从黑箱到透明：在大学课堂中利用可解释的人工智能提高自动口译评估

Zhaokun Jiang[1] **Ziyin Zhang**[1]*

[1]Shanghai Jiao Tong University

Abstract

机器学习的最新进展激发了对自动口译质量评估的越来越多的兴趣。然而，现有的研究在语言使用质量的检查方面仍显不足，由于数据的稀缺和不平衡，建模效果不尽如人意，并且缺乏对模型预测解释的努力。为了解决这些问题，我们提出了一个多维建模框架，该框架整合了特征工程、数据增强和可解释的机器学习。这种方法通过仅使用与构造相关的透明特征和进行 Shapley 值（SHAP）分析，优先考虑可解释性而不是“黑箱”预测。我们的结果显示，在一个新的英语-汉语连续口译数据集上具有很强的预测性能，识别出 BLEURT 和 CometKiwi 评分是忠实度的最强预测特征，停顿相关特征是流利度的最强预测特征，而中文特有的措辞多样性度量则是语言使用的最强预测特征。总体而言，通过特别强调可解释性，我们提供了一种可扩展、可靠、透明的替代传统人工评估的方法，有助于为学习者提供详细的诊断反馈，并支持由自动评分单独提供的自我调节学习优势。

1 介绍

口译或口头翻译是一种复杂但重要的语言能力，通过促进高级语言、交流、认知和情感能力的提升，提供广泛的教育益处（Pöchhacker, 2001; Gile, 2021）。它增强积极倾听（Lee, 2013）、口语能力（Han and Lu, 2025）、词汇习得（Chen, 2024）和跨文化交流（Stachl-Peier, 2020），同时也增强高阶认知功能（Dong and Xie, 2014）和焦虑管理能力（Zhao, 2022）。

鉴于其多方面的好处，口译逐渐被认为是一种有价值的教学工具和继听、说、读、写之后的“第五技能”（Mellinger, 2018）。口译的复杂性质需要一个连续的结构化练习、严格评估和诊断反馈的循环（Gile, 2021）。然而，传统的人为评估往往需要评分人员同时查阅源文本、口译输出和详细评分标准，这一认知要求高的过程增加了评分偏差和不一致的风险（Lee, 2019; Han et al., 2024）。

*daenerystargaryen@sjtu.edu.cn

人类评估的固有限制激发了对自动评估的极大兴趣。然而，现有的工作在主题平衡和方法论方面均存在约束。在解释质量的三个既定维度（准确性、流利性和语言使用）中，研究主要集中于前两个，而语言使用则很少受到学术关注（Yu and van Heuven, 2017; Han and Yang, 2023; Wang and Wang, 2022; Han and Lu, 2021; Lu and Han, 2022）。此外，之前的研究主要依赖于传统统计方法，比如相关和回归分析（Yu and van Heuven, 2017; Wang and Wang, 2022; Han and Lu, 2021; Lu and Han, 2022），这些方法基于线性假设，而这些假设在复杂的真实世界数据集中往往并不成立。

机器学习（ML）算法和大型语言模型（LLM）的出现为分析传统统计方法无法解释的复杂数据模式提供了新的机会。然而，在应用中面临的一个显著障碍是数据组成上的严重不平衡。例如，Wang and Yuan (2023) 发现他们的五类分类模型无法识别分布极端的表现（“非常差”和“非常好”），这直接是由于训练数据分布不平衡引起的。另一个限制是自动评分系统固有的不透明性。例如，Jia and Aryadoust (2023) 发现 GPT-4 对表演评估的解释与人类分配分数之间有适度的相关性。关键是，LLM 的内部决策过程依然不透明，只有最终的分数可以获得。这种“黑箱”性质严重限制了 LLM 分数的诊断和教育实用性。

为了应对这些挑战，我们在这项工作中提出以下问题：

1) 我们能否通过数据增强来缓解解释评估模型时的性能不佳问题？

2）忠实度、流利度和语言使用的哪些具体特征在口译评估模型中表现出最强的预测能力？

3) 哪些特定的特征组合会影响每个学生在口译质量每个维度上的得分？为了解答这些问题，我们引入了一种结合特征工程、数据增强和可解释人工智能（XAI）技术的方法（Arrieta et al., 2019; Linardatos et al., 2020），以评估口译表现的三个关键维度：忠实性、流畅性和目标语言质量。在使用变分自编码器（VAEs）增强数据后，我们提取了一组广泛的特征，包括

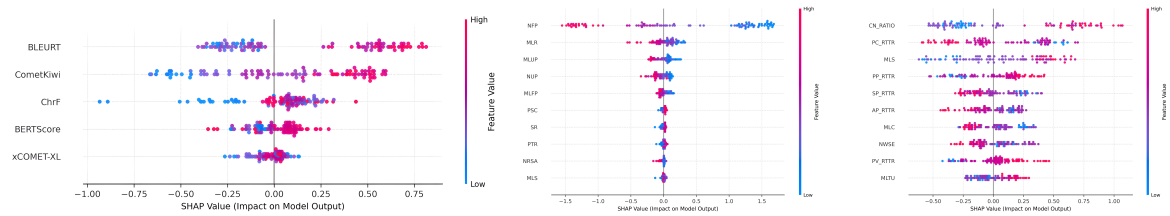

Figure 1: 基于 SHAP 的全球特征重要性，用于 InfoCom（左）、FluDel（中）和 TLQual（右）的预测。较暖的色调（例如红色）表示较高的特征值，而较冷的色调（例如蓝色）表示较低的特征值。这些特征根据其全球重要性沿着 y 轴按降序排列。FluDel 和 TLQual 特征的含义分别在表 2 ， 3 中给出。

翻译质量指标、时间测量和句法复杂性指数，以预测口译表现。基于这些特征，我们采用多维建模策略，分别预测每个维度的表现，这有助于对口译质量进行更细致的分析，并更清晰地揭示特征对每个标准的具体贡献。此外，我们应用 Shapley 值（SHAP）分析，提供全局和个体层面的可解释性解释。据我们所知，这是首次系统地努力实现口译中目标语言质量评估的自动化。

2 相关工作

2.1 自动化口译评估

自动口译评估领域正在经历范式转变，从统计方法向更复杂的神经模型转变。迄今为止，机器学习在口译质量评估中的应用仍然是一个新兴但不断增长的领域。Le et al. (2016) 的开创性工作基于自动语音识别（ASR）和机器翻译（MT）的特征开发了估计器，发现 MT 特征在预测口译质量方面最有影响力。随后，Stewart et al. (2018) 通过支持向量回归调整了 QuEst++ 质量估计管道，以预测同声传译员的表现。最近，Wang and Yuan (2023) 使用 SVM 和 KNN 算法对英中口译进行分类，而 Han et al. (2025) 通过将基于神经的方法与声学和语言指标结合，通过序数逻辑回归进一步推进了这一领域。

信息完整性，也称为忠实度，是指源信息与其翻译之间在信息、语义和语用上的对应程度。现有的自动忠实度评估指标大体可以分为两类：非神经和基于神经的。

非神经网络的指标，如 BLEU (Papineni et al., 2002) 和 chrF (Popovic, 2015) 主要依赖于统计和词汇匹配来量化候选翻译与人工参考之间的词或字符序列的重叠。尽管这些指标在过去的几十年里被广泛采用，它们也因依赖于可能无法捕捉到更深层语义等值的表层比较而受到批评 (Castilho et al., 2018) 。

相较之下，基于神经网络的度量标准源自预训练语言模型，通过比较上下文化的嵌入来超越表面匹配。著名的例子包括 BERTScore (Zhang et al., 2020) 、BLEURT (Sellam et al., 2020) 、CometKiwi (Rei et al., 2022) 和 xCOMET (Guerreiro et al., 2024) 。尽管 Han and Lu (2025) 报告这些分数与人工评价在中英口译上的总体相关性较强，Lu and Han (2022) 发现非神经网络度量标准 BLEU 和 NIST 的表现优于 BERTScore，这表明非神经网络和神经网络度量标准可能捕捉到不同的、可能互补的口译质量方面。

流利程度 流利度是口译质量的另一个关键维度，反映了口译传达的有效性和自然性 (Stenzl, 1983) 。在计算建模中，流利度特征通常被分为三类 (Tavakoli and Skehan, 2005) :(1) 速度流利度，捕捉传达的速度和密度；(2) 中断流利度，通过没有中断和停顿来衡量语音的连续性；以及 (3) 修复流利度，量化自我纠正和重复。

在解释实证研究时，大量证据强调了语速流利性特征（如语速、发声时间比例和发音速度）具有很高的预测能力 (Han and Yang, 2023; Han, 2015; Song, 2020; Yu and van Heuven, 2017) ，而其他研究也确定了中断流利性特征（例如未填充停顿的平均长度）作为强有力的预测因子 (Wang and Wang, 2022; Wu, 2021) 。相比之下，修正流利性特征较少被采用，鲜有显示出强有效预测性的情况 (Han, 2015) 。

