国际汉语教学学报（2026）
第 07 卷 · 第 03 期 5-19 https://doi.org/10.46451/ijclt.260402

人机协同下的课堂话语分析: 新加坡小学华文 AI 试点研究

胡月宝 *
贺禹婷
新加坡南洋理工大学国立教育学院，新加坡

摘要
本文针对人工智能教育中教师端工具不足的问题，提出一套智能课堂话语分析系统（AI-based Classroom Discourse Analysis System）。研究利用新加坡 12 名教师 24 节课逐字稿，结合自动语音转写与 OpenAI 的 GPT-4 模型话语功能分类，对比人工与 AI 在处理效率与分类表现上的差异。结果显示，智能课堂话语分析系统在转写与初步分类阶段将整体耗时压缩至纯人力流程的一小部分，但在话语功能分类上与人工仅呈中等偏低的一致水平，并出现稳定的“双峰”偏差：对课堂管理、社交互动等形式特征突出的话语高估，对引入、知识讲解与总结等依赖语篇结构的话语低估。研究据此将 AI 定位为预标注与趋势分析工具，由教师针对高风险类别进行复核，并指出未来需在模型中引入教学阶段与语篇结构等特征，以提升课堂话语分析的有效性。

关键词
智能课堂话语分析，人工智能教育，人机协同，教师专业发展

1 引言

近年来，人工智能在教育中的应用研究速增，但系统性综述普遍指出，人工智能教育（Artificial Intelligence in Education, AIE）研究长期以学习者端为主（如智能辅导、自动评估等），真正面向教师端的支持工具在理论与实证上都明显不足（Zawacki-Richter et al., 2019; Hwang & Chen, 2020）。

目前，虽有少数研究开始讨论“人 –AI 协同”在课堂中的潜在价值，但大多仍停留于人与 AI 混合适应机制（human–AI hybrid adaptivity）的概念层面，而非可在教师日常教学中直接部署的系统（Holstein et al., 2020）。整体来看，现有 AI 在实时识别课堂互动、支援教师进行复杂专业推理，以及提供可解释的教学诊断方面仍存在明显不足（Luckin, 2018; Holmes et al., 2022），开发能够直接面向教师、并在真实课堂情境中支持专业判断的 AI 工具便成为当前的缺口。

在笔者长期主持的课堂话语分析项目中，逐字转写与人工编码需耗费大量时间、人力与培训成本，一节课往往需要数小时才能完成分析报告，不仅难以及时回馈教师，也大幅限制了课堂研究的规模。此经验凸显了借 AI 提升效率，让课堂话语分析真正服务教师专业发展的迫切性。

基于此，本研究初步开发“智能课堂话语分析系统”（AI-based Classroom Discourse Analysis System, AI-CDAS，简称智课语析系统），旨在自动识别教师课堂话语的功能类别，并

* 通讯作者。联系电邮：guatpoh.aw@nie.edu.sg

量化比较 AI 与人工分析在效率与准确性上的差异。预期贡献包括二点：其一，提出一个融合课堂话语功能理论与智能自动化分析的教师端人工智能教育模型（Artificial Intelligence in Education, AIED）；其二，检视 AI 分类一致性与系统性偏差的分析框架，用以揭示 AI 在不同话语类型上的优势与盲区。终极目标是使课堂话语分析从难以规模化的人工工作迈向可普及的智能分析。

综上，本研究试图回应教师端人工智能教育工具发展的结构性缺口，通过智课语析系统的开发与实证检验，探索一条兼顾效率与专业判断的人机协同课堂研究路径。

2 文献综述

2.1 AI 辅助定性研究：人机协同的分析范式

在定性研究中，人工智能正逐渐从工具性角色转向与研究者协作的人机互补模式（human–AI complementarity）。AI 能在大规模与噪声较高的文本中执行初步整理、自动化编码与结构化分析，有效降低人工处理的时间成本，并提升跨研究者的编码一致性（Kasneci et al., 2023）。然而，定性研究中的情境诠释、文化判断与理论建构仍需由研究者主导，这也是诠释传统长期坚持的原则（Creswell & Creswell, 2018）。

在人机协同的框架下，AI 被视为一种“增强性认知工具”（augmenting cognitive tool），能扩展分析规模与速度，但不取代研究者对意义的最终判断（Jonassen et al., 1998）。因此，“机辅人主”成为当前较为稳健的分析范式：AI 负责初筛、预标注与趋势呈现，人类研究者则负责语篇理解与理论解释。本研究据此将自主开发的智课语析系统视为一种辅助性分析工具，并检视其与人工分类的偏差结构与一致性，以探讨更可行的人机协同模式。

2.2 课堂话语分析：理论基础与研究进展

课堂话语研究强调教师语言在知识建构、互动组织与课堂调控中的功能性作用。系统功能语言学（Systemic Functional Linguistics, SFL）将语言视为意义建构的社会符号系统，为分析教师如何通过不同话语功能实现教学目的提供基础（Halliday, 1978）。Sinclair 与 Coulthard（1975）提出的 IRF（Initiation–Response–Feedback）模型进一步揭示教学互动的序列结构，是辨识课堂话语功能与语篇推进的重要框架。

在互动研究方面，Walsh（2011）的课堂互动能力（Classroom Interactional Competence, CIC）强调教师如何调整话语以优化学习机会；会话分析则从转接、轮序与互为主体性的角度揭示课堂互动如何被参与者共同构建（Sacks, Schegloff & Jefferson，1974）。这些理论共同指出：课堂话语分析不仅是句层判断，更涉及语篇结构、教学意图与互动序列的解释。

在方法工具上，Talk Moves 与 Classroom Discourse Observation Protocol（CDOP）是目前较为成熟的结构化课堂话语编码体系。Talk Moves Dataset 通过大规模人工标注，将教师与学生的 discursive moves 系统化，为自动化课堂分析提供可量化语料基础（Suresh et al., 2022）。CDOP（Kranzfelder et al., 2020）则以 STEM 课堂为对象，发展出对解释、提问、支架化与反馈等教学行为的规范化编码，显著提升研究的可比较性。然而，这些框架仍严重依赖人工标注，难以扩展至大规模课堂数据。

自然语言处理与大型语言模型的发展为课堂话语研究带来自动化转向的可能。已有研究展示了机器学习在课堂互动 接续回应 侦测中的潜力，说明模型能够捕捉互动质量的变化；大语言模型在语篇分析、功能分类与课堂规模化研究中具有潜在突破性贡献，但其语篇理解能力尚不稳固，需要实证检验（Kasneci et al., 2023）。

