

人工智能赋能三螺旋模型驱动的网络攻防课程思政改革研究

白双杰*, 胡小明, 赵丽敏, 马 闯

上海第二工业大学计算机与信息工程学院, 上海

收稿日期: 2026年9月9日; 录用日期: 2026年10月10日; 发布日期: 2026年10月16日

摘 要

网络攻防技术课程兼具技术敏感性与伦理敏感性, 其课程思政建设长期面临思政元素与专业内容融合不畅的困境。大语言模型的快速发展为破解这一难题提供了新的技术路径。本文在三螺旋理论基础上, 引入AI作为贯穿“专业能力-思政素养-创新素质”的赋能轴, 构建了“三螺旋+AI”融合模式。该模式将AI安全与大语言模型伦理纳入教学内容, 利用AI驱动实现思政案例的精准匹配与动态生成, 借助自然语言处理技术对学生反思日志进行思政素养的多模态评估。同时, 本文分析AI评估中的隐私、数据安全和算法偏见风险, 并讨论该模式在师资、技术与资源方面的推广局限。教学实践表明, 该模式有效提升了学生的专业认同感、思政获得感与创新参与度。

关键词

网络攻防技术, 课程思政, 三螺旋理论, 人工智能

Research on Ideological-Political Reform in Network Offense and Defense Course Driven by AI-Empowered Triple Helix Model

Shuangjie Bai*, Xiaoming Hu, Limin Zhao, Chuang Ma

School of Computer and Information Engineering, Shanghai Polytechnic University, Shanghai

Received: September 9, 2026; accepted: October 10, 2026; published: October 16, 2026

*通讯作者。

文章引用: 白双杰, 胡小明, 赵丽敏, 马闯. 人工智能赋能三螺旋模型驱动的网络攻防课程思政改革研究[J]. 教育进展, 2026, 16(10): 776-783. DOI: 10.12677/ae.2026.16102168

Abstract

The course on Network Attack and Defense Technologies possesses both technical sensitivity and ethical sensitivity, and its ideological-political education construction has long faced the difficulty of poor integration between ideological-political elements and professional content. The rapid development of large language models offers a new technological pathway to address this challenge. Building on the Triple Helix theory, this paper introduces AI as an empowering axis that runs through “professional competence - ideological-political literacy - innovative qualities,” and constructs a “Triple Helix + AI” integration model. This model incorporates AI security and large-model ethics into the teaching content, utilizes AI-driven approaches to achieve precise matching and dynamic generation of ideological-political cases, and employs natural language processing techniques to conduct multimodal assessment of students’ ideological-political literacy based on their reflective journals. This paper analyzes the risks of privacy, data security, and algorithmic bias in AI assessment, and discusses the limitations of scaling up this model in terms of faculty, technology, and resources. Teaching practice demonstrates that this model effectively enhances students’ professional identity, ideological-political attainment, and innovative engagement.

Keywords

Network Attack and Defense Technologies, Curriculum Ideological-Political Education, Triple Helix Theory, Artificial Intelligence

Copyright © 2026 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

网络攻防技术作为网络安全核心课程，兼具技术敏感性与伦理敏感性，其“双刃剑”特性决定了课程必须同步传授专业技能与网络安全价值观，思政融入不是附加题而是必答题。

当前思政建设主要困境在于：内容与专业脱节、实践环节价值引导缺失，且缺乏三者统筹的系统化设计[1][2]。与此同时，ChatGPT、DeepSeek 等大语言模型(LLM)的快速发展为教育教学带来了新的可能性，AI 安全、LLM 伦理等议题天然契合网络安全课程的思政需求，而其内容生成与语义理解能力也为思政案例的动态构建和个性化推送提供了技术支撑。

针对上述问题，本文借鉴三螺旋理论，引入 AI 作为核心赋能手段，提出了“专业能力 - 思政素养 - 创新素质”三螺旋融合模式，尝试以 AI 技术贯穿三条螺旋线，实现课程思政的精准化与动态化。

2. 三螺旋理论及其 AI 赋能扩展

2.1. 三螺旋理论的课程思政适用性

三螺旋理论由 Etzkowitz 和 Leydesdorff 于 1995 年提出，最初用于描述大学、产业与政府三方在创新过程中螺旋上升、相互促进的动态耦合机制。近年来，该理论被引入教育领域。李妍等[3]将其用于信息系统安全课程的“专思创”融合改革，侯严严等[4]构建了“价值引领 - 知识建构 - 能力培养”三位一体的教学框架，刘强等[5]则基于 CC2020 胜任力模型设计了全过程思政育人模型。

