·中国一汽2022年度优秀科技论文专题·

# 车载智能语音助手综合评估模型建立及应用

道发发　丁敏　袁粲璨　陈晓军　黎小平　赵嵩

（一汽-大众汽车有限公司，长春 130011）

【欢迎引用】道发发，丁敏，袁粲璨，等. 车载智能语音助手综合评估模型建立及应用[J]. 汽车文摘，2023(4): 12-17.

【Cite this paper】DAO F F, DING M, YUAN C C, et al.Comprehensive Evaluation Model and Application of Vehicular Intelligent Voice Assistant[J].Automotive Digest (Chinese),2023(4): 12-17.

【摘要】车载语音助手是车内人机交互的一级入口，语音系统在综合场景下的业务处理能力和交互能力极大影响用户体验。阐述一套综合评价模型，对车载语音系统功能进行完全解析、导出竞品实现逻辑和生产中的阶段性评估结果。模型包含一套评估用例数据库，支持103条量化指标计算、15条非量化指标、3个综合指标生成。此外，模型包含全流程的指标分析工具及其它自动化工具。在实际生产应用过程中，该模型与人工评估的误差率约为10%，自动化效率提升约70%。

**关键词：车载语音；评估模型；自动化分析；用户体验**

**中图分类号：**U471.22　　**文献标识码：**A　　**DOI:** 10.19822/j.cnki.1671-6329.20220252

# Comprehensive Evaluation Model and Application of Vehicular Intelligent Voice Assistant

Dao Fafa, Ding Min, Yuan Cancan, Chen Xiaojun, Li Xiaoping, Zhao Song

（FAW-Volkswagen Automobile Co., Ltd., Changchun 130011）

【Abstract】The in-vehicle voice assistant is the first-level entrance for human-computer interaction in the vehicle. The business processing and interaction capabilities of the voice system in comprehensive scenarios greatly affect the user experience. This article elaborates a set of comprehensive evaluation models, which can fully analyze the function points of the vehicle voice system, derive the implementation logic of competing products, and stage evaluations in production. This model includes a set of evaluation use case databases that support the calculation of 103 quantitative indicators, 15 generate non-quantitative indicators and 3 comprehensive indicators. In addition, this model includes full-process indicator analysis tools and other automation tools. In the actual production and application process, the error rate of this model compared with manual evaluation is about 10%, and the automation efficiency is increased by about 70%.

**Key words: In-vehicle voice, Evaluation model, Automation, User experience**

## 缩略语

| | |
|---|---|
| NLP | Nature Language Processing |
| BLEU | Bilingual Evaluation Understudy |
| ROUGE | Recall-Oriented Understudy for Gisting Evaluation |
| TTS | Text To Speech |
| VPA | Virtual Personal Assistant |
| AI | Artificial Intelligence |

## 0　引言

随着物联网、车联网、自动驾驶技术的发展，汽车行业的竞争力正在从传统的以性能为核心转变为以数字化和智能化为核心，包括众多的智能座舱服务和辅助驾驶服务。车载语音助手是汽车数字化的一部分[1]，也是人机交互的一级入口，常见的车载语音助手包含任务型对话功能和闲聊功能，其中任务型对话应用于车内支持的功能操作，如车控、导航、天气等，闲聊则是通用的聊天型对话，不完成具体的任务。

车载语音助手的出现解放了驾驶员的双手和双眼，用户无需注视屏幕或操作按钮即可完成对应的需求。但同时，为了“可见即可说”，车载语音助手需要支持数百个常用的指令及无穷的说法变换，其性能的优劣直接影响用户的体验[2]。故针对车载语音助手的

