Overseas English Testing: Pedagogy and Research 国外英语考试教学与研究, 2026, 8(3), 111-122
Published Online August 2026 in Hans. https://www.hanspub.org/journal/oetpr
https://doi.org/10.12677/oetpr.2026.83014

从任务委托到协作迁移：生成式人工智能支持新托福听力备课的实践反思

——以“Listen to an Announcement”课程开发为例

吴　奇

新东方国际教育事业部青岛分中心，山东　青岛

收稿日期：2026年7月25日；录用日期：2026年8月21日；发布日期：2026年8月31日

摘　要

生成式人工智能(Agentic AI)正在进入教师备课、资源开发与课程设计等日常专业活动，但已有讨论多集中于内容生成效率、提示词技巧或使用风险，对教师与AI在持续互动中如何共同推进问题、修正知识的过程关注仍显不足。文章以笔者使用Codex准备新托福听力“Listen to an Announcement”课程的一次完整实践为案例，基于53篇不重复听力文本、105道题目、连续人机对话、AI生成的系列备课文档以及最终形成的379页课件，选取三个关键事件进行反思性分析。研究发现：第一，Agentic AI能够承担文件读取、任务拆解、语料统计和关联文档更新等长链条工作，使教师得以从逐篇处理材料转向观察整体规律；第二，AI产出并非只是问题的答案，也可能使原本隐而未显的现象变得可见，进而触发教师提出新的研究问题；第三，AI的专业误判往往具有表面合理性，并可能借助自动化能力扩散至统计、规律总结和例题选择等多个环节。文章据此提出“任务委托－结果检视－问题生成－再分析－专业审查－教学转化”的人机协作路径。

关键词

生成式人工智能，教师备课，新托福听力，人机协作，教师能动性

From Task Delegation to Collaborative Transfer: Practical Reflections on Agentic AI-Supported TOEFL iBT Listening Lesson Preparation

—A Case Study of “Listen to an Announcement” Course Development

Qi Wu

Qingdao Center, Global Education Office, New Oriental Education & Technology Group, Qingdao Shandong

文章引用: 吴奇. 从任务委托到协作迁移: 生成式人工智能支持新托福听力备课的实践反思[J]. 国外英语考试教学与研究, 2026, 8(3): 111-122. DOI: 10.12677/oetpr.2026.83014

Received: July 25, 2026; accepted: August 21, 2026; published: August 31, 2026

Abstract

Agentic AI is increasingly entering teachers' everyday professional practices, including lesson preparation, resource development, and course design. Existing discussions, however, have often focused on content-generation efficiency, prompt engineering, or potential risks, while paying less attention to how teachers and AI systems jointly advance questions and revise knowledge through sustained interaction. This article presents a reflective case study of the author's use of Codex to prepare a TOEFL iBT Listening course on "Listen to an Announcement". The analysis draws on 53 unique listening texts, 105 questions, continuous human-AI dialogues, a series of AI-generated lesson-preparation documents, and a 379-page courseware product developed from these materials. Three key events are examined during this process. The study finds that: First, Agentic AI can support long-chain work such as file reading, task decomposition, corpus-based statistics, and cross-document updating, thereby allowing teachers to move from item-by-item processing to the observation of broader patterns. Second, AI outputs may function not only as answers but also as prompts that make previously unnoticed problems visible. Third, AI errors may appear superficially reasonable and can spread quickly across statistics, pattern summaries, and teaching examples through automated updating. The article therefore proposes a human-AI collaboration pathway of "task delegation, result inspection, question generation, re-analysis, professional review, teaching transformation, demonstration feedback, and collaborative transfer".

Keywords

Agentic AI, Teacher Lesson Preparation, TOEFL iBT Listening, Human-AI Collaboration, Teacher Agency

1. 引言

在开始准备新托福听力“Listen to an Announcement”课程时，笔者首先感到的并非“没有材料”，而是材料太多，却缺少一套足以支撑教学的组织方式。当前语料包括公告文本、题目、选项和若干背景说明。逐篇阅读并不困难，真正困难的是将这些材料整合起来，回答一组相互关联的问题：公告是否存在相对稳定的话语结构？六类题型的比例如何？正确选项与干扰项有哪些可教的规律？哪些题适合作为课上例题，哪些适合课后练习？学生在没有看到题目之前，又应当怎样听懂主旨、事实、推断与话语功能？

本案例最终纳入 54 条材料记录，其中两条内容重复，按不重复文本计算为 53 篇，共包含 105 道题。若完全依靠人工逐项标注，不仅耗时，而且每次调整分类标准之后，相关统计、选项分析和教学材料都要重新核对。更重要的是，作为教师，虽然熟悉托福听力教学，却不可能在分析开始之前就预先知道这批语料中有哪些值得研究的规律。换言之，备课面对的不只是工作量问题，也是一个开放的问题发现过程。

这正是本次研究使用 Codex 的现实背景。最初的设想相当朴素：让 AI 读取文件、整理题目、完成统计，为后续课程设计节省时间。然而，随着任务推进，笔者逐渐意识到，这次实践并不能简单概括为“AI

提高了备课效率”。AI 的中间结果改变了观察材料的方式，有些结果激发了最初并未想到的问题；与此同时，一次具有迷惑性的分类错误又迫使重新确认：哪些工作可以委托，哪些判断不能让渡。

