Advances in Education 教育进展, 2025, 15(10), 848-861
Published Online October 2025 in Hans. https://www.hanspub.org/journal/ae
https://doi.org/10.12677/ae.2025.15101910

Hans汉斯

人工智能赋能的《信息隐藏》课程教学创新探讨

冯丙文，吴小天，宋婷婷，李佩雅，耿光刚

暨南大学网络空间安全学院，广东 广州

收稿日期：2025年9月9日；录用日期：2025年10月10日；发布日期：2025年10月17日

摘　要

随着人工智能技术的快速发展，大语言模型在教育教学中的应用为课程改革提供了新的思路与工具。《信息隐藏》作为网络空间安全学科的重要课程，长期以来在教学中存在理论抽象、实践资源有限、个性化不足等痛点。本文基于人工智能赋能的视角，系统探讨了《信息隐藏》课程的教学创新路径。具体包括：通过智能化教学辅助工具实现实验错误定位与反馈优化；基于学习数据的个性化学习支持与资源推荐；依托跨学科案例推动实验与实践教学创新；构建多维度的课程评价体系以实现持续改进。研究结果表明，人工智能的引入不仅能够提升学生对复杂概念的理解与掌握，还能有效缓解教师的教学压力，促进科研与教学的深度融合。最后，本文提出了课程未来的优化方向，包括技术局限突破、数据隐私保护、师资培训与跨学科融合等，为网络空间安全学科人才培养提供参考。

关键词

人工智能，信息隐藏，课程教学改革，个性化学习，网络空间安全

Discussion on the Teaching Innovation of the “Information Hiding” Course Empowered by Artificial Intelligence

Bingwen Feng, Xiaotian Wu, Tingting Song, Peiya Li, Guanggang Geng

College of Cyberspace Security, Jinan University, Guangzhou Guangdong

Received: September 9, 2025; accepted: October 10, 2025; published: October 17, 2025

文章引用: 冯丙文, 吴小天, 宋婷婷, 李佩雅, 耿光刚. 人工智能赋能的《信息隐藏》课程教学创新探讨[J]. 教育进展, 2025, 15(10): 848-861. DOI: 10.12677/ae.2025.15101910

Abstract

With the rapid development of artificial intelligence technologies, the application of large language models in education provides new perspectives and tools for curriculum reform. *Information Hiding*, as a core course in the discipline of Cyberspace Security, has long faced challenges such as abstract theoretical concepts, limited practical resources, and insufficient personalized support. This paper explores innovative teaching approaches for the *Information Hiding* course from the perspective of AI empowerment. Specifically, it introduces intelligent teaching assistants for error diagnosis and feedback optimization, personalized learning support and resource recommendation based on learning data, cross-disciplinary case integration for experimental innovation, and a multidimensional evaluation system for continuous improvement. The study shows that the integration of AI not only enhances students' comprehension of complex concepts but also alleviates teachers' workload and fosters deeper synergy between research and teaching. Finally, the paper discusses future optimization directions, including overcoming technical limitations, ensuring data privacy, enhancing faculty training, and promoting cross-disciplinary integration, providing useful insights for cultivating high-level talents in cyberspace security.

Keywords

Artificial Intelligence, Information Hiding, Curriculum Reform, Personalized Learning, Cyberspace Security

Copyright © 2025 by author(s) and Hans Publishers Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/
Open Access

1. 引言

随着大数据、人工智能、5G 等新兴技术的快速发展与普及应用，教育正进入与人工智能协作共生的新时代。教育部 2018 年工作要点提出“人工智能 + 教师队伍建设行动计划”，联合国教科文组织于 2024 年推出《面向学生的人工智能能力框架》，为全球范围内的人工智能教育提供了政策指导与实践路径。这些政策背景表明，将人工智能深度引入课堂教学已成为高等教育发展的必然趋势。

网络空间安全作为国家战略重点，要求建立完善的人才培养体系。《信息隐藏》课程作为网络空间安全学科应用安全方向的重要组成部分，涵盖隐写术、数字水印、匿名通信等核心技术，既具有深厚的学术积淀，也与现实的数字版权保护、隐私安全与人工智能生成内容(AIGC)治理紧密相关。其发展历程可追溯至古代隐蔽通信技术，20 世纪 80 年代学术界逐渐将信息隐藏与密码学区分，强调“隐藏存在性”而非“加密内容”，这一理论转型为课程体系的建立奠定了基础。

然而，在传统教学模式下，《信息隐藏》课程面临诸多挑战：学生难以理解抽象的理论知识，实验教学反馈滞后，跨学科案例匮乏，教师在课程资源准备和个性化教学方面压力较大。人工智能技术的出现，尤其是以大语言模型为代表的智能工具，为课程改革提供了新的契机。通过代码自动分析与错误定位、智能化实验任务生成、学习路径个性化定制、多模态可视化支持等手段，人工智能能够有效缓解课程教学痛点，推动信息隐藏课程实现从“知识传授”到“能力培养”的转型。

基于上述背景，本文将系统探讨人工智能赋能下《信息隐藏》课程的教学创新路径，重点关注智能化教学辅助、个性化学习支持、实验与实践创新、跨学科融合与评价体系构建，旨在为网络空间安全学

科的人才培养提供可借鉴的方案。

2. 人工智能赋能的课程发展现状

2.1. 人工智能赋能课程建设发展现状

人工智能技术，尤其是深度学习与大语言模型的发展，正在为信息隐藏(Information Hiding)领域注入新的动力。传统信息隐藏研究主要集中于隐写(Steganography)、数字水印(Digital Watermarking)以及隐写分析(Steganalysis)，重点在于提高隐藏容量、增强鲁棒性、抵御攻击。然而，近年来生成式人工智能(Generative AI)的兴起，使得信息隐藏的研究对象与应用场景呈现出显著变化。

在图像与视频领域，深度生成模型被广泛用于构造高隐蔽性载体，提升了隐写的自然性与多样性[1]-[3]。与此同时，基于深度学习的隐写分析方法不断发展，能够在复杂背景下有效检测隐藏信息[4]-[6]。在文本隐写方面，大语言模型的出现使得自然语言隐写成为新热点。研究者利用 LLM 的上下文生成能力实现语义级隐写，但同时也提出了相应的检测与防御机制[7]-[9]。

