PlanGlow: Personalized Study Planning with an Explainable and Controllable LLM-Driven System

Jiwon Chun
Texas A&M University
College Station, Texas, USA
jiwonchun@tamu.edu

Yankun Zhao
WisdomPlan Inc.
Pittsburgh, Pennsylvania, USA
yankunzhao@wisdomplan.ai

Hanlin Chen
Vanderbilt University
Nashville, Tennessee, USA
hanlin.chen@vanderbilt.edu

Meng Xia
Texas A&M University
College Station, Texas, USA
mengxia@tamu.edu

Figure 1: *PlanGlow* contains two major components. (A) is the plan generation interface where users specify their learning preferences such as subject, prior knowledge, and available time. (A1) is the background level description. (B) is the generated plan interface that presents a personalized study plan, organized into weekly segments with a detailed daily breakdown. (B1) allows in-line editing of learning goals, background knowledge level, duration, and daily availability. (B2) is the weekly breakdown including learning objectives, the reasons behind content selection, and key connections within the weekly.

Abstract

Personal development through self-directed learning is essential in today's fast-changing world, but many learners struggle to manage it effectively. While AI tools like large language models (LLMs) have the potential for personalized learning planning, they face issues such as transparency and hallucinated information. To address

this, we propose *PlanGlow*, an LLM-based system that generates personalized, well-structured study plans with clear explanations and controllability through user-centered interactions. Through mixed methods, we surveyed 28 participants and interviewed 10 before development, followed by a within-subject experiment with 24 participants to evaluate *PlanGlow*'s performance, usability, controllability, and explainability against two baseline systems: a GPT-4o-based system and Khan Academy's Khanmigo. Results demonstrate that *PlanGlow* significantly improves usability, explainability, and controllability. Additionally, two educational experts assessed and confirmed the quality of the generated study plans. These findings highlight *PlanGlow*'s potential to enhance personalized learning and address key challenges in self-directed learning.

CCS Concepts

• **Human-centered computing** → **Interactive systems and tools**; • **Applied computing** → *Education*.

Keywords

Self-directed Learning, Personalized Learning, Explainable AI, Controllable AI, Large Language Models, Study Planning

1 Introduction

Self-directed learning is a process where individuals take the initiative in their learning, with or without external assistance [11, 56]. Effective self-directed learning can enhance their overall learning experience, cultivate their habits of continuous learning, and promote personal growth [45, 53, 56, 57, 68, 73]. Learning platforms such as MOOCs, including edX[1], aim to support self-directed learners at scale with massive learning resources, but often lack individualized guidance.

It is challenging to provide personalized support at various stages of the self-directed learning process, from goal setting, planning, and selecting appropriate resources [53] to finding and effectively using interactive and multimedia resources [56]. Emerging technologies address these challenges by enabling personalized learning with access to information anytime, anywhere, making the process more accessible, autonomous, active, and independent [19, 70, 80]. Among these technologies, large language models (LLMs) have advanced self-directed learning in different stages [28, 52, 55, 56].

Most research applies LLMs during the execution stage of self-directed learning, providing resource access and task support. For instance, Ali et al. developed a GPT-4-based chatbot for outside-class Q&A and self-assessment through one-to-one interactions [3]. However, research on the planning stage, which involves goal setting, plan generation, and resource location, remains limited [53]. Although Lin et al. examined ChatGPT's role in supporting adult learners during the planning process [56], comprehensive solutions that address these challenges are still lacking.

In this context, LLM-based tools for the planning stage of self-directed learning face two major challenges that can impact user experience and effectiveness [61]. The first is the lack of transparency, which makes it difficult for learners to trust and follow recommendations without clear explanations [61, 72]. The second is the potential for incorrect information, which can confuse learners, negatively affecting their overall experience. The system can produce artificial hallucination information that lacks real-world basis [4]. These inaccuracies can further disrupt the planning process and require additional effort from learners to verify and correct.

To address the limitations of current LLM-based tools, we introduce *PlanGlow*, an LLM-assisted system for self-directed learning planning. Through a literature review, we identified explainability

and controllability as key factors. Explainability provides clear justifications for AI recommendations, fostering trust and informed interaction [72, 81, 82], while controllability allows users to adjust outputs and correct errors like hallucinations [38]. These two factors are interconnected because explainability provides users with the reasoning behind AI output, enhancing their ability to refine system behavior and reduce errors or hallucinations [38, 81]. With these features, *PlanGlow* contributes to scalable educational tools that support effective self-directed learning planning.

To validate these features, we conducted surveys and interviews with users after they generated study plans with GPT-4o. Based on this feedback, we consulted an educational expert to establish our theoretical framework such as Bloom's Taxonomy for learning objectives [65] and cognitive load theory [12, 25, 30] for content organization. These insights guided the development of *PlanGlow*, which integrates explainability and controllability as core features. Explainability offers concise summaries of the content, learning objectives, rationale for the chosen content, and connections between weekly and daily plans. Controllability lets users iteratively adjust preferences, validate resources, and refine study materials, aligning with Universal Design principles for diverse needs [13].

This dual approach improves user control, trust, and reliability by reducing misinformation and hallucinations. We conducted user studies to compare *PlanGlow* with two baselines: a GPT-4o-based system and Khan Academy's educational coach, Khanmigo. The results indicated significant improvements in both controllability and explainability, highlighting *PlanGlow*'s effectiveness.

The contributions of this work are summarized as follows.

- We propose *PlanGlow*, an LLM-based personalized study planning system designed to support learners in the initial stages of self-directed learning. By integrating explainability and controllability, *PlanGlow* provides a transparent and adaptive educational experience, encouraging learners to understand recommendations, make informed decisions, and take ownership of their learning path. The source code is available on Github[2].
- We conduct a within-subject experiment to evaluate the performance, usability, explainability, and controllability of *PlanGlow*, in comparison with two baseline systems. Experiment results show that users find *PlanGlow* more intuitive and integrated than the baselines, with controllability to create plans, verify resources, and explore alternatives. It also provides clear explanations, aligns plans with user goals, and clarifies task connections, allowing for more informed decision making.

2 Related Work

This section reviews LLM-based self-directed learning and explores explainability and controllability in educational AI systems.

2.1 Use of LLMs in Self-Directed Learning

Recent studies have explored using LLMs for self-directed learning such as progress monitoring, and reflection [3, 14, 39, 51, 56]. However, most studies focus on execution and monitoring phases, emphasizing real-time assessments, feedback, and support [3, 6, 9,

---

[1]https://www.edx.org

[2]https://github.com/dreamlab-24/PlanGlow

14, 26, 39, 56]. For instance, chatbots serve as virtual tutors, helping learners reflect on their understanding and adjust learning strategies [3, 9, 26]. TeacherGAIA, a GPT-4-based chatbot, illustrates this by guiding students through constructivist processes such as knowledge construction, inquiry-based learning, self-assessment, and peer teaching [3]. LLMs also enhance assessment by evaluating students' performance and delivering personalized feedback [6, 39]. Jones et al. developed a platform that enables teachers to design tasks and guide LLMs to provide formative feedback that identifies misunderstandings and encourages reflection [39].

However, a gap still remains in applying LLMs to the planning phase of self-directed learning. This phase, which includes goal setting, schedules structuring, and resource selection, is critical but underexplored. While previous works suggest that LLM-based tools such as ChatGPT could assist adult learners by identifying needs, setting goals, and recommending resources [29, 56], they often remain broad and risk delivering hallucination information. Consequently, more targeted research is needed to develop structured, personalized learning plans and address LLMs' limitations.

2.2 Explainability in AI for Education

LLMs have limitations in self-directed learning including biases, reliance on user input, lack of transparency, and hallucinations [14, 31, 35, 39, 56, 74]. These issues highlight the need for trustworthy AI systems that are transparent, robust, and secure to ensure ethical, legal, and safety standards are upheld [59]. Explainable AI (XAI) addresses this by clarifying AI reasoning, thus fostering trust, enhancing user satisfaction, and supporting metacognitive processes [18, 41, 59, 63].

