doi:10.19562/j.chinasae.qcgc.2025.04.001

基于大语言模型的智能汽车仿真测试*

朱 冰,汤 瑞,赵 健,张培兴,李文旭,李嘉胜,徐雪峰
(吉林大学,汽车底盘集成与仿生全国重点实验室,长春 130022)

[摘要] 针对现有智能汽车基于场景测试方法严重依赖人力、效率瓶颈凸显的问题,本文提出了一种基于大语言模型的智能汽车仿真测试方法。首先,设计基于大语言模型的智能汽车仿真测试架构,建立了对应的数据层和仿真层;在此基础上,构建了基于大语言模型的智能汽车仿真测试流程,针对知识问答型任务设计了知识挖掘、模型微调与知识库增强检索应用流程,针对场景生成任务设计了场景类型分析、场景要素生成、场景工具链调用的应用路径,针对测试评价型任务,设计了测试场景解析、评价体系构建与仿真测试执行综合应用框架;最后,对各任务进行了测试。结果证明,本文所提出的测试方法可以有效解决不同类型的测试任务,提升测试效率。

关键词:智能汽车;仿真测试;大语言模型;场景生成;自动测试

Virtual Simulation Testing Method for Intelligent Vehicle Based on Large Language Model

Zhu Bing, Tang Rui, Zhao Jian, Zhang Peixing, Li Wenxu, Li Jiasheng & Xu Xuefeng
Jilin University, National Key Laboratory of Automotive Chassis Integration and Bionics, Changchun 130022

[**Abstract**] In this paper a simulation testing method for intelligent vehicle based on a large language model is proposed to address the issues of heavy reliance on human resources and prominent efficiency bottlenecks in existing scenario based testing methods. Firstly, a simulation testing architecture for intelligent vehicle based on a large language model is designed, and corresponding data and simulation layers are established. On this basis, an intelligent car simulation testing process based on a large language model is constructed. Knowledge mining, model fine-tuning, and knowledge base enhancement retrieval application processes are designed for knowledge question answering tasks. Application paths for scenario type analysis, scenario element generation, and scenario toolchain invocation are designed for scenario generation tasks. For testing and evaluation tasks, a comprehensive application framework for testing scenario analysis, evaluation system construction, and simulation testing execution is designed. Finally, each task is tested. The results show that the testing method proposed in this paper can effectively solve different types of testing tasks and improve testing efficiency.

Keywords: intelligent vehicle; simulation testing; large language model; scenario generation; automatic testing

前言

近年来,全球智能汽车产业发展迅猛,如何确保智能汽车的安全性、可靠性、舒适性以及高效运维能力成为行业关注的核心议题[1]。相较于随机、不确定的里程测试方法,基于场景的测试方法通过参数化定义场景,具有显著的效率和成本优势,已经成为智能汽车测试评价体系中不可或缺的重要环节[2]。

目前,智能汽车测试场景设计、评价标准确定、

*国家自然科学基金(U22A20247,52172386)和中国博士后科学基金(2023M741354,GZC20230945)资助。
原稿收到日期为 2024 年 06 月 07 日,修改稿收到日期为 2024 年 07 月 22 日。
通信作者:张培兴,助理研究员,博士,E-mail:zhangpeixing@jlu.edu.cn。

模拟仿真实施及测试报告撰写等关键环节严重依赖专业人员的知识积累与实践经验,具有极高的测试成本和操作难度。大语言模型(large language model,LLM)作为一种具有强大语言理解与生成能力的深度学习模型,已经在诸多领域展现出变革性的潜力[3]。Wen 等[4]设计了利用 GPT4.0 大模型进行知识驱动的自动驾驶框架,取得了与强化学习方法相当的性能。Shao 等[5]设计了集成场景描述、场景分析和分层规划思维链的自动驾驶系统,实现了强大的空间理解和实时推理速度。将大语言模型应用于智能汽车测试,有望简化测试流程,降低测试难度。然而,大语言模型应用需要大量高质量数据,目前缺少面向智能汽车测试的数据库。同时,现有模型与基于场景的仿真测试间缺乏交互接口,限制了其与智能汽车测试的一体化应用。由于缺乏对测试理论的深入理解,现有大语言模型直接生成场景存在盲目和不确定性,无法针对待测系统准确生成所需场景。

鉴于此,本文提出一种基于大语言模型的智能汽车仿真测试方法,主要贡献如下:

(1)针对传统基于场景的智能汽车仿真测试过程过于依赖专业人员、成本与难度极高的问题,提出一种基于大语言模型的智能汽车仿真测试架构,将智能汽车测试过程细分为 3 个核心任务:语言问答、场景生成和测试评价,并以数据层、模型层、仿真层和应用层的协同作用高效率完成各任务。

(2)针对现有 LLM 对于智能汽车测试知识匮乏、难以直接生成高质量的可用文本问题,构建数据层多维度知识库,系统收集智能汽车仿真测试相关知识文本,并对模型进行对话微调,增强 LLM 在智能汽车测试领域中的表现。

(3)针对 LLM 直接生成场景的盲目、不确定性以及现有大语言模型与基于场景的仿真测试间缺乏交互接口的问题,基于测试理论设计了场景生成、仿真测试、结果评价和报告生成工具链与工具应用流程,实现了基于 LLM 的智能化场景生成与测试评价整体流程。

