A Meta-Methodology for User Evaluation of Artificial Intelligence Generated Music; Using the Analytical Hierarchy Process, Likert and Emotional State Estimations

人工智能生成音乐用户评估的元方法；使用层次分析法，李克特和情绪状态估计

Miguel Civit, Véronique Drai-Zerbib, Francisco Cuadrado & Maria J. Escalona

Miguel Civit，Véronique Drai-Zerbib，Francisco Cuadrado & Maria j

To cite this article: Miguel Civit, Véronique Drai-Zerbib, Francisco Cuadrado & Maria J. Escalona (24 Mar 2025): A Meta-Methodology for User Evaluation of Artificial Intelligence Generated Music; Using the Analytical Hierarchy Process, Likert and Emotional State Estimations, International Journal of Human-Computer Interaction, DOI: 10.1080/10447318.2025.2478265

引用这篇文章： Miguel Civit，Véronique Drai-Zerbib，Francisco Cuadrado & Maria j。 Escalona (2025 年 3 月 24 日) : 人工智能生成音乐用户评估的元方法论；使用层次分析法，李克特和情绪状态估计，国际人机交互杂志，DOI: 10.1080.10447318.2025.2478265

To link to this article: https://doi.org/10.1080/10447318.2025.2478265

链接本文： https://doi.org/10.1080.10447318.2025.2478265

A Meta-Methodology for User Evaluation of Artificial Intelligence Generated Music; Using the Analytical Hierarchy Process, Likert and Emotional State Estimations

人工智能生成音乐用户评估的元方法；使用层次分析法，李克特和情绪状态估计

Miguel Civit ❑[a] (C), Véronique Drai-Zerbib ❑[b] (C), Francisco Cuadrado ❑[a] (C), and Maria J. Escalona ❑[c] (D) Department of Communication and Education, Universidad Loyola Andalucía, Seville, Spain; LEAD - CNRS UMR5022 Université Bourgogne Institut Marey, Dijon, France; ’Department of Computer Languages and Systems, E.T.S. Ingeniería Informática, Avda. Reina Mercedes sín, Universidad de Sevilla, Seville, Spain

Miguel Civit ❑[a] (c) ，Véronique Drai-Zerbib ❑[b] (c) ，Francisco Cuadrado ❑[a] (c) 和 Maria j. Escalona ❑[c] (d) 西班牙塞维利亚 Loyola Andalucía 大学通信和教育系；LEAD-CNRS umr5022 法国第戎勃艮第大学 Marey 研究所；计算机语言与系统系，e.t.s. Ingeniería Informática，Avda◦Reina Mercedes sín，西班牙塞维利亚大学

ABSTRACT

摘要

Artificial Intelligence (AI) music generation is a trending field, and many different generators are currently under development. However, no standardized evaluation method exists that can help researchers evaluate and compare AI-based music tools. To create a meta-methodology for AI music assessment based on user evaluation, that can be both standardized and deployed as a tailored implementation model adapted to the idiosyncrasies of specific generators and their intended applications, thereby helping future researchers draw comparisons between different systems. Two different decision trees/matrices are proposed to help researchers tailor their specific evaluation studies. As evaluation tools, the paper explores Likert and analytical hierarchy process (AHP) based surveys and emotional state estimations using facial action units, self-assessment, and physiological signals. A proof-of-concept study demonstrates the viability of the proposed tools for user-based AI music generation evaluation studies. A preference for audio music over symbolic music generation was observed, and this will require future research. The implementation of the proposed methodology and tools across the field will be helpful when comparing different systems in future research and to save time in the development of user-based studies. User-based evaluation studies are needed to prevent biases from passing into future iterations of AI music generators.

人工智能 (AI) 音乐生成是一个趋势领域，许多不同的生成器目前正在开发中。然而，目前还没有标准化的评估方法可以帮助研究人员对基于人工智能的音乐工具进行评估和比较。为了创建基于用户评估的 AI 音乐评估的元方法，可以将其标准化并部署为适应特定生成器及其预期应用的特性的量身定制的实现模型，从而帮助未来的研究人员在不同系统之间进行比较。提出了两种不同的决策树 / 矩阵来帮助研究人员定制他们的具体评估研究。作为评估工具，本文探讨了基于李克特和层次分析法 (AHP) 的调查和情绪状态估计使用面部动作单位，自我评估和生理信号。概念验证研究表明，所提出的工具用于基于用户的人工智能音乐生成评估研究是可行的。观察到音频音乐优于符号音乐生成，这将需要未来的研究。建议的方法和工具在整个领域的实施将有助于比较不同的系统在未来的研究和节省时间在发展的用户为基础的研究。需要基于用户的评估研究来防止偏差传递到人工智能音乐发生器的未来迭代中。

KEYWORDS

关键词

Automatic music

自动音乐

generation; assisted music

辅助音乐

composition; artificial

作曲；人工

intelligence; user

智能；用户

evaluation; evaluation

评价；评价

studies

研究

1. Introduction

1. 简介

Automatic music generation (AMG) is the use of artificial intelligence and machine learning techniques to create music with little to no human intervention. Many techniques have been used for AMG (Civit et al., 2022; García-Peñalvo & Vázquez-Ingelmo, 2023). Some of the most common are: probabilistic methods, such as Markov models, that generate music based on the probability distribution of musical elements in a given dataset; rule-based approaches, that use predefined algorithms or grammars to generate music and which often incorporate music theory principles; and deep learning models. Thanks to advances in this technology, it has been possible to develop different music generation models, including recurrent neural networks (RNNs), long-short-term memory (LSTM) networks, generative adversarial networks (GANs) and, in recent years, transformer-based or large language model (LLM) architectures.

自动音乐生成 (AMG) 是利用人工智能和机器学习技术来创作音乐，几乎不需要人工干预。许多技术已被用于 AMG (Civit 等，2022; García-Peñalvo & Vázquez-Ingelmo，2023)。一些最常见的是：概率方法，如马尔可夫模型，基于给定数据集中音乐元素的概率分布生成音乐；基于规则的方法，使用预定义的算法或语法来生成音乐，通常结合音乐理论原则；以及深度学习模型。由于这项技术的进步，已经有可能开发出不同的音乐生成模型，包括循环神经网络 (RNNs)、长短期记忆 (LSTM) 网络、生成对抗网络 (GANs) 以及近年来的基于变压器或大型语言模型 (LLM) 架构。

These generation systems can create music either in the symbolic domain, i.e., the system output is equivalent to a musical score or in raw audio format, where the output is produced directly as a sound file. Until very recently, most generators were symbolic, and they produced outputs of higher quality than those created by audio-based generators. With the use of high-performance encoders and better language models, audio-based music generators have become a major trending alternative.

这些生成系统可以在符号域中创建音乐，即，系统输出相当于一个乐谱或原始音频格式，其中输出直接作为声音文件产生。直到最近，大多数生成器都是符号化的，它们产生的输出比那些基于音频的生成器产生的输出质量更高。随着高性能编码器和更好的语言模型的使用，基于音频的音乐生成器已经成为一个主要的趋势替代品。

The fast-evolving AMG technology can potentially offer many benefits and applications across different domains. It can, for example, enable musicians and composers to explore new ideas and musical possibilities, adapt dynamically to coincide with on-screen action in games and visual media, or democratize music creation by helping individuals with different disabilities to create their own compositions.

快速发展的 AMG 技术可以在不同领域提供许多潜在的好处和应用。例如，它可以使音乐家和作曲家探索新的想法和音乐的可能性，动态地适应游戏和视觉媒体中的屏幕动作，或者通过帮助不同的残疾人创作他们自己的作品，使音乐创作民主化。

However, despite such significant progress, it is clear that many challenges remain. As research continues, further advances can be expected in generation and evaluation techniques, along with improvements in the quality of generated music. The most common problems related to AMG include (Briot & Pachet, 2020):

然而，尽管取得了如此显著的进步，很明显，许多挑战依然存在。随着研究的继续，生成和评估技术有望取得进一步的进展，生成音乐的质量也会得到改善。AMG 最常见的问题包括 (Briot & Pachet，2020) :

- Control: Ensuring that the generated music conforms to specific tonality, rhythm, and other musical constraints can be difficult, especially for deep learning based models.

控制： 确保生成的音乐符合特定的调性、节奏和其他音乐约束可能是困难的，特别是对于基于深度学习的模型。

- Structure: Generated music often lacks a sense of direction or coherent structure, resulting in music without a clear narrative.

结构： 生成的音乐往往缺乏方向感或连贯的结构，导致音乐没有一个清晰的叙事。

- Creativity: The balance between imitation and originality is crucial to avoid plagiarism and generate new music.

创造力： 模仿和原创之间的平衡对于避免剽窃和创作新音乐至关重要。

- Interactivity: Allowing human composers to influence the music generation process is critical to ensure the uptake of AMG systems and their integration in music creation tools.

交互性： 允许人类作曲家影响音乐生成过程，对于确保 AMG 系统的吸收和它们在音乐创作工具中的整合至关重要。

- Evaluation: Assessing the quality of generated music is a complex task. As objective metrics are not able to capture all the intricacies of music, subjective evaluation methods, such as listening tests, are required to complement them. This challenge will be discussed in depth in the remainder of the paper.

评估： 评估生成音乐的质量是一项复杂的任务。由于客观的指标无法涵盖音乐的所有细节，因此需要主观的评估方法，例如听力测试，来补充这些指标。这一挑战将在本文的其余部分进行深入讨论。

- Copyright: The determination of copyright and ownership of machine-generated music is complex. In many cases AMG is used to produce public domain results but, when this is not the case, copyright ownership has to be clearly established (Frid et al., 2020).

版权： 机器生成音乐的版权和所有权的确定是复杂的。在许多情况下，AMG 是用来产生公共领域的结果，但是，当这不是情况下，版权所有权必须明确建立 (Frid 等人，2020)。

This work aims to enhance the research and resources available for addressing the evaluation problem. As AI-generated content in the arts continues to advance, the demand for standardized and easily implementable evaluation methodologies grows. Furthermore, the potential use of these AI systems in education and model training highlights the need for robust safeguards to ensure the accuracy and reliability of generated content before it is disseminated more broadly. To address this, we propose a clear three-step process for developing and implementing a user-based evaluation methodology for automatic music generators, adaptable to researchers’ specific needs.

这项工作旨在加强研究和资源可用于解决评估问题。随着艺术领域中人工智能生成内容的不断发展，对标准化和易于实施的评估方法的需求也在不断增长。此外，这些人工智能系统在教育和模型培训中的潜在用途突出表明，需要有强有力的保障措施，以确保生成的内容在更广泛地传播之前的准确性和可靠性。为了解决这个问题，我们提出了一个明确的三步骤过程，用于开发和实施基于用户的自动音乐生成器评估方法，以适应研究人员的具体需求。

When assessing AMG generators and the music they produce, a broad range of characteristics must be considered. Key evaluation criteria include the quality of the generated music, user acceptability, comprehensibility, structural coherence, emotional expressiveness, engagement, stressfulness, and applicability to specific use cases (although this can be expressed in a myriad of formulations). These aspects can be evaluated through objective methods, subjective assessments, or a combination of both.

在评估 AMG 发生器及其产生的音乐时，必须考虑广泛的特征。关键的评估标准包括生成音乐的质量，用户可接受性，可理解性，结构连贯性，情感表达，参与度，压力，以及对特定用例的适用性 (尽管这可以用无数的公式来表达)。这些方面可以通过客观方法，主观评估，或两者的结合进行评估。

1.1. Related works

1.1 相关著作

Numerous studies have explored specific tools for music generation and evaluation, yet comprehensive and standardized approaches to assess computer-generated music remain scarce. Previous methodologies can be broadly categorized into objective and subjective evaluation frameworks, each with distinct limitations and potential applicability to current challenges.

大量的研究已经探索了音乐生成和评估的具体工具，但是评估计算机生成的音乐的全面和标准化的方法仍然很少。以前的方法可以大致分为客观和主观评估框架，每个框架都有明显的局限性和对当前挑战的潜在适用性。

- Objective evaluation employs quantitative metrics to measure the quality of generated music. Cosine similarity, for instance, evaluates the resemblance of generated sequences to human-composed ones (Chuan et al., 2020). Similarly, Kullback-Leibler (KL) divergence estimates the difference between probability distributions of AI-generated and human-composed music (Jiang et al., 2020). While these metrics are helpful for quantifying quality, they often fail to capture subjective listening experiences. Other tools, such as Fréchet Audio Distance (FAD), assess perceptual quality using pre-trained deep learning models (Kilgour et al., 2019). FAD is particularly effective for general audio evaluation but requires substantial audio data for meaningful results. Additionally, symbolic music evaluation frameworks, such as heuristics based on music theory (Dervakos et al., 2021) and toolkits like MusicPy (Dong et al., 2020), focus on pitch, rhythm, harmony, and style. However, their relevance to non-symbolic music remains questionable due to their dependence on automatic transcription systems (Bittner et al., 2022). A further interesting approach, also for the symbolic music domain, is presented in Yang and Lerch (2020) that tries to create a combination of different objective metrics that provides results that try to be nearer to what human experts would expect of subjective evaluation results. This work also highlights the challenges of balancing subjective and objective approaches to achieve comprehensive assessments

客观评价采用定量指标来衡量生成音乐的质量。例如，余弦距离可以评估生成的序列与人工合成序列的相似性 (Chuan 等人，2020)。同样，Kullback-Leibler (KL) 散度估计了人工智能生成的音乐和人工创作的音乐的概率分布之间的差异 (Jiang et al。虽然这些指标有助于量化质量，但它们往往无法捕捉主观的听觉体验。其他工具，如 Fréchet 音频距离 (FAD) ，使用预先训练的深度学习模型评估感知质量 (Kilgour 等，2019)。FAD 对于一般的音频评估特别有效，但需要大量的音频数据才能得到有意义的结果。此外，象征性的音乐评估框架，如基于音乐理论的启发式 (Dervakos 等，2021) 和 MusicPy 等工具包 (Dong 等，2020) ，侧重于音高，节奏，和声和风格。然而，由于它们依赖于自动转录系统，它们与非象征性音乐的相关性仍然值得怀疑 (Bittner 等，2022)。Yang 和 Lerch (2020) 提出了另一种有趣的方法，也适用于符号音乐领域，试图创建不同客观指标的组合，提供试图更接近人类专家对主观评估结果的期望的结果。这项工作也突出了平衡主观和客观方法来实现综合评估的挑战

- Subjective evaluation involves human listeners assessing AI-generated music based on their preferences and perceptions. This approach typically uses various survey-based tools, such as Likert scales (Chu et al., 2022) and others. A common test is the usually referenced as musical Turing tests (e.g., Hernández-Orallo (2020)) which tries to discriminate human-made music from Ai-generated music. As Ariza (2009) comprehensively explains the musical Turing test is a misused term which should be replaced for "Musical Directive Toy Test" (MDtT) or "Musical Output Toy Test" (MOtT) depending on the scenario. These subjective evaluations can be used to evaluate parameters like creativity, structure, and emotional impact. While subjective evaluations provide invaluable insights into listener experiences, they often lack standardization and scalability. Hybrid approaches have also emerged, combining objective metrics with subjective user feedback. For example, Jordanous (2012) proposes a methodology to evaluate systems for creativity based on finding the most widely used concepts in creativity description (Social Interaction and Communication, Interaction and Emotional Involvement and Domain Competence...) and provide a weight for these components based on the importance given by evaluators and their musical experience. Later Four AMG improvisation generators are evaluated using a 21 point Likert scale for each component and a weighted average for each is provided.

主观评估涉及人类听众根据他们的偏好和感知评估人工智能生成的音乐。这种方法通常使用各种基于调查的工具，如李克特量表 (Chu et al。 ，2022) 等。一个常见的测试是通常被称为音乐图灵测试 (例如，Hernández-Orallo (2020)) ，它试图区分人造音乐和人工智能生成的音乐。正如 Ariza (2009) 全面解释的那样，音乐图灵测试是一个被误用的术语，应该根据场景用 “音乐指令玩具测试”(MDtT) 或 “音乐输出玩具测试”(MOtT) 来代替。这些主观评价可以用来评估参数，如创造力、结构和情感影响。虽然主观评价为听众体验提供了宝贵的洞察力，但它们往往缺乏标准化和可扩展性。混合方法也已经出现，结合客观指标和主观用户反馈。例如，Jordanous (2012) 提出了一种评估创造力系统的方法论，该方法论基于发现创造力描述中最广泛使用的概念 (社会互动与交流，互动与情感参与和领域能力...) ，并根据评估者给出的重要性和他们的音乐经验为这些组成部分提供一个权重。后来四个 AMG 即兴发生器评估使用 21 点李克特规模为每个组成部分和加权平均数提供。

- Mixed evaluation uses both objective evaluation metrics and user-based evaluation. As an example Sturm and Ben-Tal (2017) uses four different objective evaluation techniques to evaluate the transcriptions produced by a symbolic AMG system and also uses a very open expert-based evaluation in which a small set of musicians write, in free format, their opinions on the system output. Huang et al. (2021) utilized both quantitative metrics and user surveys to evaluate AI-generated music mash-ups. However, the integration of both methods remains clearly underexplored.

混合评估使用客观评估指标和基于用户的评估。例如 Sturm 和 Ben-Tal (2017) 使用了四种不同的客观评估技术来评估由符号化的 AMG 系统产生的转录，也使用了一个非常开放的基于专家的评估，其中一小组音乐家以自由格式写下他们对系统输出的意见。Huang 等人 (2021) 利用定量指标和用户调查来评估 ai 生成的音乐混搭。然而，两种方法的整合仍然明显不足。

Another set of interesting works are theoretical frameworks that discuss mainly the theoretical aspects of artistic creativity such as Wiggins (2019). In this case it is applied to development of music from the 10th to the 20th century with a very formal approach. These approaches are beyond the current state of development of AMG.

另一组有趣的作品是理论框架，主要讨论艺术创造力的理论方面，如威金斯 (2019)。在这种情况下，它被应用到从 10 世纪到 20 世纪的音乐发展中，并且是以一种非常正式的方式。这些方法超越了 AMG 目前的发展状态。

While some works like Jordanous (2012) present evaluation methodologies most of the discussed works are not based on a methodological approach but take an ad hoc approach to evaluate specific music generators. Thus, despite the umber of work that implement generated music evaluation, a standardized methodology for evaluating these systems is still lacking.

虽然一些作品，如 Jordanous (2012) 目前的评价方法，大多数讨论的作品不是基于一个方法论的方法，但采取特别的方法来评价具体的音乐生成器。因此，尽管实现生成音乐评估的工作量很大，评估这些系统的标准化方法仍然缺乏。

1.2. Human vs objective evaluation

1.2 人工评估 vs 客观评估

A potential issue associated with evaluation studies relying on objective metrics is that, while a given system consistently yields identical results with identical inputs, it may not produce comparable outcomes with inputs that users perceive as analogous. This challenge frequently arises in art-related evaluation studies, where technically disparate inputs may exhibit a significant level of similarity grounded in shared cultural experiences, emotional responses, or additional factors.

与依靠客观衡量标准的评价研究有关的一个潜在问题是，虽然一个特定系统始终以相同的投入产生相同的结果，但它可能不会以用户认为类似的投入产生可比较的结果。这种挑战经常出现在与艺术相关的评估研究中，在这些研究中，技术上不同的输入可能表现出相当程度的相似性，这种相似性基于共同的文化经历、情感反应或其他因素。

Having a wide pool of users, an expert panel, or both to validate an AI music generator can provide a range of metrics that corroborate the generator’s actual performance in a culturally and technically informed manner. In addition, they can also add an extra layer that guarantees the validity of the objective metrics that were used to train and validate the model.

有一个广泛的用户群，一个专家小组，或两者都验证人工智能音乐生成器可以提供一系列的指标，证实生成器的实际表现在文化和技术知情的方式。此外，他们还可以添加一个额外的层，保证用于训练和验证模型的客观指标的有效性。

On the other hand, evaluation studies based on human judgments are much more expensive and more difficult to scale up than those based on objective measurements, making them unsuitable for gathering data for model training (Yang & Lerch, 2020). Even though it may be interesting to be able to distinguish between human and computer-generated music using a musical output test, it has long been apparent that these types of evaluations have very little value in real-world scenarios (Chollet, 2019).

另一方面，基于人类判断的评估研究比那些基于客观测量的评估研究更昂贵，更难以扩大规模，这使得它们不适合为模型训练收集数据 (Yang & Lerch，2020)。尽管能够使用音乐输出测试来区分人类和计算机生成的音乐可能是有趣的，但是长期以来很明显，这些类型的评估在现实世界中的价值很小 (Chollet，2019)。

1.3. Aims and objectives

1.3 目的和目标

In this study, our main objective was to develop a methodology that can be adapted to the specific needs of validation and evaluation studies in the field of AI-generated music. As explained in subsection 1.2, there are several advantages to having human participants (experts or otherwise) subjectively evaluate art-related generation. Our intention was to provide a comprehensive set of tools that can help researchers perform user-centered evaluations that can be tailored in length and which focus on their own particular objectives and interests.