However, flawed explanations can lead to biases, confusion, and disrupting learning [28, 42]. Several XAI systems already support self-directed learning, such as FUMA, which delivers hierarchical explanations in open-ended environments, improving learning outcomes and trust [20, 40, 43, 46, 47]. While the importance of explainability is recognized for LLM-driven planning tools [43], research on effectively implementing and measuring XAI during the planning phase remains limited [56].

2.3 Controllability in AI for Education

Controllable AI is an emerging strategy for managing real-world AI limitations by allowing users to guide systems [44]. In recommendation systems, it allows users to refine item selections or add constraints, though excessive freedom may increase cognitive load and obscure how prompts map to outputs [10, 22, 37, 71, 75]. Approaches include static preference profiles or dynamic updates via sliders or real-time visualizations [33, 37]. For conversational AI, chaining prompts and requesting explanations help users shape responses and boost performance [76, 77]. In educational AI, controllability remains underexplored despite its recognized importance [7, 32]. Aslan et al. used pre-defined templates to manage multi-modal pedagogical agents, but the impact of granting learners more control is unclear [7]. Therefore, this study investigates how controllability supports self-directed learning's planning phase.

3 User Needs and Design Requirements

This section presents the findings from a formative study to identify user needs for LLM-based self-directed learning planning, leading to four key design requirements.

3.1 User Needs

To understand user needs, we conducted a formative study with 28 participants, each completing a 15-minute survey for a $2 Amazon gift card. Ten of them also participated in follow-up interviews lasting 45 minutes, each receiving an $8 Amazon gift card. Additionally, we interviewed an educational researcher. The study was approved by the Institutional Review Board at Texas A&M University.

3.1.1 Survey. The survey included 28 participants (19 females, 9 males), aged 19–29 years ($M = 25.37, SD = 2.38$), recruited through Slack and WeChat. The survey has 25 questions including multiple-choice, yes/no, open-ended, and 5-point Likert scale questions. It explored participants' experiences with self-directed learning, AI tools, and study plan generation. It also examined the frequency and effectiveness of AI tools like ChatGPT, control over AI-generated results, explanations, and desired improvements in study planning. Two researchers analyzed open-ended responses using affinity diagramming [34], while quantitative analysis focused on descriptive statistics and response frequencies.

The results showed that of 26 participants (92.86%) who reported trying AI tools such as ChatGPT or Notion AI, 10 (38.46%) specifically used them for generating study plans. Among these, 5 participants (50%) used the tools occasionally, while 3 participants (30%) used them exclusively before exams. However, 9 out of 10 participants (90%) encountered issues with AI-generated study plans, reporting errors or inaccuracies such as presenting incorrect information, misunderstanding questions, or providing vague explanations without evidence. Participants also noted inconsistencies in study plans such as mismatched resources that are not aligned or instances where the AI misinterpreted context or abbreviations with multiple meanings which confused them. As a result, participants felt that AI-generated plans often require human intervention, with 40% stating corrections were needed frequently, 30% always, and 30% occasionally. While 8 participants (80%) found explanations provided in their previous experiences with AI somewhat helpful, 9 participants (90%) expressed a desire for more detailed reasoning behind AI recommendations. Additionally, participants expressed mixed-to-negative opinions about the ease of using the tools, with 40% finding them not easy, 30% feeling neutral, and 30% finding them somewhat easy. They emphasized the need for greater control over modifying plans and requested features such as alternative study resources, enhanced interfaces, and iterative editing capabilities to better personalize their learning experience.

3.1.2 Semi-structured interviews. We interviewed 10 participants (9 female, 1 male), aged 22–29 years ($M = 25.7, SD = 1.95$). Participants were asked to use GPT-4o to create personalized study plans by writing their own prompts (e.g., "*I want to study GraphQL within 2 weeks using books, YouTube, and technical blogs with the goal of deploying a website.*") and could modify and re-prompt as needed. After that, they provided feedback on the generated plans

and shared views on GPT-4o's potential for personalized study systems, AI explainability, user control, and suggested improvements. Interview questions are listed in Table 1.

The authors employed a systematic qualitative analysis following established practices [62]. Interview responses were categorized into feedback on the current system and suggestions for future systems. Open-coding [16] and affinity diagramming [34] were used for analysis. Two researchers collaboratively developed a codebook[3] using responses from three randomly selected participants and independently analyzed the rest, resolving discrepancies through discussion. The current system feedback highlighted positive aspects, such as comprehensiveness, clarity, and personalization, alongside limitations like resource accessibility issues, hallucinations, inconsistent topic connections, low trust, limited control, insufficient personalization, and inadequate explanations. Users emphasized the need for better explanations of topic relevance and purpose. Future system improvements centered on controllability, reliability, materials, and information reception. For controllability, users desired features like iterative editing and flexible time adjustments. Reliability concerns included resource validation and clear reasoning, with explanations for topic sequencing. Participants sought diverse, accessible resources, including alternative videos for varied learning styles. For information presentation, users preferred customizable formats such as toggle-able sections for detailed explanations. The analysis achieved a Cohen's Kappa of 0.89, indicating strong inter-rater reliability.

3.1.3 Interview with an educational researcher. We also conducted a one-hour Zoom interview with an educational researcher at the Center for Teaching Excellence at Texas A&M University specializing in curriculum and instruction. The goal was to gather expert insights on the system's alignment with established learning theories and evaluate the educational validity of user-requested features for future development. After presenting the GPT-4o generated study plans and discussing identified issues and user-requested features, the researcher emphasized the importance of integrating established learning theories into the study plan generation process such as Bloom's Taxonomy [65] for clear learning objectives. The researcher also suggested specific learning theories for each stage of the process, which we detail in Section 4. Additionally, she supported enhanced controllability and explainability would strengthen users' ability to effectively self-direct their learning plans.

3.2 Design Requirements

From surveys and interviews with participants (P1–P10) and an educational researcher (E1), we found that explainability and controllability were top priorities. Additional design requirements also emerged to address other needs identified in the formative study.

D1: Providing Contextual Explanations of the Recommendations. Participants emphasized the need for clear explanations of the system's recommendations, including the rationale behind resource selection, video prioritization, and topic significance. Understanding the importance of each topic would help them better

---

[3]https://osf.io/mfv8j?view_only = 85c27a46d83e4d78864e26c2e0bf1fa5

User Satisfaction

Q1: How satisfied are you with the results?
Q2: Where did you find challenging?
Q3: Which parts of the recommendations are most helpful?
Q4: Which parts do you find less helpful?
Q5: Do the suggested study plans and materials align with your expectations and needs?
Q6: Do you think LLM's suggestions make sense to you?
Q7: Do you trust the results?
Q8: Do you want to take this plan?

User Experience: Controllability
Q9: Do you feel you have enough control?
Q10: How was your experience re-prompting to refine the plan?

User Experience: Explainability
Q11: Do you need explanations for suggestions in study plans?
Q12: Did you get enough explanations?
Q13: What other explanations did you want to get?
Q14: How clear were the reasons behind the suggestions?
Q15: How do you prefer explanations to be provided (e.g., concise, detailed), and in what format (e.g., icons)?

Future Development
Q16: Would explanations help you prompt better?
Q17: What types of study materials do you prefer to use for learning?
Q18: Do you want to have alternative study materials? How many of them would you like?
Q19: What criteria do you use for alternative study materials?
Q20: Do you have any suggestions for future development?

Table 1: Interviews focus on three aspects: user satisfaction (Q1-Q8), user experience (controllability: Q9-Q10, explainability: Q11-Q15), future development (Q16-Q20).

evaluate and follow the plan. P2, P3, P4, and P5 noted that as beginners in the subject, they often needed more detailed explanations, such as guidance on where to start. E1 supported these views, highlighting the necessity of clear explanations in educational reasoning. Previous studies also show that clear explanations enhance usability and help users understand and trust system recommendations [1, 67]. Participants also preferred progressive disclosure, where detailed explanations can be revealed or hidden based on their needs. For example, P6 suggested toggle-able sections to prevent cognitive overload while allowing deeper exploration when needed [20]. Therefore, contextual and progressively disclosed explanations would improve user experience and support informed decision-making for study plans.