大型语言模型能够支持教育内容生成、个性化反馈与教学资源开发，但其局限、偏差和非预期脆弱性也要求使用者保持事实核查与持续监督[1]。UNESCO 发布的生成式人工智能教育指南同样强调以人为本的使用原则，主张把伦理验证和教学设计置于技术采用过程之中[2]。在这一背景下，本文不试图证明 AI 可以“自动备课”，而是讨论三个更具体的问题：Agentic AI 如何进入一个多阶段的真实备课项目？AI 反馈如何触发教师形成新的研究问题？当 AI 出现专业误判时，教师的知识与责任体现在哪里？

2. 背景：从单次内容生成到长链条备课协作

2.1. “Listen to an Announcement”备课的三重任务

在本案例所使用的材料中，“Listen to an Announcement”题型以校园生活为主要语境，篇幅短、信息集中，题目涉及 Main Idea、Factual、Inference、Purpose、Method 和 Attitude 六类。对教师而言，完成这部分备课至少包含三个层面：第一是语料研究。教师要判断公告的交际目的、篇章结构以及重要信息的分布位置。第二是测评研究。教师不仅要确认答案，还要理解题目究竟考查什么、干扰项如何偏离原文。第三是教学转化。统计结果本身不能直接成为课堂内容，教师必须将“数据中的规律”转化为学生可以执行的听辨、判断、记笔记和排除选项的动作。

这三个层面相互牵连。例如，一道题的类型如果发生变化，六类题型的数量与占比会随之改变；按照题型归纳的正确、错误选项特点需要重算；原先选择的 Purpose 例题与 Inference 例题也可能需要调整。备课因而不是若干孤立文本的生成，而是一组存在依赖关系的知识产品。

2.2. 本文所说的 Agentic AI

本文使用的 Codex 本身是面向软件开发的智能代理，但其文件读取、任务规划、工具调用、结果检查和文件更新能力，也使其能够参与结构化的非编程知识工作。OpenAI 的 Codex 文档将其描述为可以理解项目、执行重复工作流并围绕复杂任务进行计划与验证的智能代理[3]。本研究未使用 Codex 开发软件，而是将装有文本、题目、统计表和备课文档的文件夹当作一个持续演化的“备课项目”。

从技术研究看，基于大型语言模型的自主代理通常不只生成一段回复，还会围绕目标组织规划、记忆、工具使用和行动等环节[4]。因此，本文将“Agentic”理解为一种工作方式：AI 能够在给定目标和文件环境中连续处理多个步骤，并在教师反馈后修改一组相互关联的成果。这里的“自主”并不意味着 AI 独立决定教学目标，更不意味着教师退出过程；相反，谁来设定目标、判断结果、终止错误，正是本文要讨论的核心。

近年来，教师使用生成式 AI 备课的研究开始从“能不能生成教案”转向“教师如何与 AI 互动”。Flavin、Hwang 和 Morales (2025)指出，AI 支持的备课中，知识建构会在连续提问、回应和评价中动态协商[5]。Velander (2026)对教师备课交互日志的分析则发现，教师常把 AI 置于主要内容生成者的位置，真正的协商与共同建构相对少见[6]。这一发现为本案例提供了一个有用的参照：关键并不在于使用了多少次 AI，而在于教师是否对 AI 产出进行了实质性的追问、修正与再组织。

3. 研究方法：基于真实记录的反思性案例分析

本文采用反思性单案例分析。案例中的教师既是备课任务的提出者和课程开发者，也是实践过程的回顾者。反思性实践并非在行动结束后简单总结成败，而是重新审视行动中的判断如何形成、如何发生变化，以及专业知识怎样在具体问题中发挥作用[7]。

本文的数据包括五类：第一，最初提供给 Codex 的背景资料和 PDF 任务清单；第二，53 篇不重复公告文本与 105 道题目；第三，教师与 Codex 之间的连续对话；第四，AI 生成并经多轮修改的结构分析、题型统计、选项规律、例题选择和课程框架等文件；第五，教师在这些材料基础上完成的 379 页课件。对材料的回顾以时间顺序为主，同时比较修改前后的分类、统计和文档关系。

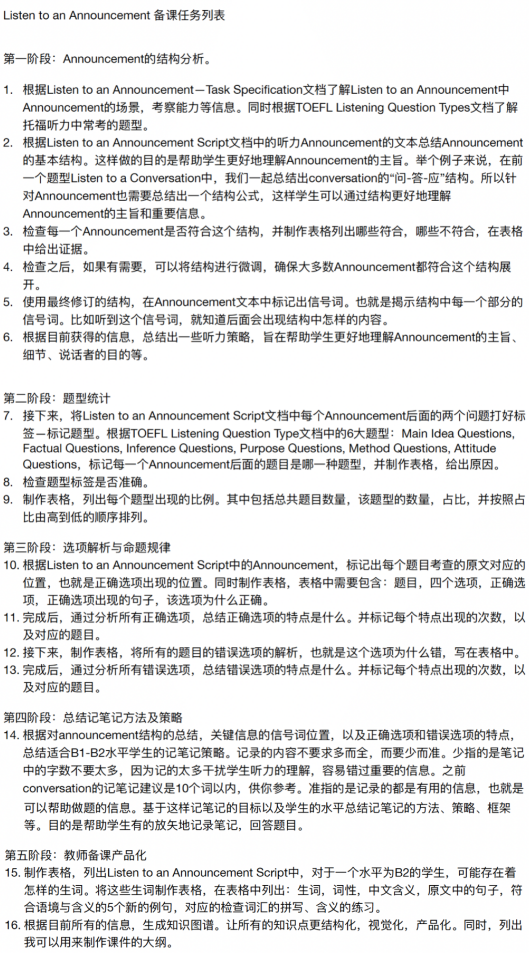
在完整记录中，本文选择三个关键事件。选择标准不是“最成功的三个结果”，而是三个事件分别呈现了不同的人机关系：AI 作为复杂任务的执行者，AI 作为新问题的触发者，以及教师作为专业判断的审定者。每个事件按照“情境－互动－结果－反思”展开。

