TransMode-LLM：基于特征的自然语言建模：通过领域增强提示进行出行行为建模

张美静
新加坡科技设计大学工程系统与设计，
新加坡索马帕路8号，meijing_zhang@sutd.edu.sg

徐莹*
新加坡科技设计大学工程系统与设计，
新加坡索马帕路8号，xu_ying@sutd.edu.sg

arXiv:2601.13763v1 [cs.CE] 20 Jan 2026

摘要

理解旅行者行为并准确预测出行方式选择，是交通规划与政策制定的核心。传统统计方法和机器学习方法难以捕捉个体决策过程背后复杂的上下文因素及其相互依赖关系。大语言模型（LLM）借助其关于人类决策的预训练知识，为弥补这些不足开辟了新途径。本研究提出了 TransMode-LLM，这一创新框架融合统计方法与基于大语言模型的技术，利用出行调查数据预测出行方式。该框架分为三个阶段：（1）通过统计分析识别关键行为特征；（2）通过自然语言编码将结构化数据转换为上下文描述；（3）采用多种学习范式，包括零样本学习、单样本/少样本学习和领域增强提示，使大语言模型适配并预测出行方式。我们使用真实出行调查数据，在不同样本量下评估了 TransMode-LLM，并采用通用模型（GPT-4o、GPT-4o-mini）和推理型模型（o3-mini、o4-mini）进行测试。大量实验结果表明，与最先进的基线分类器模型相比，基于大语言模型的方法取得了具有竞争力的准确率。此外，少样本学习显著提升了预测准确率；例如，提供 5 个示例后，o3-mini 等模型的准确率提升最高可达 42.9%。然而，领域增强提示对不同大语言模型架构的影响并不一致。具体而言，它有助于提升通用模型的性能，GPT-4o 的提升幅度为 2.27% 至 12.50%。但对于推理型模型（o3-mini、o4-mini），领域知识增强并不能普遍提升性能。本研究推动了大语言模型在出行行为建模中的应用，为未来的学术研究和交通政策制定提供了有益且有价值的见解。

关键词 大语言模型；旅行行为分析；出行方式选择预测；交通规划

1 引言

了解旅行者行为并准确预测个人出行方式选择，对交通规划和政策制定至关重要。此前大多数出行方式预测研究采用传统统计模型（如多项Logit模型）或先进的机器学习技术，这些方法都依赖结构化数据来建立旅行者特征与交通方式选择之间的关系 [1, 2]。尽管这些传统方法能够捕捉统计规律，却难以捕捉复杂的情境因素以及内在的个体决策过程背后的相互依存关系。更重要的是，这些方法无法对新情况或不断变化的情境进行推理，因此难以适应变化的环境或政策干预。由此，这些传统方法可能无法实现理想的预测准确率，也难以对出行者如何应对环境变化、基础设施改善或政策干预并调整选择进行建模。近年来，大语言模型（LLM）的迅速发展为应对这些挑战带来了革命性的机遇。大语言模型在海量文本数据上训练，展现出显著优势，能够理解上下文、推理复杂关系并生成类人回复。

近期，研究人员日益关注将大语言模型的先进技术用于理解和预测法律、经济学、政治学和社会科学等各个领域中的人类行为 [3, 4, 5, 6]。这一趋势主要源于这些领域侧重于处理具有天然文本形式的数据所涉及的语言处理任务。相比之下，交通运输研究主要使用类别变量和数值变量来表示行为信息，而非自然语言，这限制了大语言模型在该领域的直接应用。尽管存在这一差距，交通运输领域近期的研究已开始展现基于大语言模型的方法在出行行为预测方面的潜力 [7, 8, 9]。然而，针对将大语言模型应用于出行方式预测任务的系统性研究仍然缺失。受此基础启发，我们提出了 TransMode-LLM，这是一种创新框架，旨在通过将结构化行为数据转换为对出行者及其出行的自然语言描述来预测出行方式。我们的方法首先通过基于文献的变量选择和特征重要性分析进行统计分析，从而识别关键行为特征。随后，我们将选定的结构化变量转换为叙述性描述，使大语言模型能够处理并推理传统建模方法可能忽略的上下文信息。为提升大语言模型在交通运输专属任务上的表现，我们探索了多种学习范式（零样本和单样本/少样本学习），以了解它们对使用自然语言进行出行方式预测的影响。我们进一步提出一种领域增强提示策略，其中融入了标准化出行方式定义和三步结构化决策流程。为验证所提出的基于大语言模型的方法，我们在真实世界数据集上开展实验，评估模型在不同样本量下使用多种大语言模型时的表现。我们通过三项系统性比较评估性能：基于大语言模型的方法与传统统计分类器的比较、零样本与少样本学习范式的比较，以及采用与未采用领域知识增强时的性能比较。结果表明，所提出的基于大语言模型的方法有望达到与成熟基线模型相当的性能。结果还表明，少样本学习显著提升了预测准确率，例如 o3-mini 等模型在提供 5 个示例时，提升幅度稳定达到最高 42.9%。此外，领域增强提示也有助于提升性能，GPT-4o 的提升幅度为 2.27% 至 12.50%。为全面评估所提出的方法，我们进一步构建了一个分层评估体系，并将 F1 分数（宏平均 F1 和加权 F1）作为补充性能指标。F1 分数分析表明，少样本学习能够有效应对类别不平衡，使准确率与宏平均 F1 的差距最多缩小 71%。通过将传统出行行为预测方法与大语言模型的上下文智能相结合，所提出的方法本研究为突破传统局限、推进出行方式预测提供了一条新颖路径。

2 相关研究

2.1 出行方式预测

近年来，学术界对出行方式预测的兴趣日益浓厚。研究人员开发了多种出行方式预测技术，例如传统统计模型和先进的机器学习模型。早期关于出行方式预测的大多数研究都聚焦于离散选择模型，例如多项Logit模型和嵌套Logit模型。这些模型基于随机效用理论，根据出行者对不同出行方式的感知效用，估计其选择特定出行方式的概率。效用通常由出行时间、出行成本和舒适度等各类属性构建而成[10]。虽然这些模型提供了理论基础且具有可解释性，但其无关替代方案独立性假设以及参数线性效用函数限制了模型捕捉现实出行行为中复杂相互依赖关系的能力 [1], , 因此可能无法充分反映现实出行行为的复杂性。为克服传统离散选择模型的局限，研究人员日益关注将先进的机器学习技术应用于出行方式预测。多项研究证实了机器学习技术在出行方式预测中的有效性 [11, 12, 13, 2]。然而，这两类方法仍存在局限，因为它们依赖结构化特征表示，而这些结构化表示无法捕捉出行决策中丰富的情境复杂性及相互依赖关系。大语言模型有望利用其处理并推理出行信息自然语言描述的能力，弥补这些不足。

2.2 大语言模型在出行方式预测中的应用

大语言模型（LLM）是指基于 Transformer 架构、包含数十亿甚至更多参数，并在海量文本数据上训练的语言模型 [14], , 例如 GPT-3 [15], GPT-4 [16]和 LLAMA [17]。这些模型展现出强大的自然语言理解能力，并能通过文本生成解决复杂任务。大语言模型在出行方式预测中的应用是一个新兴研究领域，已有初步且令人鼓舞的成果 [7, 8, 9]。

[7] 首次将大语言模型应用于复杂的出行行为预测任务，且无需依赖传统的基于数据的训练。作者提出了一种零样本提示框架，将任务描述、出行特征、乘客属性以及结合领域知识的思维引导纳入其中。基于 Swissmetro 陈述偏好数据集的实验表明，基于大语言模型的方法在准确率和F1分数方面展现出具有竞争力的预测性能

与多项Logit模型、随机森林和神经网络等传统方法相比， [8]指出，零样本性能仍存在稳健性有限的问题。为解决这一局限，他们在零样本预测方法之外，还采用了少样本预测和基于角色设定加载的预测。在同一个 Swissmetro 陈述偏好数据集上进行测试后，他们发现，这两种方法都提升了对人类行为的预测性能，其中角色设定加载方法在三种方法中表现最佳。 [9] 进一步通过引入检索增强生成（RAG），研究了基于大语言模型的出行方式选择预测。他们评估了四种检索策略：基础 RAG、结合平衡检索的 RAG、结合交叉编码器重新排序的 RAG，以及同时结合平衡检索和交叉编码器重新排序的 RAG。他们发现，与传统方法相比，基于大语言模型的方法展现出更强的泛化能力；结合平衡检索与交叉编码器重新排序的策略取得了最高的预测准确率，显著优于传统基线方法。这些研究展现了大语言模型在出行方式选择预测中的潜力。然而，现有工作仍处于初步阶段，目前尚无针对出行方式预测任务调整大语言模型的系统框架。这一空白促使我们系统地探索如何通过自然语言转换和领域增强提示策略，将大语言模型有效应用于结构化交通数据。

尽管针对交通运输的此类研究仍然有限，但大语言模型在不同领域（即经济学、社会科学）中用于理解和预测行为的广泛应用，展现了其模拟人类决策的能力 [4, 6, 18, 19]。 [4] 提出了图灵实验，用于评估语言模型在实验环境中模拟人类行为不同方面的能力。作者成功复现了经典行为研究中的发现（即最后通牒博弈、花园路径句和米尔格拉姆电击实验）。研究结果表明，更大的模型（如 GPT-3 text-davinci-002）能够忠实再现复杂的人类行为模式，包括性别敏感型决策和对权威的服从，从而确立了大语言模型作为行为分析与预测的有前景工具的地位。 [6]进一步全面考察了大语言模型在计算社会科学中的潜力。通过在 25 项具有代表性的英文计算社会科学任务（包括行为任务）上测试 13 种语言模型，他们发现，大语言模型在约一半的分类任务中与人类表现出中等至较高的一致性，并且在生成式解释任务中表现更优。这些发现表明，大语言模型能够捕捉并预测人类行为模式；这一前提促使我们探究大语言模型在出行方式选择预测中的适用性，将其视为一种人类决策行为。

