DOI: 10.19650/j.cnki.cjsi.J2514138

基于大语言模型的示波器智能控制系统研究

张士栋,叶 芃,张沁川,杨扩军,黄 川
(电子科技大学自动化工程学院 成都 611737)

摘 要:随着示波器功能不断丰富,其操作复杂性日益提高,用户在入门阶段面临较高门槛,即使掌握基本操作知识也难以充分利用其高级功能。为降低示波器的操作复杂性,提出了一种基于大语言模型的示波器智能控制系统。首先,该系统采用领域适配技术,通过构建结构化的示波器控制知识图谱生成领域优化提示词,以增强大语言模型对用户操作指令的理解能力。其次,系统引入语义检索技术,利用向量空间建模与近似最近邻搜索从知识图谱中筛选与用户操作指令最相关的知识片段,从而压缩提示词规模并提升推理效率。最后,系统通过融合这两种技术,构建“自然语言指令-标准可编程仪器命令-操作结果反馈”闭环控制机制,实现了利用自然语言对示波器全量功能的精准控制。实验结果表明,在自构建数据集测试中,相较于直接使用大语言模型生成标准可编程仪器命令,应用领域适配技术后的 qwen-max-latest 模型的标准可编程仪器命令生成准确率从 6.20% 提升至 99.6%;相较于仅应用领域适配技术,应用语义检索技术后在单张 RTX 4090 显卡上运行 qwen2.5-32b-instruct 模型,在保证推理精度损失<7% 的情况下,平均推理时延从 296 s 降低至 23.3 s。综上所述,所提出的示波器智能控制系统能有效降低示波器使用门槛,为实验仪器的智能化操作提供了技术支持,具有良好的应用前景与推广价值。

关键词: 大语言模型;示波器;智能控制;标准可编程仪器命令;知识图谱;语义检索
中图分类号: TH7 TP273 **文献标识码**: A **国家标准学科分类代码**: 460.40

Research on LLM-based intelligent oscilloscope control system

Zhang Shidong, Ye Peng, Zhang Qinchuan, Yang Kuojun, Huang Chuan
(*The School of Automation Engineering, University of Electronic Science and Technology of China, Chengdu 611737, China*)

Abstract: As the functionality of oscilloscopes continues to expand, their operational complexity has correspondingly increased, posing a significant learning barrier for novice users. Even those with basic operational knowledge often struggle to fully utilize the advanced features. To reduce the operational complexity of oscilloscopes, this study proposes an intelligent control system for oscilloscopes based on the large language model. Firstly, the system employs a domain adaptation technique by constructing a structured knowledge graph for oscilloscope control to generate domain-optimized prompts, thereby enhancing the large language model's ability to comprehend user instructions. Secondly, the system incorporates semantic retrieval techniques, utilizing vector space modeling and approximate nearest neighbor search to filter the most relevant knowledge fragments from the knowledge graph based on user instructions. This approach compresses the prompt size and improves inference efficiency. Finally, by integrating these two techniques, the system establishes a closed-loop control mechanism of "an natural language instruction-standard commands for programmable instruments-operational feedback", enabling precise control of the full range of oscilloscope functions through natural language. Experimental results demonstrate that on a self-constructed dataset, compared to directly using a large language model to generate standard commands for programmable instruments, the generation accuracy of the qwen-max-latest model improved from 6.20% to 99.6% after applying the domain adaptation technique. Furthermore, compared to using only domain adaptation, the incorporation of semantic retrieval technique, when running the qwen2.5-32b-instruct model on a single NVIDA RTX 4090 GPU, reduced average inference latency from 296 s to 23.3 s, while maintaining a loss inference accuracy of less than 7%. In summary, the intelligent oscilloscope control system proposed in this study effectively lowers the barrier to using oscilloscopes, provides technical support for the intelligent and automated operation of laboratory instruments, and demonstrates promising application prospects.

收稿日期:2025-06-10 Received Date: 2025-06-10

Keywords: large language model; oscilloscope; intelligent control; standard commands for programmable instruments; knowledge graph; semantic retrieval

0 引　　言

示波器作为电子测量领域的核心仪器,在电路调试、信号监测与故障诊断等场景发挥着不可替代的作用[1-4],但是随着测量需求的不断升级,示波器的功能也愈发复杂。以混合信号示波器(mixed signal oscilloscope,MSO)为例,除了支持多通道波形捕获以外,MSO 还需集成协议解码、实时频谱分析和眼图分析等高级功能,导致其综合可配置参数组合达到 10^4 量级,手工控制示波器效率呈指数级下降。

