

基于 BERT 的领域知识问答系统实践研究

白毅平

(广东海启星海洋科技有限公司)

摘要: 领域知识问答系统在离线运行、数据不出域及长期调用成本等方面存在实际约束。当前主流方案倾向于采用大参数量的生成式模型,但本文认为,对于领域专用任务,大模型中大部分参数用于编码通用语言知识,真正服务于特定领域的有效参数占比极低,构成显著的模型冗余。本文以海洋领域知识服务为背景,在端侧大语言模型(LLM)实验的基础上,回归判别式架构,采用“文本分类+预设答案”思路,使用仅约 1.02 亿(102M)参数的 bert-base-chinese 构建可离线部署的问答系统。实验先在国产 ARM 开发板上对比了 Ollama CPU 推理与 RKLLM NPU 推理的性能,发现端侧 LLM 在延迟、幻觉及资源占用方面存在局限,进而转向 BERT 分类方案。本文强调“领域即专业”,训练数据来源于国家标准、行业规范及企业知识,确保答案的权威性。围绕知识提取、训练数据构造、模型训练与部署验证,形成了一套可复现的工程流程,并针对多源权威文档中同一概念存在不同表述的语义一致性问题,提出了基于分类置信度分析的训练数据去重方法,以“海温”短问题为例,置信度由 49.85% 提升至 94.16%,验证集准确率达 96.79%。对比实验表明,该方案无需 GPU,可在 x86 与 ARM 架构的 CPU 设备上流畅运行(x86 平台推理延迟约 87ms,ARM 平台约 307ms),是本地化 AI 部署的有益实践探索,在离线环境、涉密场景和标准问答等任务中具备实用价值,验证了“合适的模型比更大的模型更重要”这一选型思路。

关键词: 人工智能; BERT; 知识问答; 文本分类; 训练数据清理; 边缘部署

中图分类号: TP391

Practical Research on BERT-based Domain Knowledge Question Answering System

Bai Yiping

(Guangdong Haiqixing Marine Technology Co., Ltd.)

Abstract: Domain knowledge question answering systems face practical constraints in offline operation, data localization, and long-term invocation costs. Current mainstream solutions tend to adopt large-parameter generative models. However, for domain-specific tasks, this paper argues that most parameters in large models are used to encode general linguistic knowledge, while the proportion of parameters truly serving a specific domain is extremely low, resulting in significant model redundancy. Taking marine domain knowledge services as the background, this paper returns to the discriminative architecture based on experiments with edge-side Large Language Models (LLM), and adopts a “text classification + preset answers” approach, using bert-base-chinese with only about 102 million (102M) parameters to build an offline-deployable question answering system. Experiments first compared the performance of Ollama CPU inference and RKLLM NPU inference on domestic ARM development boards, revealing limitations of edge-side LLMs in latency, hallucination, and resource occupation, leading to the shift toward the BERT classification scheme. This paper emphasizes “domain means specialty”: the training data originate from national standards, industry norms, and enterprise knowledge to ensure authority. A reproducible engineering pipeline covering knowledge extraction, training-data construction, model training, and deployment verification is established. A confidence-analysis-based training-data deduplication method is proposed for the semantic-consistency problem where the same concept is expressed differently across multiple authoritative sources; taking the short question “sea temperature (海温)” as an example, the confidence increased from 49.85% to 94.16%, and the validation accuracy reached 96.79%. Comparative experiments show that the proposed scheme requires no GPU and runs smoothly on x86 and ARM CPUs (approximately 87ms on x86 and 307ms on ARM), serving as a useful exploration of localized AI deployment with practical value in

作者简介: 白毅平(1977—),男,高级工程师,首席科学家,主要研究方向为防灾减灾、人工智能技术应用、企业智能化。E-mail: 2727100@qq.com

offline, confidential, and standard-question-answering scenarios, validating the design philosophy that “a suitable model matters more than a larger one.”

