arXiv:2601.10691v1 [cs.CY] 15 Jan 2026

对话式考试：面向人工智能时代的可扩展评估设计

Lorena A. Barba[1] 和 Laura Stegner[2]

[1,2]乔治·华盛顿大学机械与航空航天工程系

摘要

当学生借助生成式人工智能完成作业，却没有真正投入其中时，传统评估方法便会失效，造成一种能力错觉：学生以为自己在学习，实际上并没有。本文介绍了对话式考试——一种可扩展的口头考试形式，通过让学生现场编写代码并解释其推理过程，恢复评估效度。基于人机交互原则，我们在短短两天内以小组形式考查了58名学生，证明口试能够扩展到常见的班级规模。这种形式将真实实践（学生使用文档并在监督下使用人工智能）与内在效度（实时表现无法伪造）结合起来。我们提供了详细的实施指南，帮助教师采用这种方法，为许多在全面禁止人工智能与接受有效评估不可能实现之间进退两难的教育工作者，提供一条切实可行的出路。

引言

关于生成式人工智能在教育领域的讨论，主要围绕两种论调展开：科技乐观主义者和企业大肆宣扬通过个性化服务和按需智能“革新”教育；以及对学生作弊、把课程作业转交给人工智能完成的末日式哀叹。这两种论调都无助于应对不可避免的变革。据大多数说法，学生正广泛使用人工智能获取答案、代为完成作业。问题在于，高等教育该如何适应这一现实，并确保课堂上仍然发生学习。

我们当中有一人在 2024年休完学术休假后，以十足的乐观心态开启新学期，并在一门工程计算本科课程的教学中采用人工智能作为辅助。这门课程涵盖使用 Python 进行实用线性代数计算、在自由落体和弹簧-质量系统等简单情境下数值研究微分方程，以及简要介绍声波的傅里叶分析。这是两门系列课程中的第二门，第一门是 Python 编程、探索性数据分析、可视化和线性回归入门。为这门课程编写的原创课程内容以开源形式发布在 GitHub 上。[1], 课堂采用的教学方法是 Software Carpentries 工作坊推广的实时编码风格。 [2], [3], 这种风格也受到科学计算 Python 社区许多教育工作者的青睐。多年来开发的作业使用了 `nbgrader` Jupyter 扩展 [4] 以及一个基于网络的自动评分系统，让学生可以多次尝试并立即获得反馈。这种组合不仅效果显著，也颇受学生欢迎；即使在疫情期间远程教学最艰难的时期，学生也给予了许多积极评价。随后，

生成式人工智能出现了，教师满怀善意地拥抱它，却导致了灾难性的结果。学生开始直接把作业题目复制粘贴给人工智能，再将人工智能生成的代码解答粘贴回 Jupyter 笔记本，并通过反复试错使用自动评分器，不求理解，只为拿高分。出勤率、满意度和学生的学习效果都一落千丈。

在本杂志《教育部》对那段经历的叙述中 [5], ，作者强调了学生使用人工智能所带来的一种有害影响，即能力错觉 [6]。这是一种元认知失准：面对现成答案时，学生会误判自己未来的表现；他们有一种感觉，以为自己正在学习，但实际上并没有。同样的情况也可能发生在其他形式的被动参与中，例如坐在讲座现场，或阅读并标记笔记或教材。然而，使用AI工具时，这种影响更为严重，因为AI工具极大地提升了人们在极少参与的情况下完成某些任务的能力。

遗憾的是，仅仅告诉学生某项任务不应使用人工智能，并不起作用。在无人监督时，他们会使用手边的任何辅助工具，无论我们如何要求、课程大纲政策如何规定，或我们在作业中加入何种“红绿灯”信号 [7]。这些对评估的论述性改变造成了一种执行错觉，因为我们不可能可靠地检测或核实学生是否使用了人工智能。仅靠规则无法保障评估效度；相反，我们应当着眼于对评估进行结构性改变：重新设计任务、加入互动环节、嵌入过程证据，等等。

本文介绍了一种我们为应对人工智能带来的效度挑战而设计的评估方式：对话式考试。这一构想源于我们的思考：在口试环境中，由教师逐一考查每位学生，是了解他们掌握情况的万无一失的方法。然而，在大班课堂中，口试并不实际（我们的班级有近60名学生）。一次灵感来自于我们看到一些媒体报道，提到科技公司采用群体面试，让多位候选人同时加入视频通话。我们虽然有些忐忑，但在教学大纲中写下一句话，便正式承诺采用这种新考试形式，并着手设计。我们希望本文的详细介绍能帮助其他希望根据自身情境调整这一理念的教师。