为了提升示波器控制效率,唐雷雷等[5]利用可编程仪器标准命令(standard commands for programmable instruments,SCPI)开发控制程序以实现示波器的自动化控制,但 SCPI 指令集规模庞大,以 Keysight Infiniium S 系列示波器为例,其指令手册超过 2 000 页,包含 1 200 余条独立命令,开发人员需花费大量时间研究示波器指令手册,导致控制程序开发效率低下。图形化编程工具 LabVIEW 可用于便捷开发仪器控制系统[6],Pavel 等[7]利用 LabVIEW 提升基于 SCPI 的示波器控制程序开发效率,但由于 LabVIEW 的控件功能不完整,控制程序难以覆盖所有的示波器功能;此外,不同厂商可能添加专有命令或参数[8],导致控制程序在适配不同示波器时存在兼容性问题,需要开发人员针对不同示波器指令手册对程序进行二次开发。Rao 等[9]提出的直接向示波器发送 SCPI 的控制方法虽然可以实现示波器功能的全量访问,但需要使用者频繁查阅示波器指令手册,导致人机交互体验差、操作效率低下。利用自然语言指令与 SCPI 的映射表,对自然语言进行分词处理并通过关键词匹配的方式将分词结果映射为 SCPI,可以实现基于自然语言的示波器控制[10],但该方法不具备语义容错能力,当用户输入包含模糊表达的指令时,关键词匹配机制难以将自然语言映射至精确的 SCPI。Lamaakal 等[11]提出大语言模型(large language model,LLM)具备将序列输入转换为多模态指令的潜力,但示波器领域适配瓶颈将制约 LLM 在示波器控制场景的应用[12]。使用提示词工程能约束 LLM 的输出空间[13],结构化的知识图谱能够提高 LLM 生成 SCPI 的准确率[14],但引入提示词将会增加 LLM 的推理时延[15],一方面需要大量提示词以确保 LLM 的生成精度,另一方面提示词过多会导致推理时延呈指数级增加。检索增强生成技术能有效整合外部知识库、提升生成内容的准确性和可控性[16],但将该技术应用于仪器控制领域,以利用其知识增强能力来简化提示词设计、确保控制指令的安全性与专业性,相关研究尚属空白。

针对上述挑战,本研究提出了一种基于 LLM 的示波器智能控制系统。该系统利用 LLM 的语义理解能力将用户输入的自然语言转换为符合示波器操作时序逻辑的 SCPI 序列,通过虚拟仪器扩展总线-11(virtual instrument extensions for instrumentation-11,VXI-11)协议控制示波器执行 SCPI,最后将执行结果反馈给用户,从而显著提升人机交互体验和操作效率。针对 LLM 的示波器领域适配问题,该系统通过构建示波器控制知识图谱并将其动态注入至 LLM,既可规避传统示波器控制程序繁琐的开发流程,又能支持示波器全量功能控制。实验证明相较于直接使用 LLM 生成 SCPI,使用示波器控制知识图谱灌注后的 qwen-max-latest 模型的 SCPI 生成准确率从 6.20% 提升至 99.6%。针对领域适配后 LLM 的推理时延问题,该系统通过语义检索技术根据用户输入的自然语言从示波器控制知识图谱中动态筛选出相关片段,有效缩减了提示词规模。实验证明相较于将全量示波器控制知识图谱作为提示词注入 LLM,应用语义检索技术后在单张 RTX 4090 显卡上运行 qwen2.5-32b-instruct 模型,平均推理时延从 296 s 降低至 23.3 s。

1 关键技术与系统架构

1.1 领域适配技术

对 LLM 进行领域适配的主流方法包括参数高效微调(parameter-efficient fine-tuning, PEFT)和提示词工程[17-19]。为了不受限于计算资源的限制,采用提示词工程对 LLM 进行示波器领域适配。本研究首创示波器控制知识图谱构建方法,使 LLM 利用有限的结构化提示词实现自然语言指令到 SCPI 的精准转换。

示波器控制知识图谱是基于示波器指令手册构建的三元组模板集合,由与 SCPI 一一对应的三元组模板构成,其中每个三元组模板包含指令原型、参数空间与功能描述这 3 类信息。三元组模板集合作为提示词发送给 LLM 后,LLM 基于语义解析机制,通过匹配三元组模板集合中的功能描述获取对应的指令原型,并根据参数约束将指令原型中的占位符替换为从用户指令中提取的实际值。该模板体系通过显式约束缩小 LLM 生成空间,有效抑制参数越界等语法错误,是 LLM 实现自然语言到 SCPI 精准转换的关键。示波器控制知识图谱的构建步骤如下文所示。

首先,定义三元组模板集合的表达式为:

$$\mathcal{T} = \{ t_i \mid t_i = \langle s_i, p_i, f_i \rangle \}_{i=1}^{N} \tag{1}$$

式中：N 为模板总数，是示波器指令手册中包含 SCPI 的个数；t_i 表示第 i 个三元组模板，由指令原型 s_i、参数空间 p_i 和功能描述 f_i 构成；s_i 明确 SCPI 的语法规则和组成结构；p_i 指定 SCPI 中占位符的取值范围或离散值集合；f_i 使用自然语言描述 SCPI 的功能及参数语义。

第 i 个三元组模板中指令原型 s_i 的表达式为：

$$s_i = SCPI_{param}^{(i)} \tag{2}$$

每个参数化的 SCPI 包含静态命令部分和动态占位符，如式(3)所示。

$$SCPI_{param}^{(i)} = BaseCommand^{(i)} + \sum_{k=1}^{K} \langle Wildcard_k^{(i)} \rangle \tag{3}$$

式中：$BaseCommand$ 表示静态命令部分，是字符串格式的固定命令结构；$Wildcard_k$ 表示第 k 个占位符，是待用实际值替换的参数，供 LLM 从参数空间 p_i 中取值替换；K 表示待替换占位符总数；$\sum$ 表示一个 SCPI 指令原型中可包含多个占位符；+表示静态命令片段与动态占位符的拼接。

第 i 个三元组模板中参数空间 p_i 表示指令原型 s_i 中所有占位符对应的合法参数集合的并集，如式(4)所示。

$$p_i = \bigcup_{k=1}^{K} P_k^{(i)} \tag{4}$$

参数空间 p_i 内，第 k 个合法参数集合 P_k 可分类为通道参数(channel parameter, CP)、数值参数(numerical parameter, NP)及枚举参数(enumerated parameter, EP)这 3 类，参数空间表达式如式(5)所示。

$$P_k^{(i)} \in \begin{cases} \{1,2,\cdots,N_{\max}\} & CP \\ \{p_{\min} + n \cdot \Delta \mid n \in \mathrm{N}\} & NP \\ \{e_1 \mid e_2 \mid \cdots \mid e_M\} & EP \end{cases} \tag{5}$$