Key words: AI; BERT; knowledge question answering; text classification; training data cleaning; edge deployment

0 引言

在海洋、气象等行业的业务咨询场景中，用户常以简短自然语言提问，如“海温”“潮汐怎么查”等。这类问题答案须依据行业标准与权威文献，对准确性要求极高。然而，现有大语言模型（LLM）与检索增强生成（RAG）方案^[1-2]在离线运行、数据不出域及长期成本等方面存在约束；小参数生成式模型（0.8B~1.5B）对提示词遵从性差，输出难以控制。此外，海上浮标、海岛站点、岸基野外站点等场景网络带宽有限且不稳定，难以支撑云端大模型的实时调用；涉密单位等场景数据不能外发。

当前业界普遍追求大参数量，但对于领域专用任务，大模型中绝大部分参数用于编码通用语言常识，真正服务于特定领域的有效参数占比极低，构成显著的模型冗余。这种冗余不仅带来推理延迟与资源消耗，也增加了幻觉风险[3]。因此，对于答案固定、边界清晰的判别式任务，轻量化判别式模型往往比生成式大模型更为合适。

本文前期在国产 ARM 开发板上开展了端侧 LLM 实验，发现端侧 LLM 在延迟、幻觉及资源占用方面存在局限。在此基础上，本文回归判别式架构，采用 BERT 分类方案，以“文本分类+预设答案”思路构建可离线部署的领域知识问答系统。该方案无需 GPU，可在 CPU 服务器、台式机、笔记本等海量存量设备上流畅运行，为企业事业单位利用现有硬件实现本地化 AI 部署提供了可行路径。训练数据来源于国家标准、行业规范及企业知识。主要贡献包括：(1) 建立了可复现的“知识文档—结构化条目—训练数据—分类模型”工程流程；(2) 针对多源权威文档的语义一致性问题，提出了基于分类置信度分析的训练数据去重方法；(3) 通过端侧实验与多维度对比，验证了“合适的模型比更大的模型更重要”这一选型思路，是本地化 AI 部署的有益实践探索。

1 端侧模型选型实验

1.1 实验环境

为探索端侧 AI 的可行性，本文前期在 Orange Pi 5 Plus 与 Firefly AIO-3588Q（均搭载瑞芯微 RK3588，4GB 内存）两款开发板上开展了系列实验。RK3588 配备 6 TOPS NPU 算力，支持 Ollama CPU 推理与 RKLLM NPU 推理两种模式。

1.2 端侧 LLM 实验

实验对比了三种推理方式（CPU、NPU、云端）及不同参数规模模型的性能，结果如表 1 所示。

表 1 端侧 LLM 推理性能对比

Tab. 1 Comparison of edge-side LLM inference performance

模型	推理方式	TTFT	生成速度	内存占用
Qwen2.5-1.5B	Ollama CPU	1.6 s	17 T/s	1.3 GB
Qwen2.5-3B	Ollama CPU	2.5~3.6 s	10 T/s	2.3 GB
Qwen3.5-0.8B	RKLLM NPU	~0.5 s	~30 T/s	1.4 GB
DeepSeek API	云端	~2 s	90~116 T/s	—

85 实验发现以下问题：(1) CPU 推理延迟较高^[4]，1.5B 模型首 token 延迟达数秒；(2) NPU 推理速度较 CPU 提升约 3 倍^[5]，但仍低于云端 API；(3) 小参数模型对提示词遵从性较差，格式约束与角色设定容易丢失；(4) LLM 生成式输出存在幻觉风险，答案不可控。上述问题与行业研究报告的观察结论基本一致^[6]。

1.3 方案转向

90 基于上述实验，端侧 LLM 在标准问答场景中存在局限：延迟较高、提示词遵从性差、答案不可控、资源占用大。而 BERT 分类方案无需提示词，具有推理快速（~100 ms）、答案固定、无幻觉等优势，更适合封闭域知识问答。因此，本文转向 BERT 方案，后续实验均基于此展开。

2 BERT 模型概述

95 2.1 发展历程

 BERT 由 Devlin 等于 2018 年提出，相关论文于 2019 年在 NAACL 发表^[7]。其核心做法是“预训练+微调”：先在大规模语料上学习通用的语言表示，再在少量标注数据上针对具体任务进行微调^[8]。这一范式后被 GPT 等系列模型广泛采用。

100 BERT 的名称来源于“Bidirectional Encoder Representations from Transformers”。从模型结构看，BERT 使用 Transformer 的 Encoder 部分，属于 Encoder-only 架构；GPT 系列则使用 Decoder-only 架构，两者分别侧重理解与生成任务。基于 Encoder 的分支衍生出 RoBERTa、DistilBERT、TinyBERT 等模型，基于 Decoder 的分支则衍生出 GPT 系列等生成式模型。

