基于大语言模型的自动驾驶仿真测试场景生成

陈 贞[1]，李京泰[2]，郭 煌[3]，贾贝贝[4]，李广友[2]

（1. 北京镝石数据科技有限公司，北京 100176；2. 工业和信息化部 装备工业发展中心，北京 100846；
3. 北京赛目科技股份有限公司，北京 100080；4. 北京新能源汽车股份有限公司，北京 100176）

摘　要：随着自动驾驶技术的快速发展，仿真测试场景对真实性和多样性的要求不断提高。然而，传统的自动驾驶仿真场景构建方法依赖人工编辑，不仅成本高昂，而且面临场景要素组合和复杂度受限的问题，难以满足自动驾驶系统全面测试验证的需求。为解决这一问题，提出了一种基于大语言模型（LLMs）的自动驾驶仿真测试场景生成方法。该方法基于预训练大语言模型，通过LoRA微调，结合场景语言解析，输出一种结构化的解释性语言，用于生成仿真场景文件。生成的文本经过解析器转换为可用的仿真场景文件，有效解决单个生成文本过长和模型幻觉等问题，实现了通用模型能力的专用化，显著提高了自动驾驶仿真场景生成的效率和多样性。

关键词：人工智能；自动驾驶；场景生成大模型；仿真

中图分类号：TP18　　**文献标志码：**A　　**DOI：**10.3969/j.issn.2095–1469.2025.03.06

Simulation Test Scenario Generation for Autonomous Driving Based on Large Language Models

CHEN Zhen[1]，LI Jingtai[2]，GUO Huang[3]，JIA Beibei[4]，LI Guangyou[2]

（1. Beijing Dysprosium Stone Data Technology Co.，Ltd.，Beijing 100176，China；
2. Equipment Industry Development Center， Ministry of Industry and Information Technology，Beijing 100846，China；
3. Bejing Saimo Technology Co.，Ltd.，Beijing 100080，China；
4. Beijing Electric Vehicle Co.，Ltd.，Beijing 100176，China）

Abstract: The rapid development of autonomous driving technology has increased the demand for the authentic and diverse simulation test scenarios. However, traditional methods for constructing autonomous driving simulation scenarios heavily rely on manual editing, which is not only costly but also limited by the combination and complexity of scene elements, making it difficult to meet the comprehensive testing and validation needs of autonomous driving systems. To address this issue, this paper proposes a method for generating autonomous driving simulation test scenarios based on Large Language Models (LLMs). This approach utilizes a pre-trained LLM, enhanced through LoRA fine-tuning, and integrates a scenario language parser to produce a structured interpretive language, which is used to generate scenario files. The generated text is processed by a parser to convert it into usable scenario files, effectively addressing the issues of overly

收稿日期：2025–01–22　**改稿日期：**2025–04–09

基金项目：新一代人工智能国家科技重大专项（2022ZD0116311）

参考文献引用格式：

陈贞，李京泰，郭煌，等.基于大语言模型的自动驾驶仿真测试场景生成[J].汽车工程学报，2025，15(3)：329–339.

CHEN Zhen，LI Jingtai，GUO Huang，et al.Simulation Test Scenario Generation for Autonomous Driving Based on Large Language Models[J].Chinese Journal of Automotive Engineering，2025，15(3)：329–339.(in Chinese)

long texts and model hallucinations, while also achieving the specialization of a general model's capabilities.

Keywords: artificial intelligence; autonomous driving; scenario generation model; simulation

随着自动驾驶技术的快速发展，自动驾驶系统的安全测试与验证成为产业关注的关键领域。当前，基于场景的模拟仿真、封闭场地测试、实际道路测试、安全监测等“多支柱”测试与评估方法，已成为自动驾驶系统安全测试与评估的行业共识[1-2]。其中，模拟仿真测试由于在安全性、成本效益、测试效率等方面具备显著优势，不仅是自动驾驶系统安全测试与评估的重要组成部分，也是自动驾驶系统开发流程中不可或缺的环节。仿真测试能高度逼真地模拟真实世界的驾驶场景，验证自动驾驶系统在各种复杂交通环境中的功能和性能，并评估系统的鲁棒性、安全性、可靠性等。同时，仿真测试还能在受控环境下全面验证自动驾驶系统，特别适用于现实世界中难以测试或测试成本高昂的极限和失效场景，有效降低实车测试的风险和成本[3-5]。

传统的仿真测试场景构建通常依赖人工操作，通过道路编辑器和场景编辑器手动逐一搭建路网信息和动态元素。这种方式效率低、成本高，且受到人工经验的局限，难以全面覆盖所有可能的驾驶场景，尤其是罕见但安全关键的自动驾驶长尾场景[6]。因此，如何高效地生成全面覆盖设计运行条件的测试验证场景集，已成为自动驾驶模拟仿真测试的核心技术难点。2018年，SCHULDT等[7]提出了一种基于组合测试理论的L2级自动驾驶系统的测试场景生成方法，减少了必要的测试用例数量，也适用于仿真测试场景的泛化生成。同年，KOREN等[8]基于自适应压力测试方法，提出了一种自动驾驶危险场景生成方法，结合蒙特卡洛树搜索和前馈神经网络来寻找碰撞场景。2019年，GAO Feng等[9]提出了一种考虑道路模型、交通规则、车辆行为等因素的测试矩阵方法，基于组合测试理论，实现了智能驾驶测试场景的自动生成。2023年，龚磊等[10]设计了一种能简洁描述场景路网结构的语言SceneRoad，并集成于场景描述语言中，实现了随机生成大量静态场景，用于构建仿真数据集。