式中：$N_{\max}$ 表示 SCPI 支持的最大通道数；$p_{\min}$ 表示参数最小值，如垂直刻度最小分辨率为 1 mV；n 为自然数索引，表示步数；Δ 表示参数步长；e_j 表示枚举值，如功能开关的枚举值为{OFF | ON}；M 表示枚举选项总数。

第 i 个三元组模板中功能描述 f_i 由基础描述和参数约束文本组合而成，如式(6) 所示。

$$f_i = BaseDesc^{(i)} + \sum_{k=1}^{K} \psi(Wildcard_k^{(i)}, P_k^{(i)}) \tag{6}$$

式中：$BaseDesc$ 表示功能描述的核心文本，对应 SCPI 的静态命令部分；$\psi(\cdot)$ 表示参数约束描述函数，将指令原型 s_i 的占位符 $Wildcard_k$ 与参数空间 p_i 的合法参数集合 P_k 结合，生成相应的参数约束文本。

1.2 语义检索技术

由于完整示波器控制知识图谱直接输入 LLM 会导致上下文长度激增与推理时延上升，本研究利用语义检索技术从示波器控制知识图谱中筛选与用户指令最相关的 Top-K 个知识图谱片段发送给 LLM，从而压缩提示词规模并提升推理效率。关键技术包括知识图谱向量化和 Facebook 人工智能相似性搜索[20](facebook artificial intelligence similarity search, FAISS)技术。

1)知识图谱向量化公式如式(7)所示。

$$\boldsymbol{h}_i = Pooling(SBERT(f_i)) \in \mathrm{R}^{384} \tag{7}$$

式中：$\boldsymbol{h}_i$ 为第 i 个 SCPI 对应的功能描述的向量表示；$SBERT(\cdot)$ 表示通过基于 Sentence-BERT 框架[21]的语义编码器将自然语言转换为高维语义向量；$Pooling(\cdot)$ 表示平均池化操作，将高维向量映射至 384 维。

针对示波器控制知识图谱的结构化特性，首先从知识图谱中提取所有 SCPI 的功能描述 f_i，再利用适配多种语言短文本语义相似性任务场景的预训练 paraphrase-multilingual-MiniLM-L12-v2 模型[22]，将变长功能描述文本块编码为 384 维固定维度的稠密向量，最终得到检索向量库。用户指令以同样的方式转换为稠密向量，作为查询向量。

2)对示波器控制知识图谱进行稠密向量表示后，可采用 FAISS 实现语义向量高效检索，准确获取与用户指令查询向量在隐空间内语义匹配度最高的 Top-K 个候选向量。

一般示波器控制知识图谱包含的三元组模版个数在 2 000 个以内，可采用 Flat 算法作为 FAISS 检索策略。Flat 算法的数学本质为暴力搜索，通过全量计算查询向量与检索向量库中每条向量的距离实现无精度损失的匹配。

设检索向量库包含 N 个 d 维向量，即：

$$\boldsymbol{F} = \{\boldsymbol{h}_1, \boldsymbol{h}_2, \cdots, \boldsymbol{h}_N\} \tag{8}$$

式中：$\boldsymbol{h}_i \in \mathrm{R}^d$ 表示第 i 个三元组模板中功能描述的向量化表示。

对于查询向量 $\boldsymbol{q} \in \mathrm{R}^d$，Flat 算法通过遍历计算 $\boldsymbol{q}$ 与所有 $\boldsymbol{h}_i$ 的相似度，并按降序返回 Top-K 个最相关的向量。相似度度量采用余弦相似度，即：

$$sim(\boldsymbol{q}, \boldsymbol{h}_i) = \frac{\boldsymbol{q} \cdot \boldsymbol{h}_i}{\|\boldsymbol{q}\| \|\boldsymbol{h}_i\|} \tag{9}$$

式中：分子为向量内积，分母为向量模长乘积的归一化因子。

通过最大化相似度函数确定最优匹配，即：

$$\boldsymbol{I}^* = \arg\max_{i \in \{1,\cdots,N\}} sim(\boldsymbol{q}, \boldsymbol{h}_i) \tag{10}$$

