文章编号:1671-4598(2025)10-0090-07 DOI:10.16526/j.cnki.11-4762/tp.2025.10.012 中图分类号:TP181 文献标识码:A

基于 LoRA 微调的无损检测领域大语言模型研究

任文雄[1]，贾 新[1]，吴 洋[2]，刘慧玲[1]，容 婷[3]
(1. 武汉明臣焊接无损检测有限公司，武汉 430070；
2. 中安检测集团（湖北）有限公司，武汉 430050；
3. 武汉熵减人工智能科技有限公司，武汉 430070)

摘要：通用大语言模型（LLMs）在无损检测（NDT）领域的应用普遍存在着对专业术语理解不够精准、难以准确地适配动态更新的法规标准等诸多问题，针对这些问题，对基于低秩自适应（LoRA）技术的轻量化领域适配方案进行了研究，并且对其展开评估，构建了一个含有一万余条高质量数据的 NDT 领域专属数据集，运用 LoRA 技术对参数规模在 70～90 亿的基座 LLMs 实施高效微调；研究结果表明，优化后的模型在 BLEU-4 和 ROUGE-*L* 等评估指标上，实现了显著的提升；上述模型在 NDT 专业知识问答环节均表现出优异的性能，成功证实了 LoRA 技术在 NDT 领域具备良好的适配特性，为智能化 NDT 系统的发展提出了一种资源利用高效且具有实际应用潜力的技术路径；未来研究将融合前沿技术成果，进一步提升模型的学习能力与实际应用性能。

关键词：无损检测；大语言模型；数据集；低秩自适应技术；微调

Research on Large Language Models in Non-Destructive Testing Based on LoRA Fine-Tuning

REN Wenxiong[1], JIA Xin[1], WU Yang[2], LIU Huiling[1], RONG Ting[3]
(1. Wuhan Mingchen Welding Non-destructive Testing Co., Ltd., Wuhan 430070, China;
2. Zhong'an Testing Group (Hubei) Co., Ltd., Wuhan 430050, China;
3. Wuhan Entropy Reduction Artificial Intelligence Technology Co., Ltd., Wuhan 430070, China)

Abstract: General large language models (LLMs) are applied in the field of non-destructive testing (NDT), which generally exist many problems, such as inaccurate understanding of professional terms and difficulty in accurately adapting to dynamically updated regulations and standards. In response to these problems, a lightweight adaptation scheme based on low-rank adaptive (LoRA) technology is studied and evaluated. Build a dedicated data set for the NDT domain with over ten thousand high-quality data points. Adopt the LoRA technology to efficiently fine-tune the base LLMs with parameter scales ranging from 7 to 9 billion (7-9B). The research results show that the optimized model has achieved a significant improvement in evaluation metrics such as BLEU-4 and ROUGE-L. The above models all demonstrate outstanding performance in the NDT professional knowledge Q&A session, successfully confirming that the LoRA technology has good adaptability in the NDT field, and provides a technical path in efficient resource utilization and practical application potential for the development of intelligent NDT systems. Future research will integrate cutting-edge technological achievements to further enhance the learning ability and practical application performance of the model.

Keywords: NDT; LLMs; data set; low-rank adaptation technology; fine-tuning

收稿日期:2025-07-24； **修回日期**:2025-08-19。

作者简介:任文雄(1993-),男,大学本科,工程师。

引用格式:任文雄,贾 新,吴 洋,等. 基于 LoRA 微调的无损检测领域大语言模型研究[J]. 计算机测量与控制,2025,33(10):90-96.

0　引言

无损检测（NDT，non-destructive testing）是指在不损害被检对象的情况下，对其内部和表面进行检测，以评估其完整性、结构和性能，是保障现代工业体系中重大装备服役安全与全生命周期管理效能的关键技术手段。NDT 涉及特种设备、石油化工、电力核电、航空航天、海洋工程、船舶及铁路等诸多行业，具体实施检测过程中，主要依靠人工进行操作，因此，普遍存在标准执行力不足、知识更新滞后等问题。

大语言模型（LLM，large language models）是一种基于 Transformer 架构的深度学习模型，Transformer 架构主要由两部分组成：编码器和解码器，其中编码器负责将输入文本序列转换为一系列向量表示，每个编码器层都包含一个自注意力子层和一个前馈神经网络子层。解码器则根据编码器的输出和之前的生成内容，逐步生成新的文本序列，每个解码器层除了自注意力子层和前馈神经网络子层外，还增加了一个交叉注意力子层，用于关注编码器的输出。Transformer 的核心原理是自注意力机制——它允许模型在处理一个词时，同时“关注”到序列中的所有其他词，从而捕捉到词与词之间的关系。通过这种机制，模型能够更好地理解上下文。LLM 的训练通常分为两个阶段：预训练和微调，预训练是运用模型在海量的文本数据上进行无监督学习，目标是预测下一个词，在此阶段中，模型学到了大量的语言、语法和常识。微调是在特定任务的数据集上对预训练好的模型进行有监督或无监督的训练，这使得模型能够更好地适应特定的应用场景，比如问答、摘要、翻译等等。在训练过程中，模型会不断调整其参数，使得预测结果与真实标签之间的差异最小化。