设计原则与依据

核心设计原则

对话式考试以三个理念为基础：(1) 评估理想情况下应反映真实实践；(2) 效度应源自考试形式本身，而非通过监督学生行为来保障；(3) 通过周密安排后勤事项并关注认知负荷，口头考试可以变得切实可行。让我们来探讨这些理念。

真实情境胜过人为限制。 在专业技术工作中，人们会查阅文档和在线资源，也越来越多地使用AI工具。他们与同事协作，搜索错误消息的解决方案，并参考以往项目中的示例。设置考试条件，剥夺所有这些支持——使用受限浏览器、禁止使用工具、置身隔离工作空间——或许能有效衡量学生在这些受限情境下的能力，却无法充分反映他们是否已准备好从事实际的计算工作。对话式考试采用了不同的方法。学生在熟悉的环境中完成考试：使用基于云的 Jupyter 笔记本，随手查阅文档和课程材料，并连接开放互联网。他们甚至可以使用人工智能，但须接受监督

并遵守关于何为适当使用的明确界限。评估重点在于他们能实时完成什么，以及能解释什么；其中，时间限制或许是唯一偏离实际工作情境的“考试条件”。

以表现而非禁止来保障效度。 近期关于人工智能时代评估的学术研究呼吁转变视角：与其纠结于学生行为（他们是否作弊？），不如关注评估本身的属性（它是否能有效衡量学习？）[8]。依赖学生不使用人工智能的评估很脆弱；一旦学生找到绕过限制的方法，其效度便会崩塌。对话式考试将效度融入了设计之中。现场编程并同时解释思路，无法通过死记硬背伪装，也无法委托人工智能完成。学生一边输入代码、一边阐述推理，停下来调试错误，并回答关于其方法为何有效的后续问题时，便展现了他们真正的理解。依赖人工智能完成课程作业的学生，在被要求解释自己的过程并现场编写代码时会露出破绽。表现本身便成为证据；在事先给予适当指导后，学生会意识到需要做好准备，并积极参与我们提供的练习机会。

通过周密设计实现可扩展性。 传统观点认为，口试无法大规模开展，这并非没有道理：逐一考查 58 名学生可能要耗费教师数天时间。对话式考试则以五六人为一组进行评估，学生轮流回答问题，其他人旁听。这种轮换让每个人都能投入其中，学生可以观看同伴编程，并从提出的问题中学习。它也缓解了一对一考试带来的部分焦虑，营造出更具对话氛围、较少审讯感的环境。如下文所述，小组形式也有助于提高操作效率。要让考试顺利进行，需要做好任务分配：主讲教师负责提问，协同教师在学生编程和解释时，根据预先准备的评分表记录他们的表现，助教则负责处理技术问题并协助掌控时间。通过口头考试有效评估整门课程的学生，这件看似不可能的事，如今完全可行。我们在 10 场各半小时的考试中评估了 58 名学生。

以人机交互为依据的流程设计

该设计还借鉴了人机交互（HCI）研究中的原则，尤其是认知负荷管理和评估一致性方面的原则。我们其中一人拥有丰富的用户研究经验，深知充分的准备工作对于应对现场观察和记笔记所带来的繁重多任务处理要求至关重要，因为这些工作本身就十分耗费心力。以往研究发现，如果访谈者的工作记忆负荷过重，访谈质量就会下降 [9]。我们在此采用了同样的原则。

每道题都配有一份详细的评分表：其中包含用于勾选评分标准条目的表格、预期解答、学生遇到困难时可提供的提示、学生快速答对时可进行的进阶概念检查，以及一道简化的备用题。这种做法遵循结构化观察领域成熟的最佳实践，强调必须系统地收集和记录证据，而非依赖偶然观察，以此作为表现的可靠指标 [10]。在人机交互方法论中，这种有记录的流程可作为一种外化产物，在团队中建立对有效性的共同理解；这套标准能统一不同观察者评估行为的方式，从而提高评估的一致性 [11]。此外，明确且预先设定的问答流程，体现了绿野仙踪原型法所要求的严格“行为指令” [12]，在这种方法中