D2: Providing Reliable and Alternative Study Materials. Participants valued accessible and reliable study materials to be recommended. They wanted materials with verifiable links, brief descriptions, and relevant metrics (e.g., likes, views) to ensure reliability and match their expectations. Mohammadi et al. also demonstrated that reliable resources increase user trust and engagement in learning [64]. Additionally, participants preferred alternative resources to accommodate different learning preferences and paces. P6 and P10 emphasized the need for comparative information between alternatives to make informed decisions on learning resources. E1
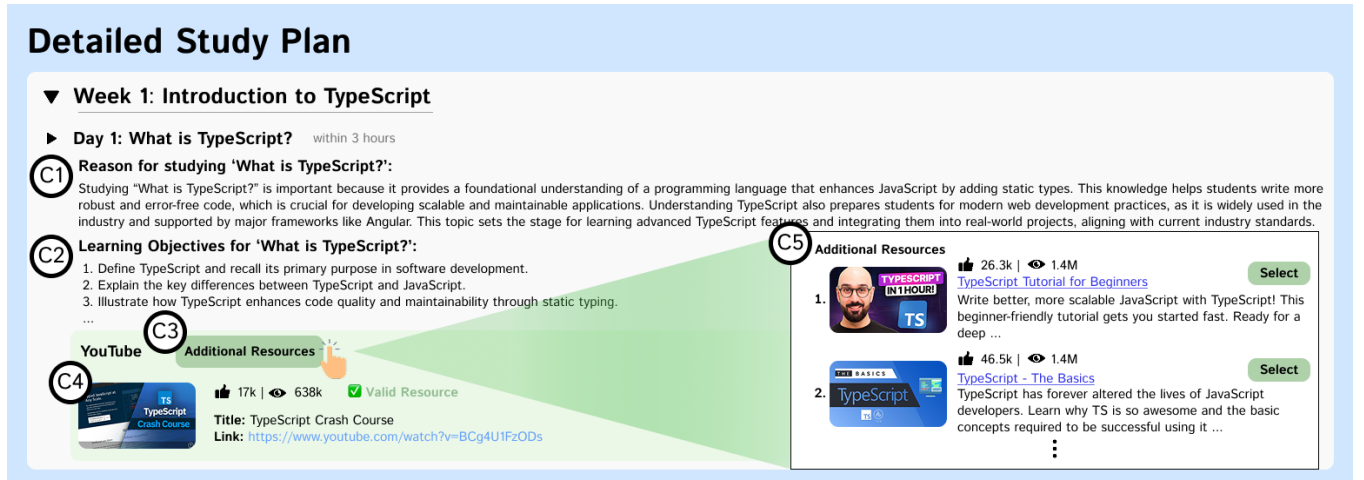
Figure 2: Detailed study plan of *PlanGlow* organizes each week into five days. (C1) explains the reasons for studying each day's topic. (C2) lists learning objectives. (C3) allows users to explore additional resources via the button connecting to (C5). (C4) displays video resources with their status. A green check icon marks 'Valid Resource', while a red icon indicates 'Invalid Resource'. (C5) displays 10 additional resources with views, likes, and video descriptions. Clicking the 'Select' button replaces the original plan's video with the selected one.

highlighted that diverse, reliable resources enhance user confidence, support learning styles, and foster trust for personalized plans.

D3: Enhancing Controllability by Allowing Easy Plan Creation and Modification. Participants prioritized the need for controllability in generating and customizing study plans to meet their specific needs. P2, P3, P4, P5, P6, and P10 expressed frustration with current systems that failed to provide satisfactory initial plans or made adjustments difficult. Research shows that study plan quality depends heavily on the input [56]. Incorporating structured input fields can guide users to provide precise responses, improving plan alignment. Participants also stressed the importance of easy, iterative modifications. Interactive features like clickable icons and editable fields empower users to refine plans in real-time, aligning them with learning goals. This aligns with Andragogy principles, which emphasize active involvement for adult learners. Empowering self-directed adults to customize plans fosters ownership, engagement, focus, and improved learning outcomes [49].

D4: Designing a User-Friendly and Clear Format for Study Plans. Clear formats and visuals improve comprehension and reduce cognitive load [66]. Participants (P1–P6, P8) emphasized breaking study plans into manageable sections, with lists to outline key topics and resources to enhance understanding and engagement. P5 suggested daily segments over weekly overviews for clarity, while P10 recommended interactive elements such as icons for easy customization. These features enhance usability and offer a more engaging and personalized experience.

4 *PlanGlow* System Design

PlanGlow is a web application consisting of two main components: the plan overview (Figure 1 (B1, B2)) and the detailed study plan (Figure 2). Its controllability features include **video replacement (D1, D2, D3)**, **input forms (D3, D4)**, **in-line editing (D3, D4)**, and a **chat feature (D3, D4)**, while its explainability features provide **background knowledge level descriptions (D1)**, **hierarchical explanations (D1, D4)**, and **video verification (D2)**. The application's front-end is developed using React.js with CSS styling, and the back-end is built with Python FastAPI. The back-end utilizes the OpenAI GPT-4o API to generate personalized study plans with explanations, such as content summaries, learning objectives, and reasons for resource selection. Detailed prompts can be found in the supplementary materials. The YouTube Data API v3 is integrated to recommend relevant and reliable resources.

To create personalized study plans, users start by filling out an input form in Figure 1 (A) with details such as the subject they want to study, learning goals, background knowledge, duration, and daily availability. This form enhances controllability by allowing users to customize their plans based on their needs and preferences. Its intuitive design improves usability by simplifying the process of providing necessary information, reducing cognitive load, and making the system accessible to a wide range of users, including those with no prior experience using prompt-based systems.

For users who are unsure about their current knowledge level, they can click the "bulb" 💡 icon in Figure 1 (A1) to see concise, informative descriptions of each level. This feature enhances usability by helping them confidently select the most appropriate level, ensuring the study plan is tailored to their abilities. The system leverages the OpenAI API to generate background knowledge insights based on the framework [8], which categorizes expertise into six levels: novice, advanced beginner, competence, proficiency, expertise, and mastery. Each level is clearly explained in the prompt, defined by specific characteristics of learning and performance progression as outlined in the framework. The prompt is fine-tuned with parameters: *temperature* (0.2) for factual accuracy, *top_p* (0.6)

for likely completions, *frequency_penalty* (0.2) to reduce repetition, and *presence_penalty* (0.1) to encourage relevant new information.

Once the form is submitted, the system generates a personalized study plan (Figure 1 (B)) using a three-step chain-of-thought [76] process with the OpenAI API, as outlined in Figure 3: the initial step, critique step, and improvement step. In the initial step, the system creates a draft study plan based on the user's form inputs. This step provides a coherent, week-by-week structure, integrating metacognitive strategies such as planning, monitoring, and evaluating learning. With these strategies, the system ensures the plan is practical, achievable, and support self-directed learning. This approach, grounded in metacognitive theory [12, 25, 30], provides a theoretical basis for creating study plans that enhance learning outcomes. To further promote the user's engagement and skill development, the system breaks the plan into daily tasks that increase in complexity over time and recommends accessible and relevant resources. Together, these elements generate a cohesive initial draft, which is refined in the critique and improvement steps. Next, in the critique step, the system evaluates the initial study plan using adult learning theories [48, 50] and study guide principles [24]. This step applies Knowles' Five Assumptions of Adult Learners and the Four Principles of Andragogy to ensure the study plan supports self-directed learning, leveraging prior experiences, emphasizing practical approaches, and encouraging internal motivation. The critique also checks the plan's clarity, relevance, and organization based on study guide principles. Finally, in the improvement step, the system refines the initial study plan using feedback from the critique. Throughout the process of generating the study plan, the system uses tuned parameters to ensure accuracy and quality: *temperature* (0.0), *top_p* (0.8), *frequency_penalty* (0.2), and *presence_penalty* (0.1). Users can further interact with *PlanGlow* through in-line editing (Figure 1 (B1)) to easily adjust key elements such as background knowledge level, study subjects, goals, duration, or daily availability. Additionally, they can utilize the chat feature to edit the plan or to ask any questions. Plan edits follow the three-step chain-of-thought process in Figure 3, while other queries receive targeted responses tailored to the user's needs.