此外，AI 水印技术(AI Watermarking)已成为生成式内容治理的重要方向。谷歌的 SynthID、Meta 的 AudioSeal 等方案[10]-[12]为大规模内容溯源提供了技术手段，这不仅拓展了传统水印的应用边界，也对《信息隐藏》课程的教学提出了新的内容更新要求。国际标准组织与 NIST 等机构也开始探索合成内容标注与可信凭证框架[13]，推动信息隐藏与合规治理的结合。

总体而言，人工智能的快速发展正在推动信息隐藏技术从“算法优化”走向“智能生成–对抗检测–内容溯源”的综合生态，既为科研带来新的机遇，也为教学改革提供了丰富的案例与实验素材。

2.2. 人工智能在信息隐藏课程中的教学应用

在教育层面，人工智能赋能的信息隐藏课程正逐步形成多元化的发展趋势。AI 技术被引入课程的实验教学。传统信息隐藏课程实验往往依赖于较复杂的算法实现与数据预处理，学生在理解过程中容易陷入“公式记忆”与“代码堆砌”的困境。人工智能的引入使得这一局面得到显著改善。例如，基于深度学习的隐写与隐写分析实验可以借助开源框架(如 HiDDeN、SteganoGAN)直接调用预训练模型，让学生通过参数可视化观察信息嵌入与提取的全过程[14]-[16]。此外，AI 还可以自动生成实验样例与数据集，降低了教师准备实验的工作量，同时也为学生提供了多样化的学习素材。

在课程内容拓展方面，人工智能推动了课程内容从传统的“图像/音频/视频隐写”扩展到“文本隐写”“多模态水印”“对抗性分析”等新兴方向。近年来，大语言模型被广泛用于生成语义自然的隐写文本[7] [8]，相应的检测与防御研究也快速发展[9]。在水印技术方面，SynthID、AudioSeal 等 AI 水印方案[10]-[12]已进入实际应用阶段，课程可将其作为案例分析，引导学生思考 AI 内容溯源的技术与治理问题[13] [17]。这些前沿内容的引入不仅增强了课程的时效性，也培养了学生追踪新技术的能力[18]。

在个性化学习与智能辅助方面，人工智能为信息隐藏课程的个性化学习提供了新的可能。大模型能够根据学生的学习水平和错误模式，动态生成实验任务与提示信息[19]。例如，当学生在图像隐写实验中遇到算法收敛问题时，系统可以给出可能的原因分析与改进建议，而不是简单返回错误结果。教师也可以利用 AI 批量生成测试题与案例，构建差异化的学习路径，提升教学效率。

在跨学科与产学研结合方面，信息隐藏课程不仅涉及算法设计与实现，还与人工智能伦理、版权保护、网络取证等议题紧密相关。随着合成内容治理逐渐成为全球关注焦点，课程教学开始尝试引入 C2PA (内容凭证标准)、AIGC 内容标注、数据隐私合规等跨学科内容[20]。部分课程还与企业或研究机构合作，设计基于真实数据的项目式实验，如“AI 生成图像的隐写取证”“短视频平台的水印鲁棒性分析”，以增强学生的实践性与应用性。

综上，人工智能赋能下的信息隐藏课程不仅改善了传统教学中的抽象性与实验难度，还为跨学科人才培养提供了新思路。但同时也带来了新挑战，如学生对 AI 工具的过度依赖、模型可解释性不足以及隐私与合规风险等，这些都需要在后续教学改革中加以关注。

3. 《信息隐藏》教学的痛点与人工智能的适配性

3.1. 实验教学面临的主要挑战

作为网络空间安全学科的重要课程，《信息隐藏》强调理论与实验的紧密结合，学生不仅需要掌握隐写、隐写分析、水印与取证等基本原理，还要在实验中实现算法设计与攻防对抗。然而，当前实验教学仍面临以下主要挑战：

(1) 实验环境复杂、实现门槛高

隐写与隐写分析实验通常涉及图像、音频、视频等多模态数据处理，同时依赖深度学习框架和大量预处理库。对于多数学生而言，环境搭建和依赖配置耗时费力，导致实验学习效率不高。

(2) 算法原理抽象、过程难以直观呈现

信息隐藏的嵌入与检测过程往往隐藏在底层矩阵运算与神经网络中，学生难以通过代码直接感知容量、鲁棒性与失真之间的动态权衡。这不仅增加了理解难度，也削弱了学习的兴趣与主动性。

(3) 课程资源有限、案例更新不足

传统实验内容多停留在 LSB 隐写、DCT 域水印等经典方法，缺乏与生成式人工智能相关的新案例，如大模型文本隐写、AI 水印对抗、AIGC 内容溯源等，难以满足学生对前沿研究的探索需求。

(4) 评价方式单一、反馈滞后

目前实验评价主要依赖人工批改报告与期末考核，反馈周期长，学生在实验过程中遇到的问题得不到即时诊断，教师也难以及时掌握学习进度与存在的短板。

3.2. 大语言模型的技术特性与教学痛点的适配性分析

近年来，涌现出多种具备强大代码生成、自然语言理解与多模态能力的开源大语言模型(LLMs)。此类模型具有参数规模大、上下文理解能力强、支持多种编程语言与模态数据等特点，在赋能《信息隐藏》课程实验教学方面展现出高度适配性。其技术特性与教学痛点的对应关系主要体现在以下几个方面：

(1) 简化实验环境搭建

此类开源大语言模型通常支持多平台部署与轻量化推理，能够与 TensorFlow、PyTorch 等常见深度学习框架无缝对接。教师可基于它们构建一体化教学实验平台，提供预配置的环境与依赖，减少学生在环境配置上的负担，让其将更多精力集中于原理理解与实验设计。

(2) 提供直观解释与多模态呈现

借助大语言模型的生成与推理能力，复杂的嵌入与检测过程可以通过可视化图表、自然语言解释甚至多模态演示呈现。例如，学生可以输入代码段或算法描述，模型能够生成过程图解或解释不同隐写容量下图像的失真变化，帮助学生直观理解算法鲁棒性与安全性的权衡。