式中：$\boldsymbol{I}^*$ 为 Top-K 个最相关向量的索引，利用该索引可以从示波器控制知识图谱中获取对应的 Top-K 个最相关三元组模版。

1.3 系统架构

本研究提出的示波器智能控制系统包含前端界面、中间件、示波器、LLM 这 4 个模块，采用多模块解耦协同

架构实现从自然语言指令到示波器硬件控制再到结果反馈的全流程闭环,系统架构图如图 1 所示。

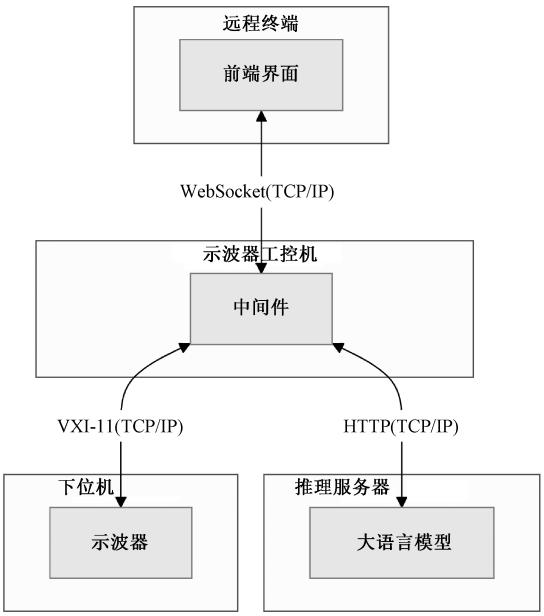


图 1　示波器智能控制系统架构

Fig. 1　Oscilloscope intelligent control system architecture

1)模块介绍

(1)前端界面

前端界面运行在示波器工控机或远程终端上,提供自然语言交互界面,支持异步消息收发以确保高并发控制场景下示波器能够按照顺序执行用户指令并反馈对应的执行结果。

(2)中间件

中间件部署于示波器工控机,作为连接前端界面、LLM 与示波器的核心枢纽,负责消息的转发和动态管理提示词,领域适配技术和语义检索技术均在此模块实现。

(3)LLM

LLM 运行在本地或云端推理服务器上,根据提示词生成推理结果。

(4)示波器

示波器作为系统的被控设备,执行中间件下发的 SCPI 并返回原始字节流数据。

2)系统工作流程

为进一步阐述各模块的动态协作机制,图 2 给出了系统工作时序流程。

(1)前端界面作为网络套接字(web socket, WebSocket)客户端向作为 WebSocket 服务端的中间件发起连接请求,建立双工通信通道;

(2)中间件动态加载示波器指令手册文档,利用特定的 SCPI(如*IDN)定位关键节点,通过标签匹配的方式从文档中筛选出指令原型、功能描述及参数空间,组织

图 2　示波器智能控制系统时序图

Fig. 2　Oscilloscope intelligent control system sequence diagram

为结构化字典列表,然后序列化为知识图谱,最终利用 paraphrase-multilingual-MiniLM-L12-v2 模型将知识图谱的功能描述集合转换为稠密向量检索底库;

(3)前端界面等待用户输入,将用户输入的自然语言指令通过 WebSocket 连接发送给中间件;

(4)中间件接受到自然语言指令后,将其转换为稠密向量,利用 Flat 算法计算其与检索底库中各向量的距离并排序,最终利用索引从知识图谱中筛选出 Top-K 条最相关知识图谱片段。为了进一步缩小 LLM 的生成空间、提升 SCPI 转换精度,中间件将系统提示词、自然语言指令、知识图谱片段和样例打包,构造混合提示词$\mathcal{H}$,即:

$$\mathcal{H}=[\ \{\text{“role”:“system”,“content”:}S\},\ \{\text{“role”:“user”,“content”:}\mathcal{T}'\},\ \{\text{“role”:“user”,“content”:}E\},\ \{\text{“role”:“user”,“content”:}q\}\ ] \tag{11}$$

式中:S 为系统提示词,定义模型行为范式;$\mathcal{T}'$ 是 Top-K 条最相关知识图谱片段;E 为(q_i,SCPI$_i$)对,是用户指令转换 SCPI 的样例;q 是用户输入的自然语指令。

(5)中间件调用 LLM 开放的应用程序编程接口(application programming interface,API)接口,将提示词通过超文本传输协议(hypertext transfer protocol,HTTP)发送给 LLM,并通过属性链式访问的方式从 LLM 返回消息中提取 SCPI 序列。LLM 返回消息包括响应标识、响应内容、模型名称、使用量统计等部分,其中 SCPI 序列包含在响应内容中。

(6)中间件对 LLM 返回的 SCPI 序列进行合法性校验后将其拆分成多条 SCPI,与示波器建立 VXI-11 协议连接,向示波器逐个发送 SCPI 并收集示波器反馈的原始字节流,最后关闭连接;

(7)中间件将示波器反馈的原始字节流发送给 LLM,从 LLM 返回消息中提取执行报告并转给前端界面,由前端界面将执行报告展示给用户。

3)核心设计原则

(1)利用标准协议实现模块解耦

为了同时支持本地和远程控制示波器,系统采用 WebSocket 协议实现前端界面与中间件的双向实时通信,利用其长连接免握手、低开销、全双工通信的特性,确保指令即时传输和执行结果实时反馈,提升交互效率与响应速度;为了兼容不同 LLM 厂商提供的 API 调用接口,系统采用 OpenAI 通用接口标准,基于 HTTP 协议实现 LLM 与中间件的通信;为了确保对主流厂商示波器的无缝支持,系统通过 VXI-11 协议实现示波器与中间件的高效互联,充分发挥其低延迟、二进制传输高效及跨品牌兼容的优势。

(2)可靠性增强设计

系统构建了从软件到硬件的双重容错防护体系,确保 LLM 生成的 SCPI 指令安全可靠执行。在软件层面,中间件通过正则表达式对 LLM 返回的 SCPI 序列进行白名单校验,该白名单为示波器控制知识图谱的指令原型集合,任何非法 SCPI 都将触发异常抛出告警,如图 3 所示。当软件校验层意外失效时,硬件层面的示波器内置安全机制将作为第 2 道防线,示波器接收到非法 SCPI 后将返回错误码,LLM 根据错误码生成执行报告提示错误信息。