1 基于大语言模型的智能汽车仿真测试架构

1.1 测试架构

本文设计的基于大语言模型的智能汽车仿真测试架构如图 1 所示,自下而上主要包含数据层、模型层、仿真层和应用层。

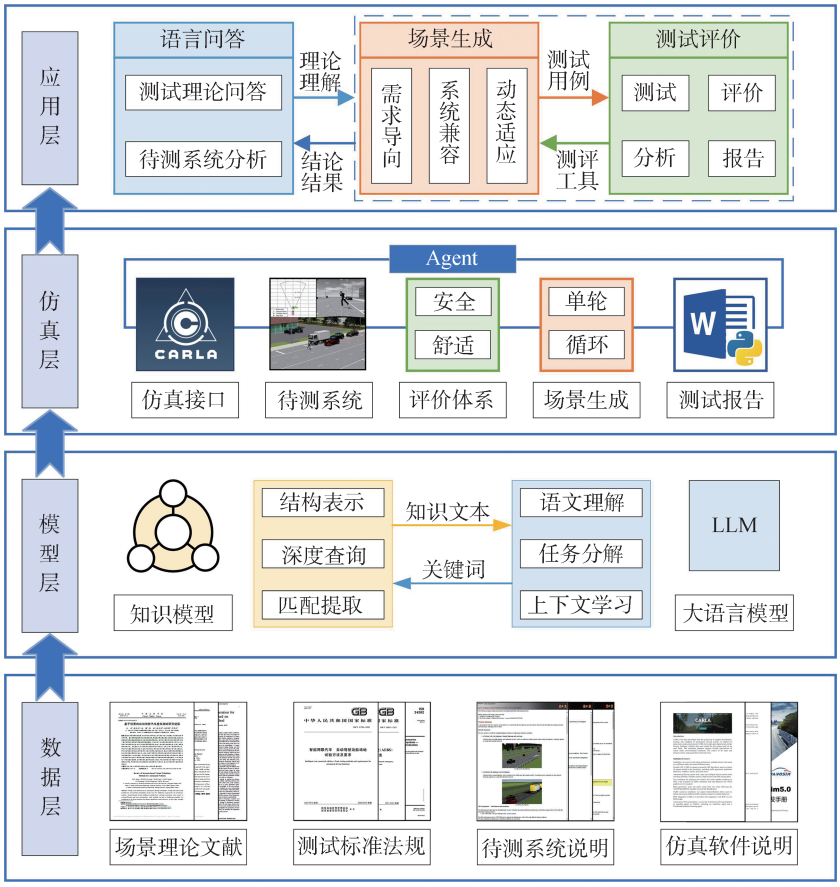

图 1 方法整体组成架构

数据层挖掘整理智能汽车测试相关文本数据,构成知识数据库。模型层定义包含知识结构模型、大语言模型的模型架构。仿真层定义用于仿真测试领域的各Agent功能,Agent是指通过语言描述任务细节并构建包含输入、输出的函数工具,以完成指定工作。应用层设计仿真测试的具体应用点,并定义各链路调取规则。各层级协同完成从输入语言指令到智能汽车测试评价的系统性任务。

本文模型层基于开源预训练大语言模型chatglm3-6b微调实现,大语言模型的微调方法有全量微调、P-Tuning、P-Tuning V2和LoRA等,P-TuningV2方法在每一层都加入Prefix前缀作为输入,在小规模任务中的性能可以达到全量微调效果,同时显著减少资源消耗[6],因此本文使用P-TuningV2方法对chatglm3-6b进行模型微调。基于本文的知识文本人工提取关键信息,构建问答对数据集,从原文档中找到相关答案。基于分词器将问答对转换为模型可理解的向量序列,并对序列进行始末位置标记与长度截断处理,设置学习率2×10^{-2},提示序列长度128,每个批次大小1,微调步骤3 000步,每1 000步保存模型。训练完成后,同时加载预训练模型权重和微调后的Prefix前缀权重,完成推理任务。本节重点介绍数据层、仿真层构建过程,应用层任务定义及语言问答中模型层知识调取准则在第2节中详细说明。

1.2　数据层知识库构建

大语言模型构建的关键在于提供丰富高质量的数据库[7]。为匹配智能汽车仿真测试任务,本文系统调研了与智能汽车测试相关的各类典型文本。根据测试过程,将文本数据概括为场景理论文献、测试标准法规、待测系统说明、仿真软件说明等文档,如图2所示。

1.2.1　场景理论文献

场景理论文献指对仿真测试场景中关键术语及定义等进行总结梳理的文献类型。本文将现有场景理论按照测试流程总结为:场景定义方法、测试场景生成、动力学与传感器建模方法、虚拟测试与评价方法。

针对测试场景定义,该类文献主要包括测试过程、测试场景的规范描述方法[8-9]。测试过程是在系统开发中,为验证其功能和性能等是否满足需求而执行的一系列有序活动,在智能汽车测试领域,包括用例设计、环境搭建、执行测试、生成报告等。测试场景是智能汽车实际应用过程中可能遇到情况的抽象和模拟,包括功能场景、抽象场景、逻辑场景和具体场景等不同层级。

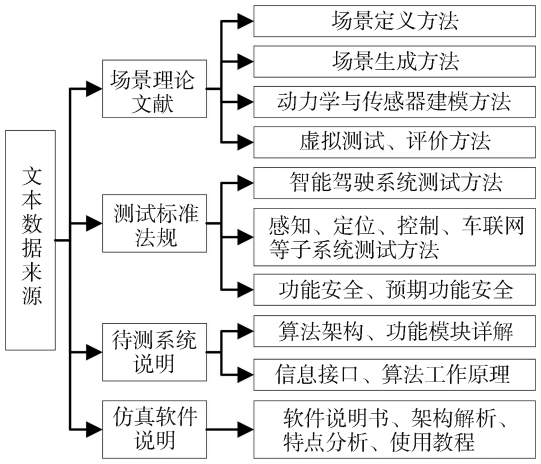

图2　收集文本数据分类

针对测试场景生成,现有文献可总结为现实世界语义抽象的场景类型生成[10]与具体测试过程量化的场景参数生成[11-12]两种。场景类型生成侧重于从复杂多样的现实世界中提炼出代表性的场景类别,每类代表一组具有相似场景特征或行为语义模式的事件序列。具体测试过程量化重点将定义好的场景类型转化为可执行的测试用例,细化各场景类型的具体条件和要素取值。

动力学与传感器建模方法用于准确模拟车辆在各种场景中的动态行为[13]以及传感器对周围环境的感知效果[14]。动力学建模可分为车辆动力学模型、场景动力学模型和交通动力学模型3部分。车辆动力学模型包括基于简化的物理模型和多体动力学模型。场景动力学模型包括道路模型和空气动力学模型。交通动力学模型包括微观交通流模型和宏观交通流模型等。传感器建模可分为相机、激光雷达、毫米波雷达、超声波雷达以及多传感器的数据融合算法,同时须考虑环境因素对传感器性能的影响。

智能汽车虚拟测试是智能驾驶技术开发过程中的关键环节[15]。按照待测系统的不同,可分为软件在环测试、硬件在环测试、车辆在环测试等。智能汽车评价方法旨在通过安全性、舒适性、智能性等多维度指标体系科学、全面地评判智能汽车在虚拟环境中的表现[16],确保智能汽车技术的应用效能。