在这项研究中，我们的主要目标是开发一种方法，可以适应在人工智能生成音乐领域的验证和评估研究的具体需要。正如第 1.2 节所解释的，让人类参与者 (专家或其他人) 主观评估艺术相关的一代有几个优点。我们的目的是提供一套全面的工具，可以帮助研究人员执行以用户为中心的评估，可以定制的长度和侧重于他们自己的特定目标和兴趣。

To the best of our knowledge, current research only covers specific evaluation tools or theoretical categorizations of strategies (Xiong et al., 2023), and this creates a situation in which designers of validation studies may need to read through several (and in many cases lengthy) articles before being able to create a tool-set suited to their intended evaluation.

据我们所知，目前的研究只涉及具体的评估工具或策略的理论分类 (Xiong et al。 ，2023) ，这就造成了验证研究的设计者可能需要阅读几篇 (在许多情况下是冗长的) 文章之前能够创建一个适合他们预期评估的工具集。

Finally, through a proof of concept study, our intention is to showcase the application of the proposed methodology and, more importantly, the usefulness of several of the proposed tools for the validation, evaluation and comparison of AI music generators. We hope that this study will serve to advance awareness of several validation tools and contribute to the standardization of validation studies for generative AI in the arts.

最后，通过一个概念验证研究，我们的目的是展示所提出的方法的应用，更重要的是，几个提出的工具的有用性验证，评估和比较人工智能音乐生成器。我们希望这项研究将有助于提高对几种验证工具的认识，并有助于艺术生成 AI 验证研究的标准化。

This research aims to address the critical gaps in current evaluation practices, fostering a more standardized and inclusive approach to assessing AI-generated music. By aligning with best practices in user-centered design and computational creativity (Sturm and Ben-Tal, 2017; Hernández-Orallo, 2020), this study contributes to advancing the field and ensuring its practical relevance.

这项研究旨在解决当前评估实践中的关键差距，培养一种更标准化和包容性的方法来评估 ai 生成的音乐。通过与以用户为中心的设计和计算创造性的最佳实践保持一致 (Sturm and Ben-Tal，2017; Hernández-Orallo，2020) ，这项研究有助于推动该领域的发展，并确保其实际意义。

2. Materials and methods

2. 材料和方法

We propose a methodology for tailoring evaluation studies to the specific needs of researchers. Figure 1 shows a general view of the method. Following a simple three-step process, researchers should first define their study by answering a series of questions with the help of the adaptability decision tree (see Figure 2). Secondly, they use the adaptability matrix (Figure 3) to choose the tools that are most appropriate for their study. Finally, they implement the selected range of tools following the guidelines and descriptions in the respective subsections in section 2 (Materials and Methods). If necessary, they can further familiarize themselves with the tools using the extended literature provided.

我们提出了一种方法，以便根据研究人员的具体需要来调整评估研究。图 1 显示了该方法的一般观点。遵循一个简单的三步过程，研究人员首先应该在适应性决策树的帮助下通过回答一系列问题来定义他们的研究 (参见图 2)。其次，他们使用适应性矩阵 (图 3) 来选择最适合他们研究的工具。最后，他们按照第 2 节 (材料和方法) 中相应小节的指导方针和描述实施选定的工具范围。如有必要，他们可以使用所提供的扩展文献进一步熟悉这些工具。

2.1. Evaluation study definition

评估研究的定义

As a prior step to any evaluation study, especially on AI generation systems, it is necessary to establish the purpose of the evaluation. A clear target for evaluation facilitates both the design and the implementation of the study. We identified two main categories of evaluation studies: those intended for model training, and those intended for confirmation and comparison of model performances.

作为任何评估研究的第一步，特别是关于人工智能生成系统，有必要确定评估的目的。明确的评估目标有助于研究的设计和实施。我们确定了两大类评估研究： 旨在进行模型培训的研究，以及旨在确认和比较模型表现的研究。

The first category requires highly replicable results, minimal variability in outcome, and a very high number of measurements and iterations. The use of objective measures that do not rely on costly user studies, but on mathematical formulas or automatic training is therefore a must. In this case, scalability is the most important thing when training the model, and problems such as the sensitivity of the measuring tools are secondary.

第一类需要高度可复制的结果，结果的可变性最小，以及非常高的测量和迭代次数。因此，使用不依赖于昂贵的用户研究，而是依赖于数学公式或自动培训的客观测量是必须的。在这种情况下，可扩展性是训练模型时最重要的事情，而测量工具的灵敏度等问题是次要的。

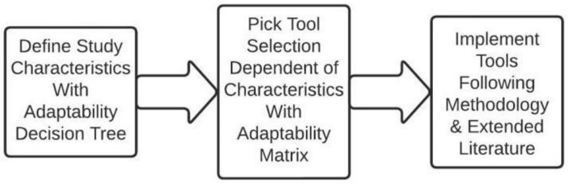
Define Study Characteristics With Adaptability Decision Tree
Pick Tool Selection Dependent of Characteristics With Adaptability Matrix
Implement Tools Following Methodology & Extended Literature

Figure 1. Meta-methodology overview.

图 1. 元方法概述。

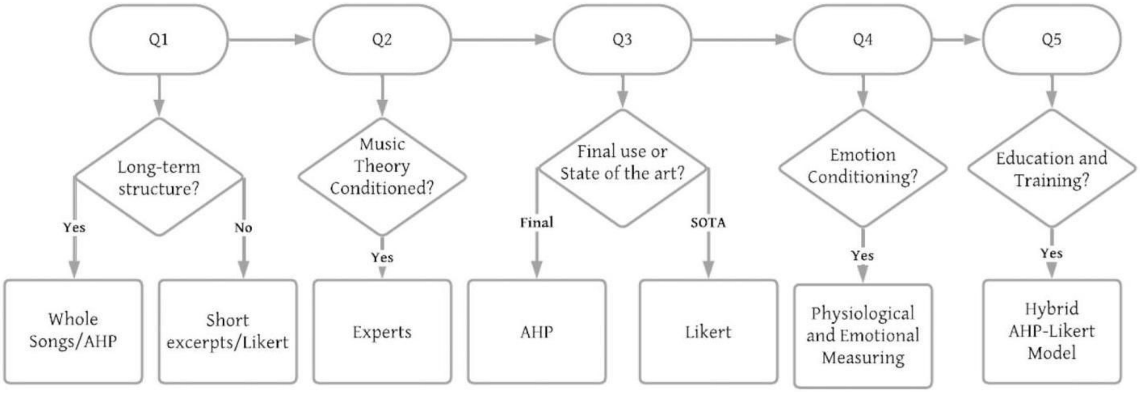

Figure 2. Adaptability decision tree. Q1. Is the generator intended to be used as a compositional tool or to produce finished pieces? Q2. Is the model conditioned? Is this conditioning dependent on music theory? Q3. Is the evaluation intended to establish the final usability of the model? Does it compare it to other state of the art systems? Q4. Is the generator intended to be able to tailor music to particular emotions? Q5. Is final quality control part of the evaluation? Is the model intended for educational purposes or for the training of other systems?.

图 2。适应性决策树。问题 1。发生器是用来作为组合工具还是用来制作成品？问题 2。模型是有条件的吗？这种条件依赖于音乐理论吗？问题 3。评估是否旨在建立模型的最终可用性？它是否与其他最先进的系统进行比较？问题 4。音乐发生器是否能够根据特定的情绪调整音乐？问题 5。最终质量控制是评估的一部分吗？该模型是为了教育目的还是为了培训其他系统？.

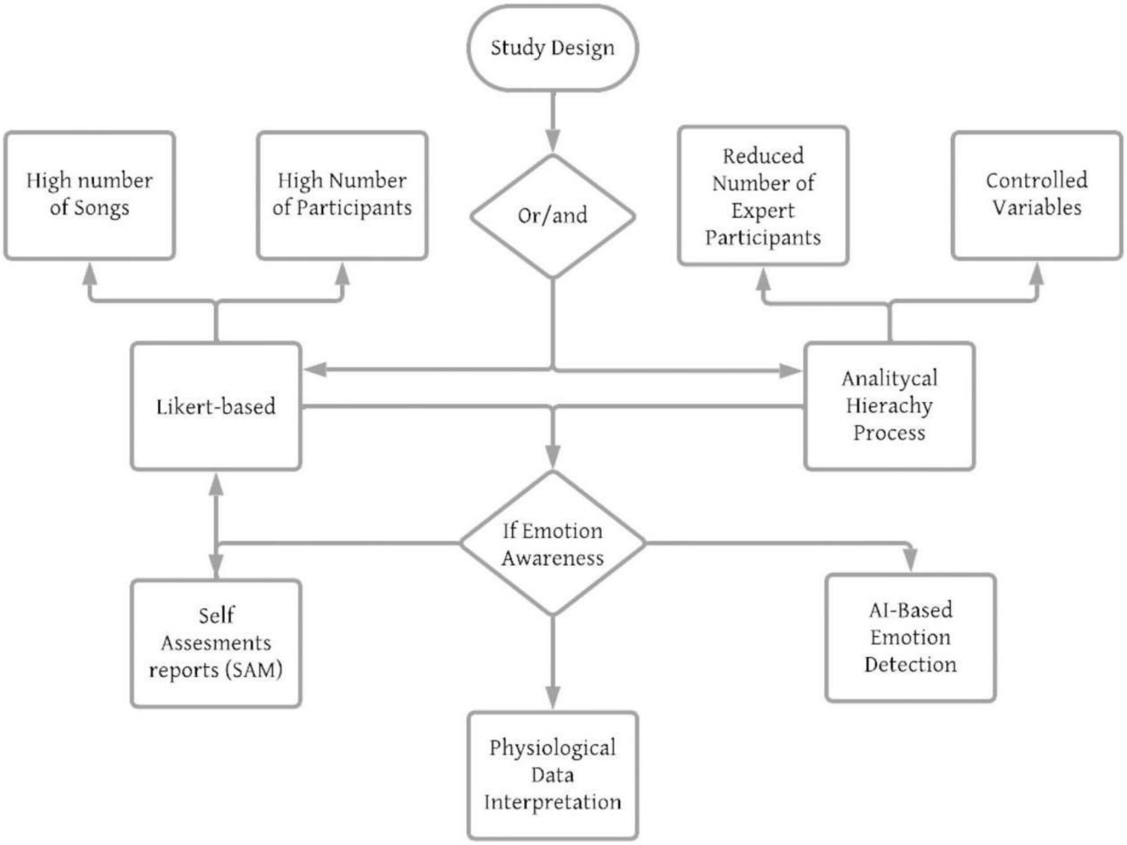
Study Design
High number of Songs
High Number of Participants
Or/and
Reduced Number of Expert Participants
Controlled Variables
Likert-based
Analitycal Hierachy Process
If Emotion Awareness
Self Assesments reports (SAM)
AI-Based Emotion Detection
Physiological Data Interpretation

Figure 3. Adaptability matrix.

图 3. 适应性矩阵。

The second category of evaluation studies can be tackled from many directions. Studies like those of Chu et al. (2022) based on surveys with 9 parameters and using a 7 point Likert scale and Ferreira et al. (2023), based on a musical output test, show how a system’s performance can be measured with many different tools with great results. In Wang et al. (2021) an objective metric for music generation is evaluated through subjective evaluation. The experiments they carried out demonstrate that objective evaluation is still less effective than subjective evaluation. This is a significant argument for continuing to develop better subjective evaluation methodologies.

第二类评估研究可以从多个方向着手。Chu 等人 (2022) 的研究基于 9 个参数的调查，使用 7 点李克特量表，Ferreira 等人 (2023) 的研究基于音乐输出测试，展示了如何使用许多不同的工具来测量系统的性能，并取得了很好的结果。在 Wang 等人 (2021) 中，通过主观评估来评估音乐产生的客观指标。他们进行的实验表明，客观评价仍然不如主观评价有效。这是继续发展更好的主观评价方法的一个重要论据。

In this article, we propose an adaptable model for designing human-based evaluation studies for the second category of the studies. This model provides the necessary tools and steps to design evaluation studies tailored to the specific needs of researchers and the music generator they are testing. Based on our previous research into the state of the art of AMG, in Civit et al. (2022), and taking into consideration several studies with different objectives and a variety of possible strategies, we designed a set of questions which, paired with a decision tree, can guide the designers of future evaluation studies in their tasks. Answering the questions in Figure 2 can help researchers define their generator and its evaluation study, and thus determine the most appropriate tools.

在本文中，我们为第二类研究提出了一个设计基于人的评价研究的适应性模型。该模型提供了设计评估研究的必要工具和步骤，以满足研究人员和他们测试的音乐发生器的特定需求。根据我们之前在 Civit 等人 (2022) 对 AMG 最新进展的研究，并考虑到几项具有不同目标和多种可能策略的研究，我们设计了一组问题，与决策树配对，可以指导未来评估研究的设计者在他们的任务中。回答图 2 中的问题可以帮助研究人员定义他们的生成器及其评估研究，从而确定最合适的工具。

The answers to the questions underline relevant characteristics that the study may require. In Q1, for example, researchers may need to use longer, complete pieces of music that require the judgment of long-term structures if they intend to generate finished pieces. This is not the case when using AI generation as a composition assistant tool, where only the quality of short excerpts of music may be relevant. Taking into account that any study made with users requires a considerable amount of time, studies related to longer pieces will benefit from the use of the analytical hierarchy process (AHP) methodology rather than the other proposed tools, because it requires more user time per piece of music, but provides significantly more data, with greater consistency, for a smaller sample pool (Saaty & Özdemir, 2014).

这些问题的答案强调了研究可能需要的相关特征。例如，在 Q1，研究人员可能需要使用更长，完整的音乐片段，需要判断的长期结构，如果他们打算生成完成的作品。但将人工智能生成作为作曲辅助工具时，情况并非如此，因为只有音乐短片段的质量才可能相关。考虑到与用户进行的任何研究都需要相当多的时间，与较长作品有关的研究将受益于使用层次分析法 (AHP) 方法，而不是其他拟议的工具，因为它需要更多的用户时间每一段音乐，但提供更多的数据，更大的一致性，为一个较小的样本池 (Saaty & özdemir，2014)。

Q2 highlights the need for the participants to be experts in music if the model is conditioned by music theory rules. The main aspects of the different evaluation methods are shown in Figure 3. The use of AHP is recommended for expert judgement as it requires a smaller pool of participants. This is a common situation when sourcing experts for evaluation studies. In Cideron et al. (2024), participants sourced via the Amazon Mechanical Turk platform, were required to have experience listening to music (over 6 years), this type of requirements are common when sourcing experienced listeners, but are much harder to accomplish when looking for expert composers or musicians. If the use of expert judgment is not required, as is the case in any general preference study that does not require music-specific knowledge, the use of Likert-based surveys could be highly beneficial. Having more than 20 participants and a good sample pool can give good results faster and with much less effort on the part of the participants. In addition, tools such as Amazon Mechanical Turk and CROWDMOS (Ribeiro et al., 2011) can be used to recruit and select a very high number of participants, producing results that are easier to generalize.

Q2 强调，如果模型受到音乐理论规则的制约，则参与者需要成为音乐专家。不同评估方法的主要方面如图 3 所示。推荐使用层次分析法进行专家判断，因为它需要较少的参与者。在为评估研究寻找专家时，这是一种常见的情况。在 Cideron et al。(2024) 中，通过 Amazon Mechanical Turk 平台采购的参与者被要求有听音乐的经验 (超过 6 年) ，这种类型的要求在采购有经验的听众时很常见，但是在寻找专业作曲家或音乐家时更难实现。如果不需要使用专家判断，就像任何不需要特定音乐知识的普遍偏好研究的情况一样，使用基于李克特的调查可能是非常有益的。拥有超过 20 个参与者和一个好的样本库可以更快地给出好的结果，而且参与者付出的努力要少得多。此外，Amazon Mechanical Turk 和 CROWDMOS (Ribeiro 等，2011) 等工具可用于招募和选择非常多的参与者，产生更容易推广的结果。

The final usability of the generation model, referenced in Q3, requires a dual approach with regard to evaluation. Firstly, many aspects of the generator have to be taken into account, from performance characteristics such as music quality evaluation to human-computer interaction characteristics such as ease of use. Secondly, the performance of the model should be measured in real-world situations, i.e., the music it generates should be compared to human-made music. With these characteristics in mind, the pairwise comparison-based AHP was considered uniquely suitable for the task, as it excels at comparing two elements (human and AI generated songs) and can take into account many variables with different statistical weightings. However, if the purpose of the study is not for the generator to be publicly released, but to advance scientific knowledge, the comparison between multiple state of the art systems and their respective generations may be too time consuming for the AHP methodology, and the much faster and simpler to use Likert-based survey will be more suitable for the task. An example of this approach can be seen in studies such as Meta (Copet et al., 2024), where a large user pool and a large sample pool from various generators are used for user evaluation with the aforementioned CROWDMOS system.

Q3 中提到的生成模型的最终可用性需要在评估方面采用双重方法。首先，生成器必须考虑很多方面，从音乐质量评价等表现特性到易用性等人机交互特性。其次，模型的表现应该在真实世界的情况下进行测量，也就是说，它生成的音乐应该与人造音乐进行比较。考虑到这些特点，基于成对比较的 AHP 被认为是唯一适合该任务的，因为它擅长比较两个元素 (人和 AI 生成的歌曲) ，并且可以考虑具有不同统计权重的许多变量。然而，如果研究的目的不是为了发生器的公开发布，而是为了提高科学知识，那么对于 AHP 方法来说，多个最先进的系统和它们各自的代之间的比较可能太耗费时间，而基于李克特的调查将更加快速和简单，更适合这项任务。在 Meta (Copet 等，2024) 等研究中可以看到这种方法的一个例子，其中使用来自各种发生器的大型用户池和大型样本池用于上述 CROWDMOS 系统的用户评估。

Many current commercial generators such as Frid et al. (2020) generate music based on emotions and music genre (e.g., happy techno or relaxing ambient music). Q4 is designed to identify whether the generator uses such features. For an affirmative answer, we propose a set of tools for self-assessment, physiological measurement, and emotion recognition that can be used with participants in the study (see subsection 2.7). These tools make it possible to measure and self-determine emotional states and should be used for both dataset tagging and study evaluations when the emotional state needs to be estimated reliably.

许多当前的商业音乐发生器，如 Frid et al。(2020) ，基于情感和音乐流派 (例如，快乐的电子音乐或轻松的环境音乐) 产生音乐。Q4 旨在确定生成器是否使用了这些特性。对于一个肯定的回答，我们提出了一套工具的自我评估，生理测量，和情绪识别，可用于与参与者的研究 (见小节 2.7)。这些工具使得测量和自我确定情绪状态成为可能，并且在需要可靠地估计情绪状态时，应该用于数据集标记和研究评估。

Q5 serves as a safeguard for future iterations of the model, future research and, in general, the passing on of misleading data to future generations. When the purpose of the model is to generate music for training other systems, or for use in some form of educational technology, there is a considerable risk that incorrect data may be passed into the future. This may have compounding effects (Osoba et al., 2017), such as misleading tagging in the dataset or generations, performance scores that do not coincide with real-world human perception. The use of the hybrid Likert-AHP methodology proposed in subsection 2.2 therefore becomes even more relevant, as the crossing of data from several different sources diminishes such risks.

Q5 为模型的未来迭代、未来的研究以及一般来说将误导性数据传递给后代提供了保障。如果该模型的目的是生成用于训练其他系统的音乐，或用于某种形式的教育技术，则存在着将不正确的数据传递到未来的相当大的风险。这可能会产生复合效应 (Osoba et al。 ，2017) ，例如在数据集或数代中进行误导性标注，性能分数与现实世界的人类感知不一致。因此，使用分节 2.2 中提出的李克特 - 层次分析法混合方法变得更加相关，因为来自几个不同来源的数据的交叉减少了这种风险。

2.2. General model and adaptability matrix

2.2 通用模型和适应性矩阵

We propose an adaptability matrix (Figure 3) that allows researchers to design different evaluation studies depending on their resources and on the ultimate purpose of the study. This methodology is intended to minimize the risks of human error-induced biases and incorrect data.

我们提出了一个适应性矩阵 (图 3) ，允许研究人员根据他们的资源和研究的最终目的设计不同的评估研究。这种方法旨在尽量减少人为错误引起的偏差和不正确的数据的风险。

After defining the evaluation study (see subsection 2.1), the choice of tools and methodologies to implement the study can be overwhelming. Whereas Figure 2, serves to define and delimit the evaluation study and can suggest some methodologies, Figure 3 shows the main tools proposed for designing a evaluation study and identifies the situations for which they are best suited.