2.2 核心原理

105 （1）**架构**：BERT 采用 Transformer 的 Encoder 结构，能够同时利用上下文信息进行理解，适合分类、匹配等任务。

 （2）**预训练任务**：主要包括掩码语言模型（MLM）和下一句预测（NSP）。MLM 随机遮盖部分词元，由模型根据上下文进行预测；NSP 判断两个句子是否在原文中相邻。

110 （3）**模型规格**：BERT-Base 为 12 层、约 1.02 亿（102M）参数、隐藏层维度 768；BERT-Large 为 24 层、约 3.4 亿参数、隐藏层维度 1024。

2.3 BERT 与 GPT 的对比

BERT 与 GPT 在架构定位上的主要差异如表 2 所示。

表 2 BERT 与 GPT 架构对比
Tab. 2 Comparison between BERT and GPT architectures

维度	BERT (Encoder-only)	GPT (Decoder-only)
注意力方向	双向（上下文）	单向（上文）
主要任务	理解（分类、检索、匹配）	生成（对话、写作、翻译）
参数规模	约 1.02~3.4 亿	数十亿至万亿级
扩展方式	相对有限	易于扩展
典型应用	搜索、文本分类	ChatGPT、Claude、Qwen

2.4 轻量化变体

为了在资源受限设备上部署，BERT 衍生出多个轻量化版本：DistilBERT^[9] 参数减少约 40%、推理速度提升约 60%，同时保留约 97% 的性能；TinyBERT^[10] 约 1,450 万参数，适合移动端；MobileBERT^[11] 针对移动设备优化，约 2,530 万参数。本文使用 bert-base-chinese^[12]（约 1.02 亿 / 102M 参数），在普通 PC 上通过 ONNX Runtime 进行 CPU 推理，无需 GPU。

3 系统架构

3.1 整体设计

系统采用“BERT 分类+预设答案”的架构：用户问题经 BERT 分类器判定所属知识类别，系统再返回该类别对应的预设答案。该设计将开放域问答转化为封闭域分类问题，答案来源固定、可人工审核，从而在流程上控制输出内容。

3.2 数据来源与构建

本文强调“领域即专业”，训练数据来源于三类权威资料：(1) 国家标准，如 GB/T 14914.2-2019《海洋观测规范 第 2 部分：海滨观测》、GB/T 39628—2020《海洋预报术语》；(2) 行业规范与科普文献，如台风科普知识、海洋灾害防御指南等；(3) 企业知识，包括公司基本信息、产品介绍、规章制度等。前两类确保专业性与权威性，第三类面向通用产品设计，拓展使用对象范围。

训练数据的生成链路如下：由 knowledge/*.md 原始知识文件经大模型辅助提取，得到 preset_qa.json 结构化问答条目；再为每个条目生成问法变体，形成 train.jsonl 与 val.jsonl 训练数据。其中 preset_qa.json 包含 82 个知识类别，每个类别定义类别名称、关键词、标准问法（10~20 条）以及预设答案。

3.3 问法变体生成

为覆盖用户问法的多样性，系统为每个类别生成 50 条问法变体。实际每条样本数约为标准问法（10~20 条）与变体（50 条）之和，即每类别约 60~70 条。最终训练集约 3,610 条样本，验证集约 903 条样本。

4 工程实践

4.1 模型训练

BERT 在中文文本分类任务中已有较多成功应用^[13-14]，在软件工程等领域的应用研究也在不断深入^[15]。模型训练的主要配置如表 3 所示。

表 3 模型训练主要配置
Tab. 3 Main configuration of model training

参数	数值
基座模型	bert-base-chinese
参数量	约 1.02 亿（102M）
类别数	82
训练轮数	5
批大小	32
学习率	5×10^{-5}
最大序列长度	64
验证准确率	96.79%