随着基于人类反馈的强化学习、指令微调等技术的发展，人工智能进入了大模型时代。大模型凭借其强大的学习能力、泛化能力、预测能力和涌现性，为自动驾驶仿真测试场景生成带来了新的可能性。2023年6月，英国WAYVE[11]发布了多模态GAIA-1世界模型，通过结合文本、图像、视频等多模态信息，生成具有智能驾驶车辆行为和场景特征的交通场景。2024年，CUI Can等[12]将大语言模型（Large Language Models，LLMs）应用于自动驾驶系统，进行了模拟仿真和实车测试，结果表明，大语言模型有助于增强自动驾驶的感知理解和决策规划等行为。TIAN Haoxiang等[13]基于LEADE多模态大模型，从真实交通视频中构建真实场景，能用于进行L4级自动驾驶系统的安全测试与验证。

但就整体来看，大模型在自动驾驶场景生成应用方面的研究还处于发展阶段，如何构建效率更高的场景生成模型，如何解决模型幻觉导致的生成内容准确性与真实性不足、场景语言文本过长等问题，仍然是学术界需要解决的问题。

传统手工编写的OpenScenario 2.0场景格式文件，其采用领域特定语言（DSL）编写场景文件（.osc格式），通常在给定场景的条件下，工程师编写场景文件的耗时为2～3 h，且还需要在实际仿真环境中进行调试验证。现有研究主要聚焦于场景文本描述生成，而本文提出的场景生成大模型，是基于微调预训练大模型，输出一种结构化的解释性语言，再经场景语言解析器将模型输出解析成OpenScenario 2.0场景文件。本文通过将大语言模型生成的场景文本转换为结构化场景文件，优化信息表达，提高仿真系统的适配性，实现了工程师利

用自然语言描述快速生成自动驾驶仿真OpenScenario 2.0结构化场景文件，支持自动驾驶仿真测试软件的直接调用，减少了人工编码、调试和验证成本。其中，本文设计的结构化语言通常通过层级化、模块化来组织信息，相比自然语言更高效。减少冗余信息，利用关键字，使模型可以更紧凑地编码场景描述，从而降低输入的Token长度。本文设计的配套场景结构化语言的解析器通过内置规则，能实现生成场景结构化语言的逻辑修正，显著降低了模型幻觉，有效解决了单个生成文本过长和模型幻觉的问题，实现了通用模型能力的专用化。本研究开发的Saimo Scenario场景生成大模型能实现快速生成多样化的测试场景，支持按需定制生成长尾场景，同时，模型具备强大的泛化能力和自学习能力，成为生成式AI赋能自动驾驶测试验证领域的重要应用。

1　自动驾驶场景生成模型构建方法

1.1　总体思路

场景生成大模型的构建方法主要分为2种。第1种是从头训练大模型，需要搭建高性能算力平台，建立完整的数据集，通过长期训练，构建场景生成大模型[14]。该方法在模型精度方面表现优异，但可能面临模型基础能力不足的风险，例如模型初始架构设计不完善或训练数据覆盖的场景多样性不足。第2种方法是基于大语言模型，采用增强训练的方式构建场景生成大模型，与第1种方法相比，其成本较低、训练时间较短，但如果选取的训练数据集质量或数量不足，可能会影响预训练大模型的原有能力。

本文构建的Saimo Scenario场景生成大模型，是一款专门针对自动驾驶行业的高级人工智能模型，能高效且准确地生成OpenScenario 2.0标准的场景文件，以提升自动驾驶模拟测试的效率和结果精度。基于应用需求与资源条件，本文采用了第2种方法，通过增强训练实现高性价比的模型构建。场景生成大模型总体思路如图1所示。

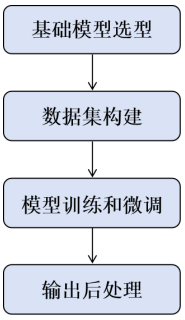


图1　场景生成大模型总体思路

模型输入、输出及后处理的具体流程如图2所示。输入来源于工程师的自然语言描述，即人工对场景要素进行描述，如“主车以60 km/h速度在左车道行驶，5 s后遭遇右侧货车切入”。该方法降低了自动驾驶仿真场景文件的生成门槛，使非专业人员也能通过自然语言快速定义测试场景。