人工操作员会谨慎地控制计算机或机器人系统的行为，以模拟自主系统。在绿野仙踪方法中，通过明确无歧义的操作员指令维持确定性系统逻辑，对于保证数据完整性至关重要；缺少这种结构，研究就可能受到随机效应或非预期即兴发挥的影响，从而损害评估。最后，结合人机交互实地工作的经验，我们努力将所有内容都安排在一页上，以免观察过程中翻动纸张。这种准备和组织工作将认知负担从考试当下转移到事前准备上。考试过程中，教师可以全神贯注于学生的表现，而不必手忙脚乱地临时编出后续问题或回忆评分标准。这种结构既提供了灵活性——我们可以根据每位学生的表现进行调整——又不需要在压力下临场发挥。

问题设计与实施

我们为考试建立了一个包含两个难度层级的题库，每个层级有 30 道题。第一个层级考查基础技能：计算工作的基本要素、语法和词汇。第二个层级则深入考查概念理解，或要求学生在具体情境中运用问题解决技能。每道题都针对一个具体的学习目标——例如计算特征值或归一化一个向量——准备充分的学生可以在一两分钟内完成。

每一轮中，学生从纸袋里抽取一道题（每个层级各有一个纸袋），这既带来了实体抽签的随机性，也增添了几分戏剧性的悬念。学生需要报出题号（这样我们就能从预先整理好的一摞评分表中找到对应的那张），然后大声读出题目；此时，我们会翻转玩具沙漏开始计时。

三级问题支架

- **第 1 级（2 分钟）：** 基础知识回忆与基本语法——最低限度的能力
- **第 2 级（3 分钟）：** 概念应用——体现更深入的理解
- **可选第 3 级：** 面向高水平学生的问题解决整合

真正的创新在于我们如何为每道题做好准备。我们没有只写题目和预期答案，而是设计了一棵决策树。如果学生很快答对，我们就准备了更深入的概念检查，例如：“如果改用 axis=1 会怎样？”或者“如果矩阵不是对称矩阵，情况会有什么变化？”这些探究性问题能揭示学生是否真正理解了概念，还是只是记住了某个程序（并让我们将其概念理解的评分从“熟练”提高到“优秀”）。如果学生遇到困难，我们可以提供提示，而不必临场发挥，例如：“记住，numpy.linalg.norm() 接受一个轴参数。”如果学生仍然一筹莫展，我们也准备了简化版本，例如：“不要归一化多个向量，只归一化一个”（并将评分下调至“发展中”）。这种支架式设计确保每位学生无论抽到哪道题，都有多次机会展示自己的能力。没有人会陷入无路可走、尴尬沉默的境地。

创建题库需要投入大量前期精力，而在这一阶段，生成式AI工具确实发挥了重要作用。我们将课程学习目标和主题对应到

具体且可评估的任务上，并着重考虑哪些内容能在几分钟内合理地展示出来。有些主题天然适合这种形式——向量运算、矩阵操作、基本的函数调用——而另一些则需要更仔细地斟酌。例如，我们无法要求学生在三分钟内实现完整的奇异值分解算法，但可以让他们使用 numpy.linalg.svd()，并解释奇异值代表什么。时间限制非但不是障碍，反而促使我们聚焦于真正重要的事：学生能否应用这一概念，并阐明其原理？我们在准备阶段对细节的关注，大大减轻了教学团队在实际考试中的认知负担，即使我们连续进行了十场考试。

实施：让它发挥作用

后勤安排

对话式考试需要协调人员、技术和实物材料。我们的团队由三人组成：负责提问的主讲教师、负责评分表并监看屏幕的协同教师，以及负责计时和处理技术问题的助教，例如选择要在投影仪上显示的学生屏幕。这种分工至关重要。任何一个人若要同时向学生提问、评估其表现、留意人工智能滥用情况并记录时间，都会不堪重负。

在安排时间方面，我们使用了 `zcal`，[1] 这是一款基于网页的预约工具，可与日历系统集成。学生从可用时间中选择自己偏好的 30 分钟时段，系统会自动向学生和考试团队发送日历邀请。这种自助式安排方式避免了与 58 人来回协调日程的麻烦。我们在两天内提供了多个时段，并安排了午餐休息时间。