综合性评价非常重要。

语音助手的实现逻辑是基于人工智能的自然语言处理(Nature Language Processing, NLP)模型。常用的模型评估方法通常是针对单个模型的点对点评估,如对于实体识别[3]、序列标注[4]的模型,采用精确率、召回率、精确率和召回率的加权平均(F1 Score)数值等评估指标;对于文本生成[5]任务使用双语评估研究(Bilingual Evaluation Understudy, BLEU)、自动摘要评价(Recall- Oriented Understudy for Gisting Evaluation, ROUGE)方法进行评估。车载语音助手通常由至少十几个不同的模型组成,每个模型在开发过程中的训练数据不一定相同,评估指标也不相同,故不能用各部分的单独技术性指标来描述整个系统的性能。在开发过程中,开发人员完成所有组件的开发并整合成语音助手之后,测试人员会根据其设计所支持的功能进行通过性测试。这种测试方式会忽略实际应用场景的复杂性、用户表达的多样性,无法深度探查语音助手的能力及其背后算法的有效性[6]。

综上所述,本文提出了一套综合性的语音助手评估模型,旨在以贴近用户的方式量化描述语音助手的综合表现,并可以反推出各个子系统的性能,用于问题定位和优化。

## 1 评估模型建立

本模型包含评价数据库、指标生成模型、可视化组件、自动化组件4个主要部分。其中评价数据库包含10 026条由人工构造的高阶用例,以矩阵形式组织,横向按语义点区分,如意图联想、语义容错、多意图识别共31个评价项,纵向按车载常用功能分为64个功能,主要涉及车控、导航、天气、多媒体、电话主要车载技能以及维保、蓝牙、计算器、油价等长尾技能[7]。指标生成模型包含完成率、意图识别率生成模型以及意图联想、语义容错、多意图识别子项指标生成模型。可视化组件包含数据载入、低代码分析、柱状图、条形图、饼图、时间序列分析功能。自动化组件主要包括自动化分析、多轮对话模拟、报告生成辅助性组件。

标准的指标生成模型可生成主要的技术性描述指标,大量的评估用例保证评估结果无偏差,本模型可应用于需求调研阶段的竞品分析,也可应用于生产阶段的需求对接,保证产品交付质量。

### 1.1 评估用例

本模型的核心设计目标是一款普适的车载语音助手评估模型,可应用于市场上常见的搭载智能语音助手的综合评估。一般的评估过程需要遵循可量化、层次性、普遍性、客观性原则。针对以上原则和实际需求,构建了一批用于评估的用例库,所有用例均由经验丰富的测试人员和产品人员编写,并通过一个审核小组逐条审核,最终形成了一个万余条的评估用例库。

评估用例库中,用例的组织方式遵从分层原则,分别从智能等级、功能点、困难程度、语义维度4个方面进行分级。为了能够得到精确的量化指标,在构造用例时确保每条用例只对应于一种语义指标。

表1示意每个维度上的详细层次结构,其中技能共有27个一级项和64个二级项,语义维度分3个一级项和31个二级项。表1中只列出一级项,省略了二级项。

表1 用例维度分级表

| | |
|---|---|
| 智能等级 | L1、L2、L3、L4 |
| 技能 | 天气、导航、音乐、电台、新闻、问答系统、车控、电话、维保、日历、日程、智能家居、航班、股票、成语、单位换算、翻译、古诗、故事、汇率、火车票、计算器、食物营养、星座运势、影讯、今日油价、万年历 |
| 困难程度 | 标准、困难 |
| 语义维度 | 理解能力、交互能力、决策能力 |

根据评估模型的特点,设计了一套用例构建标准,从语义维度对用例的构建原则、评估目标以及其智能程度进行分级,表2描述了用例构建过程中语义维度各指标的定义以及对应的评价项,由于篇幅所限,本文仅列出部分内容,全量的评价项共有31项,基本覆盖所有语义类型。本文中的所有用例都根据此表进行构建。其中,L1、L2、L3、L4分别代表4个不同的智能程度。

### 1.2 主要指标评估模型

评估模型分为主要指标和次要指标,其中主要指标为任务完成率、意图识别率,用于评估语音助手在任务型对话上的端到端能力;次要指标是语音维度的31个细分维度,用于对车载语音智能程度、语义理解能力、语义理解模型效果的分析。

