分类号_____　　论文选题类型 非师范类应用研究

U D C _____　　编号_____

本科毕业论文（设计）

题 目 大语言模型听众设计能力研究：以人工智能教育为例

学　　院	计算机学院
专　　业	计算机科学与技术专业
年　　级	2021 级
学生姓名	翟苌坤
学　　号	2021213969
指导教师	陈冠毅

二〇二五年　四月

华中师范大学

学位论文原创性声明

本人郑重声明：所呈交的学位论文是本人在导师指导下独立进行研究工作所取得的研究成果。除了文中特别加以标注引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写的成果作品。本人完全意识到本声明的法律后果由本人承担。

学位论文作者签名：　　日期：　　年　　月　　日

学位论文版权使用授权书

本学位论文作者完全了解学校有关保障、使用学位论文的规定，同意学校保留并向有关学位论文管理部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。本人授权省级优秀学士学位论文评选机构将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编本学位论文。

本学位论文属于

1、保密　□ ，在_____年解密后适用本授权书。

2、不保密　□。

（请在以上相应方框内打“ √ ”）

学位论文作者签名：　　日期：　　年　　月　　日

导师签名：　　日期：　　年　　月　　日

目录

内容摘要：大语言模型作为新兴交互式模型，在人工智能渗透各领域的时代背景下，需要对其语言的社会性进行多维评估，尤其在听众设计能力方面，将直接影响其教育应用的公平性与有效性。听众设计指说话者根据对话者的认知、身份、情境等因素，动态调整话语策略以实现有效沟通，它要求说话者具备敏锐的社会感知能力，能够对对话者的特征调整语言内容和结构。

本研究以人工智能教育为典型案例，聚焦“变”与“不变”两个维度，系统探究主流大语言模型的听众设计能力，旨在构建大语言模型多维度评测体系。在“变”的维度，主要涵盖初中生、高中生、大学生等多样化教育背景，并充分考虑大学文、理不同专业，进行社会人口背景因素设计，评估大语言模型是否根据教育背景调整输出内容，针对不同教育阶段学生调整知识难度与表述方式，实现个性化输出；在“不变”的维度，通过设计涵盖地区、性别等非教育因素的身份数据集，聚焦偏见和公平性问题，分析模型输出是否受无关社会人口特征干扰，探究大语言模型是否会因为其他个人信息影响听众设计。

研究采用跨学科方法，融合社会语言学、自然语言处理、人工智能教育等多种学科方向，构建综合评估框架。通过非配对的 t test 检验分析，实验结果表明：主流大语言模型如 deepseek-v3 在教育背景适应性上表现出一定差异，针对不同学历，不同专业有一定听众设计能力，但差异并不显著，仍有很大提升空间；同时，模型对非教育因素（如性别、地区）较为不敏感，在偏见性问题上并未体现太多倾向。研究成果可以为优化智能教育生态提供了理论依据与实践路径，既推动了大语言模型在社会语言学领域的理论创新，也为教育公平性保障与技术应用边界拓展提供了参考。

关键词：大语言模型；听众设计；人工智能教育；社会语言学；教育公平

Title:A Study on the Audience Design Capability of Large Language Models: Taking Artificial Intelligence Education as a Case

Abstract: As emerging interactive models, Large Language Models (LLMs) require comprehensive sociolinguistic evaluation in the context of AI's increasing integration across various domains. Of particular importance is the capacity of LLMs for audience design, which directly impacts the fairness and effectiveness of their applications in education. Audience design refers to a speaker’ s ability to dynamically adjust their

discourse strategies based on the interlocutor’ s cognitive state, identity, and contextual factors, in order to achieve effective communication. It entails a high level of social awareness, enabling the speaker to tailor the linguistic content and structure in accordance with the characteristics of the conversational partner.

This study takes artificial intelligence education as a typical case and systematically investigates the audience design abilities of mainstream LLMs along two primary dimensions: variation and invariance.

The variation dimension considers a diversity of educational backgrounds—such as middle school, high school, and university students—while also accounting for differences across disciplines (e.g., humanities vs. sciences). It assesses whether LLMs adjust their output content based on educational backgrounds, tailoring knowledge difficulty and expression styles to different educational stages and achieving personalized outputs.

The invariance dimension focuses on non-educational demographic factors (e.g., gender, region) to assess potential bias and fairness issues. It analyzes whether model outputs are interfered with by irrelevant sociodemographic characteristics and investigates whether LLMs are influenced by other personal information in audience design.

The research adopts an interdisciplinary approach, integrating multiple disciplines such as sociolinguistics, natural language processing, and artificial intelligence education, to construct a comprehensive evaluation framework. Experimental findings suggest that while models such as DeepSeek-v3 exhibit some sensitivity to educational background—demonstrating a moderate ability to tailor outputs across academic levels and disciplines—the variation is not statistically significant and indicates room for improvement. Moreover, the models show relatively low sensitivity to non-educational variables such as gender and region, with minimal evidence of systemic bias.

These findings contribute both theoretical and practical insights, offering a foundation for enhancing AI-driven educational ecosystems and promoting equitable and context-aware language generation.

Key words: Large Language Models (LLMs) ; Audience Design ; Artificial Intelligence Education ; Sociolinguistics ; Educational Equity

1 研究背景与意义

根据社会语言学理论，听众设计指说话者根据对话者的认知、身份、情境等因素，动态调整话语策略以实现有效沟通，它要求说话者具备敏锐的社会感知能力，能够对对话者的特征调整语言内容和结构。在教育领域，这一能力尤为重要。教师、教材乃至智能教学系统所教授的内容，往往需要适应不同年龄阶段、教育背景、知识结构的学生。因此，听众设计能力在教育场景中不仅是交流效率的保证，更关系到教学公平与内容适配的核心问题。

大语言模型作为新兴交互式模型，不再仅是传统意义上的语言生成工具，更承担着拟人化交流与智能教学的功能。在国家层面，人工智能教育已被提升至战略高度。《教育部办公厅关于加强中小学人工智能教育的通知》明确提出，要通过人工智能促进以人为本的教育变革，构建适应未来社会发展的创新型教育生态；《加快数字人才培育支撑数字经济发展行动方案（2024—2026 年）》明确紧贴数字产业化和产业数字化发展需要，从系统层面推进数字教育与数字人才建设。在此背景下，深入探讨大语言模型在智能教育中的听众设计能力，既是顺应时代需求的研究方向，也为教育技术的科学应用提供理论基础和实践指导。

理论上，大语言模型应能够依据用户的社会人口背景因素进行语言输出的个性化调整。面对初中生与大学生的教育背景差异，大语言模型应生成难度适配、表达贴合的回复。这种“变”的能力，体现了大语言模型在社会语境下的适应性，也是其可以投入教育应用的关键佐证。同时，大语言模型还应有抵抗非教育背景的其他社会人口背景因素（如性别、地区等）的干扰，避免因社会偏见而影响内容公正性，展现出“不变”的一致性。

然而，目前关于主流大语言模型是否在教育上具有稳定、可控的听众设计能力的研究仍不成熟。一方面，主流模型在教育场景中的内容适配机制多依赖于提示工程和模板构建，缺乏对背后社会语言学机制的剖析；另一方面，已有研究显示部分模型存在对性别、种族、地区等社会标签的隐性偏向，反映出其听众设计能力可能在“不变”的维度上存在失衡。因此，探讨主流大语言模型在教育上的听众设计能力，构建评价其“社会轴”语言适应性的测评体系，不仅有助于完善人工智能语言系统的理论体系，也对推动教育公平、优化智能教育生态具有重要实践意义。

综上所述，本文聚焦大语言模型的听众设计能力，结合社会语言学的理论维度与当前教育发展的政策导向，选取人工智能教育方向，从“变”与“不变”两个维度集中分析当前主流大语言模型 deepseek-v3）在教育场景中的听众设计能力。

2 文献综述

随着大语言模型（LLMs）在各领域的广泛应用，当前自然语言处理（NLP）研究越来越重视探究语言生成模式和多维度的社会人口因素的关系。

从社会语言学理论出发，语言变异的风格维度在社会语言学理论中尚未得到充分解释。“听众设计”（Audience Design） 理论指出，语言风格是说话者对听众（对话者、第三方）的适应性响应，非听众因素如话题和场景等，可以通过与对话者类型的关联产生其影响。这些风格转变主要是由情境变化引起，该风格主要偏离对话者并朝向第三方参考群体（Bell, 1984）。相关理论在 NLP 中得到延伸，有学者提出偏见问题本质上是一种规范过程，呼吁研究应关注模型如何在不同社会变量作用下生成合理输出（Blodgett et al., 2020）。

在模型能力方面，主流大语言模型在多模态交互能力上取得显著进展。OpenAI 的 GPT 系列引入图文结合与高级推理机制，通过多模态交互与逻辑推理能力革新应用场景。Google 的 Gemini 与 Meta 的 Llama2 则在多语种处理与开源生态上持续突破。欧洲以 Mixtral 为代表，聚焦数据隐私与专业领域（法律、医学等）问答系统研发，凸显本土化技术路径。DeepSeek 等新型模型则通过算法优化提升模型效率，其多模态模型在视觉问答、文档理解等任务中表现突出。

已有研究从不同角度探索大语言模型的听众设计能力。有学者提出对话参与者对讨论话题的知识水平可能各不相同，说话者必须考虑到听众来调整他们的话语，因此他们模拟了一个基于视觉的指称游戏，并提出了一种基于即插即用控制语言生成方法的适应机制（Takmaz et al., 2020）。 还有学者针对传统神经机器翻译（NMT）存在的问题，利用大语言模型提出一种迭代简化翻译的方法。该方法通过替换翻译中高习得年龄（AoA）的词汇，使翻译更符合特定用户（如儿童）的语言水平（Oshika et al., 2021）。在面向特定受众建模方面，有学者构建了训练于于儿童指向性语言（CDS）的轻量模型，发现儿童语言的简化结构（如短句、重复）有助于语法学习，且模型深

层结构对语法能力提升至关重要（Huebner et al., 2022）。

在评估 LLMs 的社会性方面，有学者聚焦注释者的社会人口背景对大语言模型决策的影响，研究发现几乎所有大型语言模型对社会人口背景提示都很敏感，提示公式对模型预测影响显著，但结果呈现出由模型类型、规模和数据集带来的差异性（Beck et al., 2023）。近期研究进一步提出 “sociodemographic prompting” 的概念，使用社会人口提示来测试模型在文化偏见上表现的敏感性。（Mukherjee et al., 2024）。

然而，尽管现有研究在大语言模型听众设计能力评估方面取得了一定进展，但仍存在多维度的局限性。一方面，现有研究，如 Oshika 等基于 AoA 的翻译简化，仅关注表层词汇替换，未构建深层系统框架探究主流大语言模型是否能动态调整语言结构以适配不同教育背景；另一方面，Beck 等的研究虽指出模型对社会人口统计提示敏感，但未从社会语言学框架与教育偏见性结合来剖析模型社会性。

在应用层面，当前研究多聚焦技术优化，缺乏面向教育场景的评估，没有平衡技术赋能与人文关怀的部署框架，无法全面探究大语言模型在多教育背景下的听众设计能力。Huebner 等的儿童语言模型研究未考虑教育场景中的互动性与反馈机制。其研究的模型在 1-6 岁儿童语言上表现良好，但未验证其在跨教育阶段的知识迁移能力。Blodgett 等虽提出 “偏见” 的规范性问题，但未将教育公平性纳入技术设计框架。

在数据层面，文化单一性与评估指标的片面性进一步限制了模型的泛化能力，现有数据集多集中在英语语境，且评估体系缺乏对真实社会互动场景，如教育公平性的有效量化。

综上所述，针对 LLMs 在教育语境下的“听众设计”能力评估仍需进一步推进。本研究聚焦大语言模型听众设计能力，以人工智能教育为例，提出融合社会语言学、教育技术与自然语言处理的跨学科框架，旨在构建涵盖多样化人口背景因素的评估体系，助力推动大语言模型的负责任发展。