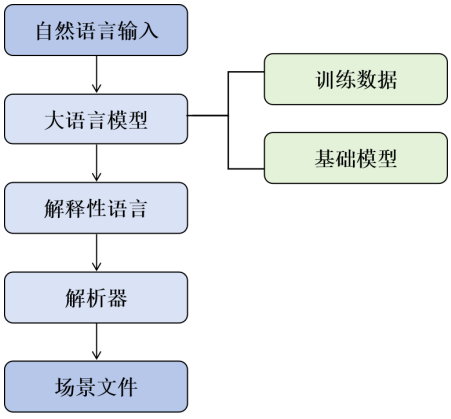


图2　输出场景文件全流程

然而，由于自然语言的表达自由度高，原始模型的输出结果难以通过解析器直接转换为符合OpenScenario 2.0规范的结构化文件，因此，需要通过微调预训练大模型，生成场景结构化解释性语言。自然语言输入经过微调训练后的大语言模型，生成结构化的自研解释性语言，有效压缩了大语言模型输出内容的长度，解决了单个生成文本过长的问题。

在与自研解释性语言适配的解析器中集成专家系统，专门对不符合逻辑和规则的场景进行自动修正，有效避免了生成场景不符合逻辑的问题，克服了大语言模型的幻觉现象。

1.2 基础模型选型

在场景生成大模型的基础模型选型过程中，主要考虑了开源和闭源两类模型。其中，开源模型如DeepSeek、BERT、LLaMA[15]，具有灵活性强、成本低的优势，但后续维护更新和迭代可能需要重新训练或微调，增加了工作复杂度。闭源模型如GPT-3和GPT-4，性能稳定，维护有保障，且具有较强的技术支持，能为基础模型的升级迭代提供保障。然而，闭源模型因其源代码不公开，相较于开源的通用大模型，存在较高的成本和数据隐私保护风险。因此，最终模型选型需要综合考虑模型的基础性能、训练效率和业务需求。

1.3 数据集构建

Saimo Scenario场景生成大模型数据集的结构设计采用了基于问答逻辑的Alpaca结构，注重问答对之间的逻辑关联性和上下文连贯性，并且具备快速适应新任务和数据集的能力，使模型能生成更加符合人类思维和表达习惯的自动驾驶仿真场景的描述语言。

数据集的主要来源分为两类。一类是在自动驾驶系统的仿真测试中积累的场景数据库。自动驾驶系统仿真测试场景数据库包括实车采集数据、基于真实交通事故数据转化的事故仿真场景、基于正向研发设计和预期功能安全风险评估的逻辑场景等。这些数据包含丰富的驾驶场景信息，如车辆行驶轨迹、道路环境、交通参与者行为等[16]。另一类是在仿真测试数据外，通过人工校注和数据增强补充的数据集。通过人工校注确保数据的准确性和可靠性，并利用大语言模型进行数据增强，如同义词替换、回译、上下文生成等技术，基于已有的数据生成更多样化的训练样本，以提高模型训练的全面性。基于这两种数据来源，本文构建了一个高质量的自动驾驶场景数据集，为场景生成大模型提供了丰富的训练素材。

1.4 模型训练和微调

为提升模型性能，大语言模型选择了适用业务需求的微调技术。常用的微调方法包括SFT（Supervised Fine-Tuning）监督微调、LoRA（Low-Rank Adaptation）微调，以及Freeze监督微调方法[17-18]。

由于计算资源有限，同时考虑效率和时间成本，在选择Saimo Scenario场景生成大模型的微调技术时，选择了LoRA微调方法。LoRA微调方法是引入低秩矩阵近似的方法，对高维参数矩阵进行分解，显著减少了需要训练的参数数量，并且能在不改变原模型结构、预训练参数的基础上，使新的LoRA参数与原参数配合使用，在不增加推理时间的前提下，实现模型对特定任务的快速适应。这在资源有限的环境，如消费级GPU上，使大模型的微调成为可能，大大提高了模型训练的效率和可行性。

在具体模型训练和微调过程[19]中，需要根据模型训练情况配置训练参数，包括学习率、批次大小、优化器等。训练初期设置较低的学习率，随后逐渐增加。根据硬件性能调整批次，通常较大的批次会更快完成训练，但也可能导致内存溢出。

通过多个迭代训练逐步优化模型。在每个训练周期开始时，随机打乱数据集，避免模型对数据顺序产生依赖，从而提高其泛化能力。将数据分批读取，每批数据通过模型进行前向传播，生成预测结果。使用预定义的损失函数计算预测结果与真实标签之间的差异。随后，通过反向传播更新模型的权重，以最小化损失值。以DeepSeek-7B作为预训练大模型为例，场景生成大模型训练过程曲线如图3所示。

