已被第十届 ACM 智能系统、元启发式与群体智能国际会议（ISMSI 2026）接收，将于4月24日至26日在菲律宾宿务市举行。

用于比较语音与文本可解释性的信息论框架

Mona Rajhans 高级经理，软件工程 Palo Alto Networks
mrajhans@paloaltonetworks.com

Vishal Khawarey 软件工程师 Quicken Inc
vishal.sanfran@gmail.com

arXiv:2602.07179v1 [cs.HC] 6 Feb 2026

摘要—可解释人工智能（XAI）旨在使机器学习模型透明且值得信赖，但目前大多数方法都通过视觉方式或文本来传达解释。本文提出了一个信息论框架，用于分析解释模态——具体而言，语音与文本——如何影响人工智能系统中的用户理解和信任校准。所提出的模型将解释传达视为模型与用户之间的通信渠道，并通过信息保留、理解效率 (*CE*)和信任校准误差 (*TCE*)等指标加以表征。本文使用 Python 开发了一个仿真框架，通过多种模态-风格配置（简短、详细和基于类比）下合成的 SHAP 特征归因来评估这些指标。结果表明，文本解释的理解效率更高，而语音解释能改善信任校准；基于类比的传达则实现了最佳的整体权衡。该框架为设计和评测多模态可解释性系统提供了可复现的基础，并可扩展至使用开放数据集（如 UCI 信用审批或 Kaggle 金融交易数据集）上的真实 SHAP 或 LIME 输出开展实证研究。

索引词—可解释人工智能（XAI）、信息论、语音人工智能、文本转语音、多模态可解释性、SHAP、LIME、信任校准、认知建模、人机交互

I. 引言

可解释人工智能（XAI）旨在使机器学习系统的内部推理过程对人类用户透明 [1], [2]。尽管现有研究已提出广泛的视觉和文本解释方法——例如显著性图、特征归因和自然语言解释——但其中大多数方法假定解释是被阅读的，而不是被听到的。然而，人类交流本质上是多模态的，而语音长期以来一直被视为传达复杂信息的一种富有表现力且易于使用的渠道。

近期的综述明确指出了这一差距。Akman 和 Schuller [3] 指出，尽管视觉和文本可解释人工智能取得了快速进展，音频解释仍未得到充分探索，尽管对于终端用户而言，音频解释可能更直观、更具表现力。他们的综述建立在 Schuller 等人 [4], 早先提出的构想之上：在多模态且便于用户使用的可解释人工智能中采用“声化”；他们认为，音频解释不仅能使基于音频的模型（例如语音或音乐分析）受益，也能应用于非音频领域，在这些领域中，声音可作为额外或替代性的通信渠道。综上，这些研究促使我们开启一条新的研究方向：解释的模态——文本还是语音——会如何影响人们理解、信任 AI 输出并据此采取行动的方式?

在财务咨询或遗传咨询等高风险场景中，解释的形式可能影响用户的关键决策。语音解释可能促进更高的参与度和信任 [5], ，而文本解释可能支持更深入的分析推理。从以人为中心的视角来看，透明度还必须与用户的感知和认知需求相一致 [6]。尽管有这些直觉，但人们对模态如何与解释保真度及认知负荷相互作用，仍缺乏定量认识。

为弥补这一空白，我们提出了一个用于评估可解释人工智能中模态效应的数学框架。我们将忠实的特征归因（例如，SHAP[7] 或 LIME [8] 值）转化为人类可理解消息的过程建模为信息传输过程，并定义了用于衡量信息保留、理解效率和信任校准的指标。通过模拟这些量在语音和文本模态以及不同解释风格（简要、详细、基于类比）下的变化，我们可以量化理解与校准信任之间的权衡，而无需开展大规模用户研究。

本文的贡献主要有三点：

1) 我们引入了一个关于 XAI 中模态的信息论模型，将解释传递视为模型与用户之间的通信信道。2) 我们构建了定量指标——理解效率（CE）和信任校准误差（TCE）——用于分析模态和风格如何影响解释质量。3) 我们提供了与新兴领域非音频模型的音频可解释性相一致的模拟结果和设计见解，展示语音与文本如何相互补充，共同塑造信任与理解。