近年来，GPT-4o、Claude 3.7 及 Gemini 2.5 等为代表的先进 LLM，凭借自身颇为强大的语义理解与文本生成方面的能力，在自然语言处理领域取得了飞跃式的发展，并在医疗、法律及编程等多个垂直领域都呈现出显著的场景适应方面的潜力[1]。LLM 虽然是处理文本的模型，但其强大的模式识别和推理能力，使其在 NDT 领域也有潜在的应用价值。即便如此，将这些通用 LLM 直接应用于高度专业化的 NDT 领域时，其可靠性往往会由于对专业术语的理解偏差和对最新标准条款的适配能力不足，进而产生一定的局限性[2]。NDT 领域有大量高度专业且上下文相关的术语，通用 LLM 可能无法准确区分在不同检测方法或材料中，同一个词语所代表的细微差别。例如，通用模型可能将“裂纹”与“裂缝”视为同义词，但在 NDT 中它们可能代表不同的缺陷形态或严重程度。而且 LLM 可能会将一个 NDT 特有概念与通用语言中的相似概念混淆。例如，通用模型可能会将“涡流”理解为水流，而不是一种电磁感应的无损检测技术，这种误解会导致模型生成不准确或完全错误的信息。NDT 领域的标准和行业规范会定期更新，通用 LLM 的训练数据通常有时间截点，因此难以获取和理解最新的标准条款，这使得模型在回答有关最新法规时，可能提供过时甚至不适用的信息。当 LLM 生成关于特定标准的信息时，它很难像专业人士那样提供具体的标准编号、版本号和章节引用，生成的内容可能缺乏可信度和可验证性，在要求严格的 NDT 领域是不可接受的。而且，模型是基于概率预测下一个词，不是基于真实的物理或知识进行逻辑推理，它可能会编造出看似合理但实际上不存在的检测结果、标准条款或技术细节，这在 NDT 这种对准确性要求极高的领域是致命的。

以上这些局限性从根本上来讲，是因为 LLM 缺乏深度嵌入的 NDT 领域专业知识，所以需要通过高效的技术手段将专业知识灌输到模型中，进而达成动态适配的效果。

传统的全参数微调方法虽然能够提升模型在特定领域的表现，但是它需要对百亿乃至千亿级别的参数进行优化，这样一方面会面临高质量领域专属数据稀缺的难题，另一方面会因算力与时间成本高昂，严重影响 LLM 在 NDT 领域的实施和更新。在这种情况下，低秩自适应（LoRA，low-rank adaptation）技术应运而生，它的运行逻辑是冻结预训练模型的主体参数，只针对灌输的、参数量远小于原模型的低秩矩阵进行优化，该技术为专业领域知识高效嵌入模型与模型轻量化适配提供了新的思路[3]。相关研究证明，LoRA 能够在明显减少微调参数量以及算力需求的同时，还能使模型性能损失控制在可以接受的范围之内[4]。尽管 LoRA 技术已经展示出巨大潜力，但它在 NDT 专业领域的适配效果、最佳实践做法以及对复杂标准遵循程度的提升，还需要更加系统性的深入研究与验证。

本研究对 NDT 领域的各类标准、专业教材与相关书籍进行全面且细致地整合，精心构建了一个高质量的 NDT 领域专属数据集，并使用 LoRA 技术，对选定的 70～90 亿参数量级（7～9 B）的基座模型[5]进行高效微调。这种方法主要目的在于充分评估并进一步优化模型对 NDT 特定知识体系的理解与应用能力，进而妥善处理现有 LLM 在该领域适配性欠佳的问题，与此同时可以避免传统全参数微调方法所引发的迭代困难与成本昂贵的问题。

1　微调策略选择

研究团队针对当下主流微调技术[6]展开深入对比，

详细优缺点对比如表 1 所示。本研究综合权衡各类技术的特点之后[7]，最终选定使用 LoRA 技术执行微调任务。

表 1 多种微调技术的优缺点比较

微调技术	优势	不足
Adapter Fusion	简单易用，训练过程稳定	需要更多计算资源和存储空间
Prefix-Tuning	训练速度快，计算资源消耗少	表达能力有限，且模型需要大量的训练数据来调整前缀
P-Tuning	能够有效地优化模型参数，从而提高模型的准确率	需要较多计算资源和较长时间进行参数优化
LoRA	计算资源和存储资源消耗较小，有良好的兼容性	需要对超参数进行更复杂的调试