1.2.2　测试标准法规

智能汽车测试标准涵盖了从整体系统到各子系统、从功能性能到安全属性的全方位评估[17]。本文以场景测试为核心目的,将现有智能汽车测试标准

总结为智能驾驶系统测试、各子系统测试、功能安全和预期功能安全测试3种。

智能驾驶系统测试遵循严格的验证与确认流程,须在实验室环境、封闭测试场地、仿真平台以及真实道路场景中进行[18]。测试方法涵盖功能测试、性能测试、耐久性测试、故障注入测试、边界条件测试、异常行为测试等。对于场景而言,重点强调场景库的建立管理,包括常见场景、边缘场景、极端场景和长尾场景的测试,确保系统在所有可能遇到的驾驶环境中具备稳健性能[19]。

对于感知、决策、规划、控制、车联网等子系统而言,各类标准详细定义了待测系统、测试内容、相关算法的联合测试体系[20]。以激光雷达为例,重点关注目标物检测性能及感知跟踪指标。测试内容包括目标分类、目标尺寸、目标位置、准确率、召回率、多目标跟踪精度、多目标跟踪准确度等[21]。通过模拟各类障碍物、交通参与者、复杂环境因素,评估感知系统对周围环境的全面、准确、及时感知能力。

功能安全[22]关注系统在设计和运行过程中因随机硬件故障、软件错误等原因导致的风险,通过故障模式与效应分析、故障树分析、故障注入测试等方法,评估系统的故障检测、诊断、隔离、缓解能力以及安全完整性等级。预期功能安全[23]关注由系统行为的不完全可预见性或外部环境不确定性引发的风险,通过场景分析、功能性能边界界定、假设分析等方法,评估系统的稳健性、鲁棒性以及对意外情况的应对能力。

1.2.3 待测系统说明

智能汽车系统是一个高度集成、复杂精密的工程系统,涵盖感知、决策、执行等多个层次以及车内、车际、车云等多种交互[1]。本文从算法架构、功能模块、信息接口、算法工作原理等方面对待测系统进行总结。

算法架构部分主要描述系统内部算法的层次结构、模块划分、数据流和控制流,揭示系统决策逻辑的全局视图。功能模块部分对待测系统进行拆解,详细说明算法内部模块的工作原理、算法实现、软硬件接口。信息接口部分表示系统内外部的信息交互标准和协议。算法工作原理部分解释核心算法(如目标检测、轨迹预测)的基本原理和数学模型。

1.2.4 仿真软件说明

智能汽车模拟仿真软件是设计、开发、测试和验证智能汽车的重要工具。本文从说明书、仿真架构、特点分析、使用教程4方面对仿真软件说明进行归纳。

软件说明书提供仿真软件的基本信息,帮助大语言模型快速入门软件的基础背景。通过对软件的仿真架构进行解析,揭示模块化设计、数据流、计算引擎、物理模型、环境模型、传感器模型、车辆动力学模型等核心组成部分,有助于理解软件的内在工作机制。进一步地,对软件特点进行分析,如高精度物理模型、丰富的场景库、强大的传感器仿真、灵活的脚本编程、与第三方工具的接口等,帮助大语言模型根据项目需求进行选择。最后,通过使用教程文档,提供详细的步骤指南,教授大语言模型理解基本操作。

1.2.5 知识库动态更新机制

由于智能汽车测试行业标准、法规快速变迁和技术迭代进步,全面且不断演进的知识库是保障模型生成结果准确性的关键。如图3所示,设计包含周期审查、专家咨询和反馈循环的知识库动态更新机制。

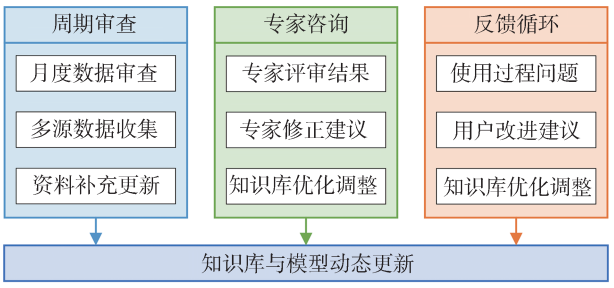

图3 知识库动态更新机制

1.3 仿真层测试模型构建

仿真层是将大语言模型应用到智能汽车测试领域的关键。本文设计包含仿真接口、待测系统、评价体系、场景生成、测试报告在内的仿真模型,待测系统集成在仿真平台中。各部分基于Agent格式实现,包含输入输出描述、工具用途描述,利于大语言模型调用。

1.3.1 仿真测试接口

仿真测试平台及其接口是模拟仿真测试的关键支撑,基于大语言模型的智能汽车功能验证需要高度可控、可重复且拟真度高的虚拟环境。CARLA[24]具有实时三维场景渲染能力、物理引擎支持以及交通元素模拟功能,同时使用与主流大语言模型相同的Python语言编程。因此本文基于CARLA仿真构建仿真测试平台,具体包含待测系统、场景搭建、场景执行、数据导出4个核心组件,如图4所示。

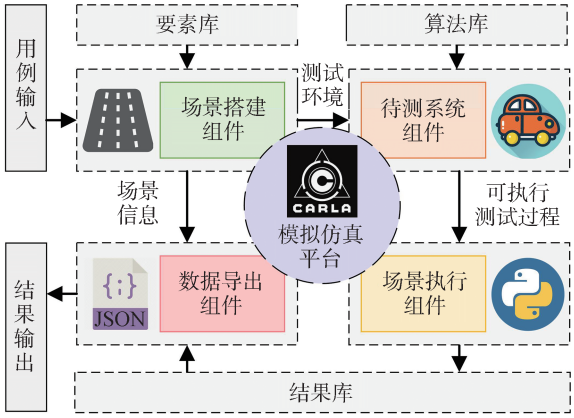

图4　仿真测试平台架构

场景搭建组件根据输入用例在要素库中选取需要的要素搭建测试场景,利用CARLA内部API,精细定制道路结构、交通标志、建筑物等静态场景元素;车辆、行人、自行车等动态交通参与者;光照、天气条件、路面状况等复杂环境特征。要素库定义了可用于环境初始化的各种要素,本文构建的元素有:车道宽度,车道id,纵向速度,横向速度,与车道的横向偏移,智能汽车与交通车之间的横向距离、纵向距离,交通车的纵向速度、横向速度,光照强度,天气条件(包含晴天、雨天、雾天,可定义各气象指标强度)。通过这些要素可以初步满足后续针对具体待测系统的测试任务。