评估过程中,为了降低评估人员主观的误差,使3名评估人员同时进行打分,当所有评估人员都认为该用例通过时,则该用例通过。

对于任务完成率和意图识别率,指标的量化计算公式如式(1)。

$$p_a = \frac{\sum_{i=0}^{|X|} a_i}{|X|} \tag{1}$$

式中，$p_a$为评估用例结果的得分；$\alpha_i$为第$i$条用例的得分；$X$为用例集合。当所有评估人员的打分都为1时，$a_i=1$，否则$a_i=0$。

表2　用例构建标准描述

| 能力类型 | 能力组成 | 评价项 | 定义 |
|---|---|---|---|
| 认知智能 | 理解能力 | 特定领域意图识别 | 系统能够支持车载常见应用领域的车主意图 |
| | | 意图联想 | 针对没有显式表达的意图，系统能够联想到关联意图并反馈车主需要的信息 |
| | | 实体抽取 | 系统能够正确提取出某个意图的指令中包含的关键词、槽位等信息 |
| | | 方言识别 | 系统可以识别用户用方言发出的指令 |
| | | … | … |
| | 交互能力 | 全双工 | 系统在播报过程中能够随时被有效指令打断并转而开始新一轮对话 |
| | | 多轮对话 | 系统能够跟踪对话状态以达成多轮对话任务 |
| | | 回复多样性 | 针对相同条件下的同一个指令，系统能够避免一成不变的回复 |
| | | … | … |
| | 决策能力 | 指代消解 | 系统能够在不同领域间借助上下文推理出用户指代的对象 |
| | | 逻辑运算 | 系统能够支持指令中的逻辑门 |
| | | 关系推理 | 系统能够对指令中的实体关系进行推理 |
| | | … | … |
| 情感智能 | 情感理解 | 文本情感识别 | 系统可以识别用户指令文本中的情感倾向 |
| | | 声音情感识别 | 系统可以识别用户指令声音中的情感倾向 |
| | | 表情情感识别 | 系统可以识别用户表情中的情感倾向 |
| | 情感响应 | 情感化-文本转语音（TTS） | 系统能够通过TTS音色的自动变化来对识别到的用户情感做出响应 |
| | | 情感化-虚拟个人助理（VPA） | 系统能够通过VPA形象的自动变化来对识别到的用户情感做出响应 |

此外，为了能够捕捉评估一致的随机性，除了上述$p_a$指标外，引入指标$p_c$，对于多个评估人员$e_j$，用例集合$X$的评估分数是集合$S$，那么$p_c$的计算公式为：

$$p_c=\sum_{s\in S}\prod p(s|e_i) \tag{2}$$

式中，$p(s|e_i)$是每个评估人员给出分数$s$的频率估计；$s$是用例；最后能够得到和评估一致性相关的结果$\sigma$：

$$\sigma=\frac{p_a-p_c}{1-p_c} \tag{3}$$

式中，当$\sigma$越靠近1，则表示多名评价者评价的一致性越强，评估结果越可靠。在本模型中，当$\sigma$>0.8时，认为当轮评价有效，采用该轮评价结果。

同时，对于完成率和意图识别率，将整个语音助手视为一个统一的机器学习模型，采用查准率、查全率和$F_1$值描述语音助手的整体表现：

$$P_r=\frac{TP}{TP+FP} \tag{4}$$

$$Recall=\frac{TP}{TP+FN} \tag{5}$$

$$F_1=2\times\frac{P_r\times Recall}{P_r+Recall} \tag{6}$$

式中，$P_r$为查准率；$TP$为语音助手成功完成的任务数；$FP$为语音助手未识别的拒识用例数；$Recall$为查全率；$FN$为语音助手成功识别到的拒识用例数；$F_1$代表语音助手的实际表现，其数值越靠近1，表示语音助手的性能越佳。