LoRA 微调方法[8]是运用低秩分解手段模拟参数的变动量，以极小的参数量达成大模型的间接训练目标，其实现原理如图 1 所示。对于一个复杂的任务，例如本研究的专业问答，模型需要学习更多细粒度的知识，这意味着参数变动量可能较大，因此需要更高的秩值来捕捉这些复杂的变动，以确保微调效果。通常，模型越大，其参数量也越大。为了有效微调，大型模型可能需要更高的秩值来捕捉其复杂的参数空间变动。例如，对于一个 7 B（70 亿参数）的模型，可能选择 $r=8$ 或 16；而对于一个 70 B 的模型，可能需要更高的秩值，如 $r=32$ 或 64，才能达到理想效果。在本研究过程中发现，选择秩值在 4～64 之间逐步增加，同时观察模型性能的变化，在保证极少参数增量的前提下，能够有效捕捉到任务所需的关键信息。大多数情况下，这个范围内的秩值可以取得与全量微调相近的性能，同时大幅节省计算资源。

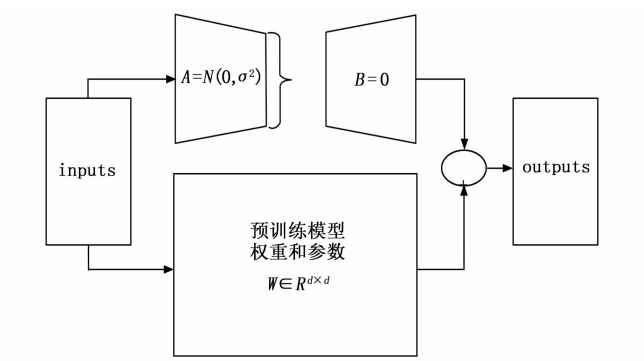

图 1 LoRA 原理

在微调预训练模型时，需更新的参数可表示为 $W_0+\Delta W$，其中，W_0 为模型的初始预训练参数，ΔW 为待学习的参数更新量。全参数微调（$W_0=\Delta W$）通常需要极高的计算资源成本。为降低微调开销，LoRA 方法被提出。LoRA 约束 ΔW 为一个低秩分解矩阵，表示为：

$$W_0+\Delta W=W_0+\boldsymbol{BA},\boldsymbol{B}\in R^{d\times r},\boldsymbol{A}\in R^{r\times k} \quad (1)$$

$\Delta W=\boldsymbol{BA}$，其中 $\boldsymbol{B}\in R^{r\times k}$，$\boldsymbol{A}\in R^{r\times k}$ 且 $r\ll min(d,k)$。

在该框架下，模型参数更新为[5]：

为在训练初始阶段最小化对预训练知识的影响，矩阵 $\boldsymbol{A}$ 使用随机高斯分布初始化，而矩阵 $\boldsymbol{B}$ 初始化为零矩阵。随后，预训练权重 W_0 被冻结，仅对低秩矩阵 $\boldsymbol{A}$ 和 $\boldsymbol{B}$ 的参数进行优化。在前向传播过程中，输入 x 同时作用于原始权重和低秩适配器：

$$h=W_0x+\Delta Wx=W_0x+\boldsymbol{BA}x \quad (2)$$

最后，通过将可训练的低秩矩阵 $\boldsymbol{A}$ 和 $\boldsymbol{B}$ 集成到 Transformer 模型的每一层，LoRA 实现了高效且低成本的参数更新，从而构建出适应特定领域的新模型。

本研究设计了一个更优化的多阶段微调流程。在第一阶段，采取领域适应性微调：首先，在海量的 NDT 文献、标准和报告数据上，使用 LoRA 对模型进行通用领域知识的适应性训练。这使得模型能够更好地理解 NDT 的专业术语和基本概念。在第二阶段：采取任务特异性微调：在针对特定任务（如缺陷报告生成）的高质量、小规模数据集上，进行进一步的微调。这种分阶段的策略可以确保模型既掌握了全面的领域知识，又具备了出色的特定任务执行能力。

2 实验与分析

2.1 模型选择

大规模多任务语言理解（MMLU，massive multitask language understanding）是评估大模型跨领域知识与推理能力的基准[18]，覆盖数学、法律等 57 个学科，采用少样本多选题的形式，其评分反映模型的知识广度和逻辑推理水平，用于衡量模型在多样化任务中的适用性[10]。