鉴于出行方式选择的数据通常为结构化的表格数据（例如家庭出行调查数据），将大语言模型（LLMs）扩展至结构化数据的方法学进展尤为相关。 [18]提出了 TabLLM 框架，该框架将表格数据序列化为自然语言表征，以用于分类任务。作者展示了大语言模型能够有效利用先验知识，在多个领域（即医疗保健、金融、消费和游戏行为任务）中预测复杂行为，且所需训练数据极少。 [19] 聚焦于实际应用，并提供了专为行为科学家设计的系统性教程框架，展示了如何通过恰当的实验设计和评估方案，有效部署开源大语言模型，以解决行为科学中的研究问题。受这些跨领域成功案例的启发，我们将这些方法学见解应用于交通运输领域，利用大语言模型已展现出的行为预测能力。

2.3 研究空白与贡献

尽管人们对将大语言模型应用于出行行为预测的兴趣日益浓厚，现有文献中仍存在若干重要空白。首先，先前研究缺乏针对自然语言提示的全面特征选择分析[7, 8, 9]。由于基于大语言模型的方法会将结构化表格数据转换为叙述性描述，因此确定应纳入哪些特征至关重要。考虑到令牌限制以及不太相关的特征可能带来的噪声，这种方法也引发了关于提示效率的问题。

其次，针对出行方式预测的提示策略尚未得到充分的系统研究。尽管已有研究通过通用推理指导 [7] 和检索增强生成策略 [9],探索了领域知识整合，但关于如何将标准出行方式定义、结构化决策框架等交通运输领域知识融入提示的不同策略，仍缺乏系统研究。

第三，现有研究缺乏在不同情境（如模型配置、数据集规模和评估指标）下开展的大规模对比实验与严谨分析。现有研究仅在单个数据集上评估大语言模型的性能，且对比分析有限 [7, 8] ；或主要聚焦于单个地理区域内的检索增强生成策略优化 [9]。这引发了关于基于大语言模型的方法的普适性及其边界条件的问题，从而限制了我们对这类方法在哪些情况下更具优势的理论认识，也限制了为其在交通运输规划中的应用提供实践指导。

为弥补这些不足，我们提出了一种名为 TransMode-LLM 的新型框架，旨在根据对出行者及其出行的自然语言描述预测出行方式。我们的方法从出行调查数据出发，随后实施统计分析，通过基于文献的变量选择和特征重要性分析识别关键行为特征。接着，我们采用自然语言编码，将这些选定的特征转换为

叙述性描述，使大语言模型能够处理上下文信息并进行推理。具体而言，我们采用多种学习范式，包括零样本和单样本/少样本学习，以及领域增强提示策略，以提升大语言模型在出行方式预测方面的预测能力。为验证所提出的基于大语言模型的方法，我们在真实世界数据集上开展实验。我们通过三项系统性比较评估其性能：基于大语言模型的方法与传统统计分类器的比较、零样本与单样本/少样本学习范式的比较，以及有无领域知识增强时的性能比较。我们根据准确率和 F1 分数（宏平均 F1 和加权 F1）评估性能，以进行全面评估。

3 问题表述

传统上，出行方式选择预测被表述为一个分类问题：将描述个人、家庭、行程特定、建成环境和经济因素的特征直接映射到预测旅行者会选择的交通方式。然而，这种从特征直接映射到出行方式的方法可能无法充分捕捉人们在选择交通方式时自然会考虑的复杂情境关系和决策过程。本研究提出一种新的问题表述方式，利用自然语言处理更好地捕捉这些情境关系，而不是将特征直接映射到出行方式。具体而言，重新表述后的问题可分解为两个连续步骤：首先，将结构化特征转换为连贯的自然语言描述；其次，基于这些描述，应用大语言模型进行出行方式预测。

形式化地说，该问题可表述如下：令 $X = \{x_1, x_2, \ldots, x_n\}$ 表示某条行程记录的一组特征，其中每个特征 x_i 代表个人社会人口特征、家庭特征、行程特征、建成环境因素或价格因素。这些特征由数值变量和分类变量构成。我们定义一个转换函数 D，将该结构化特征向量转换为连贯的自然语言描述：

$$D : X \to \mathcal{T} \tag{1}$$

其中，$\mathcal{T}$ 表示与出行方式选择相关的自然语言描述空间。该转换函数由基于模板的生成器实现，遵循一致的叙述结构，确保原始特征中的关键决策因素得到保留，并在语言表示中融入上下文。

基于这种自然语言表示，预测问题旨在估计交通方式 $m \in \mathcal{M} = \{$汽车，步行，巴士 $\ldots\}$上的条件概率分布 $P(m|D(X))$。我们定义函数 Prompt$(D(X))$，通过在自然语言描述 $D(X)$的基础上添加必要的任务说明和输出格式要求，来构建完整输入，从而引导模型生成准确且结构化的预测。该概率分布如下：

$$P(m|D(X)) = \text{LLM}_\alpha(\text{Prompt}(D(X)), \theta) \tag{2}$$

其中，α 表示具体的模型架构（例如 GPT-4o、GPT-4o-mini、o3-mini、o4-mini），θ 表示学习范式（例如零样本、单样本/少样本）。随后，预测的交通方式可表示为：

$$\hat{m} = \arg\max_{m \in \mathcal{M}} P(m|D(X)) \tag{3}$$

该方法选择预测概率最高的交通方式。

4 方法

继第3节之后，下面将解释我们的方法框架如何转换交通数据，以实现基于大语言模型的交通方式预测。如图1所示，该框架从出行调查数据出发，通过包含三个核心阶段的系统化流程运行。首先，原始出行调查数据经过特征分析，以识别对交通方式选择影响最大的特征。其次，将这些选定特征转换为大语言模型能够处理的上下文描述。最后，借助这些上下文描述，通过我们的预测框架对通用领域大语言模型进行适配，以完成出行方式预测任务。三个核心阶段正式 обозначa为如下：

- 阶段 1——特征重要性分析与选定：我们采用多种特征重要性方法并结合集成排序，以识别最具预测能力的变量。
- 阶段 2——自然语言特征表示：我们系统地将选定的结构化特征转换为描述性文本表示，使大语言模型能够对出行方式预测执行上下文推理。
- 阶段 3——通用大语言模型出行方式预测适配方法：本阶段通过实施多种学习范式（零样本和单样本/少样本）、领域增强提示策略、领域特定知识以及结构化推理处理方法，使通用大语言模型适用于出行方式预测，旨在引导通用大语言模型提供准确的出行方式预测结果。

接下来，我们将在以下各节中逐一介绍每个阶段。

4.1 特征重要性分析与选定

在利用自然语言转换使大语言模型能够进行上下文推理、深入探讨方法的核心之前，本节先进行初步

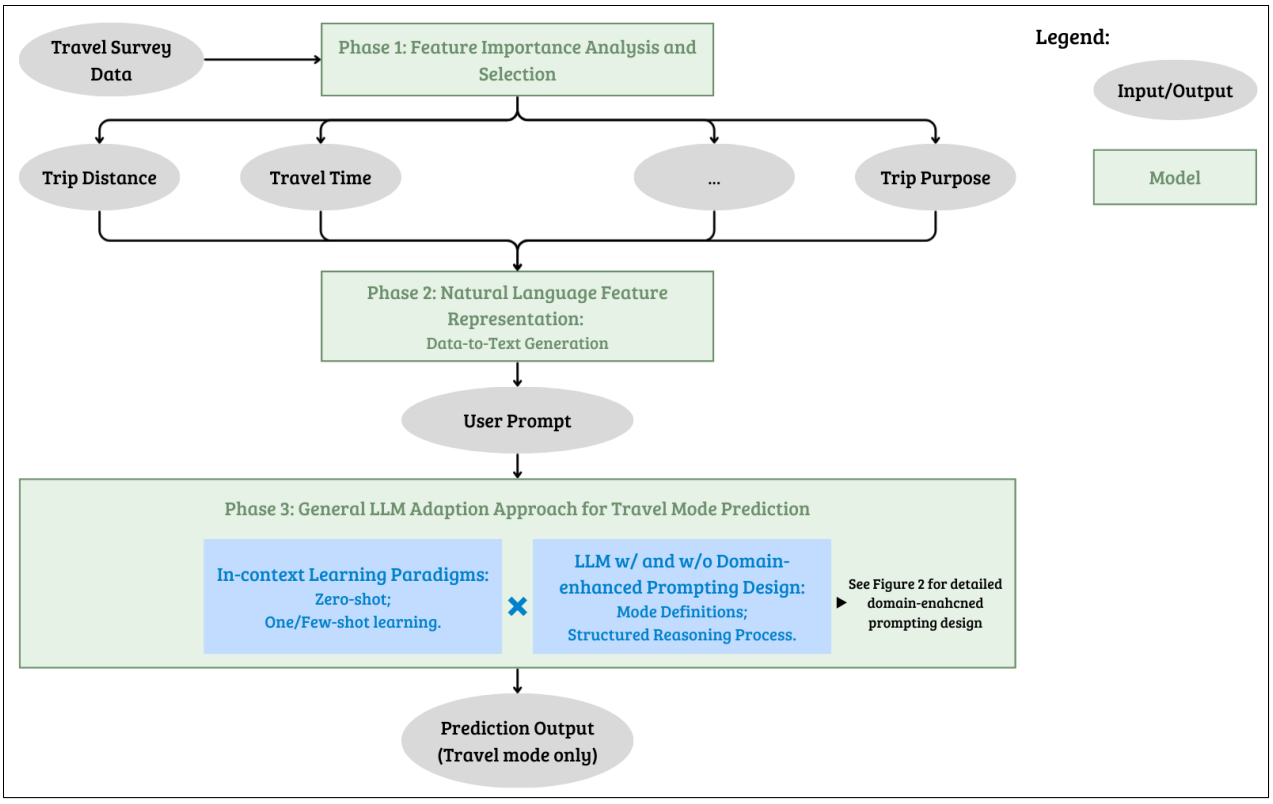


图1：基于大语言模型的出行方式预测研究框架：我们的 TransMode-LLM 模型。

分析，以从出行调查数据集中识别并确定影响变量的优先级。我们首先根据文献确定可能产生影响的因素，然后应用特征重要性分析，选出影响最大的前 15 个因素。这种两阶段筛选过程有助于降低维度，防止过拟合，并集中计算资源，以捕捉与出行方式选择行为最相关的因素。