训练过程中，损失值在前 2 轮快速下降后趋于稳定，验证准确率在第 2 轮即达到 96% 以上并保持平稳，宏观 F1 值为 0.96。训练曲线如图 1 所示。

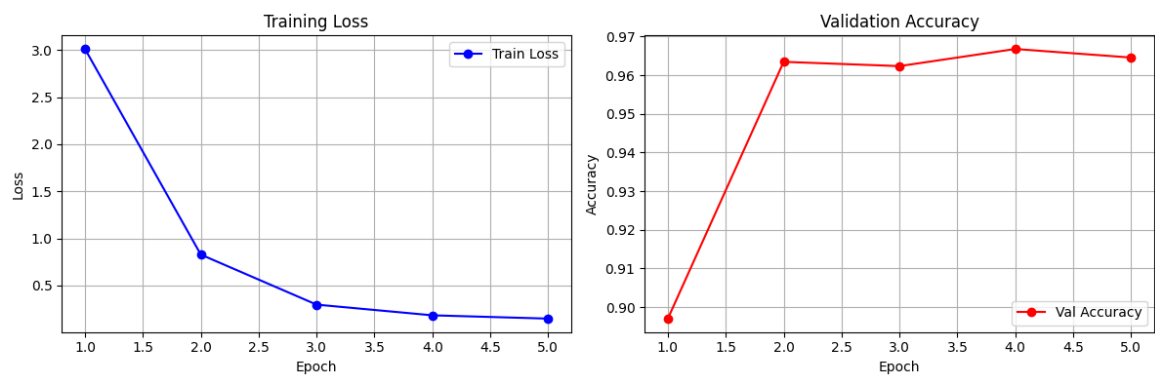


图 1 模型训练损失与验证准确率曲线

Fig. 1 Training loss and validation accuracy curves

4.2 部署方案

模型导出为 ONNX 格式^[16]，并使用 ONNX Runtime 在 CPU 上推理。为便于端到端测试，本文开发了 HXXedge 系统（HD001e v1.0.1），该系统基于 Flask 框架，集成 BERT

分类器、置信度阈值判断（阈值 0.40）、分级拒答逻辑及 SSE 流式响应，并提供推理延迟、
155 置信度等统计指标。部署环境为普通 PC，推理延迟约 100 ms。

4.3 问题发现与修复

部署后的测试中发现，“海温”“海水温度”“海表面温度”等短问题被拒答。通过调试日志定
位到具体原因：问题“海温”被分类为类别“海表面温度”（label=61），置信度仅为 0.4985，
低于预设阈值而被拒答。

160 检查 categories.json 发现，label 29（表层海水温度）与 label 61（海表面温度）在语义
上高度相近：两者答案均描述海表面约 0.5 m 水层内的温度，关键词均包含“海温”“海水温
度”等，且分别来源于《海洋观测规范》（GB/T 14914.2-2019）和《海洋预报术语》（GB/T
39628—2020）两份国家标准。由于从不同标准文档中提取时对同一概念未作合并，训练数
据中出现了语义重复，模型难以对短问题形成清晰的分类边界。

165 上述问题本质上属于训练数据清理范畴。本文提出的基于置信度分析的去重方法包含以
下步骤：(1) 构建包含短问题、歧义问题的代表性测试集；(2) 对测试集进行分类推理，筛
选置信度低于阈值的样本；(3) 分析低置信度样本的竞争类别，检测语义重叠；(4) 对语义
重复的类别进行合并，重建训练数据。

170 按上述流程，将 label 61 的关键词并入 label 29，合并两者答案，删除 label 61 并对
后续类别重新编号，类别总数由 83 减至 82。重新训练后，“海温”的分类置信度由 49.85%
提升至 94.16%，验证集准确率由 95.12% 提升至 96.79%。修复前后的对比如表 4 所示。

表 4 “海温”问题修复前后对比
Tab. 4 Comparison before and after fixing the “sea temperature” issue

问题	修复前	修复后
海温	拒答（49.85%）	正确（94.16%）
海表温度	拒答	正确（94.57%）
海水温度	拒答	正确（92.31%）

175 该案例揭示了一类具有普遍性的问题：当领域知识来源于多份标准文档、学术文献或实
测数据时，不同文献可能对同一概念使用不同术语（如“表层海水温度”与“海表面温度”），
或因研究进展对概念的界定发生调整。若未进行跨文档的语义去重与术语统一，训练集中的
语义噪声将直接干扰分类模型的决策边界，尤其影响短文本等低信息量输入的分类置信度。
该方法为多源权威文档的训练数据一致性校验提供了一种可操作的思路。

5 效果验证

180 5.1 整体性能

系统整体性能为：验证准确率 96.79%，内存占用约 400 MB。为验证方案的跨平台适

用性，本文分别在普通笔记本电脑（x86 CPU）与瑞芯微 RK3588 开发板（ARM CPU）上进行了对比测试，结果如表 5 所示。

185

表 5 跨平台推理性能实测
Tab. 5 Measured cross-platform inference performance

测试问题	平台	推理延迟	置信度
什么是 SST	笔记本（x86）	89 ms	72.3%
什么是 SST	开发板（ARM）	309 ms	72.3%
什么是海表温度	笔记本（x86）	86 ms	95.0%
什么是海表温度	开发板（ARM）	304 ms	95.0%