教室还配备了 Zoom，而我们后来发现它有一项此前并不知道的功能：多个参与者可以同时共享屏幕 [13]。这项功能非常实用。每组的五六名学生都加入同一个 Zoom 会议，同时围坐在一张大桌旁，打开各自的笔记本电脑。所有人加入时默认将麦克风静音，并共享完整屏幕。轮到某位学生回答时，助教便切换为在其连接投影仪的笔记本电脑上显示该学生的屏幕，这样我们就能共同实时观看其编写代码。协同教师监看一台大型外接显示器，其中几个共享屏幕以弹出窗口显示，其他屏幕则显示在标签页中；这样，如果学生打开人工智能聊天窗口，或离开 Jupyter 笔记本，就能迅速发现。学生在 JupyterHub 中完成任务，这是我们在整个课程中使用的云端笔记本环境；他们可以随时访问文档和课程材料。

考试流程

每场 30 分钟的考试都遵循一种可预测的节奏，这有助于管理我们的认知负荷，也能缓解学生的焦虑。开场几分钟是热身环节：所有学生同时打开 Jupyter 笔记本，导入 NumPy 和 Matplotlib，并运行一条简单的命令。这让大家放松下来，也确认了每个人的技术准备都已就绪。热身之后，我们开始第一轮提问。学生依次从一级题目袋中抽取一张卡片，报出题号，并大声朗读题目。助教翻转沙漏，随后学生便开始

[1] https://zcal.co

在 Jupyter 笔记本中输入代码，同时大声讲述自己的思路。对于准备充分的学生来说，像“创建一个包含十个从零到圆周率等间距数值的 NumPy 数组，并计算其模长”这样的问题，不到一分钟就能完成。主讲教师会观察并倾听；如果学生遇到困难，就提供提示；如果时间允许，则转入难度更高的概念检查。协同教师会在评分表上勾选，并记录学生是否解释清楚、是否出错、是否需要提示，或是否自信地顺利完成。助教负责计时，并管理事先整理好的一叠答案和评分表。

三类别评分量规（1–4 分制，不设零分）

- **技术技能（10 分）：** Python 语法、NumPy 函数、代码执行、错误处理
- **概念理解（10 分）：** 数学解释、几何诠释、概念联系、结果验证
- **问题解决与沟通（5 分）：** 逻辑方法、调试能力、对提示的回应、解释清晰度

评分表是整个流程的关键。每张评分表顶部列出题目文本，接着是包含代码的预期解答、批注空间、提示版/提高版/简化版题目变体，底部则是带复选框的评分量规网格（技术技能 1–4、概念理解 1–4、问题解决 1–4）。考试期间，协同教师会实时填写网格，并快速记录学生的表现或回答。为了计算满分 100 分的最终成绩，我们将强制选择量表（1–4 分制）的原始评分乘以特定标准乘数进行换算。技术能力和概念理解的乘数均为 10（每项最高 40 分），问题解决的乘数为 5（最高 20 分）。这种加权策略确保评估优先考查核心技术能力和数学掌握程度。随后，将两道题各自的总分取平均值，得出最终数值成绩。

第二轮使用了二级题目袋，题目稍难，需要更深入的概念理解。此时，课程已经进行了约 15 分钟，学生们看过几位同伴回答问题，也见识了这种形式的实际运作。焦虑感想必已经有所缓解。起初有些犹豫的学生可能会开始专注起来，认真观察，也许还会把看到的内容与自己可能遇到的问题联系起来。这种形式有一种奇妙的魔力：它是一场考试， stakes 很高，也要评分，但感觉更像一堂研讨课，学生们在同伴面前展示自己的成果。

整个过程中，我们努力保持鼓励的语气。学生输入正确内容时，我们会点点头或说“很好，继续。”他们犯了个小错误时，我们可能会说“检查一下这里的索引”，而不是任由他们陷入困惑。我们的目标是了解他们在尽可能少的帮助下能做到什么，但我们并不想眼看着任何人惨败。这种对话式的做法——真正把交流当作对话，而非盘问——影响了学生的表现。他们更愿意边想边说，会说“这部分我不太确定”，也更能从小错误中调整过来。

到了第三或第四场考试时，这套流程几乎已成了自动化操作，而这恰恰就是我们想要的效果。繁重的认知工作已在准备工作中完成，因此考试过程中我们可以全神贯注于学生表现。精心准备的考试试卷让我们能够在一整天里始终保持自信、快速的节奏，从而抵消观察者疲劳的影响，避免……