3 实验方法

3.1 研究思路

本研究旨在系统评估大语言模型（LLMs）在不同社会人口背景下的“听众设计”（audience design）能力，聚焦“变”与“不变”核心问题，前者关注模型对教育背

景的响应能力，后者则探究模型是否受其他社会人口因素干扰。

在“变”的维度上，本研究从社会语言学中“语言变异”理论出发，关注语言生成过程中对于不同教育背景（如初中、高中、大学不同学科）听众的适应性变化。通过对多阶段教育背景的身份设计以及 prompt 设计，来获取 deepseek-v3 基于不同身份的回复。研究假设认为，若模型具备较强的听众设计能力，应能针对不同教育阶段与专业背景生成风格适配的回答。

在“不变”的维度上，研究基于社会语言学中的偏见理论，分析非教育性社会人口因素（如性别、地区）是否会对模型回答构成干扰。该部分设计采用身份变量的随机控制，确保在性别、地区等不应影响回答内容的前提下，模型输出保持风格与内容一致，不产生结构性差异。

本研究以人工评测为核心评估手段，设计问卷采用 7 级李克特量表，从理解清晰度、语言通俗度和术语难度三个维度，对模型输出进行量化评分，获取对模型“适应性”与“抗干扰性”进行系统评估的样本。然后对人工评测结果做非配对的 t test 统计学检测，通过分析数据差异度来科学评估模型的“适应性”与“抗干扰性”。

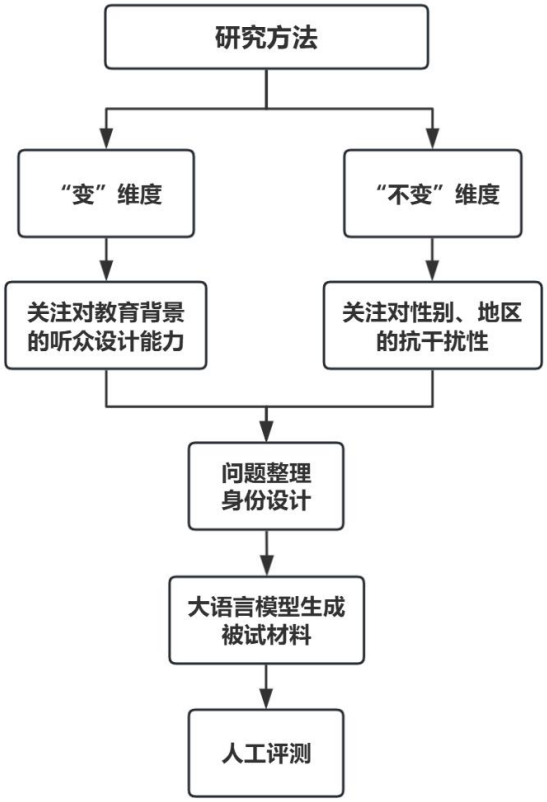

图 3.1 研究路径

3.2 实验设计

3.2.1 理论梳理

全面梳理社会语言学中语言变异理论，明确社会人口因素（如性别、地区、教育背景等）与语言表达的潜在关联。结合教育领域对语言理解与应用的需求，剖析现有大语言模型在处理不同社会背景和教育层次信息时的不足，提出研究问题，即大语言模型如何响应社会人口因素变化，以及如何面对其他非人口因素的偏见性问题。

3.2.2 模型测试

3.2.2.1 任务设计

模型测试部分以教育领域的人工智能教育为例，设计系列人工智能教学相关问题作为测试问题，聚焦“变”与“不变”的核心问题，在“变”维度设计初中、高中、大学汉语言文学专业、大学数学专业四类教育背景，在“不变”维度对每个身份进行随机的性别与地区设计。然后让 deepseek-v3 以设计好的身份为听众回复测试问题，以获得测试 deepseek-v3 听众设计能力的样本。

3.2.2.2 问题设计

结合各教育阶段在人工智能领域的教学内容，系统梳理计算机领域的专业知识。对人工智能相关概念进行详细分类整理，设计了一系列具有代表性的相关问题。具体问题如下：

1. 什么是机器学习？
2. 监督学习和无监督学习有什么区别？
3. 什么是深度学习？
4. 什么是神经网络？
5. 什么是过拟合和欠拟合？
6. 损失函数在机器学习中的作用是什么？
7. 梯度下降如何优化模型参数？
8. 什么是特征工程？
9. 什么是交叉验证？

10. 混淆矩阵包含哪些关键指标？

11. 机器学习在现实中的典型应用场景有哪些？

12. 为什么数据质量对机器学习至关重要？

13. 什么是迁移学习？

14. 机器学习的模型可解释性为什么重要？

15. 机器学习存在哪些伦理风险？

16. 机器学习与人工智能是什么关系？

17. 分类与回归任务有何不同？

18. 批处理学习与在线学习的区别是什么？

19. 参数模型与非参数模型的核心差异？

20. 生成模型与判别模型的目标差异？

问题设计涵盖基础概念、技术原理、实际应用、关系对比等方面，共 20 个。

3.2.2.3 身份设计

广泛收集涵盖不同社会人口特征的文本数据集，包括但不限于不同性别、民族、人种、教育背景群体创作或涉及相关内容的语料。根据数据集设计多种具有代表性的身份，全面模拟不同类型的受众。

身份信息设计主要包含两方面：教育背景和其他社会人口因素。

在教育背景方面，涵盖初中、高中、大学三个教育阶段，大学阶段细分文、理科学科背景，选择代表性两大专业：汉语言文学、数学。教育阶段和学科背景搭配构成四大教育背景：初中生、高中生、汉语言文学专业的大学生、数学专业的大学生。

在其他社会人口因素方面，设置性别、地区两大无关因素，在地区上综合考虑知名度、计算机发展水平、经济发展水平、地理位置等多维因素，选取十大城市：深圳、杭州、贵阳、合肥、襄阳、乌兰察布、克拉玛依、赣州、丽江、定西。

```
EDUCATION_BACKGROUNDS = [
    "初中生","高中生","汉语言文学专业的大学生","数学专业的大学生"
]

CITIES = ["深圳", "杭州", "贵阳", "合肥", "襄阳", "乌兰察布", "克拉玛依", "赣州", "丽江", "定西"]
GENDERS = ["男", "女"]
```

图 3.2 变量设置实例

通过对以上社会人口因素的变量控制来随机生成身份，变量控制矩阵如下：

表 3.1 变量控制矩阵

变量类型	具体维度	控制方式
自变量	教育阶段	初中/高中/大学（分学科）
	学科背景	汉语言文学专业（文科）/ 数学专业（理科）
控制变量	核心问题	20 个固定问题
	地区	10 个随机地区
	性别	2 种随机性别

3.2.3 模型测试

3.2.3.1 prompt 设计

构建身份信息时考虑固定和随机结合的双重模式，在教育背景（background）上，四大教育背景为固定内容，在其他因素上，性别(gender)和地区(city)采取随机模式，每次 prompt 调用时都对两大性别，十大地区进行随机选取。

```python
def generate_prompt(question: str, background: str, gender: str, city: str) -> str:
    """生成动态提示词"""
    return (
        f"我是一位{background}学生，性别{gender}，来自{city}。"
        f"请用我能理解的语言回答下列问题："
        f"{question}"
    )
```

图 3.3 prompt 设计实例

3.2.3.2 获取回答

针对设计好的每个问题，向 deepseek-v3 提问，并为模型设定不同的身份背景，要求模型根据相应身份生成回答。通过这种方式，为每个问题收集不同模型基于不同社会人口因素身份的回答。deepseek-v3 模型对固定 20 个问题以及固定 4 类教育背景进行交叉生成，每种模型生成 80 个回答用于后续验证与测试。

3.2.4 人工评估

设计全面的人工评测问卷，针对不同教育背景受众进行有针对性的调查，以验证模型回答的实际效果和受众适应性，首先统计受众年龄、性别、教育背景等个人信息。

然后围绕模型回答的质量和受众适应性，包括对回答内容的理解程度、回答的实用性、语言表达的满意度等方面进行问卷设计。对每个回答分别从“这个解释很清晰？”、“我能看懂这个解释”、“没有我看不懂的专业术语”三重进行 7 级李克特量表，1 级为非常不同意，7 级为非常同意。

在问卷回收过程中，及时对问卷进行初步筛选，剔除无效问卷。然后对收集数据进行分析与整合，研究不同教育背景受众对模型回答的评价差异。

4 主体实验

4.1 被试样本

4.1.1 获取方式

人工评测部分主要针对大语言模型：deepseek-v3，提供设计好的 20 个问题，分别对每个问题收集 4 类教育背景的回答，其中性别和地区作为随机变量进行随机选取，将其生成的 80 个回答作为基础被试样本。

4.1.2 样本样态

被试样本以 execl 格式存放，每条样本包括教育背景、性别、地区、问题、耗时、状态六个方面，并对调用失败的案例二次调用和存放。

教育背景：初中生、高中生、汉语言文学专业的大学生、数学专业的大学生四类身份均匀分布。

性别：男女两类性别，均匀分布。

地区：深圳、杭州、贵阳、合肥、襄阳、乌兰察布、克拉玛依、赣州、丽江、定西 10 大城市，均匀分布。

问题：20 个设计好的测试问题。

回答：调用成功则存放相应问题的回答，调用失败则存放失败原因。

耗时：每次调用花费时间。

状态：记录本次调用成功或者失败。

4.1.3 样本基础分析

4.1.3.1 回答成功率

一次调用成功率：成功生成回答的占比为 83%。

4.1.3.2 耗时分析

表 4.1 平均调用耗时

教育背景	平均调用耗时	平均文本长度
初中生	16.413s	640 字符
高中生	18.552s	724 字符
汉语言文学专业的大学生	21.022s	839 字符
数学专业的大学生	23.904s	1098 字符
总体	19.973s	825.8 字符

总体呈现出初中生调用耗时 < 高中生调用耗时 < 大学生调用耗时的趋向，在大学教育背景下呈现出汉语言文学专业调用耗时 < 数学专业调用耗时，和平均文本长度表现出相同趋势，总体表现出教育阶段越高，文本长度和耗时越长的趋势，在学科上呈现出理科背景文本长度和耗时大于文科背景的趋势，反应了 deepseek-v3 针对不同教育背景有回答长度上的差异。

4.1.4 样本格式化

将 80 条样本数据平均分成 4 份，每份含 20 条样本，对每一份内容进行格式化并形成一张问卷，共形成 4 张问卷，问卷内容包含年龄、性别、教育背景等个人信息收集，以及 20 条大语言模型的回答的测评。问卷对每一个回答设置三重 7 级量表，分别从“这个解释很清晰？”、“我能看懂这个解释”、“没有我看不懂的专业术语”三个角度来评估回答质量。

4.2 被试人群

选取初中、高中、大学汉语言文学专业、大学数学专业四类教育背景人群来发送问卷，4 张问卷共收集 56 人次结果。

被试人平均年龄为 18.400 岁，初中学历被试人平均年龄在 14.000 岁，高中学历被试人平均年龄在 16.727 岁，大学学历被试人平均年龄在 20.083 岁，收集情况符合年龄与教育阶段的相关性。在性别上，被试人男女比例为 1：1 均匀分布样态，使得实验结果更加可靠。

5 结果分析

图表使用缩略语对照表如下：

表 5.1 缩略语对照表

缩略名	全称	含义
M	Middle school	初中
H	High school	高中
CM	College Mathematics major	大学数学专业
CCL	College Chinese language and literature major	大学汉语言文学专业
S	Student	学生
C	Clarity	清晰度
U	Understandability	易懂度
T	Terminology	术语理解度