需要说明的是，本文分析的是一次具体备课实践，其证据主要来自过程记录和课程成果，尚不能据此推断学生成绩的变化。

4. 三个关键事件

4.1. 事件一：从一份任务清单到一套备课成果

最初，笔者先向 Codex 提供两份背景文档，随后上传整理好的公告文本、题目与选项。在确认系统可以准确读取之后，又请它完成全部题目并标记正确选项。最后，把真正需要完成的工作写成一份 PDF 任务表，要求其围绕不同阶段生成结构分析、题型统计、选项分析、笔记策略和课程框架，见图 1。



Listen to an Announcement 备课任务列表

第一阶段：Announcement的结构分析。

1. 根据Listen to an Announcement—Task Specification文档了解Listen to an Announcement中Announcement的场景，考察能力等信息。同时根据TOEFL Listening Question Types文档了解托福听力中常考的题型。
2. 根据Listen to an Announcement Script文档中的听力Announcement的文本总结Announcement的基本结构。这样做的目的是帮助学生更好地理解Announcement的主旨。举个例子来说，在前一个题型Listen to a Conversation中，我们一起总结出conversation的“问-答-应”结构。所以针对Announcement也需要总结出一个结构公式，这样学生可以通过结构更好地理解Announcement的主旨和重要信息。
3. 检查每一个Announcement是否符合这个结构，并制作表格列出哪些符合，哪些不符合，在表格中给出证据。
4. 检查之后，如果有需要，可以将结构进行微调，确保大多数Announcement都符合这个结构展开。
5. 使用最终修订的结构，在Announcement文本中标记出信号词。也就是揭示结构中每一个部分的信号词。比如听到这个信号词，就知道后面会出现结构中怎样的内容。
6. 根据目前获得的信息，总结出一些听力策略，旨在帮助学生更好地理解Announcement的主旨、细节、说话者的目的等。

第二阶段：题型统计

7. 接下来，将Listen to an Announcement Script文档中每个Announcement后面的两个问题打好标签—标记题型。根据TOEFL Listening Question Type文档中的6大题型：Main Idea Questions, Factual Questions, Inference Questions, Purpose Questions, Method Questions, Attitude Questions，标记每一个Announcement后面的题目是哪一种题型，并制作表格，给出原因。
8. 检查题型标签是否准确。
9. 制作表格，列出每个题型出现的比例。其中包括总共题目数量，该题型的数量，占比，并按照占比由高到低的顺序排列。

第三阶段：选项解析与命题规律

10. 根据Listen to an Announcement Script中的Announcement，标记出每个题目考查的原文对应的位置，也就是正确选项出现的位置。同时制作表格，表格中需要包含：题目，四个选项，正确选项，正确选项出现的句子，该选项为什么正确。
11. 完成后，通过分析所有正确选项，总结正确选项的特点是什么。并标记每个特点出现的次数，以及对应的题目。
12. 接下来，制作表格，将所有的题目的错误选项的解析，也就是这个选项为什么错，写在表格中。
13. 完成后，通过分析所有错误选项，总结错误选项的特点是什么。并标记每个特点出现的次数，以及对应的题目。

第四阶段：总结记笔记方法及策略

14. 根据对announcement结构的总结，关键信息的信号词位置，以及正确选项和错误选项的特点，总结适合B1-B2水平学生的记笔记策略。记录的内容不要求多而全，而要少而准。少指的是笔记中的字数不要太多，因为记的太多干扰学生听力的理解，容易错过重要的信息。之前conversation的记笔记建议是10个词以内，供你参考。准指的是记录的都是有用的信息，也就是可以帮助做题的信息。基于这样记笔记的目标以及学生的水平总结记笔记的方法、策略、框架等。目的是帮助学生有的放矢地记录笔记，回答题目。

第五阶段：教师备课产品化

15. 制作表格，列出Listen to an Announcement Script中，对于一个水平为B2的学生，可能存在着怎样的生词。将这些生词制作表格，在表格中列出：生词，词性，中文含义，原文中的句子，符合语境与含义的5个新的例句，对应的检查词汇的拼写、含义的练习。
16. 根据目前所有的信息，生成知识图谱。让所有的知识点更结构化，视觉化，产品化。同时，列出我可以用来制作课件的大纲。