在定义评估研究 (见 2.1 小节) 之后，实施研究的工具和方法的选择可能是压倒性的。图 2 用于定义和界定评价研究，并提出了一些方法，而图 3 显示了为设计评价研究提出的主要工具，并确定了它们最适合的情况。

Here, some important general characteristics need to be taken into account:

这里，需要考虑一些重要的一般特征：

- The possible number of participants.

参与者的可能数量。

- The time that participants will need to complete the study.

参与者完成研究所需的时间。

- The need to include expert participants.

需要包括专家参与者。

- The possibility of using a controlled environment with controlled variables.

使用受控环境和受控变量的可能性。

- The number and duration of songs.

歌曲的数量和持续时间。

- The need to measure emotional states and the time that researchers may need to interpret the data.

测量情绪状态的需要和研究人员解释数据所需的时间。

As a general guideline for each of these topics (see Figure 3) we suggest:

作为这些主题的一般指导方针 (见图 3) ，我们建议：

- Using Likert-based surveys when more participants (usually 20 or more) are available, due to its ease of use and its consistency being tied to a minimum number of participants. The possibility of using online systems for recruitment and screening is an added bonus. On the other hand, AHP is better suited to smaller studies.

当有更多的参与者 (通常是 20 个或更多) 时，使用基于李克特的调查，因为它的易用性和它的一致性与最低参与者数量有关。使用在线招聘和筛选系统的可能性是一个额外的好处。另一方面，层次分析法更适合小型研究。

- AHP surveys are much more time consuming than their Likert-based counterparts and this needs to be taken into consideration when designing the study. Nevertheless, AHP can offer more detailed comparisons between alternatives.

层次分析法调查比基于李克特的调查更耗时，在设计研究时需要考虑到这一点。然而，层次分析法可以提供更详细的替代品之间的比较。

- If expert participants are required AHP is recommended because, expert users are harder to source and it is therefore common to have a smaller reviewer pool.

如果需要专家参与，推荐使用层次分析法，因为专家用户更难寻找，因此通常会有一个较小的评论者库。

- For emotional state measurement and real-time response tracking, controlled environments can improve the overall quality and reliability of *s* study. Tools like Noldus FaceReader and HRV measuring instruments work best when conditions, such as light or sitting position, are controlled. This makes them harder to use in combination with remote evaluation protocols. If implemented, such processes need to be carefully designed and supervised.

对于情绪状态测量和实时反应跟踪，受控环境可以提高 *s* 研究的整体质量和可靠性。像诺德斯面部阅读器和 HRV 测量仪器这样的工具在光线或坐姿等条件受到控制时效果最好。这使得它们很难与远程评估协议结合使用。如果实施，这些流程需要仔细设计和监督。

- For long-term structure analysis, songs with at least 30s duration should be analysed. Such longer formats that may take up more evaluators’ time may benefit from a more thorough analysis using AHP methodology. In cases where many songs are presented, such as multi-genre generation, a faster Likert-based approach may be preferred.

对于长期结构分析，应该分析至少 30 年的时长的歌曲。这种较长的格式可能会占用更多的评估者的时间，可能会受益于使用层次分析法 (AHP) 进行更全面的分析。在有很多歌曲的情况下，例如多流派生成，基于李克特的快速方法可能是首选。

- In subsection 2.7 a range of tools are proposed for studies looking at generators with an emotion component, like those based on natural language prompts, which are able to produce "Happy folk song" type music.

在 2.7 小节中，提出了一系列工具，用于研究带有情感成分的生成器，比如那些基于自然语言提示的生成器，它们能够产生 “快乐民歌” 类型的音乐。

All these topics are addressed in greater detail in the respective sections dedicated to each tool/methodology: Likert in subsection 2.5, AHP in subsection 2.4 and emotion measurement in subsection 2.7.

所有这些主题在每个工具 / 方法的相应章节中都有更详细的介绍： 李克特在 2.5 节，层次分析法在 2.4 节，情绪测量在 2.7 节。

2.3. Samples and participants

样本和参与者

A major concern in any evaluation study is the selection of the samples and participants. When creating the sample pool, in our case pieces of music, the selection should take into account the objective of the study. Studies intended to evaluate generators which do not seek to produce any particular genre or are conditioned by theory-based criteria but merely, generate "music," may require a wide selection of musical pieces to be representative of a common generality. Music generators like DeepBach (Hadjeres et al., 2017) which are very style-specific, can be evaluated using a more representative but smaller sample pool when assessing specific criteria (for example, the quality of voice leading in a 4-part Bach-style harmony).

任何评估研究的一个主要关注点是样本和参与者的选择。当创建样本库时，在我们的案例中，选择应该考虑到研究的目的。旨在评估那些不寻求产生任何特定流派或受理论基础标准制约而仅仅产生 “音乐 “的发生器的研究，可能需要广泛选择能代表共同普遍性的音乐作品。像 DeepBach (Hadjeres 等，2017) 这样的音乐发生器是非常具体的，可以在评估具体标准时使用更具代表性但更小的样本池进行评估 (例如，声音的质量在 4 部分巴赫风格的和声)。

Another important aspect related to sample selection is commonality. To be easily compared, musical fragments should all have as many common characteristics as possible and be more or less of the same duration, regardless of whether they have lyrics. To avoid biases in the study, it is also very important for all the music in the sample pool to have the same perceived loudness. Loudness should ideally be normalized to -14 LUFS as proposed in the guidelines of Katz (2015). This normalization in volume accounts for the difference in volume that different generators or musical sources may output. It facilitates any experimental design by allowing participants to compare different musical pieces by setting a base volume for all the experiment without the need to readjust the volume of the device while changing pieces. This very conservative loudness level allows comparison of different pieces without modifying the inner dynamics of the pieces, and thus preserving musical characteristics that may be linked to emotion or preference Schubert (2004).

与样本选择相关的另一个重要方面是共性。为了便于比较，不管是否有歌词，音乐片段都应该有尽可能多的共同特征，持续时间大致相同。为了避免研究中的偏差，样本池中的所有音乐具有相同的感知响度也非常重要。按照 Katz (2015) 的指导方针，响度理想情况下应归一化为 -14 LUFS。音量的这种正常化解释了不同发生器或音乐源可能输出的音量差异。通过为所有实验设定一个基本音量，参与者可以比较不同的音乐片段，而无需在更换片段时重新调整设备的音量，这为任何实验设计提供了便利。这种非常保守的响度水平允许不同作品的比较，而不改变作品的内在动力，从而保留了可能与情感或偏好相关的音乐特征舒伯特 (2004)。

Secondly, the participant selection must be addressed taking into account the idiosyncrasy of music, both as an art form (with its technical implications) and as a cultural experience (Agawu, 2006). With this in mind, the proficiency of participants in music can be considered both from a technical point of view and from the perspective of their exposure to music: the opinion of a wide array of informed, culturally experienced listeners may be as valuable for the researchers as that of expert musicians or composers. Furthermore, when considering the emotional implications of music and its impact on listeners, it is important to address the need to establish a neutral initial emotional state to avoid possible biases. Tools such as the Profile of Mood States (POMS) (Cayrou et al., 2000) can help verify this neutrality of mood and constitute the best option for initial testing.

其次，参与者的选择必须考虑到音乐的特质，既作为一种艺术形式 (及其技术含义) ，也作为一种文化体验 (Agawu，2006)。考虑到这一点，音乐参与者的熟练程度既可以从技术角度考虑，也可以从他们接触音乐的角度考虑： 对于研究人员来说，广泛的消息灵通、有文化经验的听众的意见可能与专业音乐家或作曲家的意见一样有价值。此外，在考虑音乐的情感含义及其对听众的影响时，必须解决建立一种中立的初始情感状态的必要性，以避免可能的偏见。诸如情绪状态档案 (POMS) (Cayrou et al。 ，2000) 这样的工具可以帮助验证这种情绪的中立性，并构成初始测试的最佳选择。

2.4. Analytical hierarchy process and applicability

2.4 层次分析法和适用性

The analytic hierarchy process (AHP) is a multi-criteria decision making (MCDM) approach for structuring multiple choice criteria into a hierarchy, assessing the relative importance of those criteria, comparing alternatives for each criterion, and determining an overall ranking of the alternatives. It is particularly useful for teams working on complex problems that involve human perceptions and judgments.

层级分析法 (analytic hierarchy process，AHP) 是一种多准则决策方法 (multi-criteria decision making，MCDM) ，用于将多个选择准则构造成一个层次，评估这些准则的相对重要性，比较每个准则下的方案，并确定方案的总体排序。它特别适用于团队工作中涉及人类感知和判断的复杂问题。

AHP has been applied in several fields, including education and health, to address complex decision making and evaluation problems (Dolan, 2008). The process can be implemented with the aid of software tools like the AHP-OS online app (Goepel, 2018). AHP provides a systematic, structured approach to evaluating automatic music generators and human-composed music, or for comparing the two. Taking into account different criteria and their relative importance, and adjusting them depending on the context or in line with the preferences of the evaluators or an external group, it can be used to evaluate generators while considering many characteristics.

AHP 已经被应用在几个领域，包括教育和卫生，以解决复杂的决策制定和评估问题 (Dolan，2008)。该过程可以借助 AHP-OS 在线应用程序 (Goepel，2018) 等软件工具来实现。AHP 提供了一个系统的，结构化的方法来评估自动音乐发生器和人类作曲的音乐，或比较两者。考虑到不同的标准及其相对重要性，并根据具体情况或根据评价人员或外部群体的偏好对其进行调整，可用于评价发电机，同时考虑到许多特点。

The process follows the steps shown in Figure 4. A practical demonstration of how to implement AHP can also be found in section 3 (Results). The described implementation is illustrated in Figure 6.

流程遵循图 4 所示的步骤。一个如何实现 AHP 的实际演示也可以在第 3 节 (结果) 中找到。所描述的实现如图 6 所示。

First, it is necessary to define the criteria that are most relevant for evaluating music generators. These criteria could include musicality, creativity, originality, melody quality, harmony, rhythm, and the user-friendliness of the interface. Of the many possible criteria, particular attention should be paid to long-term structure, noise-to-signal ratio perception (in particular for non-symbolic generators), and general user preference.

首先，有必要定义与评估音乐生成器最相关的标准。这些标准包括音乐性、创造性、原创性、旋律质量、和声、节奏和界面的用户友好性。在许多可能的标准中，应该特别注意长期结构、噪声信号比感知 (特别是对于非符号发生器) 和一般用户偏好。

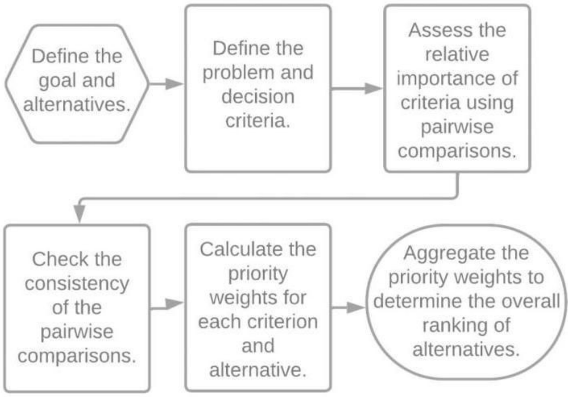
Define the goal and alternatives.
Define the problem and decision criteria.
Assess the relative importance of criteria using pairwise comparisons.
Check the consistency of the pairwise comparisons.
Calculate the priority weights for each criterion and alternative.
Aggregate the priority weights to determine the overall ranking of alternatives.

Figure 4. AHP steps.

图 4. AHP 步骤。

A second crucial step consists of organizing the criteria into a hierarchical structure. The top level is the overall objective (e.g., "Evaluate the effectiveness of automatic music generators"), and the next level specifies the main criteria (e.g., "musicality," "creativity," and "user-friendliness"). Sub-criteria can be added below the main criteria (under "musicality" sub-criteria like "melody quality" and "long-term structure" can be added).

第二个关键步骤包括将标准组织成层次结构。顶层是总体目标 (例如，“评估自动音乐生成器的有效性”) ，下一层指定主要标准 (例如，“音乐性”，“创造性” 和 “用户友好性”)。子标准可以添加在主标准之下 (在 “音乐性” 子标准下，如 “旋律质量” 和 “长期结构” 可以添加)。

For each pair of criteria at the same level, listeners are then asked to make pairwise comparisons and determine which criterion is more important. AHP typically uses a 1 to 9 scale in which odd values are significant steps (with 1 being of equal importance) and even values are intermediate steps. In this stage a comparison matrix is also created, based on the listeners’ responses. Tools such as AHP-OS provide an interface for listeners to register their comparison judgements and construct the associated comparison matrix. Although this step can be skipped if researchers prefer to use an equal criteria weight distribution, it is nevertheless one of the major advantages of the system, as it allows listeners’ criteria preference to be taken into consideration and it even allows the researcher to use two different groups (e.g., to use expert musicians’ judgements for the pairwise comparisons but with a criteria weight distribution established by a non-musician general population).

对于同一水平的每一对标准，听众被要求进行成对比较，并确定哪一个标准更重要。AHP 通常使用 1 到 9 的比例，其中奇数值是重要的步骤 (1 表示同等重要) ，偶数值是中间步骤。在这个阶段，还会根据听众的反应创建一个比较矩阵。像 AHP-OS 这样的工具为听众提供了一个接口来记录他们的比较判断并构建相关的比较矩阵。虽然如果研究人员更喜欢使用相同的标准权重分布，这一步可以跳过，但它仍然是该系统的主要优势之一，因为它允许考虑听众的标准偏好，甚至允许研究人员使用两个不同的组 (例如，使用专业音乐家的判断进行成对比较，但由非音乐家普通人群建立的标准权重分布)。

After completing the survey, the pairwise comparison results are used to calculate the weighted scores for each criterion. This is done by means of eigenvalue and eigenvector calculations on the comparison matrices carried out by the AHP software tool. The weights represent the relative importance of each criterion.

完成调查后，成对比较的结果被用来计算每个标准的加权分数。这是通过AHP软件工具对比较矩阵进行特征值和特征向量计算来完成的。权重代表每个准则的相对重要性。

It is now possible to evaluate different generators using the criteria and their weights. The assigned scores reflect how well a generator performs on each criterion. Finally, an aggregation step is executed by multiplying the scores of each criterion by their respective weights and adding them to produce a final score for each music generator. This will generate a total score that represents the overall evaluation of the generator.

现在可以使用这些准则和它们的权重来评估不同的生成器。赋予的分数反映了生成器在每个标准上的表现。最后，执行一个聚合步骤，将每个标准的分数乘以它们各自的权重，并将它们相加以产生每个音乐生成器的最终分数。这将生成一个总分，代表生成器的整体评估。

It should be noted that the AHP methodology and several AHP development tools such as AHP-OS (Goepel, 2018), include very thorough data diagnosis components, allowing for the evaluation of analysis results. Moreover, another very positive aspect of the AHP methodology is the possibility of dividing participants into clusters with high consensus among themselves, but with low consensus when compared to other clusters. This can be done thanks to the invention of a consensus indicator $\left(S^{¿}\right)$ that shows how much the members of a cluster agree with each other. This feature can be used to divide listeners into coherent subgroups based on their opinions of the different alternatives, and, as such, is potentially very useful for validating generators that may include genre transfer as part of their features, where the possible biases on the part of listeners towards particular musical genres can be decisive in the evaluation. Another benefit of this clustering methodology is that it can easily detect outliers, whose opinions may be beneficial to take into consideration, whereas other methodologies may simply discard them.

应该指出的是，AHP方法学和一些AHP开发工具，如AHP-os (Goepel，2018)，包括非常彻底的数据诊断组件，允许评估分析结果。此外，层次分析法的另一个非常积极的方面是，有可能将参与者分成小组，这些小组之间的共识很高，但与其他小组相比，共识很低。这可以通过共识指标$\left(S^{¿}\right)$的发明来实现，该指标显示集群

成员之间的一致程度。这一特征可以用来根据听众对不同选择的看法将其划分为连贯的子群，因此，对于验证可能将流派转移作为其特征的一部分的发生器来说潜在地非常有用，在这种情况下，听众对特定音乐流派可能存在的偏见可能在评价中起决定性作用。这种聚类方法的另一个好处是，它可以很容易地检测出离群点，这些离群点的意见可能值得考虑，而其他方法可能会简单地抛弃它们。

When assessing a group with a wide-ranging opinion pool, such as listeners with different musical expertise, and outlining their expectations and priorities for human or AI generated music, it is common for the collected data to indicate a good logical coherence but for the AHP group consensus indicator $\left(S^{¿}\right)$ not to reach the desired level, as there may be contradictory judgements. To address this issue, AHP implements a consensus indicator and a method for distributing the group into subclusters. This usually leads to a small subset of coherent groups with much higher consensus. The differences between these groups are always worth analyzing and should be justified in the study.

当评估具有广泛意见库的群体时，例如具有不同音乐专业知识的听众，并概述他们对人类或 AI 生成的音乐的期望和优先事项时，收集的数据通常表明良好的逻辑一致性，但是 AHP 群体共识指标 $\left(S^{¿}\right)$ 没有达到预期的水平，因为可能会有矛盾的判断。为了解决这个问题，AHP 实施了一个共识指标和将群体分配到子群集的方法。这通常导致一小部分一致的群体具有更高的共识。这些群体之间的差异总是值得分析，并在研究中得到证实。

For our proposed methodology, AHP should be prioritized over Likert-based surveys in studies with small user pools (under 20 users) or small sample pools (under or around 8). As AHP functions well in small group settings due to its high number of cross-checks between different criteria and its built-in consistency metric, it works best in the aforementioned scenarios. AHP should also be given preference or used in conjunction with Likert-based surveys in those evaluation studies where the objective is to rank different generator systems against each other. The user-weighted multiple criteria of AHP studies allow researchers to compare generators using a multilayered approach that can result in very thorough insights that go far beyond general user preference.

对于我们提出的方法，AHP 应该优先于基于李克特的调查，用于小用户池 (20 个用户以下) 或小样本池 (8 个或 8 个左右) 的研究。由于 AHP 在小组环境中运行良好，因为它在不同标准之间进行了大量的交叉检查，并且具有内置的一致性度量，所以它在上述场景中工作得最好。在那些旨在对不同发电机系统进行排序的评价研究中，也应优先使用层次分析法，或将其与基于李克特的调查结合使用。AHP 研

究的用户加权多重标准允许研究人员使用多层方法比较生成器，这种方法可以产生远远超出一般用户偏好的非常透彻的见解。

2.5. Likert survey and applicability

2.5 李克特调查和适用性

The Likert scale (Eerola et al., 2018) is a rating system widely used in surveys to estimate opinions or perceptions. Subjects can choose from a set of responses (usually 5 or 7) to a question. Table 1 shows some typical Likert scales. These scales are widely used in both social and educational research and are the basis of many software evaluation studies centered around user evaluation. When using Likert scales, the researcher must consider issues such as response categories (the values in the scale), the size of the scale, the direction of the scale, the ordinal nature of Likert-derived data, and the appropriate statistical analysis of those data.

李克特量表 (Eerola 等，2018) 是一种广泛用于调查的评级系统，用于估计观点或看法。受试者可以从一组回答 (通常是 5 或 7) 中选择一个问题。表 1 显示了一些典型的李克特量表。这些量表被广泛应用于社会和教育研究，并且是许多以用户评估为中心的软件评估研究的基础。在使用李克特量表时，研究者必须考虑诸如反应类别 (量表中的数值)、量表的大小、量表的方向、李克特衍生数据的有序性以及对这些数据的适当统计分析等问题。

Table 1. Examples of likert scales.

表 1. 李克特量表的例子。

	5	4	3	2	1
Agreement 协议	Strongly Agree 强烈同意	Agree 同意	Undecided 尚未决定	Disagree 不同意	Strongly Disagree 强烈反对
Frequency 频率	Always 一直如此	Often 经常	Sometimes 有时候	Rarely 很少	Never 从来没有
Importance 重要性	Very Important 非常重要	Important 重要事项	Moderately Important	Slightly Important 稍微重要	Unimportant 不重要

			中等重要		
Likelihood 可能性	Almost Always True 几乎总是正确的	Usually True 通常是正确的	Occasionally True 偶尔真实	Usually Not True Usually Not True 通常不对	Rarely True 很少是真的
Quality 品质	Excellent 非常好	Good 很好	Fair 公平	Poor 贫穷	Bad 糟糕
Distortion 失真	Imperceptible 难以察觉	Just perceptible not annoying 只是觉察不到而已	Perceptible slightly annoying 可察觉到的有点烦人	Annoying not objectionable 讨厌的不讨厌的	Very annoying objectionable 非常令人讨厌

Likert-derived data are usually treated statistically as interval-level data (i.e., values on the scale have directionality and are equidistant). This allows the use of parametric tests like analysis of variance (ANOVA) or Pearson’s product-moment correlation (Sullivan & Artino, 2013).