图 3 SCPI 非法校验样例

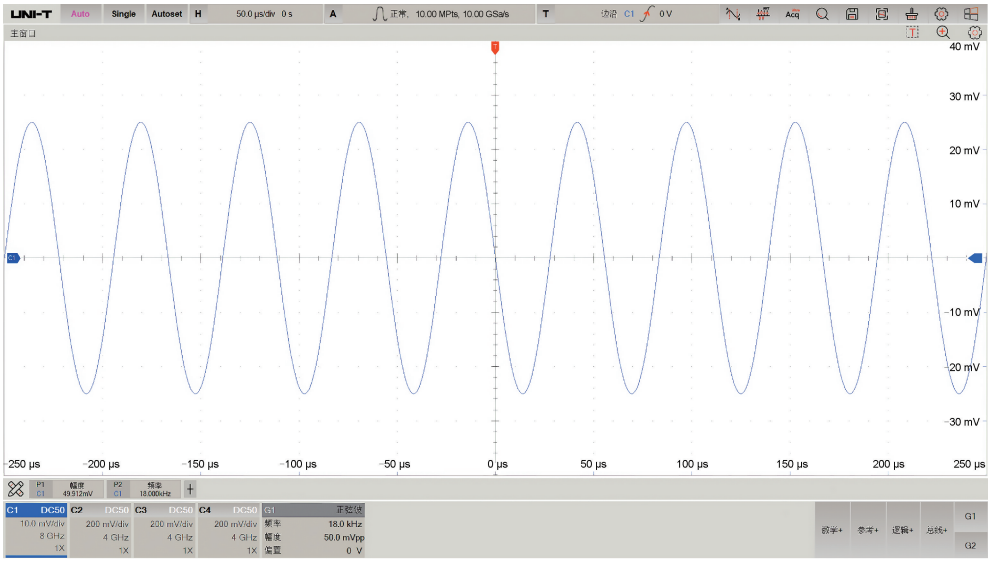
Fig. 3 SCPI syntax validation example

(3)LLM 可扩展性设计

LLM 支持灵活部署,当本地算力充足时,可自行使用多张推理卡搭建推理服务器集群,大幅提升 LLM 的响应速度;当本地算力资源不足时,也可直接调用厂商提供的云推理服务。

2 实验与结果分析

通过构建示波器智能控制系统,可以实现通过自然语言指令控制示波器,控制实例如图 4 所示。



(a) 示波器界面

(a) Oscilloscope interface

(b) 前端界面
(b) Frontend interface

图 4　示波器智能控制系统控制实例

Fig. 4　Control example of oscilloscope intelligent control system

系统的示波器控制效果主要依赖于 LLM 根据自然语言指令生成 SCPI 的准确率和推理时延。为了验证系统设计的有效性,本研究基于自构建数据集,选取 3 个参数量相对较小且支持输入 Token 长度>64 K 的 LLM,在单张 RTX 4090 显卡上进行推理,通过量化对比 LLM 生成 SCPI 的准确率和推理时延,对不同的提示词策略进行精度测试和性能测试,以此评估领域适配技术和语义检索技术对精度和性能的增益作用。由于实验平台仅配备单张 RTX 4090 消费级显卡,缺少推理服务器集群,算力资源有限,为了还原本系统在实际应用中的真实表现,本研究还选取了 7 个 LLM,通过 API 在线调用的方式对不同的提示词策略进行了精度和性能补充测试。

2.1　测试数据集构造

本研究基于 MSO7000X 示波器指令手册和专家操作经验,手工构建 500 条自然语言指令与对应的 SCPI 标签测试数据集,覆盖 3 种操作场景,如表 1 所示。

数据集包含 200 条基础操作指令、200 条参数配置指令和 100 条复合操作指令,覆盖了示波器的常用控制场景,共涉及对 729 项 SCPI 的测试,占 MSO7000X 示波器 SCPI 总数的 58.6%,能够有效反映整体统计特性。

2.2　基线对比方案

本研究将 3 种提示词策略作为对比基线,即:

1)无注入:发送给 LLM 的提示词中不包含示波器控制知识图谱,依靠 LLM 直接根据自然语言指令生成 SCPI;

表 1　操作场景及数据集示例

Table 1　Application scenarios and dataset examples

操作场景	示例指令	SCPI 标签
基础操作	显示通道 1 波形	:CHANnel1:DISPlay ON
参数配置	设置垂直刻度为 0.5 V/div	:CHANnel1:SCALe 0.5
复合操作	请保存当前波形	:FILe:WAVe:FORMat BINary; :FILe:WAVe:SOURce C1; :FILe:WAVe:NAMe Waveform; FILe:WAVe:SAVe ?

2)全量注入:应用了领域适配技术,提示词中包含完整的示波器控制知识图谱;

3)局部注入:应用了领域适配技术和语义检索技术,使用 FAISS 从示波器控制知识图谱中检索 Top-100 条相关片段注入到提示词中。

选取 qwen2.5-32b-instruct、qwen2.5-14b-instruct 和 qwen2.5-7b-instruct 模型在本地 RTX 4090 显卡上进行推理,选取 qwen2.5-32b-instruct、qwen2.5-14b-instruct、qwen2.5-7b-instruct、qwq-plus、qwen-max-latest、qwen-plus、qwen-turbo 模型调用 API 进行在线推理,记录 LLM 生成的 SCPI 与数据集 SCPI 标签完全一致的比例作为精度测试数据,LLM 生成 SCPI 的耗时作为性能测试数据。

2.3　实验结果

1)领域适配技术效果验证

为验证领域适配技术对 LLM 推理精度的提升效用,

本研究对比了无注入策略与全量注入策略在 10 个 LLM 上的 SCPI 生成准确率。实验结果如表 2 所示。

表 2 无注入策略与全量注入策略的大语言模型准确率对比

Table 2 Comparison of no-injection strategy and full-injection strategy on LLM accuracy (%)

提示词策略	本地 qwen2.5-32b-instruct	本地 qwen2.5-14b-instruct	本地 qwen2.5-7b-instruct	在线 qwen2.5-32b-instruct	在线 qwen2.5-14b-instruct	在线 qwen2.5-7b-instruct	在线 qwq-plus	在线 qwen-max-latest	在线 qwen-plus	在线 qwen-turbo
无注入	4.40	2.60	1.00	4.60	2.80	1.20	5.60	6.20	5.20	3.80
全量注入	97.40	93.00	65.20	98.60	93.60	66.60	98.80	99.60	98.00	75.60