4.1.1 出行行为的影响因素

为了进行出行方式选择预测，全面理解影响出行行为的决定因素至关重要。Zhou（2012）建立了一个影响交通方式选择因素的六类框架：物理环境与城市形态、方式特定属性、出行者个人特征、出行参数、出行需求管理（TDM）措施以及心理因素 [20]。在此基础上，并借鉴 Hörl 和 Balac（2021）提出的方法论 [21], , 本研究探讨了五类不同的影响因素：个人社会人口特征、家庭特征、出行特征、建成环境和价格因素。这种细化后的分类能够更有条理地分析个人、家庭、出行、建成环境和经济变量之间的复杂交互作用，而这些因素共同塑造了出行行为模式。

- **个人社会人口特征：**个人社会人口特征包括出行方式使用、出行距离、出行频率等。[22, 23]这些特征直接影响出行行为，也会通过影响生计活动、维持活动、休闲活动及其他活动参与而间接影响出行行为。

- **家庭特征：**
  Hu 等人（2023）最近开展的一项文献综述指出，家庭信息与个人活动及出行行为之间存在关联 [24]。其他研究也指出了类似发现 [22, 23, 25]。这些研究描述的家庭特征包括子女数量、劳动者数量、车辆数量、家庭收入、家庭类型和位置。

- **行程特征：**
  为全面捕捉出行的特征，本文纳入三个类别：地理空间背景、时间模式以及出行期间参与的活动。地理空间背景指的是出行的出发地和目的地、途中停靠点以及出行距离。时间模式包括出行日的类型（工作日或周末）、出发和到达时间、停靠

时间和出行时长。至于参与的活动，相关属性通常包括总体出行目的、途中停靠点的数量及其对应目的。

- **建成环境：**
  多项研究探讨了被认为会影响出行方式选择的建成环境因素 [26, 27, 28, 29]。Ewing 和 Cervero（2010）将影响出行方式选择的建成环境特征归纳为密度、多样性、设计、目的地可达性和距公共交通距离 [27]。
- **价格因素：**
  经济考量是影响出行方式选择的重要因素，因为直接的经济激励和限制会左右出行方式决策。交通系统中的定价因素，尤其是燃油价格和停车成本，作为市场信号，会直接影响不同出行方式的相对负担能力。停车成本被视为强有力的出行需求管理工具，在存在私家车使用替代方案的城市环境中，能够显著改变交通方式选择决策。多项研究探讨了停车成本对出行方式选择的影响 [30, 31, 32, 33]。同样，研究也表明燃油价格会影响出行方式选择 [34, 35, 36]。

4.1.2 特征重要性分析

为从文献中确定对出行方式选择影响最大的因素，我们继续开展系统性的特征重要性分析。遵循 Cherepanova 等人（2023年）对表格数据集上的多种特征选择方法进行比较的研究 [37], , 我们采用多种特征选择方法，以克服算法特有的偏差，并确保能够稳健地识别影响因素。分析涵盖六种基础方法（单变量统计检验、Lasso、随机森林、XGBoost、第一层 Lasso 和深度 Lasso）以及三种高级技术（自适应组 Lasso、LassoNet 和采用注意力图重要性的 FT-Transformer ）。

对特征选择方法的性能评估显示，XGBoost 的分类准确率最高（87.55%），其次是单变量统计检验（86.46%）和第一层 Lasso（86.01%）。通过汇总所有算法的排名，我们确定了对出行方式选择影响最大的前 15 个因素。

表1总结了所有经过考察的候选因素、其定义、均值排名以及最终的选择状态。从表中显而易见，出行距离是最稳定的重要影响特征（均值排名：4.3），其次是出行时间（8.3）、年龄（8.3）和驾驶执照状态（8.7）。在我们的排名系统中，数值越低表示特征重要性越高，因为排名代表所有特征选择方法中的平均位置，影响最大的因素排名更接近1。这种基于排名的汇总方法综合了多种特征选择算法的结果，从而增强了稳健性，并降低了对任何单一方法偏差的依赖。

4.2 自然语言特征表示：数据到文本生成

为了让大语言模型能够处理数据，需要将选定的表格特征转换为自然语言描述。按照第 4.1 节所述的特征选择流程，我们将选定的 15 个特征转换为每条行程记录对应的结构化文本描述。这一转换会将表格数据的每一行（选定特征）转化为叙述性描述，将个人和家庭特征、行程属性、建成环境及价格因素整合为一个场景。叙述形式使大语言模型能够利用其预训练过程中形成的对出行概念及特征间上下文关系的语义理解。每次出行均使用固定模板进行描述，如下所示，其中粗体文本表示选定的 15 个特征。

> 请考虑一位***44***岁女性，她是一名驾驶员，并且有工作。她住在一个有***3***人的家庭中，该家庭拥有***1***辆车，家庭收入为*125,000*至*149,999*美元，居住在一套有抵押贷款的自有住房中。她此次购物出行，距离 为*1.3*英里，预计出行时间为*10*分钟。她居住在城市地区，无法使用轨道交通，所在大都市统计区 人口为*500,000*至*999,999*，汽油价格为每加仑*4.30*美元。她最有可能选择哪种交通方式完成此次出行?

生成的描述提供了全面的叙述，既涵盖了每次出行的关键背景信息，又保留了原始特征向量的信息内容。这种方法使大语言模型能够以符合其预训练方式的格式处理交通数据，从而更细致地解读各项特征之间的关系。

4.3 出行方式预测的通用大语言模型适配方法

在第 4.2 节提出自然语言特征表征的基础上，我们现在转而探讨将通用领域大语言模型应用于出行方式预测时面临的关键挑战。尽管通用领域大语言模型具备出色的语言理解能力，但由于缺乏对领域专用术语的接触，它们并未针对出行方式预测任务进行优化，

类别	因素	定义	Mean 排名	选定
Trip 特征	出行距离	以英里计的出行距离	4.3	✓
	出行时长	以分钟计的行程时长	8.3	✓
	出行目的	出行的大致目的（例如工作、购物、社交）	11.6	✓
	具体出行目的	详细的出行目的类别	14.7	×
	日期类型	出行发生在周末还是工作日	17.9	×
	开始时间	行程开始时间（24小时制）	16.1	×
	结束时间	行程结束时间（24小时制）	15.3	×
个人 特征	Age	受访者年龄（岁）	8.3	✓
	性别	受访者性别	9.3	✓
	驾驶执照状态	受访者是否持有有效的驾驶执照	8.7	✓
	就业状态	受访者是否就业	11.0	✓
家庭 特征	家庭规模	家庭总人数	10.3	✓
	车辆数量	家庭车辆总数	11.3	✓
	住房所有权	住房是自有、租赁还是其他安排方式	11.6	✓
	家庭收入	家庭年收入	10.7	✓
	驾驶员数量	家庭持有驾驶执照的驾驶员数量	13.6	×
	劳动者数量	家庭就业人数	14.4	×
建成环境	城市/农村 设计a-tion area	家庭是否位于城市或农村地区	9.4	✓
	人口规模	五年期大都市统计区人口规模类别 ACS	13.4	✓
	铁路可用性	大都市统计区是否可使用重型轨道交通	11.6	✓
	汽油价格	出行日所在周的地区汽油价格（美分）	12.0	✓
	生命周期分类	出行日家庭生命周期分类	16.0	×
价格因素	停车成本	停车支付金额	16.3	×

表1：影响出行行为的因素及特征选择结果汇总

出行方式兼容性约束以及准确预测所需的结构化推理模式。为应对这些局限性，我们提出了一个综合框架，采用两种领域专业化技术：（1）上下文学习范式；（2）领域增强提示设计。上下文学习范式在提示中提供任务示例，包括零样本、单样本和少样本配置，以实现大语言模型的适应。领域增强提示设计提供标准化出行方式定义和结构化推理流程，将交通运输知识显式注入模型。关于这两种技术的更多细节，我们分别在第4.3.1节和第4.3.2节中介绍。此外，在第4.3.3节中，我们介绍了为进行系统比较而选择大语言模型的依据和标准。