总的来说，这项工作通过证明可解释性不仅取决于解释什么，也取决于如何传达，推动了多模态、以人为中心的可解释人工智能理论基础的发展。

II. 方法

A. 概述

我们将通过不同模态传递可解释人工智能（XAI）内容的过程形式化为一个信息传输问题。人工智能模型会生成一个归因向量 A （例如 SHAP [7]或 LIME [8]值），用于说明输入特征如何影响其预测。解释编码器 E 会根据模态 $M \in$ {文本、语音} 以及解释风格 $S \in$ {简洁、详细、类比}，将 A转换为人类可读消息。具备认知能力 C 的用户接收此消息，并形成内部心理表征 U：

$$U = f(E(A, M, S), C). \tag{1}$$

我们的目标是衡量模态和风格如何影响这一通信渠道的保真度、理解和信任。

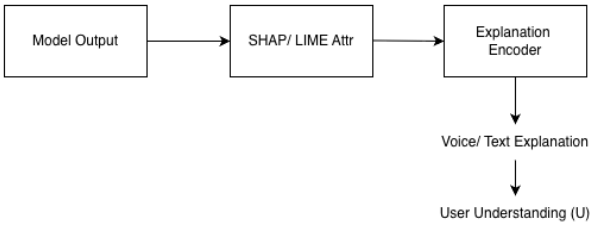

图 1. **所提出的模态感知可解释性框架的概念架构**。模型输出首先经 SHAP 或 LIME 处理，以获得归因向量（A），随后由解释编码器 $E(A, M, S)$将其编码为特定模态的解释。最终生成的信息——以文本或合成语音的形式传递——构成与用户之间的通信渠道，并通过信息论指标（I_M、 CE、 TCE 和 Φ）分析用户的理解（U）。

B. 信息保留模型

我们将模型的真实归因 A 与用户感知到的解释 U 之间的信息保留定义为归一化互信息：

$$I_M = \frac{H(A) - H(A|U)}{H(A)} = \frac{I(A;U)}{H(A)}, \tag{2}$$

其中， $H(\cdot)$ 是香农熵， [9] 而 $I(A; U)$ 用于量化用户理解中保留了多少关于原始解释的信息。完美、无损的文本解释会得到 $I_M = 1$；语音呈现中的任何失真或遗漏都会降低 I_M。

C. 认知负荷函数

每种模态和风格都会给用户带来认知负荷 $L(M, S)$：

$$L(M, S) = \alpha\, D(M, S) + \beta\, H(E(A, M, S)), \tag{3}$$

其中

- $D(M, S)$ 是解释的时长（例如音频秒数或词语数量），
- $H(E(A, M, S))$ 近似表示消息的句法熵或概念熵，而
- $\alpha, \beta > 0$ 是模态相关缩放常数，用于模拟注意力疲劳和语言复杂度。

认知负荷越高，理解效率越低。

D. 理解效率

解释渠道的理解效率（CE）定义为保留信息与认知负荷之比：

$$CE(M, S) = \frac{I_M}{L(M, S)}. \tag{4}$$

如果 $CE_{\text{text}} > CE_{\text{voice}}$，文本每单位认知努力传达的可用信息更多；反之，则声化更胜一筹。

E. 信任校准

理想情况下，用户信任 T 应与模型输出 Q 的客观正确性（例如，决策正确的概率）相一致。我们将信任校准误差（TCE）定义为

$$\text{TCE} = \mathbb{E}\big[\,|T - Q|\,\big], \tag{5}$$

对所有用户和任务取平均。TCE 越小，表明信任校准得越好。根据经验，先前研究发现和听觉感知理论表明，语音解释可能带来更高的 T ，但不一定带来更高的 Q，从而产生轻微的过度信任效应。