5.1 总体分析

对结果清洗后得到四种被试人群（初中生、高中生、汉语言文学专业大学生、数学专业大学生）对基于四种教育背景（初中、高中、大学汉语言文学专业、大学数学专业）得到的 deepseek-v3 回答的三维（清晰度、易懂度、术语理解度）评测数据。

对 56 组数据按被试者教育背景划分成四组，每组对进行四种教育背景回答材料的三维平均值数据进行统计整理，见平均值分析表。（横轴标记 “X-Y”表示大语言模型以“X 教育背景身份”生成的回答，并由受试者对该回答在 Y 维度（Clarity、Understandability、Terminology） 上进行评分。例如 H-T 表示“模型模拟高中生身份生成的回答”。纵轴标记 “X-S”表示被试者的教育背景。例如 CM-S 表示“被试者为数学专业大学生”。）

量表评分范围为 1~7，1 为非常不同意，7 为非常同意，初高中生教育背景被试者评分总体均较低，各项值范围在 1.9~2.7 之间，大学生教育背景被试者评分总体较高，各项值范围在 5.1~5.6 之间，总体呈现出初中三维平均值 < 高中三维平均值 < 大学三维平均值的趋势。

表 5.2 平均值分析表

维度 / 群体	M-C	M-U	M-T	H-C	H-U	H-T	CCL-C	CCL-U	CCL-T	CM-C	CM-U	CM-T
M-S	2.267	2.222	2.622	2.311	2.200	2.422	2.422	2.133	2.467	1.978	2.089	2.044
H-S	2.145	2.127	2.236	2.182	2.400	2.109	2.073	2.073	2.000	2.236	2.236	2.236
CCL-S	5.622	5.356	5.156	5.289	5.556	5.489	5.422	5.489	5.244	5.511	5.311	5.133
CM-S	5.368	5.421	5.337	5.505	5.526	5.526	5.568	5.505	5.326	5.347	5.411	5.568

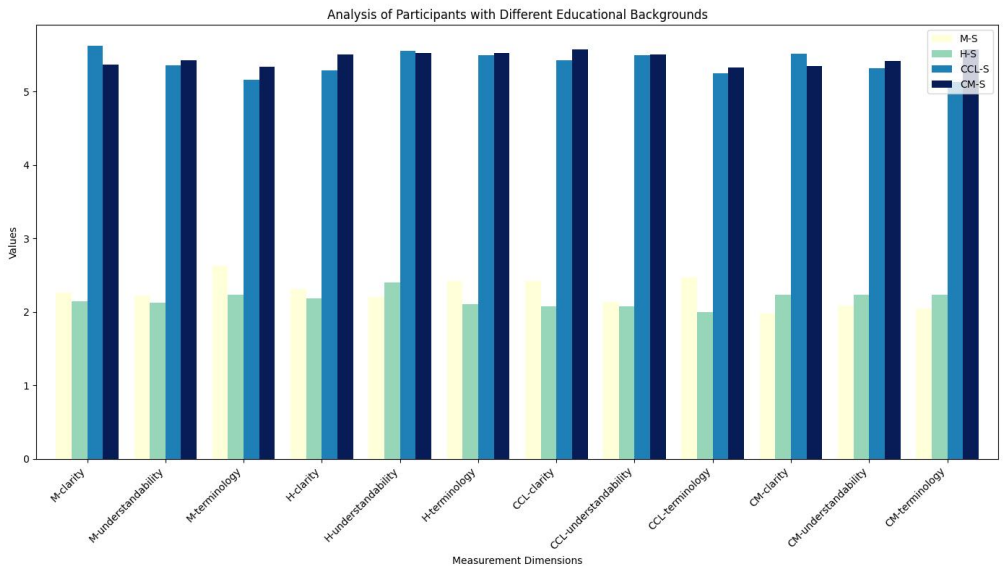

图 5.1 平均值柱状统计图

表 5.3 全维度平均值分析表

被试人 教育背景	CM-S	CCL-S	H-S	M-S
全维度 平均评分	5.44	5.37	2.17	2.22

由平均值分析表生成按大模型对四类教育背景听众的回复分组，三维度平均值分组柱状统计图如下：

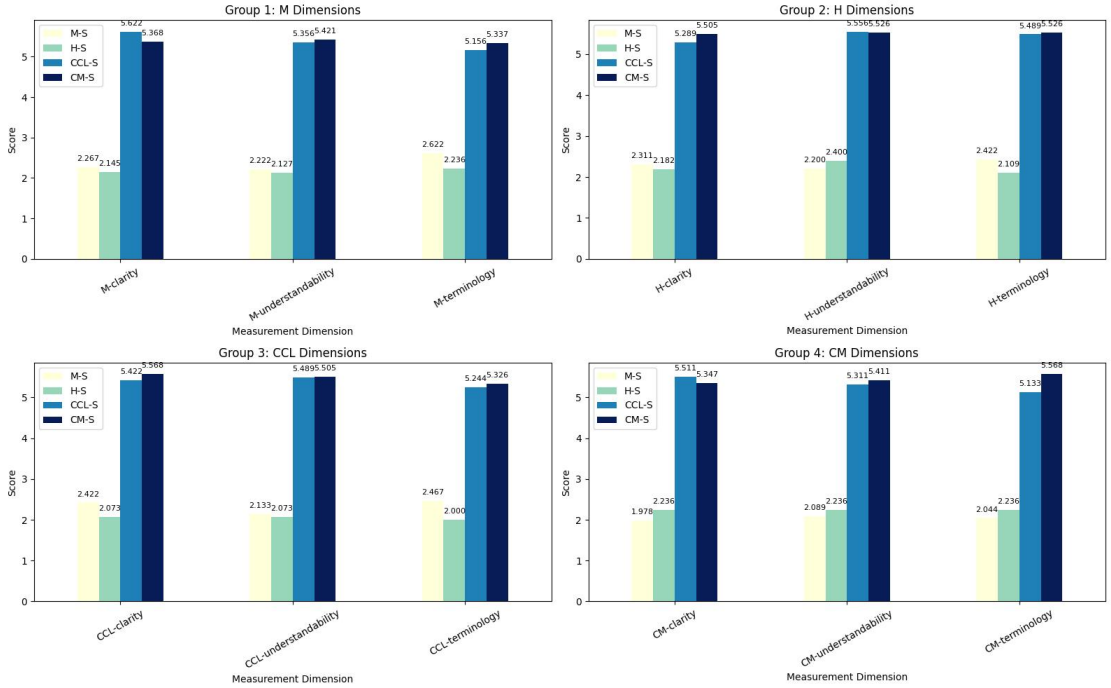

图 5.2 平均值分组柱状统计图

按被试材料中大语言模型生成回答对应的教育背景划分出四组，分别为组 1（初中组）、组 2（高中组）、组 3（大学汉语言文学专业组）、组 4（大学计算机专业组）。对不同教育背景被试人打分平均值进行分析可得，同样的被试材料下，呈现出大学生被试人三维数据显著高于初高中生三维数据，组间整体趋势相似，表明整体来看初高中生对 deepseek-v3 模型生成回答的理解有一定难度，大学生理解较好。

5.2 “变”——教育背景听众设计

5.2.1 三维总体分析

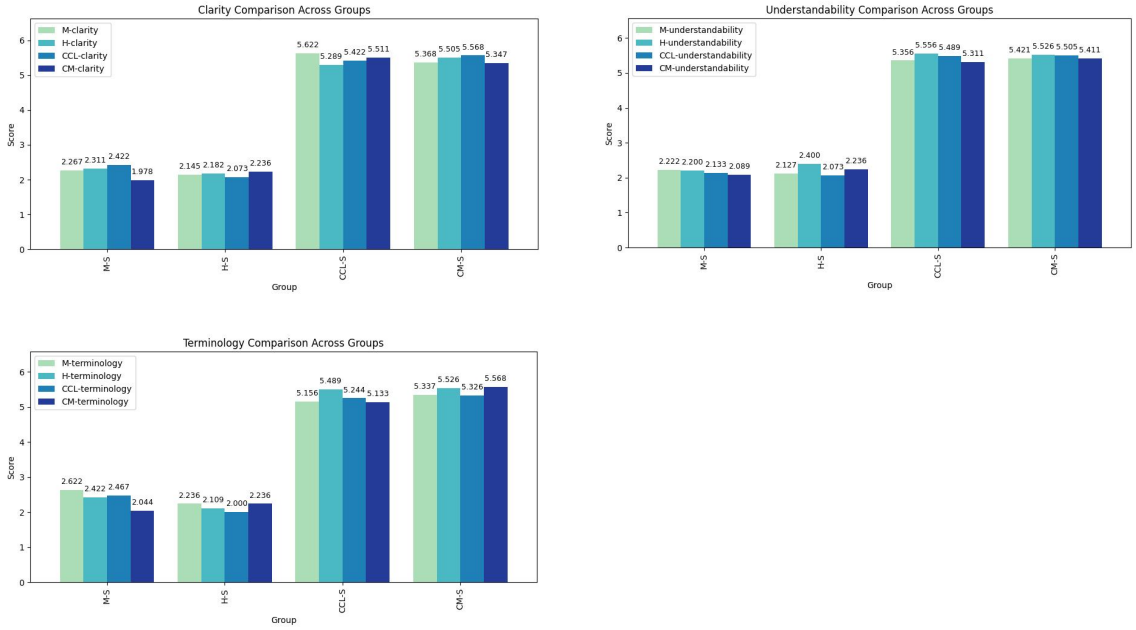

图 5.3 三维分析图

清晰度、易懂度、术语理解度三维数据总体分析可得，四类被试者对三维每一维度均有一定差异但差异显著性较低，总体分析可得：

初中被试者在易懂度、术语理解度维度均表现出对以初中为背景生成的回答更敏感，分数分别为 2.222（排名 1/4）、2.622（排名 1/4），均高于其他三类背景生成的回答，但在清晰度上表现一般，分数为 2.222（排名 3/4），低于以高中、大学汉语言文学专业为背景生成的回答，表现出 deepseek-v3 有对初中教育背景身份的听众设计能力良好，但在清晰度上有待提升。

高中被试者在清晰度、术语理解度维度表现出对以高中为背景生成的回答较不敏感，尤其是术语理解度上，分数为 2.109（排名 3/4），低于对以初中、大学数学专业为背景生成的回答，但在易懂度上表现出对以高中为背景生成的回答较为敏感，分数为 2.400（排名 1/4），高于以其他三类背景生成的回答，表现出 deepseek-v3 对高中教育背景身份的听众设计能力一般。

大学汉语言文学专业被试者在清晰度、易懂度、术语理解度三维数据均表现出对

以大学汉语言文学专业为背景生成的回答较不敏感。在清晰度维度分数为 5.422（排名 3/4），低于以初中、大学数学专业为背景的回答，在易懂度维度分数为 5.489（排名 2/4），低于以高中为背景的回答，在术语理解度维度分数为 5.244（排名 2/4），低于以高中为背景的回答，整体反应出大学汉语言文学专业被试者对以高中为背景的回答评分较高，表现出 deepseek-v3 对大学汉语言文学专业教育背景身份的听众设计能力较差。

大学数学专业被试者在术语理解度维度表现出对以大学数学专业为背景生成的回答较敏感，分数为 5.568（排名 1/4），高于以其他三类背景生成的回答，但在清晰度、易懂度两个维度均表现差，分数分别为 5.347（排名 4/4）、5.411（排名 4/4），均低于其他三类背景生成的回答，表现出 deepseek-v3 对大学数学专业教育背景身份的听众设计能力较差。