李克特派生的数据通常作为区间级数据进行统计处理 (也就是说，标度上的值具有方向性并且是等距的)。这允许使用参数测试，如方差分析 (ANOVA) 或皮尔逊的乘积矩相关性 (Sullivan & Artino，2013)。

Of the tools that are available for designing Likert-based surveys, CrowdMOS (Ribeiro et al., 2011) has become widely used in the field of audio quality evaluation, and, by extension, music software evaluation. An open-source project that can be easily deployed on a web server, it offers a way to crowd-source Mean Opinion Score (MOS) studies for subjective evaluation of audio, video, and image quality. CrowdMOS has been successfully used in several music generation studies (Copet et al., 2024) and provides a convenient and cost-effective tool to perform subjective quality evaluations, using platforms such as Amazon Mechanical Turk to recruit Internet users to participate in MOS-like listening studies. It makes it possible to filter noisy annotations and outliers by removing participants who do not listen to the full recordings or who rate references (in our example, human-composed music) with scores below 85. It also provides a set of recommended strategies (Ribeiro et al., 2011) for cleaning data. The

Likert scales for quality and distortion shown in Table 1 are widely used in CrowdMOS evaluations.

在可用于设计基于李克特的调查的工具中，CrowdMOS (Ribeiro 等，2011) 已经被广泛用于音频质量评估领域，并扩展到音乐软件评估。这是一个开源项目，可以很容易地部署在网络服务器上，它提供了一种方法来众包平均意见得分 (MOS) 研究的主观评价的音频，视频和图像质量。CrowdMOS 已经成功用于几项音乐生成研究 (Copet et al。 ，2024) ，并提供了一个方便和经济有效的工具来进行主观质量评估，使用 Amazon Mechanical Turk 等平台招募互联网用户参与类似 mos 的听力研究。它可以通过删除那些没有听完整录音的参与者或者那些给分数低于 85 的参考 (在我们的例子中是人创作的音乐) 打分的参与者来过滤噪音注释和异常值。它还提供了一套用于清理数据的推荐策略 (Ribeiro 等，2011)。表 1 中所示的质量和失真的李克特量表广泛用于 CrowdMOS 评估。

It should be noted that among the many advantages of Likert-based surveys, speed and ease of application are of the utmost importance. These characteristics make such surveys well suited to studies with many samples because the surveys themselves do not take up a significant amount of time from the participants, unlike methods such as AHP, described earlier.

值得注意的是，在基于李克特量表的众多优势中，速度和易用性是最重要的。这些特点使得这种调查非常适合于有许多样本的研究，因为调查本身不像前面描述的层次分析法等方法那样占用参与者大量时间。

For our proposed methodology, music generator evaluation studies that are not centered on the generation of a single specific genre, but may cover multiple music genres, should use Likert surveys in preference to AHP. Due to the wide variety of music sources involved, these types of studies require many more samples for a generator to be properly tested. The speed of the Likert application, in particular when users are making individual choices, makes it possible to evaluate a bigger sample pool of musical fragments. In our experience, fragment duration should usually be kept at around 30 seconds, as the best way to balance test speed with the possibility of implementing emotional measurements. This is, however, only an approximate time frame, and further research is needed to create more definitive guidelines. The use of the Likert based evaluation in this methodology can be implemented to accelerate the evaluation process when many pieces are to be evaluated. As an "accelerated" method, a smaller amount of criteria for evaluation may be preferred, which may leave behind concepts such as long-term structure, motivic development or continuous monitoring of emotional state through a piece (which would be replaced for dominant emotion classification/recognition).

对于我们提出的方法，音乐生成器评估研究，不是集中在一个单一的特定流派的生成，但可能涵盖多种音乐流派，应使用李克特调查优先于层次分析法。由于涉及音乐来源的广泛多样性，这些类型的研究需要更多的样本生成器进行适当的测试。李克特应用程序的速度，特别是当用户做出个人选择时，使得评估更大的音乐片段样本库成为可能。根据我们的经验，片段持续时间通常应该保持在 30 秒左右，这是平衡测试速度和实现情绪测量的可能性的最佳方法。然而，这只是一个大致的时间范围，需要进一步的研究来创建更明确的指导方针。在这种方法中使用基于李克特的评估可以实现加速评估过程，当许多作品要进行评估。作为一种 “加速 “方法，较少量的评估标准可能是首选，这可能会留下一些概念，如长期结构、动机发展或通过一个片段对情绪状态进行连续监测 (这将取代主导情绪分类 / 识别)。

The possibility of subsequently cross-referencing a subsample pool of Likert data with an AHP survey should also be considered, in particular in sensitive studies dealing with generators for education or model training, where failure in evaluation could produce cascading effects.

还应考虑随后将 Likert 数据的子样本池与 AHP 调查进行交叉参照的可能性，特别是在涉及教育或模型培训生成器的敏感研究中，评估失败可能产生连锁反应。

2.6. AHP, likert, and the adaptable hybrid system

2.6. AHP，likert 和适应性混合系统

Both Likert-based surveys and AHP-based studies have significant advantages and drawbacks. Ponsiglione et al. (2022) proposes a hybrid methodology to combine both approaches. Figure 3 illustrates the possibility of integration of Likert and AHP based assessments.

基于李克特的调查和基于层次分析法的研究都有显著的优点和缺点。Ponsiglione 等人 (2022) 提出了一种混合方法来结合两种方法。图 3 说明了整合基于 Likert 和 AHP 的评估的可能性。

In general, the combination of the two approaches offers a number of benefits. The use of two different data collection methods can, for example, help reduce errors and improve the accuracy of results. Complementary methods can also provide a more complete picture of the situation and can provide useful insights on the strengths and weaknesses of the studied generators that can help on their future improvements: AHP can, in general, provide guidance in more complex topics based on expert-user evaluation while Likert can benefit from a wider user pol with less expertise for simpler issues.

一般来说，这两种方法的结合提供了很多好处。例如，使用两种不同的数据收集方法可以帮助减少误差，提高结果的准确性。补充方法也可以提供一个更完整的情况，并可以提供有用的见解，研究生成器的优势和弱点，可以帮助他们未来的改进： AHP 可以，一般来说，在更复杂的主题提供指导的基础上，专家 - 用户的评估，而 Likert 可以受益于更广泛的用户政策，较少的专业知识，更简单的问题。

AHP and Likert based surveys can interact in different ways. As a first option AHP can be used to weight the Likert scale responses. In Song and Kang (2016), a method is developed where Likert scale surveys are used to reduce the number of pairwise comparisons. Ing (2021) takes a very practical approach, using differences in mean data from a previously available Likert scale survey to produce answers for the pairwise comparisons required by AHP.

基于 AHP 和李克特的调查可以以不同的方式相互作用。作为第一选择，AHP 可以用来衡量李克特量表的反应。在 Song 和 Kang (2016) 中，开发了一种方法，其中使用李克特量表调查来减少成对比较的数量。Ing (2021) 采用了一种非常实用的方法，利用先前可用的李克特量表调查的平均数据的差异，为 AHP 所要求的成对比较提供答案。

Another possible integration can use Likert scale surveys to validate AHP results (Gutknecht et al., 2018). Likert scales can be used to compare the relative importance of the criteria or alternatives obtained from both methods. In this approach, AHP is used to prioritize criteria or alternatives, while Likert scale surveys are used to directly gather user feedback on each criterion or alternative.

另一种可能的整合可以使用李克特量表调查来验证 AHP 结果 (Gutknecht 等，2018)。李克特量表可用于比较从两种方法获得的标准或替代方案的相对重要性。在这种方法中，层次分析法用于确定标准或备选方案的优先级，而李克特量表调查用于直接收集用户对每个标准或备选方案的反馈。

A variation of this implementation can provide expert supervision of music dataset tagging in music generation. Comparing the results from the larger and non-expert based Likert surveys to the shorter AHP questionaires that use expert, reaserch can detect results that have a wide variation on assesment ranges between different groups. This can indicate an error either by experts being outside their particular realm of expertise (as could be the case highlighted in subsection 4.2) or the criteria being too technical to be properly assessed by the untrained group. Rooting out errors, in particular in the dataset gathering stages is a very important topic that can highly impact music generation, thus requiring further and specific research on the topic.

该实现的一个变体可以为音乐生成中的音乐数据集标注提供专家监督。通过比较规模较大的非专家李克特调查和使用专家的较短的层次分析法调查问卷的结果，研究可以发现不同群体之间在评估范围上有很大差异的结果。这可能表明专家在其专门知识领域之外犯了错误 (如第 4.2 分节所强调的情况) ，或者标准技术性太强，未经培训的小组无法进行适当评估。根除错误，特别是在数据集收集阶段是一个非常重要的主题，可以高度影响音乐生成，因此需要进一步和具体的研究主题。

As a third option, AHP and Likert scale surveys can be used to create a dynamic, interactive decision-making process. For example, a web-based tool could be developed that allows stakeholders to input their responses to the Likert scale for different criteria. The tool could then use AHP (e.g., in the way proposed by Ing (2021)) to generate a prioritized list of alternatives. The experts could then review the results and provide feedback, and this feedback could then be used to refine the AHP model and generate a new prioritized list of alternatives. This process could be repeated until a satisfactory solution is achieved.

作为第三种选择，层次分析法和李克特规模调查可以用来创建一个动态的，交互式的决策过程。例如，可以开发一个基于网络的工具，允许利益相关者输入他们对不同标准的李克特量表的反应。然后，该工具可以使用 AHP (例如，按照 Ing (2021) 提出的方式) 来生成一个备选方案的优先列表。然后，专家们可以审查结果并提供反馈意见，这些反馈意见可以用来完善层次分析法模型，并产生一个新的优先备选办法清单。这个过程可以重复进行，直到得到一个令人满意的解决方案。

Finally, user opinions could be collected using Likert scales, which using a methodology similar to that proposed in Ing (2021), could be used to create inputs for an AHP-based analysis. Alternatively, users could directly answer AHP pairwise comparisons and then use a Likert survey as confirmation and complementary information. An approach combining direct pairwise comparisons with "virtual" comparisons obtained from a Likert survey could also be implemented and analyzed.

最后，可以使用李克特量表收集用户意见，这种方法类似于 Ing (2021) 提出的方法，可以用来为基于 ahp 的分析创建输入。或者，用户可以直接回答 AHP 成对比

较，然后使用李克特调查作为确认和补充信息。还可以实施和分析从李克特调查获得的直接成对比较和 “虚拟” 比较相结合的方法。

To further corroborate evaluation results, including also emotion / engagement / stress analysis might be very useful as a means of obtaining actual user feedback in real time.

为了进一步证实评估结果，也包括情绪 / 参与 / 压力分析，作为实时获得实际用户反馈的一种手段可能非常有用。

2.7. Emotional state measuring

2.7 情绪状态测量

Measuring emotional states in listeners can be extremely useful as a research tool, particularly in creative fields. Having listeners self-evaluate their emotional state using a survey based on self-assessment manikins (see Figure 5), conducted after having listened to each piece of music, is a very simple way to integrate emotional measurements into a evaluation study. Self-assessment manikins (SAM) are a picture-oriented approach to measuring emotional responses that have been used successfully in studies related to numerous topics, including music education (Cuadrado et al., 2020). SAMs aim to estimate three central features of emotional responses: perceived valence, perceived arousal, and perceptions of dominance. As a (usually) 9 point scale, they constitute a tool that is easy to integrate into a fast-paced Likert-survey model. Reserchers should consider that other Likert-based surveys for emotional estimation exist and been used to asses AI vs human generation (Diwanji et al., 2025) but SAMs are a widely standardized tool.

测量听众的情绪状态作为一种研究工具是非常有用的，特别是在创造性领域。让听众自我评估他们的情绪状态，使用一个基于自我评估人体模型的调查 (见图 5) ，在听完每一段音乐后进行，是一个非常简单的方法，将情绪测量整合到评估研究中。自我评估假人 (SAM) 是一种面向图片的方法，用于测量情绪反应，已经成功用于与包括音乐教育在内的许多主题相关的研究 (Cuadrado 等，2020)。Sam 旨在估计情绪反应的三个中心特征： 感知效价，感知唤醒和优势感知。作为一个 (通常) 9 分量表，它们构成了一个工具，很容易整合到一个快节奏的李克特调查模型。研究人员应该考虑到其他基于李克特的情绪评估调查存在，并被用来评估 AI 与人类生成 (Diwanji 等，2025) ，但 sam 是一个广泛标准化的工具。

To monitor emotional states in real time by means of a faster, more objective, emotional estimation, a complementary study based on the movement of face muscles or the measurement of physiological signals using wearable devices can be implemented. In this way, the researcher can compare and assess the validity of the listeners' self-assesments.

为了通过更快、更客观的情绪估计来实时监测情绪状态，可以利用可穿戴设备进行基于面部肌肉运动或生理信号测量的补充研究。通过这种方式，研究者可以比较和评估被试自我评估的有效性。

The most widely used physiological signals are parameters related to heart rate variability, such as interbeat intervals (IBI), and galvanic skin responses (GSR) parameters related to sweat gland activity. Many studies use the certified medical device Empatica E4 wristband (McCarthy et al., 2016) or its current successor Empatica EmbracePlus. A significant number of studies have used the Apple Watch (Hernando et al., 2018; Hirten et al., 2021) for HRV studies, although it lacks any medical device certification. This device and other common smartwatches are a good alternative that can help reduce study costs and improve test scal-ability. Quality measurements of HRV are very important in regards to emotion estimation, as some devices may not provide the best accuracy when being devised for sport tracking. Nevertheless, there is a significant evidence of the accuracy of some commercial sport trackers when compared to medically certified equipment (Lui et al., 2022), thus providing grounds for their use on other HRV measurement related tasks.

使用最广泛的生理信号是与心率变异分析相关的参数，如节拍间隔 (IBI) 和与汗腺活动相关的皮肤电反应 (GSR) 参数。许多研究使用经认证的医疗设备 Empatica e4 腕带 (McCarthy 等，2016) 或其当前的继任者 Empatica EmbracePlus。大量的研究使用 Apple Watch (Hernando 等，2018; Hirten 等，2021) 进行 HRV 研究，尽管它缺乏任何医疗设备认证。这种设备和其他常见的智能手表是一个很好的替代方案，可以帮助降低学习成本，提高测试的可扩展性。HRV 的质量测量在情绪估计方面是非常重要的，因为一些设备在设计用于运动跟踪时可能不能提供最佳的准确性。尽管如此，与医学认证的设备相比，一些商业运动追踪器的准确性有重要证据 (Lui et al。 ，2022) ，从而为其他 HRV 测量相关任务的使用提供了依据。

The recordings of wearable devices can be analyzed using statistical parameters calculated with tools such as the Kubios Heart Rate Variability (HRV) application (Tarvainen et al., 2014) and the MIT Medialab EDA explorer (Taylor et al., 2015) for the analysis of GSR. A comprehensive review of the works that use HR analysis to estimate emotions can be found in Ismail et al. (2024). A tool for emotion classification including a web-based interface can also be found in Bugnon et al. (2020). There are also several deep learning based tools suitable for this analysis (Muñoz-Saavedra et al., 2023). Directly measuring physiological signals usually requires a significant effort to process the data if good quality emotional information from listeners. Most researchers consider that at least 30 seconds of stimulus recording is necessary to be able to successfully carry out HRV analysis (Tanoue et al., 2023). However, it has been suggested that much shorter time analyses are suitable for valence prediction (Schippers et al., 2018). An implementation of these short-timed stimuli is further explored in section 3 (proof of concept). Nevertheless, the disadvantage of having to use longer durations to estimate general emotional states, as well as the preference for using these devices in controlled or semi-controlled environments to avoid confounding parameters may hinder the measuring of subtle emotional states that might briefly arise in music and that can also be shaped by the context of the listener. Improving these measuring technologies can improve a deeper understanding of the emotion to music detection in evaluation studies.

可穿戴设备的记录可以使用 Kubios 心率变异分析 (HRV) 应用程序 (Tarvainen 等，2014) 和 MIT Medialab EDA 探测器 (Taylor 等，2015) 等工具计算的统计参数进行分析 GSR 的分析。Ismail 等人 (2024) 对使用 HR 分析来估计情绪的作品进行了全面的回顾。Bugnon 等人 (2020) 也提供了一个包括基于网络的界面在内的情绪分类工具。还有几个基于深度学习的工具适合这种分析 (Muñoz-Saavedra 等，2023)。直接测量生理信号通常需要很大的努力来处理来自听众的高质量情感信息的数据。大多数研究人员认为，至少 30 秒的刺激记录是必要的，以便能够成功地进行 HRV 分析 (Tanoue 等，2023)。然而，有人提出，更短的时间分析适用于价格预测 (Schippers 等，2018)。这些短时间刺激的实施在第 3 节 (概念证明) 中进一步探讨。尽管如此，不得不使用较长的持续时间来估计一般情绪状态的缺点，以及在受控或半受控环境中使用这些设备以避免混淆参数的偏好可能会阻碍对音乐中可能短暂出现的微妙情绪状态的测量，并且也可以由听众的上下文决定。改进这些测量技术可以提高对评估研究中情绪对音乐检测的深入理解。

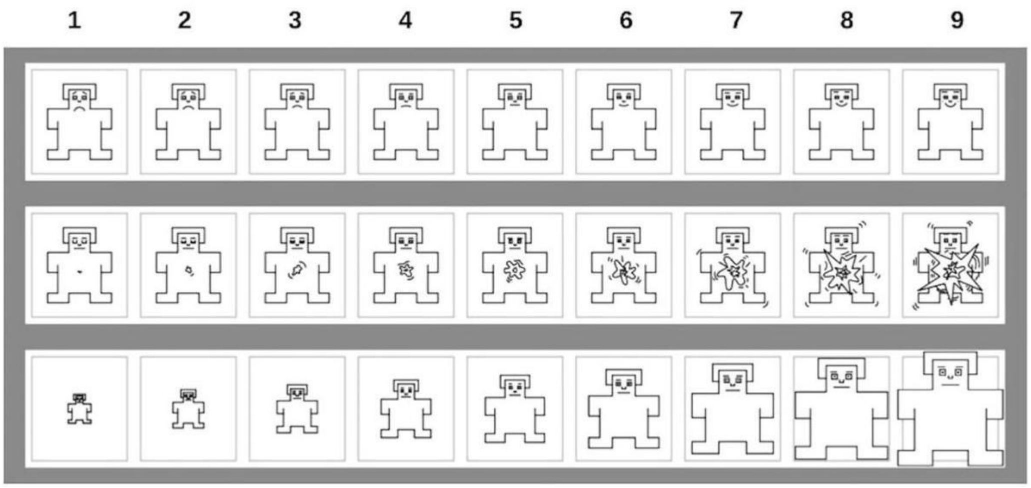

Figure 5. 9 point SAM scale (Knejzlíková et al., 2021).

图 5.9 点 SAM 量表 (Knejzlíková 等，2021)。

A consideration has to be made in regards to the relations between measured signals and emotions. The physiological responses that different individuals may have to emotions can vary significantly, adding difficulty to the development of universally applicable models (Ahmad & Khan, 2022). As the process of labeling physiological data with emotional states can be subjective, this can impact the reliability of the training data. Furthermore, Physiological signals are often noisy and contaminated with artifacts, requiring careful pre-processing to extract reliable features for emotion recognition. Different physiological signals may require different pre-processing techniques, which adds complexity to the development of robust emotion recognition systems.

必须考虑测量信号和情绪之间的关系。不同的个体对情绪的生理反应可能会有很大的差异，这给普遍适用的模型的发展增加了困难 (Ahmad & Khan，2022)。由于用情绪状态标记生理数据的过程可能是主观的，这可能会影响训练数据的可靠性。此外，生理信号往往含有噪声和伪迹，需要进行细致的预处理以提取可靠的特征用于情感识别。不同的生理信号可能需要不同的预处理技术，这增加了情感识别系统的复杂性。

As a second option, facial expression-based emotion detection systems can be implemented to further assess the emotional state of listeners. Most of these systems are based on the facial action coding system (FACS) (Rosenberg & Ekman, 2020), which encodes individual facial muscle movements based on visual observation of changes in facial expression.

This system is widely used for both expression analysis and facial expression synthesis. FACS is based on the analysis of "action units" (AU) related to specific movements in areas of the face. FACS can be used to determine listener emotions in real time, as these are directly related to the activation of specific AUs (Kodra et al., 2013).