实测结果表明：

- 190
- （1）推理延迟：x86 平台约 87 ms，ARM 平台约 307 ms，后者约为前者的 3.5 倍。但 ARM 平台的 300 ms 延迟仍满足人机交互的实时性要求（< 500 ms），用户无明显等待感。

（2）模型一致性：同一问题在两个平台上的置信度完全相同（72.3% / 95.0%），验证了 ONNX 格式的跨平台特性与模型输出的确定性。

（3）术语敏感度：英文缩写“SST”的置信度（72.3%）显著低于中文全称“海表温度”（95.0%），反映了 BERT 模型对中文术语的识别能力优于英文缩写，这也为后续训练数据的术语规范化提供了改进方向。

195

图 2 展示了系统在瑞芯微 RK3588 开发板上的实际运行效果。



图 2 HXXedge 系统在瑞芯微 RK3588 开发板上的运行界面（浏览器访问 <http://<开发板 IP>:5001>）

Fig. 2 Running interface of the HXXedge system on Rockchip RK3588 development board (accessed via browser at <http://<board IP>:5001>)

200 综上，BERT 分类方案在 x86 与 ARM 两种主流 CPU 架构上均可流畅运行，无需 GPU 支持，验证了该方案在本地化部署场景下的普适性与可行性。

5.2 对比实验：BERT 方案与 LLM 方案

为比较两类方案的实际差异，本文在同一业务场景下，将前述 BERT 方案（HXXedge）与既有 LLM 方案（HD-001 海启星数字员工系统）进行对比，技术层面对比如表 6 所示。

205 表 6 两类方案技术对比

Tab. 6 Technical comparison of the two approaches

维度	HD-001（LLM 方案）	HXXedge（BERT 方案）
核心技术	DeepSeek API + RAG + Function Calling	ONNX 本地推理+分类
模型规模	数十亿参数级	约 1.02 亿（102M）参数
部署方式	云端 API 调用	本地 CPU 推理
网络依赖	需要联网	可完全离线
答案来源	模型生成	预设答案库

性能与成本对比如表 7 所示。

表 7 两类方案性能与成本对比

Tab. 7 Performance and cost comparison of the two approaches

指标	HD-001（LLM）	HXXedge（BERT）
推理延迟	2~5 秒	约 100 ms
单次成本	约 0.004~0.02 元	本地运行，无固定单次成本
月运营成本	数百至数千元（视调用量）	主要为硬件折旧
硬件投入	无需本地算力	普通 PC
输出偏差风险	存在	较低
答案可控性	较低（生成式）	较高（预设式）

210 安全防护能力对比如表 8 所示。

表 8 两类方案安全防护能力对比

Tab. 8 Security and protection comparison of the two approaches

防护维度	HD-001（LLM）	HXXedge（BERT）
输出长度限制	需设置 max_tokens	无需（答案固定）
成本防护	需配额与频率管理	本地运行，无持续费用
敏感内容	需额外检测过滤	答案可预先审核
恶意输入	需多层防护	受限于预设类别，风险较低

适用场景对比如表 9 所示。

215 表 9 两类方案适用场景对比

Tab. 9 Applicable scenarios of the two approaches

场景	HD-001 (LLM)	HXXedge (BERT)
开放域问答	较适合	不适用
复杂推理	较适合	不适用
标准问答	可用, 但存在输出偏差	较为适合
离线环境	不可用	适合
涉密场景	不可用	适合
成本敏感	存在持续费用	一次性投入

由对比结果可得: LLM 方案在开放域和复杂推理场景中更具优势, 但存在成本较高、延迟较大、输出内容不可完全控制等问题; BERT 方案在标准问答、离线环境和涉密场景中优势明显。两者更宜按场景配合使用, 而非相互替代。

6 讨论

6.1 BERT 方案的适用场景

在本研究的实验条件下, BERT 方案在标准问答、低延迟要求和资源受限环境中表现较为合适: 答案来源固定、无自由生成偏差, 单条推理约 100 ms 可满足实时交互, 且 CPU 即可完成推理, 无需持续支付 API 调用费用。

6.2 BERT 方案的局限性

(1) 仅适用于判别式任务, 难以处理开放式对话或文本生成; (2) 需要预先准备训练数据, 类别体系的设计直接影响效果; (3) 新增类别时需重新训练, 扩展成本相对较高。