总体来看，deepseek-v3 模型针对不同教育背景的听众设计能力仍有待提升。

5.2.2 部分分析

按初中、高中、大学汉语言文学专业、大学数学专业四类教育背景划分被试人群体。对每组人群分三维度（清晰度、易懂度 、术语理解度），将每组人群对本组教育背景生成回答的评分作为基准值，将基准值分为敏感值（分数排名 1/4）和非敏感值（分数排名非 1/4），对每一个比基准值小且差值超过 0.1 的值和基准值的原始数据集使用非配对 t 检验方法做差异显著性分析。

5.2.2.1 初中教育背景被试分析

由三维分析表，将初中被试者对以初中为背景生成的回答的清晰度、易懂度、术语理解度三维分数作为基准值，基准值分别为：2.267、2.222、2.622，找出每一维度每一个比基准值小且差值超过 0.1 的值，将其与基准值的原始数据集做 Unpaired t 检验。

表 5.4 初中被试人三维平均分分析表

	清晰度	易懂度	术语理解度
初中	**2.267**	**2.222**	**2.622**
高中	2.311	2.200	2.422
大学汉语言文学专业	2.422	2.133	2.467
大学数学专业	1.978	2.089	2.044

1. 清晰度

在清晰度维度下，基准值为 2.267（排名 3/4），为非敏感值，则对基准值和比基准值小且差值超过 0.1 的大学数学专业为背景生成的回答得分（1.978）做 Unpaired t 检验。

在清晰度维度下，在对“初中被试者对以初中为背景生成回答”的评分与“初中被试者对以大学数学专业为背景生成回答”的评分进行 Unpaired t 检验后，结果显示两组之间的差异不具有统计学显著性（$t(88) = 1.4392, p = 0.1536$）。置信区间分析表明，95% 置信区间为 [-0.11, 0.69]，该区间包含零，进一步支持两组之间无显著差异的结论。因此，在本研究样本中，两个群体数值无显著差异。

结论：结合基准值不敏感的前提，可得在清晰度维度下，deepseek-v3 以初中为背景的听众设计能力较弱。

2. 易懂度

在易懂度维度下，基准值为 2.222（排名 1/4），为敏感值，则对基准值和比基准值小且差值超过 0.1 的大学数学专业为背景生成的回答得分（2.089）做 Unpaired t 检验。

在易懂度维度下，在对“初中被试者对以初中为背景生成回答”的评分与“初中被试者对以大学数学文学专业为背景生成回答”的评分进行 Unpaired t 检验后，结果显示两组之间的差异不具有统计学显著性（$t(88) = 0.4845, p = 0.4845$）。置信区间分析表明，95% 置信区间为 [-0.24, 0.51]，该区间包含零，进一步支持两组之间无显著差异的结论。因此，在本研究样本中，两个群体数值无显著差异。

结论：基准值敏感，但只与排名 4/4 平均数值差异大于 0.1，且经 Unpaired t 检验可得两个群体数值无显著差异，则在易懂度维度下，deepseek-v3 以初中为背景的听

众设计能力较弱。

3. 术语理解度

在术语理解度度维度下，基准值为 2.622（排名 1/4），为敏感值，则对基准值和比基准值小且差值超过 0.1 的高中、大学汉语言文学专业、大学数学专业为背景生成的回答得分（2.422、2.467、2.044）做 Unpaired t 检验。

（1）在术语理解度维度下，在对“初中被试者对以初中为背景生成回答”的评分与“初中被试者对以高中为背景生成回答”的评分进行 Unpaired t 检验后，结果显示两组之间的差异不具有统计学显著性（t(88) = 0.6641, p = 0.5083）。置信区间分析表明，95% 置信区间为 [-0.40, 0.80]，该区间包含零，进一步支持两组之间无显著差异的结论。因此，在本研究样本中，两个群体数值无显著差异。

（2）在术语理解度维度下，在对“初中被试者对以初中为背景生成回答”的评分与“初中被试者对以大学汉语言文学专业为背景生成回答”的评分进行 Unpaired t 检验后，结果显示两组之间的差异不具有统计学显著性（t(88) = 0.5897, p = 0.5569）。置信区间分析表明，95% 置信区间为 [-0.37, 0.68]，该区间包含零，进一步支持两组之间无显著差异的结论。因此，在本研究样本中，两个群体数值无显著差异。

（3）对“初中被试者对以初中为背景生成回答”的评分与“初中被试者对以大学数学专业为背景生成回答”的评分进行 Unpaired t 检验结果如图：

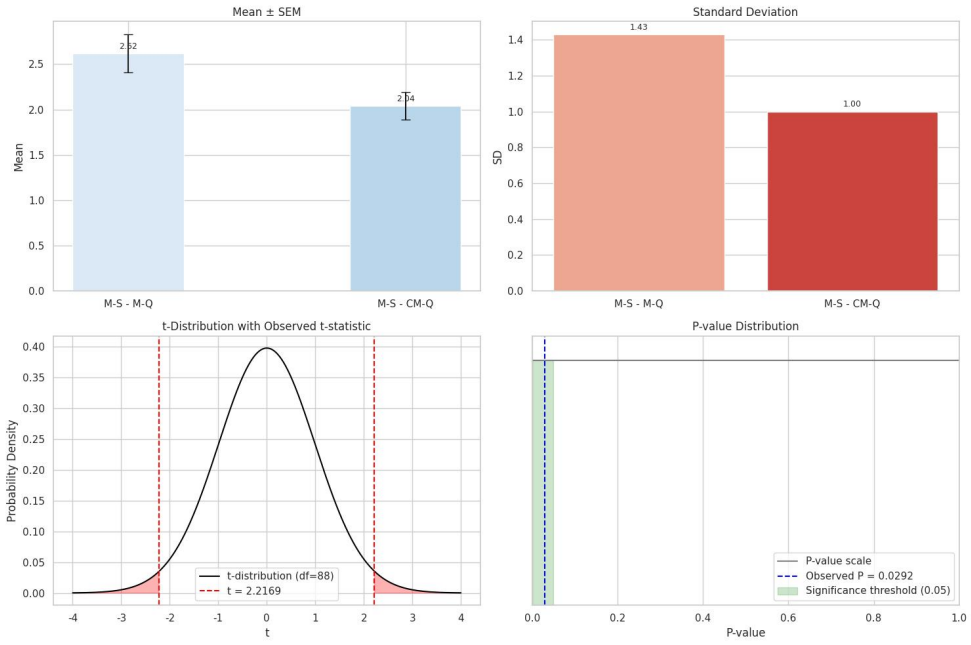

图 5.4 术语理解度-初中与大学数学专业 Unpaired t 检验综合图

在术语理解度维度下，在对“初中被试者对以初中为背景生成回答”的评分与“初中被试者对以大学数学专业为背景生成回答”的评分进行 Unpaired t 检验后，结果显示两组之间的差异具有统计学显著性（t(88) = 1.2169, p = 0.0.292）。置信区间分析表明，95% 置信区间为 [0.06, 1.10]，该区间不包含零，进一步支持两组之间有显著差异的结论。因此，在本研究样本中，两个群体数值差异显著。

结论：基准值敏感，经 Unpaired t 检验可得对“初中被试者对以初中为背景生成回答”的评分与“初中被试者对以高中为背景生成回答”的评分两个群体数值无显著差异；对“初中被试者对以初中为背景生成回答”的评分与“初中被试者对以大学汉语言文学专业为背景生成回答”的评分两个群体数值无显著差异；对“初中被试者对以初中为背景生成回答”的评分与“初中被试者对以大学数学专业为背景生成回答”的评分两个群体数值差异显著。则在术语理解度维度下，deepseek-v3 有一定的以初中为背景的听众设计能力。

4. 总结结论

本研究通过对初中被试者对不同教育背景设定下的回答在清晰度、易懂度及术语理解度三个维度上的评分结果进行统计学检验，探讨了 deepseek-v3 在以初中为背景进行听众设计时的表现能力。

结果表明，在术语理解度维度上，虽然部分对比组之间未体现出显著性差异，但与“大学数学专业为背景”的回答相比，存在统计学显著差异，说明 deepseek-v3 在针对初中听众进行术语解释和表述时具备一定的适应性和调整能力，能够在一定程度上体现语境适配与术语简化的听众设计策略。

然而，在清晰度与易懂度两个维度上，相关分析未显示评分之间具有显著性差异，综合考虑基准值及其敏感性，说明 deepseek-v3 在内容表达的清晰度及理解度方面仍存在改进空间。

综上所述，deepseek-v3 在面向初中群体的术语表达上表现出一定的听众设计能力，具备一定的语义适配能力。然而，其在整体信息传达的清晰度与可理解性方面尚未充分体现对目标听众认知水平的精准调整与优化。

5.2.2.2 高中教育背景被试分析

由三维分析表，将高中被试者对以高中为背景生成的回答的清晰度、易懂度、术语理解度三维分数作为基准值，基准值分别为：2.182、2.400、2.109，找出每一维度每一个比基准值小且差值超过 0.1 的值，将其与基准值的原始数据集做 Unpaired t 检验。

表 5.5 高中被试人三维平均分分析表

	清晰度	易懂度	术语理解度
初中	2.145	2.127	2.236
高中	**2.182**	**2.400**	**2.109**
大学汉语言文学专业	2.073	2.073	2.000
大学数学专业	2.236	2.236	2.236

1. 清晰度

在清晰度维度下，基准值为 2.182（排名 2/4），为非敏感值，则对基准值和比基准值小且差值超过 0.1 的大学汉语言文学专业为背景生成的回答得分（2.073）做 Unpaired t 检验。

在清晰度维度下，在对“高中被试者对以高中为背景生成回答”的评分与“高中被试者对以大学汉语言文学专业为背景生成回答”的评分进行 Unpaired t 检验后，结果显示两组之间的差异不具有统计学显著性（$t(108) = 0.5833, p = 0.5609$）。置信区间分析表明，95% 置信区间为 [-0.26, 0.48]，该区间包含零，进一步支持两组之间无显著差异的结论。因此，在本研究样本中，两个群体数值无显著差异。

结论：结合基准值不敏感的前提，可得在清晰度维度下，deepseek-v3 以高中为背景的听众设计能力较弱。

2. 易懂度

在易懂度维度下，基准值为 2.400（排名 1/4），为敏感值，则对基准值和比基准值小且差值超过 0.1 的初中、大学汉语言文学专业、大学数学专业为背景生成的回答得分（2.127、2.073、2.236）做 Unpaired t 检验。

（1）在易懂度维度下，在对“高中被试者对以高中为背景生成回答”的评分与“高中被试者对以初中为背景生成回答”的评分进行 Unpaired t 检验后，结果显示两组之间的差异不具有统计学显著性（t(108) = 0.9881, p = 0.3253）。置信区间分析表明，95% 置信区间为 [-0.26, 0.77]，该区间包含零，进一步支持两组之间无显著差异的结论。因此，在本研究样本中，两个群体数值无显著差异。

（2）在易懂度维度下，在对“高中被试者对以高中为背景生成回答”的评分与“高中被试者对以大学汉语言文学专业为背景生成回答”的评分进行 Unpaired t 检验后，结果显示两组之间的差异不具有统计学显著性（t(108) = 1.2072, p = 0.2300）。置信区间分析表明，95% 置信区间为 [-0.20, 0.82]，该区间包含零，进一步支持两组之间无显著差异的结论。因此，在本研究样本中，两个群体数值无显著差异。

（3）在易懂度维度下，在对“高中被试者对以高中为背景生成回答”的评分与“高中被试者对以大学数学专业为背景生成回答”的评分进行 Unpaired t 检验后，结果显示两组之间的差异不具有统计学显著性（t(108) = 0.4834, p = 0.6298）。置信区间分析表明，95% 置信区间为 [-0.45, 0.74]，该区间包含零，进一步支持两组之间无显著差异的结论。因此，在本研究样本中，两个群体数值无显著差异。

