.458、6.82、9.15。联动Ict-Retriever后, 端到端模型指标进一步提升至0.658、6.42、16.87。该方法有效提升了钻井地质设计场景下问答的准确性、专业性和实用性, 为油田钻井工程设计提供了智能化支撑。

关键词

大语言模型, 检索增强生成, 参数高效微调, 钻井地质设计, 领域知识问答

Retrieval-Augmented and Generation Optimized Method for Question Answering in Drilling Geological Design

Xiaotao Wei^{1*}, Junqing Duan², Renteng Liang³, Ruijun Chen⁴, Wei Zhang⁴, Shiming Wang⁴,

*通讯作者。

文章引用: 魏小涛, 段俊卿, 梁任腾, 陈瑞军, 张伟, 王世明, 刘永杰. 面向钻井地质设计知识问答的检索增强与生成优化方法[J]. 计算机科学与应用, 2026, 16(8): 204-219. DOI: 10.12677/csa.2026.168275

Yongjie Liu⁴¹School of Petroleum, Karamay Campus, China University of Petroleum (Beijing), Karamay Xinjiang²School of Software Engineering, Beijing Jiaotong University, Beijing³ZTE Corporation, Beijing⁴Hongyousoft Co., Ltd., Karamay Xinjiang

Received: July 16, 2026; accepted: August 18, 2026; published: August 25, 2026

Abstract

Large language model-based question-answering systems deployed in enterprise environments often face several practical challenges, including incomplete domain knowledge, generic responses, weak capability in extracting entity relations, and limited answer traceability. To address these issues, this paper proposes a retrieval-augmented and generation-enhanced framework for the drilling geological design domain in oilfields. The framework incorporates Ict-Retriever as the retrieval module. Enterprise documents are segmented into manageable passages and encoded into a vector index to support efficient retrieval of relevant external knowledge. In parallel, a domain-specific instruction-tuning dataset, named ICT-DrillGeoDesign, is constructed for drilling geological design tasks. The dataset covers key technical topics, including formation lithology, pressure prediction, drilling fluid design, and wellbore structure design. Based on this dataset, ChatGLM3-6B is fine-tuned using QLoRA to obtain IctLM-Drill, a domain-adapted language model for drilling geological design. In addition, a domain knowledge graph will be developed from core technical materials in the future, including drilling geological design specifications, reservoir characteristics, drilling engineering parameters, and safety warning standards. The knowledge graph enables the system to support multiple downstream tasks, such as domain question answering, entity and relation extraction, and drilling design parameter recommendation. Experimental results show that IctLM-Drill achieves BLEU, PPL, and Distinct scores of 0.458, 6.82, and 9.15, respectively, outperforming the original ChatGLM3-6B model. After integrating Ict-Retriever, the end-to-end system further improves the corresponding results to 0.658, 6.42, and 16.87. These results demonstrate that the proposed framework can effectively improve the accuracy, domain relevance, and practical usefulness of question answering in drilling geological design scenarios, providing intelligent support for oil-field drilling engineering design.

Keywords

Large Language Model, Retrieval-Augmented Generation, Parameter-Efficient Fine-Tuning, Drilling Geological Design, Domain Knowledge QA

Copyright © 2026 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>

Open Access

1. 引言

近年来,大语言模型在自然语言处理、智能问答和知识检索等任务中展现出较强的语义理解与文本生成能力。然而,现有大语言模型的知识主要来源于预训练阶段形成的参数记忆,当面对更新频繁、专业性较强或具有明显私域特征的企业业务知识时,往往难以及时获取准确的领域信息,容易出现知识覆盖不足、知识时效性下降以及事实性错误等问题[1]-[3]。对于企业内部管理与技术支撑场景而言,上述问

题在油田钻井地质设计业务中尤为明显。钻井地质设计涉及地层划分、岩性识别、地层压力预测、钻井液体系设计、井身结构优化、套管下深确定以及油气层保护等多个专业环节，具有知识类型多样、专业术语密集、业务约束严格等特点。相关知识通常分散存储于钻井地质设计报告、区块地质资料、钻井工程规范、行业标准及历史井史资料等不同载体中，组织形式不统一，检索与复用难度较大。

检索增强生成(Retrieval-Augmented Generation, RAG)通过在模型推理过程中引入外部知识库，将大语言模型的参数化知识与可动态更新的非参数化知识相结合，为知识密集型任务提供了有效的解决思路[4][5]。在钻井地质设计知识问答场景中，检索模块能够依据用户问题，从领域知识库中召回相关的设计规范条款、地层岩性描述、地层压力预测结果以及邻井钻探资料等内容，并将其作为生成阶段的辅助依据。该机制不仅有助于提升回答的事实准确性和领域相关性，还能够为模型输出提供相对明确的证据来源，从而增强系统结果的可追溯性。然而，钻井地质设计场景的难点并不只在于“能否检索到相关文档”，还在于“能否正确理解专业术语、抽取参数关系，并生成符合设计规范的回答”。如果生成模型缺乏领域指令跟随能力，即使引入检索结果，也可能出现术语误用、参数遗漏、结构化输出不稳定和建议泛化等问题。

与此同时，参数高效微调技术为算力资源受限条件下的大模型领域适配提供了可行方案。LoRA 通过在 Transformer 网络的部分权重矩阵中引入低秩适配参数，在保持预训练模型主体参数冻结的基础上，仅训练少量新增参数即可完成下游任务迁移[6]。在此基础上，QLoRA 进一步结合低比特量化与双重量化等技术，有效降低了模型微调过程中的显存占用，使中小规模算力条件下开展领域模型训练成为可能[6][7]。以 ChatGLM3-6B 为代表的中文开源大语言模型具有较好的中文语义理解和文本生成能力，同时具备相对较低的部署成本，适合作为钻井地质设计知识问答系统的生成模型基础[8][9]。此外，已有研究表明，将领域任务组织为统一的指令与回答形式进行微调，能够增强模型对任务要求、专业术语及结构化输出模式的理解和适应能力[10]。

基于上述分析，本文围绕检索增强与生成优化协同的技术路线，提出一种面向钻井地质设计知识问答的领域化方法。本文的主要贡献如下：

(1) 构建面向钻井地质设计领域的指令微调数据集 ICT-DrillGeoDesign，围绕地层岩性、压力预测、钻井液设计、井身结构和油气层保护等典型任务，形成“大模型预标注 - 人工复核 - 指令化转换”的数据生产流程，提升领域实体和关系标注的一致性。

(2) 基于 ChatGLM3-6B 与 QLoRA 技术构建领域生成模型 IctLM-Drill，在有限算力条件下完成钻井地质设计任务适配，增强模型对专业术语、参数关系、规范约束和结构化输出格式的理解与生成能力。

2. 相关工作

2.1. 检索增强生成

RAG 的核心思想是在文本生成阶段显式访问外部知识源，从而缓解大语言模型参数记忆静态、知识更新困难和回答缺乏溯源依据等问题。Lewis 等人将 RAG 框架集成预训练语言模型，通过检索外部文档集合，在生成阶段利用检索到的文档增强模型输出[4]。随后，系统演进为双编码器、交叉编码器等多种检索架构。Karpukhin 等人的 DPR 证明了在开放域问答场景中，稠密检索器在效果上显著优于传统的 BM25 等稀疏检索器[11]。