在解释评估时，目标语言质量通常指目标语言输出的语法性和惯用性。针对这一维度的自动化评估因 Coh-Metrix、TAASSC、L2SCA 和 CCA 等计算工具的进步而得以促进，这些工具通过计算从词汇和短语指标到句法和话语测量的一系列特征来实现语言质量的操作化。尽管这些特征已广泛应用于二语写作和口语研究，它们在翻译和口译环境中的应用仍处于初期阶段，但现有研究结果显示出相当的前景。

然而，仍然存在两个关键挑战。第一个是对更细粒度的特征设计和应用的需求。虽然像 T 单位复杂性这样的粗粒度的度量已经被长期重视 (Ortega, 2003) ，最近的研究提倡将其与细粒度、基于使用的指标相结合，以捕捉微妙的结构变化并更好地预测语言发展 (Norris and

Ortega, 2009; Kyle and Crossley, 2017) 。第二个挑战涉及语言的特异性，因为大多数 NLP 工具主要是为英语开发的，可能无法充分考虑其他语言的语言特征，比如中文中缺乏明显的形态变化和独特的短语结构 (Li and Thompson, 1989; Hu et al., 2022b) 。

随着大型语言模型（LLM）的出现，这一动态环境变得更加复杂。Zhang et al. (2024b) 最近的一项大规模研究表明，GPT-4o 在语法可接受性判断中达到了接近人类的准确性，这引发了关于如何最佳组合分析工具的问题——从已建立的语言学指数到新兴的基于 LLM 的判断——以提供最强大的预测能力来评估口译中的语言使用。

尽管在自动口译评估方面有上述结果，ML 的实际应用受到了该领域中两个基本且相互关联的数据挑战的阻碍：小样本量和数据成分不平衡。该领域主要以依赖小型数据集的研究为特征 (Yu and van Heuven, 2017; Lu and Han, 2022; Wang and Yuan, 2023; Wang and Wang, 2022) ，这大大增加了过拟合的风险。由于显著的类别不平衡，这一问题进一步恶化，因为大多数数据集往往严重偏向于平均表现，而明显较少的样本代表非常高或非常低的质量 (Wang and Yuan, 2023; Han et al., 2025) 。

为了克服这些障碍，数据扩充已成为一种关键的方法干预措施，能够增强模型的鲁棒性和有效性 (Mumuni and Mumuni, 2022) 。常见的扩充方法包括基于扰动的方法（添加高斯噪声）、SMOTE (Chawla et al., 2002) 等插值技术，以及生成模型如生成对抗网络（GANs）和变分自编码器（VAEs） (Mumuni and Mumuni, 2022) 。在这些方法中，VAE 为基于机器学习的口译评估提供了三个关键优势 (Kingma and Welling, 2014) 。首先，其概率框架能够捕捉保真度、流利度和语言使用特征中的复杂相互依赖关系。其次，连续的潜在空间使得可以在现有样本之间进行平滑插值，以创造一致的变化。第三，VAE 保持特征与标签的对应关系（即每个样本的特征与其相应的口译质量评分之间的直接链接），这对于维护评估的有效性至关重要。Zhang et al. (2024a) 也展示了该技术的实证可行性。

随着教育人工智能系统变得更加复杂，XAI 技术对于理解和验证这些系统至关重要，从而确保其可靠性、信任性和公平性。

当前的可解释人工智能技术主要分为两大类：内在方法和事后方法 (Gilpin et al., 2018; Rudin, 2018; Arrieta et al., 2019; Linardatos et al., 2020) 。内在方法通过使用透明的模型架构（如基于规则的系统、决策树和线性模型）来优先保证模型的固有可解释性，其中系数直接表明了特征的影响。相比之下，事后方法在不改变模型结构的前提下解释已训练的黑箱模型，提供对复杂模型的洞察，使其不再晦涩。最流行的事后方法，如 SHAP (Lundberg and Lee, 2017) 和 LIME (Ribeiro et al., 2016) ，提供特征归因，而其他方法还存在于基于示例的解释和反事实解释 (Arrieta et al., 2019; Linardatos et al., 2020) 。根据其范围，事后方法也可以被分类为全局解释（阐明所有实例中的整体模型行为）或局部解释（澄清个别预测）。

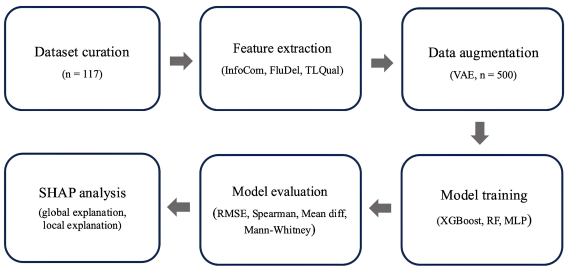

Figure 2: 本研究的方法流程。

在教育领域的 XAI 研究中，学习分析代表了最重要的领域 (Parkavi et al., 2024; Balachandar and Venkatesh, 2024) ，同时也出现了在自动语言评估中的应用，大多数研究集中在解释影响表现质量的因素 (Kumar and Boulanger, 2020; Tang et al., 2024) 。据我们所知，Wang (2024) 是唯一一个关注自动口译评估中可解释性的现有工作，它将口译质量分为 5 个等级，并使用相关分析提供特征重要性的全局解释。

如图 2 所示，本研究遵循一种结构化的方法。首先，我们编制了一个包含 117 个英语到中文方向的学生口译录音的新数据集，从中提取一系列语言学上有意义和理论驱动的特征。为了解决与小样本量和不平衡的评分分布相关的挑战，我们采用 VAE 生成新的、真实的样本 (Kingma and Welling, 2014) 。之后，训练多种机器学习模型来预测不同维度上的口译质量评分。最后，为了解释训练模型的内部决策，我们进行了一系列 SHAP 分析。

我们编制了一个新的数据集，该数据集包括 117 个中英文连续口译样本，这些样本是从中国上海某大学的 39 名英语专业本科生（平均年龄 = 18.47 岁，标准差 = 1.13 岁）中收集的。所有参与者的母语为中文，第二语言为英语，并已通过全国大学英语四级考试（CET-4），展示了良好的英语水平。在数据收集之前，他们完成了 16 周（32 个学分）的口译培训。

口译任务使用了从真实公开演讲中改编的六个段落，每个段落包含相同数量的句子，并对句子长度进行了控制（平均值为 18.14 个词，

标准差为 0.78 个词）。关于这些段落的更多详细信息，以及提取的语言特征值，见附录 A 。这些文本通过 ElevenLabs 的文本到语音技术 [1] 被转换为音频格式。生成的音频文件具有标准的发音，时长约为 2 分钟。

口译样本的评估由三位经验丰富的评分员进行，他们每人在国内或国际环境中具有超过三年的大学教学经验。评估采用 Han (2018) 的四级八分制分析性评分标准，该标准评估口译质量的三个关键维度：信息完整性（InfoCom）、流利度（FluDel）和目标语言质量（TLQual）。为确保评分的一致性，评分员在正式评估前进行了全面培训。评分员培训程序的详细说明、学生分离可靠性以及每位评分员沟合度和外合度均方统计数据在附录 B 中提供。为减少潜在的不一致性和评分员偏差，运用了多面 Rasch 测量（MFRM）分析 (Linacre, 2002) 来校准原始分数并确立最终的基准分数。

2.2 音频处理

为了处理口译的录音，我们首先使用 iFLYTEK ASR 系统 [2] 将它们转录成文本。为了增强注释的可靠性，我们实施了一个两阶段的错误检测过程。

在第一阶段，GPT-4o [3] 被用于通过适应 Rao et al. (2020) 和 Fu et al. (2018) 的框架进行语法错误诊断。设计了一种结构化的提示模板（见附录 C ）以指导 GPT-4o 的注释，针对四大主要错误类型提供明确指令：冗余词（R）、缺失词（M）、词选择错误（S）和词序错误（W）。为了增强模型的性能和可靠性，我们提供给它少量示例，并指示其明确表达对每个识别错误的决策过程，并提供相应的置信度水平。特别地，我们在指南中规定，诸如“呃”之类的填充停顿不被视为错误，分析应仅关注最终句子版本，忽略重复、错误开头或自我更正。

在第二阶段，每个转录都经过人工审核和校正。我们招募了两位语言学的研究生，独立地按照与 GPT 相同的指导方针标注了 100 个随机选取的句子。人工标注者之间的一致性产生了一个 Cohen's Kappa 系数为 0.86，而 GPT-4o 标注与人工标注之间的一致性达到了一个 Fleiss' Kappa 系数为 0.71，显示了一种相当程度的一致性。