6.3 大模型的对比

BERT 与大语言模型的总体对比如表 10 所示。

表 10 BERT 与大语言模型总体对比
Tab. 10 Overall comparison between BERT and large language models

维度	BERT	大语言模型
任务类型	判别式 (分类、匹配)	生成式 (对话、创作)
参数量	约 1.02 亿 (102M)	数十亿至万亿级
硬件要求	CPU	GPU (推理)
推理延迟	约 100 ms	数秒
输出偏差	较低	存在
答案可控性	较高	较低
开放性	较低 (封闭域)	较高 (开放域)

6.4 工程经验

(1) 数据质量直接影响模型效果, 类别定义应清晰、避免重复; (2) 短问题是分类模型的难点, 宜通过标准问法显式加入训练数据; (3) 部署阶段应记录分类标签与置信度,

235 便于快速发现类别语义重叠等训练数据问题；（4）训练数据宜按版本目录（如 v1/、v2/）独立归档，每版保留种子数据、训练集、验证集与类别定义的完整快照，便于跨版本对比与回滚；（5）当知识来源于多份标准文档时，需进行跨文档去重与术语统一，避免语义重复干扰分类边界；（6）训练数据应兼顾专业性与通用性——专业标准确保权威性，企业知识拓展使用对象范围。

7 分类标准与产品化展望

240 7.1 分类编码体系设计

本系统的分类编号采用“模型内部编号+业务编码”的双层结构：模型输出数字编号（0～81），再通过映射表转换为业务编码，以支持灵活的分类管理。建议的层级编码结构为“[大类]-[中类]-[序号]”，例如：COMP-BASIC-001（公司基本信息）、MARINE-OBS-001（海洋观测）、CS-GREET-001（客服问候）等，主要大类划分如表 11 所示。

245 表 11 分类编码大类划分
Tab. 11 Major categories of the classification coding system

大类代码	名称	说明	典型条目
COMP	公司信息	企业基础信息	地址、电话、成立时间
PROD	产品信息	产品体系介绍	海视通、外业通、智数通
MARINE	海洋专业	海洋领域知识	观测规范、预报术语
INSTR	仪器设备	仪器相关知识	校准、送检、使用
PROJ	项目管理	项目相关知识	复盘、规范、流程
CS	客服话术	通用对话处理	问候、感谢、告别、拒答
CHAT	闲聊对话	非业务对话	闲聊、身份、能力引导

7.2 端侧部署的应用场景

250 本系统面向海洋观测领域，观测站点多部署于偏远海岸或远海区域，网络条件差、环境复杂，对端侧智能问答具有较强需求。同时，BERT 轻量化、可离线的特点也使其具备向其他端侧场景拓展的潜力。典型应用场景如表 12 所示。

表 12 BERT 端侧部署典型应用场景
Tab. 12 Typical application scenarios for edge-side BERT deployment

场景	环境特点	BERT 方案的特点
岸基海洋站	偏远海岸、台风期间断网	本地持续服务，不受极端天气影响
浮标观测系统	海上无人值守、卫星通信带宽有限	完全离线运行，无需实时回传
海上观测平台	远海作业、网络不稳定	离线可用、响应迅速
水下观测站点	水下通信受限、无法接入云端	嵌入式部署、本地推理
智能家电	算力有限、无 GPU	CPU 可推理，模型较轻
工厂车间	网络较差或无网络	可完全离线运行

野外工地 / 无人站	无网络、环境复杂	本地推理、功耗较低
车载系统	移动网络不稳定	离线可用、响应较快
涉密单位	数据不能外发	全程本地处理

在数据安全要求较高的组织中，数据往往不允许发送至第三方服务器。此时即使具备本地 GPU，也无法调用云端 API（数据需经第三方），BERT 的本地部署方案因而具有实际意义：数据全程本地处理、无 API 调用记录、更易满足等保与分级保护要求。

BERT 模型参数量较小（约 1.02 亿 / 102M），推理阶段对硬件要求较低，无需昂贵的 GPU 设备，可在 x86 CPU 及各类国产 ARM SoC 上运行，适配范围较广。表 13 列出了不同硬件层级下的端侧部署方案。