待测系统组件在算法库中选取各类待测算法,并设置主车的传感器。为便于测试,本文内置基于相机感知和基于雷达感知的自动紧急制动(autonomous emergency braking,AEB)算法,并提供了说明文档用于大语言模型理解。

场景执行组件通过加载上述场景要素的具体取值初始化场景,并执行仿真模拟程序,过程中实时保存各交通参与者的运动学指标,包括车辆位置、速度、加速度、航向角、转向盘转角等信息到结果库。数据导出组件格式化在模拟过程中收集到的所有运动数据,将数据以标准化JSON格式导出合并保存。

1.3.2　过程评价体系

通过构建智能汽车场景测试结果评价体系,可以对测试过程中智能汽车算法性能进行客观评价。仿真测试平台在测试后会输出交通参与者的所有运动学信息。过程评价Agent根据这些运动学信息计算测试结果。本文以智能汽车主车为参考系,设计包含安全性和舒适性两方面的评价指标,如表1所示。

表1　测试结果评价指标体系

分类	安全性			舒适性		
名称	纵向碰撞时间	横向碰撞时间	安全距离	最大制动减速度	最大制动冲击度	平均减速度平滑度
数据	两车位置、速度			主车减速度		

安全性指标包含碰撞时间和安全距离两部分。碰撞时间表示保持当前运动状态不变,主车与其他车辆发生碰撞所需的预测时间。计算方法为当前时刻两车相对距离/相对速度,横纵向运动学分开计算,表示系统避免碰撞的反应时间。安全距离表示测试过程中主车与他车之间的瞬时相对距离,反映车辆的安全缓冲区域大小,评估算法保持安全间距的能力。

舒适性指标主要与主车的加速度信息有关,包含最大制动减速度、最大制动冲击度和平均减速度平滑度3部分。最大制动减速度表示测试过程中主车在制动中的最高减速度值,最大制动冲击度关注的是制动过程中减速度变化率的最大值,平均减速度平滑度表示整个制动过程中减速度时间序列的标准差,上述指标均关系到乘客在紧急制动情况下的身体感受。

1.3.3　场景生成模块

具体场景生成是智能汽车测试的核心,本文根据现有典型测试方法,设计单轮场景生成方法以及循环优化生成方法两种场景生成模块。

(1)单轮场景生成方法

单轮场景生成是指根据场景空间一次生成所有测试场景,主要用于随机性、覆盖性智能汽车场景测试工作。本文将单轮生成方法分为遍历生成、组合测试生成、蒙特卡洛采样生成3种。定义每种方法的应用特点如下。

遍历生成方法遍历场景空间的所有要素取值,确保每个场景至少被生成一次。随着场景空间复杂度增加,会导致场景数量呈指数级增长,变得难以实现且效率低下。组合测试假设大多数系统故障是由少数要素间的相互作用引起的,而非单要素的极端状态,通过覆盖性准则可以显著降低场景数量,同时满足覆盖性测试需求。蒙特卡洛采样通过概率统计分布,从场景空间中随机抽取样本点来拟合整个场景分布,适用于真实、随机、连续变量分布场景。

(2)循环优化生成方法

导致智能驾驶系统缺陷的关键场景具有更高的

测试价值,针对关键场景进行优化搜索研究可以提升测试效率。优化搜索流程如图 5 所示。

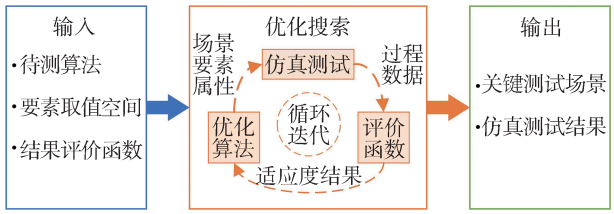

图 5 基于优化搜索的加速测试流程

给定待测算法、场景要素取值空间以及结果评价函数,利用优化搜索算法循环查找待测算法的性能缺陷,迭代搜索逐步逼近最有价值的场景组合,最终输出关键测试场景和仿真测试结果。为便于语言模型调用,设计内部搜索策略为遗传优化算法,评价函数和仿真测试过程均采用前述仿真测试工具。对于给定的关键场景数目需求进行优化生成,当某一轮生成数目大于预期数目或连续 3 轮未生成新场景,则流程结束。

1.3.4 测试报告模块

详实、准确的测试报告有助于识别并解决潜在的安全隐患,优化智能汽车算法。本文基于 python-docx 库设计测试报告生成工具。首先明确报告内容,将报告标题、测试对象、测试场景、测试执行过程、结果分析、性能改进建议等部分制成模板,定义各部分布局与样式,由此可以动态填充测试数据和相关文本信息。将工具编写为 Agent 调用格式,由此集成 LLM 完成自动填充信息、编写测试总结、分析建议等内容。

2 基于大语言模型的智能汽车仿真测试流程

2.1 测试任务划分

如图 1 中的应用层所示,本文根据智能汽车仿真测试关键环节,将基于大语言模型的智能汽车场景仿真测试任务划分为语言问答、场景生成、测试评价 3 个部分。

语言问答部分是大语言模型应用的基石,其为后续模块提供理论依据与系统理解,该模块应涵盖智能汽车测试的各方面知识及待测系统的性能缺陷、测试工具与测试重点等内容分析。

场景生成部分是测试领域的核心,为测试评价任务提供测试用例,该模块应与待测系统、测试平台、测试任务密切耦合,实现需求导向、系统兼容、动态适应的实用测试用例主动生成。

测试评价部分是大语言模型完成测试工作的关键,为场景生成任务提供测试评价平台,应包含测试、评价、分析、报告等关键环节,完成基于语言输入的完整测试评价工具链调用流程。

2.2 语言问答部分

现有预训练模型主要面向通用统一的语言类任务,难以满足智能汽车测试的各类需求。通过模型微调并结合知识库查询相关内容,并反馈到语言模型输出是解决特定领域任务的有效途径[25]。基于语言模型的知识库构建流程如图 6 所示,基于 Langchain 平台搭建。具体包含文本导入、文本分割、知识库向量化、检索向量化、相似性匹配与语言输出等步骤。

