双模型校验驱动的 C 语言智能出题系统研究

江玉珍，朱映辉
韩山师范学院 计算机与信息工程学院，广东 潮州 521041
DOI:10.61369/ASDS.2026060008

摘　　要： 针对 C 语言教学试题需求量大、传统出题效率低、单一 AI 生成质量不稳定等问题，本文基于 COZE 智能体平台设计并实现了一款智能出题系统。系统创新性地采用双 LLM 校验机制，实现生成与审核分离，结合知识库检索与 AI 生成的混合题源策略。实验结果表明，系统出题效率较传统模式有明显有提升，且有效地保障了试题质量与多样性。

关 键 词： C 语言；智能体；智能出题；双 LLM 校验

Research on Intelligent C Language Question Generation System Driven by Dual-LLM Verification

Jiang Yuzhen, Zhu Yinghui
School of Computer and Information Engineering, Hanshan Normal University, Chaozhou, Guangdong 521041

Abstract: To address the issues of high demand for C language teaching test questions, low efficiency of traditional question generation, and unstable quality of single-AI generation, this paper designs and implements an intelligent question generation system based on the COZE agent platform. The system innovatively adopts a dual-LLM verification mechanism to separate question generation from review, and combines a hybrid question source strategy of knowledge base retrieval and AI generation. Experimental results show that the system significantly improves question generation efficiency compared with traditional methods and effectively ensures the quality and diversity of test questions.

Keywords: C language; intelligent agent; intelligent question generation; dual-LLM verification

引言

C 语言程序设计是计算机专业的核心基础课，涵盖了程序设计思维、语法规则、算法实现等知识领域、也是数据结构、算法设计、操作系统等多门专业课的基础先修课，同时还承担着转专业考试、专升本考试等重要考核任务，是计算机专业人才培养中重要的一门核心课程，其教学质量直接影响着学生的专业素养及相关课程的通过率[1]。C 语言课程内容多，逻辑强，对实践及技能要求较高，其在教学、作业、考试、实训、竞赛等环节中均需要大量不同类型、知识点及难度的试题[2]。传统出题模式主要依赖教师手动完成，其存在三方面缺点：一是效率低下，教师需花费大量时间梳理知识点、设计题干，难以应对高频次需求；二是质量参差，手动出题易出现表述模糊、难度失衡；三是重复率高，受限于个人知识储备，试题同质化严重。近年来，大语言模型（LLM）在内容生成领域展现出强大能力，其在自动化命题方面的潜力已得到初步验证[3,4]，这为智能出题提供了新思路。现有的 LLM 均能快速生成试题，然而实践中发现，单一 LLM 可能会生成与现有试题相似的题目，也容易生成可读性差的试题[5]，且其缺乏独立质检环节，试题可用性难以得到保证。相关研究也指出，大语言模型在辅助命题时仍存在不可控因素，人工审核不可或缺[6]。对此，一些系统引入协同优化机制以提高 AI 应用上的准确性和实用[5,7]。本文基于COZE智能体平台，设计并实现了一个基于双LLM校验机制的C语言智能出题系统，实现生成与审核分离，确保试题质量可控。

一、相关技术基础

COZE 是字节跳动推出的一站式 AI 智能体开发与应用平台，支持可视化工作流配置、知识库管理、大模型调用、插件扩展、多渠道发布等，满足深度定制需求，用户无需编写复杂代码即可在该平台上快速搭建领域应用。现有的大语言模型大多已具备强大的自然语言理解与生成能力，既可用于创造性试题生成，也可用于对新试题的独立审核。因此，本系统引入两个模型分任两种

基金项目：2025韩山师范学院教学质量与教学改革工程项目（粤韩师教［2025］126 号 -59）；广东省示范性现代产业学院 -- 软件与智能物联产业学院；2023广东省教育厅科研项目“重点领域专项”（2023ZDZX1014）。韩山师范学院2025年度教授、博士启动项目（第二批）（QD2025202）。
作者简介：
江玉珍（1977-），女，韩山师范学院计算机与信息工程学院副教授，硕士，研究方向：深度学习、图形图像处理；邮箱：jiangyz@hstc.edu.cn；
朱映辉（1977-），男，韩山师范学院计算机与信息工程学院教授，硕士，研究方向：人工智能、对抗样本。