将三螺旋理论迁移至网络攻防技术课程思政领域，其合理性体现在三个层面。首先是目标层面的同向性，专业能力追求精技，思政素养追求明德，创新素质追求善创，三者归根结底都指向“培养什么人”这一根本问题。技术精湛但缺乏道德约束的网络安全从业者，本身就是最大的安全隐患。其次是过程层面的交织性，该课程几乎每个技术知识点都内含着思政契机，比如讲授 SQL 注入时可关联公民个人信息保护，讨论防火墙时可延伸到国产自主可控议题，进行渗透测试教学时则必然涉及白帽黑客伦理。思政教育并非外在于专业内容的附加物，而是专业教学本身的有机组成部分。第三是结果层面的增益性，三者的深度融合能够产生协同效应，思政素养为学生提供了技术学习的内生动力，创新素质为专业实践拓展了成长空间，专业能力又为思政素养的落地提供了实践载体。

2.2. AI 赋能轴的引入

现有三螺旋模型虽然阐明了专业能力、思政素养与创新素质之间的协同关系，但在实施层面仍存在瓶颈。思政案例的匹配高度依赖教师个人经验积累，教学效果因教师差异参差不齐[6]，思政素养的评价手段相对单一，过程性、动态化的评估机制尚未建立[7]。

AI 赋能轴并非分别支持三条螺旋，而是通过“任务输入、语义匹配、过程反馈、评价回流”的统一机制催化三者的动态耦合。专业任务产生学习数据，AI 据此匹配思政资源并生成具有价值冲突的教学情境；思政反思将技术行为引向责任判断，进而为创新任务提供价值约束；创新设计与攻防实践又暴露新的能力和伦理问题，回流为下一轮专业教学与思政讨论。由此形成“专业实践触发价值反思，价值反思引导创新设计，创新结果生成新的学习需求，能力提升进入下一轮螺旋”的闭环。AI 主要承担资源匹配、过程支架、冲突提示和评价回流功能，教师负责价值判断与最终审核。

3. “专思创 + AI” 三螺旋融合框架与设计

本文构建了如图 1 所示的“专思创 + AI”三螺旋融合框架。该框架以专业能力、思政素养与创新素质为三条核心螺旋，在教学目标、内容、方法、实践和评价五个维度上协同耦合，以下将围绕该框架，从这五个方面展开具体设计。课程思政教学系统首页见图 2。



Figure 1. Triple helix integration framework of specialty-ideology-innovation with AI
图 1. “专思创 + AI” 三螺旋融合框架