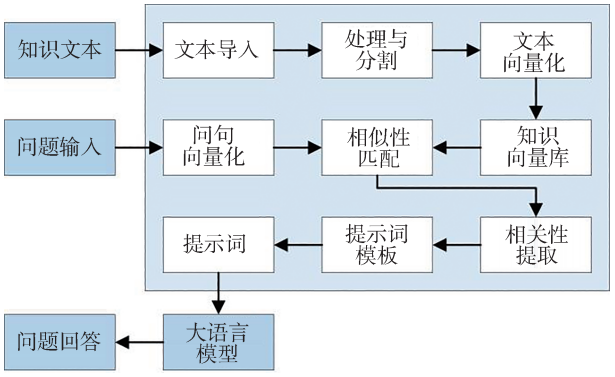

图 6 知识库构建与应用流程

将收集的各种格式文本导入 Langchain 架构;在文本处理阶段,去除无关字符、标点符号和特殊格式,在文本分割阶段,设置分割块大小限制为 200 字符、两个块之间共享的字符数量为 20;对分割后的文本进行向量化;基于 Chroma 建立知识向量库。对于输入问题而言,基于最大边际相关性的相似文本评价方法对向量化后的文本进行搜索,增加检索结果的多样性。通过提示词模板与检索结果结合构成提示词,输入语言模型即可整理输出查询结果。

2.3 场景生成部分

测试场景生成是智能汽车测试领域的核心,具体场景是模拟仿真测试的用例输入,是测试待测系统的关键。在具体场景生成过程中,大语言模型难以直接理解任务细节,会按照自我理解盲目生成不确定性场景。为使大语言模型生成准确的具体场景,本文设计如图 7 所示的工作流程,通过测试场景理论指导具体场景生成过程。

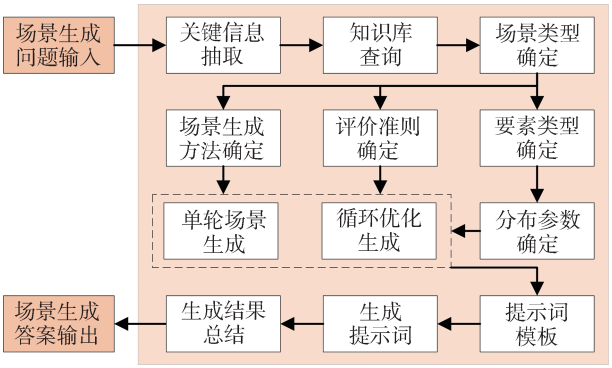

图7　场景生成工作流程模板

对于场景生成问题输入,抽取问题中的关键信息调用知识库查询,将测试场景生成需求解析为覆盖性单轮生成或关键性优化生成两种,对于各需求选择对应的场景生成方法和场景评价准则。获取待测系统的测试场景要素类型、取值范围、分布形式和离散步长,各部分要素推荐值已经存储在知识库中。为与测试评价工具链链接,考虑运动、气象和道路3部分要素对生成要素进行二次限制。结合提示词模板,将最终的场景生成结果转换为提示词提供给大语言模型,完成场景生成问题输出。

2.4　测试评价部分

测试评价部分是语言模型应用到场景测试领域的关键。对于模拟仿真测试而言,通过输入测试用例测试智驾算法,并对算法表现进行评价,发掘问题。据此设计如图8所示的工作流程。

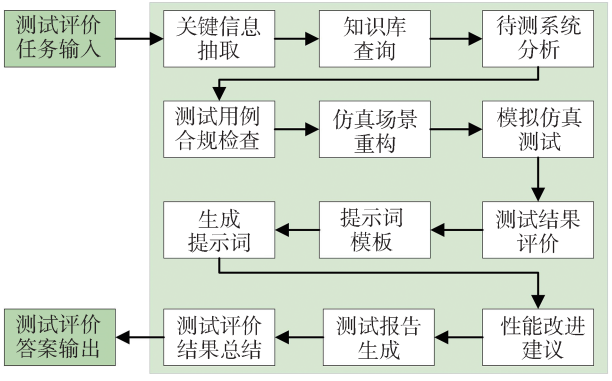

图8　测试评价工作流程模板

对于给定的待测系统和测试用例,进行关键信息抽取,进而转入知识库查询,对待测系统进行初步分析,并对测试用例是否可以在仿真平台上执行进行筛查,去除无法仿真测试的要素。将场景要素取值输入初始化的仿真场景中,实现自动化测试与过程数据输出,通过评价工具实现结果自动化评价。将测试评价结果转换为提示词,再次输入语言模型对待测系统进行改进分析,并生成测试报告。将测试评价结果反馈给用户。

3　测试实验与结果分析

3.1　语言问答型任务

3.1.1　任务实验设计

对于语言问答任务,重点考验语言模型对智能汽车场景测试领域的知识理解和逻辑推理能力。对chatglm3-6b模型[26]进行微调,集成知识库理解部分作为实验组,以原始chatglm3-6b、qwen-max[27]、GPT3.5[28]为对照组。使用同一份未参与微调训练和知识库构建的数据对所有模型进行问答测试,要求模型尽可能详细地回答。问答数据集具体涵盖场景理论、标准法规、待测系统、仿真软件、测试流程5个部分,每部分提供200个问题和推荐答案对。示例问题为:“①如何设计测试用例以覆盖所有合理的误用场景?②CARLA提供了哪些类型的车辆模型,以及它们的特点是什么?”