作为第二种选择，可以实现基于面部表情的情绪检测系统，以进一步评估听众的情绪状态。这些系统中的大多数是基于面部动作编码系统 (FACS)(Rosenberg & Ekman，2020) ，其基于面部表情变化的视觉观察对个体面部肌肉运动进行编码。该系统广泛用于表情分析和面部表情合成。FACS 是基于对 “动作单元”(AU) 的分析，这些 “动作单元” 与面部区域的特定运动相关。FACS 可以用来实时确定听众的情绪，因为这些情绪与特定 au 的激活直接相关 (Kodra 等，2013)。

FACS can be coded manually, but this process is time-consuming and requires very well trained experts. Currently, almost all experiments use automatic action unit recognition. Programs like the open-source OpenFace (Cuculo & D’Amelio, 2019), Noldus FaceReader (Skiendziel et al., 2019), and Affectiva iMotions (Kulke et al., 2020) can automatically detect individual AU activations in real time. FaceReader and iMotion both also provide real-time probabilities for seven basic emotions, valence, and arousal. The ease of use of these applications, together with the unobtrusive implementation in evaluation studies (they mostly just require a face-cam capable of recording listeners while they are doing the test), makes face-based emotion detection probably the best alternative for easily obtaining emotional data during music evaluation and, in this way, complement and contrast listeners’ self-assessments. However, as with the previous tool, the length of the stimuli is an important consideration: It is important to compensate for the traces of positive and negative expressions on neutral faces by normalizing the emotion values for individual participants as proposed in Kayser et al. (2022).

FACS 可以手动编码，但这个过程很耗时，需要训练有素的专家。目前，几乎所有的实验都使用自动动作单元识别。像开源 OpenFace (Cuculo & d’amelio，2019) ，Noldus FaceReader (Skiendziel 等，2019) 和 Affectiva iMotions (Kulke 等，2020) 这样的程序可以实时自动检测单个 AU 激活。FaceReader 和 iMotion 都提供了七种基本情绪，效价和唤醒的实时概率。这些应用程序的易用性，以及在评价研究中的低调实施 (它们大多只需要一个能够在听众做测试时记录他们的面部摄像头) ，使得基于面部的情绪检测可能是在音乐评价中轻松获取情绪数据的最佳选择，从而补充和对比听众的自我评估。然而，与以前的工具一样，刺激的长度是

一个重要的考虑因素： 重要的是通过正常化个体参与者的情绪值来补偿中性面孔上正面和负面表情的痕迹，如 Kayser 等人 (2022) 所提出的。

3. Pilot study

3. 试点研究

As a proof of concept, we decided to evaluate the real-world performance of the Meta generator (Copet et al., 2024) against human-made music from the MUSICAPS (Agostinelli et al., 2023) dataset which was used for its training. We then compared with the Perceiver Music Transformer (Hawthorne et al., 2019), a symbolic generator validated against the MAESTRO Midi dataset (Hawthorne et al., 2019) of human-composed and interpreted music.

作为概念的证明，我们决定评估 Meta 生成器 (Copet 等，2024) 针对 MUSICAPS (Agostinelli 等，2023) 数据集中用于其训练的人造音乐的现实世界性能。然后，我们将其与 perceptionr Music Transformer (Hawthorne et al。 ，2019) 进行比较，后者是针对人类作曲和解释音乐的 MAESTRO Midi 数据集 (Hawthorne et al。 ，2019) 进行验证的符号发生器。

To evaluate the feasibility of our proposed methodology for developing music validation strategies, a preliminary experiment was conducted that incorporated the majority of tools referenced in Section 2. The study employed a counterbalanced design with four participants who examined 40 short excerpts of music. The four participant were all postgraduate students, two male and two female with ages ranging from 25 to 33 years old. All had backgrounds in music research with extensive listening experience, but did not self-identify as music experts or professional musicians.

为了评估我们提出的开发音乐验证策略的方法的可行性，进行了一个初步的实验，其中包含了第 2 节中提到的大部分工具。这项研究采用了一个平衡设计，有四个参与者检查了 40 个音乐短片段。这四名参与者都是研究生，两男两女，年龄从 25 岁到 33 岁不等。他们都有音乐研究的背景，有丰富的听音经验，但并不认为自己是音乐专家或专业音乐家。

The music excerpts were equally split between those generated by the specified algorithms and those sourced from human-composed music, with all pieces adjusted to a uniform loudness of -14 LUFS (Katz & Katz, 2003).

Given the limited number of participants and the objective to compare the two generative models against human-composed music across multiple dimensions, as outlined in the recommended guideline (see Section 2), the Analytic Hierarchy Process (AHP) was selected as the principal validation instrument. Due to the generators not being designed as genre-specific, it was imperative to utilize a diverse sample pool. The broad scope of this sample pool necessitated significant user effort to evaluate the generators via AHP, thus AHP was applied only to a subset of eight pieces (two of each generator and two of each human dataset), while a simpler Likert-based survey was employed for the entire 40 music sample set, in accordance with the guidelines. All music excerpt were previously selected randomly among the produce of the generators and the pieces in the datasets.

音乐摘录被平均分配到由特定算法生成的那些和来自人类作曲的音乐的那些，所有作品调整到 -14 LUFS 的统一响度 (Katz & Katz，2003)。鉴于参与者的数量有限，以及在多个维度上比较两种生成模型与人类作曲音乐的目标，如推荐的指南 (见第 2 节) 所述，选择层级分析法 (AHP) 作为主要验证工具。由于生成器不是针对特定流派设计的，因此必须利用多样化的样本池。这个样本池的广泛范围需要用户通过 AHP 评估生成器的重大努力，因此 AHP 仅适用于八个片段的子集 (每个生成器的两个和每个人类数据集的两个) ，而对于整个 40 个音乐样本集使用更简单的基于李克特的调查，根据指导方针。先前所有的音乐摘录都是在生成器和数据集中的作品中随机选择的。

All the listening test and ratings were conducted in a controlled listening environment. All listeners used closed-back headphones. This space was a quiet and evenly illuminated room. Participants where asked to listen to the 4 AHP selected pieces in a random order and then complete the AHP assessment. Then they assessed all pieces in the Likert subset by listening and assessing one piece at a time in a random order. All four participants completed the study, both the AHP and Likert assessments.

所有的听力测试和评分都在一个可控的听力环境中进行。所有的听众都使用封闭式耳机。这是一个安静且光线均匀的房间。参与者被要求听 4 个层次分析法随机选择的片段，然后完成层次分析法评估。然后他们评估李克特子集中的所有作品，每次以随机顺序听和评估一个作品。所有的四个参与者都完成了研究，包括 AHP 和李克特的评估。

Facial expressions and physiological data (HRV) were collected while listening to each piece for the first time, before the AHP survey.

面部表情和生理数据 (HRV) 是在 AHP 调查之前，第一次听每个片段时收集的。

The incorporation of these various parameters for a comprehensive methodological assessment resulted in each participant requiring approximately two hours to complete the experience. This duration can be greatly optimized in specific experiments (Civit et al., 2024) by refining the data collection process using the Adaptability Matrix (Figure 3).

将这些不同的参数纳入全面的方法学评估导致每个参与者需要大约两个小时来完成经验。通过使用适应性矩阵 (图 3) 改进数据收集过程，可以在特定的实验中大大优化这个持续时间 (Civit 等，2024)。

Two AHP surveys were used to rate several aspects of the pieces. They were created using the design shown in Figure 6. and followed a two-tier criteria approach with three main criteria being compared for each song and including three sub-criteria for each of these criteria. This made it possible to establish comparisons and rankings between all four songs in each survey for all of those aspects.

两个 AHP 调查被用来评估这些作品的几个方面。它们的设计如图 6 所示。并遵循一个两层标准的方法，对每首歌曲有三个主要标准进行比较，每个标准包括三个子标准。这使得在每个调查中的所有四首歌之间建立所有这些方面的比较和排名成为可能。

As can be seen in Table 2, human-composed songs 1 and 2 were clearly preferred by listeners in almost all aspects. The table compares human-composed to AI-generated symbolic piano music, and includes a weighting distribution compensated by the judgement of 4 nonexperts users in music. This provides extra feedback on the pieces and generators, while also highlighting the possibility and usefulness of using two groups with different characteristics for the evaluation and weight distribution stages (note that this is not done in this pliot study). Table 3 compares Ai-generated and human-composed audio-based music. It also shows how listeners rated human-composed music higher in almost all aspects. However, listeners rated song 7 (AI) as very close to human song 5 in its overall evaluation. This demonstrates the viability of generators if enough cherry-picking is done with the generated output results. It is clear that grouping all human-composed or AI-generated songs would provide less insight into the real capabilities of generated music, as listeners clearly rated some songs above others in most aspects. This illustrates the fact that it is still widely desirable to generate several songs and let humans choose among them.

从表 2 中可以看出，人创作的歌曲 1 和 2 在几乎所有方面都受到听众的青睐。该表比较了人类创作的符号钢琴音乐与 ai 生成的符号钢琴音乐，并包括由音乐中的 4 个非专家用户的判断补偿的加权分布。这提供了关于零件和生成器的额外反馈，同时也突出了在评价和权重分配阶段使用两个具有不同特点的小组的可能性和有用性 (注意，本试点研究没有这样做)。表 3 比较了 ai 生成的和人类创作的基于音频的音乐。表 3 还显示了听众在几乎所有方面对人工音乐的评价。然而，听众对歌曲 7 (AI) 的总体评价与人类歌曲 5 非常接近。这表明，如果对生成的输出结果进行足够多的樱桃选择，生成器是可行的。很明显，将所有人类创作的或人工智能生成的歌曲分组会使人们对生成的音乐的真正能力缺乏洞察力，因为听众在大多数方面都清楚地认为某些歌曲优于其他歌曲。这说明了一个事实，即生成几首歌曲并让人类从中选择仍然是广泛可取的。

As AHP surveys are time-consuming, only eight pieces of music were evaluated with that method, the others being evaluated in a much more direct Likert-based survey seeking the "more liked" (the highest valued) subcriterion using a 9 point scale ranging from 1 (I didn’t like it at all) to 9 (I loved it). The first objective was to establish whether the participants had a clear perceived preference between human-composed music and AI-generated music, and between symbolic music and raw audio music. To answer this question, an ANOVA (analysis of variance) was carried out on the general sample of generated and composed music. A two-way ANOVA model was first tested to ascertain whether the users’ acceptance of different songs was directly linked to whether or not the music was human-composed and whether it was symbolic or raw audio. In this analysis, we obtained F values above 32 for human generation and above 14 for non-symbolic generation. In both cases p was below .0001, and therefore the null hypothesis had to be rejected. A Tukey test was performed for the two-way model, establishing a mean "like" difference between composed and generated pieces of 1.95 with a 95% confidence interval $\left[2.62, 1.27\right)\left(p<.0001\right)$. The results for non-symbolic and symbolic music showed a mean "like" difference of 1.30 with a 95% confidence interval $\left[0.63, 1.97\right)$ (p<.0001). Table 4 shows the mean values and the 95% confidence intervals for the different song groups.

由于层次分析法的调查非常耗时，只有 8 首音乐是用这种方法进行评估的，其他的则是在一个更直接的基于李克特的调查中进行评估，寻找 “更喜欢”(最高价值) 的次级标准，使用 9 分制，从 1 分 (我一点也不喜欢) 到 9 分 (我喜欢)。第一个目标是确定参与者是否在人类创作的音乐和人工智能生成的音乐之间，以及在象征性音乐和原始音频音乐之间有一个明确的感知偏好。为了回答这个问题，方差分析 (方差分析) 进行了一般样本的生成和作曲的音乐。首先测试一个双向 ANOVA 模型，

以确定用户对不同歌曲的接受程度是否与音乐是否由人创作以及是象征性音频还是原始音频直接相关。在这个分析中，我们获得了高于 32 的人类生成值和高于 14 的非符号生成值。在这两种情况下，p 都低于.0001，因此零假设必须被拒绝。对双向模型进行 Tukey 检验，建立组合和生成的片段之间的平均 “类似” 差异为 1.95,95% 置信区间 $\left[2.62, 1.27\right)\left(p<.0001\right)$ 。非象征性和象征性音乐的结果显示平均 “类” 差异为 1.30，置信区间为 $95\%\left[0.63, 1.97\right)$ (p < .0001)。表 4 显示了不同歌曲组的平均值和 95% 置信区间。

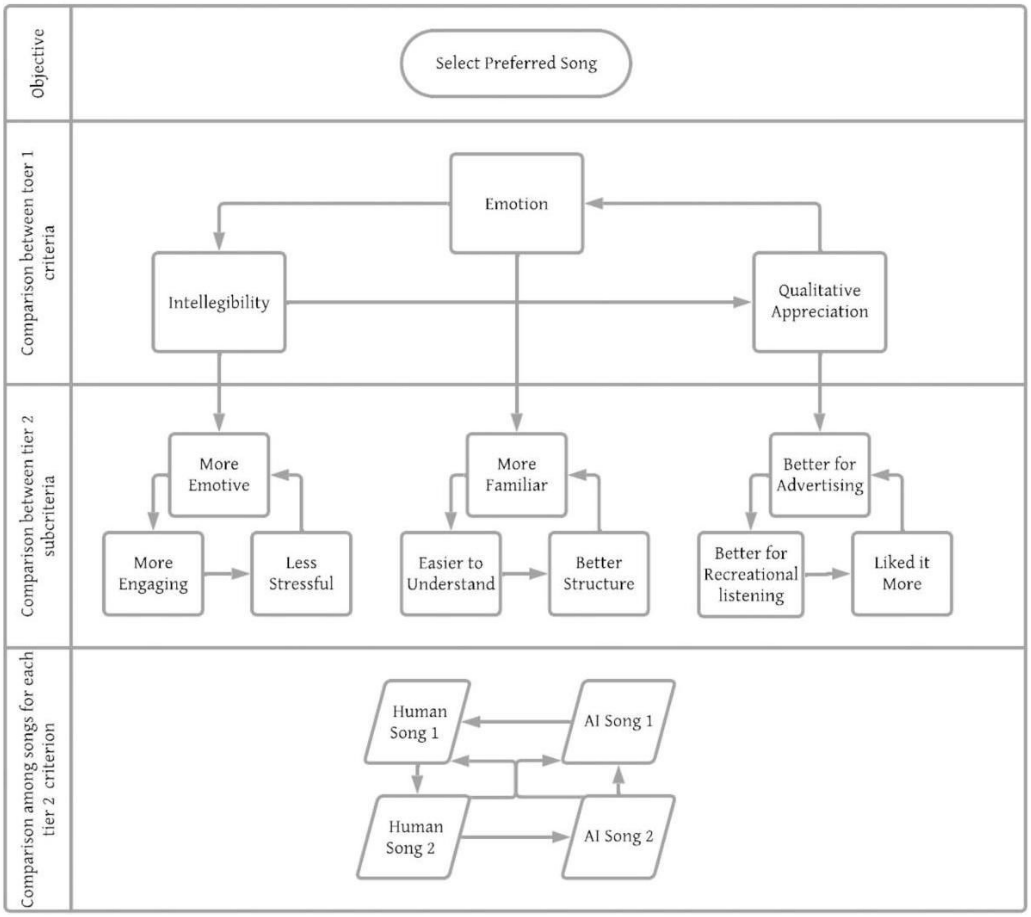

Figure 6. AHP study model.

图 6. AHP 研究模型。

Table 2. Symbolic music AHP expert evaluation with non-expert weight distribution.

表 2。符号化音乐 AHP 专家评价与非专家权重分配。

		Huma	Huma	AI	AI

				n Song 1 人类之歌 1	n Song 2 人类之歌 2	Song 3 AI Song 3 人歌 3	Song 4 艾歌 4
	Emotion 0.455 情绪 0.455	More emotive 0.414 更多情绪 0.414	18.90% 18.90%	0.495	0.296	0.09	0.119
		More engaging 0.479 更吸引人 0.479	21.80% 21.80%	0.414	0.438	0.066	0.081
		Less stressful 0.107 压力小 0.107	4.90% 4.90%	0.525	0.267	0.077	0.131
		More familiar 0.259 更熟悉 0.259	2.70% 2.70%	0.44	0.425	0.062	0.073
AI & Human Music 人工智能与人类音乐	Intelligibility 0.106 可懂度 0.106	Easier to understand 0.165 更容易理解 0.165	1.70% 1.70%	0.443	0.406	0.074	0.077

		Better structure 0.576 更好的结构 0.576	6.10% 6.10%	0.396	0.415	0.12	0.069
	Qualitative 0.439 质量 0.439	Better for video advertising 0.092 视频广告 0.092	4.00% 4.00 厘	0.4	0.459	0.079	0.062
		Better for recreational listening 0.238 休闲听力更好 0.238	10.40% 10.40%	0.407	0.429	0.1	0.064
		Liked more 0.670 更喜欢 0.670	29.40% 29.40%	0.393	0.405	0.139	0.063
			100% 100%	42.70% 42.70%	39.02% 39.02%	9.98% 9.98%	8.17% 8.17%

The numbers in the first column reflect the importance (weight) placed by the users on each of the main criteria as a fraction,   i.e., emotion 0.455

means that the importance placed on emotions is slightly over 45% while the importance placed on intelligibility is around 10%. The third column tells as the importance of each subcriteria inside the criteria,   i.e., More emotive has a weight inside the emotion criteria if about 41% while less stressful is valued only around 11%. the forth column is the global weight of the subcriteria expressed as a percentage   i.e., the product of the weight of the criteria times the weight of the subcri-teria (e.g., $18.9=0.455\times0.414\times100$). The values in the remaining columns are calculated by the AHP algorithm considering the weight in the third columns and the answers provided by the listeners for the pairwise comparison among songs. These values reflect the grade given for the considered criteria for the particular song. The total sum of the grades for the different subcriteria for a song gives the global grade of the song. The addition of the global grades, if these are expressed as a percentage, should add to 100%. In our case we can see that song 1 gets the best grade (42.7%) while song 4 gets the worst grade (8.17%).

第一栏中的数字反映了用户对每个主要标准的重视程度 (权重) ，也就是说，情绪 0.455 意味着对情绪的重视程度略高于 45% ，而对可理解性的重视程度约为 10% 。第三列说明了标准中每个子标准的重要性，也就是说，更多的情绪在情绪标准中占有权重，如果约 41% ，而较少的压力只占 11% 左右。第四列是以百分比表示的次级标准的全局权重，即标准的权重乘以次级标准的权重 (例如，$18.9=0.455\times0.414\times100$) 的乘积。其余列的值由 AHP 算法计算，考虑第三列的权重和听众提供的答案，用于歌曲之间的成对比较。这些值反映了特定歌曲所考虑的标准的等级。一首歌曲的不同次级标准的总分给出了这首歌曲的总分。如果以百分比表示的话，总分的加和应该是 100% 。在我们的例子中，我们可以看到歌曲 1 得到最好的分数 (42.7%) ，而歌曲 4 得到最差的分数 (8.17%)。

Table 3. Audio-based music AHP expert evaluation with non-expert weight distribution subcriteria weights are the same as the objective of the evaluation is the same.

表 3。基于音频的 AHP 专家评价与非专家权重分配的子准则权重是一样的，评价的目标也是一样的。

		Song 5 Human 歌曲 5 人类	Song 6 Human 歌曲 6 人类	Song 7 AI 歌曲 7 人工智能	Song 8 AI 歌曲 8 AI

	Emotion 0.455 情绪 0.455	More emotive 0.414 更多情绪 0.414	18.90% 18.90%	0.311	0.467	0.133	0.09
		More engaging 0.479 更吸引人 0.479	21.80% 21.80%	0.268	0.469	0.173	0.09
		Less stressful 0.107 压力小 0.107	4.90% 4.90%	0.327	0.364	0.264	0.046
		More familiar 0.259 更熟悉 0.259	2.70% 2.70%	0.289	0.39	0.262	0.059
AI & Human Music 人工智能与人类音乐	Intelligibility 0.106 可懂度 0.106	Easier to understand 0.165 更容易理解 0.165	1.70% 1.70%	0.246	0.372	0.326	0.056
		Better structure 0.576	6.10% 6.10%	0.375	0.341	0.243	0.04

		更好的结构 0.576					
	Qualitative 0.439 质量 0.439	Better for video advertising 0.092 视频广告 0.092	4.00% 4.00 厘	0.257	0.339	0.355	0.05
		Better for recreational listening 0.238 休闲听力更好 0.238	10.40% 10.40%	0.304	0.374	0.28	0.043
		Liked more 0.670 更喜欢 0.670	29.40% 29.40%	0.262	0.365	0.299	0.075
			100% 100%	28.70% 28.70%	40.58% 40.58%	23.45% 23.45%	7.24% 7.24%

In this case (audio-based songs) The best song is also human composed and the worst is also AI composed but there are two songs (5 and 6) which have relatively similar grades. This shows that by selecting specific AI based songs we can get results that are similar to the not preferred human-composed songs. Criteria weight distribution could also be flattened to not account for user preference giving each subcriterium a weight of 11.11%.

在这种情况下 (基于音频的歌曲) ，最好的歌曲也是人类创作的，最差的也是人工智能创作的，但有两首歌 (5 和 6) 有相对相似的等级。这表明，通过选择特定的基于人工智能的歌曲，我们可以得到类似于非首选人类创作的歌曲的结果。标准权重分布也可以平坦化，不考虑用户偏好，给每个子标准赋予 11.11% 的权重。

Our experiment follows a within-subjects design in which all participants evaluate both types of music: AI-generated and human-composed. This design controls for individual differences by having the same participants evaluate both conditions.