综上，在评价任务完成和意图识别2个主要维度时，使用了2套指标，第1套综合指标使用一致性评估方式保证评估人员的一致性，第2套指标将语音助手看做一个整体的AI模型，使用查准率、查全率和$F_1$值来评估其整体表现。

在实际研发过程中，研发人员或项目管理人员不仅关注语音助手的整体指标，更需要注意各部分子功能的具体指标，以此保证子模块算法的性能。

### 1.3　语义指标评估模型

在本评估模型中，语义方面共分3个一级语义和31个二级语义，从算法角度进行分类，可以归结为文本分类任务、匹配任务、序列标注任务和文本生成任务。

对于文本分类任务和序列标注任务，由于评估样本有限，且样本分布不完全均衡，为避免忽略小样本数据，故使用MicroAveraged方法评估：

$$P_{micro}=\frac{\overline{TP}}{\overline{TP}+\overline{FP}}=\frac{\sum_{i=1}^{n}TP_i}{\sum_{i=1}^{n}TP_i+\sum_{i=1}^{n}FP_i} \tag{7}$$

$$R_{micro}=\frac{\overline{TP}}{\overline{TP}+\overline{FN}}=\frac{\sum_{i=1}^{n}TP_i}{\sum_{i=1}^{n}TP_i+\sum_{i=1}^{n}FN_i} \tag{8}$$

$$F_{micro}=2\times\frac{P_{micro}\times R_{micro}}{P_{micro}+R_{micro}} \tag{9}$$

其中，$P_{micro}$为微平均查准率，$R_{micro}$为微平均查全率，$TP_i$为第$i$类任务里识别正确的数量，$FP_i$为第$i$类任务里识别错误的数量，$FN_i$为第$i$类里把错误类别识别成正确类别的数量，$\overline{TP}$，$\overline{FP}$，分别为$TP_i$和$FP_i$的算数平均值。

对于文本匹配任务，使用 *Top@N* 覆盖率来描述其性能，计算方式为前 *N* 项候选指标中包含正确结果的准确率。

对于文本生成任务，使用 *BLEU* 作为其评估指标：

$$BLEU = BP \times \exp\left(\sum_{n=1}^{N} w_n \times \log P_n\right) \quad (10)$$

$$BP = f(x) = \begin{cases} 1, & lc > lr \\ \exp(1 - lr/lc), & lc \leqslant lr \end{cases} \quad (11)$$

式中，*BP* 为最佳匹配长度；$w_n$ 为赋予 $P_n$ 权重；$P_n$ 为多元精度得分；*lc* 为结果的长度；*lr* 为标准答案句子的长度。

综上，描述了本模型中2个主要指标（任务完成和意图识别）以及31个二级语义指标的评估方法，主要指标将语音助手视为一个单独对象，使用查准率、查全率、$F_1$ 值来描述其性能，并采用 $\sigma$ 约束来规避评估人员主观上导致的评分不一致问题[8]。二级语义指标将语音助手视为多个子模型的集合，针对每个二级语义项，都给出单独的评价指标，开发人员可以借助这些指标进行深度的问题定位，需求分析人员可以借助这些指标完成对目标产品的多维度分析。

在实际的开发过程中，由于项目采用敏捷的工作方式，项目版本迭代次数最高可以达到每天一次，导致开发人员对于问题定位的需求频率非常高。使用人工分析来定位问题会带来大量的人力需求，为了降低对于人工的消耗，使用一个简单的算法模型来进行快速的自动化问题定位。

## 2 分析方法

### 2.1 问题描述

评估模型的输出结果是2个主要指标和31个语义指标的评分，这些指标的集合代表了语音助手各部分及整体的表现。为了适应语音助手复杂的任务型对话逻辑，如前文所述，评估模型也遵从分层的构建逻辑，并从功能点和语义的维度进行了两级划分。整体的指标体系可以分为4层，上一层级的评估指标值为下一层级指标的算数平均值（图1）。