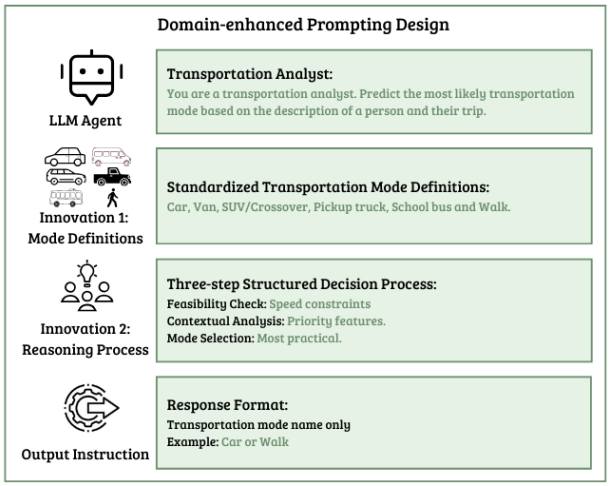


图2：基于大语言模型的出行方式预测领域增强提示设计。

4.3.1 上下文学习范式

上下文学习范式可以通过提供特定于任务的示例来应对领域适应挑战，这些示例能够弥补一般大语言模型训练语料库中对出行方式预测代表不足的问题。示例在推理时作为条件输入提供。遵循Brown等人 [15]，的做法，我们根据示例的可用性实现了三种学习范式，以研究提供示例所产生的影响：零样本学习（无示例）单样本学习（单个示例）和少样本学习（多个示例）。

零样本学习通过在不进行任何特定任务训练或提供示例的情况下进行预测，检验大语言模型的内在能力；预测仅依赖预训练知识以及对自然语言描述的上下文理解。相比之下，单样本学习和少样本学习分别在推理时提供一个和少量特定任务示例作为条件，同时不允许更新权重 [15]。我们通过单个示例实施单样本学习，并采用2、3、5和10个示例实施少样本学习，以系统评估示例数量对预测性能的影响。对于单样本和少样本配置，我们使用分层抽样从训练数据中选取示例，以确保出行方式的代表性。抽样策略优先考虑方式多样性：首先为每种出行方式至少抽取一个示例，然后根据训练数据中的方式频率按比例平衡其余示例，从而在保持代表性的同时确保行程特征多样。

4.3.2 领域增强提示设计 n

尽管上下文学习范式可以通过针对特定任务的示例提升领域适应能力，但通用大语言模型在交通运输应用中仍面临根本性局限。由于预训练数据中与交通运输相关的知识不足，它们可能缺乏出行行为术语和概念。此外，通用

大语言模型缺少用于系统评估出行决策的结构化推理框架。为应对这些挑战，我们开发了一种领域增强提示框架，将大语言模型设定为虚拟交通分析师，使其能够根据自然语言描述的行程预测出行方式。我们的框架包含两项关键技术创新：（1）标准化交通方式定义，用以提供一致的领域知识；（2）三步结构化决策流程，用以指导系统推理。图2展示了这一提示设计的完整架构。以下小节将详细介绍这两项创新。

第一项创新是引入受《国家家庭出行调查》 [38]启发的标准化出行方式定义。通过 Chatbox AI [39], 进行的预测试发现，缺乏领域知识的大语言模型会误将“汽车”理解为宽泛的车辆类别，无法区分厢式货车、SUV/跨界车和皮卡。通过引入标准化定义（见图2中的“出行方式定义”面板），我们可以提供明确的交通术语，帮助模型正确区分车辆类型及其他出行方式类别。

第二项创新是开发一种由机器学习提供信息的三步思维链推理流程，依据经验得出的特征层级来组织预测。如图2所示，推理流程首先从可行性检查（步骤1）开始，计算所需出行速度（距离/时间），并根据速度约束排除物理上不可行的出行方式。在步骤2中，情境分析引导模型按照系统性特征选择分析（第4.1.2节）得出的特征重要性排名，将各因素依照其经验确定的重要性顺序纳入思维链结构。例如，出行距离（平均排名4.3）和出行时间（平均排名8.3）是预测出行方式选择时影响最大的特征。最后，出行方式选择（步骤

3）引导模型整合所有已评估因素，并选择最实用的交通方式，同时确保遵守领域约束（例如，校车仅用于与学校相关的出行）。Wei 等人（2022年）证明，对于具有挑战性的多步骤文本到文本任务，思维链提示比其他方法更有效[40]。鉴于出行方式预测需要整合多个维度的信息，所提出的三步框架通过纳入特定领域的可行性约束和基于经验得出的特征优先级，扩展了这一概念，从而提供准确预测出行方式所需的结构化推理路径。

4.3.3 模型选择

大语言模型的快速发展催生了多样化的模型生态，这些模型在规模、架构设计和功能专门化方面各不相同。GPT-4o 等通用模型 [41] 旨在处理广泛的任务，而 GPT-4o-mini 等紧凑型变体则以较低的计算成本提供资源高效的推理 [42]。近年来，面向推理的专用架构逐渐出现，其中包括 O 系列模型（例如 o3-mini 和 o4-mini）。这些模型代表了第三代和第四代推理框架，专门针对多步骤逻辑推断进行了优化 [43, 44]。在本研究中，我们从这两类模型中选取了四个模型：（1）不同规模的通用模型 GPT-4o 和 GPT-4o-mini；（2）面向推理的专用紧凑型模型 o3-mini 和 o4-mini。通过多样化的模型选择，我们旨在回答以下研究问题：通用架构中的模型规模是否与预测性能直接相关？无论参数数量多少，面向推理的专用模型是否能够更好地捕捉出行行为背后的复杂决策过程？此外，在出行方式预测任务的应用中，综合型模型与紧凑型模型之间存在怎样的成本效益权衡？通过对这些模型进行系统比较，我们评估模型规模与预测准确率之间的关系，考察以推理为重点的架构是否能够更好地捕捉出行行为的复杂性，并为交通运输场景中的实际部署确定具有成本效益的解决方案。

5 实验

5.1 数据准备

本研究使用的数据集是由美国联邦公路管理局（FHWA）开展的 National Household Travel Survey，这是一项定期进行的全国性调查，为交通规划人员和政策制定者的决策提供支持 [45]。出行行为分析的表现取决于输入数据的质量，尤其是在开发用于出行方式选择的生成式模型和预测模型时。因此，我们首先采用全面的两阶段筛选方法，以解决 National Household Travel Survey（NHTS）数据集中的不一致问题，避免这些问题影响模型有效性。

第一步采用速度—距离—时间一致性筛选，识别物理上不合理的行程记录，例如计算速度超过 5 英里/小时的步行行程，或速度低得不切实际的机动车行程。随后，社会人口统计一致性筛选针对个人特征与所报告出行行为之间的逻辑矛盾，例如未成年驾驶员、非驾驶员驾驶车辆、幼童无陪同乘坐公共交通，以及未识别驾驶员的私人车辆行程。筛选后，最终数据集包含 22,868 条行程记录和 85 个属性。统计验证表明，筛选后的数据集在所有出行方式类别中均呈现出更加合理的分布，其中步行、骑行和公共交通类别的提升尤为显著。通过解决物理和逻辑上的不一致问题，这些预处理步骤显著提升了数据可靠性，为出行行为分析提供了支持，并能够更准确地预测出行方式选择。

5.2 实验设计与实施

我们评估了基于大语言模型的方法在不同学习范式和提示策略下执行出行方式预测任务的性能。我们在四种样本量（100、200、500 和 1,000）下，对第4.3.3节所述的四种模型进行考察。为确保实际相关性和计算效率，本研究聚焦于六种（共20种）出行方式：汽车、厢式车、SUV/跨界车、皮卡、校车和步行；这些方式占记录出行的90%以上，涵盖了典型的出行模式。作为基线，我们选取了高性能机器学习分类器 LogitBoost 和梯度提升。[46]在 National Household Travel Survey（NHTS）2017年数据集上对来自14个模型家族的86种分类器进行了评估，并发现这两种分类器始终表现优异。该数据集在结构上与本研究使用的数据集类似。此外，为确保这两种基线方法（LogitBoost 和梯度提升）与基于大语言模型的方法之间进行公平比较，我们在每种样本量配置下，为所有方法使用相同的测试集。

关于大语言模型的实现细节，对于支持确定性输出的模型（GPT-4o、GPT-4o-mini），温度设为 0；而 O 系列模型（o3-mini 和 o4-mini）不支持调整温度，因为其推理架构采用固定的采样参数。对于单样本/少样本学习，我们按照第4.3.1节所述，使用 1、2、3、5 和 10 个示例来考察其有效性。少样本示例的选择遵循分层抽样策略，旨在最大限度地提高表征多样性并覆盖各种出行方式。所有实验均使用 Python 和 OpenAI API 实现。

5.3 模型评估

出行方式建模和预测任务被视为分类问题 [11]。据此，我们使用三种标准分类指标评估模型性能：准确率、F1_{macro} 和 $\text{F1}_{\text{weighted}}$ [47]。

5.3.1 评估指标：准确率

准确率衡量正确预测的出行方式所占的总体比例。该指标为分类任务提供了一个标准基准，因为它既反映了正确预测的总体比例，也简洁地衡量了模型整体性能。

形式化地，令 $\hat{y}_i$ 表示样本 i 的预测出行方式，而 y_i 表示实际出行方式。令 $\mathcal{D}^{\text{test}}$ 表示测试数据集，M 表示所有出行方式的集合。则准确率指标形式化定义如下：

$$\text{Accuracy} = \frac{1}{|\mathcal{D}^{\text{test}}|} \sum_{i \in \mathcal{D}^{\text{test}}} \mathbf{1}[\hat{y}_i = y_i] \tag{4}$$

(5)