(3) 快速生成案例与动态更新资源

大语言模型能够结合课程大纲与最新研究文献，自动生成隐写、水印与取证的实验案例，并根据教师需求更新数据集与代码模板。这使得教学内容能够紧跟学科前沿(如大模型文本隐写、AI 水印对抗等)，避免实验案例陈旧化。

(4) 支持个性化学习与即时反馈

通过对学生实验日志、代码与 Prompt 的智能分析，大语言模型可以识别常见错误与知识薄弱点，提

供针对性的诊断与优化建议，实现“即时辅导”。同时，它还能根据学生的认知水平动态调整实验任务的提示信息与难度，满足不同层次学生的学习需求。

具备代码与多模态能力的开源大语言模型，其技术特性与《信息隐藏》实验教学的痛点高度契合，为课程改革提供了可行路径。通过“技术–教学”的精准对接，有望实现从传统实验教学向智能化、可视化与个性化的新模式转型，进一步提升学生的理解深度与创新能力。

4. 人工智能赋能的《信息隐藏》课程教学设计

4.1. 智能化教学辅助工具

人工智能技术在《信息隐藏》课程中的首要应用是构建智能化教学辅助工具。通过集成大语言模型、知识图谱和智能问答系统，教师可以为学生提供即时化的学习支持。例如，学生在学习“最小可感知失真隐写”或“对抗样本检测”时，AI 工具能够生成代码示例、实验流程图和可视化结果，帮助其快速理解复杂原理。与此同时，智能批改系统可自动分析学生实验代码的正确性与效率，并生成详细的改进建议，从而大幅提升作业批改与反馈的效率。借助这些工具，教师从繁琐的教学事务中解放出来，更加专注于课程思政引领与科研能力培养。

4.2. 个性化学习支持

《信息隐藏》课程的学生背景差异较大，部分学生具有较强的编程与数学基础，而另一些学生则更倾向于应用与实验。人工智能可以通过对学习数据的采集与分析，识别学生的个体差异，并推送个性化的学习资源。对于基础薄弱的学生，系统能够提供分步骤实验模板与基础知识回顾；而对于科研兴趣浓厚的学生，AI 可以推荐最新的隐写对抗研究、水印检测前沿论文，甚至生成实验性案例供其探索。此外，AI 驱动的虚拟助教能够提供 24 小时在线答疑与实验调试支持，避免了学生在学习过程中因问题得不到解决而产生的挫败感。

4.3. 实验与实践教学创新

实验教学是《信息隐藏》课程的重要组成部分。传统实验多以经典算法为主，如 LSB 隐写、DCT 域水印等，难以反映当前生成式人工智能背景下的新挑战。基于人工智能的课程改革强调“教学–科研一体化”，通过引入深度学习与大模型生成技术，将实验环节拓展至“隐写嵌入–对抗检测–鲁棒性测试”的全流程。例如，学生可以利用生成式模型实现文本隐写，再使用卷积神经网络或 Transformer 模型进行检测与对抗训练，形成完整的攻防实验闭环。同时，教师可以设计“AI 水印鲁棒性实验”，让学生测试不同水印方法在压缩、剪切、生成改写下的存活率，从而增强其实验创新能力与科研敏感度。

4.4. 教学效果评估与反馈

现有课程评价体系主要依赖期末考试与实验报告，反馈周期长，无法动态反映学生的学习状态。人工智能能够通过智能学习分析平台，对学生的实验日志、代码执行过程和答疑记录进行实时分析，绘制个性化学习曲线。系统不仅能够识别学生在某些知识点上的薄弱环节，还能生成改进建议。例如，当学生的检测算法准确率过低时，系统会提示其在特征提取、参数调整或网络结构上进行优化。教师则可以基于这些分析结果，及时调整课堂节奏或布置补充任务，从而形成“诊断–反馈–改进”的闭环评价机制，显著提升教学针对性。

4.5. 跨学科与创新思维培养

信息隐藏不仅是网络空间安全的核心议题，还与人工智能安全、数字版权保护、伦理治理、区块链

溯源等领域密切相关。在人工智能赋能的教学模式下，课程将通过跨学科案例与项目驱动方式，培养学生的综合创新能力。例如，可以结合 AIGC 内容治理标准(如 C2PA)，引导学生探讨水印与溯源在合成内容可信体系中的作用；或通过区块链存证实验，探索隐写技术与分布式账本的结合。跨学科小组项目的设置，则鼓励学生跨越计算机科学、法律、传媒与社会科学等学科边界，形成多维度的研究方案。这不仅有助于学生建立广阔的知识视野，也能够激发其科研创新思维与社会责任感。

5. 教学改革方式

5.1. 紧追人工智能前沿的课程内容更新

信息隐藏课程群需要结合新的人工智能技术补充各课程内容，紧跟技术前沿，促进产学合作协同育人。课程内容的更新主要涉及两方面，一是应用人工智能解决信息隐藏问题，二是人工智能本身的安全问题。

(1) 应用人工智能解决信息隐藏问题

Table 1. The design of information hiding course content partially integrated with artificial intelligence
表 1. 部分融合人工智能的信息隐藏课程内容设计