2.3 特征提取

口译的每个评分维度由一组独特的提取特征表示。流利度特征是从原始记录中提取的，而其他特征是从清理过的记录（删除填充词、错误开头和自我修复后）中衍生出来的。

对于 InfoCom，我们使用来自机器翻译质量评估领域的五个既定指标来衡量信息从源语言到目标语言的保留情况（表 1 ）。

FluDel 的特征包括来自先前研究的 14 种时间特征（表 2 ），这些特征可以分为两类：语速流利特征（1–6）和断续流利特征（7–14）。与无声停顿相关的特征使用 Python 包 librosa（v0.10.2）和 soundfile（v0.12.1）自动提取，停顿的识别基于 Wu (2021) 推荐的-18 dB 强度阈值。额外的特征来自于 iFLYTEK ASR 系统生成的时间对齐转录。

TLQual 通过与句法复杂性和语法准确性相关的 25 个特征进行评估。其中，利用专为 L2 中文文本开发的中文搭配分析器（CCA，Hu et al., 2022b,a ），提取 21 个涵盖粗粒度和细粒度测量的句法复杂性特征（表 3 ），这使其特别适合用于英汉口译研究。其余 4 个语法准确性特征来自 GPT-4o 的语法错误注释，具体包括冗余词数（NRW）、缺失词数（NMW）、词选择错误数（NWSE）和词序错误数（NWOE）。

2.4 数据增强

不同于一般的 L2 学习者，口译学生由于任务所需的高级语言能力和认知要求而组成了一个较小的群体。这种稀缺性凸显了使用数据增强技术以增加学习者数据集的数量和多样性的必要性（Mumuni and Mumuni, 2022）。根据 Zhang et al. (2024a) 提出的方法，我们使用变分自编码器（VAE）来解决原始数据集中评分分布不平衡的挑战。其主要目的是为评估的口译质量的三个不同维度生成真实的、合成的特征向量。为达到此目的，我们为每一个维度单独训练一个条件 VAE。由这些 VAE 模型生成的合成特征向量随后与原始的 117 个数据点结合，形成一个包含 500 个样本的增强数据集。

三种类型的机器学习模型——XGBoost、随机森林（RF）和多层感知器（MLP）——被用来预测 InfoCom、FluDel 和 TLQual 分数。建模过程遵循系统的步骤，包括特征提取、特征标准化、数据拆分、模型训练和验证以及模型测试。(Mienye and Sun, 2022)

所有提取的特征（如第 2.3 节中详细描述）首先使用 z-score 标准化进行标准化。然后将初始数据集分为训练集（80 %）和测试集（20 %）子集。数据分割后，通过五折交叉验证和超参数的网格搜索进行模型训练和验证，使用均方根误差（RMSE）作为验证标准。

经过交叉验证和超参数优化后，为每个模型选择表现最佳的配置。然后，使用最佳超参数重新训练每个最终模型在整个训练集上，并随

[1] https://elevenlabs.io/

[2] https://global.xfyun.cn/products/real-time-asr

[3] GPT-4o-2024-08-06，温度设置为 0。

Feature	Short description
chrF	Measures n-gram overlap between the interpreted and reference text
BLEURT-20	Assesses the semantic similarity between the interpreted text and reference text based on contextualized embeddings from BERT and RemBERT
BERTScore	Measures the similarity between interpreted and reference translations by computing cosine similarity of their contextualized embeddings using BERT
CometKiwi-da	A reference-free regression model based on the InfoXLM architecture, trained on direct assessments from WMT17-WMT20 and the MLQE-PE corpus
xCOMET-XL	An extension of COMET, designed to identify error spans and assign quality scores, achieving state-of-the-art correlation with MQM error typology-derived scores

Table 1: 用于信息通信评估的特征。

Feature	Full Name	Description
SR	Speech Rate	The overall pace of speech, calculated as the number of syllables uttered per second.
AR	Articulation Rate	The rate of syllable production, excluding pauses.
PTR	Phonation Time Ratio	The proportion of time spent vocalizing relative to the total duration.
MLS	Mean Length of Syllables	The average duration of each syllable.
MLR	Mean Length of Run	The average number of syllables produced in a continuous stream.
PSC	Pruned Syllable Count	The total syllable count after removing filled pauses.
NFP	Number of Filled Pauses	The frequency of filled pauses (e.g., “um,” “uh”).
NUP	Normalized Number of Unfilled Pauses	The frequency of silent pauses. An unfilled pause is defined as a silence of 0.35 seconds or longer, consistent with recommendations for E-C interpreting (Mead, 2005).
MLFP	Mean Length of Filled Pauses	The average duration of filled pauses.
MLUP	Mean Length of Unfilled Pauses	The average duration of silent pauses.
NRLFP	Number of Relatively Long Filled Pauses	The number of filled pauses longer than Q3 + 1.5 * Interquartile Range (IQR) and shorter than or equal to Q3 + 3 * IQR.
NRLUP	Number of Relatively Long Unfilled Pauses	The number of unfilled pauses longer than Q3 + 1.5 * IQR and shorter than or equal to Q3 + 3 * IQR.
NRSA	Number of Relatively Slow Articulations	The number of syllables longer than Q3 + 1.5 * IQR and shorter than or equal to Q3 + 3 * IQR.
NPSA	Number of Particularly Slow Articulations	The number of syllables longer than Q3 + 3 * IQR.

Table 2: 本文研究的 14 个 FluDel 特征。

后在保留的测试集上评估其对未见数据的预测性能。在此阶段，采用多种评估指标来提供模型质量的全面评估，包括：

(1) RMSE：衡量预测误差的大小。

(2) 斯皮尔曼 (ρ): 评估预测分数和实际分数之间的单调关系。

（3）平均绝对误差（MAE）：量化预测值与实际分数之间的平均绝对偏差，提供对预测准确性的直接测量。

（4）Mann-Whitney U 检验：用于确定预测得分和实际得分的分布是否存在显著差异。

(5) 精确一致率 (EAR)：量化预测值与实际评分在四舍五入到最接近的整数后完全匹配的比例。由于我们的模型预测的是连续的 MFRM 校准分数（1-8），而一致性通常是针对离散级别进行评估的，因此需要进行四舍五入。

(6) 相邻一致率 (AAR)：测量预测值在四舍五入到最近的整数后，落在实际得分附近一个单位（+1 或-1）之间的比例。

除了这些总体指标值之外，我们还对预测误差进行了案例研究，以更深入地了解模型性能的特定方面。

我们进一步采用 SHAP 从两个层次解释模型行为：整体模型（全局解释）和单个预测（局部解释）。全局解释通过总结整个数据集中特征的整体影响提供更广泛的视角。另一方面，局

Coarse-Grained	Phraseological Diversity	Phraseological complexity
Mean Length of Sentences (MLS)	Verb-Object Root Type-Token Ratio (VO_RTTR)	Verb-Object Combination Ratio (VO_RATIO)
Mean Length of T-units (MLTU)	Subject-Predicate Root Type-Token Ratio (SP_RTTR)	Subject-Predicate Combination Ratio (SP_RATIO)
Number of T-units Per Sentence (NTPS)	Adjective-Noun Root Type-Token Ratio (AN_RTTR)	Adjective-Noun Combination Ratio (AN_RATIO)
Mean Length of Clauses (MLC)	Adverb-Preposition Root Type-Token Ratio (AP_RTTR)	Adverb-Preposition Combination Ratio (AP_RATIO)
Number of Clauses Per Sentence (NCPS)	Classifier-Noun Root Type-Token Ratio (CN_RTTR)	Classifier-Noun Combination Ratio (CN_RATIO)
	Preposition-Postposition Root Type-Token Ratio (PP_RTTR)	Preposition-Postposition Combination Ratio (PP_RATIO)
	Preposition-Verb Root Type-Token Ratio (PV_RTTR)	Preposition-Verb Combination Ratio (PV_RATIO)
	Predicate-Complement Root Type-Token Ratio (PC_RTTR)	Predicate-Complement Combination Ratio (PC_RATIO)

Table 3: 21 个用于 TLQual 评估的句法复杂性特征。

部解释提供了关于单个特征如何影响单个预测结果的洞见。这些分析是使用 shap 库实现的。

由于所有学生参与者都表现出了一致且合理的口译水平，该数据集缺乏评分在 1-2 范围内的样本。然而，图 3 显示，数据增强成功实现了在剩余范围内的口译得分的近似均匀分布。表格 4 进一步揭示，与原始数据相比，增强数据在三个维度上表现出非常接近的平均值，并且标准差略有增加。本文使用的所有特征的描述性统计数据在附录 D 中提供，原始和增强数据集中特征和得分之间的两两 Spearman 相关性在图 **??** 中展示。