……导致各场考试和答案渐渐混淆。每天的考试场次结束后，我们三人会趁记忆犹新，迅速复盘，标记出表现特别出色或较弱的学生。随后，我们会把这些勾选标记换算成分数，并在需要时写下简短评语。两天的考试结束后，仅仅几小时内，我们就能公布成绩。

结果与反思

每当一组学生涌进教室，我们都会看到他们把手写笔记摊在桌上，或者在笔记本电脑的另一个窗口中打开电子笔记。他们做好了准备。有几次，他们问我们能否把错误消息输入 ChatGPT，或者用 Google 搜索（如今这 inevitably 会显示 AI 模式下的答案）。考虑到我们事先向他们说明过考试的人工智能使用政策，我们允许了这些操作。随着考试场次一场接一场地进行，谁真正投入了课程学习，谁没有，显而易见。有些学生当然会紧张，但只要做好了准备，结构化的考试形式就能让他们发挥出色。最后，这门课两次考试的平均分约为 80%。完整的体验让我们确信：这种方式行得通！

对话式考试的人工智能使用政策

允许的人工智能使用方式：

- ✓ 代码运行失败后解释错误消息
- ✓ 澄清 Python 文档内容
- ✓ 理解自己的代码为何无法运行（尝试之后）
- ✓ 讲解自己笔记中的概念

禁止使用人工智能：

- × 为他们生成代码
- × 复制粘贴人工智能生成的代码
- × 在尝试之前询问“这个问题该怎么解”
- × 使用人工智能编写函数、循环或逻辑

不过，情况也并非一片大好。对许多学生来说，这是一种陌生的体验，尽管他们事先收到了关于考试形式的书面说明，了解考试会是什么样，也参加过一项课堂活动：我们让他们两人一组，在 Google Gemini 的引导式学习模式下练习边编写代码边回答问题。少数学生答题不佳后感到非常沮丧，并认为随机抽取问题“有失公平”——考试中或许总会发生这种情况，但他们将问题归咎于这种新颖的形式。我们认为，随着口试作为应对人工智能滥用的措施日益普及，久而久之，文化也会随之改变。

总结：对话式考试

首先决定 实施真实性评估，让学生与教师面对面互动；接着设计 一种能够适用于（规模中等的）大型群体的形式；最后再统筹安排整个过程——这一经历令人倍感紧张。做出决定时心怀忐忑，设计过程十分费力，而对所有参与者来说，统筹安排也颇具挑战。我们遇到了一些意料之外的难题，例如该如何处理享有延长考试时间便利措施的学生？考虑到我们需要谨慎处理，不能向小组中的其他人透露某位学生享有便利措施，我们决定让所有人使用相同的计时器，但负责评分的教师会不动声色地观察享有便利措施的学生，并额外给他们一两分钟。（几乎所有学生都会在时间结束后继续作答，而此时投影仪已经切换到分享另一名学生的屏幕。）这一做法在第一次考试中似乎奏效了，但有一名学生提出了异议。与其讨论后，我们发现这名学生更希望能清楚地看到自己获得了额外时间。因此，第二次考试时，我们不得不逐一征得每位享有延时便利措施的学生同意，才能为他们使用明显不同的计时器。我们还多次与残障服务办公室的工作人员通话，耗费了不少时间。今后，我们将探讨正式申请“根本性变更”，以便在所申请的便利措施从根本上改变课程基本要求时获得例外处理。目前此事尚待进一步处理。

然而，尽管存在种种障碍，我们仍对结果感到满意。即使在生成式人工智能存在的情况下，对话式考试依然提供了一种有效的评估形式；它让学生对自己的学习感到更加负责，也让人感到这种考试贴近学生未来可能应用其知识的工作场景。下面我们总结对话式考试的优点。

通过自适应探查促进深度掌握： 通过探查学生心智模型的“边缘”，这种形式促使学生将学习策略从程序性记忆转向概念掌握。这种环境支持“自适应探查”，结构化的后续问题帮助教师区分浅层回忆与学生将知识应用于现实场景的真正能力。

在人工智能时代确保稳健的诚信保障： 随着生成式人工智能使传统的书面作业和程序性作业越来越难以核验，对话式形式为学术诚信提供了首要防线。通过要求学生进行实时、未经预先准备的推理，并自发地为自己的逻辑辩护，这种形式使将思考外包给人工智能变得本质上困难，并确保学生展现真正的智力成果所有权。