综合来看，课堂话语研究已具备成熟的理论框架（SFL, IRF, CIC），并有结构化标注体系支撑（TalkMoves, CDOP）; 然而，当前自动化研究虽具潜力，却缺乏对教师话语功能的系统验证。本研究以此为基础，研发智能课堂话语分析系统，找出 AI 的优势与局限，以朝向回应上述缺口的道路迈进。

3 研究设计

3.1 初始研究现状与研究问题

本研究的前期基础（2016-2020）来自新加坡南洋理工大学国立教育学院一项小学华文教师专业发展研究计划，其核心研究工具之一为课堂话语分析。研究团队采用双人听辨方式进行人工转写，由两名研究助理分别对课堂录音进行逐句辨识与核对，再依据话语功能框架进行人工编码。一小时课堂通常需约 5 小时完成逐字转写，另需 2-3 小时完成话语标注与汇总。此工作量与国际对自然情境口语转写的估计一致：口语逐字稿人工转写常需录音时长的 4-6 倍时间（MacLean et al., 2004; Halcomb & Davidson, 2006）。由此可见，传统课堂话语分析在时间成本与即时反馈之间存在结构性矛盾，也为本研究引入人工智能工具提供了方法论动机。

基于此限制，本研究于 2024 年启动课堂语料的二次分析，开发并测试由 GPT-4 驱动的智能课堂话语分析系统，并利用 24 堂课堂的一前一后测语料检验其效能。本研究旨在系统比较人工与 AI 的表现，形成可用于课堂研究与教师专业发展的实证依据。具体研究问题如下：

RQ1: 在分析效率方面，智课语析系统相较人工是否能显著提升课堂话语分析的速度与处理量？

RQ2: 在分析准确性方面，智课语析系统的分类结果与人工分析之间的一致性水平如何？其偏差结构是否具有稳定性与可预测性？

RQ3: 基于智课语析系统在效率与一致性方面的表现，本研究对教育研究方法与教师专业发展提出哪些理论与实践启示？

3.2 研究对象与教学情境说明

本研究所分析的课堂语料均来自新加坡小学华文课堂，研究对象为小学二至五年级学生所参与的常规教学活动。课堂内容涵盖基于中央教材的课文阅读与词语讲解，以及校本开发的口语、阅读理解与写作教学，属于新加坡小学华文课程中的核心教学情境。

从教学难度来看，合作学校为新加坡本地社区内的邻里学校，学生整体华文程度处于中等偏下水平。本研究涉及全校 12 名教师，未对个别班级的学生程度或教师教学风格进行人为控制，以保留学校日常教学中自然存在的差异。

尽管研究涵盖小学低至中高年级的课堂语料，但分析重点并不在于比较不同年级之间的话语差异，而在于考察课堂话语功能在真实教学情境中的整体分布特征及其被 AI 识别的稳定性。此一教学情境中，课堂话语常出现较多引导性说明、重复确认、追问与即时反馈等功能性话语，

为分析话语功能类型及其互动特征提供了合适语料，也有助于检验 AI 在较高教学支持密度情境下的分析表现。

此外，本研究不以学生语言能力或教师教学能力作为自变量进行控制或比较，而是将不同年级、班级与教师所形成的课堂差异视为真实教学情境中的自然变异来源。因此，研究结果主要反映 AI 在一般课堂话语分析中的应用潜力。

3.3 课堂话语功能分类框架说明

本研究采用课堂话语分析理论中对教学话语功能的相关界定，关注教师话语在课堂互动中所承担的不同教学功能，而非仅从语言形式或句法结构进行分类。话语功能的判定以教学意图与互动功能为核心，强调话语在具体课堂语境中的实际作用（Sinclair & Coulthard, 1975; Walsh, 2011）。

在操作层面，本研究将课堂话语功能划分为七个主要类别，涵盖课堂导入、示范、教学指令与组织、语言知识教学、总结／收束、课堂管理与社会性互动功能类型。分类方式综合借鉴了课堂话语研究中对教学推进型话语与互动支持型话语的区分思路（Flanders, 1970; Hall & Walsh, 2002），并结合新加坡小学华文课堂的教学实际进行了操作性调整。编码的话语单位以“一句话”为基本编码单位；当单一话语同时具备多重功能时，编码以其主要教学意图为准。在分析对象上，所有话语功能均以教师话语为统计单位，学生话语不纳入本研究的话语功能分类范围。编码过程中，研究者首先依据本附录所列操作性定义进行人工判定，再将结果与 AI 输出进行对照分析；当话语功能判定存在疑义时，研究者回到具体课堂语境，并依据主要教学意图进行最终裁定，以确保编码的一致性与可解释性。

为确保人工编码与 AI 分析在同一理论框架下进行，研究在编码前对各类话语功能的操作性定义与判定标准进行统一整理，并作为后续人工与自动分类的共同参照。鉴于话语功能分类的具体定义、判定要点与课堂话语示例较为详细，若分散呈现易影响方法部分的整体连贯性，相关完整编码示例已统一整理并置于附录中。

3.4 智课语析系统设计概述

本研究开发“智能课堂话语分析系统”（AI-based Classroom Dialogue Analysis System，AI-CDAS），利用大型语言模型对课堂话语数据进行自动化处理，并与人工编码结果进行系统比较，以回应 RQ1–RQ3 所关注的效率与一致性问题。系统整体遵循“教育学框架主导、技术模块化实现”的设计原则，使自动化分析流程与既有课堂话语功能理论保持一致。

在数据处理流程上，系统首先进行文本预处理。该阶段利用 Python 对课堂转写文本进行清洗、分句与角色标注，其中“角色”主要指课堂发言主体（教师或学生）。文本依据中文标点符号（如“。！？”）进行规则化分句处理，以确保每个分析单位具有相对完整的语义结构。预处理后的数据被整理为结构化格式，并通过 Pandas 等工具进行数据整理与初步统计，为后续分析提供统一的数据基础。统计分析部分则借助 R（tidyverse）与 SPSS 完成，包括描述性统计以及正态性检验与非参数检验，以比较人工与 AI 在分类结果上的差异。

在自动分类阶段，系统以 GPT-4o 为核心语言模型，通过预设的“话语功能分类提示框架”对教师课堂话语进行自动标注。提示结构通常包含由若干连续课堂话轮构成的局部语境信息以及待分析的教师话语句子，并要求模型在预设的话语功能类别集合中输出对应的分类标签。系统为每条教师话语构建局部上下文窗口，并在统一提示模板下调用模型生成分类结果。为提高