我们的实验遵循一个主题内设计，所有参与者评价两种类型的音乐： 人工智能生成的和人类创作的。这种设计通过让相同的参与者评估两种条件来控制个体差异。

To further test the songs, we implemented a set of tools to analyze emotional states. We used both HRV and GSR measurements and the FACS analyzer described in Subsection 2.7. These tools were paired with the SAM-based measurement done together with the Likert-based surveys, and their results provide further insight into the emotional state of the participants while listening to the different pieces.

为了进一步测试这些歌曲，我们实施了一套分析情绪状态的工具。我们使用 HRV 和 GSR 测量以及小节 2.7 中描述的 FACS 分析仪。这些工具与基于 sam 的测量以及基于李克特的调查结合在一起，它们的结果为参与者在听不同乐曲时的情绪状态提供了进一步的洞察。

We also recorded the facial expression of the users while listening to the selected symbolic songs (1 and 2 human-composed, 3 and 4 AI generated). These data were analyzed using Noldus Face Reader and the corrections proposed in Kayser et al. (2022) were applied. In Table 5 we present the mean basic emotions with the associated valence and arousal values detected by Noldus Face Reader for the different songs.

我们还记录了用户在听选定的象征性歌曲时的面部表情 (1 和 2 人为创作，3 和 4 人工智能生成)。使用 Noldus Face Reader 分析这些数据，并应用 Kayser 等人 (2022) 提出的修正。在表 5 中，我们呈现了平均基本情绪以及由 Noldus Face Reader 检测到的不同歌曲的相关价值和唤醒值。

In Table 6 we provide the mean and standard deviation of the valence and arousal value per song. In this table we can observe that Songs 1 and 2 exhibit high variability in valence, indicating diverse emotional perceptions among participants. This could be attributed to the complexity or ambiguity of these songs. Despite mixed valence ratings, Song 2 shows no variability in arousal, suggesting that it is universally engaging and has consistent emotional intensity across participants. Song 3, with its low valence and high arousal, appears to evoke strong reactions, potentially perceived as more unsettling or intense, likely evoking negative emotions such as anger or surprise. On the other hand, Song 4 is characterized by the highest mean valence and relatively lower arousal, suggesting that it is perceived as the most pleasant but with less emotional intensity.

在表 6 中，我们提供了每首歌的平均值和标准差。在这个表格中，我们可以观察到歌曲 1 和 2 在效价上表现出高度的可变性，表明了参与者之间不同的情绪感知。这可以归因于这些歌曲的复杂性或模糊性。尽管效价参差不齐，但是第二首歌在激发性方面没有表现出多样性，这表明它具有普遍的吸引力，并且在参与者中具有一致的情感强度。第三首歌，由于它的低效价和高唤醒度，似乎会引起强烈的反应，潜在地被认为是更加令人不安或强烈的，可能会引起诸如愤怒或惊讶等负面情绪。另一方面，宋 4 拥有属性最高的平均效价和相对较低的唤醒度，表明它被认为是最愉快的，但情绪强度较低。

Table 4. Grouped song acceptance mean rates and confidence intervals based on likert survey for all participants.

表 4: 基于李克特调查的所有参与者的分组歌曲接受率和置信区间。

generated 生成	audio 音频	Lower 低一点	Mean 平均值	Upper 上
Human 人类	Midi Midi 迷笛	5.1	5.68	6.25
Human 人类	Audio 音频	6.2	6.85	7.5
Generated 生成	Midi Midi 迷笛	2.86	3.42	3.97
Generated 生成	Audio 音频	4.06	4.92	5.77

Table 5. Face Reader results main detected emotion values for valence and arousal for each song and participant.

表 5。面孔识别结果主要检测了每首歌曲和参与者的情绪效价和唤醒值。

Song 歌曲	Part. 第二部分。	FR_EMO FR _ emo 情绪波动	FR_VAL 弗雷德里克 - 瓦尔	FR_ARO FR _ aro

1	4	Angry(6) 愤怒 (6)	3	9
2	4	Happy(8) 快乐 (8)	8	6
3	4	Surprise(6) 惊喜 (6)	5	5
4	4	Surprise(3) 惊喜 (3)	5	3
1	5	Angry(2) 愤怒 (2)	3	3
2	5	Scared(7) 害怕 (7)	3	6
3	5	Angry(6) 愤怒 (6)	2	8
4	5	Happy(8) 快乐 (8)	8	7
1	6	Sad(4) 悲伤 (4)	3	5
2	6	Disgust(4) 厌恶 (4)	4	6
3	6	Surprise(5) 惊喜 (5)	3	5
4	6	Scared(5) 害怕 (5)	4	4
1	7	Happy(7) 快乐 (7)	7	7
2	7	Disgust(3) 厌恶 (3)	3	6
3	7	Sad/4) 悲伤 / 4)	4	7
4	7	Sad(6) 悲伤 (6)	4	6

Songs 1 and 2 are human-composed while song 3 and 4 are Al-generated. All songs are symbolic.

歌曲 1 和 2 是人类创作的，而歌曲 3 和 4 是铝合金生成的，所有的歌曲都是象征性的。

Table 6. Mean and standard deviation for valence and arousal per song for all participants.

表 6: 所有参与者每首歌的效价和唤醒度的平均值和标准差。

Song 歌曲	Mean Valence 平均效价	SD Valence 瓦朗斯	Mean Arousal 平均唤醒度	SD Arousal SD 觉醒
1	4.00	2.00	6.00	2.58
2	4.50	2.38	6.00	0.00
3	3.50	1.29	6.25	1.50
4	5.25	1.89	5.00	1.83

To measure HRV (heart rate variability), participants wore an Empatica E4 medically certified device during the tests. This device captures heart interbeat intervals, galvanic skin response, body temperature, and three-axis wrist accelerometer data (McCarthy et al., 2016). It has been shown that there is a clear link between the lower frequencies in the RR spectrum and the excitation capabilities of music (Dimitriev et al., 2022). An example of those effects is shown in Figure 7 where the left-hand plot represents the spectrum of a song that the participant clearly considered surprising. The right-hand plot represents the data for a song clearly considered happy. These analyses can be automated using machine learning techniques such as those applied in Muñoz-Saavedra et al. (2023). The validity of these data in our proof of concept is limited because we used the original short music samples to extract the HRV characteristics that are then correlated to valence values. The sample window must cover at least 30s as established in Schippers et al. (2018) and therefore we had to include some data after the listening period. Future studies aiming to use this technology should be careful to take the duration of the sample into consideration when designing the experiment.

为了测量心率变异分析 (HRV) ，参与者在测试过程中佩戴了 Empatica e4 医学认证设备。该装置捕获心跳间隔，原电皮肤反应，体温和三轴手腕加速度计数据 (McCarthy 等，2016)。已有研究表明，RR 频谱中的较低频率与音乐的激发能力之间存在明显的联系 (Dimitriev 等，2022)。图 7 显示了这些效应的一个例子，其中左侧图表示参与者显然认为惊讶的歌曲的频谱。右边的图表代表了一首明显被

认为是快乐的歌曲的数据。这些分析可以使用 Muñoz-Saavedra 等人 (2023) 应用的机器学习技术实现自动化。这些数据的有效性在我们的概念证明是有限的，因为我们使用原始的短音乐样本提取 HRV 特征，然后相关的价值观。样本窗口必须覆盖 Schippers 等人 (2018) 建立的至少 30 秒，因此我们必须在听音期后包括一些数据。旨在使用这种技术的未来研究应该在设计实验时仔细考虑样本的持续时间。

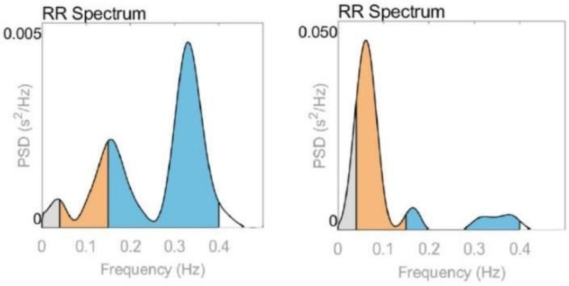

Figure 7. HRV data for surprise (song 3) and happy (song 2) example of how emotions can be represented in HRV RR spectrum.

图 7。惊喜 (歌曲 3) 和快乐 (歌曲 2) 的 HRV 数据示例情绪如何在 HRV RR 频谱中表现。

4. Discussion

4. 讨论

4.1. Implications of the pilot study

4.1 试点研究的意义

The purpose of this pilot study was to highlight the viability of the methodology and tools presented in Section 2. Although the study had a quite small participant pool, the results are nevertheless interesting and may point towards areas that require future research.

这项试点研究的目的是强调第 2 节中提出的方法和工具的可行性。虽然这项研究的参与者人数相当少，但结果仍然很有趣，可能指向需要未来研究的领域。

First, in complete contrast to our initial hypothesis, the reviewers rated music generated in audio format as being much closer to its human counterpart than midi-generated music (see Tables 3 and 2). We initially assumed that with all midi recordings being very accurately normalized classical piano solos played by the same high-quality midi instrument (Analog Lab V), they would have ratings very similar to human midi music. Compared to audio-generated pieces, it seems that the generator's higher degree of freedom (apart from not using "human" voices, it has no other constraints) made it more interesting and more able to convey emotions (one of the highest valued criterion in the AHP survey). Further research is needed into audio vs midi generation, particularly with regard to effects on emotions, as this is a new and increasingly interesting field.

首先，与我们最初的假设完全相反，评论者认为音频格式生成的音乐比 midi 生成的音乐更接近于人类对应的音乐 (见表 3 和表 2)。我们最初假设所有的 midi 录音都是由相同的高质量的 midi 乐器 (Analog Lab v) 演奏的非常准确的标准化的古典钢琴独奏，它们的等级与人类的 midi 音乐非常相似。与音频生成的作品相比，似乎生成器的更高自由度 (除了不使用 “人类” 的声音，它没有其他限制) 使它更有趣，更能够传达情感 (在 AHP 调查中最高的价值标准之一)。对于音频和中音的生成还需要进一步的研究，特别是关于对情绪的影响，因为这是一个新的和越来越有趣的领域。

The second important implication of the pilot study is that although AI music generation is rated as acceptable (especially in the audio realm), the results are below those reported in the original Meta MusicGen study (Copet et al., 2024). Even though our study is too small to draw final conclusions, it highlights the need to independently validate the most used and ubiquitous music generators that may later be used as engines for music devices, music teaching tools, etc., or as a key elements in the short-term democratization of the music making process.

试点研究的第二个重要含义是，尽管 AI 音乐生成被评为可接受的 (特别是在音频领域) ，但结果低于原始 Meta MusicGen 研究 (Copet 等，2024) 中报道的结果。尽管我们的研究规模太小，无法得出最终结论，但它强调了独立验证最常用和无处不在的音乐发生器的必要性，这些发生器以后可能被用作音乐设备、音乐教学工具等的引擎，或者作为音乐制作过程短期民主化的关键因素。

4.2. Music and culture for validation studies

4.2 音乐和文化的验证研究

As an art form that has accompanied humankind since the very beginning of our history, music has developed together with our culture, language, and forms of expression. Such a complex concept is usually the subject of whole dissertations, from historiographical research into its origin and development to cultural studies that build bridges between different interpretations. With many genres and subgenres, music is never just "music," and not considering this can pose serious problems for the future of our culture, especially when dealing with AI-generated music. From its inception, our proposed methodology is intended to be genre-blind, capable of being implemented regardless of the genre of the music that is evaluated. Nevertheless, it is important to note that our research group is of western origin, and thus can be inherently biased towards favorising characteristics that are predominant in western music. Consequently, we advise researchers that are interested in non-western music genres to source participants and, particularly experts, that are well versed in these particular genres, and to consider if any of the steps in the application of the general model (Subsection 2.2) can be fine-tuned.

音乐作为一种自人类历史之初就伴随着人类的艺术形式，与我们的文化、语言和表现形式一起发展。这样一个复杂的概念通常是整个论文的主题，从史学研究到音乐的起源和发展，再到文化研究，在不同的解释之间架起桥梁。在许多流派和子流派中，音乐永远不仅仅是 “音乐”，不考虑这一点可能会对我们文化的未来造成严重的问题，尤其是在处理人工智能生成的音乐时。从一开始，我们提出的方法论就是无视音乐类型的，无论被评估的音乐类型是什么，它都能够被实现。然而，重要的是要注意到，我们的研究小组是西方的起源，因此可以固有地偏爱的特点，主导西方音乐。因此，我们建议那些对非西方音乐流派感兴趣的研究人员去寻找参与者，尤其是那些精通这些特定流派的专家，并且考虑在应用一般模型 (第 2.2 小节) 的任何步骤是否可以进行微调。

As explained in Civit et al. (2022), many current AI music generators are designed without considering that, depending on the corpus of music used for their training, they will have an implicit bias towards a specific genre, usually western classical music or western pop music (Moysis et al., 2023). A further problem arises when the generated product of such generators is used to train new models - something which to the best of our knowledge is not currently happening, but is nevertheless a more than likely scenario as generators improve in performance and datasets of AI generated music become available (Civit et al., 2024). In this case, those cultural biases will be passed on from generator to generator, and possibly even from generators to humans if we start using AI music generation for educational purposes. With such issues in the spotlight of debate, it is now necessary for validation studies to explicitly or implicitly address the genre specificity of music generators.

正如 Civit 等人 (2022) 所解释的那样，许多目前的 AI 音乐发生器的设计没有考虑到，根据用于训练的音乐语料库，它们将对特定流派 (通常是西方古典音乐或西方流行音乐) 有隐含的偏见 (Moysis 等，2023)。当使用这种生成器的生成产品来训练新模型时，还会出现另一个问题 - 据我们所知，目前还没有发生这种情况，但是随着生成器在性能上的改善和 AI 生成音乐的数据集的可用性，这种情况可能性更大 (Civit 等，2024)。在这种情况下，如果我们开始将人工智能音乐生成用于教育目的，那么这些文化偏见将从一个发生器传递到另一个发生器，甚至可能从发生器传递到人类。随着这些问题成为争论的焦点，验证研究现在有必要明确或隐含地解决音乐生成器的流派特异性。

As an example, in our proof-of-concept study, we found that the MusicCaps dataset used to train the Google MusicLM generator (Agostinelli et al., 2023) has some inconsistencies in its tagging, having tagged a fragment of Jose Hidalgo’s "Las bragas que te compré," a sevillana-style flamenco song, as a traditional Mexican folklore song, probably due to its lyrics in Spanish. Even though out of the particular scope of the proposed methodology, implementing the comined AHP-Likert method proposed in subsection 2.6 could greatly diminish the probability of these errors appearing, by cross-referencing user assessments with expert judgments. Validation studies should therefore check for consistency with the technical and cultural attributes of the intended genre specificity, and future research is needed to develop standardized tests able to validate the similarity between the intended genre and the output of a generator, especially in text-to-music models.

例如，在我们的概念验证研究中，我们发现用来训练 Google MusicLM 生成器的 MusicCaps 数据集 (Agostinelli et al。 ，2023) 在标签上有一些不一致的地方，将 Jose Hidalgo 的 “Las bragas que te compré”(一首塞维利亚风格的弗拉门戈歌曲) 的一个片段标记为一首传统的墨西哥民歌，可能是因为它的西班牙语歌词。即使在拟议方法的特定范围之外，通过交叉引用用户评估和专家判断，实施分节 2.6 中提出的联合 AHP-Likert 方法可以大大减少出现这些错误的可能性。因此，

验证研究应该检查是否与预期的流派特异性的技术和文化属性一致，并且需要进一步的研究来开发标准化测试，以验证预期的流派和生成器输出之间的相似性，特别是在文本到音乐的模型中。

5. Conclusions and future work

5. 结论和未来工作

The methodology proposed in this study is a step towards a general model for user-based AI music generation evaluation studies. In many cases, automatic quantitative evaluation methods are still insufficient for evaluating the quality of AI-generated works (Wang et al., 2021), so there is a clear need to create and improve standardized user-based evaluation protocols. Utilizing the suggested techniques allows researchers to considerably shorten the time needed to design a new study. They are also provided with a set of tools for obtaining significant data that can be easily cross-checked with different variables and other measuring tools. The proof of concept presented in the study demonstrates the viability of the methodology and of the different tools proposed.

本研究提出的方法论是迈向基于用户的人工智能音乐生成评估研究通用模型的一步。在许多情况下，自动定量评估方法仍然不足以评估人工智能生成作品的质量(Wang et al。 ，2021) ，因此明确需要创建和改进标准化的基于用户的评估方案。利用建议的技术可以使研究人员大大缩短设计新研究所需的时间。他们还提供了一套工具来获得重要的数据，可以很容易地与不同的变量和其他测量工具交叉检查。研究中提出的概念证明表明了该方法和提出的不同工具的可行性。

The standardization of a methodology for user-based evaluation studies aimed specifically at generative AI should be a major concern for the scientific community. The implications of this rapidly evolving technology are far-reaching and require studies capable of corroborating the effectiveness of AI generators and ethically guiding and enhancing them. Standardization will improve the ability of researchers to compare and contrast different systems in independently created studies. Moreover, this fast evaluation can guide researchers in improving their learning algorithms by fine-tuning areas that human evaluators may consider important and automatic evaluation tools may not detect. This coincides with Agile Software Development guidelines (Hinderks et al., 2022) where quick evaluation allows to inform the next iteration of the development process.

专门针对生成性人工智能的基于用户的评估研究方法的标准化应该是科学界关注的主要问题。这种快速发展的技术的影响是深远的，需要研究能够证实人工智能生成器的有效性，并在伦理上指导和加强它们。标准化将提高研究人员在独立创建的研究中比较和对比不同系统的能力。此外，这种快速评估可以指导研究人员通过微调人类评估人员可能认为重要和自动评估工具可能无法检测的领域来改进其学习算法。这与敏捷软件开发指导方针 (Hinderks et al。 ，2022) 一致，其中快速评估允许通知开发过程的下一次迭代。

Our proof of concept study (Section 3) illustrates the need to develop easy-to-use emotion trackers that can estimate emotions related to short-timed events. Such trackers would improve the sensitivity of emotional state measuring devices and would make it possible to monitor emotions in numerous research settings where it is currently unfeasible.

我们的概念验证研究 (第 3 部分) 说明了开发易于使用的情绪跟踪器的必要性，该跟踪器可以估计与短时间事件相关的情绪。这样的追踪器可以提高情绪状态测量设备的灵敏度，并且可以在许多目前还不可行的研究环境中监测情绪。

As a final thought, we believe that much research is still needed to unite our current understanding of ethnomusicology, music theory, and human musical practice with the technological advances being made in generative AI, to create an ethical, inclusive, diverse technology that will pave the way for the making of music in the future.

最后，我们认为仍然需要进行大量的研究，将我们目前对民族音乐学、音乐理论和人类音乐实践的理解与生成性人工智能的技术进步结合起来，创造一种伦理的、包容的、多样化的技术，为未来的音乐创作铺平道路。

6. Limitations of the study

6. 研究的局限性

This article presents a comprehensive methodology for user-based evaluation of AI-generated music. The methodology proposed uses human evaluation, which can be tailored specifically to the object of study, and provides several alternatives for elaborating surveys, gathering emotional feedback from participants, using different kinds of generators and musical pieces, and analyzing data. Although most state-of-the-art generators use non-standardized versions of some of the evaluation tools proposed and could even fit into the proposed methodology with an adequate selection of options in the Adaptability Decision Tree and Matrices, a systematic review of all possible resources for AI music evaluation is beyond the scope of this study.

本文介绍了一种基于用户评估人工智能生成音乐的综合方法。所提出的方法使用人的评价，这种评价可以专门针对研究对象，并提供了若干备选办法，用于拟订调查、收集参与者的情感反馈、使用不同类型的发生器和音乐片段以及分析数据。尽管大多数最先进的生成器使用所提出的一些评估工具的非标准化版本，甚至可以适合所提出的方法，在适应性决策树和矩阵中有足够的选择，但是对 AI 音乐评估的所有可能资源的系统综述超出了本研究的范围。

One of the main objective of this study is to provide future researchers with the tools to formulate AI music generation evaluation studies tailored to their specific contexts and needs, while still providing common ground for standardization and comparison between studies. Nevertheless, researchers using this methodology should be aware of the potential biases towards western music these methodology may have.