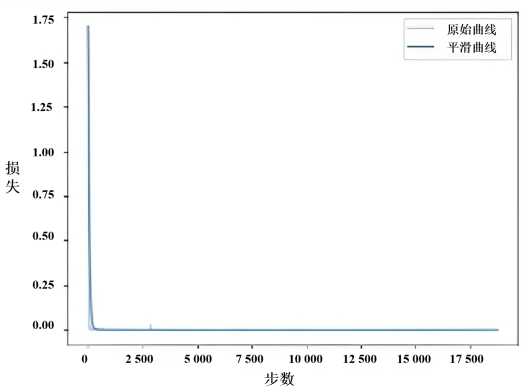


图3 场景生成大模型训练过程曲线

初始阶段损失值显著下降，符合模型快速收敛特性；后期曲线波动加剧，可能反映学习率未自适应调整或训练数据存在噪声干扰。通过滑动平均或指数加权滤波后，噪声抑制效果显著，整体呈单调递减趋势；末段平滑损失趋近于0，提示模型接近收敛状态。

1.5　模型输出及后处理

1.5.1　模型输出

传统自然语言描述方式存在冗余性高、结构松散的特点，当直接将自然语言作为大模型的输出时，会导致Token膨胀，并且由于自动驾驶场景的复杂性[20]可能超出预训练大模型的处理Token长度限制，其仿真测试场景的关键参数（如车速、触发条件、执行动作）注意力权重被自然语言的冗余连接词稀释，影响模型对场景核心要素的准确表达。同时，由于自然语言的表达自由度高，解析器难以直接将自然语言描述转换为符合OpenScenario 2.0规范的结构化文件，需要额外的规则匹配、错误修正，增加了解析成本，并降低了生成场景文件的稳定性和可用性。对此，本文设计了一种专用于生成场景文件的解释性语言，并通过动态语法适配机制优化大语言模型的输出结构。解释性语言针对场景结构分为3层，分别是道路、实体和环境。

其中，场景结构中的道路可以根据训练数据进行输出，如城市道路、高速道路、乡村道路、单行线等。在标志处也可以增加各种不同的限制标注，来进一步缩小道路的范围，如限速、十字路口、禁止掉头、公交车站等。

实体里的对象可以以列表的形式出现，每个实体包括了名称，如所在初始车道、所在相对位置、车辆动作等。其中，动作包括左变道、右变道、超车、加速、减速、横穿马路等。环境则对应描述了场景的天气（如晴、雨、雪、雾）、时间（如白天、黑夜）等。

当用户输入“我在城市道路上开车，路过十字路口时，左侧一辆大货车突然向右变道”，模型会解析并输出如图4所示的结构化场景内容。

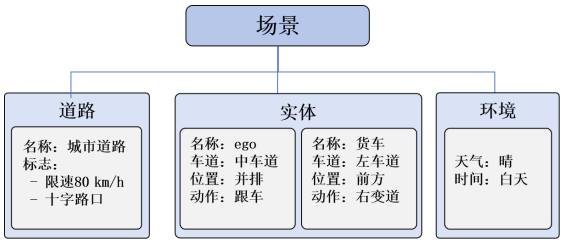


图4　结构化场景内容

1.5.2　后处理

自然语言大模型可能出现幻觉现象，生成不符合实际逻辑的内容，如在限速80 km/h的道路上，车辆运行速度为120 km/h。为了解决大语言模型的模型幻觉问题，实现从场景结构化语言到自动驾驶仿真场景文件的转化，设计了用于集成场景语言的专用解析器。解析器核心逻辑如图5所示。

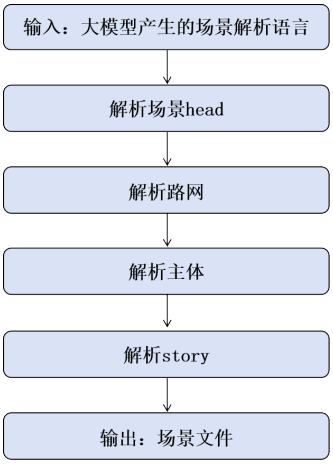


图5　解析器核心逻辑

解析器通过接收模型训练后生成的结构化文本，转换为OpenScenario 2.0格式（.osc格式）。还通过解析器内置规则，采用规则检查模块，自动纠正不合理场景设定，实现生成场景结构化语言的逻辑修正，如速度超限自动修正、车辆方向修正、无效车道切换等规则，显著降低了模型幻觉。典型规则逻辑有以下2点。

（1）速度超限自动修正。若用户在自然语言描述中设置的速度超出道路限速（如在最高限速120 km/h的道路上设置200 km/h），解析器会根据xodr地图中的限速信息，自动将速度修正为合理值120 km/h。

（2）车辆运行方向自动修正。若用户定义的车

辆方向与道路方向冲突（如在单向车道上设置逆行车辆），解析器会自动调整车辆行驶方向，用于后续逻辑处理。

2 试验结果与分析

2.1 试验设计

2.1.1 试验环境

为了验证 Saimo Scenario 场景生成大模型的性能，采 PyTorch 作为深度学习的框架，在 Ubuntu20.04 64 位操作系统上，GPU 显存 24 GB，CUDA 12.3，Python3.8 环境，Intel（R）Xeon（R）Gold 5218 CPU @ 2.30 GHz，内存 64 GB 的环境下进行对比试验分析。