输出稳定性，系统设置标签合法性校验与重试机制，并在后处理阶段对语义等价但表述不一致的标签进行统一映射。AI 分类所使用的提示结构与上下文控制方式已在方法部分作进一步说明。

在结果输出阶段，系统自动生成结构化 Excel 数据表，并输出类别统计与可视化结果，以支持研究者进行进一步分析。通过这一流程，研究能够系统比较 AI 与人工编码在课堂话语功能识别方面的效率与一致性表现。

在系统性能评估方面，本研究使用 24 节真实课堂（每节约 60 分钟）作为测试数据，并通过两阶段分析检验 AI 在不同教师风格与课堂结构下的稳定性。一致性指标采用 Cohen's κ 与 Krippendorff's α 进行评估，以量化 AI 分类结果与人工编码之间的符合程度。考虑到自动转写仍可能受到学生口语表达不规范、语音重叠等课堂互动因素的影响，本研究在数据分析过程中保留人工校对与复核环节，以确保分析结果的可靠性。

3.5 二阶段渐进式研究设计

本研究采用两阶段渐进式研究设计，旨在以“小规模试点—扩大验证”的方式检验智课语析系统的表现。此设计使系统能够在真实课堂条件下逐步校准，并确保分析结果具有稳定性与可重复性。

第一阶段（n = 4，8 节课）为试点测试阶段，主要用于初步评估 AI 在课堂话语功能自动分类中的表现，并识别可能出现的偏差模式。在此基础上，研究对提示结构与分类规则进行必要调整，以优化系统与人工编码框架之间的匹配程度。因此，该阶段的重点在于发现潜在问题并进行方法调适，而非追求统计意义上的外部效度。

第二阶段（n = 8，16 节课）在第一阶段基础上扩大样本规模，涵盖不同教师风格、课堂节奏与话语密度，以检验系统在多样化课堂情境中的稳定性。该阶段主要关注两个方面：一是检验第一阶段识别的偏差模式是否在更具差异性的样本中重复出现；二是评估系统在规模化语料条件下是否能够维持稳定的一致性表现。扩大样本的目的并不在于提升统计显著性，而在于确认相关偏差是否具有跨课堂情境的结构性特征。

通过比较两个阶段在分类趋势、偏差方向与一致性指标（如 Cohen's κ 与 Krippendorff's α）上的表现，本研究进一步区分两类可能的误差来源：一类为偶发性误差（random noise），仅在个别课堂情境中出现，缺乏稳定重复性；另一类为结构性偏差（systematic bias），在不同教师、课型或语篇条件下持续出现，可被视为模型机制所产生的稳定模式。

这种递进式研究设计使本研究能够在真实课堂语境中逐步识别系统的优势与限制，并为后续的系统优化与教学应用建议（见第 4.4 节）提供依据。

3.6 GPT-4 调用方式与人 – 机协同分析流程

在 AI-based Classroom Dialogue Analysis System（AI-CDAS）的 AI 智能分类模块中，本研究采用 OpenAI GPT-4o 作为核心语言理解与生成模型，用以辅助完成教师课堂话语在具体语境中的功能推断。在具体调用方式上，系统通过 OpenAI API 以非交互式、批处理方式调用模型，所有分析均在预先设定的固定提示框架下完成，以确保不同话语单元在分析条件上的一致性。

在输入语境控制方面，本研究针对每一条教师话轮构建受限的局部上下文窗口。所谓“话轮”（turn），指同一说话者在未被他人打断情况下完成的一次连续发言。系统以话轮为基本单位，针对目标教师话轮向前回溯最多两个连续话轮，并结合该教师话轮本身，构建最多五个连续话

轮作为输入语境窗口。该窗口可能包含教师与学生的交替发言，并以“角色：话语内容”的形式拼接为模型输入文本。若目标话轮位于课堂起始位置，则按实际可获得的话轮数量构建窗口。

此种局部语境控制策略在保留必要课堂互动信息的同时，避免长程语境对模型判断产生潜在干扰，从而增强分析对象的独立性与研究过程的可重复性。在实际数据规模下，该窗口长度远低于模型可处理的上下文上限，因而不会对模型性能构成限制。

在分析单位的设定上，系统进一步依据中文标点符号（。！？）对教师话轮进行分句处理，并将每个语义完整的句子作为最小分析单位分别输入模型进行分类。通过这一处理方式，研究在保证话语语义完整性的同时，确保了分析粒度在不同样本之间的一致性。

GPT-4o 的调用以“话语功能分类提示框架”为核心。提示内容明确限定模型的分析任务，即依据研究既定的话语功能分类体系，对给定课堂话语提出对应的功能类别判断。提示中不包含人工编码结果或统计信息，亦未向模型提供示例或进行任何形式的参数微调，且所有样本均使用完全相同的提示模板；模型仅作为通用语言模型，在结构化 zero-shot 提示约束下参与分析。模型输出被限定为预先设定的七类教师话语功能之一（课堂引入、示范、指示／组织活动、知识讲解、总结、课堂管理、社交互动）。为降低输出不稳定性，系统在模型返回结果后进行合法性校验；当输出不符合预期类别集合时，在相同提示条件下自动触发重新调用流程（最多两次），并在后处理阶段对语义等价但表述不一致的标签进行规范化映射。

在整体分析流程中，GPT-4o 的输出结果仅作为辅助参考，并不直接构成研究结论。系统首先生成 AI 话语功能分类建议，随后由研究者将 AI 输出结果与人工编码结果进行逐项比对，并通过一致性指标（Cohen’s κ）量化二者之间的符合程度。当 AI 分类结果与人工判断不一致时，研究者回到具体课堂语境与话语功能理论定义，对相关话语进行人工复核，最终编码结果始终以人工判定为准。

因此，本研究采用的是一种“人–机协同”（human-in-the-loop）的话语分析模式：GPT-4o 主要用于提升话语初步分类阶段的处理效率与系统化程度，而话语功能的理论解释权与最终判定权仍由研究者掌握。该设计在缓解传统课堂话语分析高时间成本问题的同时，也避免将教学话语分析简化为完全自动化的决策过程，从而在效率与研究解释力之间取得平衡。

4 研究结果

4.1 人工转写与 AI 转写工具效率比较

本研究采用科大讯飞与通义听悟进行自动语音转写（automatic speech recognition, ASR）。根据整体样本推估，人工转写 1 小时课堂平均需约 4.7 小时；相较之下，AI 的初转写速度约为 20-45 秒 / 分钟音频（本研究实测），整体效率提升超过 80%。基于此，研究初期采用“ASR 初转写 + 人工校对”的混合流程（MacWhinney, 2000），以兼顾效率与语义准确性。