任务名称	子任务名称	任务说明
生成更自然的隐写载体(载体生成)	文本隐写	LLMs 可以生成语法完美、语义连贯且主题一致的文本(如文章、评论、社交媒体帖子)，在其中嵌入秘密信息(如修改特定词汇、语法结构、或利用模型本身的参数)。这种由 AI 生成的载体比人工编写的更自然，更难被传统统计分析方法检测。
	图像/音频描述隐写	利用 LLMs 生成描述图像的 Alt 文本或音频的字幕，将秘密信息嵌入到这些描述中。载体本身是正常且需要的元数据。
	代码隐写	生成功能正常但包含隐藏信息(如特定变量命名模式、注释风格、代码结构)的源代码或脚本。
开发更鲁棒和隐蔽的嵌入算法(自适应嵌入)	理解载体内容	LLMs 可以深度理解文本、图像描述、甚至音频/视频的语义内容。利用这种理解，可以设计自适应嵌入算法，将信息隐藏在语义上最不敏感或最不容易被修改的部分。
	优化修改策略	模型可以学习对载体进行最小修改以达到嵌入目的的策略，使得修改痕迹更难被察觉。
生成式隐写	--	直接利用 LLMs 的生成能力，根据秘密信息“按需”生成全新的、看起来完全正常的载体内容(文本、图像描述、甚至简单图像/音频的元数据指令)。秘密信息作为生成过程的“种子”或条件输入。这种方法隐蔽性极高，因为载体本身没有任何“修改”痕迹。
数字水印中的应用	生成带有固有水印的内容	训练 LLMs 时，将特定的、难以移除的水印模式(作为模型参数或生成过程中的约束)植入其生成的内容中(文本、图像描述、代码等)。这有助于追踪 AI 生成内容的来源和版权。
	鲁棒水印嵌入	利用 LLMs 理解内容的重要性，将水印信息嵌入到内容中语义关键的部分(这些部分在修改或传输过程中更可能被保留)，提高水印抵抗攻击(如压缩、裁剪、格式转换)的能力。
	水印检测与提取	开发基于 LLM 的检测器，即使内容经过一定程度的修改或处理，也能识别和提取出嵌入的水印信息。

将先进人工智能技术引入课堂，让学生体验这些技术强大的分析能力，并学习这些技术的基本应用技巧。例如大语言模型正在深刻变革信息隐藏领域。它们极大地提升了生成自然载体的能力，启发了更智能、自适应的嵌入策略，并推动了多模态融合的隐写方法。同时，它们也提供了更强大的检测工具，

引发了对抗性训练的新范式。在数字水印方面，LLMs 有助于源头追溯和提高鲁棒性。因此，《信息隐藏》课程的更新已成为当务之急，迫切需要引入最新的人工智能方法，以应对信息隐藏领域日益增长的需求和挑战。部分拟补充的实验内容如表 1 所示。

同时，课程需与企业合作安排实践教学，提升学生的实际操作能力和解决问题的能力。企业在应用安全领域拥有丰富的实践经验和资源，能够为学生提供真实的案例、实际的工作场景以及专业的指导，帮助学生将理论知识与实际工作相结合。

(2) 人工智能本身的安全问题

Table 2. The design of information hiding course content partially integrated with artificial intelligence
表 2. 部分融合人工智能的信息隐藏课程内容设计

涉及内容		课程板块	教学目标	教学内容
数据安全与隐私问题评价	数据泄露	信息内容安全	体会大模型训练时数据在传输和存储过程中可能面临的泄露风险。攻击者可能通过窃取数据或利用漏洞来访问敏感信息。	复现 ChatGPT 发生的用户隐私数据泄露事件，OpenAI 声明由于开源代码库中存在一个漏洞，使得部分用户能够看到另一个用户的聊天标题记录。 复现大模型提示词可能含有的隐私信息
	数据滥用	信息内容安全	体会攻击可能通过分析大模型的输出结果来推断出原始数据的信息，进而滥用这些数据。	复现基于 GSM8K 训练 loss 的泄漏检测，使用官方 GSM8K 训练集/测试集；由 GPT-4 生成的类似 GSM8K 的样本集。如果一个语言模型在预训练期间没有接触过这三个数据集中的任何一个，那么三个损失应该大致相等。但是如果模型已经在训练集上进行了预训练，或者在预训练过程中无意中暴露了测试数据，就会发现损失之间存在明显的差异。
	隐私侵犯	大数据安全	体会大模型在处理用户数据时可能侵犯了用户的隐私权。	复现使用 GPT-4、Meta 等大模型通过分析用户的文本输入，推断出用户的身份、兴趣、习惯等敏感信息。
模型流转/部署过程中的安全问题	对抗攻击	人工智能安全	体会大模型可以被精心构造的输入所欺骗，产生错误的输出。	复现对抗样本生成利用 FGSM 算法，实现 ResNet 框架的图像分类网络的攻击。使得图像分类网络产生错误的输出。
	后门攻击	人工智能安全	体会训练大模型时攻击者可嵌入特定的“后门”，使得在不破坏模型整体性能的情况下，通过特定的输入来操纵模型的输出结果。	复现后门攻击。构造带有触发器的特定训练集来训练 ResNet 框架的图像分类网络。使其在正常样本上能够给出正确结果，对带有触发器样本会给出异常结果。
	prompt 攻击	人工智能安全	体会通过构造特定的 prompt 来诱导大模型产生错误的输出或泄露敏感信息。	复现“奶奶漏洞”、套取提示词、改变系统设定等攻击。
AIGC 的内容合规问题	版权侵权	信息隐藏技术	了解 AIGC 的生成内容可能会侵犯他人的版权。了解应用数字水印可以追述版权信息。	案例讲述：未经授权使用他人的作品作为训练数据，或者生成的内容直接复制了他人的作品。 实验复现：复现 sepmark，使用数字水印保护训练数据集
	虚假信息	信息隐藏技术	了解 AIGC 伪造的虚假信息会误导用户，甚至对社会造成不良影响。了解对生成内容添加水印标记。	案例讲述：利用 AIGC 造谣的典型案例 实验复现：利用数字水印技术实现对生成图像的标定。

人工智能的发展本身带来了许多新的安全威胁。常见的安全威胁有数据安全与隐私问题、模型流转/部署过程中的安全问题、AIGC 的内容合规问题等。这些问题给信息隐藏技术的利用提供了广阔的空间。

他们相关的知识和技术将穿插到信息隐藏的各个课程板块内，表 2 列出了常见的安全威胁涉及的内容，及对应的课程板块。

(3) 以人工智能为枢纽的跨课程实验设计

人工智能的安全问题及其解决方案需要多个技术的融合，这促使《信息隐藏》课程需要结合多个学科，如计算机科学、法学、心理学、艺术等形成大培养目标，并在《信息隐藏》课程内相互交织，遵循网络空间安全知识体系的自然内在联系，开设跨课程的人工智能安全实验。跨课程实验开设思路如下：