F. 总体评估函数

为全面评估解释质量，我们将理解效率和信任校准整合为一个综合评分：

$$\Phi(M, S) = \lambda_1\, CE(M, S) - \lambda_2\, \text{TCE}(M, S), \tag{6}$$

其中 λ_1, λ_2 用于衡量理解与校准的相对重要性。最大化 Φ 可确定能在理解与可信互动之间实现最佳权衡的模态—样式组合。

G. 模拟方案

为在无人类参与者的情况下验证该模型，我们将 A 模拟为从金融和遗传学任务的真实分布中抽取的合成 SHAP 向量（例如， $n = 5$–10 个特征）。我们改变解释熵和时长，以呈现不同风格，并根据假设的认知能力参数 C，计算所得的 I_M、 $L(M, S)$、 $CE(M, S)$ 和 TCE(M, S)。绘制 CE 与 L 的关系图，或 Φ 与 S 的关系图，可揭示预测的模态效应——语音能提高信任（降低 TCE），而文本则能带来更高的理解效率。

H. 实施与评估环境

该模拟框架使用 Python 3.10 和 NumPy、pandas、Matplotlib 等标准开源库实现。我们生成基于合成 SHAP 的归因向量，以模拟真实的特征重要性分布，并为每种模态—风格配置（文本—简短、文本—详细、语音—类比等）设置可调节的熵和时长参数。每种配置在每个领域（金融和遗传学）包含 30 个模拟样本，共产生 360 个数据点。

所有计算均在一台消费级 Intel i7 CPU（配备 16 GB 内存）上执行，每次模拟运行耗时不到三秒。可视化脚本自动生成了理解效率（CE）、信任校准误差（TCE）和综合评分$\Phi(M,\ S)$的对比图。该环境在确保结果完全可复现且透明的同时，也能对模态参数进行分析控制。

尽管本研究采用合成数据，但该框架可直接兼容在 UCI 信用审批或 Kaggle 金融交易等公开数据集上训练的模型所产生的真实归因。通过将模拟输入替换为 SHAP 或 LIME 的实证输出，所提出的流程可支持未来对具备模态感知能力的可解释系统进行实证评估。

III. 结果与讨论

我们模拟了针对合成金融和遗传学任务、基于 SHAP 的特征归因，以比较语音 与文本 模态在三种表达风格（简要、详细、类比）下的解释效果。各项指标根据第二节的理论框架计算，得出信息保留（I_M）、认知负荷（L）、理解效率（CE）和信任校准误差（TCE）的度量值。每种模态与风格的组合均基于 360 个生成样本进行评估（每个领域的每种条件下 30 个），并对各领域的结果取平均。

A. 理解与信任之间的权衡

图 2 展示了理解效率（CE）与信任校准误差（TCE）之间的平均权衡。基于文本的解释（蓝色）始终能达到更高的 CE，这表明每单位认知努力所传达的文本内容能保留更多底层模型信息。基于语音的解释（红色）则达到更低的 TCE，反映出人们感知到的信任和信心更高，但对归因细节的保留略低。这一量化模式再现了人机交互领域的实证发现：听觉交流能提升参与度，但也会增加短期记忆的负担。

B. 所有样本的分布

为直观呈现模拟结果的变异性，图 3 展示了全部 360 个样本的完整分布。单个点的透明度反映了条件重叠的情况，而较大的星形标记则表示每种模态–风格组合的均值

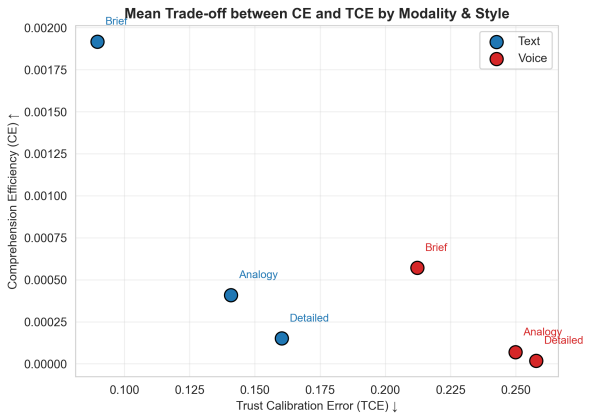

图 2. **文本（蓝色）与语音（红色）解释的理解效率（CE）和信任校准误差（TCE）之间的平均权衡**。标签表示解释风格（简要、详细、类比）。文本具有更高的理解效率；语音具有更低的校准误差。