3.1.2　实验结果评价

为评估各模型在场景测试领域的认知效果,对所有组别进行问答测试,评价模型在领域问题回答的准确性。由于问答数据存在大部分非标准答案,基于人工打分结合大语言模型打分的方式计算,将问题、推荐答案、所有模型的输出一并提供给10名智能汽车测试专业人员和GPT3.5,要求各测试人员和GPT3.5输出各问题的难度评分(1-5)与模型回答的准确度评分(1-10),按照式(1)对各维度进行加权评价,结果如图9所示。计算方法如式(1)和式(2)所示。

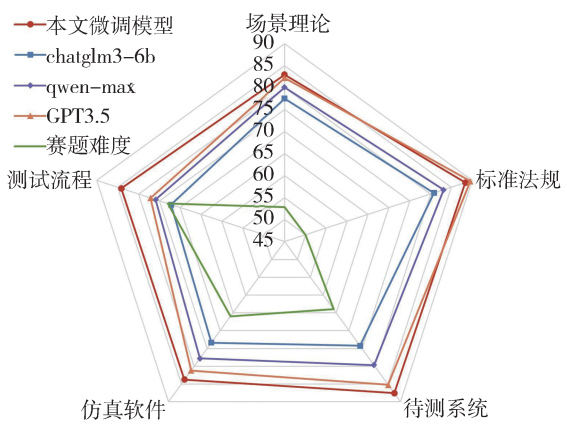

图9　模型语言问答任务评价结果

$$H=\sum_{1}^{11}(\omega_j\times\frac{\sum_{1}^{n}h_i}{n}) \tag{1}$$

式中:H 为该维度问题难度加权评价结果;h_i 为测评人员给出的难度评分;ω_j 为测评人员和 GPT3.5 的评分权值,测评人员取 0.05,GPT3.5 取 0.5;n 为该维度的问题数目。

$$S=\sum_{1}^{11}(\omega_j\times\frac{\sum_{1}^{n}a_i\times h_i}{\sum_{1}^{n}h_i\times 10}) \tag{2}$$

式中:S 为该维度回答准确性加权评价结果;a_i 为测评人员给出的回答准确性评分。

由图 9 可以看出,经过本文微调并集成知识库的语言模型在各部分均有较好的问答效果,在难度最高的测试流程部分中取得了明显性能提升。例如,在“如何利用仿真工具对 LKA 系统进行全面的测试评价?”问题中,本文模型可以在知识库测试标准的基础上,以 CARLA 软件为依托进行 LKA 系统测试流程、仿真接口、评价方法解析,相较于其他模型仅介绍基础流程的答案表现更好。在测试标准法规部分,GPT3.5 由于模型参数巨大,同时可以进行外部搜索资料,评分最高。在今后应用中,可以继续补充数据,并选用更多参数的模型进行优化设计。

3.2 场景生成型任务

3.2.1 任务实验设计

对于场景生成型任务,重点考察语言模型对场景理论的理解能力和场景生成工具链的应用效果。设计如下测试任务:“发掘 3 类用于基于激光雷达测距的 AEB 算法的测试场景,要求可在仿真测试平台上执行,考虑天气要素,①对于第 1 类,生成覆盖性为 1 的测试场景,②对于第 2 类,生成 100 个关键性场景。”

场景生成过程中,本文的语言模型通过匹配知识库找到了 CCRs(前车静止测试)、CCRm(前方慢行测试)、CCRb(前车制动测试)、VRU_Ped(行人紧急制动测试)等标准项目。基于仿真平台特点(直道双车场景,可模拟雨、雾)选择了前 3 类。场景要素设置为:主车速度、前车速度、两车间距、降雨强度,并基于知识库获取了对应要素的具体取值和分布情况。

对于任务①,本文的语言模型基于组合测试工具生成的覆盖性测试场景如表 2 所示。可以看出,大语言模型可以成功理解测试任务,确定场景要素,并调用组合测试工具生成了覆盖度为 1 的测试场景。

表 2 组合测试场景列表

序号	主车速度/(km·h^{-1})	前车速度/(km·h^{-1})	两车间距/m	降雨强度/(mm·h^{-1})
1	65	0	45	0
2	80	0	25	60
3	35	0	35	40
4	40	0	55	100
5	50	0	20	80
6	70	0	50	20
7	20	0	30	40
8	60	0	60	0
9	75	0	40	80
10	30	0	25	60
11	55	0	35	20
12	45	0	25	60
13	25	0	50	80

对于任务②,将本文的大语言模型的场景生成过程数据绘制如图 10 所示。可以看出,大语言模型成功理解测试任务,确定场景要素,以知识库中的碰撞时间作为关键性阈值,生成了满足要求的测试场景。在 43 次迭代时搜索完成 100 个场景,即结束搜索任务。

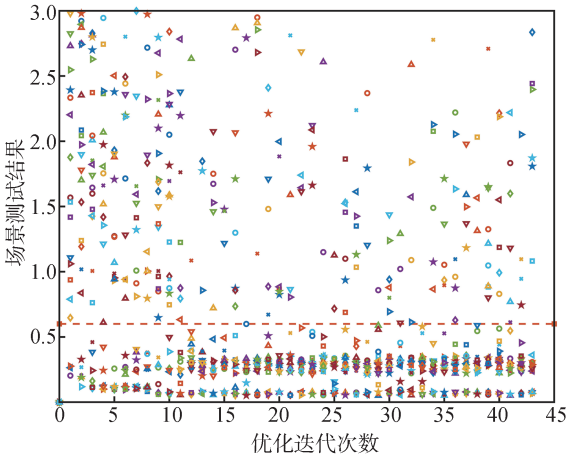

图 10 优化生成结果图像

为评价本文场景生成方法的有效性,以本文模型为实验组 A,以原始未经调整、没有集成知识库和场景生成 Agent 的大语言模型 chatglm3-6b 为对照组 B,为确保生成场景可执行,将文中构建的仿真平台描述集成到任务对话中。重复执行 100 次场景生成过程。

3.2.2　测试结果分析

为评价语言模型在不同场景生成任务上的应用效果,设计如下评价指标:

(1)知识索引成功率,表示大语言模型成功索引知识库查询对应测试项目的成功率;

(2)要素设计成功率,表示大语言模型根据待测系统与仿真平台特点合理设计场景要素、取值区间,确保场景可以成功在本文的仿真平台上执行的成功率;

(3)场景生成成功率,表示大语言模型在要素设计的基础上,按照方法生成测试场景的成功率,A组为调用Agent生成,B组为语言模型直接生成,关键场景有具体生成数量要求,满足数量要求才定义为成功生成;

(4)场景需求满足率,判断生成场景是否满足指定的覆盖性需求和关键性需求,覆盖性需求要求该轮次生成的场景最终覆盖度计算准确,以覆盖度达标的轮次比例评判,将最终仿真测试确定为关键的场景数目除以整体生成的关键场景数目定义为关键性需求满足率。