如表 2 所示，此次微调选取 MMLU 评分较高的 Qwen2.5－7 B[11]、Yi1.5－9 B、Ministral-8 B 总共 3 个基座模型，对其进行数据集中专业知识预测并测试模型问答示例交由专家评估，选取最适用于 NDT 领域的模型。

表 2 不同模型在 MMLU 上的表现

模型名称	MMLU
Qwen2.5－7 B	74.2
Yi1.5－9 B	69.5
Ministral-8 B	65.0

2.2 数据收集及处理

本研究用于构建 NDT 专业数据集的资料主要包括专业培训资料、法规标准文件、专业技术文献。专业培

训资料作为长期逐步积累的教学素材，涵盖了完整理论框架，经由大量工程实践案例不断沉淀，给模型微调打下了颇为扎实的专业基础，对于来自非权威渠道（如论坛问答）的信息，我们会与至少两个权威来源（如行业标准或知名教科书）进行交叉比对，验证其准确性；法规标准文件通过对现行的国家标准以及行业规范加以整合，优先选择最新版本（近 5 年内）的标准和文献，去除已被新规范取代的过时内容，确保知识的时效性，为模型微调提供标准化依据；专业技术文献则精心挑选了和 NDT 技术关联紧密的专业著作，内容涉及到特种设备设计、安装、制造及检验检测等环节，确保数据集涵盖不同检测方法、材料类型（金属、复合材料）、缺陷种类（裂纹、气孔、夹渣）和应用行业，以保证模型的通用性和泛化能力，相当程度上拓展了知识体系的深度与广度。

收集到的原始数据形态各异，需要经过一系列预处理步骤，才能形成适合模型训练的结构化数据集。在数据处理环节中，研究团队实施了一套严格的清洗流程[12]，将 PDF、Word 文档和网页数据转换成纯文本格式，使用 OCR 技术从图片和扫描件中提取文本，同时移除文档中的页眉、页脚、版权声明、广告等与内容无关的冗余信息。通过对原始数据展开多轮细致的筛选，将自动化查重与人工复核这两种方式有机结合起来，将重复数据、错误信息及非相关内容剔除，以此确保数据质量能够达到相应要求。随后针对筛选完毕的数据，进一步进行结构化处理，将其系统性地转化为标准化的知识问答对形式[13]。最终构建完成的 NDT 专业数据集包含 1 万余条高质量数据（数据格式见下方）：

" instruction":" 编制渗透检测工艺文件时应遵循哪些基本原则?",

" input":"",

" output":" 编制渗透检测工艺文件时应遵循以下基本原则：

\ n \ n1. 检测工作质量保证：所编制的工艺文件必须确保达到相关法规、安全技术规范和技术标准等各项技术要求的检测灵敏度。

\ n2. 高效性：工艺过程需具有较高的检测效率，以提高整体工作效率。

\ n3. 成本控制：尽量减少费用，降低检测成本，实现经济效益最大化。

\ n4. 劳动强度减轻：通过优化工艺流程和采用适当的设备器材来减轻检测人员的工作负担。

\ n5. 环保与安全：在保证检测效果的同时，应尽可能减少对环境的影响，并确保操作过程的安全性。",

" system":""

所有数据全部采用规范的 JSON 格式进行存储，如此一来，便能够为后续的模型训练以及应用开发等相关工作提供极大的便利[10]。

2.3　实验评估指标

本次实验选取数据集中的 500 条数据构建专业知识测试集，以此对模型进行精准预测分析[14]，并借助 BLEU-4、ROUGE-1、ROUGE-2、ROUGE-L [15]等专业指标严谨验证 NDT 专业知识领域的模型预测效果。相关原理如下：

1）双语评估替换（BLEU，bilingual evaluation understudy）：

BLEU 指标通过计算生成文本与参考文本之间的 n-gram 匹配度以及整体长度的适配性来评估翻译质量。其核心计算公式为：

$$BLEU\text{-}N = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{3}$$

其中：p_n 表示 n-gram 精确度；w_n 是对应 n-gram 精确度的权重；N 是考虑的最长 n-gram 长度（通常取值为 4）。

BP（Brevity Penalty）为简洁性惩罚因子，用于惩罚过短的输出，其计算公式如下：

$$BP = \begin{cases} 1, & \text{if} \quad a > b \\ e^{(1-b/a)}, & \text{if} \quad a \leqslant b \end{cases} \tag{4}$$