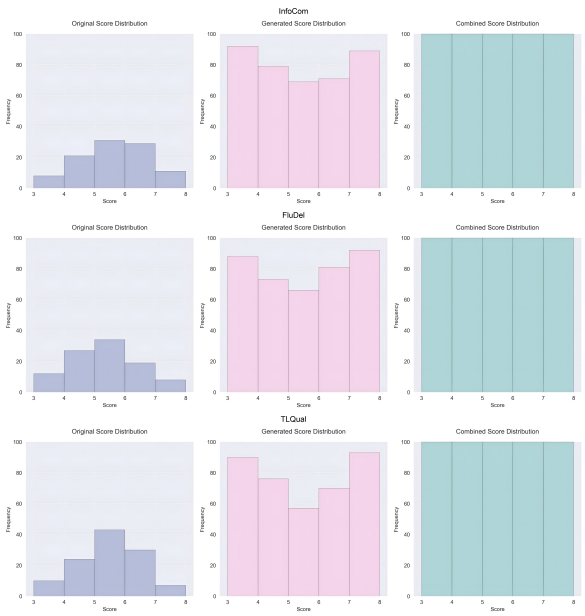

Figure 3: 原始数据（左）、生成数据（中）和扩充数据（右）的分布。

如表 5 所示，在增强数据集上训练的 XG-Boost 在预测 FluDel 和 TLQual 分数时取得了最高性能，比其在原始数据集上的稳健表现有所提升。对于 InfoCom 预测，在增强数据上训练的 RF 回归器取得了最佳结果，明显优于在原始数据上训练的同一模型。相比之下，MLP 一直表现最差，尽管在增强数据上训练时也表现出了显著的改进。在附录 E 中，我们提供了关于模型预测结果与人工评分差异较大的实例进行详细分析，提供了对模型性能特征的细致洞察。

Score		Mean	SD	Skewness	Kurtosis
InfoCom	Raw	5.32	1.35	-0.37	2.25
	Aug.	5.33	1.47	-0.05	-0.51
FluDel	Raw	4.93	0.77	-0.31	2.94
	Aug.	4.95	0.98	-0.10	-0.67
TLQual	Raw	5.21	0.95	-0.23	3.38
	Aug.	5.24	1.06	0.06	-0.85

Table 4: 原始数据和增强数据分数的描述性统计。

2.5 模型预测的全局解释

图 1（左）展示了用于预测 InfoCom 得分的最佳 RF 回归器的全局特征重要性。在这些特征中，BLEURT（M = 0.32, 95 % CI [4] = [0.25, 0.37]）、CometKiwi（M = 0.17, 95 % CI = [0.08, 0.26]）和 chrF（M = 0.07, 95 % CI = [0.04, 0.09]）的平均 SHAP 值最高。换句话说，这些指标的较高值与较高的预测 InfoCom 分数正相关。

[4]为了评估特征贡献的稳定性，进行了一个自举程序，从增强数据集中提取了 1,000 个重采样。对于每个自举样本，使用表现最佳的机器学习模型计算 SHAP 值。记录每个特征在迭代过程中的平均 SHAP 值，以估计其对预测的平均影响。95 % 置信区间（CI）通过自举分布的第 2.5 和第 97.5 百分位数来计算，捕捉每个特征影响的方向和大小。

Score	Model	Data	RMSE	Spearman	MAE	Mann-Whitney U	EAR	AAR
InfoCom	XGBoost	raw	1.36	0.49 **	0.95	259 (p = 0.70)	0.63	0.83
		aug.	1.17	0.62 **	0.49	5751 (p = 0.12)	0.71	0.86
	RF	raw	1.42	0.51 **	0.87	209 (p = 0.45)	0.67	0.88
		aug.	1.05	0.68 **	0.41	5693 (p = 0.15)	0.77	0.90
	MLP	raw	2.43	0.43 *	1.21	215 (p = 0.53)	0.54	0.75
		aug.	1.25	0.58 **	0.79	5744 (p = 0.12)	0.68	0.77
FluDel	XGBoost	raw	0.84	0.69 **	0.65	272 (p = 0.49)	0.69	0.83
		aug.	0.68	0.87 **	0.41	5375 (p = 0.36)	0.72	0.91
	RF	raw	0.70	0.65 **	0.68	274 (p = 0.46)	0.71	0.83
		aug.	0.61	0.86 **	0.43	5302 (p = 0.46)	0.75	0.93
	MLP	raw	1.74	0.39 **	1.17	274 (p = 0.46)	0.54	0.71
		aug.	1.20	0.53 **	0.89	4621 (p = 0.36)	0.64	0.82
TLQual	XGBoost	raw	0.87	0.66 **	0.72	267 (p = 0.41)	0.67	0.83
		aug.	0.75	0.79 **	0.45	5386 (p = 0.33)	0.76	0.91
	RF	raw	0.97	0.58 **	0.86	232 (p = 0.42)	0.63	0.79
		aug.	0.92	0.73 **	0.54	5522 (p = 0.20)	0.78	0.89
	MLP	raw	1.58	0.45 *	1.10	206 (p = 0.40)	0.58	0.75
		aug.	1.04	0.62 **	0.83	4973 (p = 0.95)	0.69	0.85

Table 5: 在原始和增强数据上训练的机器学习回归模型的性能。$^{**}p < 0.01$ ； $^{*}p < 0.05$ 。

如图 1 （中间）所示，NFP（M = -0.17，95 % CI = [-0.27, -0.10]）对 FluDel 评分表现出最强的负面影响，较高的 NFP 值导致 XGBoost 回归预测值更低。同样，其他中断流利度特征，包括 MLUP、NUP 和 MLFP 也对预测结果产生负面影响。速度流利度特征如 PSC、SR、PTR 和 MLS 对模型的预测有积极但非常小的影响，而 MLR 则产生负面影响。

图 1 （右）表明，语法准确性指数 NWSE（M = -0.09, 95 % CI = [-0.15, -0.04]）与模型预测呈反向关系，这表明更高频率的选词错误与更低的预测分数相对应。在短语复杂性特征中，CN_RATIO（M = 0.25, 95 % CI = [0.18, 0.31]）具有最显著的影响，较高的值会导致预测增加。此外，一组短语多样性指标也对模型输出产生正面贡献，包括 PP_RTTR 和 PV_RTTR。相比之下，AP_RTTR 和 PC_RTTR 表现出负面影响。对于粗粒度特征，较高的 MLC 值与模型预测的降低相关，而 MLS 则对预测结果有正面影响。

图

2.6 模型预测的局部解释

展示了样本 25 的 InfoCom 预测的 SHAP 力图，提供了个体特征贡献的详细描述。该图围绕基值（大约 5.4）为中心，代表训练数据集中模型输出的平均值。InfoCom 特征的累积贡献使预测值略微提升至 5.66。在这些特征中，BLEURT 和 COMET-Kiwi 施加了最显著的正面影响，而 chrF 则产生了负面贡献。相对较高的 BLEURT 和 COMET-Kiwi 分数表明样本 25 保留了大部分源信息，尽管有一些损失，而显著低的 chrF 分数表明与参考文本相比在词汇和句法上有显著偏差。

在图 5 中，展示了样本 50 的 FluDel 预测的 SHAP 瀑布图。期望值 E[f(x)] = 4.991 代表了训练数据集中的模型输出均值。特征贡献将预测值集体降低到 f(x) = 4.746。其中，暂停相关特征 - NFP、MLUP 和 NUP - 表现出最显著的负面影响，分别将预测值减少了 0.22、0.16 和 0.1。相反，MLR 具有最强的正面影响，将预测值增加了 0.2。这些发现表明，解释者可能需要通过减少暂停的频率和时长，同时努力实现更长时间的不间断语言输出，来增强暂停管理能力。

图 6 展示了样本 87 的 TLQual 预测的 SHAP 瀑布图。模型的预期值为 E[f(x)] = 5.258，特征贡献共同将预测值增加到 6.466。其中，CN_Ratio 是最具影响力的正因素，将预测值增加了 0.47。其他贡献特征包括 PC_RTTR、AP_RTTR、PV_RTTR 和 AP_Ratio。相反，PP_RTTR 施加了最显著的负面影响，使预测值减少了 0.44，还有 PV_Ratio 和 MLC 带来的额外的负面贡献。这些结果表明，多样和复杂的 CN、PC、AP、PV 和 AP 结构的使用符合该上下文中的典型语言模式。然而，PP 结构