各指标计算结果如表3所示。

表3　场景生成任务结果评价表

生成方式	知识索引成功率	要素设计成功率	场景生成成功率	场景需求满足率
覆盖性需求-A	96.0%	93.8%	94.4%	100%
关键性需求-A	96.0%	93.8%	82.2%	100%
覆盖性需求-B		86.0%	77.9%	7.5%
关键性需求-B		86.0%	15.1%	11.2%

分析结果可知,A组中不同需求下的场景生成成功率均大于80%,由于生成过程基于Agent工具实现,成功执行工具即保证满足场景需求。A组的关键性生成过程中,需要仿真测试、结果评价、优化搜索等组件的联合调用,成功率低于覆盖性生成。B组由于未集成场景生成工具,生成过程仅靠LLM的自我理解与生成。根据待测系统特点设计要素,覆盖性需求下成功生成场景的比例较高,但大部分轮次下生成的场景都无法满足组合测试的覆盖性需求,最终完成率仅有7.5%。在关键性生成下,生成指定数量的关键场景存在较大困难,且由于基于语言的关键场景生成过程不与仿真平台相互耦合,最终需求完成率仅有11.2%。分析内部原因,LLM在生成文本时可能存在"幻觉"现象,生成不准确或虚构的信息,同时将数字信息进行向量转化时难以精准识别数学特征;此外,由于缺乏训练,在场景生成的过程中,LLM偶尔会用省略号代替部分内容,或需要人工介入。多因素的共同作用,导致纯LLM的生成结果与实际情况存在偏差。综上,本文构建的Agent工具流可以有效避免纯LLM生成场景的弊端,有效提升整体成功率。

3.3　测试评价型任务

3.3.1　任务实验设计

对于测试评价型任务,评估语言模型对测试任务的理解能力和整体测试评价工具链的使用效果。

将具体任务设定为:"根据指定的测试用例输入,完成某内置算法的虚拟仿真测试,并生成测试报告。分别针对2种待测算法进行测试,每种算法输入10个测试用例,要求模型完成测试和评价工作,并输出测试报告。重复执行50次,每次实验的测试用例取值均不同,每种算法累计需要测试500次。待测智能驾驶算法为待测系统组件中的2套AEB算法。"

测试过程中,语言模型接收任务需求后首先对测试用例进行了信息抽取与分析,将规范化后的测试用例输入仿真测试平台,并对测试结果进行了评价,最终系统性地分析了系统性能改进建议,并生成了测试报告,如图11(a)所示。测试过程的部分CARLA环境图像如图11(b)所示,说明语言模型可以根据指令控制仿真测试平台并执行测试过程。

(a) 测试报告　　(b) 测试过程图像

图11　模型测试评价任务结果

语言模型对于①算法的总结为:"系统在高速接近时表现紧张,安全距离不足,存在潜在风险。两级制动AEB系统能运行,但决策速度与安全缓冲须提升。优化算法减少决策延迟,提升纵向TTC;设定更

大安全距离标准,确保行车安全。”,说明语言模型可以理解算法核心原理(两级制动模型),分析测试结果并给出优化建议。

3.3.2 测试结果分析

重复执行了5次测试过程,为客观评价基于语言模型的场景测试评价任务的整体应用效果,针对各种Agent执行效果,定义如下指标:

(1)知识索引成功率,表示语言模型成功解析待测系统,调用知识库进行知识检索,优化测试用例的成功率;

(2)仿真执行成功率,表示语言模型成功解析优化后的测试用例,并通过模拟仿真Agent调用CARLA进行测试的成功率;

(3)结果评价成功率,表示语言模型成功使用评价Agent对CARLA导出的测试运动学信息进行性能评价的成功率;

(4)报告生成成功率,可准确无误地反映测试场景和评价数据的报告比例,当生成的报告能够完全匹配场景数据并提供有效的结果分析与性能改进建议时,才被视为成功的生成。

5次测试的统计结果如表4所示。

表4 测试评价任务成功率

AEB算法	知识索引成功率	仿真执行成功率	结果评价成功率	报告生成成功率
雷达测距	96.0%	91.3%	93.2%	80.0%
相机测距	94.0%	92.3%	94.7%	82.0%

分析结果可知,整体任务具备80%以上成功率,说明本文的测试评价应用路线具有一定意义。知识检索成功率很高,可有效利用知识库优化测试用例。执行CARLA仿真时,存在部分超出软件等待时长的导致失败的案例。由于软件输出评价结果为断续集合,导致评价Agent出现部分故障,难以准确计算指标数值,成功率下降。LLM可能存在幻觉现象,可以不经过原始数据的情况下直接生成测试报告,我们实施了严格的筛选机制以剔除此类不准确的报告。整体报告生成成功率大于等于80%,说明了本文构建的报告生成工作流程的可行性和实用价值。

4 结论

通过多Agent协同机制与知识库资源调用,初步实现了语言模型在智能汽车场景测试领域的应用实践。实验结果表明,本文提出的复合式应用架构可以初步完成测试领域核心的语言问答型、测试评价型和场景生成型任务,为语言模型在智能汽车测试领域的深度应用提供了思路。随着语言模型的继续深度发展,未来研究可以在此基础上拓展应用领域,在更复杂的智驾功能测试评价、精细化场景理解生成、深度知识融合推理和实时交通交互决策等方面进行深入研究与技术创新,语言模型有望在智能汽车场景测试领域发挥更大作用,成为推动自动驾驶技术进步的重要工具。

参考文献

[1] 崔明阳,黄荷叶,许庆,等.智能网联汽车架构、功能与应用关键技术[J].清华大学学报(自然科学版),2022,62(3):493-508.
CUI Mingyang, HUANG Heye, XU Qing, et al. Survey of intelligent and connected vehicle technologies: architectures, functions and applications[J]. Journal of Tsinghua University (Science and Technology), 2022, 62(3): 493-508.

[2] 朱冰,张培兴,赵健,等.基于场景的自动驾驶汽车虚拟测试研究进展[J].中国公路学报,2019,32(6):1-19.
ZHU B, ZHANG P X, ZHAO J, et al. Review of scenario-based virtual validation methods for automated vehicles[J]. China Journal of Highway and Transport, 2019, 32(6): 1-19.

[3] ZHAO W X, ZHOU K, LI J, et al. A survey of large language models[J]. arXiv preprint arXiv:2303.18223, 2023.

[4] WEN L, FU D, LI X, et al. Dilu: a knowledge-driven approach to autonomous driving with large language models[J]. arXiv preprint arXiv:2309.16292, 2023.