首先，需要清晰定义跨课程实验的目标。这包括希望学生通过实验掌握哪些跨学科的知识和技能，以及实验应达到什么样的教学效果。明确目标有助于后续实验内容的选择和设计。

跨课程实验设计的核心在于整合不同课程的内容。这需要充分了解各学科之间的关系，并找到它们的交汇点。例如，可以选择与多个学科都相关的主题或问题，然后围绕这个主题或问题，整合不同学科的知识和技能，形成一个综合性的实验内容。

在设计跨课程实验时，还需要关注学生的需求和兴趣。可以通过问卷调查、小组讨论等方式了解学生对哪些跨学科主题感兴趣，然后根据这些信息来调整和完善实验设计。满足学生的兴趣和需求可以激发他们的学习热情，提高实验的教学效果。

为了培养学生的跨学科思维能力，应设计具有跨课程性质的任务和活动。这些任务和活动应能促使学生在不同课程背景下进行问题分析、整合和解决。例如，可以设计需要运用多课程知识来完成的实验项目。

最后，跨课程实验设计应包含有效的评估机制。通过实验后的测试、问卷调查或学生反馈等方式，了解学生对跨学科知识和技能的掌握情况，以及实验的教学效果。根据评估结果，可以及时调整实验设计，以更好地满足学生的需求和达到预期的教学目标。

5.2. 智慧协同的课程实践设计

(1) 实验平台设计

课程实验平台为实验教学开展提供必备的环境。在人工智能渗透到各个课程的过程中。建设人工智能+的课程实验平台可确保涉及的知识能够应用到具体的实际问题上。而平台建设所依托的底层软硬件环境不仅需要满足算力、共享能力以及成本效益，还需要契合产教融合和国产化需要。为此，项目拟基于恒电的科研大数据平台，设计人工智能安全实训平台。平台架构如图 1 所示。架构需具备国产化支持，确保实验平台的核心技术和组件均采用国产化方案，支持国产 CPU、操作系统、数据库等，以促进国内人工智能技术的自主可控发展。同时满足《信息隐藏》课程的教学实验需求，提供丰富的实验案例和教学资源，支持理论与实践相结合的教学模式。

在此平台核心模块包含以下部分：

第一部分是人工智能安全靶场。该靶场作为综合性的实训平台基石，为教师与学生打造了一个高度集成的人工智能安全虚拟仿真环境。此外，该靶场还扮演着人工智能安全实训体系的核心支撑角色，集成了人工智能模型安全评估实训系统、人工智能安全对抗实训系统以及人工智能模型安全防护系统，为深入探索人工智能安全威胁和防护技术提供资源、应用场景以及基础运行环境。

第二部分是人工智能模型安全评估实训系统。从稳定性、交互性、应用性、安全性、鲁棒性五个角度出发构建人工智能模型测评体系，通过攻击模式分析、自身风险分析、内容审核分析等，获得测评报告。通过该实训平台教师和学生可系统了解人工智能模型常见风险，掌握测评方法。

第三部分是人工智能安全对抗实训系统。复现常见的针对人工智能模型的攻击手段和经常被利用的漏洞，掌握利用不同漏洞构建测试用例的方法。涉及的攻击方式包括但不限于直接注入类攻击、语言逻

辑类攻击、FewShot 类攻击、绕过类攻击、Fuzz 类攻击、语义对抗等。

第四部分是人工智能模型安全防护系统。学生可以学习并实践多种防护机制，例如对抗训练、输入预处理、特征压缩与模型水印嵌入等。在实验过程中，系统能够对比不同防护策略的效果，帮助学生形成“攻击–评估–防御–再评估”的完整闭环认知。通过该模块，学生不仅掌握防御性技术，还能理解模型可信性建设的重要意义。

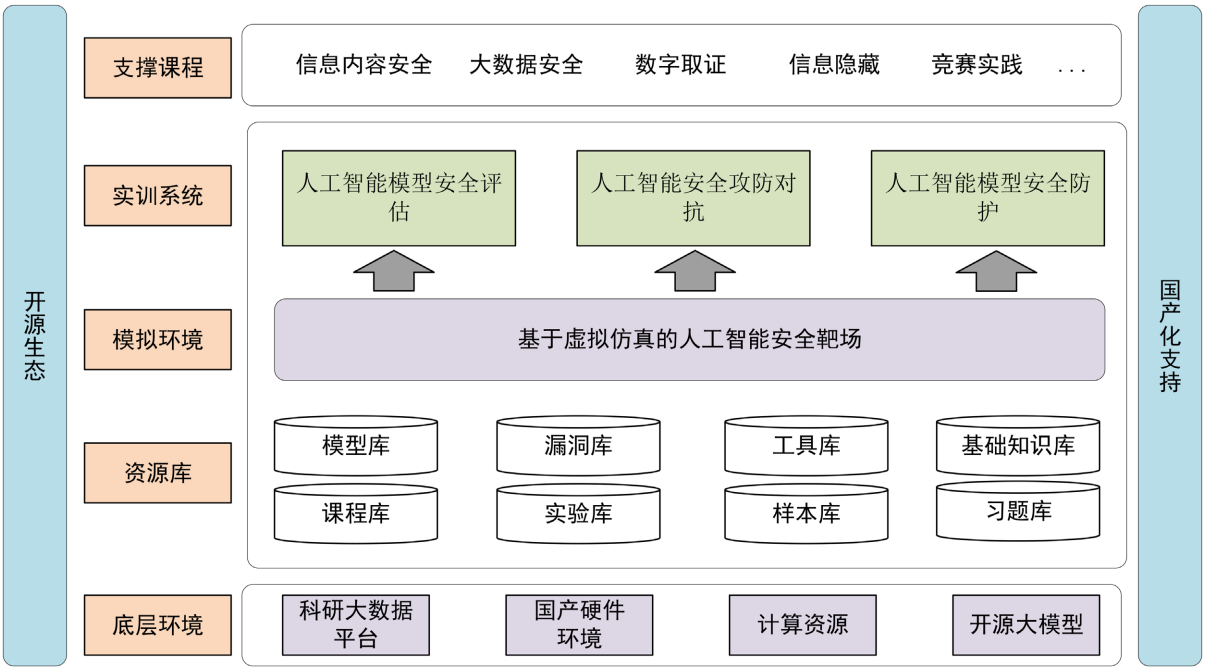

Figure 1. Course practice platform architecture design
图 1. 课程实践平台架构设计

(2) 实验流程设计

大语言模型(LLM)已全面渗透至多个行业领域。在教育领域，LLM 可以极大加速学生查阅文献、优化实验、完善报告的效率。鉴于此，本项目革新传统实验流程，通过融合 LLM 交互，增强学生的自主学习能力，培养他们的实际问题解决能力和实践创新能力。