表 13 端侧部署硬件要求

Tab. 13 Hardware requirements for edge-side deployment

硬件层级	典型配置	推理延迟	适用场景
x86 CPU	普通 PC / 工控机（4 GB 内存）	约 100 ms	岸基站点、室内部署
ARM SoC + NPU	国产开发板（4-8 GB 内存）	约 50 ms	嵌入式设备、观测平台
国产 AI 芯片	瑞芯微、昇腾、全志等	视具体型号	国产化替代、信创场景

本研究团队已在普通 PC（x86 CPU）和国产 ARM 开发板（RK3588）上完成 CPU 推理部署验证，推理延迟分别约为 100 ms 和 300 ms，均可满足实时交互需求。若启用 NPU 加速，延迟有望进一步降至 50 ms 量级。

7.3 产品化方向

未来产品化可按三个层级推进，如表 14 所示。

表 14 产品化层级规划

Tab. 14 Productization roadmap at three levels

层级	内容	目标客户
SaaS 服务	分类模型+管理后台	中小企业
行业解决方案	预置行业知识库	行业客户
私有化部署	完整系统+定制训练	大型企业

产品化的核心在于：一是建立可复用的知识分类标准，支持知识库的快速扩展；二是保持轻量化部署，无需 GPU 即可运行；三是保证答案可控，适用于对准确性有要求的正式业务场景；四是控制长期使用成本，避免持续的 API 调用费用。

与 LLM 方案相比，二者并非替代关系：LLM 适合开放式对话、内容生成与复杂推理；BERT 适合标准问答、知识检索与流程化问答。在答案准确性要求较高的场景中，BERT 分类方案具有更稳定的表现。

7.4 模型轻量化方向

BERT 各轻量化变体的参数规模与适用场景如表 15 所示。

表 15 BERT 轻量化变体对比
Tab. 15 Comparison of lightweight BERT variants

模型变体	参数量	相对 Base	适用场景
BERT-Base	约 1.02 亿 (102M)	基准	普通 PC / 服务器
DistilBERT	约 6,600 万	约 -40%	资源受限环境
MobileBERT	约 2,530 万	约 -77%	移动端
TinyBERT	约 1,450 万	约 -87%	嵌入式 / IoT

BERT 等预训练语言模型的压缩与轻量化技术已有系统性研究综述^[17]。常用的轻量化手段包括知识蒸馏^[18]（以教师模型指导学生模型学习）、模型剪枝（移除不重要的权重或层）以及量化^[19]（由 FP32 转为 INT8，减小体积并加速推理）等。

8 结论

本文围绕海洋领域知识问答场景，在端侧 LLM 实验的基础上，采用仅约 1.02 亿（102M）参数的 BERT 分类模型构建了可离线部署的问答系统。前期在国产 ARM 开发板上的端侧实验表明，LLM 在延迟、幻觉及资源占用方面存在局限，而 BERT 分类方案具有推理快速、答案固定、无幻觉等优势。

本文强调“领域即专业”，训练数据来源于国家标准、行业规范及企业知识，确保答案的权威性与专业性。在数据构建过程中，发现多源权威文档中同一概念存在不同表述的语义一致性问题的，并提出了基于分类置信度分析的训练数据去重方法，使“海温”等短问题置信度由 49.85% 提升至 94.16%，验证集准确率达 96.79%。

本文的核心观点是：对于领域专用任务，模型选型应追求“小而精”而非盲目追求参数规模。大模型中大部分参数用于编码通用语言知识，对领域任务构成冗余；而轻量化的判别式模型配合高质量的领域训练数据，能够在特定任务上达到甚至超越大模型的效果，同时在部署成本、推理延迟、数据安全等方面具有显著优势。更重要的是，跨平台实测表明，该方案在 x86（笔记本）与 ARM 开发板两种主流 CPU 架构上均可流畅运行（推理延迟分别约 87ms 与 307ms），无需 GPU 支持，可在 CPU 服务器、台式机、笔记本、嵌入式开发板等海量存量设备上部署，为企事业单位利用现有硬件实现本地化 AI 应用提供了可行路径。实践表明，在答案相对固定、对准确性要求较高的专业场景中，该类方案能够提供稳定、低成本的自动答复，可作为生成式大模型之外的有效补充。后续工作将围绕以下方向持续深入：(1) 领域知识分类体系构建方法论——将本文的分类编码体系与训练数据构建流程推广至其他垂直行业，研究可迁移的领域分类标准设计方法；(2) 多源知识融合与数据质量控制——将本文提出的基于置信度分析的语义去重方法扩展至更大规模的多源文档融合场景，研究跨文档语义对齐与术语规范化的自动化机制；(3) 面向领域分类模型的增量学习策略——探索参数高效的微调方法，使模型在不全量重训练的前提下持续吸收新领域知识，同时缓解