随着大语言模型技术的发展，RAG 研究已逐步由早期的检索结果拼接模式，扩展至召回、上下文增强、内容压缩、重排序以及生成约束等多个环节的协同优化[12]。现有研究表明，知识增强系统的性能不仅取决于检索结果的相关性，还受到文档切片粒度、上下文组织策略、Prompt 模板设计、生成约束机制及评测方法等因素的共同影响[13]。对于企业知识问答场景而言，RAG 能够将持续更新的私有知识库引入模型推理过程，为生成结果提供外部依据，从而在一定程度上降低闭卷生成带来的知识缺失和幻觉风险。

现有 RAG 方法主要面向开放域问答和通用企业知识库。钻井地质设计文档包含大量地层、深度、压力、钻井液和套管参数，且不同参数之间存在较强工程约束。针对这类专业场景，仅依赖通用语义检索和简单上下文拼接，仍可能出现证据遗漏、参数错配和回答泛化。因此，需要结合领域元数据、专业任务指令和生成模型适配构建面向钻井地质设计的检索增强方法。

2.2. 参数高效微调与领域指令对齐

除检索增强外，参数高效微调也是推动大语言模型在企业场景中落地的重要技术路径。Hu 等人提出的 LoRA 方法将模型权重更新表示为低秩矩阵乘积，在冻结预训练模型主体参数的基础上，仅训练少量适配参数即可完成下游任务迁移，从而有效降低训练成本，同时保持较好的生成性能[6]。在此基础上，Dettmers 等人提出的 QLoRA 方法结合 4-bitNormalFloat 量化和双重量化等技术，进一步降低了微调过程中的显存占用，使中等规模大语言模型能够在企业常见的有限 GPU 资源条件下完成领域适配[7]。

ChatGLM 系列模型在中文语义理解、问答生成和多轮交互等方面具有较好的基础能力，同时具有相对较低的部署门槛，适合作为企业级知识问答系统的基础模型[8][9]。但钻井地质设计场景包含大量专业术语、工程参数和规范表达，例如地层压力系数、破裂压力、坍塌压力、钻井液密度、套管下深及油气层保护要求等。通用模型未经领域适配时，容易在术语理解、参数关联、规范引用和方案生成等方面产生偏差，难以满足实际工程应用需求。

ChatGLM 系列模型因其在中文场景下的良好表现和较低的数量，成为钻井地质设计领域应用的首选[8][9]。指令微调能够将领域任务组织为从任务描述到输入内容。最后到标准回答的统一形式，使模型学习特定任务的响应方式、专业表达及结构化输出模式。对于钻井地质设计知识问答而言，领域指令数据除应覆盖常规问答任务外，还应包含参数提取、实体关系生成、规范查询及设计方案建议等多类任务。由于通用指令数据集难以保证钻井地质术语、文档表达方式和参数关系的一致性，本文构建了面向钻井地质设计场景的 ICT-DrillGeoDesign 指令数据集，并采用 QLoRA 方法对 ChatGLM3-6B 进行领域微调，以提升生成模型在专业知识理解、任务响应和工程应用方面的表现。

3. 方法

3.1. 问题定义与系统架构

钻井地质设计领域知识问答的核心任务是：给定企业地质设计知识库 D 和用户问题 q ，系统首先从 D 中检索与 q 相关的 Top-k 个文档片段，然后由大模型 G 结合检索到的证据生成专业、准确的地质设计回答 y ，同时支持地层参数、压力预测、钻井液配方等结构化实体关系抽取。

钻井地质设计知识库构建：(1) 数据来源：包括钻井地质设计报告、区块地质资料、钻井工程规范、行业标准文件、钻井事故案例等。(2) 文档类型：涵盖地层描述、岩性分析、压力预测、钻井液设计、井身结构设计、油气层保护方案等。(3) 预处理流程：文档清洗、分块处理、实体识别、向量索引。(4) 检索增强模块：Ict-Retriever 支持多轮对话上下文补全，检索 Top-5 最相关证据片段，支持地层、深度以及岩性等复合查询。(5) 生成优化模块：IctLM-Drill 基于 QLoRA 微调，支持地质设计问答、参数提取、方案推荐等任务。

本文面向钻井地质设计知识问答场景的核心任务可形式化定义为：给定钻井地质设计领域知识库 D 和用户问题 q ，检索模块首先从 D 中召回与 q 语义最相关的 Top-k 个证据片段 $C_k = \{c_1, c_2, \dots, c_k\}$ ；随后生成模型 G 在问题、对话上下文、检索证据的联合约束下，输出自然语言业务问答回复 y ，或在实体关系任务中输出符合预定义 JSON 模式的结构化结果 z 。核心目标是在不显著增加训练和部署成本的前提

下，同时提升系统对钻井地质设计知识利用能力、专业术语表达能力和结构化生成能力。
系统总体架构分为离线知识组织与在线推理两部分，整体框架如图 1 所示。

检索增强与生成优化总体框架

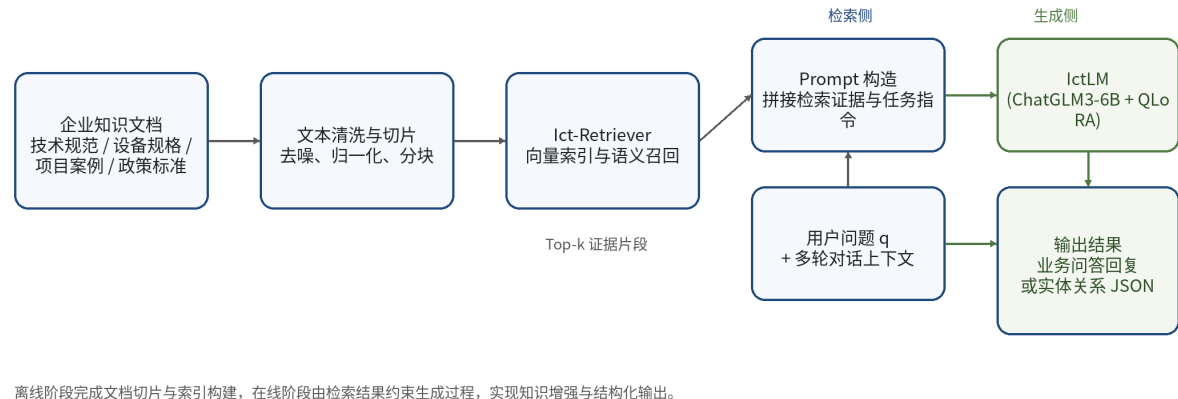


Figure 1. Overall framework of the retrieval-augmented and generation-optimized method
图 1. 检索增强与生成优化总体框架

在离线阶段，首先对钻井地质设计相关文档进行采集、清洗、切片及格式规范化处理。针对不同来源的文档，为文本片段补充井号、区块、地层、深度段、参数类型及文档来源等元数据信息，以增强知识片段的可检索性与可定位性。随后，将处理后的文本片段转换为向量表示，并构建领域知识索引，为后续相关知识召回提供支持。