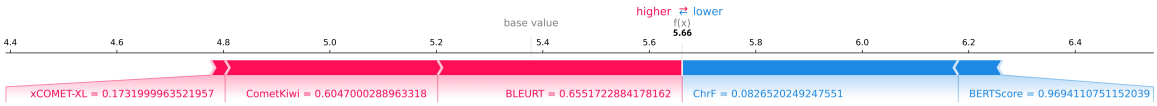

Figure 4: 样本 25 的 InfoCom 预测的 SHAP 力量图。

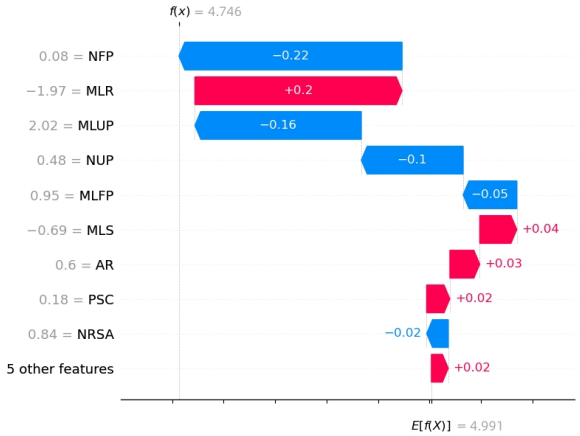

Figure 5: 样本 50 预测 FluDel 分数的 SHAP 瀑布图。

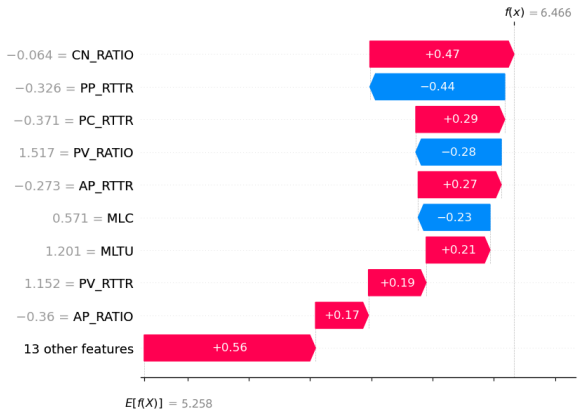

Figure 6: 样本 87 预测 TLQual 分数的 SHAP 瀑布图。

的过度使用（例如 UTF8gkai 在... 上, 当... 时）似乎是有害的。此外，MLC 的负面影响表明，复杂的从句可以重组为较简单的句子或采用主题评论结构进行重新表达，这是中文中常见的语法模式。

我们的分析表明，所选的机器学习算法在增强数据集上表现出稳健的性能，其中随机森林（RF）在 InfoCom 得分估计中产生了最佳结果，而 XGBoost 在 FluDel 和 TLQual 上表现最佳。与之前仅在中间范围得分表现良好但在低端和高端范围表现欠佳的模型相比，我们的结果强调了数据增强在提高模型性能方面的重要性，特别是对于极端范围得分的预测。

我们的 SHAP 分析识别出两个基于神经网络的指标，BLEURT 和 CometKiwi，对 InfoCom 评分的全局预测有最大的影响，这与 Han and Lu (2025) 的先前研究一致。BLEURT 表现优异可能归因于它在合成数据上的广泛预训练以及其结合多样化词汇和语义信号的能力，这使该指标能够比 BERTScore (Han and Lu, 2025) 捕捉更细微的语言模式。相反，XCOMET 的相对较低性能可能源于其训练模式（错误注释）与评估情境（分析量表评分）之间的不一致。

表达流畅性 我们的研究结果显示，NFP 对模型的 FluDel 评分的全局预测有最显著的负面影响，其次是其他与停顿相关的特征，包括 MLUP、NUP 和 MLFP，这与之前的研究 (Yu and van Heuven, 2017) 一致。相反，大多数速率流利度特征（如 PSC、PTR、SR）表现出轻微的正面影响，尽管较高的 MLR 值与预测的降低有关。我们假设 MLR 的负面角色源于这样一种现象：过长的连续话语并不反映出有控制的流畅表达，而是一种“跑题讲话”。译者在高认知负荷下，可能会急于输出信息，而没有战略性地停顿以强调或帮助听众理解 (Lennon, 1990; Mead, 2005)，导致人工评分者将其感知为管理不善且难以处理的讲话。

目标语言质量 在 GPT-4o 注释特征中，NWSE 对模型预测产生显著的负面影响，强调了语法准确性在人类语言质量判断中的基础性作用，并反映了在 L2 口语评估中的发现 (Li et al., 2024)。

关于与长度相关的特征，MLS 对预测有积极作用，这与 Zechner et al. (2017) 一致。然而，MLC 则产生负面影响，这与其他语境中如德语和英语作为第二语言的研究结果形成鲜明对比 (Neary-Sundquist, 2017; Bulté and Roothooft, 2020)。这种差异可能源于类型学上的不同: 汉语的话题-评论结构优先考虑话语连贯性，而英语中常见的句法详细说明更依赖于复杂的从句依赖关系 (Li and Thompson, 1989)。这表明，在中文口译的语境中，句法密度较低但较长的句子被视为更高质量。

另一个关键发现是细粒度特征相比粗粒度

特征具有更高的预测重要性。在这一类别中，反映短语多样性的特征（PC_RTTR, PP_RTTR, SP_RTTR, AP_RTTR, PV_RTTR）比单一的短语复杂性特征（CN_RATIO）更具影响力。此外，我们的结果显示，中国特有的短语特征（CN, PC, PP, PV）比其语言无关的对应项（SP, AP）显示出更大的重要性。总之，这些发现表明，对于英中交替传译来说，对语言使用的稳健评估更少依赖于传统的分句复杂性测量，而更多依赖于语言特有的短语单位的多样性和准确性使用。

2.7 局部解释在自动口译评估中的关键作用

自动口译评估中的局部解释为教学和学习实践提供了重要价值（Kumar and Boulanger, 2020; Tang et al., 2024; Gilpin et al., 2018; Rudin, 2018; Linardatos et al., 2020）。对于教育工作者来说，这些解释通过突出影响预测分数的正面或负面因素，提供了关于学生个人表现的具体优劣的可操作性洞察。这使教师能够量身定制反馈和教学策略，以针对需要改进的具体领域。对于学生而言，局部解释使他们能够掌握自己的学习，专注于需要关注的具体表现方面。

以样本 50 的 FluDel 预测为例，基于 SHAP 分析的局部解释，值得注意的是，与停顿相关的特征成为主要的影响因素：NFP 将预测降低了 0.22，MLUP 降低了 0.16，而 NUP 降低了 0.1，这表明学生在犹豫管理方面存在困难。为此，教师可以实施有针对性的练习，如跟读练习，让学生以最小的延迟再现源语言（Christoffels and de Groot, 2004）。教师还可以实施专注的训练，要求学生在无犹豫的情况下进行短语段的复述，并逐步延长段落长度，同时监控停顿的减少。为特别减少无声停顿，预测练习可以帮助学生预测即将到来的内容元素，从而减少处理延迟（Chmiel, 2020）。此外，教授切分策略 - 将信息组织成易于处理的单元 - 可以减轻经常表现为延长停顿的认知负荷（Thalmann et al., 2019）。

此外，SHAP 值的定量特性也让教师能够有效地优先安排干预措施。对于这个特定的学生来说，处理填充停顿应该优先于长时间的未填充停顿，因为其负面影响更大（0.22 对 0.16）。此外，在多次表现中纵向跟踪这些 SHAP 贡献，使教师能够监控学习进展和干预效果，从而便于根据需要及时调整教学方法。

3 结论

在这项工作中，我们提出了一个有效的框架，集成了特征工程、机器学习模型、数据增强和 XAI，用于对口译质量的多维评估。一个关键发现是，基于 VAE 的数据增强显著提高了模型性能。全局 XAI 分析表明，忠实度的预测最受神经嵌入指标（如 BLEURT）的影响，而流利度得分主要受到分解特征的影响，其中 NFP 产生最强的负面影响。目标语言质量则在很大程度上依赖于特定语言的短语学特征，特别是 CN_RATIO。这些全局见解通过深入的局部解释得到了补充，这些解释能够有效诊断个别表现的优点和缺点。展望未来，我们的方法为将 XAI 驱动的见解转化为提供可操作反馈的教学工具提供了一个有前途的方向，从而弥合自动化评估与学生学习之间的差距。

References

Alejandro Barredo Arrieta, Natalia Díaz Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, A. Barbado, Salvador García, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2019. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Inf. Fusion*, 58:82–115.

V. Balachandar and K. Venkatesh. 2024. A multi-dimensional student performance prediction model (mspp): An advanced framework for accurate academic classification and analysis. *MethodsX*, 14.

Bram Bulté and Hanne Roothooft. 2020. Investigating the interrelationship between rated l2 proficiency and linguistic complexity in l2 speech. *System*.