Figure 1. Initial task list and teacher instructions
图 1. 最初 PDF 任务清单与教师指令截图

收到任务后，Codex 没有只返回一篇长回复，而是将结果写入多个文件，并建立索引，见图 2。第一步：分析公告结构、语境和交际目的；第二步：标注六类题型并计算数量与占比；第三步：核对 105 道题，分析正确、错误选项特征，并按教学用途选择例题；第四步：整理笔记策略；第五步：形成知识图谱与课程框架。此后每次修改，系统都可以定位并更新相关文件。

```
# Listen to an Announcement 备课成果索引

本文件夹按照任务表的五个阶段组织成果。

## 第一阶段：Announcement 结构分析

- `01_Announcement_Structure_Analysis.md`：修订后的 Opening/Greeting–Purpose–Details–Action 结构公式、篇章类型分类、信号词频次与教学策略。
- `01_structure_fit_table.csv`：54 条 announcement 的结构符合度检查表，含四个结构部分的证据。
- `01_genre_classification.csv`：54 条 announcement 的篇章类型、子类型、分类依据与原文摘要。
- `01_signal_word_statistics.csv`：Opening/Greeting、Purpose、Details、Action 四部分的信号词与出现次数。
- `01_signal_marked_texts.csv`：每条 announcement 中对应四个结构部分的信号词标记。
- `01_Listening_Comprehension_Strategies.md`：从不看题目的角度，总结学生听录音时理解主旨、事实、推断、句子目的、说明方式和态度的策略。
- `01_Purpose_Signal_Patterns.md`：按照三大公告目的，总结揭示目的的原文句型、关键词和教学使用建议。
- `01_purpose_signal_patterns.csv`：三大公告目的、目的信号句型、关键词、原文例句和典型材料的表格版。

## 第二阶段：题型统计

- `02_question_type_labels.csv`：105 道题逐题题型标签与原因。
- `02_question_type_stats.csv`：六大题型数量与占比。
- `02_Question_Type_Statistics.md`：题型统计摘要与判断原则。

## 第三阶段：选项解析与命题规律

- `03_question_option_analysis.csv`：105 道题的题干、四个选项、正确答案、证据句、正确/错误选项解析。
- `03_question_option_analysis_B1.csv`：面向 B1 中国初高中学生的逐题选项解析，语言更适合课堂讲解。
- `03_Answer_Audit.md`：105 道题正确选项复核记录与特殊题目说明。
- `03_wrong_option_feature_table.csv`：315 个错误选项的错误类型与解析。
- `03_Option_Patterns.md`：正确选项特点、错误选项特点、命题规律总结。
- `03_Type_Based_Option_Patterns.md`：按照六大题型分别总结正确选项和错误选项特点。
- `03_Question_Selection_For_Teaching.md`：按题型整理课上例题、课后练习和拔高难题。
- `03_question_selection_by_type.csv`：课上例题、课后练习、拔高难题的表格版。

## 第四阶段：记笔记方法及策略

- `04_Note_Taking_Strategies.md`：适合 B1–B2 学生的 P–D–A 笔记框架、10 词以内示范、听力策略。

## 第五阶段：教师备课产品化

- `05_B2_vocabulary_table.csv`：44 个可能影响 B2 学生理解的词汇，含词性、中文含义、原文句、5 个例句和检查练习。
- `05_Knowledge_Map_and_Course_Outline.md`：知识图谱与课件大纲。
```

Figure 2. Codex-generated file index
图 2. Codex 生成文件索引

这一阶段最明显的变化，是材料第一次获得了“全局可见性”。过去备课时，往往以单篇文本或单道题为单位思考；当 AI 将 53 篇文本和 105 道题放进统一表格后，可以看到不同题型的比例、不同公告之间的共性，以及哪些例子偏离一般规律。AI 在这里承担的是大规模读取、初步分类和重复整理。它没有替教师决定课程怎样教，却改变了教师能够看到什么。

从备课劳动角度看，Agentic AI 的优势并非单纯“写得快”，而是维持一套关联成果。传统聊天机器人往往在一次回复后结束任务，本案例中的 Codex 则持续读写项目文件，根据新要求修改表格和说明文档。这种能力释放了教师的一部分操作性劳动，使注意力能够转向分类是否成立、规律是否值得教等较高层次的问题。

不过，第一批文件仍只是研究底稿。它们看起来完整，甚至带有精确的数量和百分比，但完整并不意味着可靠，更不意味着能够直接进入课堂。此后的两个事件正是从“检视结果”开始的。

4.2. 事件二：从 5 道 Purpose Questions 到 53 篇文本

第一轮题目与选项分析完成后发现，原有总结过于宽泛。于是，要求 Codex 按照六大题型分别归纳

正确选项和错误选项的特点，并重新站在英语水平为 B1 的中国初高中生角度检查解析；与此同时，还要将各题型材料进一步分为课上例题、课后练习和拔高难题。这次追问并不是预先写在任务表里的，而是阅读初始成果之后产生的。

新的分类结果显示，当时 105 道题中只有 5 道被标为 Purpose Questions。这个数字又引出了一个此前没有被明确提出的问题：这 5 道题实际考查了说话者怎样的局部目的？笔者随后向 Codex 提出了一个范围更大的要求：“请放眼全部的 53 篇文本，如果推断说话者目的，比如‘Why does the speaker mention...’，还可能存在哪些不同的目的？”

系统重新扫描全部文本，将潜在的话语功能概括为 9 类，包括解释行动建议或安排变化的原因、帮助听众定位地点、说明活动要求、强调行动的重要性或后果、鼓励参加、介绍活动内容、说明项目意义、安抚听众以及说明资格或责任分工，并为不同类别匹配原文句子和上下文。