在在线推理阶段，Ict-Retriever 根据用户问题及多轮对话上下文，从领域知识库中检索与当前任务相关的证据片段。随后，系统通过统一的 Prompt 模板，将用户问题、检索证据和任务指令共同输入 IctLM-Drill，由模型完成钻井地质设计问答、设计参数提取或实体关系生成等任务。

相较于仅依赖通用大模型参数化知识的闭卷生成方式，本文方法通过引入外部知识证据，降低了知识缺失、知识陈旧及幻觉生成带来的影响；通过领域指令微调，增强了模型对钻井地质设计专业术语、参数关系和结构化任务的理解能力；同时，利用统一 Prompt 模板约束问题、证据与输出结果之间的对应关系，进一步提升系统在钻井地质设计知识问答场景中的准确性、可解释性和实际应用价值。

3.2. 面向钻井地质设计的检索增强模块

面向钻井地质设计的检索增强模块钻井地质设计是油气田勘探开发过程中的关键环节，其核心内容包括：在检索侧，本文面向钻井地质设计业务构建领域知识库，知识源主要包括钻井地质设计报告、区块地质资料、钻井工程规范、行业标准文件及典型井史案例等。与通用企业文档相比，钻井地质设计文档具有术语密集、参数关联强和规范约束明显等特点，常见信息包括地层划分、岩性描述、孔隙压力与破裂压力预测、钻井液性能参数、井身结构设计、套管下深及油气层保护要求等。因此，系统在文档预处理阶段不仅进行常规文本清洗、切片和向量化，还在片段元数据中记录井号、区块、地层、深度段、参数类型和规范来源等字段，以支持后续证据定位与结果追溯。

在在线问答阶段，Ict-Retriever 以用户问题为核心查询条件，并结合多轮对话上下文对省略信息和指代信息进行语义补全，从领域知识库中召回 Top-(k)个相关证据片段。针对钻井地质设计中参数多、关联关系复杂的特点，检索模块重点关注地层与深度以及岩性、压力系数与钻井液密度、井身结构与套管下深以及油气层及保护措施等典型知识关联。

例如,当用户询问某一深度段的钻井液密度建议时,系统不仅检索对应深度段的地层压力预测结果,还结合相邻井资料、地层岩性描述及相关设计规范等信息,为生成模型提供较为完整的证据上下文。相较于仅依赖模型内部参数知识的闭卷生成方式,该检索机制能够降低模型对参数化记忆的依赖,并支持钻井地质设计规范、区块资料和工程参数的动态更新,无需在知识发生变化后重复训练大语言模型。

本研究的重点并非单纯扩充钻井领域知识条目,而是将检索证据、领域任务指令与生成模型整合至统一的推理链路中。在外部证据约束下,IctLM-Drill 能够完成地质设计问答、工程参数提取及结构化实体关系生成等任务,从而在一定程度上降低术语误用、参数遗漏和回答泛化等问题。

3.3. ICT-DrillGeoDesign 数据集构建

ICT-DrillGeoDesign 数据集构建为增强模型对钻井地质设计任务的指令跟随能力,构建了面向该领域的指令微调数据集 ICT-DrillGeoDesign。该数据集以钻井地质设计报告、区块地质资料、钻井工程规范、行业标准文件及典型井史案例等文档为基础,覆盖地层岩性识别、压力预测、钻井液设计、井身结构设计、套管下深确定、油气层保护及事故风险分析等任务类型。数据集构建流程如图 2 所示。

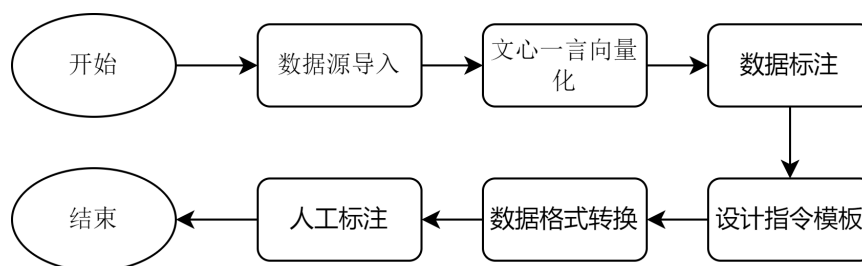


Figure 2. Construction pipeline of the instruction-tuning dataset

图 2. 指令数据集构建流程

3.3.1. 数据来源

数据来源包括专有钻井地质设计文档与公开中文工程技术语料两部分。专有文档包含钻井地质设计报告、钻井工程规范、区块地质资料和行业标准文件等共计 1500 篇文档。公开技术语料选用与地质、能源、工程技术和规范文本相关的中文语料作为补充[14],以增强模型的语言泛化能力和工程表达能力。数据集在进入训练流程前首先进行脱敏处理,然后进行数据清洗、章节切分和片段规范化流程,保证数据的安全性及样本质量。

3.3.2. 标注流程

ICT-DrillGeoDesign 采用模型预标注、人工复核以及指令化转换的构建流程。首先,利用通用大语言模型依据预设标签体系,对原始文档中的实体、工程参数及其关联关系进行初步标注。随后,由具备钻井地质设计相关知识背景的标注人员对预标注结果进行补充、修正和一致性校验,重点核查地层名称、深度范围、压力类型、钻井液参数、套管层次及油气层保护措施等关键信息。最后,将经审核的结构化标注结果转换为指令与回答的形式的训练样本,使模型能够学习面向知识问答、参数提取和实体关系生成等任务的统一输出方式。图 3 展示了 ICT-InstructTech 的构建与 IctLM 的训练流程。

3.3.3. 标签体系

基于实体关系抽取的主流技术框架[15][16],本文将钻井地质设计领域中的核心知识要素抽象为实体类型与关系类型,并采用统一的 JSON 输出模式对模型结果进行规范化表示。实体标签主要包括地层、岩性、压力、钻井液、套管、井身结构、油气层保护措施及设计参数等。关系标签主要描述不同实体之间

的工程关联，例如地层对应岩性、地层对应压力参数、压力预测约束钻井液密度、套管层次对应下深范围等。具体标签体系如表 1 所示。

ICT-InstructTech 构建与 IctLM 训练流程

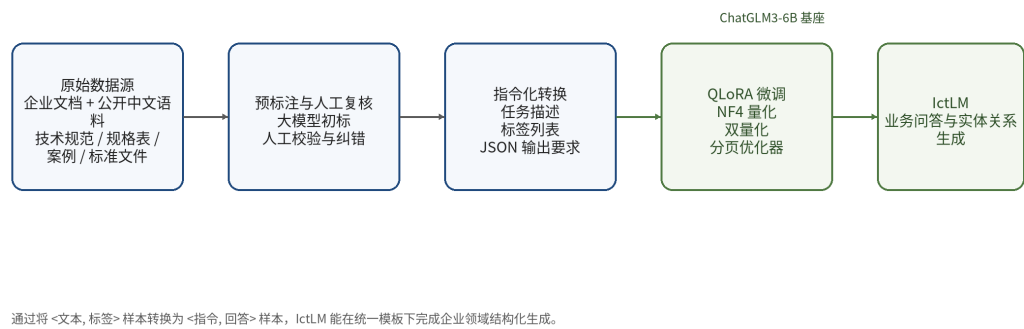


Figure 3. Construction of ICT-InstructTech and training pipeline of IctLM
图 3. ICT-InstructTech 构建与 IctLM 训练流程