角色，实现出题与审核分离，旨在提升出题质量。

二、系统设计

（一）总体架构

系统采用分层架构：设置知识库层用于存储已有优质试题；搭建两个工作流层分别实现题库检索（search_question）与大模型新题生成（generate_test）两种功能；交互层解析用户需求并输出结果。系统定位为资深 C 语言出题人，参考教材是谭浩强著的《C 语言程序设计（第5版）》，试题指定为 C99 及以后版本，支持6类题型（选择、填空、判断、看程序写结果、程序完形填空、编程题），所有题目分5级难度，覆盖语法、数组、指针、函数、结构体等25个知识点。

（二）知识库设计

知识库包含题型、知识点、难度、题干、答案5个字段，原有试题以 Excel 格式批量导入新建知识库 C_test，支持以知识点、题型、难度为检索关键词，能确保精准匹配，也支持模糊检索。COZE 知识库支持后续的补充及更改。

（三）相关工作流

1.search_question 工作流：接收知识点、题型（选用）、难度（选用）及题量4个参数，在知识库中进行精确检索，检索结果直接展示并存入变量 {{current_set}}，供后续去重校验使用。{{current_set}} 代表了本次生成的题目集。由于入库试题均已经过人工审核，检索结果可直接使用。search_question 工作流的架构图如图1所示。

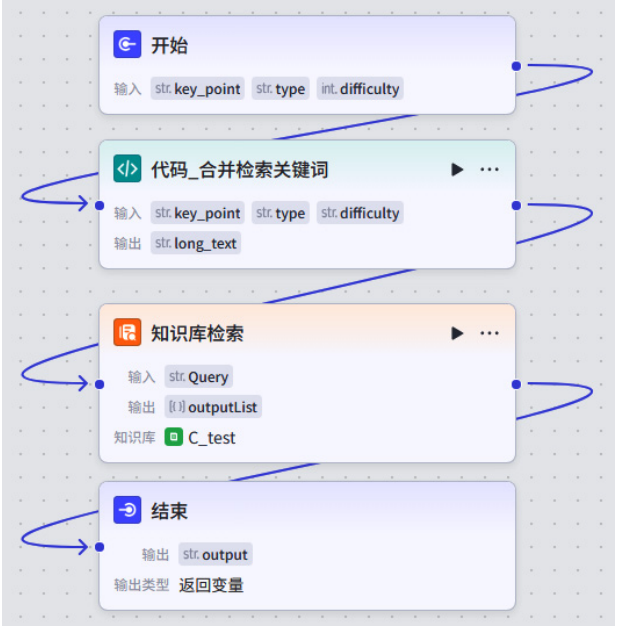

图1：search_question 工作流

2.generate_test 工作流：采用双 LLM 校验机制，两个大模型使用不同系列模型，一个作为出题 LLM，另一个作为审核 LLM。该工作流首先调用出题 LLM，根据参数输出符合要求的试题，随后调用审核模型对刚生成的新题进行0–100分的独立评分，并将分数存入变量 {{score}} 中。评分维度主要包括质量维度（代码正确性、难度匹配度、表述清晰度）和创新维度（与 {{current_set}} 的相似度）。审核 LLM 的评分标准如表1所示，评分≥80分视为合格，合格题目将以对话内容输出并更新至变量 {{current_set}} 之中；评分 <80则触发循环重新生成试题。generate_test 工作流的架构图如图2所示。多模型协同驱动的思路在教育智能体领域已有初步探索[8]，本文在此基础上，将这一思想引入 C 语言出题场景，通过“出题－挑错”分离机制，确保每道试题经过独立质检，避免单一模型的自我确认偏差。

表1：审核 LLM 的评分标准

维度	子指标	分值	扣分条件
题目正确性	代码语法正确、逻辑正确、边界处理正确等	40分	有明显错误导致该题不可用的全扣（该维度得0分）；多余表述等不影响题目使用每处扣10分
创新性	与已有试题相似度	20分	相似度 >30% 扣10分；>50% 全扣（该维度得0分）
表述清晰度	无歧义、格式规范	20分	每处模糊、造成理解困难扣10分
难度匹配度	与指定级别相符	20分	偏差1级扣10分