这一结果没有结束讨论。阅读 9 类功能后，笔者进一步提出：整篇公告已经分为介绍类、变动类和提醒类，那么三类公告中的局部话语目的是否具有不同倾向？进一步分析后，区分出两个层次：其一是整篇公告的宏观交际目的，即“为什么发布这则公告”；其二是某句话的微观功能，即“为什么在这里提到这个细节”。由此形成了更适合课堂讲解的判断方式：介绍类公告中的局部信息往往服务于“让听众更想参加、了解或使用”；变动类公告中的信息往往帮助听众“理解为什么改变以及如何调整”；提醒类公告中的信息则更多说明“必须做什么、怎样做以及为什么要做”。具体过程如下：

1) 互动阶段：当时可见的结果；新产生的问题；新形成的认识。

2) 六类题型分析：Purpose 题数量较少；5 道题分别考查什么功能；现有 Purpose 题的功能分类。

3) 扩展语料分析：53 篇不重复文本；还可能形成哪些 Why mention 题；9 类潜在局部话语功能。

4) 交叉分析：公告分为介绍、变动、提醒三类；不同公告类型中的局部目的有何差异；宏观目的与微观功能的两层框架。

回看这段过程，最值得注意的并不是 AI 一次性给出了 9 个类别，而是研究问题本身发生了变化。最初的任务是“总结已有题目”，随后变成“利用全部语料推测可考查的话语功能”，最后又变成“将局部功能放回不同类型公告的整体目的中理解”。后一个问题建立在前一个回答提供的新视野之上。

AI 参与备课不失为一种动态协商的知识建构。本案例进一步显示，协商并不总表现为对同一个答案反复润色，也可能表现为 AI 先压缩大量材料，使某种分布或缺口变得可见；教师再凭借教学经验判断其中什么值得追问。AI 没有替教师提出最终的研究方向，教师也并非独自完成全部语料归纳。二者形成的是“产出–理解–追问–再分析”的循环。

这与单纯训练提示词有所不同。提示词当然影响结果，但更关键的是教师能否从结果中识别尚未回答的问题。Velander (2026)发现，实际备课互动中，教师往往将 AI 当作主要生成者，协商和共同建构较少[6]。本事件恰好说明，要从“委托生成”走向“认知共建”，教师必须保留提出新问题的主动权。

4.3. 事件三：“Imply”不一定是“Inference”

事件二发生时，我们认为题库中有 5 道 Purpose Questions。后来的逐题检查注意到一处分类：

> What does the speaker imply when he says, “Don’t take a chance on missing out on critical information”?

Codex 最初将它归入 Inference Questions。乍看之下这一归类似乎合理，毕竟题干中出现了“imply”。但从托福听力题型的考查对象看，这道题引用了说话者的一句具体话语，要求考生理解这句话在当前语境中承担什么功能；它并不是要求考生根据时间、地点、变化或其他事实推导未明说的信息。因此，该题的本质是 Purpose，而不是 Inference。和 Codex 的对话见图 3：

Figure 3. Misclassification between Purpose and Inference questions, teacher correction, and AI revision
图 3. Purpose/Inference 题目误分类、教师纠错及 AI 修改回复

纠正之后，Purpose Questions 由 5 道变为 6 道，Inference Questions 由 7 道变为 6 道。更重要的是，这不只是改动表格中的一个单元格，题型标签、数量与占比、按题型总结的正确和错误选项特征、教学例题选择、Purpose Questions 专题分析以及文件索引都需要同步调整。Codex 完成了关联文件更新，并将判断原则修改为：不能仅凭题干中的“imply”分类；如果题目引用原文中的具体话语并考查其上下文功能，应归入 Purpose。

这个错误之所以值得写入本文，恰恰因为它不是一眼就能发现的荒谬答案，而是一个具有语言表面依据的专业误判。大型语言模型生成的内容流畅、结构完整，容易给人以“已经分析过”的印象。Kalenda 等(2025)发现，职前教师在真正分析 AI 生成教案之后，对其完成完整、详细教案的评价反而下降，并普遍认为输出仍需修改[8]。这说明接触 AI 结果与审查 AI 结果是两种不同的活动。

在本案例中，教师专业知识发挥作用的地方，不只是知道“正确标签是什么”，还包括解释原标签

为什么错。若修正停留在“把 Inference 改成 Purpose”，类似错误仍可能出现；只有将判断对象从题干中的表面词汇转回题目实际要求的认知操作，新的分类标准才具有迁移价值。

这一事件还揭示了 Agentic AI 能力的双面性。系统能够在短时间内同步修改多个文件，这是长链条备课的便利；然而，如果基础标签错误，同一种联动能力也可能将错误扩散到统计、规律归纳和课堂例题中。自动化并不会天然提高知识质量，它只是提高了某个判断被执行和传播的速度。因此，越靠近数据源头、影响范围越大的标签，越需要教师优先复核。

5. 讨论：AI 改变了什么，教师又承担了什么

5.1. 备课效率的变化，本质上是注意力重新分配

如果仅从时间上评价本次实践，AI 确实显著提高了处理速度。它可以读取多份材料、生成表格、统计频率、维护文件索引，并在标准改变后批量更新相关成果。但效率并非本文最重要的发现。更有意义的变化是教师注意力的重新分配：从复制、计数和整理，转向定义类别、解释差异、选择研究方向和设计课堂活动。