The generated plan follows a layered hierarchical approach, starting with a high-level overview of the study plan. Users can expand weekly sections to view the learning objectives, reasoning behind the selected content, and the connections between daily materials in Figure 1 (B2). Further expansion reveals more detailed daily plans, balancing a broad understanding with in-depth exploration. This structure enhances usability by presenting complex information in manageable layers, from the overall plan to the finer details. The system employs Bloom's Taxonomy [65] to create clear and structured learning objectives. Through the OpenAI API, each objective from foundational skills such as remembering and understanding, to advanced skills such as analyzing, evaluating, and creating, ensures a well-rounded and progressive learning experience. The system generates reasoning for resource selection based on Vygotsky's Zone of Proximal Development (ZPD) [15] to match the learner's abilities and gradually increase complexity. It ensures diversity, accessibility, interactivity, and quality to support learners' growth and align with their goals. To enhance comprehension, the system connects relevant prior knowledge structures when introducing new content. This approach, grounded in schema theory [5], shows how learners use existing knowledge to interpret and integrate new information, enhancing understanding and retention.

For daily study plans, the system outlines the rationale for selecting topics (Figure 2 (C1)), specific learning objectives (Figure 2 (C2)). The prompt guides the system to explain the reasons for studying each topic by addressing learners' needs for clarity, orientation, and goal-setting. It highlights each topic's relevance within the study plan, its connection to broader course objectives, and its role in achieving learning goals [24]. The system also presents reliable and relevant video resources with likes, views, and validation (Figure 2 (C4)). It validates videos using the YouTube Data v3 API, includes valid videos in the study plan, and replaces invalid videos automatically with suitable alternatives matched to learner's topic, proficiency, and daily study time, sorted by ratings. To enhance usability, the system provides a clear visual indicator provides real-time feedback on video validity. In Figure 2 (C4), a green check icon indicates valid resources labeled as 'Valid resources,' while a red icon signifies invalid resources, ensuring reliability and enabling users to make adjustments as needed. In addition to the primary videos, users can click on Figure 2 (C3) to view up to 10 additional video suggestions (Figure 2 (C5)), ranked by relevance, with details such as views, likes, and descriptions. Users can select their preferred video to replace the original one in the study plan, offering flexibility to customize their learning experience. This enables users to tailor the plan to their individual preferences and goals, ensuring it effectively supports their learning process.

5 Evaluation of *PlanGlow* System

This section presents a within-subject user study evaluating *PlanGlow*'s performance, usability, and user experience, with a focus on controllability and explainability. The protocol was approved by the Texas A&M University's Institutional Review Board.

5.1 Participants

We recruited 24 current students from Texas A&M University (9 males, 15 females; 15 Bachelor's, 7 Master's, 2 Ph.D.), all of whom had experience with self-directed learning. A power analysis confirmed the sample size was sufficient for a large effect ($d = 0.8$, $\alpha_{\text{Bonferroni}} = 0.05/3$, power = 0.9). Of the participants, 21 were familiar with LLMs like ChatGPT, and 10 had used AI for study planning. The study, conducted via Zoom, lasted 1–1.5 hours per session, with participants accessing the systems as web applications. Each received a \$20 Amazon gift card as compensation.

5.2 Baselines

The study compares *PlanGlow* to two baseline systems: a GPT-4o-based system [27] and Khan Academy's Khanmigo[4]. GPT-4o is the commonly used system for generating personalized feedback and has been used as the baseline for generating curriculum for teachers [27]. Khanmigo is a popular commercial tool for generating study plans. The GPT-4o-based system has a simple interface where users are asked to "*Please input the subject you want to learn and goals you hope to achieve with your background, duration, and available time per day.*" in a text box (Figure 4). For Khanmigo, we use the 'Coach my academic and career growth' feature in Figure 4. To
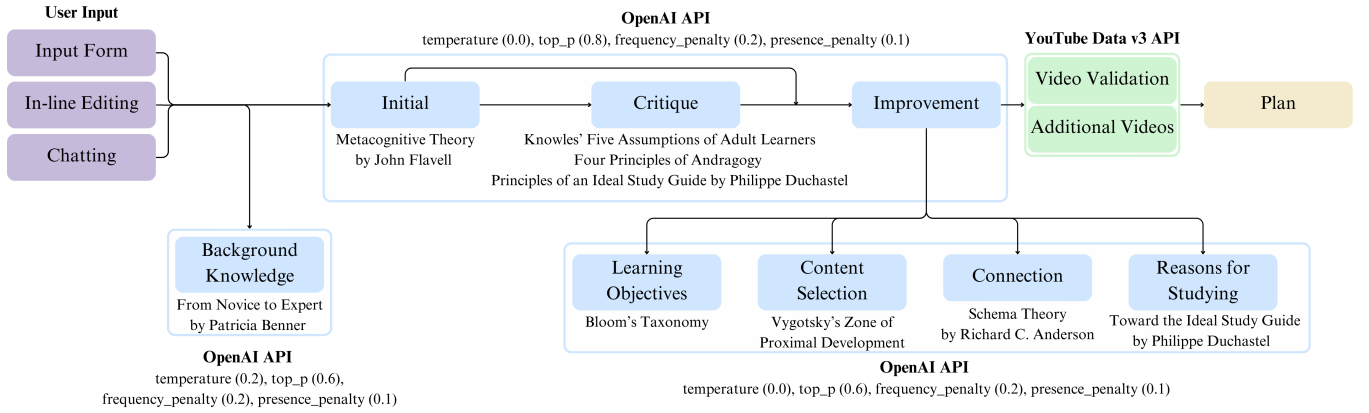
---

[4]https://www.khanmigo.ai/

Figure 3: The workflow of _PlanGlow_ progresses from left to right. The system begins by collecting user inputs through an input form, in-line editing, or chat interface to create an initial plan and describe the background knowledge level. The study plan is generated through three sequential steps using the OpenAI API: Initial Generation, Critique, and Improvement. The final plan incorporates comprehensive elements, including learning objectives, content selection rationales, conceptual connections across daily and weekly units, and explanations for studying each topic. All video resources are validated and supplemented by the YouTube Data v3 API. Each block details the specific learning theories applied and parameters tuned.
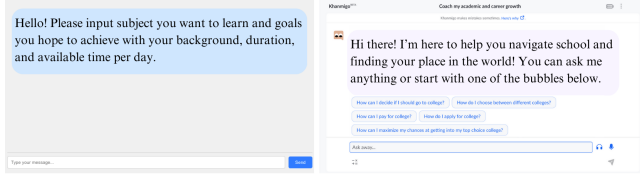

Figure 4: The interface of two systems are compared with _PlanGlow_ in the evaluation: a GPT-4o-based system (left) and Khan Academy's Khanmigo for the 'Coach my academic and career growth' feature (right).

ensure balanced testing, six possible system usage orders were assigned, with four participants per order.

5.3 Tasks

Before using each system, participants were introduced to its main features. They were asked to create personalized study plans on subjects of their choice using each system, without field restrictions. Participants iteratively create and refine their study plans as needed. Once they were satisfied with the generated plan from each system, they completed a post-task survey, explaining the reasons for their responses to each question. The survey included 22 identical questions, all based on a 7-point Likert scale (1 = Strongly Disagree, 7 = Strongly Agree) assessing performance, usability, and user experience, with a focus on controllability and explainability. A full list of the survey questions is provided in Table 2. Usability questions were adapted from the System Usability Scale and user experience questions were based on the previous framework [75]. After using all three systems, participants were asked which system they preferred the most.