[5] SHAO H, HU Y, WANG L, et al. Lmdrive: closed-loop end-to-end driving with large language models[J]. arXiv preprint arXiv: 2312.07488, 2023.

[6] LIU X, JI K, FU Y, et al. P-tuning v2: prompt tuning can be comparable to fine-tuning universally across scales and tasks[J]. arxiv preprint arxiv:2110.07602, 2021.

[7] YANG J, JIN H, TANG R, et al. Harnessing the power of LLMs in practice: a survey on ChatGPT and beyond[J]. ACM Transactions on Knowledge Discovery from Data, 2024, 18(6): 1-32.

[8] MENZEL T, BAGSCHIK G, MAURER M. Scenarios for development, test and validation of automated vehicles[C]. 2018 IEEE Intelligent Vehicles Symposium (IV). IEEE, 2018: 1821-1827.

[9] International Organization for Standardization. Road vehicles—test scenarios for automated driving systems-vocabulary: ISO 34501:2022[S]. Geneva: ISO, 2022.

[10] CORSO A, DU P, DRIGGS-CAMPBELL K, et al. Adaptive stress testing with reward augmentation for autonomous vehicle validatio[C]. 2019 IEEE Intelligent Transportation Systems Conference (ITSC). IEEE, 2019: 163-168.

[11] 朱冰,汤瑞,赵健,等.基于代理遗传优化的智能驾驶系统加速测试方法[J].同济大学学报(自然科学版),2024,52(4):501-511.
ZHU B, TANG R, ZHAO J, et al. Accelerated test method of intelligent driving system based on surrogate genetic optimization model [J]. Journal of Tongji University (Natural Science), 2024, 52(4): 501-511.

[12] ZHU B, ZHANG P, ZHAO J, et al. Hazardous scenario enhanced generation for automated vehicle testing based on optimization searching method [J]. IEEE Transactions on Intelligent Transportation Systems, 2021, 23(7): 7321-7331.

[13] 方培俊, 蔡英凤, 陈龙, 等. 基于车辆动力学混合模型的智能汽车轨迹跟踪控制方法[J]. 汽车工程, 2022, 44(10): 1469-1483.
FAHG P J, CAI Y F, CHEN L, et al. Trajectory tracking control method for intelligent vehicles based on a hybrid vehicle dynamics model [J]. Automotive Engineering, 2022, 44(10): 1469-1483.

[14] ZHU B, SUN Y, ZHAO J, et al. Millimeter-wave radar in-the-loop testing for intelligent vehicles[J]. IEEE Transactions on Intelligent Transportation Systems, 2021, 23(8): 11126-11136.

[15] 赵树廉, 来飞, 李克强, 等. 基于数字孪生技术的智能汽车测试方法研究[J]. 汽车工程, 2023, 45(1): 42-51.
ZHAO S L, LAI F, LI K Q, et al. Research on intelligent vehicle test methods based on digital twin technology [J]. Automotive Engineering, 2023, 45 (1): 42-51.

[16] 陈君毅, 李如冰, 邢星宇, 等. 自动驾驶车辆智能性评价研究综述[J]. 同济大学学报 (自然科学版), 2019, 47(12): 1785-1790,1824.
CHEN J Y, LI R B, XING X Y, et al. Survey on intelligence evaluation of autonomous vehicles[J]. Journal of Tongji University (Natural Science), 2019, 47(12): 1785-1790,1824.

[17] 朱冰,范天昕,张培兴,等.智能网联汽车标准化建设进程综述[J].汽车技术,2023(7):1-16.
ZHU B, FAN T X, ZHANG P X, et al. Review of the standardization construction process of intelligent connected vehicles [J]. Automotive Technology, 2023 (7): 1-16.

[18] 胡大林, 何丰, 薛晓卿, 等. 自动驾驶车辆模拟仿真测试平台技术要求: T/CMAX 121—2019[S]. 北京: 中关村智通智能交通产业联盟, 2019.
HU D L, HE F, XUE X Q, et al. Technical requirement for automatic driving vehicle simulation test platform in Beijing: T/CMAX 121—2019[S]. Beijing: CMAX, 2019.

[19] International Organization for Standardization. Road vehicles—test scenarios for automated driving systems—scenario based safety evaluation framework: ISO 34502: 2022 [S]. Geneva: ISO, 2022.

[20] 中国汽车工程学会. 智能网联汽车V2X系统预警应用功能测试与评价方法: T/CSAE 246—2022[S]. 北京: 中国汽车工程学会, 2022.
China-SAE. Functional test and evaluation method of V2X system warning application of intelligent connected vehicle: T/CSAE 246—2022[S]. Beijing: China-SAE, 2022.

[21] 中国汽车工程学会.《智能网联汽车激光雷达感知评测要求及方法》标准立项[EB/OL]. (2021-07-14)[2022-12-01]. http://www.caicv.org.cn/index.php/newsInfo? id=381.
China-SAE. Technical requirement and testing method of intelligent connected vehicle lidar perception standard project establishment[EB/OL]. (2021-07-14)[2022-12-01]. http://www.caicv.org.cn/index.php/newsInfo? id=381.

[22] International Organization for Standardization. Road vehicles—functional safety:ISO26262:2018 [S]. Geneva: ISO, 2018.

[23] International Organization for Standardization. Road vehicles—safety of the intended functionality: ISO21448: 2018 [S]. Geneva: ISO, 2018.

[24] DOSOVITSKIY A, ROS G, CODEVILLA F, et al. CARLA: An open urban driving simulator[C]. Conference on Robot Learning. PMLR, 2017: 1-16.

[25] LUO H, TANG Z, PENG S, et al. ChatKBQA: a generate-then-retrieve framework for knowledge base question answering with fine-tuned large language models [J]. arXiv preprint arXiv: 2310.08975, 2023.

[26] ZENG A, LIU X, DU Z, et al. GLM-130B: an open bilingual pre-trained model[J]. arXiv preprint arXiv:2210.02414, 2022.

[27] BAI J, BAI S, CHU Y, et al. Qwen technical report[J]. arXiv preprint arXiv:2309.16609, 2023.

[28] OUYANG L, WU J, JIANG X, et al. Training language models to follow instructions with human feedback[J]. Advances in Neural Information Processing Systems, 2022, 35: 27730-27744.