这并不意味着低层次工作从此无需核查。相反，教师将操作性劳动委托给 AI 之后，需要设计新的审查节点。教师不必手工完成每一次计数，却要确认计数所依据的标签；不必亲自给每份文档改数字，却要确认修改是否覆盖了全部关联文件。备课劳动没有简单消失，而是从“逐项生产”转向“管理一套人机协作的知识流程”。

5.2. AI 的认知价值在于让问题浮现

事件二表明，AI 并非只在教师已经知道答案时才有用。当语料规模超出人脑可以随时整体把握的范围时，机器的整理和聚合能够制造一种新的观察位置。教师看到“只有 5 道 Purpose Questions”之后，才进一步追问 5 道题的功能、53 篇文本的潜在功能以及不同公告类型的局部目的。问题不是凭空由 AI 创造的，也不是教师在任务开始前完整预设的，而是在互动中逐步生成的。

这种作用可以称为“认知催化”，但不宜将其理想化为 AI 拥有与教师相同的教学理解。AI 善于快速提出候选分类、搜索跨文本的相似表达；教师则判断分类是否符合考试构念、是否具有教学价值、是否适合学生理解。二者贡献不同，却在连续互动中彼此依赖。

5.3. 教师能动性体现为对 AI 说“不”

教师能动性不等于所有内容都由教师亲自完成。在 AI 支持的教学设计中，能动性更集中地表现为教师能否设定目标、质疑输出、保留自己的判断并承担结果责任。Krushinskaia、Elen 和 Raes (2026)在职前教师使用生成式 AI 进行教学设计的研究中发现，即使互动时间较长，参与者仍可能缺少对 AI 建议的实质反馈；与系统持续交谈并不自动意味着更强的协作性[9]。

本案例中的关键节点恰好都与教师的“不接受”有关：不接受过于宽泛的选项规律，于是要求按六类题型重做；不满足于 5 道现有目的题，于是把问题扩展到 53 篇文本；不接受“imply”必然对应“Inference”的表面判断，于是重建分类标准。由此可见，高质量人机协作的标志不应只是生成文件的数量或对话轮数，而应包括教师是否真正改变了问题、标准和最终成果。

5.4. 从数据底稿到教师示范：教学逻辑的显性化

备课项目后期，在 AI 生成的统计和分析基础上制作了 379 页课程课件。题型占比影响了各教学模块的篇幅；公告目的与信号句被转化为抓主旨的课堂练习；“Greeting、Purpose、Detail、Closing”结构被

转化为听力理解和笔记框架；正确、错误选项规律被转化为“圈证据、辨析干扰项和总结策略”的活动；按难度选择的题目则进入“课前测评、课上例题、课后练习和综合测评”。

这一过程说明，“数据能够支持教学”与“数据已经成为教学”之间仍有距离。AI 给出的是频率、类别、候选例句和初步解释；教师需要根据学生水平重新排序、删减、改写，并决定课堂上先呈现什么、让学生做什么、何时给出规律。教师的不可替代性，不应停留在抽象宣言上，它具体存在于这些微小而连续的教学决定中。

课件完成后，再将整套 379 页课件发回 Codex，并明确要求它注意三个方面：课件在什么位置使用了此前生成的数据、这些数据被转化成了怎样的课堂活动，以及每节课的授课结构和推进逻辑。这个动作与一般的“请评价我的课件”并不相同。它实际上是教师向 AI 提供了一份完整的教学示范：同一批统计数据经过怎样的选择、排序和再表达，才会成为学生可以参与的课堂。

在阅读课件后，Codex 识别出题型占比与课程模块权重、Purpose 信号与主旨训练、“Greeting、Purpose、Detail、Closing”结构与笔记框架、选项规律与课堂辨析活动、例题选择与测评环节之间的对应关系，并进一步生成“06_Courseware_Data_Reference_Map.md”，将“备课数据文件－课件页面－课堂用途”连接起来。原本主要存在于教师经验中的教学转化逻辑，由此被整理成一份可检索、可复用的说明文档。

这里所说的“让 AI 学习教师思路”，并不是对基础模型进行参数训练，也不意味着 AI 获得了脱离材料的永久记忆。更准确地说，教师通过课件、解释和映射文档，将隐性的教学判断外化为当前项目环境中的显性上下文。AI 此后不仅知道“有哪些数据”，还知道“教师通常怎样使用这些数据”。教师反馈因而不再只是纠正对错，也成为对 AI 的示范性输入。

5.5. 从“Announcement”到“Academic Talk”题型：协作惯例的迁移

这种示范是否真正产生了作用，在下一轮备课中得到了更直观的验证。在准备“Listen to an Academic Talk”时，新开一个对话，第一句话只有：“announcement 我已经完成备课了，我想现在开始准备托福听力中的 listen to an academic。”随后，附上官方题型介绍、Academic Talk 简介图片以及整理好的文本和题目，只增加了一句承接性要求：“请你按照之前 announcement 一样，首先帮我生成我备课所需要的各种统计文档。”Codex 生成的文档见图 4。

这一次，没有重新列出此前任务表中的 16 项要求，也没有再次解释每一类成果应该怎样命名和组织。Codex 却能够在相同项目环境中参照已经留存的 Announcement 成果，迅速建立 Academic Talk 备课目录，并生成结构分析、题型统计、逐题选项分析、答案审校、笔记策略和 README 索引等一组对应文件。第一轮结果覆盖 40 个候选材料条目和 154 道题，还主动将 16 道文本与题目明显不匹配的题目列入审校记录，为后续人工修正保留了入口。