本研究的主要目的之一是为未来的研究者提供工具，以制定符合其特定背景和需求的人工智能音乐生成评估研究，同时仍然为研究之间的标准化和比较提供共同基础。尽管如此，使用这种方法的研究人员应该意识到这些方法可能对西方音乐有潜在的偏见。

Moreover, the provided strategies and tools to asses emotion in music following guidelines that suggest the need for emotion validation using multiple approaches (Eerola & Vuoskoski, 2013) (through self-assesment, emotion detection and physiological measurements). These approaches can coincide with both continuous and discrete emotion models. They are very useful when mixed with objective validation methods (Subsection 1.1) to improve the reliability of datasets labeling and model refinement, as these approaches are easily translated into numerical formulas. On the other hand, researcher are advised against taking these emotional estimations as absolute measurements. Music emotions depends on large myriad of factors that are very subjective and are not completely covered by the basic emotion or the affective circumplex models (Eerola & Saari, 2025). Context, where the music is made and received outside the laboratories, is also to be accounted for.

此外，提供的策略和工具来评估音乐中的情绪，遵循指导方针，表明需要使用多种方法进行情绪验证 (Eerola & Vuoskoski，2013)(通过自我评估，情绪检测和生理测量)。这些方法可以与连续和离散的情绪模型一致。当与客观验证方法 (1.1 小节) 相结合时，这些方法非常有用，可以提高数据集标记和模型改进的可靠性，因为这些方法很容易转化为数值公式。另一方面，建议研究者不要将这些情绪估计作为绝对测量。音乐情感取决于大量的因素，这些因素是非常主观的，并不能完全被基本情感或情感环绕模型所涵盖 (Eerola & Saari，2025)。音乐在实验室之外制作和接收的环境也是需要考虑的。

The article also provides a proof-of-concept in which the methodology is applied and in which many of the proposed tools are used. This proof of concept throws light on the applicability of the tools and can aid future researchers in designing a benchmark for their evaluation studies. Due to the small number of participants, however, the potentially interesting implications of its results should be further studied. This pilot study shows how the mothodology can be applied for AI generation. It should be noted that AI generation is a very rapidly evolving field, and that the generators used in the study may become obsolete in the near future. Nevertheless, the necessity for constant evaluation and supervision remain; while the perspective gained through the study can point towards a much wider acceptance for the technology, with the use of next-generation systems in the near future.

本文还提供了一个概念验证，其中应用了该方法，并使用了许多建议的工具。这种概念证明阐明了这些工具的适用性，可以帮助未来的研究者为他们的评估研究设计一个基准。然而，由于参与者的数量较少，其结果的潜在有趣的含义还有待进一步研究。这个试点研究显示了如何将方法学应用于 AI 生成。应该指出的是，人工智能生成是一个非常迅速发展的领域，研究中使用的发生器可能在不久的将来会过时。尽管如此，仍然需要不断进行评价和监督；而通过研究获得的观点可以指出，随着不久的将来下一代系统的使用，这项技术将得到更广泛的接受。

Disclosure statement

披露声明

No potential conflict of interest was reported by the author(s).

作者没有报告任何潜在的利益冲突。

Ethics statement

道德声明

This study was reviewed and approved by the ethical Committee of Universidad Loyola.

本研究经洛约拉大学伦理委员会审查通过。

Funding

拨款

This research was supported by the EQUAVEL project PID2022- 137646OB-C31, funded by MICIU/AEI/10.13039/501100011033 and by ERDF, EU.

本研究由 equivel 项目 PID2022-137646ob-c31 资助，由 MICIU/AEI/10.13039/501100011033 和欧盟 ERDF 资助。

ORCID

ORCID

Miguel Civit D http://orcid.org/0000-0003-4310-6377 Véronique Drai-Zerbib [D http://orcid.org/0000-0002-5623-6229 Francisco Cuadrado [D http://orcid.org/0000-0003-2307-3846 Maria J. Escalona [1] http://orcid.org/0000-0002-6435-1497

Véronique Drai-Zerbib [d http://orcid.org/0000-0003-4310-6377 Véronique Drai-Zerbib [d http://orcid.org/0000-0002-5623-6229 Francisco Cuadrado [d http://orcid.org/0000-0003-2307-3846 Maria j。 Escalona [1] http://orcid.org/0000-0002-6435-1497

Data availability statement

数据可用性声明

Music Excerpts from the pilot study are available in their respective datasets and in: Music Folder

试点研究的音乐摘录可以在各自的数据集和音乐文件夹中找到

References

参考文献

Agawu, K. (2006). Structural analysis or cultural analysis? competing perspectives on the "standard pattern" of west african rhythm. Journal of the American Musicological Society, 59(1), 1-46. https://doi.org/10.1525/jams.2006.59.1.1

阿加乌，k。(2006)。结构分析还是文化分析？关于西非节奏 “标准模式 “的不同观点。美国音乐学会杂志，59 (1) ，1-46。Https://doi.org/10.1525/jams

Agostinelli, A., Denk, T. I., Borsos, Z., Engel, J., Verzetti, M., Caillon, A., Huang, Q., Jansen, A., Roberts, A., Tagliasacchi, M., Sharifi, M., Zeghidour, N., & Frank, C. (2023). Musiclm: Generating music from

Agostinelli，a. ，Denk，T.i. ，Borsos，z. ，Engel，j. ，Verzetti，m. ，Caillon，a. ，Huang，q. ，Jansen，a. ，Roberts，a. ，Tagliasacchi，m. ，Sharifi，m. ，Zeghidour，n. ，& Frank，c. (2023).Musiclm: 从

text. arXiv preprint arXiv:2301.11325.

文本。 arXiv 预印本 arXiv: 2301.11325。

Ahmad, Z., & Khan, N. (2022). A survey on physiological signal-based emotion recognition. Bioengineering, 9(11), 688. https://doi.org/10.3390/bioengineering9110688

艾哈迈德，z。和可汗，n。(2022)。基于生理信号的情绪识别研究综述。生物工程，9 (11) ，688。Https://doi.org/10.3390/bioengineering9110688

Ariza, C. (2009). The interrogator as critic: The turing test and the evaluation of generative music systems. Computer Music Journal, 33(2), 48-70. https://doi.org/10.1162/comj.2009.33.2.48

阿里扎，c。(2009)。作为批评家的质询者： 图灵测试与生成音乐系统的评估。计算机音乐杂志，33 (2) ，48-70。Https://doi.org/10.1162/comj

Bittner, R. M., Bosch, J. J., Rubinstein, D., Meseguer-Brocal, G., & Ewert, S. (2022). A lightweight instrument-agnostic model for polyphonic note transcription and multipitch estimation. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP).

Bittner，r。，Bosch，j。，Rubinstein，d。，Meseguer-Brocal，g。，& Ewert，s。复调音符转录和多音高估计的轻量级乐器不可知模型。发表于 IEEE 声学、语音和信号处理国际会议 (ICASSP)。

Briot, J.-P., & Pachet, F. (2020). Deep learning for music generation: Challenges and directions. Neural Computing and Applications, 32(4), 981-993. https://doi.org/10.1007/s00521-018-3813-6

布里奥，j.-p。 & 帕切特，f。(2020)。音乐生成的深度学习： 挑战与方向。神经计算与应用，32 (4) ，981-993。Https://doi.org/10.1007/s00521-018-3813-6

Bugnon, L. A., Calvo, R. A., & Milone, D. H. (2020). Dimensional affect recognition from HRV: An approach based on supervised SOM and ELM. IEEE Transactions on Affective Computing, 11(1), 32-44. https://doi.org/10.1109/TAFFC.2017.2763943

Bugnon，L.a。，Calvo，R.a。，& Milone，D.h。基于 HRV 的维度情感识别： 一种基于监督 SOM 和 ELM 的方法。IEEE 情感计算汇刊，11 (1) ，32-44。https://doi.org/10.1109/TAFFC.2017.2763943

Cayrou, S., Dickes, P., Gauvain-Piquard, A., Dolbeault, S., Callahan, S., & Roge, B. (2000). Validation de la traduction française du poms (profile of mood states). Psychologie et psychométrie, 21(4), 5-22.

Cayrou，s。，Dickes，p。，gauvan-piquard，a。，Dolbeault，s。，Callahan，s。，& Roge，b。(2000)。法语翻译的验证 (情绪状态概况)。心理学与心理测量学，21 (4) ，5-22。

Chollet, F. (2019). On the measure of intelligence. arXiv preprint arXiv: 1911.01547.

(2019)《论智力的测量》，arXiv 预印本，arXiv: 1911.01547。

Chu, H., Kim, J., Kim, S., Lim, H., Lee, H., Jin, S., & Ko, S. (2022). An empirical study on how people perceive ai-generated music. In Proceedings of the 31st ACM international conference on information & knowledge management (pp. 304-314).

(2022).一项关于人们如何感知人工智能产生的音乐的实证研究。第 31 届 ACM 国际信息与知识管理会议论文集 (304-314 页)。

Chuan, C.-H., Agres, K., & Herremans, D. (2020). From context to concept: Exploring semantic relationships in music with word2vec. Neural Computing and Applications, 32(4), 1023-1036. https://doi.org/10.1007/s00521-018-3923-1

川，c.-h。，Agres，k。 & Herremans，d (2020)。从语境到概念： 用 word2vec 探索音乐中的语义关系。Neural Computing and Applications，32 (4) ，1023-1036 神经计算与应用，32 (4) ，1023-1036。Https://doi.org/10.1007/s00521-018-3923-1

Cideron, G., Girgin, S., Verzetti, M., Vincent, D., Kastelic, M., & Borsos, Z. (2024). Musicrl: Aligning music generation to human preferences. arXiv preprint arXiv:2402.04229.

Cideron，g。，Girgin，s。，Verzetti，m。，Vincent，d。，Kastelic，m。，& Borsos，z。Musicrl: 调整音乐生成以适应人类的喜好。arXiv 预印本 arXiv: 2402.04229。

Civit, M., Civit-Masot, J., Cuadrado, F., & Escalona, M. J. (2022). A systematic review of artificial intelligence-based music generation: Scope, applications, and future trends. Expert Systems with Applications, 209, 118190. https://doi.org/10.1016/j.eswa.2022.118190

Civit，m。，Civit-Masot，j。，Cuadrado，f。，& Escalona，m。(2022)。基于人工智能的音乐生成系统综述： 范围、应用与未来趋势。专家系统与应用，209,118190。Https://doi.org/10.1016/j.eswa.2022.118190

Civit, M., Drai-Zerbib, V., Lizcano, D., & Escalona, M. J. (2024). Sunocaps: A novel dataset of text-prompt based ai-generated music with emotion annotations. Data in Brief, 55, 110743. https://doi.org/10.1016/j.dib.2024.110743

Civit，m. ，Drai-Zerbib，v. ，Lizcano，d. ，& Escalona，M.j. (2024).Sunocaps: 一个新颖的基于文本提示的人工智能音乐数据集，带有情感注

释。Data in Brief，55,110743 简要数
据，55,110743。Https://doi.org/10.1016/j.dib.2024.110743

Civit, M., Escalona, M. J., Cuadrado, F., & Reyes-de Cozar, S. (2024). Class integration of chatgpt and learning analytics for higher education. Expert Systems, 41(12), e13703. https://doi.org/10.1111/exsy.13703

Civit，m. ，Escalona，M.j. ，Cuadrado，f. ，& Reyes-de Cozar，s. (2024).Chatgpt 与高等教育学习分析的课堂整合。专家系统，41 (12)，e13703。Https://doi.org/10.1111/exsy.13703

Copet, J., Kreuk, F., Gat, I., Remez, T., Kant, D., Synnaeve, G., & Défossez, A. (2024). Simple and controllable music generation. Advances in Neural Information Processing Systems, 36, 47704- 47720. https://doi.org/10.48550/arXiv.2306.05284

Copet，j。，Kreuk，f。，Gat，i。，Remez，t。，Kant，d。，Synnaeve，g。，& Défossez，a。(2024)。简单可控的音乐生成。神经信息处理系统的进展，36,47704-47720。https://doi.org/10.48550/arXiv.2306.05284

Cuadrado, F., Lopez-Cobo, I., Mateos-Blanco, T., & Tajadura-Jiménez, A. (2020). Arousing the sound: A field study on the emotional impact on children of arousing sound design and 3d audio spatialization in an audio story. Frontiers in Psychology, 11, 737. https://doi.org/10.3389/fpsyg.2020.00737

Cuadrado，f。，Lopez-Cobo，i。，Mateos-Blanco，t。，& Tajadura-Jiménez，a。唤醒声音： 一项关于在有声故事中唤醒声音设计和 3d 音频空间化对儿童的情感影响的实地研究。心理学前沿，11,737。Https://doi.org/10.3389/fpsyg

Cuculo, V., & D’Amelio, A. (2019). Openfacs: An open source facs-based 3d face animation system. In Image and graphics: 10th international conference, ICIG 2019, Beijing, China, August 23-25, 2019, proceedings, part ii 10 (pp. 232-242).

Cuculo，v. & d’amelio，a. (2019)。Openfacs: 一个开源的基于 facs 的 3d 人脸动画系统。图像与图形： 2019 年 ICIG 第 10 届国际会议，中国北京，2019 年 8 月 23-25 日，论文集，第 10 部分 (第 232-242 页)。

Dervakos, E., Filandrianos, G., & Stamou, G. (2021). Heuristics for evaluation of ai generated music. In 2020 25th international conference on pattern recognition (ICPR) (pp. 9164-9171).

Dervakos，e. ，Filandrianos，g. 和 Stamou，g. (2021)。启发式评估ai生成的音乐。2020年第25届模式识别国际会议 (ICPR)(第9164-9171页)。

Dimitriev, D., Indeykina, O., & Dimitriev, A. (2022). The effect of auditory stimulation on the nonlinear dynamics of heart rate: The impact of emotional valence and arousal. bioRxiv, 2022-2003.

Dimitriev，d. ，Indeykina，o. & Dimitriev，a. (2022).听觉刺激对心率非线性动力学的影响： 情绪效价和唤醒的影响。bioRxiv，2022-2003.

Diwanji, V. S., Geana, M., Pei, J., Nguyen, N., Izhar, N., & Chaif, R. H. (2025). Consumers' emotional responses to ai-generated versus human-generated content: The role of perceived agency, affect and gaze in health marketing. International Journal of Human-Computer

Diwanji，V.s. ，Geana，m. ，Pei，j. ，Nguyen，n. ，Izhar，n. ，& Chaif，R.h. (2025).消费者对人工智能产生的内容与人工产生的内容的情绪反应：感知代理的作用，影响和凝视在健康营销。国际人机杂志

Interaction, 1-21. https://doi.org/10.1080/10447318.2025.2454954

相互作用，1-21。 https://doi.org/10.1080.10447318.2025.2454954

Dolan, J. G. (2008). Shared decision-making-transferring research into practice: The analytic hierarchy process (AHP). Patient Education and Counseling, 73(3), 418-425. https://doi.org/10.1016/j.pec.2008.07.032

杜兰 (2008)。共享决策 —— 将研究转化为实践： 层级分析法 (AHP)。病人教育和咨询，73 (3) ，418-425。Https://doi.org/10.1016/j.pec. 2008.07.032

Dong, H.-W., Chen, K., McAuley, J., & Berg-Kirkpatrick, T. (2020). Muspy: A toolkit for symbolic music generation. In Proceedings of the 21st International Society for Music Information Retrieval conference (ISMIR).

Dong，h.-w。 ，Chen，k。 ，McAuley，j。 ，& Berg-Kirkpatrick，t。Muspy: 象征性音乐生成的工具箱。发表于第21届国际音乐信息检索会议论文集。

Eerola, T., & Saari, P. (2025). What emotions does music express? structure of affect terms in music using iterative crowdsourcing paradigm. PLoS One, 20(1), e0313502. https://doi.org/10.1371/journal.pone.0313502

Eerola，t。 & Saari，p。音乐表达什么样的情感？基于迭代众包范式的音乐情感术语结构。PLoS One，20(1) ，e0313502 PLoS One，20(1) ，e0313502.Https://doi.org/10.1371/journal.pone.0313502

Eerola, T., & Vuoskoski, J. K. (2013). A review of music and emotion studies: Approaches, emotion models, and stimuli. Music Perception, 30(3), 307-340. https://doi.org/10.1525/mp.2012.30.3.307

埃罗拉，t。 & 沃斯科斯基，j。(2013)。音乐与情绪研究综述： 方法、情绪模型和刺激。音乐知觉，30 (3) ，307-340。Https://doi.org/10.1525/mp. 2012.30.3.307

Eerola, T., Vuoskoski, J. K., Peltola, H.-R., Putkinen, V., & Schäfer, K. (2018). An integrative review of the enjoyment of sadness associated with music. Physics of Life Reviews, 25, 100-121. https://doi.org/10.1016/j.plrev.2017.11.016

Eerola，t。 ，Vuoskoski，j。 ，Peltola，h.-r。 ，Putkinen，v。 ，& Schäfer，k。(2018)。与音乐相关的悲伤享受的综合评论。生命物理学评论，25,100-121。Https://doi.org/10.1016/j.plrev. 2017.11.016

Ferreira, P., Limongi, R., & Fávero, L. P. (2023). Generating music with data: Application of deep learning models for symbolic music composition. Applied Sciences, 13(7), 4543. https://doi.org/10.3390/app13074543

费雷拉，p. ，利蒙吉，r. & Fávero，L.p. (2023)。用数据生成音乐： 深度学习模型在符号音乐创作中的应用。应用科学，13 (7) ，4543。Https://doi.org/10.3390/app13074543

Frid, E., Gomes, C., & Jin, Z. (2020). Music creation by example. In Proceedings of the 2020 Chi conference on human factors in computing systems (pp. 1-13).

弗里德，e。 ，戈麦斯，c。 ，& Jin，z。(2020)。音乐创作实例。In Proceedings of the 2020 Chi conference on human factors In computing systems (pp. 1-13) 2020 年智能计算系统中的人为因素会议论文集 (第 1-13 页)。

García-Peñalvo, F., & Vázquez-Ingelmo, A. (2023). What do we mean by genai? A systematic mapping of the evolution, trends, and techniques involved in generative ai. International Journal of Interactive Multimedia and Artificial Intelligence, 8(4), 7. https://doi.org/10.9781/ijimai.2023.07.006

García-Peñalvo，f。 & Vázquez-Ingelmo，a。(2023)。我们说的 genai 是什么意思？对生成 ai 所涉及的进化、趋势和技术的系统性描绘。国际交互式多媒体和人工智能杂志，8 (4) ，7。Https://doi.org/10.9781/ijimai.2023.07.006

Goepel, K. D. (2018). Implementation of an online software tool for the analytic hierarchy process (ahp-os). International Journal of the Analytic Hierarchy Process, 10(3), 590. https://doi.org/10.13033/ijahp.v10i3.590

Goepel，K.d. (2018).层级分析法 (ahp-os) 在线软件工具的实现。国际期刊的层级分析法，10 (3) ，590。Https://doi.org/10.13033/ijahp.v10i3.590

Gutknecht, M., Danner, M., Schaarschmidt, M.-L., Gross, C., & Augustin, M. (2018). Assessing the importance of treatment goals in patients with psoriasis: Analytic hierarchy process vs. likert scales. The Patient, 11(4), 425-437. https://doi.org/10.1007/s40271-018-0300-1

Gutknecht，m. ，Danner，m. ，Schaarschmidt，m.-l. ，Gross，c. ，& Augustin，m. (2018).评估银屑病患者治疗目标的重要性： 层级分析法与李克特量表。病人，11 (4) ，425-437。Https://doi.org/10.1007/s40271-018-0300.1

Hadjeres, G., Pachet, F., & Nielsen, F. (2017). Deepbach: A steerable model for bach chorales generation. In International conference on machine learning (pp. 1362-1371).

哈杰雷斯，g. ，Pachet，f. ，& Nielsen，f. (2017)。Deepbach: 巴赫合唱团生成的可控模型。In International conference on machine learning (pp. 1362-1371) 机器学习国际会议 (1362-1371 页)。

Hawthorne, C., Stasyuk, A., Roberts, A., Simon, I., Huang, C.-Z. A., Dieleman, S., & Eck, D. (2019). Enabling factorized piano music modeling and generation with the maestro dataset. In Proceedings of the International Conference on Learning Representations (ICLR 2019).