5.4 Hypothesis

We propose the following hypotheses based on prior work [78].

H1. _PlanGlow_ outperforms baseline systems in performance by improving study plan clarity (H1a), alignment with user goals and needs (H1b), and user intention to adopt the plan (H1c).

H2. _PlanGlow_ demonstrates better usability over baseline systems by encouraging frequent use (H2a), reducing complexity (H2b), enhancing ease of use (H2c), integrating functions effectively (H2d), ensuring consistency (H2e), supporting quick learnability (H2f), and boosting user confidence (H2g).

H3. _PlanGlow_ offers greater controllability by enabling effortless study plan generation (H3a), efficiently achieving desired plans (H3b), effective resource search (H3c), reliable resource validation (H3d), and confident plan revisions to meet goals and needs (H3e).

H4. _PlanGlow_ provides more effective explanations, offering concise (H4a), accurate (H4b), and relevant (H4c) information. It helps users understand the rationale behind recommendations (H4d), aligns plans with user goals (H4e), clarifies connections between daily and weekly study tasks (H4f), and empowers informed decisions on following or revising plans (H4g).

5.5 Evaluation of Study Plans

We recruited two educators: E1, a female assistant professor in special education specializing in learning development at the Nanjing Normal University of Special Education, and E2, a male educational consultant at the Texas A&M University Center for Teaching Excellence. E1 and E2 collaboratively developed evaluation criteria for five standardized questions that incorporate a 5-point Likert scale (1 = Strongly Disagree, 5 = Strongly Agree), adapted from prior works [23, 54]. The full questions are in Table 3 and the evaluation criteria are available online[5]. Based on these criteria, E1 evaluated all generated plans from the user study. The plans were

---

[5]https://osf.io/bwjqg?view_only=85c27a46d83e4d78864e26c2e0bf1fa5

Performance
Q1: The study plan was clearly presented.
Q2: The study plan met my goals and needs.
Q3: I want to use this study plan in my learning process.

Usability
Q4: I would like to use this system frequently.
Q5: The system is unnecessarily complex.
Q6: The system was easy to use.
Q7: The various functions in the system were well integrated.
Q8: There was too much inconsistency in the system.
Q9: I would imagine that most people would learn to use this system very quickly.
Q10: I felt very confident using the system.

User Experience: Controllability
Q11: The system allows me to generate the study plan easily.
Q12: The system efficiently provides the desired study plan.
Q13: The process of searching for additional or alternative study resources is straightforward.
Q14: The system ensures that the process of verifying the validity of study resources is reliable.
Q15: I feel confident and able to revise the study plan easily to suit my specific goals and needs.

User Experience: Explainability
Q16: The system provides concise explanations.
Q17: The explanations offered by the system are accurate.
Q18: The explanations provided by the system are relevant.
Q19: The system helps me understand why certain recommendations or resources were included.
Q20: The system explains how the study plan aligns with my goals and preferences.
Q21: The system clearly explains the connection between daily and weekly study tasks.
Q22: The system offers explanations that help me make informed decisions about following or revising the study plan.

Table 2: Survey focuses on three dimensions: performance (Q1-Q3), usability (Q4-Q10), user experience (controllability: Q11-Q15, explainability: Q16-Q22).

Q1: Does the plan outline clear and specific learning objectives?
Q2: Does the plan include a timeline or estimated completion time?
Q3: Does the plan detail practical, well-supported learning activities that are feasible and make use of available resources?
Q4: Does the plan provide methods to monitor and measure progress toward achieving the learning goals?
Q5: Is the plan and its explanation pedagogically sound?

Table 3: Evaluation matrix for assessing generated plans.
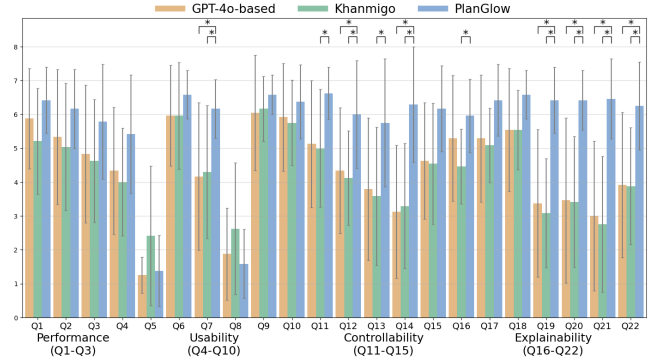
Figure 5: Means and standard errors of GPT-4o-based system, Khanmigo, and *PlanGlow* on a 7-point Likert scale (∗: $p < .05$).
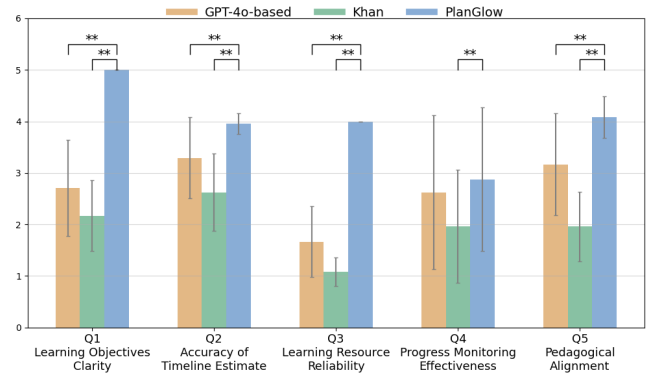
Figure 6: Means and standard errors for each evaluation assess the educational quality of generated plans on a 5-point Likert scale (∗∗: $p < .01$).

presented as text in a counterbalanced order along with each participant's context (subject, background knowledge, study duration, daily availability), while *PlanGlow* results were simplified to include only the Week 1 overview (Figure 1 (B1, B2)) and Day 1 detailed plan (Figure 2). E1 was informed that full explanations covered all weeks and days, and that resource availability had been confirmed but was not shown in the presented plans. Further details on how the plans were presented are available online[6].

5.5.1 Hypothesis. We hypothesize **H5** that *PlanGlow* enhances educational quality with clearer learning objectives (H5a), more accurate timelines (H5b), better-supported activities with reliable resources (H5c), more effective progress monitoring (H5d), and stronger pedagogical foundations (H5e) compared to the baselines.

6 Results

This section presents findings from participant feedback from the user study (P1–P24) and educational evaluations of *PlanGlow* and two baseline systems. Figure 5 details survey results, with statistical significance assessed using ANOVA with Bonferroni post-hoc tests. Figure 6 shows experts' evaluations of generated plans, analyzed using One-way ANOVA.

Participant Behavior and Learning Preferences. Participants' learning subjects fell into six categories: 29.2% focused on language

---

[6]https://osf.io/q3uf4?view_only=85c27a46d83e4d78864e26c2e0bf1fa5

learning (e.g., Spanish, Japanese), 20.8% on programming or technology (e.g., Typescript, Python), 20.8% on academic topics (e.g., UAVs, WJ IV intellect assessment), 12.5% on research and writing, and 8.3% on creative skills (e.g., video editing, dancing). Background knowledge levels included 29.2% novices, 41.7% advanced beginners, 20.8% competent, 4.2% experts, and 4.2% mastery. Interaction data showed all participants viewed the background knowledge description once, submitted plans 1.67 times, and used chat features 0.63 times on average. Two participants used in-line editing, making two modifications each. They accessed additional resources 6.13 times, selecting 3.5 on average, and viewed weekly and daily explanations 13.042 and 10.33 times, respectively.