Sheila Castilho, Stephen Doherty, Federico Gaspari, and Joss Moorkens. 2018. Approaches to human and machine translation quality assessment.

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.*, 16:321–357.

Sijia Chen. 2024. Effects of subtitles on vocabulary learning through videos: An exploration across different learner types. *The Journal of Specialised Translation*.

Agnieszka Chmiel. 2020. Effects of simultaneous interpreting experience and training on anticipation, as measured by word-translation latencies.

Ingrid Christoffels and Annette M.B. de Groot. 2004. Components of simultaneous interpreting: Comparing interpreting with shadowing and paraphrasing. *Bilingualism: Language and Cognition*, 7:227 – 240.

Yanping Dong and Zhilong Xie. 2014. Contributions of second language proficiency and interpreting experience to cognitive control differences among young adult bilinguals. *Journal of Cognitive Psychology*, 26:506 – 519.

Ruiji Fu, Zhengqi Pei, Jiefu Gong, Wei Song, Dechuan Teng, Wanxiang Che, Shijin Wang, Guoping Hu, and Ting Liu. 2018. Chinese grammatical error diagnosis using statistical and prior knowledge driven features with probabilistic ensemble enhancement. In *Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications, NLP-TEA@ACL 2018, Melbourne, Australia, July 19, 2018*, pages 52–59. Association for Computational Linguistics.

Daniel Gile. 2021. The effort models of interpreting as a didactic construct. *Advances in Cognitive Translation Studies*.

Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael A. Specter, and Lalana Kagal. 2018. Explaining explanations: An overview of interpretability of machine learning. In *5th IEEE International Conference on Data Science and Advanced Analytics, DSAA 2018, Turin, Italy, October 1-3, 2018*, pages 80–89. IEEE.

Nuno Miguel Guerreiro, Ricardo Rei, Daan van Stigt, Luísa Coheur, Pierre Colombo, and André F. T. Martins. 2024. xcomet : Transparent machine translation evaluation through fine-grained error detection. *Trans. Assoc. Comput. Linguistics*, 12:979–995.

Chao Han. 2015. (para)linguistic correlates of perceived fluency in english-to-chinese simultaneous interpretation. *International Journal of Comparative Literature and Translation Studies*, 3:32–37.

Chao Han. 2018. Using analytic rating scales to assess english/chinese bi-directional interpretation: A longitudinal rasch analysis of scale utility and rater behavior. *Linguistica Antverpiensia, New Series – Themes in Translation Studies*.

Chao Han and Xiaolei Lu. 2021. Can automated machine translation evaluation metrics be used to assess students' interpretation in the language learning classroom? *Computer Assisted Language Learning*, 36:1064 – 1087.

Chao Han and Xiaolei Lu. 2025. Beyond bleu: Repurposing neural-based metrics to assess interlingual interpreting in tertiary-level language learning settings. *Research Methods in Applied Linguistics*.

Chao Han, Xiaolei Lu, and Shirong Chen. 2025. Modeling rater judgments of interpreting quality: Ordinal logistic regression using neural-based evaluation metrics, acoustic fluency measures, and computational linguistic indices. *Research Methods in Applied Linguistics*.

Chao Han and Liuyan Yang. 2023. Relating utterance fluency to perceived fluency of interpreting. *Translation and Interpreting Studies. The Journal of the American Translation and Interpreting Studies Association*, 18(3):421–447.

Chao Han, Binghan Zheng, Mingqing Xie, and Shirong Chen. 2024. Raters' scoring process in assessment of interpreting: an empirical study based on eye tracking and retrospective verbalisation. *The Interpreter and Translator Trainer*, 18:400 – 422.

Renfen Hu, Jifeng Wu, and Xiaofei Lu. 2022a. Chinese collocation analyzer (cca).

Renfen Hu, Jifeng Wu, and Xiaofei Lu. 2022b. Wordcombinationbased measures of phraseological diversity, sophistication, and complexity and their relationship to second language chinese proficiency and writing quality. *Language Learning*.

Yichen Jia and Vahid Aryadoust. 2023. The utility of generative artificial intelligence in rating interpreters' accuracy: A case study of chatgpt-4.

Diederik P. Kingma and Max Welling. 2014. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

Vivekanandan Kumar and David Boulanger. 2020. Explainable automated essay scoring: Deep learning really has pedagogical value. In *Frontiers in Education*.

Kristopher Kyle and Scott Andrew Crossley. 2017. Assessing syntactic sophistication in l2 writing: A usage-based approach. *Language Testing*, 34:513 – 535.

Ngoc-Tien Le, Benjamin Lecouteux, and Laurent Besacier. 2016. Joint ASR and MT features for quality estimation in spoken language translation. In *Proceedings of the 13th International Conference on Spoken Language Translation, IWSLT 2016, Seattle, WA, USA, December 8-9, 2016*. International Workshop on Spoken Language Translation.

Sang-Bin Lee. 2019. Holistic assessment of consecutive interpretation. *Interpreting. International Journal of Research and Practice in Interpreting*.

T. Lee. 2013. Incorporating translation into the language classroom and its potential impacts upon l2 learners.

P. Alan Lennon. 1990. Investigating fluency in efl: A quantitative approach. *Language Learning*, 40:387–417.

C.N. Li and S.A. Thompson. 1989. *Mandarin Chinese: A Functional Reference Grammar*. Linguistics: Asian studies. University of California Press.

Wenchao Li, Zhentao Zhong, and Haitao Liu. 2024. A computer-assisted tool for automatically measuring non-native japanese oral proficiency. *Computer Assisted Language Learning*.

John M. Linacre. 2002. What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions*, 16:878.

Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris B. Kotsiantis. 2020. Explainable ai: A review of machine learning interpretability methods. *Entropy*, 23.

Xiaolei Lu and Chao Han. 2022. Automatic assessment of spoken-language interpreting based on machine-translation evaluation metrics. *Interpreting. International Journal of Research and Practice in Interpreting*.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Neural Information Processing Systems*.

Peter Mead. 2005. Methodological issues in the study of interpreters' fluency.

C. Mellinger. 2018. Translation, interpreting, and language studies: Confluence and divergence. *Hispania*, 100:241 – 246.

Ibomoiye Domor Mienye and Yanxia Sun. 2022. A survey of ensemble learning: Concepts, algorithms, applications, and prospects. *IEEE Access*, 10:99129–99149.

Alhassan G. Mumuni and Fuseini Mumuni. 2022. Data augmentation: A comprehensive survey of modern approaches. *Array*, 16:100258.

Colleen A. Neary-Sundquist. 2017. Syntactic complexity at multiple proficiency levels of l2 german speech. *International Journal of Applied Linguistics*, 27:242–262.

John M. Norris and Lourdes Ortega. 2009. Towards an organic approach to investigating caf in instructed sla: The case of complexity. *Applied Linguistics*, 30:555–578.

Lourdes Ortega. 2003. Syntactic complexity measures and their relationship to l2 proficiency: A research synthesis of college-level l2 writing. *Applied Linguistics*, 24:492–518.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

R. Parkavi, P. Karthikeyan, and A. Sheik Abdullah. 2024. Enhancing personalized learning with explainable ai: A chaotic particle swarm optimization based decision support system. *Appl. Soft Comput.*, 156:111451.

Franz Pöchhacker. 2001. Quality assessment in conference and community interpreting. *Meta: Translators' Journal*, 46:410–425.

Maja Popovic. 2015. chrf: character n-gram f-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation, WMT@EMNLP 2015, 17-18 September 2015, Lisbon, Portugal*, pages 392–395. The Association for Computer Linguistics.

Gaoqi Rao, Erhong Yang, and Baolin Zhang. 2020. Overview of nlptea-2020 shared task for chinese grammatical error diagnosis. *Proceedings of the 6th Workshop on Natural Language Processing Techniques for Educational Applications*.

Ricardo Rei, Marcos V. Treviso, Nuno Miguel Guerreiro, Chrysoula Zerva, Ana C. Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte M. Alves, Luísa Coheur, Alon Lavie, and André F. T. Martins. 2022. Cometkiwi: Ist-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation, WMT 2022, Abu Dhabi, United Arab Emirates (Hybrid), December 7-8, 2022*, pages 634–645. Association for Computational Linguistics.

Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144. ACM.

Cynthia Rudin. 2018. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1:206 – 215.

Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020. BLEURT: learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7881–7892. Association for Computational Linguistics.

Shuxian Song. 2020. Fluency in simultaneous interpreting of trainee interpreters : the perspectives of cognitive, utterance and perceived fluency.

U. Stachl-Peier. 2020. Translating, interpreting, mediating: The cefr and advanced-level language learning in the digital age.

Catherine Stenzl. 1983. Simultaneous interpretation: Groundwork towards a comprehensive model.