实验结果显示,全量注入策略对 LLM 推理精度产生了显著的提升效应,在测试的 10 个不同规模的本地及在线模型上,全量注入策略的准确率均实现了数量级的提升:

在本地模型中,qwen2.5-32b-instruct 模型的准确率从无注入时的 4.40% 提升至全量注入后的 97.4%,提升超 22 倍;即便是参数量最少的 qwen2.5-7b-instruct 模型,其准确率也从 1.00% 显著提升至 65.2%。在在线模型中,qwen-max-latest 模型的准确率达到了 99.6%,相较于其无注入时的基线水平(6.20%)也有显著提升;参数量最少的在线 qwen2.5-7b-instruct 模型准确率从 1.20% 显著提升至 66.6%。无注入策略下所有模型的 SCPI 生成准确率均处于极低水平(1.00% ~ 6.20%),其原因主要在于 LLM 不具备 SCPI 先验知识,生成的 SCPI 与实际的 SCPI 存在格式、结构和参数上的差异。

通过两种策略的准确率对比可知,将结构化的示波器控制知识图谱全量注入提示词是解决 LLM 在示波器控制领域产生“幻觉”、输出错误指令的核心技术。实验结果证明,领域适配技术保证了 LLM 生成 SCPI 指令的准确性,是系统实现示波器精准控制的关键。

值得注意的是,本地部署模型的 SCPI 生成精度略低于在线调用相同模型的 SCPI 生成精度。其原因在于,为降低显存需求,本地部署的 qwen2.5 系列模型采用 bit 量化技术[23],将权重和激活值压缩为 4 位整数,从而牺牲了部分精度。

为验证领域适配技术对推理效率的影响,本研究还对比了无注入与全量注入两种策略的时延表现。实验结果如表 3 所示,全量注入策略以超线性增长的计算开销为代价换取 SCPI 生成精度。

表 3 无注入策略与全量注入策略的大语言模型推理时延对比

Table 3 Comparison of no-injection strategy and full-injection strategy on LLM inference latency (s)

提示词策略	本地 qwen2.5-32b-instruct	本地 qwen2.5-14b-instruct	本地 qwen2.5-7b-instruct	在线 qwen2.5-32b-instruct	在线 qwen2.5-14b-instruct	在线 qwen2.5-7b-instruct	在线 qwq-plus	在线 qwen-max-latest	在线 qwen-plus	在线 qwen-turbo
无注入	19.2	6.24	4.89	4.13	4.24	4.62	2.84	4.7	4.85	5.58
全量注入	296.0	95.00	46.00	15.40	63.00	60.90	18.90	28.9	26.60	19.50

在本地部署的 qwen2.5-32b-instruct 模型上,时延从 19.2 s 激增至 296 s,增幅达 15.4 倍;参数量最少的 qwen2.5-7b-instruct 模型推理时延亦从 4.89 s 升至 46.0 s,增幅达 9.41 倍。在线模型虽凭借云端算力维持了较低的绝对时延,但仍面临倍数级增长,其中在线 qwen2.5-14b-instruct 模型时延从 4.24 s 增至 63.0 s,增幅高达 14.9 倍;增幅最低的 qwen-turbo 模型亦从 5.58 s 增至 19.5 s,增幅达 3.49 倍。

这一结果表明,尽管全量注入策略能极大提升 SCPI 生成精度,但每次推理都需要将整个示波器控制知识图谱发送给 LLM,其中与自然语言指令无关的知识图谱片段会极大增加提示词规模,进而增加 LLM 的推理时延。该结果凸显了引入语义检索机制的必要性——需在精度损失可控的前提下,大幅压缩提示词规模以适配实时交互场景。

2)语义检索技术效果验证

为了验证语义检索技术对 LLM 推理效率的影响,本研究进一步评估了局部注入策略的推理效率。如表 4 所示,与全量注入策略相比,局部注入策略在所有测试模型上均实现了推理时延的数量级降低。

在本地模型中,参数量最多的 qwen2.5-32b-instruct 模型推理时延从全量注入的 296 s 骤降至 23.3 s,降幅高达 92.2%;参数量最少的 qwen2.5-7b-instruct 模型推理时延亦从 46.0 s 降至 5.67 s,降幅达 87.7%。下降趋势在

表 4　局部注入策略与全量注入策略的大语言模型推理时延对比

Table 4　Comparison of partial-injection strategy and full-injection strategy on LLM inference latency　(s)

提示词策略	本地 qwen2.5-32b-instruct	本地 qwen2.5-14b-instruct	本地 qwen2.5-7b-instruct	在线 qwen2.5-32b-instruct	在线 qwen2.5-14b-instruct	在线 qwen2.5-7b-instruct	在线 qwq-plus	在线 qwen-max-latest	在线 qwen-plus	在线 qwen-turbo
全量注入	296.0	95.0	46.00	15.40	63.00	60.90	18.90	28.90	26.60	19.50
局部注入	23.3	10.3	5.67	4.09	4.38	4.03	2.98	4.81	3.78	4.13

在线模型中表现一致:在线 qwen2.5-7b-instruct 模型的推理时延从 60.9 s 优化至 4.03 s,降幅达 93.4%;降幅最低的在线 qwen2.5-32b-instruct 模型亦从 15.4 s 增至 4.09 s,降幅达 73.4%。