其中，$\mathbf{1}[\cdot]$ 是示性函数。

然而，当类别不平衡时，准确率的表现并不理想。本研究使用的数据集类别极不平衡：超过三分之一的出行是汽车出行（37.9%），其次是 SUV/跨界车（33.5%）、皮卡（10.7%）和厢式货车（6.0%），合计占所有观测值的近 88%。相比之下，步行（6.8%）和校车（2.1%）等其他出行方式各自仅占数据的一小部分。这种不平衡现象在交通运输研究中很常见，某些出行方式往往主导出行行为。为应对这一问题，我们结合使用宏平均 F1 和加权F1分数，以衡量模型在少数和多数出行方式上的性能。与总体准确率不同，后者仅通过预测占主导地位的出行方式就可能显得虚高；结合使用宏平均 F1 和加权F1分数，则能全面评估模型在所有出行方式上的性能。

5.3.2 宏平均 F1 和加权 F1 分数

宏平均 F1 对每种出行方式赋予相同权重，因此特别适合评估占比较低但与政策相关的出行方式（如步行、骑行）。相比之下，加权 F1 根据出行方式占比加权计算表现，因此反映了总体人口层面的预测准确率。令 M 表示出行方式集合，F1_m 表示方式 m 的 F1 分数。总体宏平均 F1 和加权F1分数计算如下：

$$\text{F1}_{\text{macro}} = \frac{1}{|M|} \sum_{m \in M} \text{F1}_m \tag{6}$$

$$\text{F1}_{\text{weighted}} = \sum_{m \in M} \frac{|M_m|}{|\mathcal{D}^{\text{test}}|} \text{F1}_m \tag{7}$$

其中，M_m 表示出行方式为 m 的测试样本子集。

每个 F1_m 均根据特定出行方式的精确率和召回率计算得出：

$$\text{F1}_m = \frac{2 \cdot \text{Precision}_m \cdot \text{Recall}_m}{\text{Precision}_m + \text{Recall}_m} \tag{8}$$

$$\text{Precision}_m = \frac{\text{TP}_m}{\text{TP}_m + \text{FP}_m} \tag{9}$$

$$\text{Recall}_m = \frac{\text{TP}_m}{\text{TP}_m + \text{FN}_m} \tag{10}$$

其中，TP_m、FP_m和 FN_m 分别表示出行方式 m的真正例、假正例和假负例。

精确率可 p取 p预测准确率 y（正确 p预测在所有预测结果中正确的比例），而召回率衡量完整性（在所有实际实例中被正确识别的比例）。例如，步行的精确率为 80% 意味着，被预测为步行的出行中有 80% 是正确的，另有 20% 是假正例。

鉴于我们的数据集存在类别不平衡，精确率与召回率之间会出现权衡：

- **高精确率、低召回率：** 保守的预测可减少假正例，但会漏掉许多真实情况，从而低估需求。
- **低精确率、高召回率：** 宽松的预测可捕捉大多数真实实例，但会因假正例而高估需求。

总之，宏平均 F1 和加权F1分数提供了互补的视角：宏平均 F1 揭示模型在所有出行方式上的表现是否公平，而加权 F1 则反映与观察到的出行方式分布相一致的整体预测准确率。在后续分析中同时使用这两项指标，有助于评估准确率的提升是否转化为均衡性能，还是主要使占主导地位的出行方式受益。

6 结果

本节采用两项性能指标评估基于大语言模型的出行方式预测方法：准确率和 F1 分数（宏平均 F1 和加权 F1）。对于每项指标，我们通过三项关键比较展示基于大语言模型的方法的预测性能：（1）基于大语言模型的方法与基线分类器（LogitBoost 和梯度提升）的比较；（2）零样本

与单样本/少样本学习范式的比较；（3）领域知识增强与未进行领域知识增强时的性能比较。每项比较均涵盖四种语言模型（GPT-4o、GPT-4o-mini、o3-mini、o4-mini）和四种样本量（100、200、500 和 1000）。第 6.1 节报告准确率指标的结果，随后第 6.2 节报告 F1 分数指标的结果。

6.1 准确率

本研究的主要评估指标是准确率，即数据集中预测的出行方式与实际出行方式相符的总体比例。以下各节通过三组比较展示准确率结果：基于大语言模型的方法与基线分类器、零样本与单样本/少样本学习范式，以及有无领域知识增强时的性能。

6.1.1 性能比较：大语言模型与基线分类器

为考察基于大语言模型的方法在出行方式预测任务中是否具有竞争力，我们将零样本大语言模型的表现与第 5.2 节所述的基线分类器（Logit-Boost 和梯度提升）进行比较。预测准确率结果汇总于表 2。

这些结果表明，基于大语言模型的方法，尤其是 o4-mini 等先进推理模型，无需进行任何特定任务训练，便能达到与成熟基线分类器相当或更优的表现。这标志着交通运输应用中的预测建模范式发生了根本性转变，即从基于训练的预测转向基于大语言模型提示的预测。

传统分类器可能需要进行大量模型训练、特征工程、超参数优化，并配备持续维护所需的基础设施；而大语言模型仅凭出行的自然语言描述，通过零样本推理，就能在出行方式选择预测任务中达到相当或更高的准确率。值得注意的是，o4-mini 尽管仅使用通用网络规模数据进行训练，且不具备交通运输领域知识，仍能取得这一表现。这表明，大规模预训练能够赋予模型可迁移推理能力，使其适用于专业领域。

6.1.2 性能比较：零样本与单样本/少样本学习范式

在确认大语言模型在零样本条件下优于基线分类器后，我们现在转而考察少样本学习对出行方式选择预测性能的影响。由于大语言模型是在广泛的数据集上训练的，因此在出行方式预测等领域特定任务上仍可能表现不佳，因为其注意力机制针对的是通用知识，而非专业知识。借鉴 [15], 少样本学习通过提供上下文示例，以示范的方式引导大语言模型进行预测，而非进行参数更新，这可能弥补通用模型在领域特定知识方面的不足。

表 3 系统评估了多个大语言模型在不同样本量下的少样本学习效果。对于每种配置，我们确定最佳示例数量（从 1 到 10），并将表现最佳的少样本设置与相应的零样本基线进行比较。显然，在大多数情况下，与零样本基线相比，少样本学习显著提升了预测准确率。该方法在 16 种实验配置中的 13 种带来了提升，幅度从 3.6% 到令人瞩目的 42.9% 不等。值得注意的是，最佳性能通常只需少量示例（1–5 个）即可达到，这表明即使少量上下文引导也能显著提升模型性能。然而，我们也发现，少样本学习并非总能像预期那样优于零样本方法。一个特别值得借鉴的案例是：在样本量为 1,000 时，使用 o4-mini 模型加入一个示例后，准确率从 57.50% 降至 54.00%，性能下降了 6.1%。这表明，能力强大的推理模型或许已经具备足够的领域知识，即使不提供额外示例也能达到最佳性能。过多的示例反而可能以某些方式限制模型的推理路径，削弱其内在能力。

这一模式也与先前文献中的发现一致[15, 48]。 [48] 指出，看似细微的方法学决策（例如提示格式、示例选择和排序）会对少样本学习性能产生重大影响，准确率可能从接近随机猜测的水平变化到最先进的水平。这种差异表明，系统化的提示工程策略，尤其是基于模型能力进行自适应示例选择，可能有助于提高少样本学习的可靠性，值得专门研究；未来研究可进一步探讨这一问题。

6.1.3 性能比较：经领域知识增强与未经增强的大语言模型

在探讨少样本学习对基于大语言模型的出行方式预测性能的影响后，我们接下来研究领域知识整合对预测性能的影响。如第4.3节所述，我们提出了一种领域增强提示策略，将标准交通方式定义以及按重要性排序的因素纳入提示设计。表4展示了四种大语言模型在不同样本量下的对比结果，比较了标准少样本学习方法与领域增强方法。

方法	模型	样本量			
		100	**200**	**500**	**1000**
基线分类器	LogitBoost	**0.5100±0.058**	0.4400±0.030	0.4580±0.034	0.4810±0.022
	梯度提升	0.3700±0.060	0.4050±0.058	0.5040±0.035	0.5040±0.049
基于大语言模型（零样本）	GPT-4o	0.4000	0.4750	0.4400	0.4500
	GPT-4o-mini	0.3500	0.4750	0.4000	0.4350
	o3-mini	0.3500	0.4500	0.4400	0.4500
	o4-mini	0.4500	**0.5000**	**0.5600**	**0.5750**

表 2：基线分类器与基于大语言模型的零样本性能对比

模型	样本量	零样本准确率	最佳少样本准确率	数量示例	提升 (%)
GPT-4o	100	0.4000	0.4000	1	0.0
	200	0.4750	**0.5250**	**1**	**+10.5**
	500	0.4400	0.4400	1	0.0
	1000	0.4500	**0.4700**	**2**	**+4.4**
GPT-4o-mini	100	0.3500	**0.4500**	**2/3**	**+28.6**
	200	0.4750	0.4750	2	0.0
	500	0.4000	**0.4300**	**5**	**+7.5**
	1000	0.4350	**0.4550**	**2/3/5**	**+4.6**
o3-mini	100	0.3500	**0.5000**	**5**	**+42.9**
	200	0.4500	**0.4750**	**1/3/5/10**	**+5.6**
	500	0.4400	**0.4800**	**2**	**+9.1**
	1000	0.4500	**0.4900**	**5**	**+8.9**
o4-mini	100	0.4500	0.5000	**5**	**+11.1**
	200	0.5000	0.5500	**3**	**+10.0**
	500	0.5600	0.5800	**5**	**+3.6**
	1000	**0.5750**	0.5400	1	-6.1