结论：基准值敏感，且经 Unpaired t 检验可得群体数值都无显著差异，则在易懂度维度下，deepseek-v3 以高中为背景的听众设计能力较弱。

3. 术语理解度

在术语理解度度维度下，基准值为 2.109（排名 3/4），为非敏感值，则对基准值和比基准值小且差值超过 0.1 的大学汉语言文学专业为背景生成的回答得分（2.000）做 Unpaired t 检验。

在术语理解度维度下，在对“高中被试者对以高中为背景生成回答”的评分与“高中被试者对以大学汉语言文学专业为背景生成回答”的评分进行 Unpaired t 检验后，结果显示两组之间的差异不具有统计学显著性（t(108) = 0.5188, p = 0.6050）。置信区间分析表明，95% 置信区间为 [-0.31, 0.53]，该区间包含零，进一步支持两组之间无显著差异的结论。因此，在本研究样本中，两个群体数值无显著差异。

结论：基准值不敏感，经 Unpaired t 检验可得两个群体数值无显著差异，则在术语理解度维度下，deepseek-v3 以高中为背景的听众设计能力较弱。

4. 总结结论

本研究通过对高中被试者对不同背景设定下的回答在清晰度、易懂度及术语理解度三个维度上的评分结果进行统计学检验，探讨了 deepseek-v3 在以高中为背景进行听众设计时的表现能力。

在清晰度维度，基准值不敏感，尽管“以大学汉语言文学专业为背景”生成回答的得分低于“以高中为背景”的回答，且二者差值超过 0.1，但统计分析可得不存在显著差异，表明 deepseek-v3 在内容表达的清晰性方面仍存在改进空间。

在易懂度维度，各组对比（包括初中、大学汉语言文学专业及大学数学专业背景）虽均与高中背景存在超过 0.1 的得分差异，但统计检验均未显示显著差异。表明 deepseek-v3 在内容表达的易懂性方面仍存在改进空间。

在术语理解度维度，基准值不敏感，且“以大学汉语言文学专业为背景”生成的回答评分与“以高中为背景”回答评分的差值由统计分析可得不存在显著性差异。由此可见，deepseek-v3 在对高中听众进行术语解释时的听众设计能力仍存在改进空间。

综上所述，deepseek-v3 在面向高中群体进行语言内容生成时，无论在信息的清晰呈现、内容的易于理解，还是术语的使用等方面，听众设计能力均存在不足。

5.2.2.3 大学汉语言文学专业教育背景被试分析

由三维分析表，将大学汉语言文学专业被试者对以大学汉语言文学专业为背景生成的回答的清晰度、易懂度、术语理解度三维分数作为基准值，基准值分别为：2.182、2.400、2.109，找出每一维度每一个比基准值小且差值超过 0.1 的值，将其与基准值的原始数据集做 Unpaired t 检验。

表 5.6 大学汉语言文学专业被试人三维平均分分析表

	清晰度	易懂度	术语理解度
初中	5.622	5.356	5.156
高中	5.289	5.556	5.489
大学汉语言文学专业	**5.422**	**5.489**	**5.244**
大学数学专业	5.511	5.311	5.133

1. 清晰度

在清晰度维度下，基准值为 5.422（排名 2/4），为非敏感值，则对基准值和比基准值小且差值超过 0.1 的高中、大学数学专业为背景生成的回答得分（5.289、5.511）做 Unpaired t 检验。

对“大学汉语言文学专业被试者对以高中为背景生成回答”的评分与“大学汉语言文学专业被试者对以高中为背景生成回答”的评分进行 Unpaired t 检验结果如图：

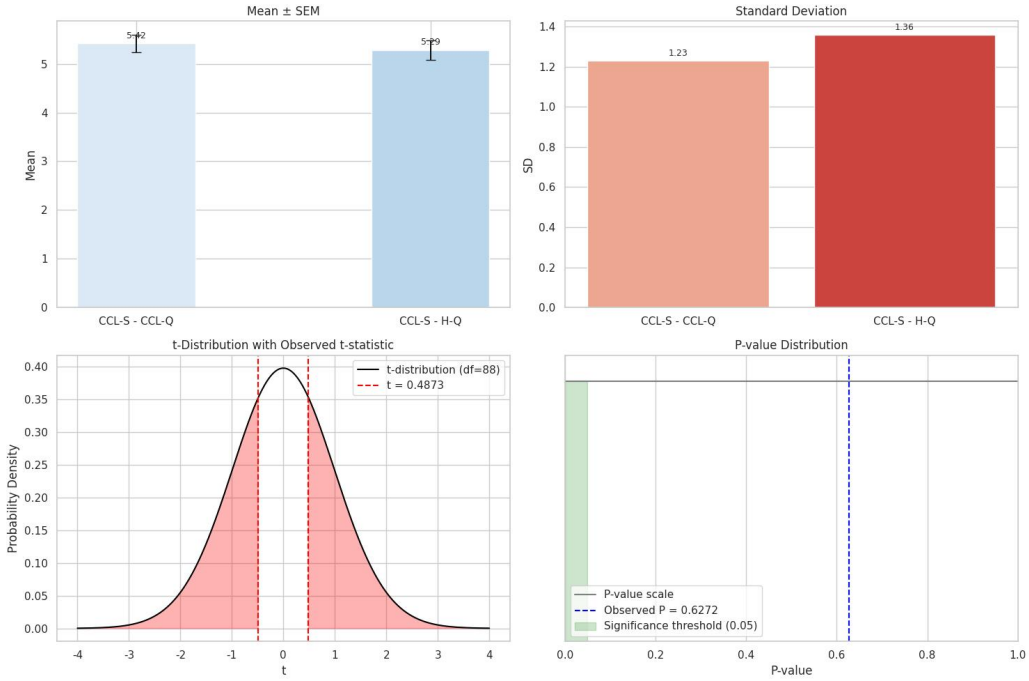

图 5.5 清晰度-大学汉语言文学专业-高中 Unpaired t 检验综合图

在清晰度维度下，在对“高中被试者对以高中为背景生成回答”的评分与“高中被试者对以大学汉语言文学专业为背景生成回答”的评分进行 Unpaired t 检验后，结果显示两组之间的差异不具有统计学显著性（$t(88) = 0.4873$, $p = 0.6272$）。置信区间分析表明，95% 置信区间为 [-0.41, 0.68]，该区间包含零，进一步支持两组之间无显著差异的结论。因此，在本研究样本中，两个群体数值无显著差异。

结论：结合基准值不敏感的前提，可得在清晰度维度下，deepseek-v3 以大学汉语言文学专业为背景的听众设计能力较弱。

2. 易懂度

在易懂度维度下，基准值为 5.489（排名 2/4），为非敏感值，则对基准值和比基准值小且差值超过 0.1 的初中、大学数学专业为背景生成的回答得分（5.356、5.311）

做 Unpaired t 检验。

（1）在易懂度维度下，在对“大学汉语言文学专业被试者对以大学汉语言文学专业为背景生成回答”的评分与“大学汉语言文学专业被试者对以初中为背景生成回答”的评分进行Unpaired t检验后，结果显示两组之间的差异不具有统计学显著性（t(88) = 0.5721, p = 0.5687）。置信区间分析表明，95% 置信区间为 [-0.33, 0.60]，该区间包含零，进一步支持两组之间无显著差异的结论。因此，在本研究样本中，两个群体数值无显著差异。

（2）在易懂度维度下，在对“高中被试者对以高中为背景生成回答”的评分与“高中被试者对以大学汉语言文学专业为背景生成回答”的评分进行 Unpaired t 检验后，结果显示两组之间的差异不具有统计学显著性（t(88) = 0.7023, p = 0.4844）。置信区间分析表明，95% 置信区间为 [-0.33, 0.68]，该区间包含零，进一步支持两组之间无显著差异的结论。因此，在本研究样本中，两个群体数值无显著差异。

结论：基准值不敏感，且经 Unpaired t 检验可得群体数值都无显著差异，则在易懂度维度下，deepseek-v3 以高中为背景的听众设计能力较弱。

3. 术语理解度

在术语理解度度维度下，基准值为 5.244（排名 2/4），为非敏感值，则对基准值和比基准值小且差值超过 0.05 的初中、大学数学专业为背景生成的回答得分（5.156、5.133）做 Unpaired t 检验。

（1）在术语理解度维度下，在“大学汉语言文学专业被试者对以大学汉语言文学专业为背景生成回答”的评分与“大学汉语言文学专业被试者对以初中为背景生成回答”的评分进行 Unpaired t 检验后，结果显示两组之间的差异不具有统计学显著性（t(88) = 0.3364, p = 0.7373）。置信区间分析表明，95% 置信区间为 [-0.44, 0.61]，该区间包含零，进一步支持两组之间无显著差异的结论。因此，在本研究样本中，两个群体数值无显著差异。

（2）在术语理解度维度下，在“大学汉语言文学专业被试者对以大学汉语言文学专业为背景生成回答”的评分与“大学汉语言文学专业被试者对以初中为背景生成回答”的评分进行 Unpaired t 检验后，结果显示两组之间的差异不具有统计学显著性（t(88) = 0.4247, p = 0.6721）。置信区间分析表明，95% 置信区间为 [-0.41, 0.63]，该区间包含零，进一步支持两组之间无显著差异的结论。因此，在本研究样本中，两

个群体数值无显著差异。

结论：基准值不敏感，经 Unpaired t 检验可得两组群体数值无显著差异，则在术语理解度维度下，deepseek 以大学汉语言文学专业为背景的听众设计能力较弱。

4. 总结结论

本研究通过对大学汉语言文学专业被试者对不同背景设定下的回答在清晰度、易懂度及术语理解度三个维度上的评分结果进行统计学检验，探讨了 deepseek-v3 在以大学汉语言文学专业为背景进行听众设计时的表现能力。

在清晰度维度，基准值不敏感，对“以大学汉语言文学专业为背景”生成回答和“以高中为背景”的回答、“以大学数学专业为背景”的回答做得分上的差值统计分析，结果显示均不存在显著差异，表明 deepseek-v3 在内容表达的清晰性方面仍存在改进空间。

在易懂度维度，基准值不敏感，对“以大学汉语言文学专业为背景”生成回答和“以初中为背景”的回答、“以大学数学专业为背景”的回答做得分上的差值统计分析，结果显示均不存在显著差异，表明 deepseek-v3 在内容表达的易懂性方面仍存在改进空间。

在术语理解度维度，基准值不敏感，对“以大学汉语言文学专业为背景”生成回答和“以初中为背景”的回答、“以大学数学专业为背景”的回答做得分上的差值统计分析，结果显示均不存在显著差异由此可见，deepseek-v3 在对大学汉语言文学专业听众进行术语解释时的听众设计能力仍存在改进空间。

综上所述，deepseek-v3 在面向大学汉语言文学专业群体进行语言内容生成时，无论在信息的清晰呈现、内容的易于理解，还是术语的使用等方面，听众设计能力均存在不足。

5.2.2.4 大学数学专业教育背景被试分析

由三维分析表，将大学数学专业被试者对以大学数学专业为背景生成的回答的清晰度、易懂度、术语理解度三维分数作为基准值，基准值分别为：2.182、2.400、2.109，找出每一维度每一个比基准值小且差值超过 0.1 的值，将其与基准值的原始数据集做 Unpaired t 检验。

表 5.7 大学数学专业被试人三维平均分分析表

	清晰度	易懂度	术语理解度
初中	5.368	5.421	5.337
高中	5.505	5.526	5.526
大学汉语言文学专业	5.568	5.505	5.326
大学数学专业	**5.347**	**5.411**	**5.568**