试验中设置 Dropout 率为 0.2，防止模型过拟合，选择较小的 Batch_Size，设置为 2，以提升模型泛化能力。梯度累积的步数设置为 8，以模拟较大的 Batch_Size 的效果。学习率调节器采用余弦调节的方法，并使用 BAdam 优化器进行优化。采用依赖层数的递增切换方式，每 50 层进行一次参数切换，切换更新比例为 0.05，逐步调整模型参数，避免过大的参数变化。设置初始学习率为 0.000 001，以确保模型在训练初期能稳定收敛。

2.1.2 验证集

对评测维度准备具体领域的数据集和对比模型。本文选用的均为自研数据集，训练集数据 10 348 条，验证数据 3 274 条，包括了场景脚本语言所涵盖的所有内容。

2.2 基础模型选型

2.2.1 对比模型

在构建 Saimo Scenario 场景生成大模型过程中，需要综合考虑模型基础性能、训练效率及业务需求，选择适合的基础模型。

本文选用了 3 种语言模型进行对比试验，分别为 LLaMa2-7B、Qwen1.5-7B、DeepSeek-7B 模型。3 种基础模型均运用 LoRA 微调方法和相同的数据集进行训练，通过分析模型的 BLEU、ROUGE 指标以及推理耗时，选取用于构建 Saimo Scenario 场景生成大模型的预训练基础模型。3 种基础语言模型如下：

（1）LLaMa2-7B 是 Meta 公司推出的开源预训练大语言模型，是一个预先训练和微调的生成文本模型的集合；

（2）Qwen1.5-7B 是阿里 Qwen 系列中的一种规模的预训练和指令微调模型；

（3）DeepSeek-7B 是基于 DeepSeek-Coder-v1.5 7B 基础架构，并经过特殊的指令调教和强化学习（RL）训练的大语言模型。

2.2.2 对比指标

由于场景生成大模型经过后处理生成的 Open Scenario 2.0 场景文件本质上是以领域特定语言（DSL）编写的结构化文本，其语法规则在形式化表达上与自然语言存在一定共性。因此，对场景生成大模型开展性能评估分析时，能采用双语评估替换（Bilingual Evaluation，BLEU）指标和自动文摘评价（Recall-Oriented Understudy for Gisting Evaluation，ROUGE）指标评估生成语句，衡量生成文本与参考文本的相似度[21-22]。BLEU 分数是用于评估模型机器翻译质量的 1 项评价指标，它会根据模型生成的结果与验证集中答案的匹配程度给出分数，这个分数在 0～1 之间，BLEU 值越接近 1 则翻译质量越高。ROUGE 是一种用于衡量自动文摘生成质量的指标，它根据生成的文摘与参考摘要之间的匹配程度给出分数，同样在 0～1 之间，1 表示最匹配，0 表示最不相关。

2.2.3 选型结果分析

本文对比了 3 种基于 LoRA 微调的，使用同一硬件环境和数据集训练的场景生成大语言模型，对比评估分析 BLEU、ROUGE 指标以及推理耗时，试验结果见表 1。

试验结果表明，3 种基础模型训练生成的场景语言大模型准确度基本一致，BLEU、ROUGE 指标值相近，且均超过 99，推理耗时均小于 2 s，表明 3 种该场景生成大模型具备较高的生成准确度和效率。其中，推理速度相对较快的原因是，在构建场景生成大模型过程中，通过结构化语言的设计，降

表 1　基于 LoRA 微调的场景生成大语言模型 BLEU 和 ROUGE 分数对比分析

基础模型	BLEU	ROUGE-1	ROUGE-2	ROUGE-I	推理耗时/s
LLaMa2-7B	99.89	99.84	99.80	99.87	1.26
Qwen1.5-7B	99.89	99.83	99.79	99.86	0.99
DeepSeek-7B	99.89	99.82	99.78	99.86	0.53

低了嵌套结构的复杂性，增强了结构拓展性，极大提升了推理速度。

在准确度均达标的前提下，由于 DeepSeek-7B 基础模型的推理耗时更短。考虑应用领域特点、数据质量和任务需求，在 Saimo Scenario 场景生成大模型构建过程中，选择了综合效果最好的 DeepSeek-7B 大语言模型作为预训练大模型构建场景生成大模型，基于 LoRA 微调训练方法，训练输出 1 种用于生成场景文件的结构化的解释性语言。

DeepSeek-7B 大语言模型基于 MLA 架构和 DeepSeekMoESparse 结构，大幅降低显存占用和计算量。不仅显著提升了模型的性能和效率，降低了计算成本，同时还具备训练的稳定性，可以避免梯度爆炸、梯度消失等问题，同时在中文任务中具备显著优势。

为了进一步验证 LoRA 微调对场景结构化解析能力的提升作用，设计对比试验，在输入相同自然语言的条件下，分别测试两类模型的输出质量。对照组，未经 LoRA 微调的 DeepSeek-7B 大语言模型的输出结果；试验组，经过加载秩 $r = 8$ 的 LoRA 适配器微调（基于自研场景数据集微调）的模型的输出结果。具体输入输出示例对比结果如下。

用户输入：“我在下雪的城市道路上开车，前方货车突然向右做强行变道。”