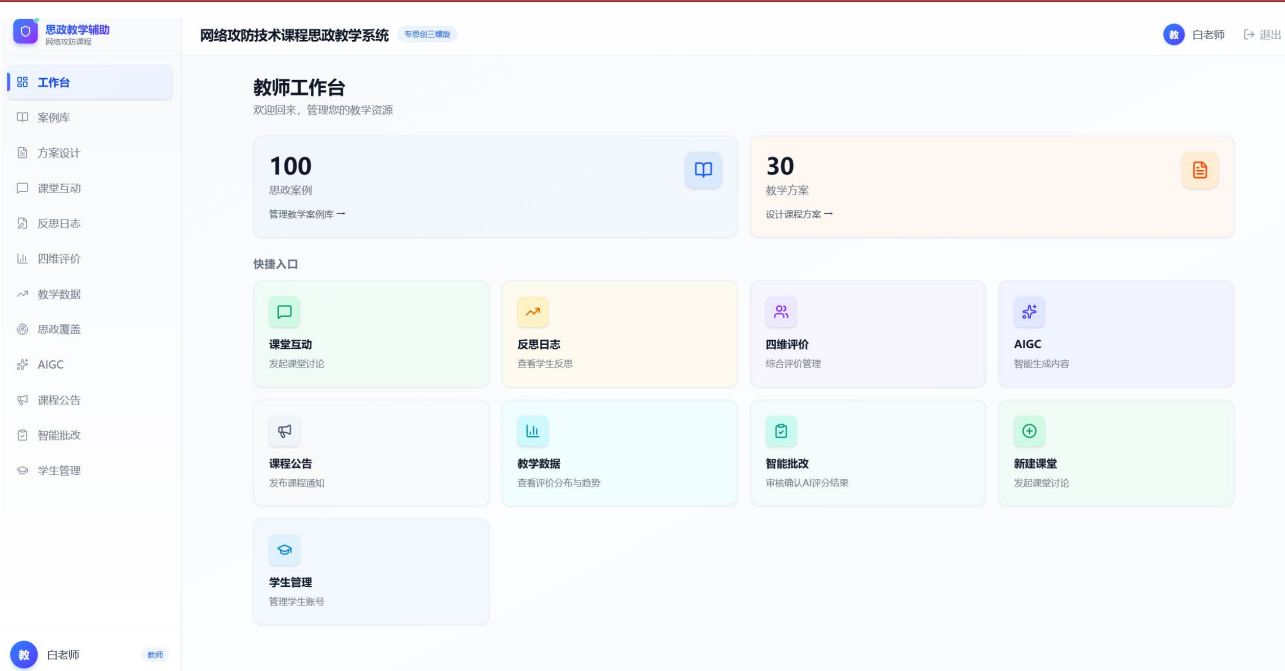


Figure 2. Curriculum ideological and political teaching system homepage
图 2. 课程思政教学系统首页

3.1. 教学目标重构

对照布鲁姆教育目标分类与教育部《高等学校课程思政建设指导纲要》¹，将教学目标重组为三个维度，并在各维度内嵌入 AI 赋能目标，如表 1 所示。

Table 1. Teaching objectives of Professional-Ideological-Innovative-AI integration
表 1. “专思创 + AI” 教学目标

维度	核心目标	AI 赋能目标
专业能力	掌握网络扫描、漏洞检测、渗透测试等核心技术	理解 LLM 安全风险
思政素养	树立网络安全法治观念、科技报国情怀、工程伦理意识	建立 AI 伦理判断力、算法偏见识别能力
创新素质	具备 CTF 赛事参与能力、攻防对抗创新思维	能利用 LLM 辅助安全工具开发与漏洞分析

3.2. 教学内容重构

结合 AI 赋能需求，将教学内容重构为四个模块。
模块一聚焦思政教育，以《中华人民共和国网络安全法》²与网络空间主权为基础，结合网络攻防博弈的典型案例展开分析；围绕《中华人民共和国数据安全法》³《中华人民共和国个人信息保护法》⁴，引入 AI 训练数据中的合规争议进行研讨；聚焦白帽黑客的伦理规范与职业道德，探讨 AI 辅助漏洞挖掘所面临的伦理边界。此外，本模块新增 AI 伦理与大模型安全法规模块，引导学生深入研读《生成式人工智能服务管理暂行办法》⁵，在夯实法律认知的同时，强化科技向善的责任意识。

¹http://www.moe.gov.cn/srcsite/A08/s7056/202006/t20200603_462437.html
²https://www.cac.gov.cn/2025-12/29/c_1768735112911946.htm
³http://www.npc.gov.cn/c2/c30834/202106/t20210610_311888.html
⁴http://www.npc.gov.cn/npc/c2/c30834/202108/t20210820_313088.html
⁵https://www.cac.gov.cn/2023-07/13/c_1690898327029107.htm

模块二为理论基础，在常规的网络攻防发展史、安全模型、风险评估方法论、国内外安全态势等内容基础上，新增了大语言模型安全基础，涵盖提示注入攻击原理与模型输出幻觉的安全风险等内容。

模块三是核心技术与实践应用，内容覆盖信息收集与网络扫描并嵌入隐私保护意识、Web 安全与漏洞挖掘并融入技术向善理念、密码学基础与国密算法并强化自主可控意识、防火墙与入侵检测并引入国产化替代讨论，同时新增 LLM 辅助安全分析环节，让学生尝试使用大模型进行日志分析与漏洞报告撰写。

模块四注重综合实战与创新拓展，包括 CTF 竞赛并设置 AI+ 思政 + 技术融合题型、虚拟仿真靶场对抗并设计 AI 安全攻防叙事场景，同时新增 AI 安全专题实训，涵盖智能体漏洞扫描和 AI 异常流量分析等。

3.3. 教学方法创新

3.3.1. LLM 驱动的智能助教

本课程引入基于大语言模型的智能助教系统，定位为贯穿教学全程的“思政副导师”。设计初衷在于解决传统课程思政中素材查找滞后、融入生硬的问题，通过人机协同和个性化推送，实现思政教育智能精准的效能提升。