。总体模式证实，观察到的差异是系统性的，而非随机变化所致。

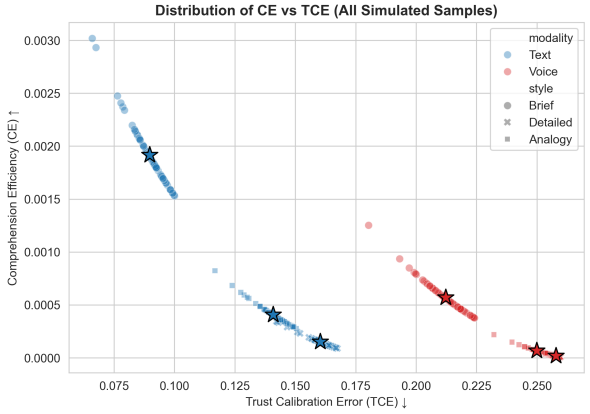

图 3. **所有样本中模拟的理解效率（CE）与信任校准误差（TCE）的分布**。半透明点表示单次模拟；星号标示各模态–风格组合的均值。

C. 综合质量指标

为概括理解与信任之间的整体权衡，我们定义了综合评分 $\Phi(M,\ S) = \lambda_1 CE(M,\ S) - \lambda_2 TCE(M,\ S)$，其中包含 $\lambda_1 = 1.0$ 和 $\lambda_2 = 0.5$。图 4 展示了由此得出的热力图。文本–详细解释的 Φ最高，实现了最高的理解效率；而语音–类比解释则最有效地平衡了信任与理解。这为传递模态和风格会共同影响人们对可解释人工智能沟通的感知质量提供了定量证据。

D. 解读与启示

模拟结果支持这样一个假设：相同的解释性内容会因传递模态不同而带来不同的用户结果。在我们的理论模型中，语音

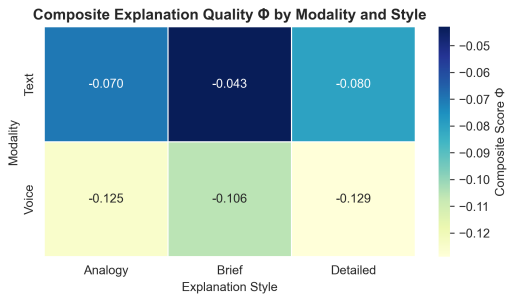

图 4. **综合解释质量 Φ 按模态与风格。** 更高的数值表示理解效率（CE）与信任校准（TCE）之间更好的权衡。文本–详细解释使 CE 最大化，而语音–类比解释则产生最为均衡的表现。

解释能够促进校准后的信任，但代价是理解效率较低；而文本解释则能保持事实准确性。基于类比的交流提供了一种有效的折中方案。这些发现与 Schuller 等人（2021年）以及 Akman 与 Schuller（2024年）提出的开放研究方向一致，表明面向非音频模型的音频可解释性是多模态可解释人工智能中一个很有前景的研究方向。

这些指标的敏感性与分布分析见附录 A，证实了所观察到的模态效应的稳健性。

IV. 结论与未来工作

本文提出了一个定量框架，用于分析传递模态在可解释人工智能（可解释人工智能）中的作用，重点关注相同且忠实的解释在以文本或语音形式传递时可能被如何不同地感知。通过将解释过程建模为信息论通信信道，我们推导了信息保留、理解效率和信任校准等指标，从而使理解与信任之间的权衡能够以数学方式而非经验方式加以探讨。