Table 1. Label taxonomy of ICT-InstructTech
表 1. ICT-InstructTech 标签体系

层级	标签	说明	层级	标签	说明
实体类型	Scene (场景)	描述业务应用发生的行业场景或任务背景。	关系类型	solution_of	表示方案与任务/场景之间的对应关系。
实体类型	Solution (方案)	描述针对某类业务问题的技术方案、方法或实现路径。	关系类型	requires	表示方案/方法对原理、条件或关键资源的依赖关系。
实体类型	Metric (指标)	描述性能指标、约束条件或评估标准。	关系类型	has_result	表示方案/方法带来的结果或效果。
实体类型	Result (结果)	描述实施后的效果、输出或结论。			

ICT-DrillGeoDesign 的标签体系涵盖以下核心实体和关系，实例化标签体系如表 2 所示：

Table 2. Instantiation of the ICT-InstructTech label taxonomy
表 2. 实例化 ICT-InstructTech 标签体系

层级	标签	说明	层级	标签	说明
实体类型	Formation (地层)	描述钻井设计中涉及的地层单元，如沙河街组、东营组等。	关系类型	has_lithology	表示地层与岩性之间的对应关系。
实体类型	Lithology (岩性)	描述地层岩石类型及特征，如砂岩、泥岩、灰岩、白云岩等。	关系类型	has_pressure	表示地层、深度段与压力参数之间的关系。
实体类型	Pressure (压力)	描述孔隙压力、破裂压力、坍塌压力等压力参数。	关系类型	requires_fluid	表示地层或压力条件对钻井液类型及性能参数的要求。
实体类型	DrillingFluid (钻井液)	描述钻井液类型、密度、黏度、滤失量等设计内容。	关系类型	casing_depth	表示套管层次与设计下深之间的关系。
实体类型	Casing (套管)	描述导管、表层套管、技术套管、生产套管等套管层次	关系类型	protects_zone	表示保护措施与目标油气层或敏感地层之间的关系。

续表

实体类型	WellStructure (井身结构)	描述二开、三开、四开等井身结构设计方案。	关系类型	depth_range	表示地层、压力或工程参数对应的深度范围。
实体类型	ProtectionMeasure (保护措施)	描述油气层保护、井壁稳定、防漏堵漏等措施。	关系类型	has_result	表示方案实施后的结果、效果或风险控制结论。
实体类型	DesignParameter (设计参数)	描述井深、井径、套管下深、压力系数、密度窗口等参数。			

3.3.4. 指令类型分类

组织训练样本时，本文将传统的文本、标签监督方式转换为指令、回答形式。每个指令样本由任务描述、输入文本、标签约束和输出要求四部分构成，以明确模型的任务目标、输入范围及结果格式。对于自然语言问答任务，模型需结合检索证据生成符合钻井地质设计规范的专业回答；对于实体关系生成任务，模型需按照预定义的 JSON 模式输出识别出的实体及其关系；对于参数提取任务，模型需从文本中提取地层、深度段、压力参数、钻井液性能参数和套管参数等关键信息。

通过统一的指令化表示方式，IctLM-Drill 能够在学习钻井地质设计领域知识表达的同时，掌握不同任务对应的结构化输出规则，从而提升模型在知识问答、参数提取和实体关系生成等任务中的一致性与可用性。示例模板如下：(1) 地质特征查询：如“某地层的岩性特征是什么？”(2) 压力参数提取：如“某深度段的孔隙压力系数是多少？”(3) 钻井液推荐：如“针对某地层应选用何种钻井液？”(4) 井身结构设计：如“某区块推荐采用何种井身结构？”(5) 油气层保护：如“如何保护某油气层？”(6) 规范标准查询：如“SY/T 5615-2004 对某项设计有何要求？”[17][18](7) 事故案例分析：如“某类事故的成因和预防措施是什么？”

3.4. 基于 QLoRA 的领域优化模型 IctLM-Drill

本文选择 ChatGLM3-6B 作为基座模型，该模型继承了 GLM 系列统一理解与生成任务的技术路线，具备优秀的中英文理解与生成能力，同时在对话能力、工具调用支持与部署门槛之间取得了较好平衡，适合作为企业级业务模型的基座。面向企业场景，直接进行全参数微调会带来较高的显存与训练成本，因此本文采用 QLoRA 进行参数高效微调，其核心结构如图 4 所示。

LoRA 的核心思想是对大型模型的权重矩阵进行低秩转换，仅训练新增低秩参数而冻结原模型权重，从而显著减少训练参数量。在涉及矩阵相乘的 Transformer 模块旁新增一条低秩适配通路。设预训练模型线性层的权重矩阵为 $W_0 \in R^{d \times k}$ ，其增量更新表示为 $\Delta W = BA$ ，其中 $B \in R^{d \times r}$ ， $A \in R^{r \times k}$ ，且 r 远小于 d 与 k 。经过缩放后，适配后的权重可写为：

$$h = w_0 + \frac{\alpha}{r} \cdot ABx$$

其中 α 为缩放因子，训练时保持 W_0 冻结，仅优化低秩矩阵 A 和 B 的参数，大幅降低训练数量与显存开销。

与全参数微调相比，LoRA 只需训练低秩矩阵中的少量参数即可完成任务迁移。QLoRA 在 LoRA 的基础上引入三项核心优化：(1) 4-bit NormalFloat (NF4)量化，针对正态分布权重实现信息最优的 4-bit 量化；(2) 双量化(Double Quantization)，对量化常量再次量化，进一步降低内存占用；(3) 分页优化器(Paged Optimizers)，通过 CPU-GPU 分页内存管理避免训练时的显存溢出。最终实现了在单张 T4 GPU 上完成 6 B 规模模型的高质量微调，不同微调方法的显存占用对比如图 5 所示。基于这一策略，本文得到领域化生成模型 IctLM，并将其用于两类任务：一是面向科技文档的实体关系生成，二是面向系统对话场景的业务问答。