（1）原始预训练模型（DeepSeek-7B）输出为：

“在下雪的城市道路上遇到前方货车强行变道时，安全操作需兼顾路面特性和应急反应。以下是专业应对流程，1. 动态风险评估（0.5 s 内）。立即扫视后视镜及两侧环境……”

（2）微调后模型输出为：

场景：

道路：

名称：城市道路

标志：

-限速 80 km/h 的标志

-十字路口

-禁止掉头的标志

实体：

-名称：ego

车道：中车道

位置：并排

动作：跟车

-名称：货车

车道：有车道

位置：前方

动作：右变道

环境：

天气：雪

时间：白天

从输出对比可以看出，微调前的原始模型输出难以用于直接生成结构化场景文件。微调后的原始模型输出的解释性语言结构更适合用于构建自动驾驶仿真场景文件。

2.3　第三方评价

2.3.1　场景逻辑合理性评估

由于自动驾驶仿真场景的逻辑合理性难以仅通过文本相似度度量（如 BLEU、ROUGE）来衡量。针对场景逻辑合理性，基于选择的 DeepSeek-7B 大语言模型作为预训练大模型构建场景生成大模型，采用专家评审评估方法，邀请 9 位来自北京赛目科技股份有限公司的仿真工程师独立评估，人工评分基于 10 分制，评估角度包括场景因果关系合理性、驾驶行为合理性、环境约束符合性。第三方评价在评测者有经验的情况下，试验方法更简单且能得到更准确的评测结果。基于 10 分制，由工程师进行打分，按 1～10 分来表示场景逻辑合理性从劣到优，见表 2。测试结果见表 3。

表 2　第三方评价等级

分值	主观评价	分值	主观评价
1	无法接受	6	一般
2	极不满意	7	较满意
3	非常不满意	8	满意
4	不满意	9	非常满意
5	较不满意	10	极满意

表 3　场景逻辑合理性分数对比

工程师序号	场景逻辑合理性分数	工程师序号	场景逻辑合理性分数
1	9	6	9
2	8	7	9
3	9	8	9
4	9	9	10
5	9	10	9

根据测试结果，由人工评分的场景逻辑合理性得分均达到 8 分以上，可以较好地满足工程师对生成仿真场景文件的要求。

2.3.2　推理性能评估

本文研究的 Saimo Scenario 场景生成大模型基于 LangChain 和 VLLM 进行部署。在模型部署后，需要重点关注推理性能，特别是高并发场景下的准确性和推理速度。因此，还需要对生成场景文件效率、生成场景的适用度 2 个维度进行评估。本研究邀请了 18 位来自北京赛目科技股份有限公司的工程师，将他们分为 3 组，与 Saimo Scenario 场景生成大模型进行 10 轮对话，最终对生成的答案进行评估，对比分析生成场景文件效率、生成场景的适用度，基于 10 分制，由工程师进行打分，按 1～10 分来表示效率、适用度从劣到优，具体见表 2。

由 3 组第三方具有专业经验的自动驾驶仿真工程师，分别对场景生成大模型进行输入，基于生成的场景文件进行仿真测试，并对生成的场景文件进行主观评分。评分标准主要基于场景文件在建模仿真中的可用度，包括语法合规性、逻辑完整性、可执行性以及对测试需求的匹配度，最终分别对生成场景文件效率、生成场景的仿真适用度进行打分。生成场景文件用于自动驾驶仿真测试的操作页面，如图 6 所示。

图 6　自动驾驶仿真面板

3 组仿真工程师根据 Saimo Scenario 场景生成大模型生成的答案，搭建的仿真场景如图 7 所示。评价结果如图 8 所示。

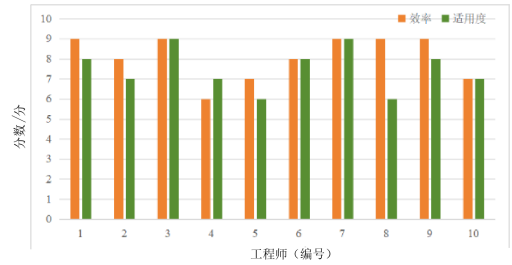
（a）　前方行人横穿场景

（b）　前方自行车缓行场景

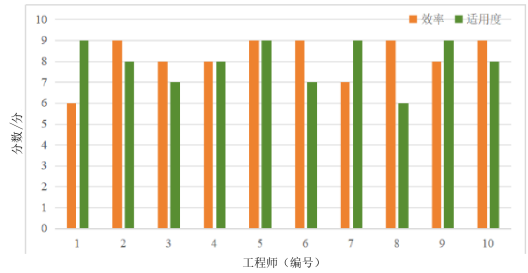
（c）　公交车切入场景

图 7　根据 Saimo Scenario 场景生成大语言模型生成答案搭建的仿真场景

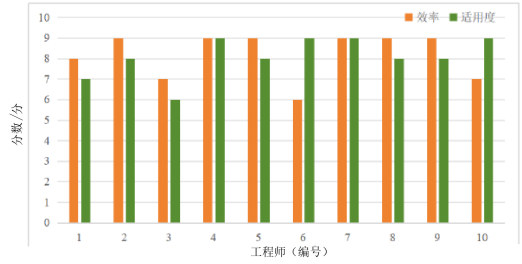
根据 3 组仿真工程师使用中的反馈情况，Saimo Scenario 场景生成大模型生成场景效率和适