1. 清晰度

在清晰度维度下，基准值为 5.347（排名 4/4），为非敏感值，没有比基准值小的回答得分。

结论：在清晰度维度下，deepseek 以大学数学专业为背景的听众设计能力弱。

2. 易懂度

在易懂度维度下，基准值为 5.411（排名 4/4），为非敏感值，没有比基准值小的回答得分。

结论：在易懂度维度下，deepseek-v3 以大学数学专业为背景的听众设计能力弱。

3. 术语理解度

在术语理解度度维度下，基准值为 5.568（排名 1/4），为敏感值，则对基准值和比基准值小且差值超过 0.1 的初中、大学汉语言文学专业为背景生成的回答得分（5.337、5.326）做 Unpaired t 检验。

（1）在术语理解度维度下，在“大学数学专业被试者对以大学数学专业为背景生成回答”的评分与“大学数学专业被试者对以初中为背景生成回答”的评分进行 Unpaired t 检验后，结果显示两组之间的差异不具有统计学显著性（$t(188) = 1.2928$, $p = 0.1977$）。置信区间分析表明，95% 置信区间为 [-0.12, 0.58]，该区间包含零，进一步支持两组之间无显著差异的结论。因此，在本研究样本中两群体数值无显著差异。

（2）在术语理解度维度下，在“大学数学专业被试者对以大学数学专业为背景生成回答”的评分与“大学数学专业被试者对以大学汉语言文学为背景生成回答”的评分进行 Unpaired t 检验后，结果显示两组之间的差异不具有统计学显著性（$t(188) = 1.3970$, $p = 0.1640$）。置信区间分析表明，95% 置信区间为 [-0.10, 0.58]，该区间包

含零，进一步支持两组之间无显著差异的结论。因此，在本研究样本中，两个群体数值无显著差异。

结论：基准值不敏感，经 Unpaired t 检验可得两组群体数值无显著差异，则在术语理解度维度下，deepseek=v3 以大学数学专业为背景的听众设计能力较弱。

4. 总结结论

本研究通过对大学数学专业被试者对不同背景设定下的回答在清晰度、易懂度及术语理解度三个维度上的评分结果进行统计学检验，探讨了 deepseek-v3 在以大学数学专业为背景进行听众设计时的表现能力。

在清晰度维度，基准值不敏感，且排名最末，表明 deepseek-v3 在内容表达的清晰性方面仍存在改进空间。

在易懂度维度，基准值不敏感，且排名最末，表明 deepseek-v3 在内容表达的易懂性方面仍存在改进空间。

在术语理解度维度，基准值不敏感，对“以大学数学专业为背景”生成回答和“以初中为背景”的回答、“以大学汉语言文学专业为背景”的回答做得分上的差值统计分析，结果显示均不存在显著差异由此可见，deepseek-v3 在对大学数学专业听众进行术语解释时的听众设计能力仍存在改进空间。

综上所述，deepseek-v3 在面向大学数学专业群体进行语言内容生成时，无论在信息的清晰呈现、内容的易懂度，还是术语的使用等方面，听众设计能力均存在不足。

5.2.3 “变”维度结论

本研究系统考察了大语言模型 deepseek-v3 在不同教育背景条件下的听众设计能力，从清晰度、易懂度与术语理解度三个维度开展人工评估与统计分析，人工评估结果显示部分维度中本教育背景回答得分甚至低于其他背景下回答得分。进一步分析，通过 unpaired t 检验，表明尽管模型在部分情境中展现出一定的听众设计能力，尤其是在面向初中受众时，在术语理解度维度表现出与其他背景之间的显著性差异。然而，这一能力并未在更高教育阶段背景下稳定表达。在清晰度、易懂度两个维度更是均未表现出具有统计显著性的风格适配能力。

综上所述，在“变”的维度上，尽管 deepseek-v3 在某些低语境负载条件下（如初中背景）展现出初步的语言适应性机制，但整体而言，其听众设计能力仍不具备系

统性与稳定性。模型未能形成对不同教育背景的风格敏感调节能力，表现出在教育分层语境下“变异性”语言设计的有限性。该结果反映出当前大语言模型在语言生成中的听众建模仍以默认语言样式为主，对社会人口身份的响应仍偏向静态，缺乏深度语用意识与动态生成机制。未来模型优化应强化对语言社会情境的编码机制，提升其在高维度语境中执行语言适配策略的能力，从而实现更高阶的个性化对话生成水平。

5.3 “不变”——抗非教育因素干扰

5.3.1 性别分析

按初中、高中、大学汉语言文学专业、大学数学专业四类教育背景划分deepseek-v3 的生成回答。对每组回答分三维度（清晰度、易懂度、术语理解度），对每一教育背景的每一维度做被试人性别的非配对 t 检验方法差异显著性分析。

5.3.1.1 初中教育背景材料分析

在对“初中女”组（N = 145）与“初中男”组（N = 135）的三维评分分别进行独立样本 t 检验后，结果如下：

1. 清晰度

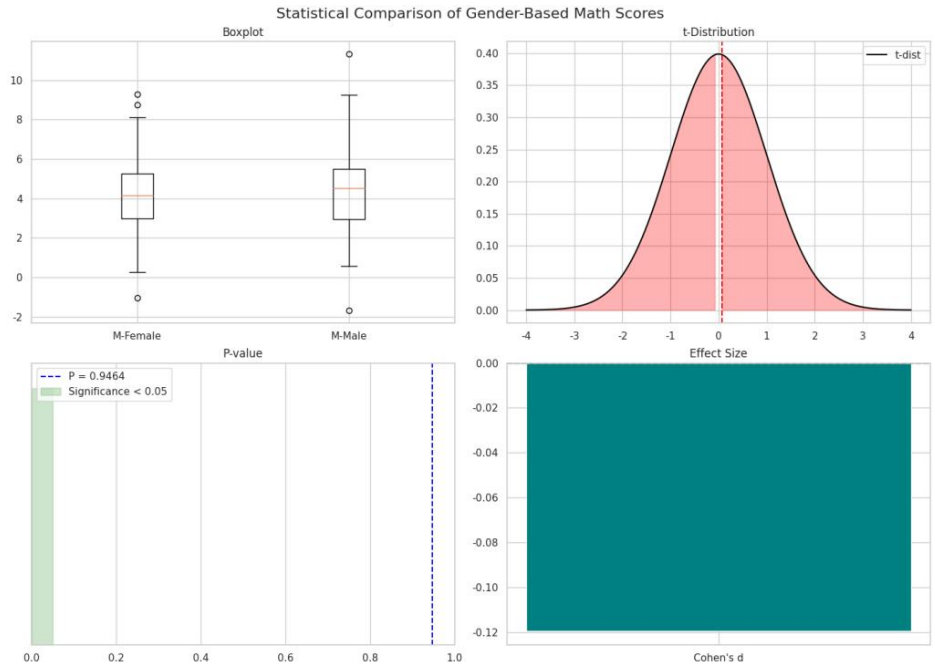

图 5.6 清晰度-初中-女与男 Unpaired t 检验综合图

两组之间的差异不具有统计学显著性（t(278) = 0.0673, p = 0.9464）。其中，女生组的平均成绩为 4.29（SD = 2.03），男生组的平均成绩为 4.27（SD = 1.83），两组的均值差为 0.02，表明在评分上的表现几乎一致。置信区间分析表明，95% 置信区间为 [-0.44, 0.47]，该区间包含零，进一步支持两组之间无显著差异的结论。因此，在本研究样本中，性别因素并未对听众设计产生显著影响。

2. 易懂度

两组之间的差异不具有统计学显著性（t(278) = 0.1331, p = 0.8926）。其中，女生组的平均成绩为 4.24（SD = 1.90），男生组的平均成绩为 4.27（SD = 1.93），两组的均值差为 -0.03，表明在评分上的表现几乎一致。置信区间分析表明，95% 置信区间为 [-0.48, 0.42]，该区间包含零，进一步支持两组之间无显著差异的结论。因此，在本研究样本中，性别因素并未对听众设计产生显著影响。

3. 术语理解度

两组之间的差异不具有统计学显著性（t(278) = 0.4574, p = 0.6478）。其中，女生组的平均成绩为 4.30（SD = 1.85），男生组的平均成绩为 4.20（SD = 1.94），两组的均值差为 0.10，表明在评分上的表现差异小。置信区间分析表明，95% 置信区间为 [-0.34, 0.55]，该区间包含零，进一步支持两组之间无显著差异的结论。因此，在本研究样本中，性别因素并未对听众设计产生显著影响。

4. 总结结论

在本研究样本中，deepseek-v3 对初中教育背景下的听众设计具有对性别因素的抗干扰能力。

5.3.1.2 高中教育背景材料分析

在对“高中女”组（N = 110）与“高中男”组（N = 169）的三维评分分别进行独立样本 t 检验后，结果如下：

1. 清晰度

两组之间的差异不具有统计学显著性（t(278) = 0.4300, p = 0.6675）。其中，女生组的平均成绩为 4.35（SD = 1.90），男生组的平均成绩为 4.24（SD = 1.99），两组的均值差为 0.10，表明在评分上的表现差异较小。置信区间分析表明，95% 置信区间为 [-0.37, 0.57]，该区间包含零，进一步支持两组之间无显著差异的结论。因此，

在本研究样本中，性别因素并未对听众设计产生显著影响。

2. 易懂度

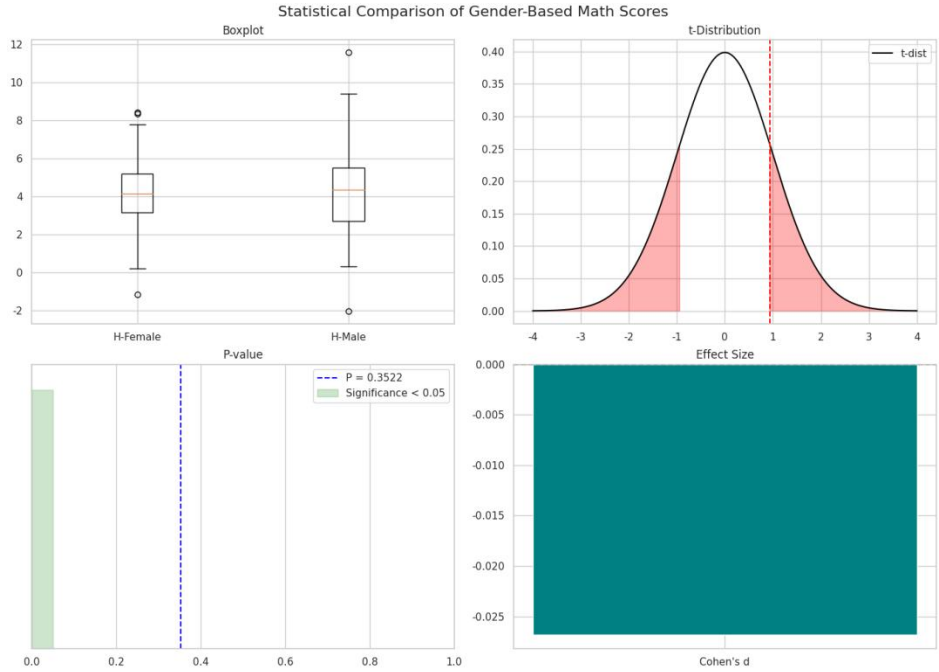

图 5.7 易懂度-高中-女与男 Unpaired t 检验综合图

两组之间的差异不具有统计学显著性（t(278) = 0.9320, p = 0.3522）。其中，女生组的平均成绩为 4.42（SD = 2.13），男生组的平均成绩为 4.19（SD = 1.92），两组的均值差为 0.23，表明在评分上的表现差异较小。置信区间分析表明，95% 置信区间为 [-0.25, 0.71]，该区间包含零，进一步支持两组之间无显著差异的结论。因此，在本研究样本中，性别因素并未对听众设计产生显著影响。