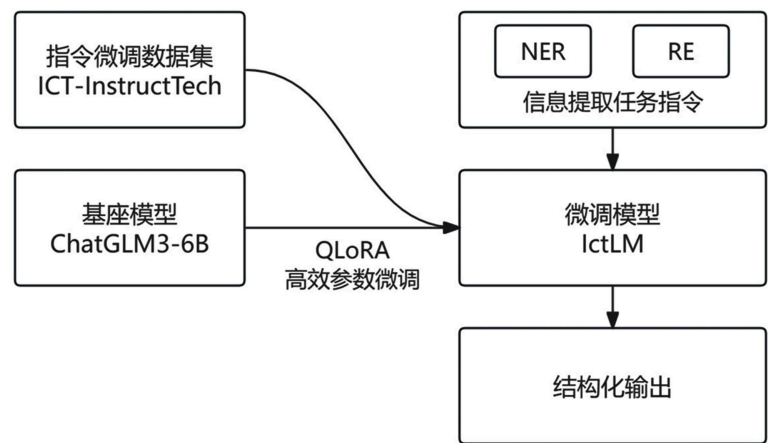


Figure 4. Core architecture of QLoRA fine-tuning
图 4. QLoRA 微调核心结构示意图

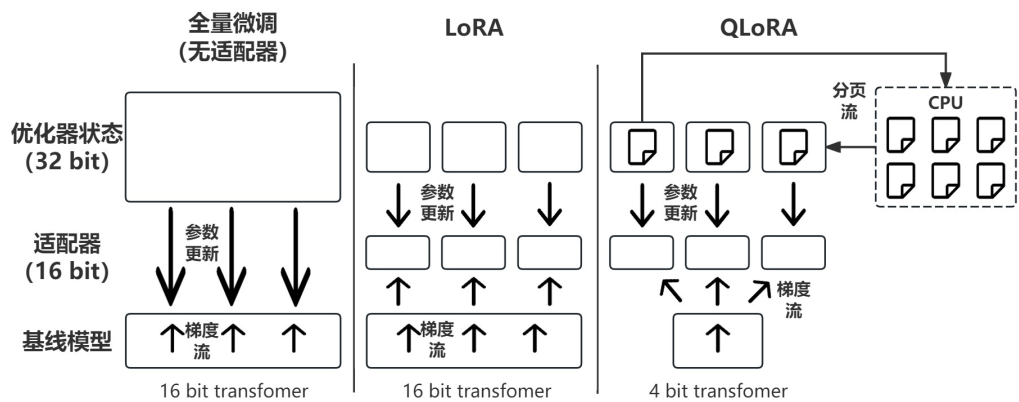


Figure 5. Memory requirements of different fine-tuning methods
图 5. 不同的微调方法及其内存需求

本文在 ICT-DrillGeoDesign 数据集上对 ChatGLM3-6B 进行领域指令微调，使模型学习钻井地质设计中的专业术语、参数关系、规范表达方式和结构化输出模式，最终得到面向钻井地质设计业务的领域生成模型 IctLM-Drill。该模型主要面向三类任务开展训练和推理。

第一类为钻井地质设计知识问答任务。模型结合检索模块返回的领域证据，对地层特征、压力预测、钻井液设计、井身结构及规范要求等问题进行回答。第二类为领域参数提取任务，主要从钻井地质设计报告、工程规范及相关技术资料中识别并抽取地层名称、深度区间、压力参数、钻井液密度、套管下深等关键信息。第三类为结构化实体关系生成任务，即按照预定义的 JSON 模式，输出地层、岩性、压力、钻井液、套管及油气层保护措施等实体，并识别实体之间的工程关联关系。

在推理过程中，系统首先根据不同任务类型预设相应的 Prompt 模板，再将用户输入与 Ict-Retriever 返回的检索证据共同输入 IctLM-Drill。对于业务问答任务，Prompt 重点约束用户问题与检索证据以及生成回答之间的对应关系，引导模型基于外部证据生成专业回复。对于实体关系生成任务，Prompt 进一步给出实体标签集合、关系标签集合及 JSON 输出格式要求，以提高结构化结果的一致性和规范性。

通过上述设计，IctLM-Drill 不仅能够生成与业务知识库内容相一致的自然语言回答，还能够输出可被后续业务系统直接解析和调用的结构化结果，从而满足钻井地质设计场景下知识问答、信息抽取和知识组织等多类应用需求。

4. 实验

4.1. 实验环境与数据划分

实验在配备 NVIDIA T4 GPU 的服务器上开展，操作系统为 Ubuntu 22.04，深度学习框架为 PyTorch 2.0.1，CUDA 版本为 11.7，Transformers 版本为 4.30.2。为了兼顾泛化评估与实现成本，本文采用固定的 8:1:1 比例划分训练集、验证集和测试集，而非多折交叉验证。数据集划分情况如表 3 所示。

Table 3. Partition of the experimental dataset

表 3. 实验数据集划分

数据集	训练集	验证集	测试集
ICT-InstructTech	1200	150	150
WuDaoCorpora 子集	1200	150	150
总计	2400	300	300

4.2. 对比模型与训练配置

本文设置三组基础生成模型进行实验对比，包括原始 ChatGLM3-6B 模型、经 LoRA 微调后的 ChatGLM3-6B 模型以及经过 QLoRA 微调后 IctLM 模型。在端到端问答评测中，进一步考察 Ict-Retriever 与 IctLM 的搭配实验效果。模型训练时采用 AdamW 优化器，并根据输入数据的长度动态调整输入与输出的最大 token 长度，确保技术文档中的较长文本被完全覆盖。实验过程的主要训练参数如表 4 所示。

Table 4. Key training hyperparameter settings

表 4. 主要训练参数设置

参数	设置值
微批次大小(batch_size)	16
梯度累积步数(gradient_accumulation_steps)	8
训练轮数(epochs)	30
学习率(learning_rate)	1e-5
优化器	AdamW
LoRA 秩(lora_r)	8
缩放因子(lora_alpha)	32
Dropout (lora_dropout)	0.1

4.3. 评估指标

本文采用 BLEU、PPL 和 Distinct 三个评价指标来综合评估模型的输出质量。BLEU 是机器翻译与文本生成任务中常用的核心评价指标，其通过比较生成文本与参考文本之间的 n-gram 重叠程度来衡量输出结果的相似性[19]。计算公式如下所示：

$$\text{BLEU} = \text{BP} \times \exp \left(\sum_{n=1}^N w_n \log p_n \right)$$

其中 BP 为长度惩罚因子, p_n 为 n-gram 精度, w_n 为对应权重, 本文采用主流的 BLEU-4 指标。

PPL (Perplexity) 反映语言模型在测试集上的困惑度, 数值越低表明模型对真实数据分布的拟合越好。计算公式如下所示:

$$\text{PPL}(D) = 2^{-\frac{1}{N} \sum_{i=1}^N \log P(w_i)}$$

其中, D 为测试数据集, N 为句子总数, $P(w_i)$ 为模型对句子 w_i 的预测概率。

Distinct 则通过统计生成文本中不同 n-gram 的比例来衡量输出的多样性, 数值越大代表生成结果越不容易陷入模式化回复[20], 计算公式如下所示:

$$\text{Distinct-n} = \frac{\sum_{i=1}^N n_{\text{unique}}}{\sum_{i=1}^N n_{\text{total}}}$$

其中, n_{unique} 为独特 n-gram 的数量, n_{total} 为 n-gram 的总数量。考虑到实体关系生成与业务问答需要同时具备正确性和可读性, 如果仅依赖自动指标并不足以全面反映模型效果。因此, 本文结合少量人工分析, 对输出语法的流畅性、上下文一致性以及业务表达准确性来进行辅助判断。

5. 结果与分析

5.1. 微调模型性能比较

Table 5. Performance comparison of baseline generative models

表 5. 基础生成模型性能比较

模型	BLEU	PPL	Distinct	Epoch
ChatGLM3-6B	0.302	10.06	6.15	30
ChatGLM3-6B (LoRA)	0.385	7.65	8.32	30
IctLM (QLoRA)	0.413	7.14	8.47	30