该系统可以实现知识点与思政点的语义级自动匹配。比如教师输入“SQL 注入”等技术知识点后，系统不再依赖人工预设的关键词查表，而是依托 Embedding 语义检索技术，自动推送关联的思政素材，可能是《中华人民共和国个人信息保护法》相关条款或是典型数据泄露事件，并且会进行汇聚融合。

通过分析学生在实验平台的操作日志及实验反思总结的语义特征，动态绘制学生的思政画像。定向推荐与其认知水平相匹配的思政拓展阅读。

3.3.2. 案例驱动式教学

本课程在传统案例教学基础上引入 AI 增强环节，构建“技术原理 - 安全事件 - 思政反思 - AI 伦理”四环节案例链。以提示注入与 LLM 越狱攻击为例。在技术原理环节，教师讲解提示注入的基本机理，包括直接注入、间接注入以及越狱提示词等典型攻击方式，使学生理解大模型安全护栏的技术脆弱性。在安全事件环节，引入国内外公开报道的 LLM 越狱案例，如“祖母漏洞”和“DAN 越狱”等典型样本，演示攻击者如何通过精心构造的提示词诱导模型输出违法或有害内容。进入思政反思环节，教师引导学生围绕以下问题进行讨论：当 AI 可以被轻易诱导生成虚假信息、仇恨言论甚至犯罪指导时，技术开发者应承担怎样的安全责任？由此引出《生成式人工智能服务管理暂行办法》的立法逻辑与制度设计，使学生深刻理解发展与安全并重、创新与治理协同的治理思路。最后在 AI 伦理维度，将讨论 AI 对齐这一根本性问题：如何让人工智能系统的目标与人类的价值观保持内在一致？此外，本模式同样适用于 SQL 注入、社会工程学等经典安全领域的案例教学，体现了方法的可迁移性与普适性。

3.3.3. 以赛促学机制

本课程将 CTF 竞赛作为三螺旋融合框架的实践检验平台，构建以赛验学、以赛促思、以赛创融的教学闭环。

在专业能力维度，竞赛题目覆盖 Web 安全、逆向工程、密码学及 AI 安全等方向，全面检验学生对网络安全技术的掌握程度与实践操作能力。

在思政素养维度，竞赛并非简单设置法规知识问答附加题，而是将思政触发点内嵌于赛题情境之中。例如，在竞赛中设置 AI 生成内容安全审计挑战题，题目向学生提供一段由大模型生成的营销文案，文案中隐含了违法信息，比如“本保健品可治愈所有癌症”。学生的任务是扮演内容安全审核员，利用课程所学知识，找出其中违反《生成式人工智能服务管理暂行办法》的具体问题点，并提交修正后的合规版本，正确完成即可获取 Flag。学生答题完毕后，引导学生进行反思，刚刚查出的违规文案，如果未经审

核直接发布在互联网上，可能对哪些群体造成伤害？平台应承担怎样的审核责任？使学生在完成内容安全审查任务之后即刻面对社会后果的追问，将法治意识内化为职业本能。

在创新素质维度，鼓励学生利用 LLM 辅助开发自动化漏洞挖掘脚本与安全检测工具，并自主设计 AI 安全方向的 CTF 赛题，比如构造安全的 LLM 问答环境、设计对抗性提示词防御方案等，推动学生从参赛者向创造者转变。

3.4. 考核评价体系——AI 赋能的多模态评估

本课程从知识掌握、实践能力、思政素养、创新能力四个维度出发构建评价体系，对学生进行全过程、多模态的综合评估。各维度均嵌入 AI 辅助评价手段，具体构成如表 2 所示。

思政素养维度的 AI 辅助评价是本体系的核心。传统的课程思政评价往往依赖期末考试的简答题或教师的主观印象，评价标准难以统一。难以捕捉学生在整个学期中价值观认知的动态变化。本方案通过持续采集学生不同教学环节中的文本数据，自动生成思政素养的时序发展曲线，帮助教师精准识别学生的认知转折点与关键触发事件。

本课程的评价体系基于 LLM 的语义分析接口作为底层驱动引擎，对学生的反思日志、课堂讨论记录与实验报告进行文本分析。在人工预设的价值观词库基础上，通过语义相似度计算辅助识别“安全责任”、“法治意识”、“科技报国”等关键词及其上下文语境，通过大模型的推理能力，判断学生表达的情感倾向和价值立场。系统据此生成每位学生的思政素养画像。需要指出的是，AI 分析结果仅作为教师评阅的参考依据，最终评分须由教师结合学生综合表现进行人工复核与校准，确保评价的科学性与教育温度的统一。