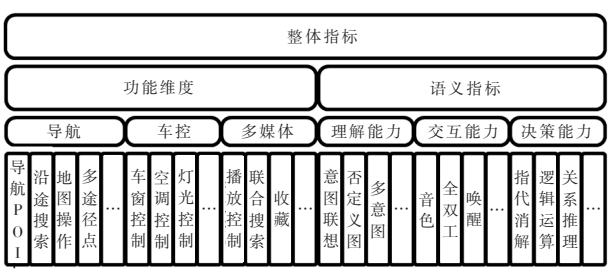


**图1 评估指标层级划分**

表3为使用评估模型对上述组织方式的用例集合进行打分后的结果，其中3项打分 $S_1$、$S_2$、$S_3$ 分别为语义识别、意图识别、任务完成情况。由于篇幅限制，此处省略了一些其它辅助字段的信息。

**表3 模型评估结果**

| 编号 | 技能 | 功能 | 语义 | $S_1$ | $S_2$ | $S_3$ |
|---|---|---|---|---|---|---|
| 1 | 导航 | 途经点 | 语义容错 | 1 | 0 | 1 |
| 2 | 导航 | 沿途搜索 | 意图联想 | 0 | 1 | 1 |
| 3 | 车控 | 灯光控制 | 多意图 | 1 | 1 | 1 |
| 4 | 多媒体 | 播放控制 | 逻辑运算 | 1 | 0 | 0 |
| … | … | … | … | … | … | … |

通常，一次评估后会得到30 000条以上的评估结果数据。在研发生产过程中，伴随着敏捷迭代，需要进行高频的模型评估和问题定位分析，使用人工的方式进行分析会带来极大的人力需求。为了提高问题分析和定位的能力，设计了一套自动化的分析算法，用于研发过程中快速分析。

### 2.2 分析算法

模型的评估结果是一个多维分层指标体系，先构建数据模型，如图2所示。*S* 为整体评估分数，由其下层指标合并而成（如整体的任务完成率由功能维度和语义维度得分合并而成）。因此，对于一个二级指标体系来说，分析算法的任务是从 *S* 得分的波动中找出造成这种波动的下级节点，且结果必须具有原子特性，即节点组合的最简约形式，如（A1B1、A1C1）的最简约形式为（A1B1C1）[9]。

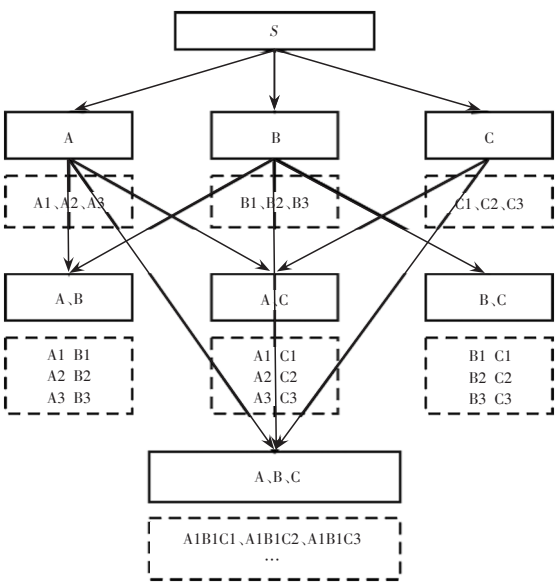


**图2 数据模型**

在问题定位过程中，需要结果能够准确反应波动出现的原因，即异常点[10]。异常点的查找需要满足3

个主要条件，也是算法设计过程中的难点[11]。

（1）对于每一个维度，结果需要尽可能解释主要指标波动原因；

（2）对于每一个维度，结果需要符合最简原则，即不可再分；

（3）在所有维度中，需要找出和预期结果相差最大的元素。

针对以上问题，参考 Adtributor 方法[12]，设计分析算法。$S$ 值为当前指标的惊喜度，代表该指标偏离预期的距离，距离越远，惊喜度越高[13]，算法如下。