Craig Stewart, Nikolai Vogler, Junjie Hu, Jordan L. Boyd-Graber, and Graham Neubig. 2018. Automatic estimation of simultaneous interpreter performance. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers*, pages 662–666. Association for Computational Linguistics.

Xiaoyi Tang, Hongwei Chen, Daoyu Lin, and Kexin Li. 2024. Incorporating fine-grained linguistic features and explainable ai into multi-dimensional automated writing assessment. *Applied Sciences*.

Parveneh Tavakoli and Peter Skehan. 2005. Strategic planning, task structure and performance testing.

Mirko Thalmann, Alessandra S. Souza, and Klaus Oberauer. 2019. How does chunking help working memory? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 45:37–55.

Xiaoman Wang. 2024. Developing an automated graded assessment system for english/chinese interpreting.

Xiaoman Wang and Binhua Wang. 2022. Identifying fluency parameters for a machine-learning-based automated interpreting assessment system. *Perspectives*, 32:278 – 294.

Xiaoman Wang and Lu Yuan. 2023. Machine-learning based automatic assessment of communication in interpreting. In *Frontiers in Communication*.

Zhiwei Wu. 2021. Chasing the unicorn? the feasibility of automatic assessment of interpreting fluency.

Wenting Yu and Vincent J. van Heuven. 2017. Predicting judged fluency of consecutive interpreting from acoustic measures: Potential for automatic assessment and pedagogic implications. *Interpreting*, 19:47–68.

Klaus Zechner, Su-Youn Yoon, S. Bhat, and Chee Wee Leong. 2017. Comparative evaluation of automated scoring of syntactic competence of non-native speakers. *Comput. Hum. Behav.*, 76:672–682.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Yidi Zhang, Margarida Lucas, Pedro Bem-haja, and Luís Pedro. 2024a. The effect of student acceptance on learning outcomes: Ai-generated short videos versus paper materials. *Comput. Educ. Artif. Intell.*, 7:100286.

Ziyin Zhang, Yikang Liu, Weifang Huang, Junyu Mao, Rui Wang, and Hai Hu. 2024b. MELA: multilingual evaluation of linguistic acceptability. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 2658–2674. Association for Computational Linguistics.

Nan Zhao. 2022. Speech disfluencies in consecutive interpreting by student interpreters: The role of language proficiency, working memory, and anxiety. *Frontiers in Psychology*, 13.

A 源材料的更多细节

Passage	Theme	DESWC	DESSL	DESWLlt	LDTTRa	RDFRE	RDFKGL	RDL2
1	Migration	185	19.32	5.11	0.72	35.25	15.05	13.37
2	Migration	193	18.91	5.42	0.65	41.12	16.23	8.35
3	Festival	182	19.75	5.16	0.75	29.74	16.51	8.90
4	Festival	191	18.86	5.32	0.68	42.18	14.66	10.20
5	Social equality	179	19.44	5.23	0.74	45.36	13.28	7.39
6	Social equality	185	19.06	5.28	0.66	33.33	15.73	11.46

Table 6: 用于口译任务的六个段落的基本信息。DESWC：字数；DESSL：句子长度（字数）；DESWLlt：词长（均值）；LDTTRa：词汇密度（类型-词形比）；RDFRE：Flesch 易读性指数；RDFKGL：Flesch-Kincaid 年级等级；RDL2：二语可读性。

B 评估者训练程序

为了让评分员熟悉评估程序，我们通过视频会议软件安排了一次在线培训课程。本研究的两位作者介绍了源文本和相应的参考译文，并对分析评分标准中的某些关键术语（例如“填充停顿”、“长时间沉默”和“过多修正”）进行了说明。我们积极鼓励评分员对评分任务的任何方面提出疑问，以确保对评估标准有共同的理解。为了提高评分一致性，播放和集体分析了每个等级预先评分的有代表性的口译，这有助于说明与不同表现水平相关的典型特征。随后，评分员独立完成了五个附加口译的试评分。之后，他们进行了协同讨论，比较他们的分数并为其评分决策提供合理性。正式评分也在远程进行，每位评分员通过安全的在线访问获取所有必要材料，包括源文本、参考译文和匿名的口译。为了确保充足的时间进行全面评估，评分员被给予两周的时间来完成他们的评估。

Dimension		Infit MnSq	Outfit MnSq	Rater reliability	Person Separation reliability
InfoCom	Rater 1	1.02	1.01	0.97	0.83
	Rater 2	0.84	0.79		
	Rater 3	0.78	0.65		
FluDel	Rater 1	1.15	1.08	0.98	0.81
	Rater 2	1.04	1.01		
	Rater 3	0.89	0.75		
TLQual	Rater 1	1.07	0.99	0.96	0.76
	Rater 2	0.90	0.93		
	Rater 3	0.82	1.04		

Table 7: 来自 MFRM 分析的题内、题外、评分员信度和人物分离信度统计。

C 中文语法错误诊断提示

Prompt for Chinese grammatical error diagnosis

** 指令 ** 您是一位中文语法专家。您的任务是诊断并纠正中文句子或较长文本中的语法错误。请严格遵循以下步骤和指南：
1. 错误检测和分析顺序按照以下优先级顺序分析输入文本中可能存在的错误：- 冗余（R）：重复的词或字符使句子不必要地臃肿。- 缺词（M）：遗漏的词或颗粒使句子不完整或含糊不清。- 词选择（S）：不当或不准确的词选择应该被更符合上下文的术语替换。- 词序（W）：不正确的词或短语排列扭曲了原意。
2. 错误描述和纠正对于每个检测到的错误：- 描述错误性质。- 提出一个纠正方案，明确意义同时保留原意。- 分配一个代表纠正准确性的信心分数（0-1）。分数越接近 1 表示信心越高。
3. 对低信心的再检查如果错误得到的信心分数低于 0.7，请通过以下问题重新检查：- “这个纠正提升了句子而不引入歧义吗?” - “错误类型是否正确归类?” 在必要时修订纠正内容然后最终确定您的输出。
4. 处理特殊情况以下特殊情况应予以解决：- 填充停顿：像 “UTF8gkai 呃 ”、“UTF8gkai 额 ” 和 “UTF8gkai 嗯 ” 等词被视为填充词，应在错误分析期间忽略。不要将这些报告为语法错误。- 重复短语、错误启动和自我纠正：只分析句子的最终输出。忽略因重复或自我纠正而产生的任何多余部分。
5. 输出格式对于每个检测到的错误，使用以下格式输出一个条目：[sentence_id, start_index, end_index, error_type, corrected_text, confidence] - sentence_id: 在分析中对句子（或文本片段）的唯一标识符。- start_index 和 end_index: 错误发生的位置，基于句子的索引。- error_type: 以下代码之一：R（冗余）、M（缺词）、S（词语选择）或 W（词序）。- corrected_text: 建议的修正。- confidence: 表示您确定性的 0 到 1 之间的数值。
6. 多个错误请注意，一个句子或文本段落可能包含多个错误。在这种情况下，将每个错误作为单独的条目输出。
7. 说明性示例 - 示例 1：简单冗余校正
   - 输入: UTF8gkai 我昨天去学校学校了。
   - 预期输出: [1, 6, 7, R, UTF8gkai 学校 , 0.95]
   - 推理: 语句中 “UTF8gkai 了 ” 重复不必要（位置 6–7）。额外的 “UTF8gkai 了 ” 应被移除。由于不含糊的冗余，给出较高的确定性。- 示例 2：词序校正
   - 输入: UTF8gkai 他跑得快比我还。
   - 预期输出: [2, 4, 6, W, UTF8gkai 比我还快 , 0.85]
   - 推理: 语句 “UTF8gkai 跑得快比我还 ” 排序不当。重新排序为 “UTF8gkai 比我还快 ” 符合自然的中文词序。- 示例 3：词语选择改善
   - 输入: UTF8gkai 不受监管的移民活动会造成移民进入许多危险的路线，也会遭到人口贩卖者的残忍魔爪。
   - 预期输出:
     条目 1: [3, 10, 11, S, “UTF8gkai 让 ” , 0.95]
     条目 2: [3, 25, 28, S, “UTF8gkai 移民会落入 ” , 0.90]
   - 推理: 条目 1：“UTF8gkai 造成 ” 不是合适的动词。条目 2：“UTF8gkai 遭到 ” 与 “UTF8gkai 魔爪 ” 不是自然的搭配。动词 “UTF8gkai 落入 ” 更好地传达了移民 “陷入” 人贩子（UTF8gkai 魔爪 ）手中的意思。此外，额外的副词 “UTF8gkai 也 ” 是不必要的。
- 示例 4：处理特殊情况 - 输入: UTF8gkai 呃，我觉得今天的会议，嗯，没啥大问题。
   - 预期输出: 无错误项。
   - 推理: 忽略了 “ UTF8gkai 呃 ” 或 “ UTF8gkai 嗯 ’’。只应检查经过自我修正和填充停顿后的最终措辞，以发现真正的语法问题。