Performance (H1). We do not notice significant performance differences across the systems. *PlanGlow* had higher means and lower standard deviations for study plan clarity ($M = 6.42, SD = 0.97$), alignment with user goals and needs ($M = 6.17, SD = 1.17$), and adoption ($M = 5.79, SD = 1.69$) compared to GPT-4o and Khanmigo, but results were not statistically significant. Thus, **H1a-H1c are rejected**. Participants described *PlanGlow*'s comprehensive and well-organized plans, highlighting clear goals and progressive learning paths (P9, P13, P15, P19, P24). P10 also noted alignment with a previous class she had taken. However, limitations such as slow response time (P6, P9, P16) and broad resources (e.g., full tutorials) reduced its advantage over baselines.

Usability (H2). Most usability hypotheses showed no significant differences between *PlanGlow* and the baselines ($p > .05$, **H2a-H2c, H2e-H2g rejected**). However, *PlanGlow* significantly outperformed both GPT-4o ($p < .01$) and Khanmigo ($p < .05$) in functional integration, **supporting H2d**. This improvement is attributed to *PlanGlow*'s comprehensive interface, offering multiple interaction methods beyond the chat-based focus of the baselines. The rejections of other hypotheses may reflect similar underlying LLMs across systems and the short-term study duration, limiting the exploration of nuanced usability differences.

User Experience: Controllability (H3). Participants reported greater controllability with *PlanGlow* compared to the baselines. *PlanGlow* facilitated easier study plan generation than Khanmigo($p < .05$, **H3a supported**) and was significantly more efficient in delivering desired plans than both GPT-4o and Khanmigo ($p < .05$, **H3b supported**). It also simplified searching for additional resources compared to Khanmigo ($p < .05$, **H3c supported**) and showed greater reliability in resource validation than both GPT-4o and Khanmigo ($p < .01$, **H3d supported**). However, no significant difference was found in capabilities to revise study plans to suit their specific goals and needs (**H3e rejected**). Participants mentioned *PlanGlow*'s structured input form for easier plan generation (P1, P9, P16, P19, P22) and its flexibility in customizing plans especially replacing video resources to fit their learning preferences and goals, P10 noted, "*The ability to search and replace study resources is very helpful. Previously, when studying this topic, I had a favorite YouTuber, and now I am glad that I can include their video that fits perfectly here.*" In contrast, Khanmigo's chat-based input was noted as challenging (P5, P10, P23). Despite these strengths, similar basic editing features across systems likely explain the lack of significant differences in plan revision.

User Experience: Explainability (H4). *PlanGlow* showed a significant advantage in providing concise explanations compared to Khanmigo ($p < .05$, **H4a supported**) but no significant differences in accuracy or relevance (**H4b, H4c rejected**). *PlanGlow* outperformed both GPT-4o and Khanmigo in explaining the rationale behind recommendations ($p < .01$, **H4d supported**), aligning study plans with user goals ($p < .01$, **H4e supported**), and clarifying connections between daily and weekly tasks ($p < .01$, **H4f supported**). *PlanGlow* also empowered users to make informed decisions about study plans compared to both baselines ($p < .01$, **H4g supported**). P24 stated, "*The explanations are concise, easy to read, understand, and intuitive.*" P3 added that, "*I like the concise explanations in the overview.*" Similarly, P1 commented, "*I like learning objectives in the overview and detailed plan because they help me understand what I need to know.*" P11 highlighted, "*I can easily understand why these are recommended to me, and they align well with my intentions.*" While all participants reviewed and approved the accuracy and relevance of the explanations, no significant differences were observed. Supporting this, P10 trusted Khanmigo's explanations due to its reputation as a reliable educational website.

Overall Preferences. A total of 83.3% of participants (20) ranked *PlanGlow* as their top choice, compared to 12.5% (3) for Khanmigo and 20.8% (5) for GPT-4o. Three participants provided tied rankings: P8 and P20 preferred both *PlanGlow* and GPT-4o, while P10 ranked all three systems equally.

Study Plan Quality (H5). *PlanGlow* significantly outperformed the baselines. Compared to Khanmigo, *PlanGlow* excelled in all aspects ($p < .01$, **H5a-H5e supported**). When compared to GPT-4o, it showed improvements in learning objectives, timeline estimates, reliable resources, and pedagogical principles ($p < .01$, **H5a-H5c, H5e supported**), while progress monitoring showed no significant difference ($p > .05$, **H5d rejected**).

7 Discussion

In this section, we discuss the design considerations, limitations, and future work.

7.1 Design Considerations

DC 1: The Need for Enhanced Explainability. Our study demonstrates strong user demand for explainable AI in educational planning. The interaction logs reveal substantial engagements with explanation features. Participants viewed weekly explanations 13.04 times and daily explanations 10.33 times on average, accessed additional resources 6.13 times, and selected 3.5 resources per session. These high engagement rates with explanatory features align with prior work highlighting the importance of AI explainability in education [17, 79]. The frequent access to explanations suggests that users value understanding the AI's reasoning, particularly in plan structure and resource recommendations.

DC 2: The Need for Enhanced Controllability. The study revealed important insights about user preferences regarding control in AI-assisted learning plan generation. Participants valued the structured input form (Figure 1 (A)) for straightforward plan generation, aligning with Satyanarayan et al. [71], which suggests user preference for guided control over full autonomy in AI interactions.

However, participants made limited use of additional control features. Specifically, only two participants utilized in-line editing, two edits each, and chat-based modifications averaging 0.63 per person. This low usage may reflect the one-time nature of the study, where participants may not have needed to significantly modify their plans. Future longitudinal studies could better evaluate the utility of plan modification features by examining how users adapt and update their learning plans over extended periods.

DC 3: Accountability and Pedagogical Responsibility. The integration of AI in educational guidance raises important ethical considerations that *PlanGlow* explicitly addresses. First, we maintain human agency by designing the system as a supportive tool rather than a replacement for educational expertise [36, 69]. *PlanGlow*'s explanations enable learners to make informed decisions about accepting or modifying recommendations which ensure accountability [58]. Second, we address pedagogical responsibility by incorporating established educational principles and maintaining transparency about its capabilities and limitations [36]. *PlanGlow* informs that its recommendations are suggestions rather than definitive instructions, encouraging critical evaluation by users. By allowing iterative refinement of plans, *PlanGlow* supports a balanced approach where AI assists but does not dictate educational decisions, preserving the essential role of human judgment in learning. While our current evaluation has not identified specific ethical concerns, we acknowledge potential future challenges, particularly when dealing with sensitive subjects or specialized domains. As future work, we plan to investigate safeguards and guidelines for handling potentially sensitive educational content and ensure responsible AI-guided learning across diverse subject matters [2].

7.2 Limitation and Future Work

However, this work still has several limitations. *PlanGlow* faces challenges in generalizing effectively across diverse domains due to the limitations of the underlying LLMs. These models are trained on vast datasets that may embed inherent biases or lack domain-specific knowledge, potentially leading to inaccurate plans in certain areas. A particular challenge emerges with specialized college-level subjects, where the system's performance is hindered by the scarcity of structured learning resources and detailed study plans in public datasets. For example, during user studies, highly domain-specific topics' plans such as UAVs or WJ IV intellect assessment lacked concrete implementation steps and sufficient domain expertise. While general educational content exists, resources for advanced specialized topics were limited. To address these limitations, we will implement two main solutions. First, we will utilize specialized academic datasets from MIT OpenCourseWare [60] and expert-validated study guides from Coursera [21]. Second, we will develop domain-specific prompting strategies in collaboration with subject matter experts. These combined strategies aim to mitigate biases, improve scalability, and ensure *PlanGlow* can provide accurate and adaptable plans across various domains.

Another limitation is that *PlanGlow* relies solely on YouTube videos as the primary resource. This design choice reflects our focus on enhancing the system's controllability and explainability in its current iteration, rather than developing a comprehensive tool that integrates multiple resources. While this approach is effective for certain users, it limits the diversity of available resources, particularly for those who prefer other formats such as reading materials and exercises. Evaluation results from two educators highlighted this limitation by rejecting H5d against GPT-4o, which questioned whether *PlanGlow* can monitor and measure progress toward achieving learning goals. Since *PlanGlow* is designed primarily to assist with the planning phase of self-directed learning, it lacks features to support learners during the execution and monitoring of their progress. To address these issues, we aim to include diverse resources, such as assessments and interactive exercises, and to implement methods for monitoring and measuring learners' progress. These enhancements will involve strategies to track progress toward achieving learning goals and to strengthen learners' confidence in their self-directed learning abilities [54].