Table 2. Course assessment dimensions and AI-assisted methods design
表 2. 课程考核评价维度与 AI 辅助手段设计

评价维度	权重	评价方式	AI 赋能评价手段
知识掌握	40%	AI 自适应测验 + 期末笔试+	LLM 自动生成个性化测试题
实践能力	40%	实验报告 + 实操考核	自动化代码质量分析 + AI 辅助分析，教师复核
思政素养	10%	案例分析 + 反思日志 + 讨论表现	文本语义特征提取，教师综合研判
创新能力	10%	CTF 成绩 + 自主命题 + 项目成果	LLM 辅助查新与证据整理，教师或同行评议

其他三个维度的 AI 赋能手段同样贯穿评价全程：在知识掌握维度，LLM 根据学生的学习行为数据自动生成个性化测试题，实现自适应考核；在实践能力维度，AI 对实验报告进行格式规范检查与代码质量分析，辅助教师批阅；在创新能力维度，LLM 对学生自主设计的 CTF 赛题与项目成果进行新颖性评估，提供创新性评分参考，从而确保四个维度均有 AI 赋能支撑，实现评价的全面性与精准性。

4. 实践案例：大模型安全提示注入攻击与防御

本章以“大模型安全提示注入攻击与防御”为例，具体说明本文提出的框架在实践教学中的完整实施过程。

4.1. 教学目标

在专业能力维度，要求学生理解提示注入的基本原理，掌握直接注入、间接注入与越狱提示词三种典型攻击方式的构造方法；在思政素养维度，要求学生能够准确阐述《生成式人工智能服务管理暂行办法》对 AI 生成内容的安全要求，并形成对技术能力与安全责任对等的价值认同；在创新素质维度，鼓励学生在掌握基础攻击手法后，尝试设计抗提示注入的防御提示词模板。

4.2. 教学设计

本课程安排为 3 学时，其中理论讲授 1.5 学时，实验操作 1.5 学时。实验采用 Gandalf AI 在线闯关平台(<https://gandalf.lakera.ai/intro>)作为核心实践环境，该平台由瑞士 AI 安全公司 Lakera 推出，是一款免费的提示注入攻防演练工具，针对多个关卡配置了不同强度的安全防御机制。此外，课程同步在本地部署开源大模型 deepseek-r1:32b 作为辅助演示环境，用于课堂拓展演示环节。

4.3. 教学实施

4.3.1. 课前预习

学生观看教师提供的提示注入原理视频，视频内容涵盖提示注入基本概念、不同攻击方式的区别以及经典案例，初步建立专业认知。学生阅读《生成式人工智能服务管理暂行办法》中关于 AI 生成内容安全相关条款，并结合教师提供的“LLM 越狱攻击导致虚假新闻传播”等案例材料，完成一份简短的阅读笔记。

4.3.2. 课中教学

课中教学分为 4 个环节，各环节内容设计及 AI 赋能方式如表 3 所示。

Table 3. In-Class teaching session design for the “LLM Security Prompt Injection Attack and Defense”
表 3. “大模型安全提示注入攻击与防御”课中教学环节设计

环节	时长	内容	对应螺旋	AI 赋能方式
课程导入	10 min	讲解 LLM 越狱攻击案例，提问：“为什么一个看起来听话的 AI 会被几句精心设计的话骗过？”，引发学生兴趣和思考。	思政 + 专业	-
理论讲授	35 min	系统讲解提示注入的基本原理，辅以实际攻击样本展示	专业	LLM 根据知识点实时生成攻击示例，辅助理解
实验操作	70 min	学生两人一组，轮流扮演攻击者与记录者。攻击者登录 Gandalf AI 平台，逐关挑战，尝试通过构造各类提示词诱导 AI 泄露密码。记录者负责记录每关的防御机制类型、攻击所用的提示词原文及 AI 的响应结果。小组讨论防御策略弱点，记录分析，并将结果提交给教师。	专业 + 创新	Gandalf AI 平台自动记录每次攻击尝试与 AI 响应，关卡通过后展示简要的防御机制说明。LLM 辅助分析攻击失败的原因，实时生成调整建议。
反思讨论	20 min	教师组织全班讨论：“刚刚成功使用的攻击技巧，如果被用于生成虚假内容，谁应该承担责任？”各组分享成功攻破的攻击思路，教师引导归纳不同防御策略的优劣。	思政 + 创新	根据各小组分析结果，LLM 现场生成多角度伦理分析框架，供讨论参考。