进一步比较人工与 AI 在本研究课堂实录中的整体耗时差异，发现两者在转写及后续话语分析阶段的效率差距极大。以 45 分钟课程为例，人工转写平均需 280 分钟，人工话语分析平均需 180 分钟，总耗时约 460 分钟。相比之下，AI 初转写阶段平均用时 15 分钟，配合人工校对的分析阶段约 5 分钟，总耗时仅 20 分钟，整体效率提升超过 20 倍（见表 1）。需要说明的是，前述“1 小时 ≈ 4.7 小时人力转写”为本研究整体资料的平均估算；而表 1 中的 45 分钟数据则基于若干节课堂的个别观测，因教师语速、课堂互动密度与环境噪音差异而有所浮动，但整体量级一致，充分反映两种方式在效率上的显著差距。

表 1. 人工与 AI 转写时间对比

工具类型	一堂课的转写时间（分钟）	一堂课的分析时间（分钟）	总耗时 (分钟)
纯人力分析	280	180	460
AI 初转 + 人力校核阶段	90	60	150
智课语析系统（稳定以后）	15	5	20

总体而言，AI 在转写与初步话语分析阶段展现出远高于人工的效率优势，使其特别适用于需要即时或近即时反馈的课堂场景。人工分析一小时课堂往往需耗时四至六小时，而 AI 则可在数秒至数分钟内完成初步处理。即便后续仍需人工校对，其总体工作量仍大幅减少，使教师得以在课后迅速获得课堂语言使用的初步反馈，有助于专业发展、教学反思与行动研究。因此，AI 在本研究中不仅承担了“前置处理”的角色，也显著减轻了人工在初步编码阶段的负担，为后续更精细的语篇分析奠定基础。

4.2 人工与 AI 在课堂话语功能分类上的比较

4.2.1 统计差异概览

表 2 显示，阶段一共有 6 个话语功能类别在人工与 AI 之间达到显著差异（$p < 0.05$），而阶段二的显著类别数量增加至 11 个，范围更广，涵盖示范、语言知识讲解、总结、课堂管理与社交互动等多个子类。值得注意的是，其中有 5 类在两个阶段均呈现显著差异（包括“语言知识讲解 R2”“总结 R1”“课堂管理 R1”和“社交互动 R1/R2”），显示这些话语功能类别中 AI 与人工之间的差异具有跨语料重复性。整体而言，显著差异主要集中在“教学类”（如语言知识讲解、总结）和“管理 / 互动类”（如课堂管理、社交互动），呈现出稳定的偏差方向：教学类多为低估，管理与互动类多为高估。其余类别虽偏差方向一致，但未达到统计显著水准。此部分统计检验旨在确认差异的存在，并为后续的偏差结构分析提供基础，具体偏移的幅度、方向与跨阶段一致性将在表 3 中进一步讨论。

4.2.2 偏差结构与跨阶段稳定性（Bias Structure & Replicability）

在统计检验确认一些话语类别存在显著差异的基础上（见表 2），本研究进一步结合偏差比例与偏差强度等级（见表 3）可以更清楚看出人工与 AI 在课堂话语分析中的结构性偏移模式。偏差比例揭示 AI 相对人工的高估或低估幅度，而偏差强度等级则有助于辨识该偏差在教育解释上的重要性。

综合阶段一与二的偏差结果可发现，AI 的分类偏差方向在 14 个类别中达到 100% 的一致性（低估 / 高估方向完全一致），显示其差异并非样本偶然，而是模型层面的系统性特征。具体而言，在依赖教学语篇结构、教师意图与阶段性组织的类别中——如“引入 R1/R2”“语言知识讲解 R1/R2”与“总结 R1/R2”——AI 在两阶段均呈现大幅低估，其中总结类话语的偏差比例达 −80% 至 −91%，属于“显著”或“严重低估”等级。这反映 AI 缺乏识别教学活动阶段转换与收束功能的能力，也佐证其在语篇层级判断上的弱势（Sinclair & Coulthard, 1975; Walsh，2011）。

与此相对，AI 在“课堂管理”“社交互动”“示范”等表层特征显著的类别中则呈现持续而大幅的高估，偏差比例常超过 100%，甚至在社交互动类别中达到 +394% 至 +1485% 的严

重高估。这显示 AI 的分类机制对情绪词、指令词、重复词、填充语等显性语言线索高度敏感，倾向将所有具有互动性或指令性的语句归入相关类别，而缺乏处理多功能混合语句的能力（Mercer，2004; Kranzfelder et al., 2019）。这一偏差亦在两个阶段完全复现，进一步验证 AI 的分类是基于词层与句层特征，而非语篇结构。

表 2. 人工与 AI 在课堂话语功能分类上的比较

阶段一（n=4）				
指标	人工组 平均值 ± 标准差 / 中位数（P25，P75）	AI 组 平均值 ± 标准差 / 中位数（P25，P75）	t / Z 值	p 值
引入 R1	9.27 ± 4.14	3.68 ± 0.80	2.652	0.038*
引入 R2	5.30 ± 2.70	4.69 ± 3.03	0.301	0.774
示范 R1	9.77 ± 8.61	14.63 ± 8.38	−0.808	0.450
示范 R2	4.10（0.5，7.2）	5.535（5.4，6.9）	−0.289	0.773
组织 R1	46.83 ± 14.75	29.42 ± 11.42	1.866	0.111
组织 R2	32.30 ± 12.60	21.46 ± 7.30	1.489	0.187
语言知识讲解 R1	18.25 ± 10.32	14.04 ± 7.49	0.661	0.533
语言知识讲解 R2	42.27 ± 21.00	11.53 ± 8.40	2.719	0.035*
总结 R1	6.97 ± 2.60	0.94 ± 0.95	4.361	0.014*
总结 R2	2.13 ± 2.10	0.19 ± 0.16	1.842	0.162
课堂管理 R1	4.75（0.4，14.9）	25.08（11.5，28.9）	−2.021	0.043*
课堂管理 R2	12.13 ± 13.50	27.64 ± 13.78	−1.608	0.159
社交互动 R1	0.35（0.1，2.5）	15.36（12.8，18.2）	−2.309	0.021*
社交互动 R2	2.00 ± 2.64	28.55 ± 8.19	−6.173	0.005**
阶段二（n＝8）				
指标	人工组 平均值 ± 标准差 / 中位数（P25，P75）	AI 组 平均值 ± 标准差 / 中位数（P25，P75）	t / Z 值	p 值
引入 R1	2.80（1.1，6.5）	1.58（0.8，6.5）	−0.053	0.958
引入 R2	7.50（3.7，16.8）	1.76（0.2，2.9）	−2.521	0.012*
示范 R1	1.65（0.0，7.0）	3.43（2.0，6.1）	2.007	0.010**
示范 R2	6.00（0.0，3.5）	7.12（5.2，8.7）	−2.530	0.010**
组织 R1	35.95 ± 12.22	29.26 ± 14.71	1.008	0.331
组织 R2	29.24 ± 18.22	28.11 ± 9.95	0.154	0.881
语言知识讲解 R1	44.89 ± 29.47	17.65 ± 11.39	3.008	0.008**
语言知识讲解 R2	45.89 ± 29.41	18.21 ± 10.29	2.997	0.014*
总结 R1	3.60（2.8，6.1）	1.06（0.2，1.9）	−4.363	0.020*
总结 R2	2.90（1.4，8.5）	0.41（0.1，1.3）	−2.317	0.020*
课堂管理 R1	4.90（1.5，10.2）	25.00（15.2，28.9）	−2.009	0.009**
课堂管理 R2	1.35（0.4，3.1）	26.05（12.5，28.5）	−2.403	0.030**
社交互动 R1	1.65（0.3，7.17）	19.49（15.3，28.9）	−2.943	0.003**
社交互动 R2	0.60（0.1，1.5）	20.33（14.6，29.8）	−3.361	0.001**