这一现象最容易被描述为“一句话就让 AI 懂了”，但从专业角度看，真正发挥作用的并不是某句具有魔力的提示词。简短指令之所以有效，是因为此前的协作已经被外化为文件结构、命名规则、分析维度、审校方式和课件映射。所谓“像之前一样”，实际上指向了一套系统可以重新读取的工作范式。提示词变短，不代表任务变简单，而表明双方已经形成了共享的任务语法。

当然，迁移并不等于复制。Academic Talk 具有不同于 Announcement 的篇章组织方式，后续仍需围绕宏观行文逻辑、学科分类和错误文本进行多轮校准。它首先被归纳为八类宏观逻辑，随后又在教师参与下调整为四类一级结构与八类二级结构。由此可见，可复用的是分析与协作的程序，而不是将上一题型的结论原封不动地套入下一题型。Agentic AI 带来的“强大感”，正在于复杂工作流可以被保存和迁移；教师专业性的价值，则在于决定迁移什么、拒绝照搬什么。

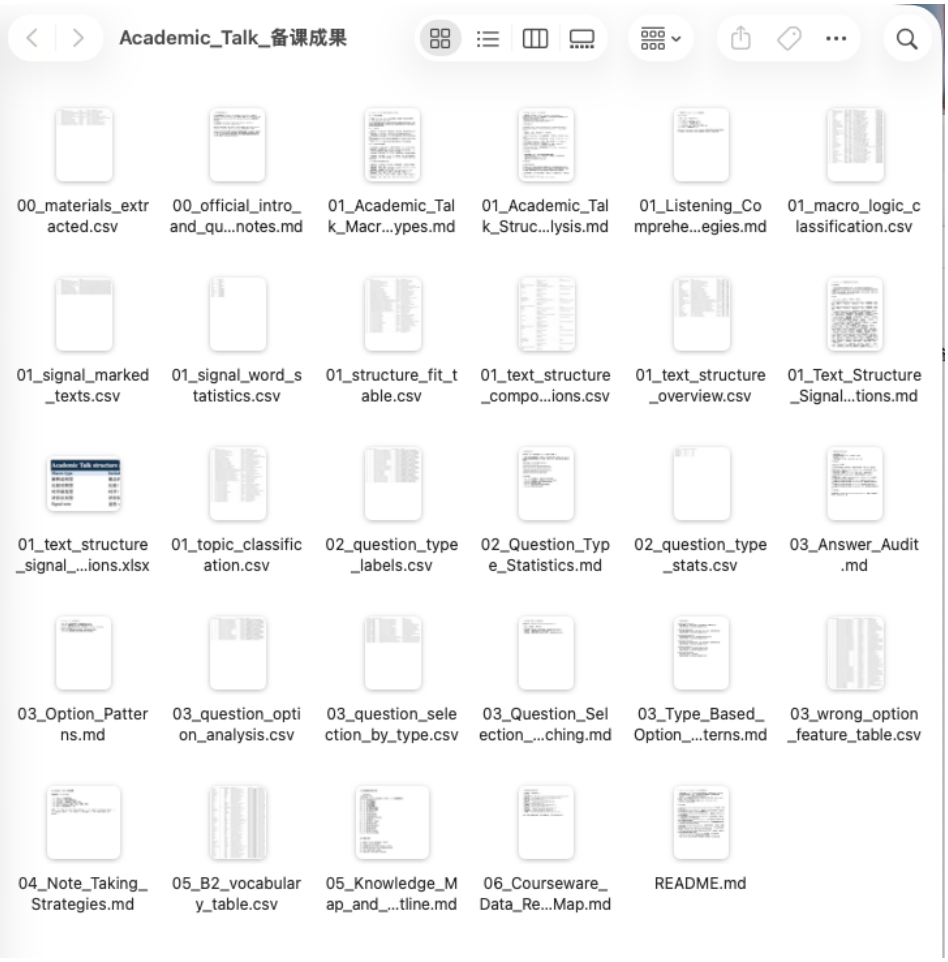


Figure 4. Transfer of collaborative routines from “Announcement” to “Academic Talk”
图 4. 从“Announcement”到“Academic Talk”的协作惯例迁移

综合三个事件，本文将本次协作概括为以下路径：界定任务 - 委托执行 - 检视结果 - 发现新问题 - 重新分析 - 专业审查 - 教学转化 - 示范回传 - 迁移复用。

其中，前几步解决“AI 能否参与”和“AI 是否促进认识”，专业审查与教学转化回答“谁对知识和课堂结果负责”，最后的示范回传与迁移复用则说明：一次成功协作能否沉淀为下一次工作的起点。

6. 边界与启示

6.1. 流畅性不能替代证据

AI 产出常常同时具备清晰标题、完整表格和精确数字，这种形式上的完成度很容易削弱使用者的警惕。事件三的启发：专业错误未必表现为胡言乱语，也可能是“看起来很像正确答案”的近似判断。教师需要要求 AI 给出分类依据，并回到原始文本、题干和选项核查，而不是只检查语言是否通顺。

6.2. 对高影响标签设置人工审查

并非所有内容都需要同等强度的复核。可以优先审查会影响大量下游成果的基础标签，例如题型、正确答案、语料是否重复、统计口径和关键术语定义。在这些节点确认无误之后，再让 AI 完成频率统计、表格重排和跨文件同步，可降低错误传播的成本。