此外,对比表 4 中局部注入策略与表 3 中无注入策略的推理时延可以发现,时延变化呈现有增有降的态势,且波动幅度普遍较小(多数在±20%范围),qwen-turbo 等在线模型的局部注入策略推理时延甚至更低,这主要源于云端模型性能的动态波动以及检索优化后提示词更加精准所带来的潜在加速效应。实验结果证明,语义检索技术通过过滤无关的知识图谱片段,使得向 LLM 发送提示词规模大幅降低,从而显著提升了系统的示波器控制效率。

为进一步评估检索机制对推理精度的影响,本研究对比了全量注入与局部注入两种策略的 SCPI 生成精度差异。实验结果如表 5 所示。

表 5　局部注入策略与全量注入策略的大语言模型准确率对比

Table 5　Comparison of partial-injection strategy and full-injection strategy on LLM accuracy　(%)

提示词策略	本地 qwen2.5-32b-instruct	本地 qwen2.5-14b-instruct	本地 qwen2.5-7b-instruct	在线 qwen2.5-32b-instruct	在线 qwen2.5-14b-instruct	在线 qwen2.5-7b-instruct	在线 qwq-plus	在线 qwen-max-latest	在线 qwen-plus	在线 qwen-turbo
全量注入	97.4	93.0	65.2	98.6	93.6	66.6	98.8	99.6	98.0	75.6
局部注入	92.6	88.6	58.8	93.0	89.6	60.4	93.6	95.6	91.6	70.6

具体来讲,SCPI 生成准确率降幅最高的本地 qwen2.5-7b-instruct 模型的准确率从全量注入的 65.2% 降至局部注入的 58.8%,仅下降 6.4 个百分点;降幅最低的在线 qwen-max-latest 模型的准确率从 99.6% 降至 95.6%,降幅为 4 个百分点,仍保持在较高水准。实验结果表明,局部注入策略导致的精度下降在可接受范围内,且在不同模型上表现稳定。

综合以上 3 种策略在准确率和推理时延上的对比可知,局部注入策略所导致的有限精度损失,与其带来的时延显著降低的收益相比,是一次极具价值的工程权衡。该策略成功地将计算资源集中于最相关的知识图谱片段上,证明了系统的优越性。

3　结　　论

针对示波器控制场景,本研究创新性地提出了示波器智能控制系统。该系统利用 LLM 的语义解析能力,结合领域适配和语义检索两大核心技术,实现了通过自然语言对示波器的快速精准控制,有效解决了传统手工操作模式的效率缺失、SCPI 编程模式的语义解析瓶颈和 LLM 在示波器控制场景中的领域适配难题。领域适配技术在 10 个测试模型中均显著提升了 SCPI 生成准确率(如 qwen-max-latest 从 6.20% 提升至 99.6%)。这表明注入结构化的示波器控制知识图谱能有效引导 LLM 理解复杂指令语义,准确生成 SCPI。语义检索技术在仅牺牲<7% 准确率的前提下,将提示词规模大幅压缩,推理时延最高降低 93.4%。领域适配和语义检索技术带来的准确率和推理效率的提升趋势在本地模型和在线模型上均得到一致体现,这表明本研究提出的提示词优化策略对于不同架构、不同规模的 LLM 均具有适用性和增益效果,后续研究者在搭建示波器智能控制系统时可根据硬件计算资源自行选择本地部署或在线调用合适的 LLM。为进一步提升系统的准确率和操作效率,本研究将尝试通过参数微调的方式训练 LLM,使其能够在更少的模型参数和更小的提示词规模下依然能够根据自然语言指令准确推理出 SCPI 序列。

参考文献

[1] 李承阳,田书林,杨扩军, 等. 高速数据采集系统中基于 EMD 的异常信号捕获方法研究[J]. 仪器仪表学报,2024,45(12):98-106.

LI CH Y, TIAN SH L, YANG K J, et al. Research on the anomaly detection method based on EMD in the high-speed data acquisition system[J]. Chinese Journal of Scientific Instrument, 2024,45(12):98-106.

[2] SÁNCHEZ-RODRÍGUEZ T, GÓMEZ-GALÁN J A, HINOJO-MONTERO J, et al. Simple power-efficient preamplifier-shaper channel for readout interface of silicon detectors[J]. AEU, 2025, 188: 155577.

[3] 江茹,李国超,郑浩, 等. 基于多传感器的加工过程智能监测系统设计[J]. 光学与光电技术,2024,22(1):67-76.
JIANG R, LI G CH, ZHENG H, et al. Intelligent monitoring system design for machining process based on multi-sensor[J]. Optics & Optoelectronic Technology, 2024,22(1):67-76.

[4] 董海迪,张瑞,王淅娜, 等. 基于时域反射法的电力电缆网络故障诊断方法研究[J]. 计算机测量与控制, 2024,32(3):30-36.
DONG H D, ZHANG R, WANG X N, et al. Research on fault diagnosis method of power cable network based on time domain reflectometry[J]. Computer Measurement & Control, 2024,32(3):30-36.

[5] 唐雷雷,卢平,孙葆根, 等. 合肥光源逐束团相位测量及纵向不稳定性诊断[J]. 强激光与粒子束,2021, 33(10):99-105.
TANG L L, LU P, SUN B G, et al. Bunch-by-bunch phase measurement and longitudinal instabilities diagnostics at Hefei Light Source[J]. High Power Laser and Particle Beams, 2021,33(10):99-105.

[6] 俞宙,李静,魏亚峰, 等. 基于虚拟仪器的高速混合信号自动测试系统设计[J]. 仪器仪表学报,2016, 37(S1):94-101.
YU ZH, LI J, WEI Y F, et al. Automatic test system of high-speed mixed-signal based on virtual instruments[J]. Chinese Journal of Scientific Instrument, 2016,37(S1): 94-101.