3. 术语理解度

两组之间的差异不具有统计学显著性（t(278) = 0.3892, p = 0.6974）。其中，女生组的平均成绩为 4.30（SD = 1.96），男生组的平均成绩为 4.40（SD = 2.06），两组的均值差为 -0.10，表明在评分上的表现差异小。置信区间分析表明，95% 置信区间为 [-0.58, 0.39]，该区间包含零，进一步支持两组之间无显著差异的结论。因此，在本研究样本中，性别因素并未对听众设计产生显著影响。

4. 总结结论

在本研究样本中，deepseek-v3 对高中教育背景下的听众设计具有对性别因素的抗干扰能力。

5.3.1.3 大学汉语言文学专业教育背景材料分析

在对“大学汉语言文学专业女”组（N = 148）与“大学汉语言文学专业男”组（N = 132）的三维评分分别进行独立样本 t 检验后，结果如下：

1. 清晰度

两组之间的差异不具有统计学显著性（t(278) = 0.4405, p = 0.6599）。其中，女生组的平均成绩为 4.34（SD = 1.85），男生组的平均成绩为 4.24（SD = 2.03），两组的均值差为 0.10，表明在评分上的表现差异较小。置信区间分析表明，95% 置信区间为 [-0.35, 0.56]，该区间包含零，进一步支持两组之间无显著差异的结论。因此，在本研究样本中，性别因素并未对听众设计产生显著影响。

2. 易懂度

两组之间的差异不具有统计学显著性（t(278) = 0.5599, p = 0.5760）。其中，女生组的平均成绩为 4.36（SD = 2.13），男生组的平均成绩为 4.23（SD = 1.97），两组的均值差为 0.14，表明在评分上的表现差异较小。进一步的置信区间分析表明，95% 置信区间为 [-0.35, 0.62]，该区间包含零，进一步支持两组之间无显著差异的结论。因此，在本研究样本中，性别因素并未对听众设计产生显著影响。

3. 术语理解度

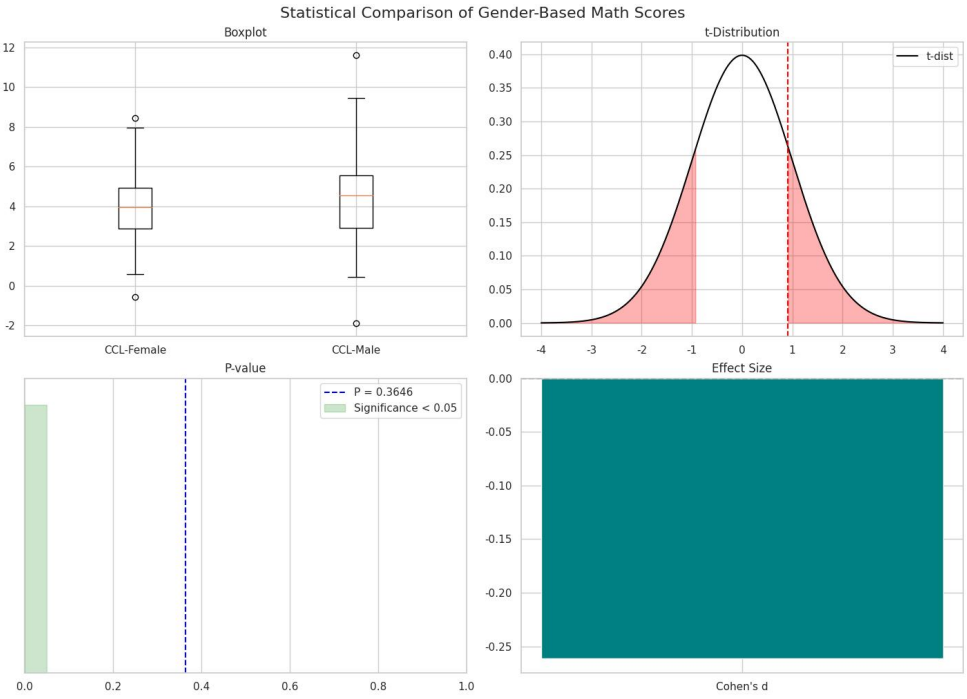

图 5.8 术语理解度-大学汉语言文学专业-女与男 Unpaired t 检验综合图

两组之间的差异不具有统计学显著性（t(278) = 0.9080, p = 0.3646）。其中，女生组的平均成绩为 4.08（SD = 1.77），男生组的平均成绩为 4.28（SD = 1.90），两组的均值差为 -0.20，表明在评分上的表现差异小。进一步的置信区间分析表明，95% 置信区间为 [-0.63, 0.23]，该区间包含零，进一步支持两组之间无显著差异的结论。因此，在本研究样本中，性别因素并未对听众设计产生显著影响。

4. 总结结论

在本研究样本中，deepseek-v3 对大学汉语言文学专业教育背景下的听众设计具有对性别因素的抗干扰能力。

5.3.1.4 大学数学专业教育背景材料分析

在对“大学数学专业女”组（N = 168）与“大学数学专业男”组（N = 112）的三维评分分别进行独立样本 t 检验后，结果如下：

1. 清晰度

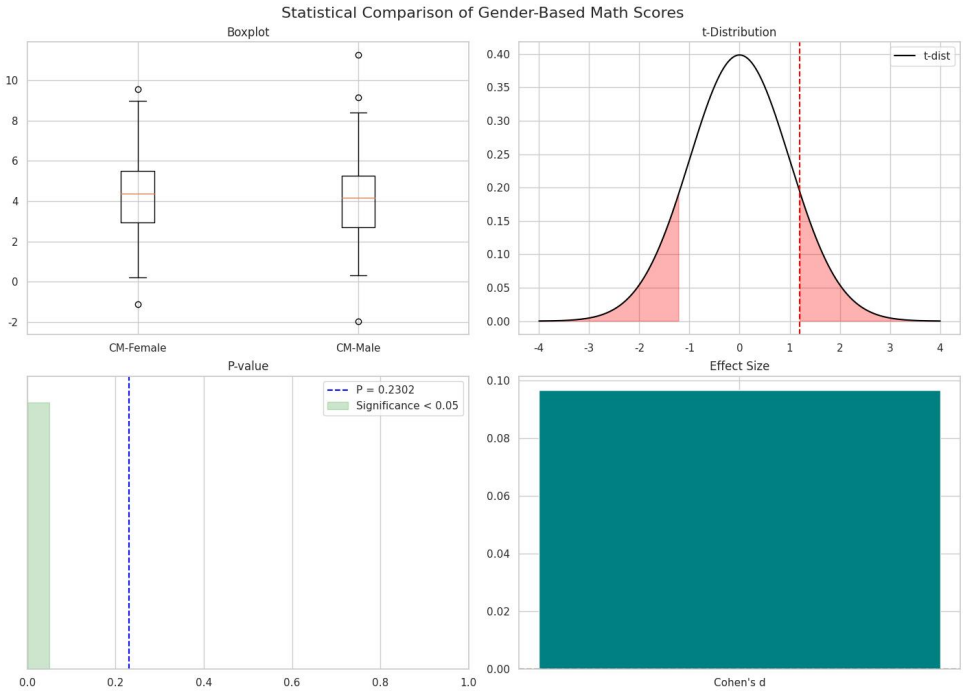

图 5.9 清晰度-大学数学专业-女与男 Unpaired t 检验综合图

两组之间的差异不具有统计学显著性（t(278) = 1.2025, p = 0.2302）。其中，女生组的平均成绩为 4.38（SD = 2.10），男生组的平均成绩为 4.08（SD = 1.86），两组的均值差为 0.29，表明在评分上的表现差异较小。置信区间分析表明，95% 置信区

间为 [-0.19, 0.78]，该区间包含零，进一步支持两组之间无显著差异的结论。因此，在本研究样本中，性别因素并未对听众设计产生显著影响。

2. 易懂度

两组之间的差异不具有统计学显著性（$t(278) = 0.5841, p = 0.5597$）。其中，女生组的平均成绩为 4.20（SD = 1.95），男生组的平均成绩为 4.34（SD = 2.08），两组的均值差为 -0.14，表明在评分上的表现差异较小。进一步的置信区间分析表明，95% 置信区间为 [-0.62, 0.34]，该区间包含零，进一步支持两组之间无显著差异的结论。因此，在本研究样本中，性别因素并未对听众设计产生显著影响。

3. 术语理解度

两组之间的差异不具有统计学显著性（$t(278) = 0.0500, p = 0.9601$）。其中，女生组的平均成绩为 4.20（SD = 1.99），男生组的平均成绩为 4.21（SD = 1.89），两组的均值差为 -0.01，表明在评分上的表现相似。进一步的置信区间分析表明，95% 置信区间为 [-0.48, 0.46]，该区间包含零，进一步支持两组之间无显著差异的结论。因此，在本研究样本中，性别因素并未对听众设计产生显著影响。

4. 总结结论

在本研究样本中，deepseek-v3 对大学数学专业教育背景下的听众设计具有对性别因素的抗干扰能力。

5.3.1.4 总结结论

按大语言模型面对初中、高中、大学汉语言文学专业与大学数学专业四类教育背景身份对所生成的回答分组，对每组回答的被试人分男女来做面对不同教育背景回答得分上的性别差异分析，在“清晰度”“易懂度”“术语理解度”三项评分维度中，均未发现显著性差异。从统计学层面表明，在多元教育背景样本中，性别并未对受众对 deepseek-v3 生成文本的理解和评价产生显著影响。

综上所述，deepseek-v3 在听众设计上表现出较强对性别因素的抗干扰能力，增强了其作为普适性语言生成工具的适用潜力。

5.3.2 地区分析

按初中、高中、大学汉语言文学专业、大学数学专业四类教育背景划分deepseek-v3的生成回答。随机选取三对地区的任一维度的可选教育背景下使用非配对t检验方法做差异显著性分析。

5.3.2.1 定西-杭州材料分析

在清晰度维度上，在全部教育背景下的“定西”组（N = 81）与“杭州”组（N = 72）分别进行独立样本 t 检验后，结果如下：

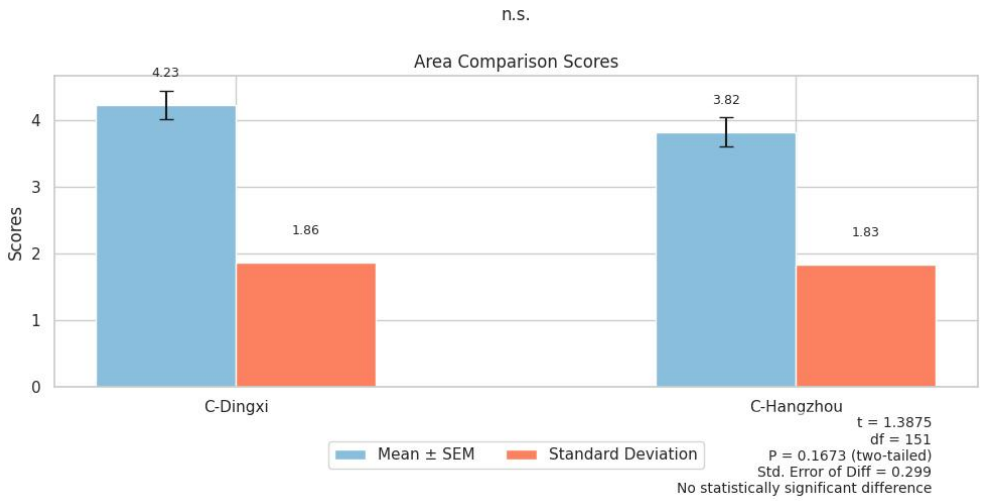

图 5.10 清晰度-定西与杭州 Unpaired t 检验综合图

两组之间的差异不具有统计学显著性（$t(151) = 1.3875$, $p = 0.1673$）。其中，定西组的平均成绩为 4.23（SD = 1.86），杭州组的平均成绩为 3.82（SD = 1.83），两组的均值差为 0.42，表明在评分上的表现差异较小。进一步的置信区间分析表明，95%置信区间为 [-0.18, 1.01]，该区间包含零，进一步支持两组之间无显著差异的结论。因此，在本研究样本中，地区因素并未对听众设计产生显著影响。