（a） 第 1 组评分结果



（b） 第 2 组评分结果



（c） 第 3 组评分结果

图 8　三组仿真工程师评价结果

用度都获得较好的第三方评价。这是因为传统场景的构建主要是基于编辑器人工搭建的，通过道路编辑器构建路网信息，通过场景编辑器构建动态场景，费时费力，人工成本较高，而且对于场景要素组合的丰富化也受到人工的限制。而通过 Saimo Scenario 场景生成大模型生成场景，仅需要输入场景文字需求，将以前仿真工程师需要 2 h 才能完成的场景构建工作压缩到 10 多秒，极大提升了仿真测试场景的生成效率。但是根据评分结果，也可以看出不同仿真工程师的评价存在一定差异，主要原因如下。

（1）工程师主观评估因素。受工程师对场景生成大语言模型的使用经验、描述的自动驾驶仿真场景复杂度等因素影响，存在主观评判偏差。

（2）自然语言描述到场景文件的映射误差。自然语言描述（输入）到场景文件（输出）的映射存在多义性歧义，因此，存在人工描述准确性的传递损失。

3　结论

随着人工智能技术的快速发展，大模型在自动驾驶汽车领域的应用正逐渐成为重点发展方向之一，自动驾驶有望成为具身智能的商业化落地应用实践。

本文利用真实交通场景数据，构建专属自动驾驶领域的数据集，并对大语言模型进行微调，显著提升了模型训练的针对性和质量，实现了从通用大模型到领域大模型的成功转化。试验表明，本文提出的场景生成大语言模型比其他基础模型的训练结果生成自动驾驶场景更快，同时，通过自主研发的场景解析器，不仅有效解决了大语言模型在自动驾驶场景生成中的模型幻觉问题，还有效解决了场景文本过长的限制，为自动驾驶测试验证提供了强有力的技术支撑。

目前，本文的主要工作是基于对话系统中单轮的对话数据进行评测，提出的场景生成大语言模型主要聚焦于提升自动驾驶仿真场景生成效率。由于当前本文自动驾驶仿真测试训练数据集存在规模和分布上的局限性，所以仍未解决场景覆盖度不足的问题，当前暂时不能生成极端场景。未来的研究工作将进一步通过增量学习机制，不断注入新的场景数据，模型可以持续优化，从而逐步提升场景覆盖范围，最终增强自动驾驶仿真测试场景覆盖的全面性。今后的研究工作还将集中于推动生成式 AI 与自动驾驶系统的深度整合，构建自动化、智能化的测试验证流程，同时将本文的场景生成大模型应用到构建覆盖中国特色交通流的自动驾驶仿真测试场景库中，提升危险场景和边缘场景的生成与泛化能力，为自动驾驶算法研发提供高效、高覆盖度的安全验证资源，加速其商业化应用进程。

参考文献（References）

[1] European Commission. Commission Implementing Regulation (EU) 2022/1426 of 5 August 2022 Laying Down Rules for the Application of Regulation (EU) 2019/2144 of the European Parliament and of the Council as Regards Uniform Procedures and Technical Specifications for the Type-Approval of the Automated Driving System (ADS) of Fully Automated Vehicles (Text with EEA Relevance)[Z].Europe:European Commission,2022.

[2] 刘法旺,徐晓庆,陈贞,等.搭载自动驾驶功能的智能网联汽车安全测试与评估方法研究[J].汽车工程学报,2022,12(3):221-227.
LIU Fawang, XU Xiaoqing, CHEN Zhen, et al.Research on Safety Testing and Assessment Methods for Intelligent and Connected Vehicles with Autonomous Driving Functions [J].Chinese Journal of Automotive Engineering, 2022,12(3):221-227.(in Chinese)

[3] VISHNUKUMAR H J, BUTTING B, MULLER C, et al. Machine Learning and Deep Neural Network — Artificial Intelligence Core for Lab and Real-World Test and Validation for ADAS and Autonomous Vehicles: AI for Efficient and Quality Test and Validation [C]//2017 Intelligent Systems Conference (IntelliSys), Sept. 7-8, 2017,London,UK.Piscataway NJ:IEEE,c2017:714-721.

[4] BRUTO DA COSTA A A, IRVINE P, ZHANG Xizhe, et al. Ontology-Based Scenario Generation for Automated Driving Systems Verification and Validation Using Rules of the Road[J].IEEE Transactions on Intelligent Vehicles, 2024:3377534.1-3377534.11.