霍桑、斯塔素克、罗伯茨、西蒙、黄。A，Dieleman，s. & Eck，d. (2019).使用 maestro 数据集启用因子化钢琴音乐建模和生成。In Proceedings of the International Conference on Learning Representations (ICLR 2019) 在国际学习表征会议论文集 (ICLR 2019)。

Hernández-Orallo, J. (2020). Twenty years beyond the turing test: Moving beyond the human judges too. Minds and Machines, 30(4), 533-562. https://doi.org/10.1007/s11023-020-09549-0

Hernández-Orallo，j。(2020)。图灵测试之后的 20 年： 也超越了人类的判断。Minds and Machines，30 (4) ，533-562《思维与机器》30 (4) ，533-562。Https://doi.org/10.1007/s11023-020-09549-0

Hernando, D., Roca, S., Sancho, J., Alesanco, Á., & Bailón, R. (2018). Validation of the apple watch for heart rate variability measurements during relax and mental stress in healthy subjects. Sensors, 18(8), 2619. https://doi.org/10.3390/s18082619

Hernando，d。，Roca，s。，Sancho，j。，Alesanco，á。，& Bailón，r。(2018)。验证苹果手表在健康受试者放松和精神压力期间的心率变异分析。传感器，18 (8) ，2619。Https://doi.org/10.3390/s18082619

Hinderks, A., Mayo, F. J. D., Thomaschewski, J., & Escalona, M. J. (2022). Approaches to manage the user experience process in agile software development: A systematic literature review. Information and Software Technology, 150, 106957. https://doi.org/10.1016/j.infsof.2022.106957

Hinderks，a。，Mayo，f。，Thomaschewski，j。，& Escalona，m。敏捷软件开发中管理用户体验过程的方法： 一个系统的文献综述。信息与软件技术，150,106957。Https://doi.org/10.1016/j.infsof.2022.106957

Hirten, R. P., Danieletto, M., Tomalin, L., Choi, K. H., Zweig, M., Golden, E., Kaur, S., Helmus, D., Biello, A., Pyzik, R., Calcagno, C., Freeman, R., Sands, B. E., Charney, D., Bottinger, E. P., Murrough, J. W., Keefer, L., Suarez-Farinas, M., Nadkarni, G. N., & Fayad, Z. A. (2021). Factors associated with longitudinal psychological and physiological stress in health care workers during the covid-19 pandemic: Observational study using apple watch data. Journal of Medical Internet Research, 23(9), e31295. https://doi.org/10.2196/31295

Hirten，R.p. ，Danieletto，m. ，Tomalin，l. ，Choi，K.h. ，Zweig，m. ，Golden，e. ，Kaur，s. ，Helmus，d. ，Biello，a. ，Pyzik，r. ，Calcagno，c. ，Freeman，r. ，Sands，B.e. ，Charney，d. ，Bottinger，E.p. ，Murrough，J.w. ，Keefer，l. ，Suarez-Farinas，m. ，Nadkarni，G.n. ，& Fayad，Z.a. (2021).新型冠状病毒肺炎疫情期间医护人员纵向心理生理应激影响因素研究 —— 基于苹果手表数据的观察性研究。医学互联网研究杂志，23 (9) ，e31295。Https://doi.org/10.2196/31295

Huang, J., Wang, J.-C., Smith, J. B., Song, X., & Wang, Y. (2021). Modeling the compatibility of stem tracks to generate music mash-ups. In

Proceedings of the AAAI conference on artificial intelligence (Vol. 35, pp. 187-195).

Huang，j。，Wang，j.-c。，Smith，J.b。，Song，x。，& Wang，y。(2021)。模拟干音轨的兼容性以生成音乐混搭。美国人工智能协会会议论文集 (第 35 卷，第 187-195 页)。

Ing, E. B. (2021). A survey-weighted analytic hierarchy process to quantify authorship. Advances in Medical Education and Practice, 12, 1021-1031. https://doi.org/10.2147/AMEP.S328648

英格，e。(2021)。量化作者身份的调查加权层级分析法。医学教育与实践进展，12,1021-1031。https://doi.org/10.2147/AMEP.S328648

Ismail, S. N. M. S., Aziz, N. A. A., Ibrahim, S. Z., & Mohamad, M. S. (2024). A systematic review of emotion recognition using cardio-based signals. ICT Express, 10(1), 156-183.

Ismail，S.N.M.s。，Aziz，N.A.a。，Ibrahim，S.z。，& Mohamad，M.s。(2024)。使用基于心脏的信号的情绪识别的系统综述。ICT Express，10(1)，156-183.

Jiang, J., Xia, G. G., Carlton, D. B., Anderson, C. N., & Miyakawa, R. H. (2020). Transformer vae: A hierarchical model for structure-aware and interpretable music representation learning. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 516-520).

Jiang，j。，Xia，G.g。，Carlton，D.b。，Anderson，C.n。，& Miyakawa，R.h。(2020)。Transformer vae: 结构感知和可解释的音乐表征学习的分层模型。在 ICASSP 2020-2020 IEEE 声学，语音和信号处理国际会议 (ICASSP)(第 516-520 页)。

Jordanous, A. (2012). A standardised procedure for evaluating creative systems: Computational creativity evaluation based on what it is to be creative. Cognitive Computation, 4(3), 246-279. https://doi.org/10.1007/s12559-012-9156-1

Jordanous，a. (2012).评估创造性系统的标准程序： 基于什么是创造性的计算创造性评估。认知计算，4 (3) ，246-279。Https://doi.org/10.1007/s12559-012-9156-1

Katz, B. (2015). Sound board: Can we stop the loudness war in streaming? Journal of the Audio Engineering Society, 63(11), 939-940. https://aes2.org/publications/elibrary-page/?id=18053

卡茨，b。(2015)。声板： 我们能阻止流媒体中的声音大战吗？音频工程学会杂志，63 (11) ，939-940。Https://aes2.org/publications/elibrary-page/

Katz, B., & Katz, R. A. (2003). Mastering audio: The art and the science. Butterworth-Heinemann.

Katz，b。和 Katz，r。(2003)。掌握音频： 艺术与科学。巴特沃斯 - 海涅曼。

Kayser, D., Egermann, H., & Barraclough, N. E. (2022). Audience facial expressions detected by automated face analysis software reflect emotions in music. Behavior Research Methods, 54(3), 1493-1507. https://doi.org/10.3758/s13428-021-01678-3

Kayser，d，Egermann，h，& Barraclough，N.e.自动人脸分析软件检测到的观众面部表情反映了音乐中的情绪。行为研究方法，54 (3) ，1493-1507。Https://doi.org/10.3758/s13428-021-01678-3

Kilgour, K., Zuluaga, M., Roblek, D., & Sharifi, M. (2019). Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms. In Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH) (pp. 2350-2354). https://doi.org/10.21437/ Interspeech.2019-2219

Kilgour，k。 ，Zuluaga，m。 ，Roblek，d。 ，& Sharifi，m。Fréchet 音频距离： 用于评估音乐增强算法的无参考度量。In Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)(pp. 2350-2354) 国际语音交流协会年会论文集 (2350-2354 页)。Https://doi.org/10.21437/Interspeech.2019-2219

Knejzlíková, T., Světlák, M., Malatincová, T., Roman, R., Chládek, J., Najmanová, J., Theiner, P., Linhartová, P., & Kašpárek, T. (2021). Electrodermal response to mirror exposure in relation to subjective emotional responses, emotional competences and affectivity in adolescent girls with restrictive anorexia and healthy controls. Frontiers in Psychology, 12, 673597. https://doi.org/10.3389/fpsyg.2021.673597

Knejlíková，t. ，sv tlák，m. ，Malatincová，t. ，Roman，r. ，Chládek，j. ，Najmanová，j. ，Theiner，p. ，Linhartová，p. ，& ka párek，t. (2021).患有限制性厌食症和健康控制的青春期女孩对镜子暴露的皮肤反应与主观情绪反应、

情绪能力和情感性的关系。心理学前沿，12,673597。Https://doi.org/10.3389/fpsyg.2021.673597

Kodra, E., Senechal, T., McDuff, D., & El Kaliouby, R. (2013). From dials to facial coding: Automated detection of spontaneous facial expressions for media research. In 2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG) (pp. 1-6).

Kodra，e。，Senechal，t。，McDuff，d。，& El Kaliouby，r。(2013)。从拨号到面部编码： 为媒体研究自动检测自发面部表情。2013 年第 10 届 IEEE 国际会议和研讨会关于自动面部和手势识别 (FG)(第 1-6 页)。

Kulke, L., Feyerabend, D., & Schacht, A. (2020). A comparison of the affectiva imotions facial expression analysis software with emg for identifying facial expressions of emotion. Frontiers in Psychology, 11, 329. https://doi.org/10.3389/fpsyg.2020.00329

Kulke，l。，Feyerabend，d。 & Schacht，a。(2020)。Affectiva imotions 面部表情分析软件与肌电图识别情绪面部表情的比较。心理学前沿，11,329。Https://doi.org/10.3389/fpsyg.2020.00329

Lui, G. Y., Loughnane, D., Polley, C., Jayarathna, T., & Breen, P. P. (2022). The apple watch for monitoring mental health-related physiological symptoms: Literature review. JMIR Mental Health, 9(9), e37354. https://doi.org/10.2196/37354

Lui，g。，Loughnane，d。，Polley，c。，Jayarathna，t。，& Breen，p。苹果手表用于监测心理健康相关的生理症状： 文献综述。JMIR Mental Health，9 (9) ，e37354 JMIR 精神健康，9 (9) ，e37354。Https://doi.org/10.2196/37354

McCarthy, C., Pradhan, N., Redpath, C., & Adler, A. (2016). Validation of the empatica e4 wristband. In 2016 IEEE Embs International Student Conference (ISC) (pp. 1-4).

McCarthy，c. ，Pradhan，n. ，Redpath，c. ，& Adler，a. (2016).Empatica e4 腕带的验证。2016 年 IEEE Embs 国际学生会议 (ISC)(第 1-4 页)。

Moysis, L., Iliadis, L. A., Sotiroudis, S. P., Kokkinidis, K., Sarigiannidis, P., & Nikolaidis, S. (2023). The challenges of music deep learning for traditional music. In 2023 12th international conference on modern circuits and systems technologies (Mocast) (pp. 1-5).

Moysis，l. ，Iliadis，L.a. ，Sotiroudis，S.p. ，Kokkinidis，k. ，Sarigiannidis，p. ，& Nikolaidis，s. (2023).音乐深度学习对传统音乐的挑战。2023 年第 12 届现代电路与系统技术国际会议 (Mocast)(第 1-5 页)。

Muñoz-Saavedra, L., Escobar-Linero, E., Miró-Amarante, L., Bohórquez, R., & M, D.-M. (2023). Designing and evaluating a wearable device for affective state level classification using machine learning techniques. Expert Systems with Applications, 219, 119577. https://doi.org/10.1016/j.eswa.2023.119577

Muñoz-Saavedra，l。，Escobar-Linero，e。，Miró-Amarante，l。，Bohórquez，r。，& m，d。(2023).使用机器学习技术设计和评估用于情感状态水平分类的可穿戴设备。专家系统与应用，219,119577。Https://doi.org/10.1016/j.eswa.2023.119577

Osoba, O. A., Welser, W. IV., & Welser, W. (2017). An intelligence in our image: The risks of bias and errors in artificial intelligence. Rand Corporation.

Osoba，O.a. ，Welser，w. iv. ，& Welser，w. (2017).我们想象中的智能： 人工智能中的偏见和错误的风险。兰德公司。

Ponsiglione, A. M., Amato, F., Cozzolino, S., Russo, G., Romano, M., & Improta, G. (2022). A hybrid analytic hierarchy process and likert scale approach for the quality assessment of medical education programs.

Ponsiglione，A.m。，Amato，f。，Cozzolino，s。，Russo，g。，Romano，m。，& Improta，g。医学教育项目质量评估的混合层级分析法和李克特量表方法。

Mathematics, 10(9), 1426. https://doi.org/10.3390/math10091426

数学，10 (9) ，1426. https://doi.org/10.3390/math10091426

Ribeiro, F., Florêncio, D., Zhang, C., & Seltzer, M. (2011). Crowdmos: An approach for crowdsourcing mean opinion score studies. In 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 2416-2419).

Ribeiro，f。，Florêncio，d。，Zhang，c。，& Seltzer，m。(2011)。Crowdmos: 众包平均意见得分研究的一种方法。In 2011 IEEE International Conference on Acoustics，Speech and Signal Processing (ICASSP)(pp. 2416-2419) 2011 年 IEEE 声学、语音和信号处理国际会议 (ICASSP)(第 2416-2419 页)。

Rosenberg, E. L., & Ekman, P. (2020). What the face reveals: Basic and applied studies of spontaneous expression using the facial action coding system (FACS). Oxford University Press.

罗森伯格，e。 & 埃克曼，p。(2020)。面部表现： 使用面部动作编码系统 (FACS) 的自发表情的基础和应用研究。牛津大学出版社。

Saaty, T. L., & Özdemir, M.S. (2014). How many judges should there be in a group? Annals of Data Science, 1, 359-368. https://doi.org/10.1007/s40745-014-0026-4

萨蒂，T.l。 & 厄兹德米尔，m.s。(2014)。一个小组应该有多少个法官？数据科学年鉴，1,359-368。Https://doi.org/10.1007/s40745-014-0026-4

Schippers, A., Aben, B., Griep, Y., & Van Overwalle, F. (2018). Ultrashort term heart rate variability as a tool to assess changes in valence. Psychiatry Research, 270, 517-522. https://doi.org/10.1016/j.psychres.2018.10.005

Schippers，a。 ，Aben，b。 ，Griep，y。 ，& Van Overwalle，f。(2018)。超短期心率变异分析作为评估价值变化的工具。精神病学研究，270,517-522。Https://doi.org/10.1016/j.psychres.2018.10.005

Schubert, E. (2004). Modeling perceived emotion with continuous musical features. Music Perception, 21(4), 561-585. https://doi.org/10.1525/mp.2004.21.4.561

舒伯特，e。(2004)。用连续的音乐特征建模感知情绪。音乐知觉，21 (4) ，561-585。 https://doi.org/10.1525/mp。2004.21.4。561

Skiendziel, T., Rösch, A. G., & Schultheiss, O. C. (2019). Assessing the convergent validity between the automated emotion recognition software noldus facereader 7 and facial action coding system scoring. PLoS One, 14(10), e0223905. https://doi.org/10.1371/journal.pone.0223905

斯基恩齐尔，t。 ，Rösch，A.g。 ，& Schultheiss，O.c。(2019)。评估自动情绪识别软件 noldus facereader 7 和面部动作编码系统评分之间的收敛有效性。PLoS One，14(10) ，e0223905.Https://doi.org/10.1371/journal.pone.0223905

Song, B., & Kang, S. (2016). A method of assigning weights using a ranking and nonhierarchy comparison. Advances in Decision Sciences, 2016, 1-9. https://doi.org/10.1155/2016/8963214

宋，b。 & Kang，s。(2016)。一种使用排名和非等级比较来分配权重的方法。决策科学进展，2016,1-9。Https://doi.org/10.1155/2016/8963214

Sturm, B. L., & Ben-Tal, O. (2017). Taking the models back to music practice: Evaluating generative transcription models built using deep learning. Journal of Creative Music Systems, 2(1), 32-60. https://doi.org/10.5920/JCMS.2017.09

Sturm，B.l。 ，& Ben-Tal，o. (2017)。将模型带回音乐实践： 评估使用深度学习构建的生成转录模型。Journal of Creative Music Systems，2 (1) ，32-60 创意音乐系统杂志，2 (1) ，32-60。https://doi.org/10.5920/JCMS.2017.09

Sullivan, G. M., & Artino, A. R. Jr, (2013). Analyzing and interpreting data from likert-type scales. Journal of Graduate Medical Education, 5(4), 541-542. https://doi.org/10.4300/JGME-5-4-18

Sullivan，G.m. ，& Artino，A.r. Jr，(2013).分析和解释李克特型量表的数据。研究生医学教育杂志，5 (4) ，541-542。https://doi.org/10.4300/JGME-5-4-18

Tanoue, Y., Nakashima, S., Komatsu, T., Kosugi, M., Kawakami, S., Kawakami, S., Michishita, R., Higaki, Y., & Uehara, Y. (2023). The validity of ultra-short-term heart rate variability during cycling exercise. Sensors, 23(6), 3325. https://doi.org/10.3390/s23063325

Tanoue，y. ，Nakashima，s. ，Komatsu，t. ，Kosugi，m. ，Kawakami，s. ，Kawakami，s. ，Michishita，r. ，Higaki，y. ，& Uehara，y. (2023).自行车运动中超短期心率变异分析的有效性。传感器，23 (6) ，3325。Https://doi.org/10.3390/s23063325

Tarvainen, M. P., Niskanen, J.-P., Lipponen, J. A., Ranta-Aho, P. O., & Karjalainen, P. A. (2014). Kubios hrv-heart rate variability analysis software. Computer Methods and Programs in Biomedicine, 113(1), 210-220. https://doi.org/10.1016/j.cmpb.2013.07.024

Tarvainen，M.p. ，Niskanen，j.-p. ，Lipponen，J.a. ，Ranta-Aho，P.o. ，& Karjalainen，P.a. (2014).Kubios hrv - 心率变异性分析软件。生物医学中的计算机方法和程序，113 (1) ，210-220。Https://doi.org/10.1016/j.cmpb.2013.07.024

Taylor, S., Jaques, N., Chen, W., Fedor, S., Sano, A., & Picard, R. (2015). Automatic identification of artifacts in electrodermal activity data. In 2015 37th annual international conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (pp. 1934-1937).

Taylor，s. ，Jaques，n. ，Chen，w. ，Fedor，s. ，Sano，a. ，& Picard，r. (2015).电皮肤活动数据中伪影的自动识别。2015 年第 37 届 ieee 医学和生物学工程学会 (EMBC) 国际年会 (1934-1937 页)。

Wang, S., Bao, Z., & Jingtong, E. (2021). Armor: A benchmark for meta-evaluation of artificial music. In Proceedings of the 29th ACM International Conference on Multimedia (pp. 5583-5590).

Wang，s。 ，Bao，z。 ，& Jingtong，e。(2021)。Armor: 人工音乐元评估的基准。第 29 届 ACM 国际多媒体会议论文集 (5583-5590 页)。

Wiggins, G. A. (2019). A framework for description, analysis and comparison of creative systems. In T. Veale & F. Cardoso (Eds.), Computational creativity. Computational synthesis and creative systems. Springer. https://doi.org/10.1007/978-3-319-43610-4_2

Wiggins，G.a (2019).一个描述、分析和比较创造性系统的框架。Veale & f。Cardoso (编辑)，计算创造性。计算综合和创造性系统。斯普林格。Https://doi.org/10.1007/978-3-319-43610-4 _ 2

Xiong, Z., Wang, W., Yu, J., Lin, Y., & Wang, Z. (2023). A comprehensive survey for evaluation methodologies of ai-generated music. arXiv preprint arXiv:2308.13736.

Xiong，z。 ，Wang，w。 ，Yu，j。 ，Lin，y。和 Wang，z。人工智能生成音乐评价方法综述。arXiv 预印本 arXiv: 2308.13736。

Yang, L.-C., & Lerch, A. (2020). On the evaluation of generative models in music. Neural Computing and Applications, 32(9), 4773-4784. https://doi.org/10.1007/s00521-018-3849-7

Yang，l.-c. ，& Lerch，a. (2020).关于音乐生成模型的评价。Neural Computing and Applications，32 (9) ，4773-4784 神经计算与应用，32 (9) ，4773-4784。Https://doi.org/10.1007/s00521-018-3849-7

About the authors

关于作者

Miguel Civit has a PHD in Computer Science from the Université de Bourgogne and University of Seville with a focus on quantitative and qualitative validation of AI in Music. His research focuses on the effects of emotion in AI music generation. He is a member of the ES3 research group.

Miguel Civit 拥有勃艮第大学和塞维利亚大学的计算机科学博士学位，主要研究音乐中人工智能的定量和定性验证。他的研究重点是情感在人工智能音乐生成中的作用。他是 es3 研究小组的成员。

Véronique Drai-Zerbib is a professor of cognitive psychology at the University of Bourgogne Europe, specializing in expertise development, musical cognition, expert memory, multimodal information integration, digital reading, musical reading. She uses behavioral and neurophysiological approaches such as eye-tracking and also virtual reality and machine learning.

Véronique Drai-Zerbib 是欧洲勃艮第大学的认知心理学教授，专门研究专业知识发展，音乐认知，专家记忆，多模态信息整合，数字阅读，音乐阅读。她使用行为和神经生理学的方法，比如眼球追踪，还有虚拟现实和机器学习。

Francisco Cuadrado is PhD in Communication, researcher, professor, composer and sound designer. His research fields are music and sound creation and perception in media, and social and emotional development through music. He has been the IP for different research projects, like "Learning To Be" and "The Unconscious Listening."

Francisco Cuadrado 是传播学博士、研究员、教授、作曲家和音响设计师。他的研究领域是音乐和声音在媒体中的创造和感知，以及通过音乐的社会和情感发展。他是不同研究项目的知识产权人，如 “学习成为” 和 “无意识倾听”

María José Escalona, PhD in Computer Engineering from the University of Seville, is a full professor and director of the ES3 research group. Specializing in web engineering and software quality, she developed the NDT methodology, widely used in industry. She has an extensive research career, with numerous publications and projects.

María José Escalona，来自塞维利亚大学的计算机工程博士，是 es3 研究小组的全职教授和主任。专攻网络工程和软件质量，她开发了广泛应用于工业的无损检测方法。她有着广泛的研究生涯，发表过无数的论文和项目。