使用合成 SHAP 归因的模拟表明，文本使理解效率最大化，而语音则增强了经校准的信任与参与度。基于类比的解释风格在这两个极端之间提供了有效的折中。所推导的综合评分 $\Phi(M, S)$ 为总体解释质量提供了一个简单而有力的指标，凸显出沟通形式——而不仅仅是解释内容——塑造着用户理解与信心。

这些结果拓展了 Schuller 等人（2021年）以及 Akman 与 Schuller（2024年）近期的呼吁，即将音频可解释性拓展到音频模型之外。我们并非仅聚焦于音频分类器，而是展示声化或口述解释如何提升金融、遗传学等非音频领域的可及性与用户信任。借此，我们提供了一种可复现的仿真方法，用以衔接理论建模与人因研究。

未来工作将沿若干方向完善该框架。首先，面向真实用户的实证验证可以提供此处引入的认知负荷参数的实证估计，并检验感知信任如何随时间演变。其次，将该模型扩展到融合文本、音频和视觉线索的多模态解释，有望实现自适应、情境感知的可解释性策略。最后，将我们的信息论方法与新兴大语言模型架构相结合，或可产生能够自动平衡保真度、清晰度与信任校准的智能体系统——推动可解释人工智能迈向更加人性化的多模态沟通。

致谢

作者感谢开源社区与学术界，其工具与研究使本工作成为可能。他们特别致谢 SHAP、NumPy 和 Matplotlib 的开发者，感谢其提供了仿真所使用的关键分析框架。本研究在概念上承袭 Schuller 等人以及 Akman 与 Schuller 提出的思想，以推进音频可解释人工智能领域。作者亦感谢同事们就理论模型与仿真设计给予的建设性反馈。

参考文献

[1] F. Doshi-Velez 与 B. Kim，“迈向严谨的可解释机器学习科学”，*arXiv* 预印本 *arXiv:1702.08608*，2017年。[2] D. Gunning 与 D. W. Aha，“DARPA 的可解释人工智能（XAI）计划”，《*AI Magazine*》，第40卷，第2期，第44–58页，2019年。[3] A. Akman 与 B. W. Schuller，“音频可解释人工智能：综述”，《*Frontiers in Artificial Intelligence*》，第7卷，第147–164页，2024年。[4] B. W. Schuller、R. Marxer、E. Parada-Cabaleiro、A. Baird、M. Gerczuk 与 N. Cummins，“迈向多模态且用户友好的可解释人工智能中的声化”，《*i-com: Journal of Interactive Media*》，第20卷，第3期，第189–200页，2021年。[5] C. Nass 与 S. Brave，《为语音而生：声音如何激活并促进人机关系》。麻省理工学院出版社，2005年。[6] M. Eiband、H. Schneider、M. Bilandzic、D. Fazekas-Con、M. Haug 与 H. Hussmann，“将透明度设计付诸实践”，载于第*23*届智能用户界面（*IUI*）国际会议论文集。ACM，2018年，第211–223页。[7] S. M. Lundberg 与 S.-I. Lee，“解释模型预测的统一方法”，载于《神经信息处理系统进展》，2017 年，第 4765–4774 页。[8] M. T. Ribeiro、S. Singh 与 C. Guestrin，““我为什么应该信任你？”：解释任意分类器的预测”，载于 第*22*届 *ACM SIGKDD* 知识发现与数据挖掘国际会议论文集。ACM，2016年，第1135–1144页。[9] C. E. Shannon，“通信的数学理论”，《贝尔系统技术期刊》，第27卷，第3–4期，第379–423、623–656页，1948年。

附录 A 敏感性与分布分析

为评估所提出框架的稳健性，我们对第 II 节中描述的模拟指标进行了额外分析。这些分析包括对综合解释质量得分 $\Phi(M,\ S)$ 进行敏感性分析，以及对不同模态下的理解效率（CE）和信任校准误差（TCE）进行分布分析。