注：本表合并呈现阶段一（n = 4，8 节课）与阶段二（n = 8，16 节课）的分析结果。R1 与 R2 分别表示第一轮与第二轮分析结果。正态分布数据采用独立样本 *t* 检验，非正态分布数据采用 Mann–Whitney *U* 检验（报告 *Z* 值）。数据以均值 ± 标准差或中位数（P25，P75）表示。* $p < .05$，** $p < .01$。

总体而言，AI 的表现呈现出一种典型的“双峰结构”：一方面，在词汇线索清晰、形式特征强的类别中，AI 的分类高度敏感且跨语料一致；另一方面，在依赖教学语篇理解、阶段性逻辑与教师意图推断的类别中，AI 持续显著低估。此“双峰结构”反映了 AI 在课堂话语分析上的优势（高效率、高一致性、易用于趋势分析）与明显限制（缺乏语篇理解能力、无法处理教学阶段、难以识别多功能语句）。

表 3. 人工与 AI 在课堂话语功能分类上的偏差整合分析

话语类别	阶段一人工平均	阶段一 AI 平均	阶段一偏差比例	阶段一偏差强度	阶段二人工平均	阶段二 AI 平均	阶段二偏差比例	阶段二偏差强度	偏差方向一致与否
引入 R1	9.27	3.68	−60.3%	显著低估	3.80	3.23	−15.0%	轻度低估	一致低估
引入 R2	5.30	4.69	−11.5%	轻度低估	10.71	3.19	−70.2%	显著低估	一致低估
示范 R1	9.78	14.63	+49.6%	中度高估	3.78	6.83	+80.7%	显著高估	一致高估
示范 R2	4.10	5.54	+35.1%	中度高估	2.79	6.11	+119.0%	严重高估	一致高估
组织 R1	46.83	29.42	−37.2%	中度低估	35.95	29.26	−18.6%	轻度低估	一致低估
组织 R2	32.30	21.46	−33.6%	中度低估	28.56	27.11	−5.1%	轻度低估	一致低估
语言知识讲解 R1	18.25	14.05	−23.0%	中度低估	44.89	22.90	−49.0%	中度低估	一致低估
语言知识讲解 R2	42.27	11.53	−72.7%	显著低估	46.11	18.22	−60.5%	显著低估	一致低估
总结 R1	6.97	0.94	−86.5%	显著低估	4.45	0.86	−80.7%	显著低估	一致低估
总结 R2	2.13	0.19	−91.1%	显著低估	5.21	0.75	−85.6%	显著低估	一致低估
课堂管理 R1	7.90	21.85	+176.6%	严重高估	3.73	20.57	+451.5%	严重高估	一致高估
课堂管理 R2	12.13	27.64	+127.9%	严重高估	3.33	21.07	+532.7%	严重高估	一致高估
社交互动 R1	0.98	15.45	+1476.5%	严重高估	4.14	20.47	+394.4%	严重高估	一致高估
社交互动 R2	2.00	28.55	+1327.5%	严重高估	1.60	21.87	+1266.9%	严重高估	一致高估

本研究依据 AI 相对人工的偏差比例（%）划分偏差强度：$|\Delta| < 20\%$ 为轻度、20–50% 为中度、50–100% 为显著、≥100% 为严重，此方法常用于课堂互动量化与语篇功能比对研究，以呈现分类偏移的方向性与幅度（Kranzfelder et al., 2019; Walsh, 2011）。

4.2.3 一致性水平检验（Cohen’s Kappa 与一致率分析）

为检验两阶段中人工与 AI 课堂话语分析结果的一致性，本研究分别对阶段一与阶段二进行一致性检验，结果见表 4 与表 5。

从表 4，表 5 一致性检查结果来看，AI 在不同话语类别上的表现呈现出清晰的结构性差异，并在跨阶段中展现出一定的稳定性与可预测性。首先，在“课堂管理”类别中，AI 的一致率在两个阶段均维持在高水平（阶段一 79.5%，阶段二 76.0%），显示其对显性指令、程序性语

言以及管理性语气具有高度敏感度。类似地，“示范”类别在阶段一表现良好（77.4%），但在阶段二显著下降（41.8%），可能反映教师示范方式、语言风格或课堂节奏的差异，使 AI 对此类操作性话语的侦测能力受到影响。

“组织活动”类别的变化最为显著，从阶段一的 51.1% 降至阶段二的 11.3%。这表明阶段二语料中的组织类语句呈现更高的语用混合度（例如同一句同时包含提醒、解释与任务提示），导致 AI 倚赖表层词汇的策略难以应对更复杂的功能重叠现象。“社交互动”类别在阶段一达到 100%，但阶段二下降至 57.1%，主要原因是阶段一样本量极小，使一致率被动拉高；当样本量扩大后，AI 的真实表现趋于中等水平。

属于教学核心功能的类别则呈现跨阶段一致的低表现，包括“知识讲解”（阶段一 29.8%，阶段二 19.4%）与“引入”（13.3% 与 5%）。这两类话语通常依赖语篇逻辑、教学目的与概念推进，而非表层语言特征，AI 因缺乏对教学结构的识别能力，因而在两阶段均无法有效分类。同样地，“总结”类别在两个阶段的一致率均为 0%，显示 AI 完全无法辨识具有“收束功能”的语句结构，而这类语句往往需要对课程进程、教师意图以及语篇阶段性进行整体判断，是 AI 在当前模型条件下的最薄弱类别。