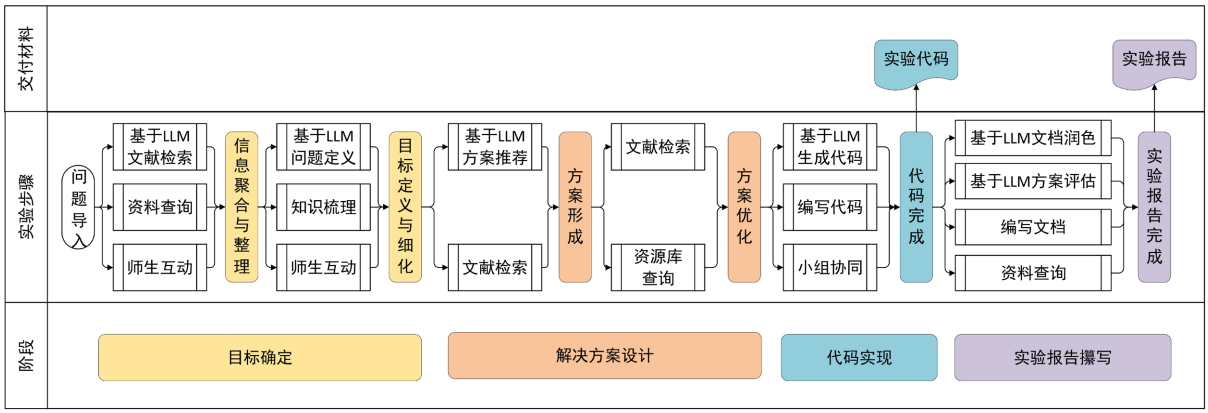

Figure 2. Integrating the experimental process design with LLM
图 2. 融入 LLM 的实验流程设计

可将各类实验的实验流程系统性地划分为成四个阶段：问题分析，解决方案设计，代码实现，实验报告攥写。LLM 可以助力各个阶段任务的完成，如图 2 所示。

在问题分析阶段，LLM 可以快速检索和整理大量相关文献、研究论文和在线资源，帮助学生全面了解问题背景、现有解决方案及其局限性。通过 LLM 交互，学生能够更精确地分析问题，寻找与理论知识的契合点，定义问题。

在解决方案设计阶段，LLM 可以生成多种可能的技术解决方案，包括算法选择、模型架构、数据处理方法等。结合 LLM 的推荐，学生查找最新文献，并利用开源或共享的工具、库、数据集等资源，可以设计出更有效、更便捷的问题解决方式。

在代码实现阶段，对于简单的或标准化的编程任务，LLM 可以直接生成代码片段，减少重复劳动。LLM 也能够帮助学生识别代码潜在的错误、性能瓶颈和可优化的地方，提供改进建议。

在实验报告攥写阶段，LLM 可以帮助评估现有实验设计和实验结果的优缺点，规划实验报告的结构，并能够提供语言的润色功能，确保实验报告逻辑清晰、语言准确、格式规范。

5.3. 纳入人工智能的课程评价体系构建

在构建《信息隐藏》课程时，将人工智能安全作为核心培养目标，并同时强调人工智能的运用能力，不仅反映了当前信息技术领域的发展趋势，也符合未来社会对安全专业人才实践能力的需求。

(1) 人工智能安全知识能力培养目标

Table 3. Design of training objectives for artificial intelligence security knowledge and ability
表 3. 人工智能安全知识能力培养目标设计

能力目标	子目标	能力要求
工程知识	知识与表达	学生应掌握人工智能安全领域的核心概念、原理、技术框架及最新发展动态，能够清晰、准确地使用专业术语和图表表达复杂工程问题。
	建模与求解	学生能够运用数学、统计学、计算机科学等工具对人工智能安全问题进行建模，设计并实施有效的算法和系统安全评估方法，分析安全漏洞并制定攻防策略。
	推演与分析	学生能够深入理解系统架构、安全协议、模型框架等，运用这些知识对人工智能安全问题进行深度推演和全面分析。
	综合与比较	培养学生跨学科知识整合能力，能够综合计算机、法学、心理学等多学科知识，对多种人工智能安全解决方案进行比较和优化。
问题分析	识别与判断	学生能够敏锐识别人工智能安全中的关键问题，准确判断其复杂性和潜在影响。
	建模与描述	学生能够利用数学建模和仿真技术，对复杂的人工智能安全问题进行精确建模和描述。
	解决方案探索	通过文献研究、案例分析等方法，学生能够探索多种解决方案，评估其可行性和有效性。
	验证与评估	学生能够运用专业知识分析各影响因素及其关联性，验证解决方案的合理性，并进行全面评估。
设计/开发解决方案	设计方法与技术	掌握全周期、全流程的设计方法和技术，确保解决方案的完整性和高效性。
	软件模块设计	设计高效、安全的软件模块，满足特定功能和安全要求。
	创新意识	在设计和实现过程中融入创新元素，提升解决方案的先进性和实用性。
	可持续性	设计时充分考虑社会、法律、环境等外部因素，确保解决方案的可行性和可持续性。