4.3.3. 课后训练

要求学生以小组为单位继续挑战平台上的高阶关卡，并撰写实验报告。另外要求每位学生撰写一篇关于 AI 安全治理的短文，论述技术创新与安全治理之间的平衡关系。

5. 伦理考量与局限性分析

5.1. AI 评估的伦理风险与治理

AI 评估的伦理风险主要集中在隐私、安全与偏见。操作日志、讨论记录和反思文本包含学习行为与

个人表达,若用途过宽、保存过久或缺乏权限控制,可能造成隐私泄露和数据滥用;大模型的训练语料和语言模式差异也可能产生算法偏见,将学生的表达能力误作价值认知水平。为此,本模式坚持最小必要、知情同意、去标识化、人机协同和可追溯原则:仅采集评价所必需的数据,明确用途和保存期限,优先采用本地或受控环境部署并限制访问权限;AI 分析结果只作为教师复核的证据线索,不作为思政素养的直接评分依据;学生有权了解评价依据、提出异议,并在不影响成绩的前提下选择退出 AI 分析。

5.2. 模式局限与推广挑战

该模式的实施首先受到师资与组织条件约束。教师既要具备网络攻防专业能力,又要掌握人工智能工具、数据分析和课程思政设计方法,人工复核还会增加备课与评价负担。技术层面,大语言模型存在输出幻觉、结果不稳定和误判风险,教学平台还可能面临接口变化、系统兼容及维护困难。资源层面,算力、软件采购、系统开发和持续运维均需要稳定投入,不同院校之间的技术与经费条件差异会直接影响模式的实施效果。

其次,本文的实践基础仍较有限。研究主要基于单门课程和单校教学环境,实践周期较短,尚缺少对照实验、多轮教学数据和长期跟踪证据,现有反馈只能说明该模式具有可行性,尚不足以支持普遍效果的因果推断。思政素养具有内隐性,文本分析也可能受到学生表达能力和写作风格的影响。因此,该模式不宜作为统一化模板直接复制,而应先在师资、技术和数据治理条件相对成熟的课程中试点,再根据课程特点逐步调整和推广。

6. 结论与展望

本研究围绕网络攻防课程思政教学改革,构建并实践了“专思创 + AI”融合模式,以 AI 赋能轴连接专业能力、思政素养与创新素质三条螺旋线。初步实践显示,该模式有助于改善思政案例与专业内容的匹配,增强学生的专业认同、伦理意识和创新参与,并推动思政素养评价由单一主观判断转向多源证据支持的过程性反馈。

针对本文分析的实施边界,后续研究将从三个方面展开:一是在多课程、多院校环境中开展对照或准实验研究,检验模式的稳定性和适用条件;二是完善教师复核机制,对 AI 辅助评价的公平性、可解释性和评价效度进行验证;三是逐步建设共享案例资源,并探索面向网络安全课程思政的垂直领域教学模型。

基金项目

本工作由上海高校青年教师培养资助计划项目支持,编号 EGD2024020。

参考文献

- [1] 鲁艳蓉. 网络空间安全专业课程思政融合路径探究[J]. 计算机教育, 2022(6): 1-4+9.
- [2] 曹腾飞, 牛国清, 刘薇, 等. 融合思政元素的网络攻防实验教学研究与探索[J]. 软件导刊, 2022, 21(12): 204-208.
- [3] 李妍, 罗棋, 高启浩, 等. 基于“专思创”融合模式的信息系统安全课程改革[J]. 计算机教育, 2025(9): 114-119.
- [4] 侯严严, 杨贞萍, 宋柱芹. 课程思政理念下应用型本科“计算机网络安全”课程教学改革探索[J]. 教育进展, 2025, 15(4): 1172-1178.
- [5] 刘强, 赵文涛, 柳林, 等. 面向胜任网络安全岗位的全过程思政育人模型设计与实践[J]. 计算机教育, 2024(5): 111-117.
- [6] 罗岚, 孙安, 蒋水华, 等. 基于结构方程模型的课程思政教学效果影响因素分析[J]. 中国教育技术装备, 2024(12): 66-71.
- [7] 李军祥. 数字化背景下高校课程思政考核评价实施成效监控[J]. 高教学刊, 2026, 12(14): 57-60.