为了综合检验跨阶段一致性，本研究将两阶段数据整合后再次计算 Kappa（κ =0.210），整体处于一般一致（fair）水平，显示尽管 AI 与人工之间的一致性有限，但其偏差结构在不同语料中具有一定稳定性。整体来看，AI 的跨阶段一致性呈现出明确的结构性特征：

(1) 对显性形式特征敏感，例如管理、部分示范，能保持较高一致性；

(2) 在语篇功能混合类别上容易受语料变化影响，例如组织活动、社交互动；

(3) 无法持续识别教学语篇关键结构，尤其是引入、讲解、总结。

此模式与前述偏差结构分析完全一致，进一步验证当前 AI 辅助课堂话语分析工具在语篇理解、教学阶段识别与意图推断方面仍存在根本性限制，但在处理具有明显表层特征的语言类别时可展现相对稳定的效能。

表 4. 阶段一（n=311）一致性检查表（Kappa $\kappa_1 = 0.252$；一般一致）

话语类别	人工样本数	AI 一致样本数（TP）	一致率 (%) = TP/人工	说明
课堂管理	39	31	79.5%	高一致，表层语言特征明显
示范	31	24	77.4%	高一致，动作 / 程序性语句易识别
指示 / 组织活动	133	68	51.1%	中等一致，仍有大量误判
社交互动	8	8	100%	高一致，但样本极少
知识讲解	57	17	29.8%	低一致，AI 对此话语功能的识别能力明显不足
引入	30	4	13.3%	极低一致，AI 缺乏教学语篇意识
总结	12	0	0%	严重低估，模型无法识别“收束性语句”
总体	311	152	48.9%	中低一致，显示在仅依赖局部语言特征、未纳入课堂教学流程与宏观语篇结构信息的情况下，AI 的整体分类表现仍存在明显不足。

表 5. 阶段二一致性检验结果（n=379，$\kappa_2 = 0.161$，轻度一致）

话语类别	人工样本数	AI 一致样本数（TP）	一致率 (%)	说明
课堂管理	25	19	76.0%	与阶段一一致均偏高
示范	55	23	41.8%	一致率比阶段一更低
指示 / 组织活动	133	15	11.3%	大幅下降（阶段一为 51%）
社交互动	7	4	57.1%	样本偏少
知识讲解	196	38	19.4%	AI 持续严重低估
引入	40	2	5%	非常低，与阶段一一致
总结	12	0	0%	两阶段皆为零
总体	379	101	26.6%	低于阶段一（48.9%）

5 讨论与结语

本研究从效率、偏差结构与一致性三方面系统检验了智课语析系统在课堂话语分析中的表现，对三个研究问题可作如下回应：

（1）效率

AI 在转写与初步分类上的效率远高于人工。一小时课堂的人工转写与话语分析通常需 7-8 小时，而 AI 仅需数分钟即可完成初步处理，即便加上必要的人工校对，总体人力投入仍大幅降低（见表 1）。因此，AI 尤其适合作为课堂研究与行动研究的前置工具，用于扩大可分析的课堂样本、缩短从录制到回馈的时间线，并支撑更高频率的课堂观察与教学反思。

（2）准确性与偏差结构

在语篇功能分类上，AI 与人工之间的一致性仅处于轻微一致 slight–fair 水平（$\kappa_1 = 0.252$; $\kappa_2 = 0.161$; 整体 $\kappa = 0.210$），但偏差方向在 14 个类别中呈现 100% 的跨阶段一致。具体而言，AI 对依赖显性词汇与语气的类别（如课堂管理、部分示范及部分社交互动）高度敏感，往往出现显著或严重高估；而对依赖教学阶段性与概念推进的结构性类别（如引入、知识讲解及总结）则持续低估，甚至在“总结”类别完全无法识别。这种稳定的“双峰结构”偏差说明：AI 更擅长侦测表层形式特征，却缺乏对教学语篇结构与教师意图的理解能力。为进一步说明“总结类话语”误判的具体机制，表 6 列举若干具有代表性的语例：

表 6. 总结类话语的典型误判示例

原始语句	人工分类	AI 分类	误判原因
“今天我们学的是什么？”（课堂教学编号 8Z）	总结	知识讲解	句型为疑问句
“所以，谁能告诉我这是五何法里的哪一何？”（课堂教学编号 8Z）	总结	社交互动	“谁能…?”被视为互动
“那评量表你先收着，我们今天上课就到这里。”（课堂教学编号 10S）	总结	课堂管理	被识别为程序性话语

（3）对研究方法与教师专业发展的启示：

1 AI 适合作为高效预处理与趋势分析工具（教育研究方法层面）：鉴于其效率优势与偏差的可预测性，AI 更适宜被定位为“高效预处理者”和“趋势可视化引擎”，而非取代人工判断

的最终评估者。研究者可以据此设计“局部人工覆核”的混合流程：例如仅对高偏差或高风险类别（如引入、总结及知识讲解）进行人工复查，而将管理、部分示范类的话语比例作为相对可靠的趋势指标，从而在成本与效度之间取得平衡。

2 需从词面模式走向语篇建模（课堂话语分析理论与模型发展层面）：本研究的一致性与偏差结果表明，当前模型仍主要依赖词汇与句层特征，难以处理教学阶段、IRF 结构与多功能语句等语篇层次现象。若要推动 AI 参与课堂话语理论与模型的进一步发展，未来系统需要在输入与建模层面引入明确的语篇结构信息（如课堂阶段、活动类型及互动序列），并将“先判教学目的，再判功能类别”的目的驱动与分层分类机制纳入推理框架。这将有助于缩小 AI 在结构性教学话语上的“盲区”，使其由“词面匹配”逐步迈向“语篇理解”。