续表

研究	理论与实践结合	将理论学习与工程实践紧密结合，通过研究提升问题解决能力。
	实验数据采集与分析	构建合理实验环境，采集准确数据，进行科学分析。
	综合与比较	对实验结果进行综合分析，通过比较不同方法的效果，得出最优解决方案。
使用现代工具	工具理解与应用	熟悉并掌握常用的人工智能安全工具，理解其局限性和适用场景。
	软件开发与选择	熟悉主流开发工具，用于分析、设计和实现安全解决方案。
	建模与分析	运用专业软件进行建模、仿真、预测和分析。
工程与社会	政策法规与标准	了解并遵守相关法律法规、技术标准和产业政策。
	社会责任与风险意识	评估人工智能对社会、伦理、安全等方面的影响，树立责任意识，具备风险防控能力。

随着人工智能技术的快速发展和广泛应用，其安全性问题也日益凸显。从数据隐私保护、算法偏见到系统对抗性攻击，人工智能安全面临着诸多挑战。因此，将人工智能安全作为课程群的培养目标，在工程知识、问题分析、设计/开发解决方案、研究、使用现代工具、工程与社会等方面设计培养目标，表3列出了这些培养目标可对应的能力要求。

(2) 人工智能运用能力作为实践能力培养目标

理论与实践相结合是人才培养的关键。在强调人工智能安全的同时，提升学生的人工智能运用能力同样重要。以大语言模型为代表的人工智能使知识和浅层经验的重要性日益降低，转而强调人工智能辅助的设计创新能力，因此实践能力培养目标必然需要对应调整。结合教育心理学家本杰明·布鲁姆的六级认知层次思维模型(记忆、理解、应用、分析、评价、创新)，项目把人工智能运用能力界定为分析、评价和创新3个高阶思维能力层次，并依次设计实验评测方法，具体如表4所示。

Table 4. Design of objectives for cultivating the ability to apply artificial intelligence
表4. 人工智能运用能力培养目标设计

学习能力	对应人工智能运用能力	实验评测方法
分析	数据解读	能够准确解读人工智能处理的数据，识别数据中的模式、趋势和异常，为决策提供数据支持。
	算法理解	深入理解不同人工智能算法的原理、适用场景、优缺点及性能表现，能够选择合适的算法解决问题。
	系统诊断	对人工智能系统进行故障排查和性能分析，识别系统瓶颈和潜在问题，提出改进建议。
	场景适应	分析特定应用场景下人工智能结果的适用性、可行性及潜在风险，为技术选型和应用部署提供依据。
评价	评测	通过量化指标(如文本分类、情感分析、机器翻译等等)评估人工智能模型或系统的性能表现，判断其是否满足预期目标。
	伦理	评估人工智能应用对社会、伦理、隐私等方面的影响，甄别技术应用的合法性和道德性。
创新	集成	将大语言模型等人工智能辅助工具与其他信息检索能力与信息储备能力进行集成，提升整体任务完成能力。
	优化	能够通过与大语言模型等人工智能辅助工具的交互反馈，改造现有系统，提升效能
	主观能动性	不会过度依赖人工智能辅助工具，保持自我反思与批判性思考，始终作为具备高阶思维能力的学习主体。

6. 挑战与优化方向

尽管人工智能赋能的《信息隐藏》课程展现出显著优势，但其深度融合仍面临诸多挑战。这些挑战并非单纯的障碍，而是推动教学范式创新的契机。下文将系统分析五大核心挑战，并提出具体、可实施的优化方案。

6.1. 技术依赖与学生能力培养的平衡问题

人工智能工具能够极大提升学习效率，但也容易造成学生过度依赖 AI，出现“工具替代思考”的现象。例如，部分学生可能依赖大模型直接生成隐写或检测代码，而忽视算法原理与数学推导的学习。若长期如此，学生的自主创新能力和科研素养将受到削弱。

优化方向：为从根本上杜绝“工具替代思考”，课程需构建一个“基于 AI 协作能力的进阶式评价框架”。该框架将学习过程划分为三个阶段：在基础夯实阶段，严格限定核心算法的实现必须独立完成，以筑牢知识根基；在项目探索阶段，强制要求学生提交详实的“AI 交互日志”，其评价权重不低于最终代码，日志需完整呈现其 Prompt 设计迭代、对模型生成结果的批判性验证与筛选，以及最终决策的逻辑链条，从而将教学评价从静态的“结果评价”彻底转向动态的“元认知与创新过程评价”；在综合创新阶段，则鼓励学生将大模型视为“科研助理”，专注于探索其在新型对抗样本生成、多模态隐写系统联调等复杂任务中的赋能潜力，培养其利用尖端工具解决前沿问题的创新能力。

6.2. 教学资源动态更新与持续维护的挑战

人工智能与信息隐藏领域发展迅速，新方法、新对抗和新应用不断涌现。若课程资源不能及时更新，容易出现教材与实验案例滞后的问题，导致学生学习内容与科研前沿脱节。

优化方向：破解资源更新难题的关键在于打造一个“可持续的众智–智能双轮驱动资源生态”。一方面，建立“贡献–认证”机制的课程案例众包平台，将学生优秀的课程设计、论文复现、创新想法转化为经过助教审核的标准化教学案例，并给予贡献者学分认证，形成可持续的“众智”来源。另一方面，深度赋能教师，开发“科研论文一键转实验”智能工具，教师只需输入前沿论文链接，AI 即可自动解析其核心思想、生成算法流程图、并提供一个可运行的基础代码框架，极大降低将科研成果转化为实验素材的技术门槛与时间成本，实现“智能”化生成。两者结合，共同保障教学资源与科研前沿的同步迭代。

6.3. AI 评价体系的公平性与可解释性不足

目前，人工智能可以自动批改实验报告、评分代码和分析学习轨迹，但其算法决策过程存在一定“黑箱性”，可能引发对评分公平性和一致性的质疑。

优化方向：构建一个“透明、多维，且具教育意义的人机协同评价系统”是消除质疑的核心。系统应由 AI 首先完成初评，但其输出不能仅有一个分数，而必须是一份“可解释性诊断报告”，具体指出代码的效率瓶颈、算法逻辑的潜在缺陷，并与最佳实践进行比对。随后，系统自动依据评分置信度、分数分布异常点(如超高或超低分)以及学生申诉，智能筛选出需复核的样本，推送至教师端进行最终仲裁。此举不仅将教师从繁重的批改中解放出来，更使其能聚焦于有争议、有价值的创造性评价工作上，同时确保每一份评价都有据可查、有疑必复，从根本上提升评价的公正性与教学价值。