[7] PAVEL I, BRÂNZILĂ M, SĂRMĂSANU C, et al. LabVIEW based control and monitoring of a remote test-bench experiment for teaching laboratories[C]. 2021 International Conference on Electromechanical and Energy Systems, 2021: 398-402.

[8] CHANG Y Y, ZHANG Y ZH, KIOURTI A, et al. MPADA: Open source framework for multimodal time series antenna array measurements[C]. 2024 Antenna Measurement Techniques Association Symposium, 2024: 66-71.

[9] RAO A S, SAI D, MAHAJAN A T, et al. Development of python-based applications for virtual instrument control using PyQt5, PyVISA, and SCPI protocol[C]. 2024 Second International Conference on Emerging Trends in Information Technology and Engineering, 2024: 1-7.

[10] 中电科思仪科技股份有限公司. 一种用于数字示波器的智能语音识别与控制方法及系统:中国, CN119479626A[P]. 2025-02-18.
Ceyear Technologies Co., Ltd. Intelligent speech recognition and control method for digital oscilloscopes and system: China, CN119479626A[P]. 2025-02-18.

[11] LAMAAKAL I, MALEH Y, EI MAKKAOUI K, et al. Tiny language models for automation and control: Overview, potential applications, and future research directions[J]. Sensors, 2025, 25(5): 1318.

[12] KUMAR P. Large language models (LLMs): Survey, technical frameworks, and future challenges[J]. Artificial Intelligence Review, 2024, 57(10): 260.

[13] 张玲玲,黄务兰. 基于 ChatGPT API 和提示词工程的专利知识图谱构建[J]. 情报杂志,2025,44(3):180-187.
ZHANG L L, HUANG W L. Construction of patent knowledge graphs based on ChatGPT API and prompt engineering[J]. Journal of Intelligence, 2025,44(3): 180-187.

[14] IBRAHIM N, ABOULELA S, IBRAHIM A, et al. A survey on augmenting knowledge graphs (KGs) with large language models (LLMs): Models, evaluation metrics, benchmarks, and challenges[J]. Discover Artificial Intelligence, 2024, 4(1): 76.

[15] ZHANG W X, HUANG M, SONG ZH Y, et al. DimA: A parameter-efficient fine-tuning method with knowledge transfer based on transformer[C]. 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, 2024: 4922-4934.

[16] 刘雪颖,云静,李博, 等. 基于大型语言模型的检索增强生成综述[J]. 计算机工程与应用,2025,61(13): 1-25.
LIU X Y, YUN J, LI B, et al. Survey of retrieval-augmented generation based on large language

models[J]. Computer Engineering and Applications, 2025,61(13):1-25.

[17] KIM S, YANG H, KIM Y, et al. Hydra: Multi-head low-rank adaptation for parameter efficient fine-tuning[J]. Neural Networks, 2024, 178: 106414.

[18] LIN W, LIAO L CH. Lexicon-based prompt for financial dimensional sentiment analysis[J]. Expert Systems with Applications, 2024, 244: 122936.

[19] KHOBOKO P W, MARIVATE V, SEFARA J. Optimizing translation for low-resource languages: Efficient fine-tuning with custom prompt engineering in large language models [J]. Machine Learning with Applications, 2025, 20: 100649.

[20] DOUZE M, GUZHVA A, DENG CH Q, et al. The faiss library[J]. ArXiv preprint arXiv:2401.08281, 2024.

[21] REIMERS N, GUREVYCH I. Sentence-BERT: Sentence embeddings using Siamese BERT-networks[C]. Conference on Empirical Methods in Natural Language Processing 2019 and the 9th International Joint Conference on Natural Language Processing, 2019: 3980-3990.

[22] REIMERS N, GUREVYCH I. Making monolingual sentence embeddings multilingual using knowledge distillation[C]. 2020 Conference on Empirical Methods in Natural Language Processing, 2020: 4512-4525.

[23] YAO ZH W, AMINABADI R Y, ZHANG M J, et al. ZeroQuant: Efficient and affordable post-training quantization for large-scale transformers[C]. Proceedings of NeurIPS, 2022: 27168-27183.

作者简介

张士栋,2018 年于山东大学获得学士学位,2019 年于加州大学圣地亚哥分校获得硕士学位,现为电子科技大学在读博士生,主要研究方向为人工智能在示波器领域的应用、示波器智能控制与分析。

E-mail:202411060914@ std. uestc. edu. cn

Zhang Shidong received his B. Sc. degree from Shandong University in 2018, his M. Sc. degree from University of California, San Diego in 2019. He is currently pursuing his Ph. D. degree at University of Electronic Science and Technology of China. His main research interests include the application of artificial intelligence in the field of oscilloscopes, as well as intelligent control and analysis of oscilloscopes.

杨扩军(通信作者),2007 年于电子科技大学获得学士学位,2010 年于电子科技大学获得硕士学位,2015 年于电子科技大学获得博士学位,现为电子科技大学教授,主要研究方向为超带宽超高速数据采集系统、智能高速数据采集与处理、高速数字信号处理。

E-mail:yangkuojun@ uestc. edu. cn

Yang Kuojun (Corresponding author) received his B. Sc., M. Sc, and Ph. D. degrees all from University of Electronic Science and Technology of China in 2007, 2010, and 2015, respectively. He is currently a professor at University of Electronic Science and Technology of China. His main research interests include ultra-bandwidth and ultra-high-speed data acquisition system, intelligent high-speed data acquisition and processing, and high-speed digital signal processing.