三组基础模型在统一数据集上的评测结果如表 5 所示。从结果可以看出, 经过 LoRA 微调后, ChatGLM3-6B (LoRA) 已在三项指标上全面优于原始 ChatGLM3-6B, 说明参数高效微调对企业领域任务具有明显增益。进一步采用 QLoRA 后, IctLM 在 BLEU、PPL 与 Distinct 上分别达到 0.413、7.14 和 8.47, 整体表现最佳。

以原始 ChatGLM3-6B 模型为基线, IctLM 在 BLEU、PPL 和 Distinct 指标上分别提升或优化约 36.8%、29.0% 和 37.7%。其中, BLEU 的提升表明模型生成结果与参考答案之间具有更高的一致性。PPL 的下降反映出模型对领域文本分布具有更好的拟合能力。Distinct 的提升则说明模型生成内容的词汇多样性和表达丰富性得到增强。上述结果说明, 结合领域指令数据与参数高效微调方法, 能够进一步提升模型在钻井地质设计场景下的生成质量和任务适配能力, 使其更适用于地层岩性分析、压力预测、钻井液设计及井身结构设计等专业任务。

训练过程中的 PPL 变化也反映出类似趋势, 如图 6 所示。根据训练日志, 在前 10 个 epoch 内, IctLM 的困惑度下降更平稳, 波动幅度小于另外两组模型, 说明量化参数高效微调在本任务中具有较好的收敛特性。综合自动指标可以认为, IctLM 已具备作为企业领域生成模型的可用性。

5.2. 检索 - 生成联动结果

为了验证检索增强与生成优化的协同效果, 本文进一步对 Ict-Retriever 与 IctLM 的联动模型进行端

到端问答评测，结果如表 6 所示。Ict-Retriever + IctLM (QLoRA)在三项指标上均取得最佳结果，BLEU 达到 0.612，PPL 为 6.76，Distinct 为 15.39。

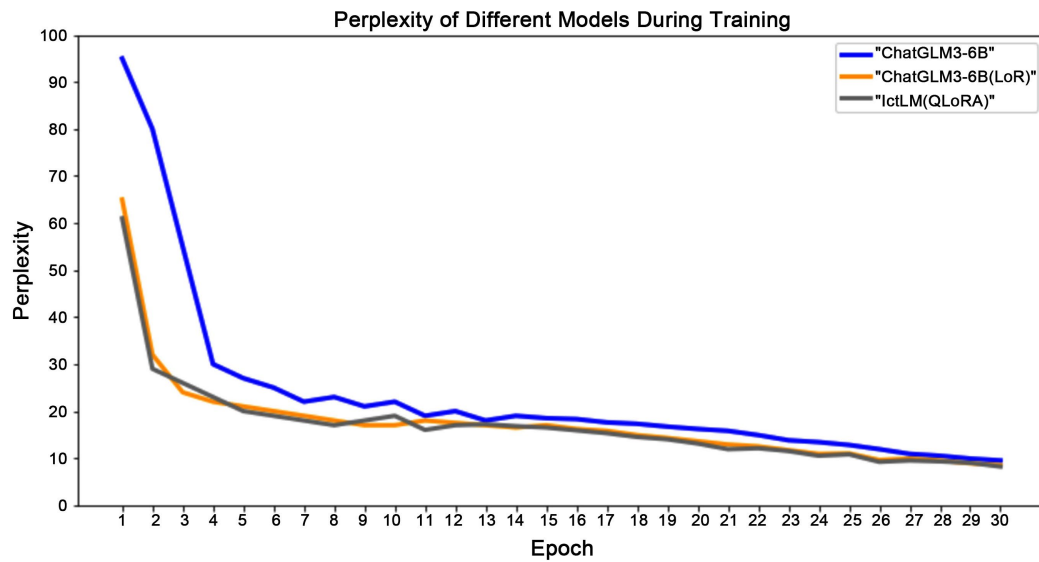


Figure 6. Training results of Perplexity for the three model groups

图 6. 三组模型 Perplexity 指标训练结果

Table 6. Performance comparison of end-to-end question answering models

表 6. 端到端问答模型性能比较

模型	BLEU	PPL	Distinct	Epoch
ChatGLM3-6B	0.394	6.87	8.16	30
ChatGLM3-6B (LoRA)	0.552	7.23	13.47	30
Ict-Retriever + IctLM (QLoRA)	0.612	6.76	15.39	-

为更直观地展示不同模型在 BLEU、PPL 和 Distinct 三项指标上的变化趋势，本文进一步绘制端到端问答模型性能对比图，如图 7 所示：

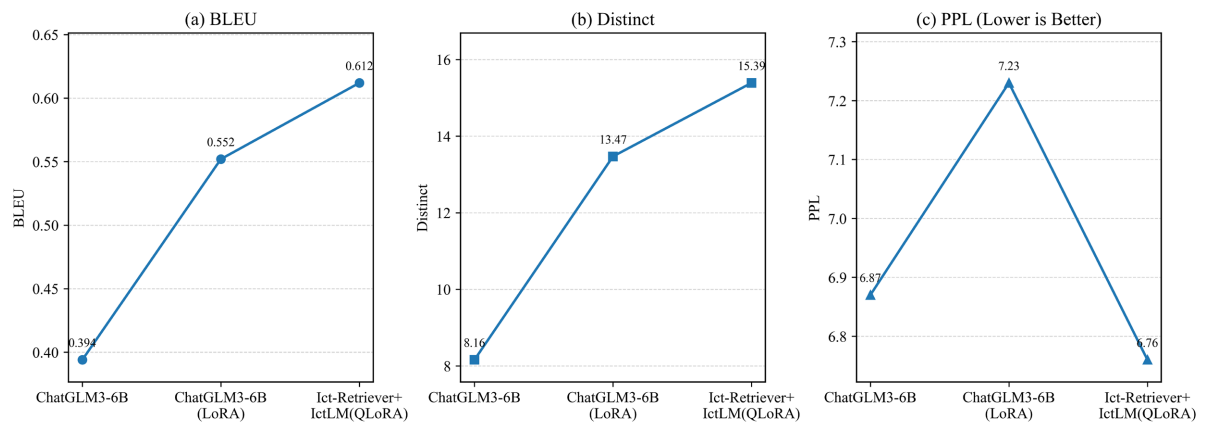


Figure 7. Performance comparison of end-to-end question answering models

图 7. 端对端问答模型性能比较

与表 5 中的 ChatGLM3-6B 基线模型相比,融合 Ict-Retriever 的端到端模型在 BLEU、PPL 和 Distinct 指标上分别提升或优化约 55.3%、1.6%和 88.6%。相较于 ChatGLM3-6B (LoRA)模型, BLEU、PPL 和 Distinct 指标亦分别进一步改善约 10.9%、6.5%和 14.3%。实验结果表明,检索模块与领域化生成模型之间具有较好的协同作用:检索模块能够为模型提供与当前问题相关的外部知识证据,缓解模型知识覆盖不足和知识更新滞后的问题;领域微调模型则能够更好地理解钻井地质设计中的专业术语、参数关系和任务输出要求,从而提升回答结果的准确性、专业性和表达质量。