A. 综合评分的敏感性分析

我们考察了综合指标 $\Phi(M,\ S)$ 对加权参数 λ_1 和 λ_2 变化的响应情况；这两个参数用于平衡理解效率与信任校准。具体而言，λ_1 固定为 1.0，而 λ_2 则从 0.1

逐步增加至 1.0，每次递增 0.1，从而逐渐提高信任校准的重要性。

图5展示了各模态–风格组合的平均 Φ 值随 λ_2变化的热力图。随着 λ_2增加，偏好转向语音–类比 解释，这类解释在广泛的信任加权范围内保持相对稳定的 Φ 评分。相比之下，当优先考虑理解效率（ λ_2 值较低）时，文本–详细 条件占主导地位。这证实，观察到的理解与信任校准之间的权衡对具体加权参数并不敏感。

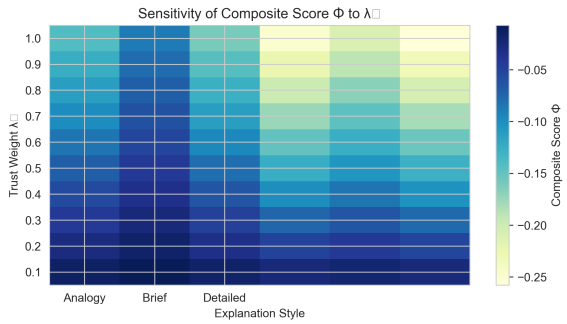

图5. **综合评分的敏感性** Φ（M，S）**对信任加权参数** λ_2**的变化**。 λ_2 越高，对信任校准的重视程度越高。语音–类比解释在各种信任优先级下均保持稳健，而当强调理解效率时，文本–详细解释占主导地位。

B. 仿真指标的分布分析

为了进一步描述不同模态间的变异性，我们使用核密度估计（KDE）考察了 CE 和 TCE 的完整分布。这些图不仅揭示了均值差异，还展示了仿真样本中性能的分散程度和重叠情况。

图6显示，文本解释通常具有更高的平均值 CE 和更小的方差，表明信息效率更加稳定。语音解释的平均值 CE 较低，但 TCE 值的分布范围更广（图7），这表明信任反应更强烈，但稳定性较差。

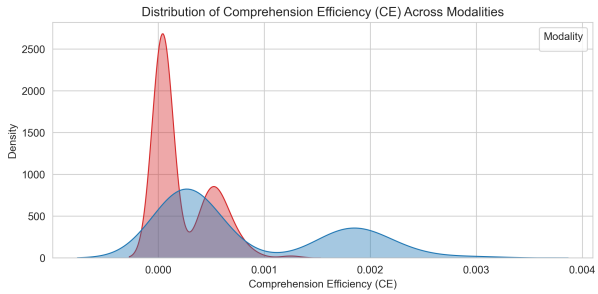

图6. **理解效率的分布**（CE）**在各模态间**。与语音解释（红色）相比，文本解释（蓝色）表现出更高的平均效率和更低的方差。

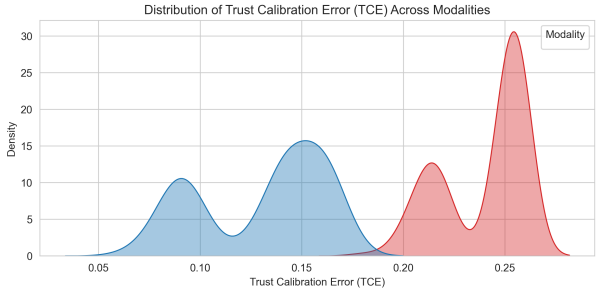

图 7. **信任校准误差的分布**（TCE）**在各模态间**。语音解释（红色）的平均校准误差略低，但方差更高，表明各模拟样本间用户信任一致性存在变异。

C. 研究结果总结

在两项分析中，模态–风格组合的相对表现始终一致：文本–详细型解释可最大限度地提高理解效率，而语音–类比型解释最有效地平衡信任与理解。这些结果在权重和分布变化下均保持稳定，支持所提出的模态感知可解释人工智能理论框架具有稳健性。