6.4. 教师能力提升与教学模式转型压力

人工智能赋能课程不仅要求学生转变学习方式，也对教师提出了更高要求。目前部分教师对 AI 工具的掌握程度有限，难以将其与教学目标有机融合；同时，传统的授课思路也难以完全适应智能化与个性

化教学的转型需求。

优化方向：教师的成功转型需要体系化的支持，而非零散的培训。学校应牵头成立“AI 赋能教学创新中心”，提供一个实体化运作的支持平台。该中心的核心职能包括：第一，提供“AI 教学设计师”一对一咨询，帮助教师将课程目标转化为可执行的 AI 赋能活动；第二，组织“跨学科课程设计工作坊”，打破院系壁垒，促成不同学科教师组建教学团队，共同开发融合性项目与案例库；第三，开发并共享一批“高可用性教学模版”(如标准化的 Prompt 库、AI 评价量规、跨学科项目任务书)，极大降低教师的试错成本；最后，将教学创新成果明确纳入职称晋升与绩效考核体系，从制度上激励并认可教师在教学模式转型上付出的巨大努力。

基金项目

本研究获得以下基金项目的资助：

广东省本科高校教学质量与教学改革工程项目：“思政融通、虚实相济、智慧赋能、面向产出”的高级语言程序设计课程改革与实践；

教育部产学合作协同育人项目：网络空间安全专业“人工智能+”应用安全课程群改革探索；

2025 年暨南大学人工智能赋能研究生课程项目重点项目：信息隐藏；

暨南大学实验教学改革研究专项：网络空间安全实验教学标准化建设研究。

参考文献

[1] Zhu, J., Kaplan, R., Johnson, J. and Fei-Fei, L. (2018) HiDDeN: Hiding Data with Deep Networks. In: Ferrari, V., Hebert, M., Sminchisescu, C. and Weiss, Y., Eds., *Lecture Notes in Computer Science*, Springer International Publishing, 682-697. https://doi.org/10.1007/978-3-030-01267-0_40

[2] Zhang, R., Dong, S. and Liu, J. (2019) Invisible Steganography via Generative Adversarial Networks. *Multimedia Tools and Applications*, **78**, 8559-8575. https://doi.org/10.1007/s11042-018-6951-z

[3] Tancik, M., Mildenhall, B. and Ng, R. (2020) Stegastamp: Invisible Hyperlinks in Physical Photographs. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 13-19 June 2020, 2117-2126. https://doi.org/10.1109/cvpr42600.2020.00219

[4] Qian, Y., Dong, J., Wang, W. and Tan, T. (2015) Deep Learning for Steganalysis via Convolutional Neural Networks. *Media Water-Marking, Security, and Forensics*, 1-11.

[5] Wu, H., Zhou H, Zheng W, *et al*. (2024) A Comprehensive Survey on Image Steganalysis Using Deep Learning. *Information Sciences*, **674**, Article 119574.

[6] Liu, Z., Xu, X., Qiao, P. and Li, D. (2025) Acceleration for Deep Reinforcement Learning Using Parallel and Distributed Computing: A Survey. *ACM Computing Surveys*, **57**, 1-35. https://doi.org/10.1145/3703453

[7] Yang, W., Wang, Y. and Zhang, X. (2024) LLM-Stega: Generative Text Steganography with Large Language Models. arXiv:2403.12345.

[8] Fang, Y., Wang, X. and Li, Z. (2023) Neural Linguistic Steganography and Its Detection. *IEEE Transactions on Information Forensics and Security*, **18**, 1123-1136.

[9] Yang, Y., Li, M. and Xu, J. (2024) Adversarial Attacks and Defenses in Text Steganography. *Computers & Security*, **139**, Article 103615.

[10] Kirchenbauer, J., Geiping, J., Wen, Y., *et al*. (2023) A Watermark for Large Language Models. *International Conference on Machine Learning*.

[11] Zhao, Z., Bansal, M. and Durmus, E. (2023) On the Reliability of Watermarks for Large Language Models. *The Twelfth International Conference on Learning Representations*, 123-135.

[12] Roman, R.S., Fernandez, P., Elsahar, H., Défossez, A., Furon, T. and Tran, T. (2024) AudioSeal: Proactive Detection of Voice Cloning with Localized Watermarking. In: *Forty-First International Conference on Machine Learning, ICML*.

[13] NIST (2024) Reducing Risks Posed by Synthetic Content: NIST AI 100-4 Draft Report.

[14] Lin, J., Chen, Y. and Zhang, H. (2024) Deep Learning-Based Steganography Experiments in Cybersecurity Education. *Proceedings of the* 2024 *IEEE Frontiers in Education Conference*, Urbana, IL, 23-26 October 2024, 455-462.

[15] Lee, D. and Park, J. (2023) Teaching Steganalysis with Adversarial Examples in Undergraduate Security Courses. *Proceedings of the ACM Conference on Computer Science Education*, Toronto, 15-18 March 2023, 98-105.

[16] Zhang, K., Wu, Z. and Chen, T. (2019) SteganoGAN: High Capacity Image Steganography with GANs. *ACM Multimedia Conference*, Nice, 21-25 October 2019, 75-83.

[17] Li, H., Sun, Z. and Liu, X. (2024) Teaching Digital Watermarking with AI-Generated Content Cases. *Journal of Information Security Education*, **15**, 35-47.

[18] Chen, W. and Zhao, X. (2025) Curriculum Design for AI Watermarking and Content Provenance in Cyberspace Security Programs. *Computers & Security*, **142**, Article 103811.

[19] Wang, Y. and Zhao, Q. (2025) Personalized Learning Support for Information Hiding Courses with Large Language Models. *Proceedings of the* 2025 *International Conference on Advanced Learning Technologies*, Paris, 30-31 October 2025, 250-258.

[20] Xu, J., He, Y. and Zhang, P. (2025) Integrating Ethics and Legal Aspects into AI-Powered Steganography Teaching. *ACM Transactions on Computing Education*, **25**, 1-20. https://doi.org/10.1145/3727987