在钻井地质设计知识问答场景中,检索模块的作用主要体现在以下三个方面:其一,能够从钻井地质设计报告、工程规范及相关技术资料中召回与用户问题相关的证据片段,为模型生成结果提供依据;其二,能够围绕地层、深度、压力、钻井液及井身结构等复合信息组织上下文,增强模型对关键参数及其关联关系的利用能力;其三,能够支持知识库内容的持续更新,使新增或修订后的企业私有资料无需重新参与模型训练即可被检索和使用,从而提高系统对最新设计资料和业务知识变化的适应能力。

通过少量人工分析,我们进一步发现, ChatGLM3-6B (LoRA)在语法和流畅性方面已具有较好表现,但在涉及多轮上下文和企业内部概念时仍会出现泛化式回答。Ict-Retriever 与 IctLM 组合在多轮对话过程中能够更好地利用检索到的业务证据,语义表达更准确,回复也更加自然、具体且具有说服力。由此可见,单独依赖检索或单独依赖微调都难以达到最优效果,二者的协同才是企业知识问答场景的关键。

5.3. 消融实验

为验证本文核心模块的独立贡献,本文开展消融实验,结果如表 7 所示。同时为更直观地比较不同模块对模型性能的影响,本文进一步绘制核心模块消融实验结果图,如图 8 所示。

Table 7. Ablation results of core modules
表 7. 核心模块消融实验结果

模型	BLEU	PPL	Distinct
ChatGLM3-6B	0.394	6.87	8.16
仅 QLoRA 微调	0.358	8.22	7.63
仅 ICT-InstructTech 微调	0.376	7.89	7.95
IctLM(QLoRA)	0.413	7.14	8.47
Ict-Retriever + IctLM (QLoRA)	0.612	6.76	15.39

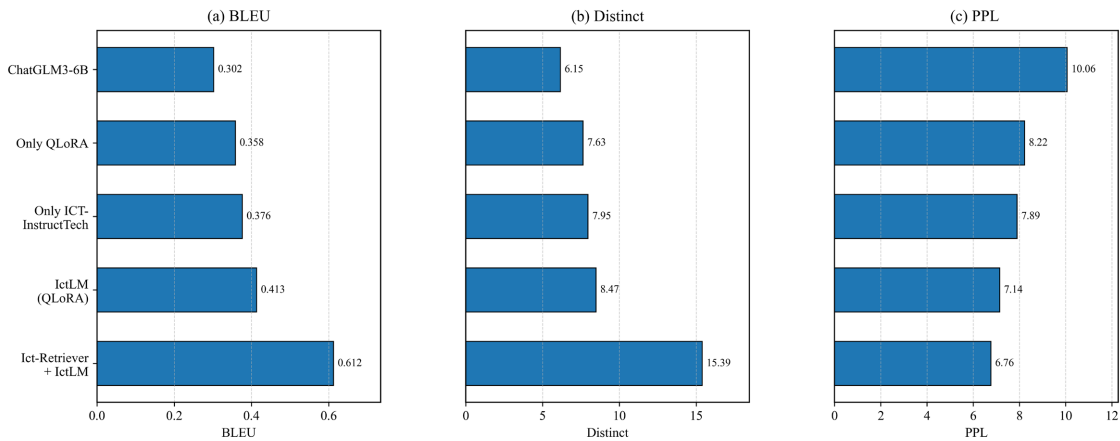


Figure 8. Comparison of ablation experiment results for different model variants
图 8. 模型消融实验结果对比

消融实验结果表明, QLoRA 微调策略与 ICT-InstructTech 领域指令数据集均能够对模型性能产生积极影响。与基础 ChatGLM3-6B 相比, 仅采用 QLoRA 微调后, 模型的 BLEU 由 0.302 提升至 0.358, PPL 由 10.06 降至 8.22, Distinct 由 6.15 提升至 7.63, 说明参数高效微调能够在较低训练成本下改善模型的语义匹配能力、语言建模能力和回复多样性。仅采用 ICT-InstructTech 数据集进行领域适配时, 模型的 BLEU、PPL 和 Distinct 分别达到 0.376、7.89 和 7.95, 整体表现略优于仅使用 QLoRA 微调的模型。该结果表明, 高质量领域指令数据能够有效增强模型对钻井地质设计专业术语、典型问答模式及参数关联关系的理解能力。相较于仅调整模型参数, 领域指令数据能够为模型提供更贴近实际业务场景的任务描述、输入语境和目标输出形式, 从而提升其对领域语义和任务要求的适应能力。

在此基础上, 完整的 IctLM (QLoRA) 模型进一步结合参数高效微调策略与 ICT-InstructTech 领域指令数据集, 在各项评价指标上均优于仅采用单一优化模块的模型。其中, BLEU 提升至 0.413, PPL 降至 7.14, Distinct 提升至 8.47。该结果说明, QLoRA 微调与领域指令数据之间并非简单的性能叠加, 而是具有一定的协同作用。一方面, QLoRA 为有限计算资源条件下的大语言模型领域适配提供了较为高效的训练方式。另一方面, ICT-InstructTech 为模型提供了与企业知识问答场景相匹配的指令样本、专业表达方式及结构化任务模式。二者结合后, 模型能够更好地学习钻井地质设计领域的知识表达和任务响应方式, 从而提高回答质量。

进一步引入 Ict-Retriever 后, Ict-Retriever 与 IctLM (QLoRA) 的组合在全部评价指标上取得最优结果, BLEU 达到 0.612, PPL 降至 6.76, Distinct 提升至 15.39。与完整 IctLM (QLoRA) 相比, 检索增强带来了更为显著的性能提升, 尤其在 BLEU 和 Distinct 指标上的改善较为明显。这表明, 检索模块能够为生成模型补充与用户问题相关的外部知识证据, 缓解企业知识问答中常见的知识覆盖不足、知识更新滞后及回答内容泛化等问题。在检索证据的支撑下, 模型生成结果在语义准确性、信息覆盖度和表达多样性等方面均得到进一步提升。

总体而言, 单独采用参数高效微调或领域指令数据适配, 均能够在一定程度上改善基础模型在钻井地质设计场景中的表现, 但其提升幅度仍受到模型内部知识覆盖范围的限制, 难以充分应对知识密集、术语复杂及上下文依赖较强的企业知识问答任务。相比之下, 检索增强机制与领域化生成模型的协同结合, 能够同时发挥外部知识补充和领域表达优化的作用: 前者为模型提供可动态更新的业务证据, 后者提升模型对专业术语、参数关系和任务输出要求的理解能力, 从而使系统能够生成更准确、更专业且更符合企业业务语境的回答。因此, 消融实验进一步验证了本文所提出的检索增强与领域生成优化协同框架的有效性, 也说明该协同机制是提升企业知识问答系统整体性能的关键因素。