通过组织创新与人机交互严谨性实现可扩展性： 教师采用人机交互指导的组织设计——例如“绿野仙踪”确定性逻辑和详尽的评分表——便能克服口试历来面临的后勤障碍和高昂人力成本。这种“通过结构实现可扩展性”的做法，使高保真评估能够高效地应用于更多学生群体，将繁重的认知工作转移到准备阶段，并确保人工考官保持一致性，避免无意间即兴发挥或产生观察者疲劳。

结语

对话式考试应势而生——面对学生不受限制地使用人工智能，眼看传统评估逐渐崩塌，这一考试形式应运而生。起初，它是一次绝望之举，后来却变成了

更有价值的东西：它证明，我们能够根据世界的真实面貌来设计评估，而不是依照我们希望它仍然保持的样子。该形式承认，AI工具无处不在，而且日益强大，却拒绝屈从于这样一种看法：有了这些工具，真实评价就不可能实现。

我们犯过错误，遇到过意料之外的问题，也花了无数时间反复打磨那些起初看似微不足道、后来却至关重要的细节。但这种形式行之有效，更重要的是，它能够扩展。两天内考查了58名学生，所得结果有效，能够区分真实能力与表面表现。这是一项概念验证，表明只要我们认真关注结构和流程，有效评估依然可以实现。

更广泛的挑战依然艰巨。每个学期，都有学生在依赖人工智能完成课程作业、却未能培养真实能力的情况下毕业。每个学期，各机构都推迟调整评估实践，或许还寄希望于人工智能颠覆会自行化解或逐渐消退。事实并非如此。这项技术发展过于迅速，正日益融入知识工作的方方面面。问题在于，教育将引领这场变革，还是仅仅被动应对。

我们分享这项工作，并非将其视为万灵药，而是想证明存在其他选择。对话式考试尤其适用于计算课程，但其基本原则——真实性、内在效度和可扩展结构——适用范围更广。其他学科也需要根据自身情境和限制，探索各自的创新方案。重要的是，我们要开始尝试，诚实地记录成果，并共同积累有关哪些方法行之有效的知识。许多教师感到无所适从，是因为他们眼中只有两条路：要么全面禁止人工智能，要么接受有效评估已不可能。其实还有别的路，我们需要共同探索。

对话式考试的意义超越了它直接的实用价值。它提醒我们，有意义的评估离不开人类判断与亲身参与，而这些无法自动化，也无法外包。这种形式之所以成功，是因为它以对话为核心——教师熟悉相关领域，学生则是初次探索，双方之间这种不可替代的人际交流正是关键。经过恰当设计并认真开展的交流，能够创造一个让真正理解显现出来的空间。或许，评估本该一直如此：不是监控和排名，而是让学生展示自身能力、让教师见证这一过程的机会。如果我们有勇气尝试，并以谦逊的态度从失败中汲取经验，或许就能在这场变革之后，形成比以往更好的评估实践。

作者简介

Lorena A. Barba 是乔治·华盛顿大学机械与航空航天工程学教授。她是IEEE计算机学会会员，曾任*Computingin Scienceand Engineering*主编。联系邮箱：labarba@gwu.edu

Laura Stegner 是乔治·华盛顿大学机械与航空航天工程系助理教授，研究领域为机器人学。可通过以下地址联系她：laura.stegner@gwu.edu

采用者实施清单

考前准备（提前 2-3 周）：

- 设计题库（根据轮数设置30至60道题）
- 为每道题制作评分表
- 制定评分细则，并与共同授课教师校准
- 设置日程安排系统（如 zcal）
- 向学生说明考试形式和预期要求
- 与学生进行练习／样题环节

提前一周：

- 敲定题目选择并打印题卡
- 测试技术准备（Zoom 屏幕共享、监考画面可见性）
- 准备实体材料（纸袋、评分表资料包）
- 向学生发送提醒，并附上技术设置说明

考试当天：

- 在第一场考试前 30 分钟测试技术准备
- 向考试团队简要说明职责和规程
- 保持稳定的精神状态和鼓励性的语气
- 在各场考试之间进行复盘，以便校准

考试后：

- 在 24 小时内完成评分（趁记忆犹新）
- 与团队校准任何临界案例
- 记录经验教训，以便未来迭代

B 考试　考试期间使用的完整支架式试题卷示例　m.