D 完整特征统计

Feature	Mean		SD		Skewness		Kurtosis	
	Raw	Aug.	Raw	Aug.	Raw	Aug.	Raw	Aug.
InfoCom features								
CometKiwi	0.51	0.51	0.10	0.06	0.13	0.22	-0.53	0.82
BertScore	0.96	0.96	0.01	0.00	-0.73	-1.20	-0.32	1.56
chrF	0.11	0.11	0.02	0.02	0.14	0.16	-0.55	1.23
BLEURT-20	0.51	0.50	0.13	0.07	1.14	1.87	2.85	1.52
XCOMET	0.18	0.17	0.11	0.06	1.06	1.66	1.77	0.96
FluDel features								
NUP	34.05	34.57	14.95	15.26	0.78	0.73	1.24	2.16
MLUP	1.00	0.94	0.61	0.46	2.01	2.53	5.35	7.62
MLFP	0.35	0.35	0.14	0.08	-0.06	-0.12	0.80	5.01
NFP	15.72	15.40	8.41	6.03	0.68	0.51	0.89	1.29
MLR	16.99	17.00	2.60	1.58	0.53	0.58	1.04	2.35
PSC	197.78	196.12	55.36	34.18	0.76	0.84	0.55	0.97
PTR	0.63	0.59	0.12	0.09	0.18	0.19	-0.76	0.80
MLS	0.26	0.26	0.04	0.02	0.96	1.17	3.64	1.72
SR	1.73	1.72	0.48	0.39	0.81	0.87	1.58	1.91
AR	3.87	3.86	0.53	0.32	0.03	0.07	2.08	1.61
NRSA	3.75	3.76	2.99	1.82	1.25	1.57	1.72	0.58
NPSA	0.78	0.80	1.28	1.16	2.28	2.27	5.75	2.33
NRLFP	0.18	0.17	0.54	0.42	3.51	5.37	9.03	7.39
NRLUP	1.05	0.99	1.33	1.21	1.81	1.98	4.02	3.26
TLQual features								
NRW	1.68	1.70	0.51	0.55	0.44	0.12	1.23	2.57
NMW	2.17	2.15	0.62	0.67	-0.43	-0.30	2.26	1.95
NWSE	4.13	4.16	1.15	1.48	1.33	0.68	2.34	3.39
NWOE	0.98	1.02	0.34	0.36	0.88	0.57	1.95	2.64
MLC	16.87	16.84	2.70	2.46	0.79	0.79	1.47	5.38
MLTU	19.57	20.04	3.46	3.87	0.98	1.05	1.34	2.32
NCPS	3.69	3.68	1.28	1.64	1.49	2.35	3.58	2.64
NTPS	3.20	3.27	1.11	1.55	1.35	2.17	2.70	1.44
TOTAL_RTTR	5.41	5.45	0.94	0.81	0.18	-0.10	1.18	3.61
VO_RATIO	0.21	0.22	0.08	0.04	0.18	-0.11	1.15	2.76
VO_RTTR	2.55	2.58	0.62	0.54	-0.22	-0.70	2.01	2.23
SP_RATIO	0.22	0.23	0.09	0.11	0.81	0.68	3.94	3.50
SP_RTTR	2.54	2.52	0.60	0.52	-0.06	-0.40	-0.19	1.24
AN_RATIO	0.08	0.09	0.04	0.02	0.24	-0.14	3.08	1.12
AN_RTTR	1.48	1.51	0.65	0.47	-0.64	-1.27	2.29	1.16
AP_RATIO	0.37	0.39	0.09	0.05	-0.02	-0.48	-0.69	2.86
AP_RTTR	3.18	3.19	0.77	0.62	0.21	-0.06	3.40	1.09
CN_RATIO	0.01	0.01	0.02	0.01	1.71	1.63	-0.30	0.82
CN_RTTR	0.40	0.42	0.58	0.44	0.98	0.70	2.39	1.18
PP_RATIO	0.03	0.03	0.03	0.02	1.61	2.18	-1.45	1.32
PP_RTTR	0.67	0.71	0.56	0.39	-0.15	-0.69	4.48	2.71
PV_RATIO	0.04	0.05	0.04	0.02	1.78	2.01	5.64	2.63
PV_RTTR	0.89	0.89	0.57	0.41	-0.41	-1.05	-0.79	1.96
PC_RATIO	0.04	0.04	0.04	0.03	1.22	1.32	1.41	3.62
PC_RTTR	0.88	0.91	0.64	0.71	-0.26	-0.82	-1.01	2.78

Table 8: 原始数据和增强数据中所有提取特征的描述性统计。

在

E 模型预测错误的案例研究

Sample 47	From the original dataset; RF model True score: 6.34; Predicted score: 5.29
Key features	BLEURT: 0.66; CometKiwi: 0.62; chrF: 0.07; BERTScore: 0.97; xCOMET: 0.35
Key features (M±SD) for Score 6 samples	BLEURT (0.54±0.13); CometKiwi (0.54±0.10); chrF (0.13±0.02); BERTScore (0.96±0.01); xCOMET (0.21±0.12)
Error analysis	The model underestimates the InfoCom score of Sample 47 by 1.05. Upon examining samples within the 5.5–6.5 score range, we observe that Sample 47 exhibits a particularly low chrF score (0.07). This value is more than one standard deviation below the mean (0.11) for this feature among samples in this range. Analysis of the corresponding student transcript reveals a tendency to reorder sentence components during interpretation, though key information in the source speech is interpreted faithfully into the target language. For instance, when interpreting an "if...then..." sentence, the student processes the "then" clause before the "if" clause, which results in reduced n-gram matching and consequently a lower chrF score for this sample.

Table 9: 机器和人工评分在 InfoCom 中的显著分歧案例。

Sample 95	From the original dataset; XGBoost model True score: 4.73; Predicted score: 3.48
Sample features	NFP: 13; MLR: 20.64; MLUP: 1.18; NUP: 42; MLFP: 0.26; PSC: 185; SR: 1.53; PTR: 0.41; NRSA: 2; MLS: 0.25
Features (M±SD) for Score 5 samples	NFP (18.16±5.66); MLR (17.13±1.11); MLUP (1.02±0.11); NUP (30.4±6.73); MLFP (0.38±0.12); PSC (195.96±13.44); SR (1.72±0.25); PTR (0.45±0.24); NRSA (4.4±3.55); MLS (0.27±0.04)
Error analysis	For Sample 95, the model underestimates the FluDel score by 1.25 points. Analysis of this sample's features reveals notably high values for MLUP (1.18) and NUP (42), both approximately two standard deviations above their respective means. Also, the speech rate (1.53) is lower than the mean (1.72). Collectively, these feature values likely lead the model to interpret this sample as having more significant breakdowns and reduced speaking speed. However, qualitative examination of the corresponding student recording offers a contrasting perspective. While the student does exhibit longer and more frequent pauses than average, these disfluencies predominantly occur at boundaries between semantic units within sentences. For human rates, this placement of pauses does not hurt perceived fluency as much as within-phrase disfluencies, which may explain why the actual perceived score is higher than the model's prediction based on these automated features.

Table 10: 机器和人工评分在 FluDel 上的显著分歧案例。

Sample 62	From the original dataset; XGBoost model True score: 6.22; Predicted score: 5.01
Key features	CN_RATIO: 0; PC_RTTR: 0; MLS: 19.57; PP_RTTR: 1; SP_RTTR: 0.71; AP_RTTR: 2; MLC: 14; NWSE: 0.26; PV_RTTR: 0.89; MLTU: 17.11
Key features (MśSD) for Score 6 samples	CN_RATIO (0.01ś0.01); PC_RTTR (0.99ś0.39); MLS (21.36ś7.18); PP_RTTR (0.81ś0.29); SP_RTTR (2.54ś0.35); AP_RTTR (3.23ś0.44); MLC (17.08ś1.30); NWSE (1.69ś0.74); PV_RTTR (0.98ś0.34); MLTU (19.73ś1.56)
Error analysis	The predicted score is 1.21 points lower than that assigned by human raters. A contributing factor to this discrepancy may be the notable absence of two specific Chinese structures, CN and PC expressions, in the student's interpretation. Instead, the students frequently employ expressions characteristic of Westernized Chinese, a style influenced by Western language structures. While human raters appear to find these alternative expressions acceptable within the context of the task, the model likely penalizes the lack of the expected native Chinese forms, leading to the observed lower scores.

Table 11: 机器评分与人工评分对于 TLQual 的显著分歧案例。