6.3. 保存过程，而不只保存最终答案

如果只保留最后一版文件，教师很难解释某个规律如何形成，也难以定位错误从何处进入。本案例中，原始指令、AI 回复和修改记录使“5 道变 6 道”的过程得以追溯。对话与版本记录不仅是工作痕迹，也是反思性研究的证据。对于希望总结 AI 备课经验的教师而言，保存关键中间版本与纠错理由，可能比收集“高质量提示词模板”更有价值。

6.4. 保持以教师和学习者为中心

UNESCO 强调，生成式 AI 进入教育应坚持以人为本，并关注隐私、适切性与教学设计[2]。本案例未向 AI 提供学生个人数据，但教师在其他备课情境中仍应避免上传可识别的学生信息，并核查试题、教材和课件的版权边界。若期刊或机构要求披露 AI 使用，作者还应如实说明 AI 参与了哪些环节、最终内容由谁核查。

基于本案例，可提出五点实践建议：第一，先说明目标、材料、约束和完成标准，再委托 AI 处理复杂任务；第二，将 AI 产出视为可检验的备课底稿，而不是自动成立的结论；第三，让 AI 说明判断依据并保留可追溯记录；第四，对基础标签和高影响结论设置人工审查；第五，由教师完成面向具体学生的教学转化。

本文也存在明显局限。其一，案例来自一位教师和一类考试材料，所得经验不能直接推广到所有学科。其二，本文关注备课过程，尚未使用课堂观察或学习成绩检验课程效果。其三，教师同时是案例参与者和文章作者，选择哪些事件、如何解释事件不可避免地带有主体性。后续研究可在不同教师之间比较人机协作路径，并结合课堂实施数据考察 AI 支持的备课是否真正改善学习体验。

7. 结论：从使用工具到管理协作

本次备课开始于一个现实而有限的愿望：借助 AI 更快地处理大量材料。项目结束后，笔者对 Agentic AI 的认识发生了变化。它确实可以承担文件读取、统计、分类和关联更新，也能够通过整理大量语料帮助教师看到原先不易发现的现象。但它既不是可靠答案的自动来源，也不会自行判断什么最值得教。

三个事件呈现了一些变化：第一，复杂任务可以整体委托，教师由逐项处理者转向工作流设计者；第二，AI 结果可以触发新问题，备课由材料加工转向循环式知识建构；第三，表面合理的错误可能沿着自动化链条扩散，教师必须成为知识质量的最终审定者。课件回传和“Academic Talk”的后续实践又补充了另一层认识：教师不仅在 AI 出错时负责纠偏，也可以通过展示自己的最终作品，将“我怎样把数据变成教学”教给 AI，并使一次协作沉淀为下一次工作的参照。

因此，真正成熟的人机协作，不是让教师将专业判断完全交给 AI，也不是每次都靠写更长的提示词从零开始，而是借助 AI 扩大观察与行动的范围，同时将教师的判断标准、教学方法沉淀为可复用的工作语境。新任务中的简短指令之所以能够撬动大量工作，背后依靠的是此前共同建立、持续修订并被明确记录下来的协作惯例。

Agentic AI 可以延伸教师的视野与双手，却不能替教师决定什么值得相信、什么适合学生、什么应当进入课堂。或许这正是此次备课带给我们的主要认识：AI 参与得越深，教师的专业性越不应隐身；相反，它需要在提问、纠错与教学转化中变得更加清晰可见并且可以被检验。

参考文献

[1] Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., *et al*. (2023) ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education. *Learning and Individual Differences*, **103**, Article 102274. https://doi.org/10.1016/j.lindif.2023.102274

[2] Miao, F. and Holmes, W. (2023) Guidance for Generative AI in Education and Research. UNESCO. https://www.unesco.org/en/articles/guidance-generative-ai-education-and-research

[3] OpenAI (2026) Using Codex with Your ChatGPT Plan. https://help.openai.com/en/articles/11369540-using-codex-with-your-chatgpt-plan

[4] Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., *et al*. (2024) A Survey on Large Language Model Based Autonomous Agents. *Frontiers of Computer Science*, **18**, Article ID: 186345. https://doi.org/10.1007/s11704-024-40231-1

[5] Flavin, E., Hwang, S. and Morales, M. (2025) "Let's Ask the Robot!": Epistemic Stance between Teacher Candidates toward AI in Mathematics Lesson Planning. *Journal of Teacher Education*, **76**, 262-279. https://doi.org/10.1177/00224871251325079

[6] Velander, J. (2026) Beyond Operational Skills: Teachers' AI Knowledge and Interactions with Generative AI in Lesson Planning. *Computers and Education Open*, **10**, Article 100371. https://doi.org/10.1016/j.caeo.2026.100371

[7] Schön, D.A. (1983) The Reflective Practitioner. Basic Books.

[8] Kalenda, P.J., Rath, L., Abugasea Heidt, M. and Wright, A. (2025) Pre-Service Teacher Perceptions of ChatGPT for Lesson Plan Generation. *Journal of Educational Technology Systems*, **53**, 219-241. https://doi.org/10.1177/00472395241301388

[9] Krushinskaia, K., Elen, J. and Raes, A. (2026) Pre-Service Teachers' Agency during Their Interactions with Generative AI While Designing for Learning—A Process View on Intelligent-TPACK. *Computers and Education Open*, **10**, Article 100325. https://doi.org/10.1016/j.caeo.2025.100325