3 尽管 AI 无法取代教师在语篇层面的专业判断，但其在管理性话语、部分示范与互动类话语上的相对高一致性，仍可为教师提供具有参考价值的趋势性信息。例如，教师可利用系统生成的时间序列图观察“课堂管理语句占比随教学阶段的变化”，或比较前后测课堂在管理与讲解比例上的差异，而无需假定每一句标签都绝对准确。关键前提在于，教师必须理解不同类别结果的可靠程度，能够区分相对稳定的输出与需要谨慎解读的类别，从而对 AI 产出进行批判性筛选与责任判断。由此可见，对 AI 结果的判读能力本身，正逐渐成为教师专业发展与教育研究方法训练中的重要素养。然而，本研究主要从技术表现与偏差结构层面进行分析，尚未系统探讨教师在真实情境中的使用体验、接受度及培训需求。未来研究可结合技术接受模型与教师访谈，分析教师对不同类别结果的信任程度与使用意愿，并通过培训干预设计检验系统化学习是否能提升教师对 AI 偏差结构的理解与专业反思能力。同时，有必要通过行动研究或纵向追踪，验证 AI 反馈是否真正促进课堂话语结构的调整与教学行为的改变。唯有在技术效能、教师理解与实践转化之间建立稳定联系，AI 辅助课堂话语分析工具方能从“研究辅助工具”迈向“教师专业成长工具”。

4 系统优化的方向：基于本研究发现的系统性偏差，智课语析系统的后续优化可聚焦于三个方向。输入端：在转写文本中显式标注课堂阶段与活动类型，并提供结构化上下文窗口，以强化模型对课堂语篇结构的感知；推理端：采用“先判教学目的，再判功能类别”的分层与目的驱动架构，减少对情绪词、指令词等表层线索的过度依赖；输出端：叠加规则化校验与人机协同机制，对高风险类别触发二次判断或置信度提示，使教师能有针对性地进行人工校对。

6 综合结语

综合而言，本研究表明，AI 在课堂话语分析中的潜力与限制并存：它在效率上具有难以忽视的优势，却仍无法像经验丰富的教师与研究者那样，理解教学语篇的结构与意图。更恰当的定位并非“取代者”，而是作为一种人机协同框架中的高效助手——由 AI 负责规模化处理与趋势可视化，由教师与研究者负责语篇层次的诠释与教学意义的判断。未来的课堂分析系统若能在这一分工基础上，进一步将语篇建模、可解释性与教师专业学习机制结合起来，才有可能真正推动“人机共研、智化课堂”的长期发展。

注释

关于本文中偏差比例与偏差强度的计算与划分方法，作者在此进行说明解释。偏差比例的计算公式如下：

偏差比例（%）=（M_{AI} − M_{Human}）/ M_{Human} × 100%

根据偏差比例绝对值的大小，将偏差强度划分为以下四个等级：

（1）轻度偏差（$|\Delta| < 20\%$）：表示 AI 的估计与人工接近，仅有轻微高估或低估，通常不影响整体分类趋势。

（2）中度偏差（$20\% \leq |\Delta| < 50\%$）：表示 AI 在特定类别中出现稳定偏移，但仍维持部分与人工一致的判断模式。

（3）显著偏差（$50\% \leq |\Delta| < 100\%$）：表示 AI 的判断明显偏离人工，推断为模型策略（如依赖表层特征）造成的结构性差异。

（4）严重偏差（$|\Delta| \geq 100\%$）：表示 AI 的分类结果与人工显著不同，通常源自 AI 对该功能类别缺乏语篇意识或过度依赖显性语言线索，属于系统性高估或系统性低估的模式。

采用此分类方式的目的在于提供一种具可解释性（interpretability）且适用于语篇功能分析的偏差衡量方法，使研究者能辨识 AI 在不同话语类别中的优势区（例如表层特征强的管理性话语与互动性话语）与弱势区（例如依赖教学语篇结构的引入、讲解与总结）。此偏差强度等级亦用于跨阶段比较，以检验偏差结构是否具有可重复性（replicability）。

基金资助说明

本研究为新加坡南洋理工大学研究项目第二阶段 Teachers as Adult Learners: Effecting Professional Development and Teacher Change in Primary Chinese Language Teaching and Learning through Variation Theory and Multi-Perspectival Reflective Dialogue 的研究成果。

伦理审查批准说明

该研究已通过新加坡南洋理工大学伦理审查委员会审批（IRB-2024-717，2024），并依照相关研究伦理规范开展。

人工智能使用说明

关于生成式人工智能工具的使用说明如下：本文在论文撰写与修订过程中，有限度地使用了生成式人工智能工具（ChatGPT，OpenAI）作为辅助工具。其使用范围仅限于作者原有文本的语言表达优化与行文结构连贯性检查，并未用于研究问题的提出、研究设计、数据收集、数据分析或研究结论的生成。所有研究构想、方法选择、数据解释与学术判断均由作者独立完成并承担全部学术责任。作者已对所有经人工智能工具辅助修订的内容进行人工审校，以确保其准确性、原创性与学术诚信，符合学术出版的伦理规范。

附录

课堂话语功能分类框架与编码说明

话语功能类别	操作性定义	主要判定要点	课堂话语示例（节录）
导入	用于引出新课主题、激活学生背景知识并建立学习情境的课堂话语	出现在课堂初段；引出主题而非具体教学任务	我们先来看，小乐一家人要去什么地方？谁要来猜一猜？（课堂教学编号 1C）

示范	教师通过朗读、板书、示例或操作展示和示范学习方法	教师主动呈现范例，此时学生以观察为主	跟我一起读：“竹竿”。（课堂教学编号 3B ）
教学指令与组织	用于布置任务、说明学习步骤、组织课堂活动或引导学生操作的课堂话语	明确指令性；推动课堂流程	现在呢我要请你们呢先看这一个录像，然后呢告诉我，他们这一家人，他们去野餐的时候呢天气怎么样？（课堂教学编号 1C）
语言知识教学	直接讲解词语意义、句式结构、语法规则或文本语言特点的课堂话语	以语言知识建构为核心目标	你可以叫杆子，也可以叫竹竿。（课堂教学编号 11L）
总结 / 收束	对课堂学习内容进行回顾、概括或评价学习重点的课堂话语	多出现在课堂后段；具有概括或收束功能	我们学了用这几个来叙述，来描述事情的先后顺序。（课堂教学编号 2W）
课堂管理	用于维持课堂秩序、提醒行为规范或调节课堂状态的课堂话语	与教学内容无直接关系；以秩序管理为目的	Brandon 来，你跟着 Brandon 的后面。Huibin 来，Guangming, Kailei 出来。（课堂教学编号 7H）
课堂 / 社会性互动	用于建立情感联系、表达鼓励、回应学生情绪或进行非学科性的互动话语	情感取向明显；非知识或任务导向	非常棒！（课堂教学编号 4N）

参考文献

Creswell, J. W., & Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches* (5th ed.). SAGE Publications.

Flanders, N. A. (1970). *Analyzing teaching behavior*. Addison-Wesley.