此处 a 代表生成文本的长度，b 代表参考文本的有效长度（通常取最接近 a 的参考文本长度）。

2）面向召回的要点评估替换（ROUGE，recall-oriented understudy for gisting evaluation）：

ROUGE 是一组基于生成文本与参考文本之间重叠单元（如 n-gram 或最长公共子序列）的召回率来评估文本摘要（或机器翻译）质量的指标。本研究选用 ROUGE-1、ROUGE-2 和 ROUGE-L 进行评估：

（1）ROUGE-N：衡量 n-gram 的召回率：

$$ROUGE\text{-}N = \frac{\sum_{S\in\{\text{Reference Summaries}\}} \sum_{gram\ m_n \in S} Count_{\text{match}}(gram_n)}{\sum_{S\in\{\text{Reference Summaries}\}} \sum_{gram\ m_n \in S} Count(gram_n)} \tag{5}$$

其中：$Count_{\text{match}}$（$gram_n$）是参考文本和生成文本中共同出现的 $gram_n$ 的数量，$Count_{\text{match}}$（$gram_n$）是指在参考文本中出现的的 $gram_n$ 总数。

（2）ROUGE-L：基于最长公共子序列（LCS）的 F 值（F-score）：

$$ROUGE\text{-}L = \frac{\sum_{S\in\{\text{Reference Summaries}\}} LCS_{\text{Reference,Candidate}}}{\sum_{S\in\{\text{Reference Summaries}\}} length_{\text{Reference}}} \tag{6}$$

其中：$LCS_{\text{Reference,Candidate}}$ 是参考摘要 Reference 和生成摘要 Candidate 之间最长公共子序列的长度，$length_{\text{Reference}}$

是参考摘要 Reference 的长度。

2.4 实验设置

本研究采用的配置见表 3[16]。

表 3 系统与环境配置一览表

分类	项目	配置说明
硬件	CPU	Intel® Core™ i9-14900K @ 3.20 GHz
	GPU	NVIDIA RTX 4090,24 GB 显存
软件	操作系统	Ubuntu 24.04.2 LTS
	Python 版本	Python 3.10
	深度学习框架	PyTorch 2.6.0
	CUDA 版本	CUDA 12.2
	Transformers 库	Transformers 4.50.0
	LLM 运行工具	vllm 0.8.2

本次实验微调采用 LoRA＋技术[17]。LoRA＋是对参数高效微调方法 LoRA 的改进，其核心创新在于，研究人员为矩阵 ***B***（将低维向量映射到原始高维空间的矩阵）设置一个比矩阵 ***A***（将原始高维向量投影到低维空间的矩阵）高得多的学习率。这种设置的理论依据在于，在训练初期，***A*** 和 ***B*** 都是随机初始化的，它们的梯度幅值存在天然的不平衡。具体来说，当反向传播时，***B*** 矩阵的梯度会乘以一个大得多的权重矩阵，导致其梯度幅值远大于 ***A*** 矩阵。在一系列实验中，研究人员对比了标准 LoRA 和 LoRA＋训练过程中矩阵 ***A*** 和 ***B*** 的梯度幅值。

标准 LoRA 在整个训练过程中，***B*** 矩阵的梯度幅值始终远大于 ***A*** 矩阵。这导致 ***B*** 的参数更新幅度过大，容易造成训练不稳定甚至发散，而 ***A*** 的参数更新则过于缓慢，无法有效学习。LoRA＋通过为 ***B*** 设置更高的学习率，我们观察到在训练初期，虽然 ***B*** 的梯度幅值仍高于 ***A***，但其参数更新能够更快地进入一个稳定的状态。更重要的是，通过调整学习率，我们可以使 ***A*** 和 ***B*** 的有效更新幅度在训练后期趋于平衡。这使得两个矩阵都能以合适的步长进行学习，从而提升训练的效率和稳定性。通过差异化学习率，LoRA＋能够使模型更快地收敛。这得益于更协调的参数更新，***B*** 矩阵的参数快速调整到正确的方向，而 ***A*** 矩阵则能够稳定地微调，两者协同工作，使得训练路径更平滑、更直接地通向最优解。实验表明，LoRA＋可以在更少的训练步骤内达到与标准 LoRA 相同的性能水平。

关键参数设置如下：基础学习率为 3×10^{-5}；训练轮数（epoch）为 5；LoRA＋的学习率比例因子取值为 16；学习率调度（scheduler）采用余弦衰减（cosine）[18]。