问题定位算法（根因分析）

$Foreach m \in M$ // 计算当前评估数据的S值

$Foreach E_{ij}$ //

$p = F_{ij}(m)/F(m)$ //

$q = A_{ij}(m)/A(m)$

$S_{ij}(m) = D_{JS}(p,q)$

$ExplanatorySet = \{\}$

$Foreach i \in D$

$SortedE = E_{i}.SortDescend(S_{ij}(m))$

$Candidate = \{\}, Explains = 0, Surprise = 0$

$Foreach E_{ij} \in SortedE$ //筛选可疑节点加入候选集合

$EP = (A_{ij}(m) - F_{ij}(m))/(A(m) - F(m))$

$if(EP > T_{EEP})$

$Candidate.Add += E_{ij}$

$Surpprise += S_{ij}(m)$

$Explains += EP$

$if(Explains > T_{EP})$ // 忽略影响较小的节点

$Candidate.Surprise = Surprise$

$ExplanatorySet += Candidate$

$break$

$Final = ExplanatorySet.SortDescend(Surprise)$

$ReturnFinal.take(n)$ //按照S值降序排列，输出结果

### 2.3 试验结果

根据上述分析算法，使用真实的打分数据进行了相关试验，以验证该算法在数据集上的有效性。试验之前，使用前述用例集合对一款自研语音助手进行了全量的打分，生成原始打分数据并计算各个维度的打分以及整体的任务完成指标打分。此外，对原始数据集随机添加不同数量的异常点，通过统计该算法的识别效果验证上述算法的有效性。

结果如表4所示，可以发现在精确度方面和人工分析的差距约为10%，且当异常和数据量增加时，算法性能有所下降。这样的性能在生产过程中是可接受的，同时，结合一些规则工具，实际的问题定位精确度可以进一步提升。本文只介绍纯算法的性能。

表4 Adtributor算法试验结果 %

| 数据集 | P-Top@3 | P-Top@5 | 人工-Top3 | 人工-Top5 |
|---|---|---|---|---|
| C1-3 | 87.2 | 95.1 | 98.1 | 99.2 |
| C5-3 | 83.6 | 91.0 | 96.1 | 98.5 |
| C10-3 | 80.9 | 82.2 | 89.2 | 92.0 |
| C1-5 | 82.3 | 86.7 | 90.0 | 91.5 |
| C5-5 | 81.0 | 84.0 | 87.6 | 90.0 |
| C10-5 | 65.6 | 71.2 | 77.6 | 80.5 |

表4中，C1-3代表在1 000条数据中注入3条异常，C10-5代表在10 000条数据中注入5条异常，以此类推。其中，P-Top为使用本模型进行评估的得分，人工-Top为使用人工评估后的得分。

## 3 相关工作

本模型已应用于正常的研发过程，使用本模型对市场上的车型进行了多次全量竞品分析，下面列出部分分析数据。表5所示为整体评估指标，表6所列为语义部分评估指标。可以看出，本模型可以对语音助手整体做出量化的评估，也可以按语义功能进行评估，维度更多更深，能够充分分析市场上车载语音产品的表现。

表5 整体指标评估结果 %

| 车型 | 任务完成 | 意图识别 |
|---|---|---|
| 竞品1 | 45.1 | 62.4 |
| 竞品2 | 62.0 | 83.7 |
| 竞品3 | 44.8 | 57.6 |
| 竞品4 | 38.3 | 67.1 |
| 竞品5 | 28.9 | 58.9 |

表6 部分语义指标评估结果 %

| 车型 | 否定意图 | 语法纠正 | 指代 | 多意图 |
|---|---|---|---|---|
| 竞品1 | 19.1 | 54.4 | 29.4 | |
| 竞品2 | 28.0 | 73.7 | 45.7 | 11.7 |
| 竞品3 | | 57.6 | 17.6 | |
| 竞品4 | 21.3 | 61.1 | 39.1 | 12.1 |
| 竞品5 | | 43.9 | 29.9 | |