Student Name:

Question 1

Create a 2x2 NumPy matrix that scales the x-axis by a factor of 4 and the y-axis by 0.5. Apply this transformation to the vector $v = \begin{bmatrix} 5 \\ 20 \end{bmatrix}$ and print the resulting vector.

HINT	ELEVATED	SIMPLIFIED
A scaling matrix only has non-zero values on its main diagonal. Think about which diagonal element corresponds to the x-axis and which to the y-axis.	**Concept Check:** What matrix would reverse this transformation? Explain what would happen if you applied this new matrix to your resulting vector. **Answer:** The inverse matrix, which has the reciprocals on the diagonal: $\begin{bmatrix} 1/4 & 0 \\ 0 & 2 \end{bmatrix}$. Applying it to the result would transform the vector back to its original state.	**Code Adjustment:** Can you first create just the vector $v = \begin{bmatrix} 5 \\ 20 \end{bmatrix}$ as a NumPy array?

```
import numpy as np

# Define the scaling matrix
S = np.array([
    [4, 0],
    [0, 0.5]
])

# Define the vector
v = np.array([5, 20])

# Apply the transformation
v_transformed = S @ v
print(v_transformed)
```

Category	Excellent (4)	Proficient (3)	Developing (2)	Needs Improvement (1)
Technical Skills				
Conceptual Understanding				
Problem-Solving & Communication				

参考文献

[1] L. A. Barba，《工程计算》（EngComp）课程材料，https://github.com/engineersCode/EngComp，工程计算开源课程，2017年。[2] G. Wilson 等人，“Software Carpentry：经验教训”，F1000Research，第3卷，2014年。DOI：10.12688/f1000research.3-62.v2[3] A. Nederbragt、R. M. Harris、A. P. Hill 和 G. Wilson，“参与式实时编程教学的十条实用建议”，PLOS计算生物学，第16卷，第9期，e1008090，2020年。DOI：10.1371/journal.pcbi.1008090[4] D. S. Blank 等人，“Nbgrader：用于在 Jupyter 笔记本中创建和评分作业的工具”，《开放源代码教育期刊》，第2卷，第11期，2019年。DOI：10.21105 / jose.00032[5] L. A. Barba，“在工程计算课程中拥抱生成式人工智能的经验：哪里出了问题，接下来怎么办”，《Computer》，第58卷，第8期，第136–141页，2025年。DOI：10.1109/ MC.2025.3572654[6] A. Koriat 和 R. A. Bjork，“学习过程中监控自身知识时的能力错觉”，《实验心理学杂志：学习、记忆与认知》，第31卷，第2期，第187–194页，2005年。DOI：10.1037/0278-7393.31.2.187[7] T. Corbin、P. Dawson 和 D. Liu，“空谈无济于事：为什么生成式人工智能时代需要评估结构变革”，《高等教育评估与评价》，第50卷，第7期，第1087–1097页，2025年。DOI：10.1080/02602938.2025.2503964[8] P. Dawson、M. Bearman、M. Dollinger 和 D. Boud，“效度比作弊更重要”，《高等教育评估与评价》，第49卷，第7期，第1005–1016页，2024年。DOI：10.1080/02602938.2024.2386662[9] D. Giorgianni、A. Vrij、S. Leal 和 H. Deeb，“访谈者的认知负荷对司法访谈质量的影响”，《欧洲法律情境应用心理学期刊》，第17卷，第2期，第101–110页，2025年。DOI：10.5093/ejpalc2025a9[10] P. Croll，《系统课堂观察》（社会研究与教育研究）。纽约：Taylor & Francis，1986年，第3卷，第202页。[11] R. Hartson 和 P. S. Pyla，《UX Book：打造优质用户体验的敏捷用户体验设计》，第2版。Morgan Kaufmann，2018年。[12] J. F. Kelley，“面向用户友好的自然语言办公信息应用的迭代设计方法”，《ACM信息系统汇刊》，第2卷，第1期，第26–41页，1984年。[13] Zoom Video Communications, Inc.，允许多位参与者同时共享屏幕，https : / / support . zoom . com / hc / en / article ? id = zm_kb & sysparm_article = KB0064237，访问日期：2026-01-15，2024年。