6. 结论

本文针对通用大语言模型在钻井地质设计知识问答中存在领域知识不足、工程参数生成不准确及回答依据难以追溯等问题, 提出一种检索增强与生成优化协同方法。该方法以 Ict-Retriever 为检索前端, 对钻井地质设计报告、区块资料、工程规范和行业标准进行清洗、语义切片、元数据标注与向量化处理, 构建领域知识库; 同时, 基于 ICT-DrillGeoDesign 数据集, 采用 QLoRA 对 ChatGLM3-6B 进行领域适配, 得到 IctLM-Drill。系统通过统一提示模板组织用户问题、检索证据和输出约束, 可完成知识问答、关键参数提取及结构化实体关系生成等任务。

实验结果表明, 领域指令微调与检索增强具有互补作用。IctLM-Drill 能够学习钻井地质设计中的专业术语、参数关系和结构化输出规则, 提高模型对领域任务的理解与响应能力; Ict-Retriever 则将设计报告、规范文件和区块资料中的相关证据引入生成过程, 减少仅依赖参数化知识造成的知识遗漏、更新滞后和无依据生成。统一提示模板进一步明确了问题、证据与输出目标之间的对应关系, 提高了模型对检

索内容的利用效率。与基础模型相比,完整系统在实体关系抽取、问答事实正确性及文本生成辅助指标方面总体表现更优,验证了所提方法的有效性。

从应用角度看,该方法可为设计资料查询、规范问答、地层参数抽取、钻井液设计辅助及井身结构方案检索提供支持,并促进分散文档的统一管理与知识复用。检索证据能够增强回答的可追溯性,结构化实体关系结果也可后续知识组织提供数据基础。但本文尚未在真实生产环境中量化验证其对设计效率、错误率和决策质量的影响,相关工程价值仍需进一步评估。

本文仍存在一定局限。现有实体关系体系主要覆盖常见实体及直接关系,对跨井、跨区块、跨文档和多跳关联知识的表达能力有限;当前检索流程主要依赖向量相似度与元数据约束,尚未引入显式重排序、证据压缩和复杂查询分解,在长尾问题和多条件问题下仍可能受到初始召回质量的影响;现有测试数据和人工评测规模相对有限,模型在不同区块及真实业务环境中的泛化能力与稳定性仍需进一步验证。此外,当前系统仅实现实体和关系三元组输出,尚未完成实体消歧、关系融合、图存储及图查询,因此知识图谱不作为本文已实现和评估的功能。

后续研究将重点开展检索重排序、证据压缩和多跳检索,提升复杂问题下的证据召回与利用能力;在现有实体关系抽取结果基础上构建领域知识图谱,并通过独立对比与消融实验验证其实际增益;同时融合测井曲线、岩心照片等多模态资料,探索实时钻井数据接入,并在更多油田区块开展应用验证,以提高系统的事实可靠性、泛化能力和工程适用性。

基金项目

项目克拉玛依市科技计划资助(项目编号: 2025BA0066)。

参考文献

- [1] Gao, Y., Xiong, Y., Gao, X., *et al.* (2023) Retrieval-Augmented Generation for Large Language Models: A Survey. *arXiv:2312.10997*.
- [2] 熊凯, 杜理, 丁效, 刘挺, 秦兵, 付博. 面向文本推理的知识增强预训练语言模型[J]. *中文信息学报*, 2022, 36(12): 27-35.
- [3] 袁乐, 刘绍华, 王禹, 等. 大语言模型检索增强生成优化技术研究综述[J]. *计算机学报*, 2026, 49(2): 383-422.
- [4] Lewis, P., Perez, E., Piktus, A., *et al.* (2020) Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *2020 Advances in Neural Information Processing Systems*, Vancouver, 6-12 December 2020, 9459-9474.
- [5] 张犬俊, 谢杨, 房春荣, 等. 检索增强生成在软件工程中的应用综述[J]. *软件学报*, 2026, 37(3): 1316-1339.
- [6] Hu, E.J., Shen, Y., Wallis, P., *et al.* (2022) LoRA: Low-Rank Adaptation of Large Language Models. *2022 International Conference on Learning Representations*, Online, 25-29 April 2022, 1-26.
- [7] Dettmers, T., Pagnoni, A., Holtzman, A. and Zettlemoyer, L. (2023) QLoRA: Efficient Finetuning of Quantized LLMs. *Advances in Neural Information Processing Systems* 36, New Orleans, 10-16 December 2023, 10088-10015. <https://doi.org/10.52202/075280-0441>
- [8] Du, Z., Qian, Y., Liu, X., Ding, M., Qiu, J., Yang, Z., *et al.* (2022) GLM: General Language Model Pretraining with Autoregressive Blank Infilling. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Dublin, May 2022, 320-335. <https://doi.org/10.18653/v1/2022.acl-long.26>
- [9] Team GLM, Zeng, A., Xu, B., Wang, B., *et al.* (2024) ChatGLM: A Family of Large Language Models from GLM-130B to GLM-4 All Tools. *arXiv:2406.12793*.
- [10] Wang, Y., Kordi, Y., Mishra, S., Liu, A., Smith, N.A., Khashabi, D., *et al.* (2023) Self-Instruct: Aligning Language Models with Self-Generated Instructions. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Toronto, July 2023, 13484-13508. <https://doi.org/10.18653/v1/2023.acl-long.754>
- [11] Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., *et al.* (2020) Dense Passage Retrieval for Open-Domain Question Answering. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online, November 2020, 6769-6781. <https://doi.org/10.18653/v1/2020.emnlp-main.550>
- [12] 冯掌印. 浅谈大模型时代下的检索增强[C]//中国计算语言学大会. 2024.

-
- [13] 王亮. 检索增强生成(RAG)驱动的知识服务: 原理、范式及评估[J]. 科技与出版, 2025(4): 37-46.
- [14] Yuan, S., Zhao, H., Du, Z., Ding, M., Liu, X., Cen, Y., *et al.* (2021) WuDaoCorpora: A Super Large-Scale Chinese Corpora for Pre-Training Language Models. *AI Open*, **2**, 65-68. <https://doi.org/10.1016/j.aiopen.2021.06.001>
- [15] 李冬梅, 张扬, 李东远, 等. 实体关系抽取方法研究综述[J]. 计算机研究与发展, 2020, 57(7): 1424-1448.
- [16] 鄂海红, 张文静, 肖思琪, 等. 深度学习实体关系抽取研究综述[J]. 软件学报, 2019, 30(6): 1793-1818.
- [17] 中华人民共和国石油天然气行业标准. SY/T 5615-2004 石油天然气地质编图规范及图式[S]. 北京: 石油工业出版社, 2004.
- [18] 中国石油集团地质研究院. 钻井地质设计手册[M]. 北京: 石油工业出版社, 2020.
- [19] Papineni, K., Roukos, S., Ward, T., *et al.* (2002) BLEU: a Method for Automatic Evaluation of Machine Translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, 7-12 July 2002, 311-318.
- [20] Li, J., Galley, M., Brockett, C., Gao, J. and Dolan, B. (2016) A Diversity-Promoting Objective Function for Neural Conversation Models. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, San Diego, June 2016, 110-119. <https://doi.org/10.18653/v1/n16-1014>