[5] 刘法旺,何丰,周时莹,等.基于场景的智能网联汽车模拟仿真测试评估方法与实践[J].汽车工程学报,2023,13(2):135-145.
LIU Fawang, HE Feng, ZHOU Shiying, et al. Scenario-Based Virtual Testing and Assessment Method and Practice for Intelligent and Connected Vehicles [J]. Chinese Journal of Automotive Engineering, 2023, 13 (2):135-145.(in Chinese)

[6] DING Wenhao,XU Chejian,ARIEF M,et al.A Survey on Safety-Critical Driving Scenario Generation—A Methodological Perspective[J]. IEEE Transactions on Intelligent Transportation Systems,2023,24(7):6971-6988.

[7] SCHULDT F,RESCHKA A,MAURER M.A Method for an Efficient,Systematic Test Case Generation for Advanced Driver Assistance Systems in Virtual Environments[J]. Automotive Systems Engineering II,2017:147-175.

[8] KOREN M,ALSAIF S,LEE R,et al.Adaptive Stress Testing for Autonomous Vehicles[C]//2018 IEEE Intelligent Vehicles Symposium(IV),June 26-30,2018,Changshu, China.Piscataway NJ:IEEE,c2018:1-7.

[9] GAO Feng, DUAN Jianli, HE Yingdong, et al. A Test Scenario Automatic Generation Strategy for Intelligent Driving Systems[J].Mathematical Problems in Engineering: Theory, Methods and Applications, 2019, 2019: 3737486.1-3737486.10.

[10] 龚磊,孙新雨,张昱,等.嵌入路网图模型的自动驾驶场景描述语言[J].软件学报,2023,34(9):3981-4002.
GONG Lei, SUN Xinyu, ZHANG Yu, et al. Scenario Description Language of Autonomous Driving Embedded with Road Network Graph Model [J]. Journal of Software,2023,34(9):3981-4002.(in Chinese)

[11] WAYVE.Introducing GAIA-1: A Cutting-Edge Generative AI Model for Autonomy[EB/OL].(2023-06-17)[2024-08-12].https://wayve.ai/thinking/introducing-gaia1.

[12] CUI Can, MA Yunsheng, YANG Zichong, et al. Large Language Models for Autonomous Driving (LLM4AD): Concept, Benchmark, Simulation, and Real-Vehicle Experiment[J]. Journal of Latex Class Files, 2015, 14 (8):1-19.

[13] TIAN Haoxiang,HAN Xingshuo,WU Guoquan,et al.An LLM-Enhanced Multi-Objective Evolutionary Search for Autonomous Driving Test Scenario Generation [Z]. arXiv:2406.10857.2024.

[14] FAN Haolin, FUH J, LU Wenfeng, et al. Unleashing the Potential of Large Language Models for Knowledge Augmentation: A Practical Experiment on Incremental Sheet Forming [J]. Procedia Computer Science, 2024, 232:1269-1278.

[15] TOUVRON H, MARTIN L, STONE K, et al. LlaMa 2: Open Foundation and Fine-Tuned Chat Models [Z]. arXiv:2307.09288,2023.

[16] GIUNCHIGLIA E,STOIAN M C,KHAN S,et al.ROAD-R: The Autonomous Driving Dataset with Logical Requirements [J]. Machine Learning, 2023, 112 (9): 3261-3291.

[17] LIU Xiao, JI Kaixuan, FU Yicheng, et al. P-Tuning v2: Prompt Tuning Can Be Comparable to Fine-Tuning Universally Across Scales and Tasks[Z].arXiv:2110.07602, 2021.

[18] DETTMERS T, PAGNONI A, HOLTZMAN A, et al.

QLoRA：Efficient Finetuning of Quantized LLMs（2023）[Z].arXiv：2305.14314，2023.

[19] AMIR K，ABDOLLAH A，MASOUD M T，et al. Multi-Domain Autonomous Driving Dataset：Towards Enhancing the Generalization of the Convolutional Neural Networks in New Environments[J].IET Image Processing，2023，17（4）：1253-1266.

[20] POSEDARU B S，PANTELIMON F，DULGHERU M N ，et al. Artificial Intelligence Text Processing Using Retrieval-Augmented Generation：Applications in Business and Education Fields[J]. Proceedings of the 18th International Conference on Business Excellence，2024，18（1）：209-222.

[21] PAPINENI K，ROUKOS S，WARD T，et al. BLEU：A Method for Automatic Evaluation of Machine Translation[C]//Proceedings of the 40th Annual Meeting of Association for Computational Linguistics，July 6，2002，Philadelphia，Pennsylvania，USA.2002：311-318.

[22] LIN Chinyew. ROUGE：A Package for Automatic Evaluation of Summaries[C]//ACL Proceeding of Workshop on Text Summarization Branches Out，July 25-26，2004，Barcelona，Spain.2004：74-81.

作者简介

陈贞（1989-），女，湖北荆门人，博士，主要研究方向为智能网联汽车安全测试与评估方法。

E-mail：chenzhen@dystech.cn

通信作者

李京泰（1994-），男，四川营山人，硕士，主要研究方向为智能网联汽车政策法规及测试评价技术。

E-mail：lijingtai@eidc.org.cn