2.5 结果与分析

为切实验证模型的训练成效，观察训练过程中详实记录的 *Loss* 值变化态势、训练完成时模型的最终 *Loss* 值以及模型对问题的推理表现，通过三者协同作用，全方位实现对模型准确性与稳定性的科学评价。模型的 *Loss* 值运用交叉熵损失进行精准计算，其在微调过程中的变化曲线横坐标清晰标注训练步数，纵坐标精准反映模型在对应训练步数下的损失值[19]。当 *Loss* 值趋于稳定且最终 *Loss* 值低于 0.3 之后，通过精心构造的包含 500 道问答题的数据文件对模型进行验证性测试。结果如表 4 所示，与原始模型相比，LoRA 方法带来了大幅度的改进，使模型在处理 NDT 知识任务时表现更加优越，具备更高的准确性和专业性[20]。

表 4 模型微调前后预测评分

模型名称	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-*L*
Qwen2.5-7 B	2.86	17.79	4.63	9.01
Qwen2.5-7 B(LoRA)	25.33	48.25	32.03	37.92
Yi1.5-9 B	5.21	28.71	14.54	12.50
Yi1.5-9 B(LoRA)	48.16	65.38	52.10	58.73
Ministral-8 B	9.14	32.19	12.93	18.33
Ministral-8 B(LoRA)	34.27	55.02	38.46	46.08

2.6 模型问答示例及专家评估

2.6.1 Qwen2.5-7B 问答示列

问：简述渗透检测的基本原理及适用范围。

答：渗透检测是一种无损检测方法，主要用于发现材料表面开口的缺陷，如裂纹、气孔等。其基本原理是利用润湿作用和毛细管效应，通过渗透剂渗入这些缺陷中，并再经过显像剂显现出来。具体步骤包括预清洗、施加渗透剂、去除多余的渗透剂、施加显像剂以及观案和评估。

渗透检测的适用范围广泛，适用于多种材料如金属、非金屋和复合材料。它可以用于检查工件表面的各种缺陷，包括但不限于裂纹、气孔、折叠、分层等。

问：超声波斜角探伤时，焊缝中的近表面缺陷不容易探测出来，其原因是什么？

答：在超声波斜角探伤中，焊缝近表面缺陷（通常指距离工件表面小于 10 mm 的缺陷）难以探测，主要与超声波的传播特性、缺陷信号的干扰及仪器响应限制等因素密切相关，具体原因如下：

1）近表面盲区（死区）的影响超声波探伤时，探头发射的初始脉冲（电脉冲转换为声波的瞬间）具有一定宽度（通常称为“始脉冲宽度”）。对于近表面缺陷，其反射信号返回探头的时间极短，会与始脉冲重叠，导致缺陷信号被始脉冲“覆盖”或“淹没”，无法与始脉冲区分 。

2）波型转换与干扰：

斜角探伤中，声波（通常为横波，由斜探头折射产生）在近表面区域易发生波型转换：当声波倾斜入射到

工件表面或近表面缺陷时，会同时产生纵波、横波甚至表面波（如瑞利波）。不同波型的传播速度、反射规律不同，其反射信号会相互干扰，导致近表面缺陷的特征信号被杂波掩盖。

3）缺陷尺寸与超声波波长的匹配问题：

超声波的探测能力与缺陷尺寸和波长的比例密切相关：当缺陷尺寸接近或小于超声波波长时，反射信号会显著减弱（遵循“瑞利散射”规律）。

2.6.2　Yi1.5-9B 问答示例

问：简述渗透检测的基本原理及适用范围。

答：渗透检测的基本原理及适用范围如下：

基本原理：渗透检测是一种无损检测方法，它通过利用液体（渗透剂）渗入材料表面的微小缺陷或裂纹来识别这些缺陷。具体来说，在进行渗透检测时，首先会将含有着色染料或荧光染料的渗透剂涂抹在待检工件表面上，让渗透剂充分渗入到工件表面的缝隙中，包括裂纹、气孔等微小缺陷处。随后，使用清洗步骤去除多余的渗透剂，并施加显像剂以增强缺陷处的渗透剂显示效果。最后通过观察显微镜下的工件表面来识别和评估这些缺陷。

适用范围：渗透检测适用于发现非多孔材料表面开口的缺陷，如裂纹、气孔、冷隔、折叠等。它特别适合于锻件、铸件、焊缝及机加工件的表面检测，能够有效识别这些部位的细微缺陷，并且具有高灵敏度、成本低、操作简便的特点。

问：超声波斜角探伤时，焊缝中的近表面缺陷不容易探测出来，其原因是什么？

答：在超声波斜角探伤时，焊缝中的近表面缺陷（通常指距离工件表面 10 毫米以内的缺陷）确实很难被有效地探测出来，这主要是由以下几个原因造成的：

1）探头盲区与声束扩散：近表面盲区：超声波探头在发射声波后，在其近场区内的声压分布非常复杂且不均匀，声束的能量波动剧烈。这个区域无法用于有效的缺陷探测，因此被称为盲区。近表面缺陷恰好位于这个区域内，导致其回波信号非常微弱或被强烈的声压波动所掩盖，难以被接收。