5.3.2.2 深圳-赣州材料分析

在易懂度维度上，对初中教育背景下的在对“深圳”组（N = 71）与“赣州”组（N = 46）进行独立样本 t 检验后，结果如下：

图 5.11 易懂度-深圳与赣州 Unpaired t 检验综合图

两组之间的差异不具有统计学显著性（$t(115) = 0.2589$, $p = 0.7962$）。其中，定西组的平均成绩为 4.27（$SD = 1.98$），杭州组的平均成绩为 4.17（$SD = 1.80$），两组的均值差为 0.09，表明在评分上的表现相似。进一步的置信区间分析表明，95% 置信区间为 [-0.62, 0.81]，该区间包含零，进一步支持两组之间无显著差异的结论。因此，在本研究样本中，地区因素并未对听众设计产生显著影响。

5.3.2.3 合肥-襄阳材料分析

在术语理解度维度上，对大学数学专业教育背景下的在对“合肥”组（$N = 46$）与“襄阳”组（$N = 45$）进行独立样本 t 检验后，结果如下：

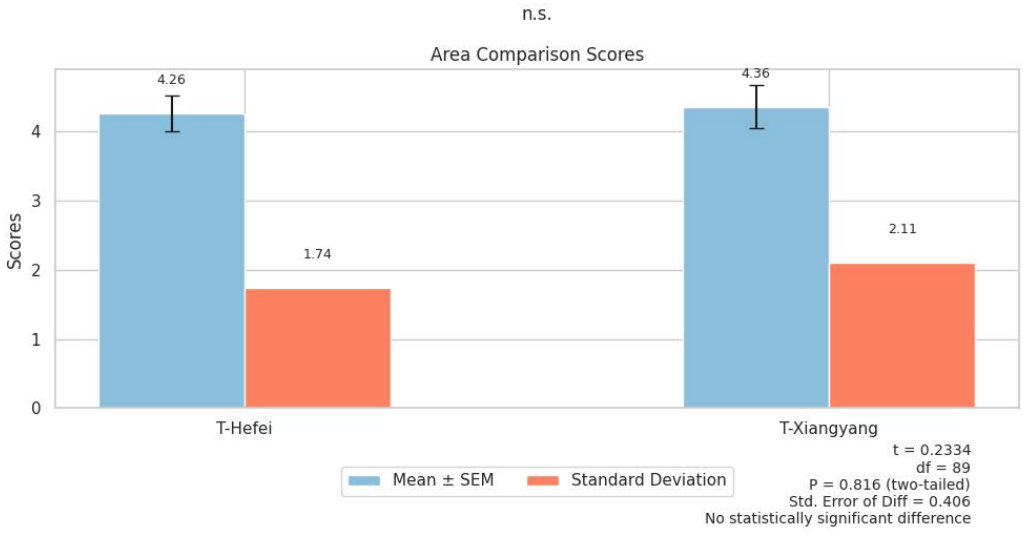

图 5.12 术语理解度-合肥与襄阳 Unpaired t 检验综合图

两组之间的差异不具有统计学显著性（$t(89) = 0.2334, p = 0.8160$）。其中，定西组的平均成绩为 4.26（$SD = 1.74$），杭州组的平均成绩为 4.36（$SD = 2.11$），两组的均值差为 -0.09，表明在评分上的表现相似。进一步的置信区间分析表明，95% 置信区间为 [-0.90, 0.71]，该区间包含零，进一步支持两组之间无显著差异的结论。因此，在本研究样本中，地区因素并未对听众设计产生显著影响。

5.3.2.4 总结结论

基于对“定西-杭州”“深圳-赣州”“合肥-襄阳”三组地区对比样本的非配对 t 检验结果，在“清晰度”“易懂度”“术语理解度”三项评分维度中，均未观察到具有统计学意义的差异。表明 deepseek-v3 所生成文本的听众感知质量未因生成回答时所提示的地理因素而产生系统性偏差，反映出 deepseek-v3 在听众设计过程中对地区语境因素具备较强的抗干扰能力。

考虑到所覆盖地区在经济发展水平、教育资源分布与语言环境等方面存在显著异质性，此种稳定性为大语言模型在公共信息传播、教育内容生成等多场景跨区域应用提供了实践依据与技术支撑。

6 结论

本研究立足于社会语言学关于“听众设计”的理论基础，聚焦于教育场景下大语言模型对不同社会人口背景用户的语言适应能力，以人工智能教育为例，采用人工评估、量化实证的方法对主流大语言模型 deepseek-v3 的听众设计能力进行系统性评估。通过构建“变”与“不变”核心问题分析框架，本研究将受众的教育背景与性别、地区纳入比较体系，结合清晰度、易懂度、术语理解度三维度指标，采用独立样本 t 检验对人工评估结果展开量化实证研究，旨在从语言适应性与内容公正性的角度，评估大语言模型 deepseek-v3 的听众设计能力。在构建包含“教育背景”（可变维度）与“地区因素”（不可变维度）在内的双轴变量体系后，研究以三项核心维度（清晰度、易懂度、术语理解度）为评估指标，通过多组独立样本 t 检验，对模型输出文本的用户适配表现进行了分析。

研究结果表明，在“变”的维度，deepseek-v3 展现出一定的听众设计能力，但仍有待提升。模型在术语理解度维度上对初中教育背景的做出了统计学显著的响应，表现出一定的语体调节能力。但在面向高中、大学汉语言文学专业与大学数学专业三类受众生成文本时，其术语使用呈现出一定的分化倾向，但统计学分析该差异不够显著，在清晰度、易懂度维度同样未对不同教育背景受众表现出明显的听众设计差异，表明大语言模型 deepseek-v3 针对教育背景的听众设计能力仍有待提升。

在“不变”的维度，研究对性别与地区两个典型社会人口背景因素变量展开分析，结果显示模型在这两个维度下生成内容的评分差异均不具统计学显著性。在多个评分维度（清晰度、易懂度、术语理解度）上，无论是性别分组（如男性被试人 vs 女性被试人），还是地区分组（如定西 vs 杭州、深圳 vs 赣州、合肥 vs 襄阳），deepseek-v3 的生成输出均保持高度一致性。这种结果说明，模型的语言生成过程并未受到无关的社会身份变量干扰，具备较强的抗干扰能力，有助于维护教学内容的普适性与教育传播的公正性。

总体而言，deepseek-v3 在教育场景中的听众设计能力表现出“在不该变处保持一致”的特征，但未表现出“在该变处变”特征，体现了大语言模型具备面向教育的适应潜力，但仍处于机制不完善、调节能力有限的发展阶段，仍需进一步完善。

参考文献

[1] Bell A. Language style as audience design[J]. Language in Society, 1984, 13(2): 145-204.

[2] Blodgett S L, Barocas S, Daumé III H, et al. Language (Technology) is Power: A Critical Survey of “Bias” in NLP[J]. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020: 5454–5476.

[3] Takmaz E, Zarrieß S, Bos J. Audience Design in Referring Expressions: A Game-Theoretic Approach[J]. Proceedings of the 24th Conference on Computational Natural Language Learning, 2020: 16–30.

[4] Oshika R, Nishino M, Komachi M. Simplify the Translation with Large Language Models Using Age of Acquisition[J]. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP 2021), 2021: 943–953.

[5] Huebner S, Warstadt A, Bowman S R. Learning Child-Directed Syntax with Large Language Models[J]. Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP 2022), 2022: 12285–12300.

[6] Beck J, Ping J, Gehman S, et al. Sociodemographic Prompting: Evaluating the Effects of Demographic Impressions on LLM Behavior[J]. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023), 2023: 2345–2357.

[7] Mukherjee S, Cho K, Belinkov Y. Prompting Sociodemographic Sensitivities in Large Language Models[J]. Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024), 2024: 8567–8577.

[8] Pan L, Leng Y, Xiong D. Can Large Language Models Learn Translation Robustness from Noisy-Source In-context Demonstrations?[C]. Torino: ELRA and ICCL, 2024: 2798–2808.

[9] Lyu C, Du Z, Xu J, et al. A Paradigm Shift: The Future of Machine Translation Lies with Large Language Models[C]. Torino: ELRA and ICCL, 2024: 1339–1352.

[10] Acerbi A, Stubbersfield J M. Large language models show human-like content biases in transmission chain experiments[J/OL]. Proceedings of the National Academy of Sciences of the United States of America, 2023, 120(44): e2313790120.

致　谢

感谢母校与老师的栽培和指引。

感谢爸爸妈妈的爱和呵护。

感谢朋友们的支持和鼓励。

感谢那些从未放弃的时刻。

世上本无常照月，天边还有再来春，还看今朝；

落子无悔，无问西东，人间本无路，宁作我。

大学期间，感受最深的其实是人与人的关系，和老师，相处都很融洽，无论是生活上还是项目科研上老师们都提供了很多的帮助。

和家人，更不用多说，没有家人的支持和爱，我无法成长健全完备的人格，他们在人生道路上给我指引，在生活上给我提供物质帮助，在情感上永远支持我所做的决定，他们爱是不需要宣之于口的，就已经溢出来了。

和朋友，有更多话要讲，四年荏苒，新朋友慢慢变成老朋友，老朋友慢慢变成一辈子的朋友，高中的朋友们，虽然天南海北，但心在一起，距离隔绝不了共鸣，在难过低落的时候，会有他们站在我身边，在我高兴开心的时候，也会和他们分享喜悦，我不知道友谊该如何描绘，但我记得每一个瞬间的感动。

大学的朋友们，那要说到一个我在大学从未后悔加入的团体——辩论队，因为热爱和对高中一场失败的辩论的不甘，这是加入辩论队的初心，在备赛期间，几乎每个夜晚都在一起沟通，打完比赛的周末被学长姐带着去团建吃饭，慢慢的从大一的懵懂，接过学长姐手中的登山杖，带着新的小孩向上攀登，一攀就是 3 年，我们当然聊辩论，在一起探寻黑夜中微微渺茫的白点，为每一场辩论找公正的平台，为每一句话寻严密的逻辑，更让每一个成员，教自己如何秉持心中公平的天平，去不吝惜自己的理想主义光芒，但也不忽视现实中或近或远的哭声。我们当然也不只聊辩论，我们谈天、谈地、谈情感，去剖析自己的心，然后和别人的做交换，我们爱团建，聊着聊着，冒出一个说走就走的点子，于是拿上东西就出门，在便利店、在快餐厅、在 KTV、在饭店、在轰趴馆，去烧烤、去买菜做饭、去唱歌、去桌游、去看日出、去打羽毛球，迎着清晨的风和光，散步回来，这是辩论队的常态，我爱辩论队，这是大学里给我归属感的最大甚至是唯一的来源。大学里人来人往，停留的少见，本来带队一年就该退休的我，在这儿留了一年又一年，当然是因为爱。

除了辩论队，还有其他团体，比如学生会，我们也度过了美好欢乐的时光，在一起工作，加班，做策划，举办活动，然后去团建，那些快乐的日子流水般流走了，成为大学里程上的一个刻度，我将永远记得我们如何奔波于合唱比赛，如何去布置音乐节场地，如何把南湖音乐节做成系列活动，如何去参与策划评选，不算薄的策划书一本本，那些成就感难以言喻。

寝室，作为必在之所，因为很好的室友们变得舒适和谐，我们彼此包容，也一起出门玩，吐露烦心事时会有室友拉着我的手安慰我，生病会有室友照顾我，我们曾经素未相识，但我们共住同一个屋檐 4 年，那些回忆会成为心海上沉沉浮浮的贝壳一般，每一个都藏着一段珍珠般的回忆。

而我自己，成为了想成为的人，这是最幸福的事，至此。