Tab表 3：无领域知识增强时最佳少样本性能与零样本基线的对比 t

结果显示，在领域知识的利用方面，模型架构存在显著差异。从表4可以看出，领域知识增强对 GPT-4o 和 GPT-4o-mini 等通用模型的影响主要为正面或中性，却会降低 o3-mini 和 o4-mini 等推理系列模型的性能。具体而言，通用模型的成功率为 75%（8 种情形中有 6 种呈现正面或中性影响）；其中，GPT-4o 成功体现了领域知识整合的效果，在所有样本量下均实现了稳定提升（+2.27% 至 +12.50%）。这种普遍收益表明，GPT-4o 的架构旨在广泛适用于各类任务，并具备协调多种知识来源的内在机制。相较之下，o3-mini 和 o4-mini 等推理系列模型则呈现出相反的模式，成功率仅为 25%；在 8 种情形中，有 5 种因提供领域知识而产生负面影响。引入领域知识后，o3-mini 在样本量为 100 时的预测降幅最大，达到 -20.00%。这表明，推理系列模型的架构可能依赖特定的计算路径，而任何外部领域信息都可能干扰这些路径。

这些推理模型中观察到的负迁移，可以从架构专业化的角度加以解释。针对逻辑推理和问题求解进行优化的模型，可能会形成高度结构化的内部表征和推理机制。引入领域知识可能会干扰这些经过精细调校的推理路径，产生相互冲突的信号，导致性能下降而非提升。 [49], 指出，将外部知识注入推理能力较强的模型，可能会使其偏离原有的思维链推理路径，进而导致性能下降；这一解释与该研究的发现一致。同样， [50] 表明，即便语义相关的信息也可能成为强干扰项，非但不能增强大语言模型推理，反而会产生不利影响。例如，针对高效推理进行优化的 o4-mini，可能已经能够在没有额外上下文引导的情况下从数据中提取相关模式。添加领域知识可能会给模型的内部表征引入噪声或造成不匹配，从而对性能（即预测准确率）产生负面影响。因此，领域知识增强对于受益于明确引导的通用模型（即 GPT-4o、GPT-4o-mini）而言是一种有价值的补充，但对于推理专用架构而言，却可能成为干扰源；这类架构已发展出复杂的内部机制，能够自主进行模式发现和逻辑推断。

6.2 宏平均 F1 和加权 F1 分数

在第6.1.2节和第6.1.3节中考察了少样本学习和领域知识增强对预测准确率及准确率最优配置的影响后，我们现在进一步分析相应的宏平均 F1 和加权 F1 分数，以评估类别均衡性能。这项分析十分重要，因为在类别不均衡的分类场景中，仅凭准确率可能会产生误导：模型可能取得较高的准确率，却在少数类上的表现不佳。宏平均 F1 对每种出行方式赋予相同权重，而加权 F1 则根据各出行方式的占比进行加权，因此能够反映总体人群层面的预测准确率。

首先，我们考察准确率优化的少样本配置与其零样本基线（表 5）的 F1 分数，以评估最大化准确率的配置是否也能提升类别均衡的性能。其次，我们比较有领域知识增强和无领域知识增强的准确率优化少样本配置的 F1 分数（表 6），以确定能够提高准确率的领域知识增强是否也有助于提升 F1 分数。通过这种方法，我们可以探究准确率优化与 F1 优化是否一致，或二者是否存在差异，从而揭示整体正确性与类别均衡预测之间可能存在的权衡。

6.2.1 F1 分数性能比较：零样本与单样本/少样本

继第 6.1.2 节基于准确率的分析之后，我们进一步考察这些准确率优化配置的宏平均 F1 和加权 F1 分数，以评估高准确率是否能转化为所有类别上的均衡性能。表 5 展示了不同大语言模型和样本量下宏平均 F1 与加权 F1 分数的对比结果。

表 5 显示，在所有零样本配置中，准确率与 F1 分数之间都存在差距：宏平均 F1 比准确率低约 15–20 个百分点，而加权 F1 也呈现类似但不那么明显的差距。这一趋势表明，类别分布严重不均衡，模型偏向多数类，虽然总体准确率尚可，但在少数类上的表现不佳。值得注意的是，表中清楚显示，少样本学习能够持续缩小准确率与 F1 之间的差距，这说明上下文示例有望提升类别均衡性能。例如，样本量为 100 时，GPT-4o-mini 的宏平均 F1 提升了 124%，加权 F1 提升了 72%，而准确率仅提升了 28.6%，从而将准确率与宏平均 F1 的相对差距从 49% 缩小到 11%。这种 F1 的显著提升表明，少样本学习的主要价值在于让模型了解少数类的特征，而不只是提高整体正确性。值得注意的是，样本量为 200 时，GPT-4o-mini 更清楚地展现了这一机制。在零样本和少样本设置下，预测准确率均保持在 0.4750，但宏平均 F1 提升了 10%，加权 F1 提升了 19.8%。这证实了少样本学习能够改善类别均衡

模型	样本量	无领域 最佳少样本（示例）	有领域 最佳少样本（示例）	提升（%）
GPT-4o	100	0.4000 (1)	**0.4500 (10)**	**+12.50**
	200	0.5250 (1)	**0.5500 (5)**	**+4.76**
	500	0.4400 (1)	**0.4500 (2)**	**+2.27**
	1000	0.4700 (2)	**0.4850 (3)**	**+3.19**
GPT-4o-mini	100	**0.4500 (2/3)**	0.4000 (1)	-11.11
	200	0.4750 (2)	0.4750 (5)	0.00
	500	0.4300 (5)	**0.4900 (5)**	**+13.95**
	1000	0.4550（2/3/5）	**0.4600 (1)**	**+1.10**
o3-mini	100	**0.5000 (5)**	0.4000 (1)	-20.00
	200	0.4750（1/3/5/10）	**0.5000（3）**	**+5.26**
	500	0.4800（2）	0.4800（3）	0.00
	1000	**0.4900（5）**	0.4750（1）	-3.06
o4-mini	100	**0.5000 (5)**	0.4500 (1)	-10.00
	200	0.5500 (3)	**0.5750 (3)**	**+4.55**
	500	**0.5800 (5)**	0.5500 (2)	-5.17
	1000	**0.5400 (1)**	0.5250 (10)	-2.78

表4：有无领域知识增强条件下的小样本性能

而不受总体准确率影响。在这种情况下，加权 F1 提升幅度更大，这表明当准确率已达到最优时，少样本学习既能改善少数类预测，也能改善多数类预测。

然而，高级模型表现出截然不同的行为。例如，在样本量为 1000 时，o4-mini 在所有指标上的差距都相对较小（准确率与宏平均 F1 的差距为 22%，准确率与加权 F1 的差距为 6%）。虽然少样本提示缩小了宏平均 F1 的差距，但零样本提示的预测准确率和加权 F1 分数更高。这表明，高级模型具备对类别分布的内在理解能力：对于较大的类别，零样本提示已能取得出色表现（体现为较高的加权 F1），而少样本提示对少数类的额外收益有限（体现为宏平均 F1 略有提升）。

总而言之，结果表明，少样本学习能够有效缓解传统模型中少数类和多数类的类别不平衡问题；但对于那些具备强大预训练知识、能够进行平衡预测的高级架构而言，少样本学习就不再那么必要。

6.2.2 有无领域知识增强的大语言模型在 F1 分数上的性能比较

继第 6.1.3 节基于准确率的分析之后，我们进一步分析这些准确率优化配置的宏平均 F1 和加权 F1 分数，以评估准确率的提升是否能转化为所有类别上的均衡性能。表 6 展示了宏平均 F1 和加权 F1 分数随大语言模型和样本量变化的比较结果。

表 6 表明，领域知识增强的效果因模型而异，其中通用模型（即 GPT-4o）与高级推理模型（即 O4-mini）之间的对比尤具启发性。从表中可以看出，GPT-4o 从领域增强中获益最为稳定，在大多数配置下，宏平均 F1 和加权 F1 均有所提升。例如，在样本量为 100 时，领域知识增强使宏平均 F1 提高了 21.0%（从 0.2020 增至 0.2444），使加权 F1 提高了 41.9%（从 0.2303 增至 0.3267）。值得注意的是，这些提升源于少样本学习与领域知识的相互作用；二者结合所带来的性能提升，是任一方法单独使用都无法实现的。这种协同效应表明，少样本示例提供特定任务模式，而领域知识则提供上下文理解；将二者结合起来，比单独采用任一方法都能实现更稳健的预测。

然而，对于高级推理模型 o4-mini，我们发现了一个有趣且相反的结果：领域知识增强基本无效，甚至会产生不利影响，尤其是在样本量较大时。其中一个原因是，具备复杂预训练知识的高级推理模型已经拥有足够的任务理解能力，而额外的领域信息可能会向模型固有的推理路径引入冲突信号或噪声。