2）焊缝余高与表面反射：焊缝通常存在余高（焊缝凸起部分），这在斜角探伤时会产生强烈的干扰。表面反射干扰：超声波斜探头在扫查时，探头与工件表面之间形成一个角度。当声束经过焊缝余高时，会发生强烈的反射、折射和模式转换。这些表面反射产生的伪缺陷信号可能非常强，会遮盖住来自近表面缺陷的微弱回波信号，导致无法识别或误判。

2.6.3　Ministral-8B 问答示列

问：简述渗透检测的基本原理及适用范围。

答：渗透检测的基本原理是利用润湿作用和毛细管效应，通过渗透剂渗入材料表面开口的缺陷中，并再经过显像剂显现出来。具体步骤包括预清洗、施加渗透剂、去除多余的渗透剂、施加显像剂以及观察和评估。渗透检测的适用范围广泛，适用于多种材料如金属、非金属和复合材料，能够检查工件表面的各种缺陷，包括但不限于裂纹、气孔、折叠、分层等。

问：超声波斜角探伤时，焊缝中的近表面缺陷不容易探测出来，其原因是什么？

答：超声波斜角探伤时，焊缝近表面缺陷难以探测的原因主要包括以下几点：

1）盲区效应：超声波探头在发射脉冲后需要短暂恢复时间才能接收反射信号，导致近表面区域（通常为 5～10 mm）形成检测盲区。斜角探伤时，声波需通过楔块折射进入工件，进一步扩大了盲区范围。

2）表面波干扰：近表面缺陷的反射信号容易与探头直接激发的表面波（爬波）叠加，干扰缺陷识别。表面波能量集中在表层，可能掩盖微小缺陷的反射。

为了确保模型问答结果评估的客观性和可靠性，我们制定了一套详细的专家评估标准和流程。评估小组由十名拥有 10 年以上 NDT 实践经验的资深专家组成，针对不同模型给出的回答，他们独立对模型的回答进行打分和评论。研究团队设计了四个核心评估指标（准确性、专业性、完整性、可读性和结构），并为每个指标分配了权重，以反映其在 NDT 领域的关键性。所有指标均采用 5 分制（1 分代表非常差，5 分代表非常好），评分表见表 5。

表 5　专家评估指标与权重

评估类别	权重/%	评估标准:1 分	评估标准:3 分	评估标准:5 分
准确性	40	回答包含重大事实性错误，或内容完全虚假。	回答基本正确，但存在一些次要的、非关键性的错误或不精确之处。	回答完全正确，无任何事实性错误。
专业性	30	术语使用不当，概念混淆，回答缺乏专业性。	术语使用基本正确，但可能存在一些不常用的表述或欠缺深度。	术语使用精准无误，回答逻辑严谨，体现出专家级的理解深度。
完整性	20	回答过于简略，仅触及问题的表面。	回答涵盖了主要方面，但遗漏了部分重要信息或细节。	回答详尽且全面，提供了所有相关信息，无需额外补充。
可读性与结构	10	回答语言不通顺，逻辑混乱，难以理解其核心内容。	回答语言略显生硬，或结构稍显混乱，但仍可理解。	回答语言流畅，段落划分合理，要点突出，便于快速阅读和理解。

评估指标定义如下：

准确性：模型回答中的信息是否与 NDT 领域的理论、标准和最佳实践完全相符，不包含任何错误、误导性或虚假信息（“幻觉”）。

专业性：模型回答是否使用了正确的 NDT 专业术语，并以严谨、专业的口吻进行阐述。回答是否能体现对 NDT 复杂概念的深刻理解，而非简单的知识罗列。

完整性：模型回答是否全面地回答了用户的问题，涵盖了所有关键方面。

可读性与结构权重：模型回答的语言是否通顺、易于理解，并且结构清晰、逻辑分明。

评估过程遵循严格的步骤，最大限度地减少主观性偏差。过程中由十位专家独立对每个模型的回答进行评估，每位专家根据上述四个指标，为每个回答打分并撰写简要评论。专家在评估时不被告知回答来自哪一个模型，以排除对特定模型的偏见。评估后计算每个回答的加权总分：总分＝准确性×0.4＋专业性×0.3＋完整性×0.2＋可读性×0.1。取 3 位专家的评分平均值作为该回答的最终得分。将所有模型的最终得分进行比较，识别出表现最佳的模型，如果 10 位专家的评分存在显著差异（例如，分数离散度过大），评估小组将召开会议，共同讨论分歧原因，并重新审视回答。在达成共识后，给出最终评分。