灾难性遗忘问题；(4) 判别式-生成式混合问答架构——研究 BERT 与生成式模型的协同机制，由判别式模型承担快速意图分类与置信度守门，生成式模型处理开放式或复杂推理查询；(5) 模型量化与海洋观测端侧部署——探索 BERT 模型的量化压缩方案及其在海洋观测站点（岸基、浮标、海上平台、海底观测站等）的实际部署与性能评估。

[参考文献] (References)

- [1] LEWIS P, PEREZ E, PIKTUS A, et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks[C]//NeurIPS. 2020: 9459-9474.
- [2] GAO Y, XIONG Y, GAO X, et al. Retrieval-Augmented Generation for Large Language Models: A Survey[J]. arXiv preprint arXiv:2312.10997, 2023.
- [3] HUANG L, YU W, MA W, et al. A Survey on Hallucination in Large Language Models[J]. ACM Transactions on Information Systems, 2025, 43(2): 1-55.
- [4] NA S, JEONG G, AHN B H, et al. Understanding Performance Implications of LLM Inference on CPUs[C]//IEEE International Symposium on Workload Characterization (IISWC). IEEE, 2024: 169-180.
- [5] RAHA A, KUNDU S, SRIDHAR S N, et al. LLM-NPU: Towards Efficient Foundation Model Inference on Low-Power Neural Processing Units[C]//IEEE International Conference on Omni-Layer Intelligent Systems (COINS). IEEE, 2025: 1-8.
- [6] 中国信息通信研究院, 中国人工智能产业发展联盟. 大模型推理优化关键技术及应用实践研究报告 (2026 年) [R/OL]. 北京: 中国信通院, 2026[2026-09-13]. https://www.caict.ac.cn/kxyj/qwfb/ztbg/202604/t20260415_719838.htm.
- [7] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[C]//Proceedings of NAACL-HLT. Minneapolis: ACL, 2019: 4171-4186. DOI: 10.18653/v1/N19-1423.
- [8] SUN C, QIU X, XU Y, et al. How to Fine-Tune BERT for Text Classification?[C]//China National Conference on Chinese Computational Linguistics (CCL). Berlin: Springer, 2019: 194-206.
- [9] SANH V, DEBUT L, CHAUMOND J, et al. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter[J]. arXiv preprint arXiv:1910.01108, 2020.
- [10] JIAO X, YIN Y, SHANG L, et al. TinyBERT: Distilling BERT for Natural Language Understanding[C]//Findings of EMNLP. Online: ACL, 2020: 4163-4174. DOI: 10.18653/v1/2020.findings-emnlp.372.
- [11] SUN Z, YU H, SONG X, et al. MobileBERT: a Compact Task-Agnostic BERT for Resource-Limited Devices[C]//Proceedings of ACL. Online: ACL, 2020: 2158-2170. DOI: 10.18653/v1/2020.acl-main.195.
- [12] HUGGING FACE. bert-base-chinese[EB/OL]. (2019)[2026-09-01].
- [13] 沈立力, 姜鹏, 王静. 基于 BERT 模型的中文期刊文献自动分类实践研究[J]. 图书馆杂志, 2022, 41(5): 109-118.
- [14] 涂远来, 周家乐, 王慧锋. 基于 BERT 预训练模型的事故案例文本分类方法[J]. 华东理工大学学报 (自然科学版), 2023, 49(4): 576-582.
- [15] 宫丽娜, 周易人, 乔羽, 等. 预训练模型在软件工程领域应用研究进展[J]. 软件学报, 2025, 36(1): 1-26.
- [16] ONNX RUNTIME. Deploying Transformer Models with ONNX Runtime[EB/OL]. (2024)[2026-09-01].
- [17] XU C, McAULEY J. A Survey on Model Compression and Acceleration for Pretrained Language Models[C]//AAAI. 2023.
- [18] HINTON G, VINYALS O, DEAN J. Distilling the Knowledge in a Neural Network[J]. arXiv preprint arXiv:1503.02531, 2015.
- [19] ZAFIR O, BOUDOUKH G, IZSAK P, et al. Q8BERT: Quantized 8Bit BERT[J]. arXiv preprint arXiv:1910.06188, 2019.