模型	样本量	零样本			最佳少样本		
		准确率	宏平均 F1	加权 F1	准确率	宏平均 F1	加权 F1
GPT-4o	100	0.4000	0.2020	0.2303	0.4000	0.2020	0.2303
	200	0.4750	0.2354	0.3096	**0.5250**	**0.3746**	**0.3991**
	500	0.4400	0.2172	0.2828	0.4400	**0.2194**	**0.3109**
	1000	0.4500	0.2825	0.2991	**0.4700**	**0.3618**	**0.3506**
GPT-4o-mini	100	0.3500	0.1786	0.2071	**0.4500**	**0.4007**	**0.3561**
	200	0.4750	0.2354	0.3096	0.4750	**0.2589**	**0.3708**
	500	0.4000	0.1849	0.2555	**0.4300**	**0.2216**	**0.3480**
	1000	0.4350	0.2724	0.2890	**0.4550**	**0.3399**	**0.3220**
o3-mini	100	0.3500	0.1911	0.2107	**0.5000**	**0.4483**	**0.4333**
	200	0.4500	0.2250	0.3000	**0.4750**	**0.2708**	**0.3427**
	500	0.4400	0.2109	0.2974	**0.4800**	**0.2988**	**0.3626**
	1000	0.4500	0.3228	0.3029	**0.4900**	**0.4569**	**0.3920**
o4-mini	100	0.4500	0.2887	0.4058	**0.5000**	**0.2952**	**0.4552**
	200	0.5000	0.3308	0.4745	**0.5500**	**0.4579**	**0.5339**
	500	0.5600	0.4030	0.5261	**0.5800**	**0.4743**	**0.5341**
	1000	**0.5750**	0.4479	**0.5416**	0.5400	**0.4647**	0.4913

表 5：比　　较未进行领域知识增强时的零样本与最佳少样本性能

模型	样本量	无领域			有领域		
		准确率（示例）	宏平均 F1	加权 F1	准确率（示例）	宏平均 F1	加权 F1
GPT-4o	100	0.4000 (1)	0.2020	0.2303	**0.4500 (10)**	**0.2444**	**0.3267**
	200	0.5250 (1)	0.3746	0.3991	**0.5500 (5)**	**0.3803**	**0.4423**
	500	0.4400 (1)	0.2194	0.3109	**0.4500 (2)**	**0.2317**	**0.3437**
	1000	0.4700（2）	0.3618	0.3506	**0.4850（3）**	0.2469	**0.3909**
GPT-4o-mini	100	**0.4500（2/3）**	**0.4007**	**0.3561**	0.4000（1）	0.2203	0.2365
	200	0.4750（2）	**0.2589**	**0.3708**	0.4750 (5)	0.2175	0.3060
	500	0.4300 (5)	0.2216	0.3480	**0.4900 (5)**	**0.2586**	**0.4075**
	1000	0.4550（2/3/5）	**0.3399**	**0.3220**	**0.4600 (1)**	0.2342	0.3208
o3-mini	100	**0.5000 (5)**	**0.4483**	**0.4333**	0.4000 (1)	0.2203	0.2365
	200	0.4750（1/3/5/10）	0.2708	0.3427	**0.5000 (3)**	**0.3613**	0.3518
	500	0.4800 (2)	0.2988	0.3626	0.4800 (3)	**0.3214**	0.3448
	1000	**0.4900 (5)**	**0.4569**	**0.3920**	0.4750 (1)	0.2537	0.3178
o4-mini	100	**0.5000 (5)**	0.2952	**0.4552**	0.4500 (1)	**0.3278**	0.4445
	200	0.5500 (3)	0.4579	0.5339	**0.5750 (3)**	**0.4742**	0.5168
	500	**0.5800 (5)**	**0.4743**	**0.5341**	0.5500 (2)	0.3931	0.4934
	1000	**0.5400 (1)**	**0.4647**	**0.4913**	0.5250 (10)	0.3414	0.4384

表 6：领域知识增强对最佳少样本性能的影响

7 讨论

7.1 面向自然语言转换的特征工程

本研究提出了一种由特征重要性引导的转换方法，可在进行语言转换前，将 80 多个变量缩减至少于 15 个关键预测变量。降维解决了从高维数据集中生成连贯自然语言描述这一实际难题。通过让叙述聚焦于预测能力最强的变量，该方法构建出可解释描述，既能体现重要的行为关系，又便于语言模型理解。这种方法的应用范围或可超越出行方式预测，惠及其他需要通过自然语言界面处理高维结构化数据的领域。关键在于，作为一项重要预处理步骤，特征重要性分析有助于使叙述性描述聚焦于真正具有预测性的关系，而非用边缘信息淹没模型。

7.2 提示策略

实验结果表明，提示策略在出行方式预测任务中的有效性与模型架构和数据可用性相关。少样本学习范式通常能够提升预测性能，但提升幅度因模型架构而异。o3-mini 模型在少样本学习下的表现持续改善。相比之下，o4-mini 模型在零样本学习下的表现往往优于少样本学习。这种性能差异表明，具备先进推理能力的模型无需额外示例即可提取模式，而少样本示例可能造成干扰，而非提供引导。

除了提升预测准确率外，少样本学习范式在解决出行调查数据固有的类别不平衡问题方面也具有独特价值。在少样本设置下，模型的准确率与宏平均 F1 分数之间的差距显著缩小，这表明示例有助于模型关注少数类，而不仅仅是提高整体预测准确率。

领域知识增强也呈现出复杂的模式。GPT-4o 始终受益（提升幅度为 2.27% 至 12.50%），而 o4-mini 和 o3-mini 在领域知识增强下往往表现更差，尤其是在样本量较大时。这表明，能力更强的模型可能已通过预训练编码了交通相关关系，而明确的领域提示会干扰现有表征。这些发现表明，领域知识整合应采用自适应且考虑模型特性的方式。通用语言模型可能受益于领域知识增强，而先进的推理模型则可能在采用最简提示配置时表现最佳。

8 结论

本研究提出了 TransMode-LLM，这是一种新颖的框架，利用大语言模型的上下文推理能力，根据自然语言描述预测出行方式选择。该方法以出行调查数据为起点，开展特征重要性分析以识别影响最大的因素，随后将选定特征转化为描述性文本输入。我们系统地探索了（零样本和单样本/少样本）设置，并提出了一种领域增强提示策略，其中融入了标准化出行方式定义和三步结构化决策流程。（零样本和单样本/少样本）设置，并提出了一种领域增强提示策略，其中融入了标准化出行方式定义和三步结构化决策流程。我们利用多种架构（GPT-4o、GPT-4o-mini、o3-mini、o4-mini）和不同样本量，在美国国家家庭出行调查（NHTS）数据上进行了广泛实验。

实验结果与我们的系统性比较相对应，得出三项主要发现。首先，所提出的基于大语言模型的方法有潜力达到与成熟基线分类模型（LogitBoost 和梯度提升）相当的性能。这一发现表明，自然语言表征能够有效编码出行行为信息，并暗示交通运输应用中的预测范式正从基于训练转向基于提示。其次，单样本/少样本学习范式能够有效提升预测性能。通常仅使用 1–5 个示例，模型相较于零样本学习基线的准确率提升幅度介于 3.6% 至 42.9% 之间。第三，通用模型（GPT-4o、GPT-4o-mini）能够从领域知识增强提示中受益；而对于推理型模型（o3-mini、o4-mini），领域知识增强并不能普遍提升其性能。

鉴于出行调查数据固有的类别不平衡问题，我们构建了一个全面的评估框架，在准确率之外，还纳入了 F1-宏平均分数和加权F1分数。分析表明，少样本学习范式的主要价值在于应对类别不平衡，而不仅仅是提升总体准确率；该范式最多可将准确率与 F1-宏平均分数之间的差距缩小 71 %。

总之，这些发现兼具理论与实践意义。从理论上看，本研究将传统定量建模方法与新兴的基于大语言模型的出行方式预测技术相结合，表明语言模型的情境智能能够突破传统局限，推动出行方式预测的发展。从实践上看，这些发现凸显了大语言模型在交通研究中的双重潜力。大语言模型不仅是强大的预测工具，也是可解释的框架，能够深入揭示

出行决策的多面性。这也为交通规划从业者和政策制定者带来了显著优势，因为他们需要可解释的模型输出，以支持决策过程。

若干局限性指明了未来研究方向。首先，在交通领域数据上对语言模型进行领域特定预训练，或可得到更专门的模型，使其更能把握出行行为建模的细微特点。其次，本研究未考察不同人口统计群体之间的表现差异。未来研究应调查不同年龄、收入和家庭构成群体的预测准确率。第三，本研究使用的当前数据集来自美国，这是一个以汽车为主导的国家。为提高可推广性，未来研究应探索其他国家的数据，尤其是以公共交通为主导的国家。

参考文献

[1] Xilei Zhao、Xiang Yan、Alan Yu 和 Pascal Van Hentenryck。出行方式选择的预测与行为分析：机器学习与逻辑回归模型的比较。《*Travel Behaviour and Society*》，20：22–35，2020年。

[2] Mohammad Tamim Kashifi、Arshad Jamal、Mohammad Samim Kashefi、Meshal Almoshaogeh 和 Syed Masiur Rahman。采用可解释机器学习技术预测出行方式选择：一项比较研究。《*Travel Behaviour and Society*》，29：279–296，2022年。

[3] Shohreh Shaghaghian、Luna Yue Feng、Borna Ja- farpour 和 Nicolai Pogrebnyakov。为法律文档审查定制上下文化语言模型。载于*2020*年*IEEE*国际大数据会议（*Big Data*），第2139–2148页。IEEE，2020年。

[4] Gati V Aher、Rosa I Arriaga 和 Adam Tauman Kalai。使用大语言模型模拟多个人类并复现人类受试者研究。载于 国际机器学习会议，第337–371页。PMLR，2023年。