结合专家的文字评论，分析每个模型的优势和劣势。通过这一严谨的评估流程，能够确保对模型的性能评估是基于客观、可量化的标准，为后续的模型改进提供了可靠的数据支持。

对前述问答示例进行评估后，相关情况总结如下：这 3 个模型给出的答案均展现出较高的精准度，但风格上各具特色，Qwen2.5-7 B 比较适合用于深度学习，Ministral-8 B 便于快速参考，Yi1.5-9 B 则在专业性和可读性两者之间取得较好的平衡状态。

3 结束语

本研究构建了 NDT 领域专业知识数据集，采用 LoRA 技术针对多个 LLM 进行 NDT 领域适配性微调。实验结果得出，经过微调的 LLM 展现出诸多方面的优势，在维持出色的语言泛化本领的同时，还使得专业术语表述的精准程度以及知识体系的完整程度都得到了提升。

参考文献：

[1] 巫彤宁. 国内垂直行业大模型发展现状及趋势 [J]. 信息化建设，2024 (10)：31－32.

[2] 白 硕. 语控万数，建设大模型垂直应用生态体系 [J]. 信息化建设，2024 (10)：38.

[3] 董子冰，王海虹，徐加祥. 浅谈人工智能中大模型微调技术和应用 [J]. 电信快报，2024 (11)：35－38.

[4] 张钦彤，王昱超，王鹤羲，等. 大语言模型微调技术的研究综述 [J]. 计算机工程与应用，2024，60 (17)：17－33.

[5] 王 浩，王 珺，胡海峰，等. PMoE：在 P-tuning 中引入混合专家的参数高效微调框架 [J]. 计算机应用研究，2025，42 (7)：1956－1963.

[6] 吴春志，赵玉龙，刘 鑫，等. 大语言模型微调方法研究综述 [J]. 中文信息学报，2025 (2)：1－26.

[7] 韩霄龙，曾 曦，刘 锟，等. 基于 LoRA 高效微调通用语言大模型的文本立场检测 [J]. 计算机与现代化，2025 (1)：1－6.

[8] 伍 洁. 面向大语言模型的参数高效微调方法的对比研究 [D]. 广州：广东外语外贸大学，2024.

[9] 贺 李，刘兴红，贾鹏宇. 基于 RAG 架构的智能知识库设计与应用研究 [J]. 信息记录材料，2025，26 (6)：127－128.

[10] 杨赟辉，程 虎，魏敬和，等. 面向 Transformer 模型边缘端部署的常用激活函数高精度轻量级量化推理方法 [J]. 电子学报，2024 (10)：3301－3311.

[11] 骆仕杰，金日泽，韩抒真. 采用低秩编码优化大语言模型的高校基础知识问答研究 [J]. 计算机科学与探索，2024，18 (8)：2156－2168.

[12] 汪 伦，艾斯卡尔·艾木都拉，张华平，等. 基于大语言模型的开源情报摘要生成研究 [J]. 情报理论与实践，2025，48 (5)：43－48.

[13] 王 菁. 大语言模型驱动的医院财务数据智能检索方法 [J]. 商业会计，2025 (7)：27－31.

[14] 国家市场监督管理总局，国家标准化管理委员会. 人工智能大模型第 2 部分：评测指标与方法：GB/T 45288.2－2025 [S]. 北京：中国质量标准出版传媒有限公司，2025：1－21.

[15] 陈建海，毛雨璐，沈智康，等. 赋能人工智能教学的问答数据集、微调大模型系统及优化研究 [J]. 广播电视网络，2024 (2)：14－20.

[16] 姜胜耀，袁 铖，朱立峰，等. 基于大语言模型微调的出院小结生成“幻觉”抑制方法 [J]. 医学信息学杂志，2025 (2)：14－21.

[17] 朱永康，高彦杰. 基于 LoRA 及其变体的通用大语言模型微调方法 [J]. 上海电力大学学报，2025，41 (1)：90－95.

[18] 王 浩，陈广磊，王 涵，等. 海关知识问答场景下的大语言模型应用研究 [J]. 中国口岸科学技术，2025，7 (2)：4－9.

[19] 黄星晨. 基于自适应量化的大语言模型微调方法 [J]. 信息技术与信息化，2024 (9)：9－12.

[20] 柴景贤，郎许锋，李红岩，等. 基于 LoRA 微调的轻量化中医药古籍大语言模型研究 [J]. 世界科学技术－中医药现代化，2025，27 (3)：823－831.