The last limitation was the skewed gender distribution during interviews, with most participants being female, potentially biased feedback. Future studies should ensure balanced gender representation for more inclusive insights into system design.

8 Conclusion

In this paper, we present *PlanGlow*, a personalized study planning system offering explainable and controllable recommendations. Through a within-subject user study with 24 participants, we found that *PlanGlow* significantly outperformed two baselines. In controllability, *PlanGlow* demonstrated great performance in plan generation efficiency, resource management, and validation. For explainability, *PlanGlow* significantly improved users' understanding of recommendation rationales, goal alignment, and task connections, enabling more informed decision-making. *PlanGlow* also improved in study plan quality, including objectives, timelines, resources, monitoring, and pedagogy. While usability differences were minimal, *PlanGlow* achieved notably better functional integration, with 83.3% of participants ranking it as their preferred system. These findings highlight the transformative potential of integrating explainable and controllable AI in self-directed learning and education, offering a path toward more personalized, effective, and user-centered learning environments.

References

[1] Hasan Abu-Rasheed, Mohamad Hussam Abdulsalam, Christian Weber, and Madjid Fathi. 2024. Supporting student decisions on learning recommendations: An llm-based chatbot with knowledge graph contextualization for conversational explainability and mentoring. *arXiv preprint arXiv:2401.08517* (2024).

[2] Mohammad Amir Khusru Akhtar, Mohit Kumar, and Anand Nayyar. 2024. Socially Responsible Applications of Explainable AI. In *Towards Ethical and Socially Responsible Explainable AI: Challenges and Opportunities*. Springer, 261–350.

[3] Farhan Ali, Doris Choy, Shanti Divaharan, Hui Yong Tay, and Wenli Chen. 2023. Supporting self-directed learning and self-assessment using TeacherGAIA, a generative AI chatbot application: Learning approaches and prompt engineering. *Learning: Research and Practice* 9, 2 (2023), 135–147.

[4] Hussam Alkaissi and Samy I McFarlane. 2023. Artificial hallucinations in ChatGPT: implications in scientific writing. *Cureus* 15, 2 (2023).

[5] Richard C Anderson. 2018. Role of the reader's schema in comprehension, learning, and memory. In *Theoretical models and processes of literacy*. Routledge, 136–145.

[6] Nursultan Askarbekuly and Nenad Aničić. 2024. LLM examiner: automating assessment in informal self-directed e-learning using ChatGPT. *Knowledge and Information Systems* (2024), 1–18.

[7] Sinem Aslan, Lenitra M Durham, Nese Alyuz, Eda Okur, Sangita Sharma, Celal Savur, and Lama Nachman. 2024. Immersive multi-modal pedagogical conversational artificial intelligence for early childhood education: An exploratory case

study in the wild. *Computers and Education: Artificial Intelligence* 6 (2024), 100220.

[8] Patricia Benner. 1982. From novice to expert. *AJN The American Journal of Nursing* 82, 3 (1982), 402–407.

[9] Chantelle Bosch and Donnavan Kruger. 2024. AI chatbots as Open Educational Resources: Enhancing student agency and Self-Directed Learning. *Italian Journal of Educational Technology* (2024).

[10] Svetlin Bostandjiev, John O'Donovan, and Tobias Höllerer. 2012. TasteWeights: a visual interactive hybrid recommender system. In *Proceedings of the sixth ACM conference on Recommender systems*. 35–42.

[11] Stefanie L Boyer, Diane R Edmondson, Andrew B Artis, and David Fleming. 2014. Self-directed learning: A tool for lifelong learning. *Journal of marketing education* 36, 1 (2014), 20–32.

[12] AL Brown. 1978. Knowing when, where and how to remember. A problem of metacognition. In. R. Glaser (Ed.), Advances in instructional psychology (Vol. I). *Hillsdale NJ Erlbaum* 77 (1978), 165.

[13] Sheryl Burgstahler. 2009. Universal Design of Instruction (UDI): Definition, Principles, Guidelines, and Examples. *Do-It* (2009).

[14] William Cain. 2024. Prompting change: exploring prompt engineering in large language model AI and its potential to transform education. *TechTrends* 68, 1 (2024), 47–57.

[15] Seth Chaiklin et al. 2003. The zone of proximal development in Vygotsky's analysis of learning and instruction. *Vygotsky's educational theory in cultural context* 1, 2 (2003), 39–64.

[16] Kathy Charmaz. 2006. *Constructing grounded theory: A practical guide through qualitative analysis.* sage.

[17] Blerta Abazi Chaushi, Besnik Selimi, Agron Chaushi, and Marika Apostolova. 2023. Explainable artificial intelligence in education: A comprehensive review. In *World Conference on Explainable Artificial Intelligence*. Springer, 48–71.

[18] Li Chen and Feng Wang. 2017. Explaining recommendations based on feature sentiments in product reviews. In *Proceedings of the 22nd international conference on intelligent user interfaces*. 17–28.

[19] Constanţa Aurelia Chiţiba. 2012. Lifelong learning challenges and opportunities for traditional universities. *Procedia-social and behavioral sciences* 46 (2012), 1943–1947.

[20] Cristina Conati, Oswald Barral, Vanessa Putnam, and Lea Rieger. 2021. Toward personalized XAI: A case study in intelligent tutoring systems. *Artificial intelligence* 298 (2021), 103503.

[21] Coursera, Inc. 2024. Coursera. Online Learning Platform. https://www.coursera.org/ Accessed: 2024-12-10.

[22] Cecilia di Sciascio, Peter Brusilovsky, and Eduardo Veas. 2018. A study on user-controllable social exploratory search. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces*. 353–364.

[23] Fengning Du. 2012. Using Study Plans to Develop Self-Directed Learning Skills: Implications from a Pilot Project. *College Student Journal* 46, 1 (2012).

[24] Philippe Duchastel. 1983. Toward the ideal study guide: An exploration of the functions and components of study guides. *British Journal of Educational Technology* 14, 3 (1983), 216–231.

[25] Orpha K. Duell. 1986. MC Skills. In *Cognitive Classroom Learning*, Gary D. Phye and Thomas Andre (Eds.). Academic Press, Orlando, Florida.

[26] Elif Esiyok, Sahin Gokcearslan, and Kemal Gurkan Kucukergin. 2024. Acceptance of Educational Use of AI Chatbots in the Context of Self-Directed Learning with Technology and ICT Self-Efficacy of Undergraduate Students. *International Journal of Human–Computer Interaction* (2024), 1–10.

[27] Haoxiang Fan, Guanzheng Chen, Xingbo Wang, and Zhenhui Peng. 2024. Lesson-Planner: Assisting Novice Teachers to Prepare Pedagogy-Driven Lesson Plans with Large Language Models. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*. 1–20.

[28] Krzysztof Fiok, Farzad V Farahani, Waldemar Karwowski, and Tareq Ahram. 2022. Explainable artificial intelligence for education and training. *The Journal of Defense Modeling and Simulation* 19, 2 (2022), 133–144.

[29] Mehmet Firat. 2023. How chat GPT can transform autodidactic experiences and open education? (2023).

[30] JH Flavell. 1976. Metacognitive aspects of problem solving. *The nature of intelligence/Erlbaum* (1976).

[31] Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics* (2024), 1–79.

[32] Xiu Guan, Xiang Feng, and AYM Islam. 2023. The dilemma and countermeasures of educational data ethics in the age of intelligence. *Humanities and Social Sciences Communications* 10, 1 (2023), 1–14.

[33] F Maxwell Harper, Funing Xu, Harmanpreet Kaur, Kyle Condiff, Shuo Chang, and Loren Terveen. 2015. Putting users in control of their recommendations. In *Proceedings of the 9th ACM Conference on Recommender Systems*. 3–10.