## 4 结束语

本文介绍了一个车载语音助手评估模型，该模型

的设计背景来源于实际的生产项目。解决了车载语音助手研发过程中，设计开发人员在产品分析和问题定位过程中的问题。在构建大量模拟真实交互环境的数据集合的基础上，设计了分层指标评估模型和问题定位算法，并应用于实际研发过程，有效提高了产品质量以及研发效率。此外，本文仅阐述评估模型的核心思路及算法，实际生产过程中会用到一些自动化的辅助工具以提升系统工作效率和规范化输出。

随着需求的不断变化，本模型也在不断迭代更新，如计划在功能维度和语义维度之外新增环境维度，通过还原车辆和用户所处的环境，如设计高速行驶、城区道路行驶、车窗状态、车内噪声环境等，使评估过程更贴切拟合实际场景。

基于单独Adtributor算法的模型问题定位能力比人工定位能力弱，计划额外引入HotSpot方法，通过投票决策的方式进行问题定位，以提升成功率。

## 参 考 文 献


[1] 杨超. 2021年中国车企数字化转型趋势研究报告[J]. 数字经济, 2021(12): 60-68.

[2] 赵婷婷, 宋亚静, 李贵喜, 等. 基于深度强化学习的文本生成研究综述[J]. 天津科技大学学报, 2022, 37(2): 71-80.

[3] 韩伟华. 智能车载信息娱乐系统交互设计研究[D]. 青岛：青岛大学, 2021.

[4] 张笛, 杨婷婷, 沙通. 智能终端语音助手标准化研究[J]. 广东通信技术, 2019, 39(12): 10-15.

[5] 赵一鸣, 朱奕蓉, 吴林容. 智能语音助手的知识服务能力评价研究[J]. 图书与情报, 2019(4): 132-140.

[6] 冯开来. 基于特征对齐的中文分词和用户标识识别研究[D]. 重庆：重庆邮电大学, 2019.

[7] 史亚楠, 代天文. 浅谈车载语音的现状与发展[C]//第十五届河南省汽车工程科技学术研讨会论文集, 2018: 32-33.

[8] 秦颖. 机器生成语言的质量评价方法综述[J]. 计算机工程与科学, 2022, 44(1): 138-148.

[9] 冯鲁汉. 智能运维中多维监测指标的异常定位研究[D]. 西安：西安电子科技大学, 2019.

[10] 王鑫, 张涛, 金映谷. 异常检测算法综述[J]. 现代计算机, 2020(30): 21-26.

[11] 刘正望. 基于用户行为的根因分析方法研究与设计[D]. 北京：北京邮电大学, 2021.

[12] 肖开发. 多维时序数据根因定位关键技术的研究[D]. 大连：大连理工大学, 2020.

[13] SUN Y Q, ZHAO Y J, SU Y, et al. HotSpot: Anomaly Localization for Additive KPIs with Multi-Dimensional Attributes[J]. IEEE Access. 2018, 6(2): 2169-3536.


**【作者简介】**


道发发（1997-），男，就职于一汽-大众汽车有限公司，算法工程师，机器学习方向。

丁敏（1992-），男，就职于一汽-大众汽车有限公司，算法工程师，自然语言处理方向。

袁粲璨（1989-），女，就职于一汽-大众汽车有限公司，测试工程师，语音测试方向。

陈晓军（1989-），男，就职于一汽-大众汽车有限公司，算法工程师，自然语言处理方向。

黎小平（1984-），男，就职于一汽大众汽车有限公司，架构师，应用架构方向。

赵嵩（1978-），男，就职于一汽-大众汽车有限公司，车联网部部长兼摩斯智联科技有限公司总经理。