Halcomb, E. J., & Davidson, P. M. (2006). Is verbatim transcription of interview data always necessary?. *Applied Nursing Research, 19*(1), 38-42. https://doi.org/10.1016/j.apnr.2005.06.001

Hall, J. K., & Walsh, M. (2002). Teacher–student interaction and language learning. *Annual Review of Applied Linguistics*, *22*, 186-203. https://doi.org/10.1017/S0267190502000107

Halliday, M. A. K. (1978). *Language as social semiotic: The social interpretation of language and meaning*. Edward Arnold.

He, Yuting(贺禹婷). (2025). 人工智能信息技术如何改善定性研究的数据收集与分析：基于变式理论 - 多视角反思性对话教师专业发展模式的案例研究 [How AI information technology can improve qualitative research data collection and analysis: A case study of the variation theory–multiple perspectival reflective dialogic professional development model], 南洋理工大学硕士学位论文 [Unpublished master’s dissertation, Nanyang Technological University].

Holmes, W., Porayska-Pomsta, K., Holstein, K., Sutherland, E., Baker, T., Shum, S. B., Santos, O. C., Rodrigo, M. T., Cukurova, M., Bittencourt, I. I., & Koedinger, K. R. (2022). Ethics of AI in Education: Towards a Community-Wide Framework. *International Journal of Artificial Intelligence in Education, 32*(3), 504-526. https://doi.org/10.1007/s40593-021-00239-1

Holstein, K., Aleven, V., & Rummel, N. (2020). A conceptual framework for human-AI hybrid adaptivity in education. In *International conference on artificial intelligence in education* (pp. 240-254). Springer International Publishing.

Hwang, G. J., & Chen, N. S. (2020). Effects of digital game-based learning on students' learning performance, motivation, and problem-solving skills: A meta-analysis. *Computers & Education, 148*, 103786. https://doi.org/10.1016/j.compedu.2020.103786

Jonassen, D. H., Carr, C., & Yueh, H. P. (1998). Computers as mindtools for engaging learners in critical thinking. *TechTrends, 42*(2), 24-32. https://doi.org/10.1007/BF02818172

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., & Hüllermeier, E. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences, 103*, 102274. https://doi.org/10.1016/j.lindif.2023.102274

Kranzfelder, P., Bankers-Fulbright, J. L., García-Ojeda, M. E., Melloy, M., Mohammed, S., & Warfa, A. R. M. (2019). The classroom discourse observation protocol (CDOP): A quantitative method for characterizing teacher discourse moves in undergraduate STEM learning environments. *PloS One*, *14*(7), e0219019. https://doi.org/10.1371/journal.pone.0219019

Luckin, R. (2018). *Machine learning and human intelligence: The future of education for the 21st century*. UCL IOE Press.

MacLean, L. M., Meyer, M., & Estable, A. (2004). Improving accuracy of transcripts in qualitative research. *Qualitative Health Research, 14*(1), 113-123. https://doi.org/10.1177/1049732303259804

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk* (3rd ed.). Lawrence Erlbaum Associates.

Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, *50*(4), 696-735. https://doi.org/10.1353/lan.1974.0010

Sinclair, J. M., & Coulthard, R. M. (1975). *Towards an analysis of discourse: The English used by teachers and pupils*. Oxford University Press.

Suresh, A., Jacobs, J., Harty, C., Perkoff, M., Martin, J. H., & Sumner, T. (2022). The TalkMoves dataset: K-12 mathematics lesson transcripts annotated for teacher and student discursive moves. In *Proceedings of the thirteenth language resources and evaluation conference* (pp. 4654-4662). European Language Resources Association.

Walsh, S. (2011). *Exploring classroom discourse: Language in action*. Routledge.

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education: Where are the educators?. *International Journal of Educational Technology in Higher Education, 16*(1), 39. https://doi.org/10.1186/s41239-019-0171-0

投稿：2025 年 12 月 14 日；接受：2026 年 3 月 8 日；出版：2026 年 4 月 6 日

作者简介

胡月宝，新加坡南洋理工大学国立教育学院副教授，现任亚洲语言文化学部副主任。其研究专长涵盖国际双语教育、人工智能教育与中文教学，并长期参与国际汉语教育相关的教师专业发展与培训研究工作。

贺禹婷，新加坡南洋理工大学国立教育学院国际汉语教育硕士研究生，2025 年毕业，其研究领域为国际中文教育及人工智能教育。

International Journal of Chinese Language Teaching (2026)
Vol. 7 (3) 5-19 https://doi.org/10.46451/ijclt.260402

Classroom Discourse Analysis in Human–AI Collaboration: Evidence from a Pilot Study in Singapore Primary Chinese Classrooms

Guat Poh Aw
Yuting He
Nanyang Technological University, Singapore

Abstract
Addressing the underdevelopment of teacher-oriented tools in Artificial Intelligence in Education (AIED), this study proposes an AI-based Classroom Discourse Analysis System (AI-CDAS) designed to support instructional diagnosis and classroom interaction analysis. The system was developed and evaluated using verbatim transcripts from 24 lessons taught by 12 primary Chinese language teachers in Singapore. Automated speech recognition was employed for transcription, and OpenAI's GPT-4 model was applied to classify discourse functions. System performance was examined by comparing AI-assisted and fully human workflows in terms of processing efficiency and classification consistency.

Results indicate that the AI-CDAS substantially reduced overall processing time at the transcription and preliminary coding stages, compressing the workflow to a fraction of the time required by manual analysis. However, agreement between AI and human coders on discourse function classification remained at a moderate-to-low level. A stable bimodal bias was observed: the model tended to overestimate utterances with salient formal features (e.g., classroom management and social interaction) while underestimating discourse functions dependent on broader pedagogical sequencing and discourse structure (e.g., lesson introduction, knowledge explanation, and summarization).

Based on these findings, the study positions AI not as a substitute for human judgment but as a pre-annotation and trend-detection tool, with teachers responsible for reviewing high-risk categories. The findings further suggest that future iterations should incorporate instructional phase indicators and discourse-structural features to enhance classification validity in classroom discourse analysis.

Keywords
Classroom discourse analysis, AI in education, human–AI collaboration, teacher professional development

Guat Poh Aw, Associate Professor, National Institute of Education, Nanyang Technological University, Singapore; Deputy Head, Asian Languages and Cultures Academic Group. Research interests: International Bilingual Education, Artificial Intelligence in Education, and Chinese Language Teaching.

Yuting He, Master's Graduate in Teaching Chinese as an International Language, National Institute of Education, Nanyang Technological University, Singapore. Research interests: International Chinese Language Education and Artificial Intelligence in Education.