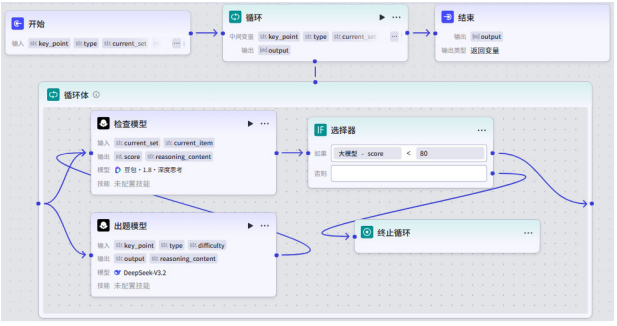

图2：generate_test 工作流

（四）自动召回与弹性补偿机制

选择智能体作为出题系统的优势是支持自然语言指令并能自动、灵活地处理。在知识库检索应用中，COZE 智能体提供了两种对用户非常友好的机制：自动召回机制与弹性补偿机制。

1. 自动召回机制是一种默认开启的“自动查知识库”功能。当需要做知识库检索时，系统首先启动“自动召回”对知识库进行简单相似度匹配，这是一种粗糙又快速的检索方法。只有当“自动召回”不能获得有效知识库反馈时，才启动 search_question 工作流进行精细检索。对于数据庞大的题库而言，“自动召回”大多情况下均能匹配到试题，即“自动召回”的响应成功率高，因此能有效地提高系统对知识库的检题效率。

2. 弹性补偿机制是一种动态资源调度策略，用于解决知识库检索结果不足时的资源缺口问题。对于出题系统而言，当某类题知识库检索出的数量不能满足需求时，系统会自动计算缺口。例如，需求5道题库试题和5道 AI 新题，但系统在知识库中仅找到3道，缺口为2道，则智能体能自动将差额转由 AI 生成补充，即

search_question 工作流停止后会自动启动 generate_test 工作流生成7道新题，使总题数仍为10道。弹性补偿机制实现了检索与生成的动态协同，保障了用户的特定用题需求。

三、实验与分析

以下通过3类实验数据验证该出题系统的实用性和有效性。

实验一：不同出题方式的综合评估

该实验按全知识库、全 AI 新题、知识库 +AI 新题，弹性补偿的知识库 +AI 新题4种方式测试出题效果，并从出题效率、试题质量（按表1的评分标准）两个维度对系统进行综合评估。测试计算机为 Intel(R) Core(TM) i7-8550U CPU @ 1.80GHz，内存8.00 GB。COZE 智能体的知识库预先导入了524道优质 C 语言试题，涵盖6类题型、1至5级难度及各类知识点。针对10道指定类型题的生成任务，经16次试验，4种生成方式的平均出题时间和质量分均值如表2所示。

表2：不同出题方式的综合评估

出题方式	生成时间	质量分均值
全知识库（含自动召回）	9.5s	97.1
全 AI 生题	83.7s	93.5
知识库 +AI 生题（各50%）	72.1s	96.8
弹性补偿的知识库 +AI 生题	91.8s	94.3

实验二：双 LLM 低分题分析

为验证双 LLM 系统的特性，在 AI 生题过程中提取部分 <80 的题目（即被弃用的题目），并结合审核 LLM 的评分进行分析。以下为两道错题的代表案例。

案例1：以下对二维数组 a 的初始化语句中，错误的是（ ）

A. int a[2][3] = {{1,2}, {3,4}};

B. int a[][3] = {1,2,3,4,5,6};

C. int a[2][3] = {1,2,3,4,5,6,7};

D. int a[2][] = {1,2,3,4,5,6};

答案 D

提取审核 LLM 分析：该题为数组类难度2级单项选择题，score 值为55，低分原因是题目存在两个错误选项（C 和 D），C 选项初始化值个数超过数组元素个数，编译无法通过；D 选项二维数组定义时第二维长度不可缺省。题目所给答案为 D，但 C 选项也明显错误。这类题目在出题上存在明显问题，不可使用。

案例2：分析以下程序，最终输出结果是（ ）

```
#include
int main()
{    int arr[5] = {1, 2, 3, 4, 5};
   int *p = arr;
   *(p++) += 12;
   *++p = *(p + 2);
   printf( “%d %d\n”, arr[2], p[2]);
   return 0;
}
```