[5]Lisa P Argyle、Ethan C Busby、Nancy Fulda、Joshua R Gubler、Christopher Rytting 和 David Wingate。以一化众：使用语言模型模拟人类样本。《*Political Analysis*》，31(3):337–351，2023年。[6] Caleb Ziems、William Held、Omar Shaikh、Jiaao Chen、Zhehao Zhang 和 Diyi Yang。大语言模型能否改变计算社会科学？《*Computational Linguistics*》，50(1):237–291，2024年。[7] Baichuan Mo、Hanyong Xu、Dingyi Zhuang、Ruoyun Ma、Xiaotong Guo 和 Jinhua Zhao。用于出行行为预测的大语言模型。*arXiv* 预印本 *arXiv:2312.00819*，2023年。[8] Tianming Liu、Manzi Li 和 Yafeng Yin。大语言模型能否捕捉人类出行行为？出行方式选择的证据与见解。出行方式选择的证据与见解（*2024*年*8*月*26*日），2024年。[9] Yiming Xu 和 Junfeng Jiao。评估用于出行方式选择预测的大语言模型检索增强生成策略。*arXiv* 预印本 *arXiv:2508.17527*，2025。[10] Moshe E Ben-Akiva 和 Steven R Lerman。离散选择分析：理论及其在出行需求中的应用，第9卷。麻省理工学院出版社，1985年。[11] Julian Hagenauer 和 Marco Helbich。用于出行方式选择建模的机器学习分类器比较研究。《*Expert Systems with Applications*》，78:273–282，2017年。

[12] Fangru Wang 和 Catherine L Ross。机器学习出行方式选择：比较极端梯度提升模型与多项式逻辑回归模型的性能。《Transportation Research Record》，2672(47):35–45，2018年。

[13] Yixuan Ma 和 Zhenji Zhang。使用带有实体嵌入的深度神经网络进行出行方式选择预测。《IEEE Access》，8:64959–64970，2020年。

[14] Wayne XinZhao、Kun Zhou、Junyi Li、Tianyi Tang、Xiaolei Wang、Yupeng Hou、Yingqian Min、Be- ichen Zhang、Junjie Zhang、Zican Dong 等。《大语言模型综述》。*arXiv* 预印本 *arXiv:2303.18223*，1(2)，2023年。

[15] Tom Brown、Benjamin Mann、Nick Ryder、Melanie Subbiah、Jared D Kaplan、Prafulla Dhariwal、Arvind Neelakantan、Pranav Shyam、Girish Sastry、Amanda Askell 等。《语言模型是少样本学习者》。神经信息处理系统进展，33：1877–1901，2020年。

[16] Josh Achiam、Steven Adler、Sandhini Agarwal、Lama Ahmad、Ilge Akkaya、Florencia Leoni Aleman、Diogo Almeida、Janko Altenschmidt、Sam Altman、Shyamal Anadkat 等。GPT-4 技术报告。*arXiv* 预印本 *arXiv:2303.08774*，2023年。

[17]Hugo Touvron、Thibaut Lavril、Gautier Izacard、Xavier Martinet、Marie-Anne Lachaux、Timothée Lacroix、Baptiste Rozière、Naman Goyal、Eric Ham-bro、Faisal Azhar 等。《Llama：开放且高效的基础语言模型》。*arXiv* 预印本 *arXiv:2302.13971*，2023年。

[18]Stefan Hegselmann、Alejandro Buendia、Hunter Lang、Monica Agrawal、Xiaoyi Jiang 和 David Sontag。《TabLLM：利用大语言模型对表格数据进行少样本分类》。载于国际人工智能与统计学会议，第5549–5581页。PMLR，2023年。[19] Zak Hussain、Marcel Binz、Rui Mata 和 Dirk U Wulff。《开源大语言模型在行为科学中的教程》。《行为研究方法》，56(8):8214–8237，2024年。[20] Jiangping Zhou。《汽车主导型城市中的可持续通勤：影响替代出行方式的因素

选择：以大学生为例》。《交通研究A辑：政策与实践》，46(7):1013–1029，2012年。[21] Sebastian Hörl 和 Milos Balac。《基于开放且公开可用的数据构建巴黎和法兰西岛的人口合成数据与出行需求》。《交通研究C辑：新兴技术》，130:103291，2021年。[22] Susan Hanson 和 Perry Hanson。《城市居民的出行活动模式：维度及其与社会人口特征的关系》。《经济地理学》，57(4):332–347，1981年。[23] Xuedong Lu 和 Eric I Pas。《社会人口特征、活动参与和出行行为》。《交通研究A辑：政策与实践》，33(1):1–18，1999年。[24] Yang Hu、Bert van Wee 和 Dick Ettema。《家庭内部决策以及建成环境对活动—出行行为的影响：文献综述》。《交通地理学杂志》，106:103485，2023年。[25] Sumeeta Srinivasan 和 Joseph Ferreira。《家庭层面的出行行为：理解其与居住地选择的关联》。《交通研究D辑：交通与环境》，7(3):225–242，2002年。[26] Cynthia Chen、Hongmian Gong 和 Robert Paaswell。《建成环境在交通方式选择决策中的作用：密度影响的补充证据》。《交通运输》，35:285–299，2008年。[27] Reid Ewing 和 Robert Cervero。《出行与建成环境：元分析》。《美国规划协会杂志》，76(3):265–294，2010年。[28] Chuan Ding、Donggen Wang、Chao Liu、Yi Zhang 和 Jia wen Yang。《考虑汽车拥有量和出行距离的中介效应，探讨建成环境对出行方式选择的影响》。《交通研究A辑：政策与实践》，100:65–80，2017年。[29] Joonho Ko、Sugie Lee 和 Miree Byun。《探讨与通勤方式选择相关的因素：城市级综合社会调查数据的应用》。《交通政策》，75:36–46，2019年。[30] Daniel Baldwin Hess。《免费停车对通勤者交通方式选择的影响：来自出行日记数据的证据》。《Transportation Research Record》，1753(1):35–42，2001年。[31] Kevin Washbrook、Wolfgang Haider 和 Mark J accard。《通勤者交通方式选择估计：道路定价和停车收费影响的离散选择分析》。《交通运输》，33:621–639，2006年。[32] Donald Shoup。《免费停车的高昂代价》。劳特利奇出版社，纽约，2021年。[33] Natasa Vidovic 和 Jelena Simicevic。《停车定价对交通方式选择的影响》。《交通研究论文集》，69(3):297–304，2023年。[34] Felix Creutzig。燃油价格如何决定公共交通基础设施、出行方式份额和城市形态。城市气候，10:63–76，2014年。[35] Ashley R Haire 和 Randy B Machemehl。燃油价格上涨对美国公共交通客流量的影响。《运输研究记录》，1992(1):11–19，2007年。[36] Boris Jäggi、Alexander Erath、Christoph Dobler 和 Kay W Axhausen。利用多重离散—连续选择模型，将家庭车队选择建模为燃油价格的函数。《*Transportation Research Record*》，2302(1):174–183，2012年。[37]Valeriia Cherepanova、Roman Levin、Gowthami Somepalli、Jonas Geiping、C Bayan Bruss、Andrew G Wilson、Tom Goldstein 和 Micah Goldblum。面向表格深度学习中特征选择的性能驱动基准。神经信息处理系统进展，36:41956–41979，2023年。[38] Stacey Bricka、Timothy Reuscher、Paul Schroeder、Mitchell Fisher、Justina Beard、Xiaoyuan Layla Sun 等。出行趋势摘要：2022年全国住户出行调查。技术报告，美国联邦公路管理局政策办公室……，2024年。[39] ChatBox AI。Chatbox AI：您在任何设备上的终极人工智能副驾驶，未注明日期。[40] Jason Wei、Xuezhi Wang、Dale Schuurmans、Maarten Bosma、Fei Xia、Ed Chi、Quoc V Le、Denny Zhou 等。思维链提示能够激发大语言模型的推理能力。神经信息处理系统进展，35:24824–24837，2022年。[41]OpenAI。GPT-4o。https://platform.openai. com/docs/models/gpt-4o，2024年。访问日期：2025-04-22。[42]OpenAI。GPT-4o mini。https://platform.openai.com/docs/models/gpt-4o-mini，2024年。访问日期：2025-04-22。[43]OpenAI。o3-mini。https://platform.openai. co m/docs/models/o3-mini，2025年。访问日期：2025-04-22。[44]OpenAI。o4-mini。https://platform.openai. co m/docs/models/o4-mini，2025年。访问日期：2025-04-22。[45] 联邦公路管理局。2022年下一代全国住户出行调查核心数据。http: //nhts.ornl.gov，2022年。访问日期：2024-12-09。[46] S Wang、B Mo 和 J Zhao。使用86种机器学习分类器预测出行方式选择：一项实证基准研究。载于第*99*届交通研究委员会年会论文集，第279–296页，2020年。[47] Oona Rainio、Jarmo Teuho 和 Riku Klén。机器学习的评估指标与统计检验。科学报告，14(1):6086，2024年。[48] Zihao Zhao、Eric Wallace、Shi Feng、Dan Klein 和 Sameer Singh。使用前先校准：提升语言模型的少样本性能。载于

机器学习国际会议论文集，第12697–12706页。PMLR，2021。

[49] Changyue Wang、Weihang Su、Qingyao Ai、Yujia Zhou 和 Yiqun Liu。解耦推理与知识注入，实现上下文内知识编辑。载于 Wanxiang Che、Joyce Nabende、Ekaterina Shutova 和 Mohammad Taher Pilehvar（编），计算语言学协会*ACL 2025*年会论文集，第24543–24562页，奥地利维也纳，2025年7月。计算语言学协会。

[50] Chen Amiraz、Florin Cuconasu、Simone Filice 和 Zohar Karnin。干扰效应：理解检索增强生成中的无关段落。载于第*63*届*ACL*年会论文集，第18228–18258页，2025年。