A.33　B.54　C.35　D.55

答案 D

提取审核 LLM 分析：该题为指针类难度3级单项选择题，score 值为72，低分原因是“*++p = *(p + 2);”中，p 被多次修改和多次使用，存在未定义行为（UB），不同编译器结果可能不同。C 语言标准规定：在同一表达式中，如果对一个变量既有修改又有读取，且读取不是为了计算新值，则行为未定义。虽然在大多数编译器中可能得到 D 选项的结果，但这种一行代码完成过多操作的方式存在刻意复杂化嫌疑，实际工程中不推荐这种写法，不建议作为试题使用。

以上案例说明审核 LLM 能细致地发现出题 LLM 所给题目中不易觉察的错误，能较合理地对题目进行评分，并有效拦截不可用题目。

实验三：单 LLM 与双 LLM 出题效果比较

为验证双 LLM 相较于单 LLM 在出题上的优势，本实验参照 generate_test 工作流架构模式，分别创建了无审核的单 LLM 出题工作流和自审核的单 LLM 出题工作流。三组方案各生成100道题目，每组方案的题目要求相同，合格阈值仍按表1评分标准执行，所得数据如表3所示。

表3：单 LLM 与双 LLM 出题效果对比

方案	合格率	质量分均值	重生成次数
单 LLM（无审核）	89%	85.3	0
单 LLM（自审）	94%	91.0	8
双 LLM（本文方案）	100%	95.1	14

上述数据说明，单 LLM 方案无论是否带有自审机制，其生题时均有一定概率产出不可用但容易被忽视的题目，即单 LLM 存在审核缺乏或审核不力的问题。即使带有自审机制的单 LLM，其自我审核仍存在一定的不可靠性，其根源在于目前大多数 AI 存在“确认偏误”，即模型倾向于认可自己生成的内容。而引入独立的审核模型后，就能明显规避这一问题。这也是本文在出题系统工作流中引入不同类双 LLM 的原因：利用不同模型间的能力互补，提升出题与审核效果。

四、结论与展望

本文设计了基于双 LLM 校验机制的 C 语言出题智能体，该系统通过生成与审核分离、设置质量门禁、同批次去重等策略，有效解决了传统出题模式的效率与质量问题。在该系统中，生成与审核分离机制确保了两个 LLM 各司其职——生成模型专注于创造性内容产出，审核模型则严格把控质量，二者形成良性互补；质量门禁的设置通过多维度评分体系对试题进行筛选，只有达到预

设标准的题目才能存入当前题目集，从而保证了输出内容的可靠性；同批次去重策略则避免了相似或重复试题的冗余出现，提升了题库的整体质量与多样性。

未来工作主要包括：细化审核评分维度，引入区分度等教育测量学指标，构建更加科学完善的试题评价体系；扩展知识库规模与覆盖范围，提升系统的通用性与专业深度；支持更复杂的用户需求，如智能批量组卷、试卷难度梯度调控等高级功能，进一步提升系统的智能化水平与用户体验，使其成为真正赋能教育场景的智能化出题助手。

参考文献

[1] 刘阳，陈峰 .C 语言程序设计课程教学改革探索 [J]. 安徽工业大学学报 (社会科学版),2025,42(03):55-57.

[2] 赵雪梅，张宏，王如刚，等 . 互联网 + 教学模式下 C 程序设计题库平台建设 [J]. 电脑知识与技术 ,2023,19(12):177-180.

[3] 陈大建，胡杰辉 . 基于大语言模型的自动化命题研究——以英语阅读理解试题为例 [J]. 外语教学 ,2025,46(02):40-47.

[4] 杨志明，徐庆树，祁长生 .ChatGPT 命题潜力的实证研究 [J]. 教育测量与评价 ,2023(04):3-13.

[5] 景宏伟，徐娟 . 三类大语言模型自动出题的效能测评——以 HSK6 级阅读理解题为例 [J]. 考试研究 ,2026,22(01):44-53.

[6] 李峰，郭嘉悦，胡新雨，等 . 大语言模型辅助情境化命题模式探索——以创造性思维测评为例 [J]. 中国考试 ,2025(09):76-86.

[7] 师曼飞，钱玉航，杨淑琪，等 . 知识库与多模型协同驱动的循证护理知识问答智能体的构建与评价研究 [J]. 中华护理教育 ,2025,22(09):1036-1042.

[8] 刘嘉，李楠，梁晓娜，等 . 基于多智能体系统的中文学术问答与应用 [J]. 昆明理工大学学报 (自然科学版),2025,50(03):58-69.