[34] Rex Hartson and Pardha S Pyla. 2012. *The UX Book: Process and guidelines for ensuring a quality user experience.* Elsevier.

[35] Jianxing He, Sally L Baxter, Jie Xu, Jiming Xu, Xingtao Zhou, and Kang Zhang. 2019. The practical implementation of artificial intelligence technologies in

medicine. *Nature medicine* 25, 1 (2019), 30–36.

[36] Wayne Holmes, Kaska Porayska-Pomsta, Ken Holstein, Emma Sutherland, Toby Baker, Simon Buckingham Shum, Olga C Santos, Mercedes T Rodrigo, Mutlu Cukurova, Ig Ibert Bittencourt, et al. 2022. Ethics of AI in education: Towards a community-wide framework. *International Journal of Artificial Intelligence in Education* (2022), 1–23.

[37] Dietmar Jannach, Sidra Naveed, and Michael Jugovac. 2017. User control in recommender systems: Overview and interaction challenges. In *E-Commerce and Web Technologies: 17th International Conference, EC-Web 2016, Porto, Portugal, September 5-8, 2016, Revised Selected Papers 17*. Springer, 21–33.

[38] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *Comput. Surveys* 55, 12 (2023), 1–38.

[39] Erik Jones and Jacob Steinhardt. 2022. Capturing failures of large language models via human cognitive biases. *Advances in Neural Information Processing Systems* 35 (2022), 11785–11799.

[40] Samad Kardan and Cristina Conati. 2015. Providing adaptive support in an interactive simulation for learning: An experimental evaluation. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 3671–3680.

[41] Judy Kay. 2001. Learner control. *User modeling and user-adapted interaction* 11 (2001), 111–127.

[42] Panayiota Kendeou and Victoria Johnson. 2023. The nature of misinformation in education. *Current Opinion in Psychology* (2023), 101734.

[43] Hassan Khosravi, Simon Buckingham Shum, Guanliang Chen, Cristina Conati, Yi-Shan Tsai, Judy Kay, Simon Knight, Roberto Martinez-Maldonado, Shazia Sadiq, and Dragan Gašević. 2022. Explainable artificial intelligence in education. *Computers and Education: Artificial Intelligence* 3 (2022), 100074.

[44] Peter Kieseberg, Edgar Weippl, A Min Tjoa, Federico Cabitza, Andrea Campagner, and Andreas Holzinger. 2023. Controllable AI-An Alternative to Trustworthiness in Complex AI Systems?. In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*. Springer, 1–12.

[45] Rosemary Kim, Lorne Olfman, Terry Ryan, and Evren Eryilmaz. 2014. Leveraging a personalized system to improve self-directed learning in online educational environments. *Computers & Education* 70 (2014), 150–160.

[46] Simon Knight. 2020. Augmenting assessment with learning analytics. *Re-imagining university assessment in a digital world* (2020), 129–145.

[47] Simon Knight, Sophie Abel, Antonette Shibani, Yoong Kuan Goh, Rianne Conijn, Andrew Gibson, Sowmya Vajjala, Elena Cotos, Ágnes Sándor, and Simon Buckingham Shum. 2020. Are you being rhetorical? a description of rhetorical move annotation tools and open corpus of sample machine-annotated rhetorical moves. *Journal of Learning Analytics* 7, 3 (2020), 138–154.

[48] Malcolm Knowles. 1977. Adult learning processes: Pedagogy and andragogy. *Religious Education* 72, 2 (1977), 202–211.

[49] Malcolm S Knowles. 1978. Andragogy: Adult learning theory in perspective. *Community College Review* 5, 3 (1978), 9–20.

[50] Malcolm S Knowles, Elwood F Holton III, and Richard A Swanson. 2014. *The adult learner: The definitive classic in adult education and human resource development.* Routledge.

[51] Soila Lemmetty and Kaija Collin. 2020. Self-directed learning as a practice of workplace learning: Interpretative repertoires of self-directed learning in ICT work. *Vocations and Learning* 13, 1 (2020), 47–70.

[52] Belle Li, Curtis J Bonk, Chaoran Wang, and Xiaojing Kou. 2024. Reconceptualizing self-directed learning in the era of generative AI: An exploratory analysis of language learning. *IEEE Transactions on Learning Technologies* (2024).

[53] Huiyong Li, Rwitajit Majumdar, Mei-Rong Alice Chen, Yuanyuan Yang, and Hiroaki Ogata. 2023. Analysis of self-directed learning ability, reading outcomes, and personalized planning behavior for self-directed extensive reading. *Interactive Learning Environments* 31, 6 (2023), 3613–3632.

[54] Su-Ting T Li, Daniel J Tancredi, John Patrick T Co, and Daniel C West. 2010. Factors associated with successful self-directed learning using individualized learning plans during pediatric residency. *Academic pediatrics* 10, 2 (2010), 124–130.

[55] Fernando Antonio Flores Limo, David Raul Hurtado Tiza, Maribel Mamani Roque, Edward Espinoza Herrera, José Patricio Muñoz Murillo, Jorge Jinchuña Huallpa, Victor Andre Ariza Flores, Alejandro Guadalupe Rincón Castillo, Percy Fritz Puga Peña, Christian Paolo Martel Carranza, et al. 2023. Personalized tutoring: Chat-GPT as a virtual tutor for personalized learning experiences. *Przestrzeń Społeczna (Social Space)* 23, 1 (2023), 293–312.

[56] Xi Lin. 2024. Exploring the role of ChatGPT as a facilitator for motivating self-directed learning among adult learners. *Adult Learning* 35, 3 (2024), 156–166.

[57] Huijin Lu, Maria Limniou, and Xiaojun Zhang. 2024. Exploring the metacognition of self-directed informal learning on social media platforms: taking time and social interactions into consideration. *Education and Information Technologies* (2024), 1–28.

[58] Rose Luckin and Wayne Holmes. 2016. Intelligence unleashed: An argument for AI in education. (2016).

[59] Aniek F Markus, Jan A Kors, and Peter R Rijnbeek. 2021. The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies. *Journal of biomedical informatics* 113 (2021), 103655.

[60] Massachusetts Institute of Technology. 2024. MIT OpenCourseWare. Open Educational Resource. https://ocw.mit.edu/ Accessed: 2024-12-10.

[61] Bahar Memarian and Tenzin Doleck. 2023. ChatGPT in education: Methods, potentials and limitations. *Computers in Human Behavior: Artificial Humans* (2023), 100022.

[62] Matthew B Miles, A Michael Huberman, and Johnny Saldaña. 2014. Qualitative data analysis: A methods sourcebook. 3rd.

[63] Brent Mittelstadt. 2019. Principles alone cannot guarantee ethical AI. *Nature machine intelligence* 1, 11 (2019), 501–507.

[64] Vahid Mohammadi, Amir Masoud Rahmani, Abdullah Mohammad Darwesh, et al. 2019. Trust-based recommendation systems in Internet of Things: a systematic literature review. *Human-centric Computing and Information Sciences* 9 (2019), 21. doi:10.1186/s13673-019-0183-8

[65] University of Arkansas. n.d.. Using Bloom's Taxonomy. https://tips.uark.edu/using-blooms-taxonomy/ Accessed: 2024-06-19.

[66] Linnea Öhlund. 2020. Valuable Visuals: Defining a design space for presenting medical results.

[67] Jeroen Ooge, Shotallo Kato, and Katrien Verbert. 2022. Explaining recommendations in e-learning: Effects on adolescents' trust. In *Proceedings of the 27th International Conference on Intelligent User Interfaces*. 93–105.

[68] Hyanghee Park and Daehwan Ahn. 2024. The Promise and Peril of ChatGPT in Higher Education: Opportunities, Challenges, and Design Implications. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–21.

[69] Kaśka Porayska-Pomsta, Wayne Holmes, and Selena Nemorin. 2023. The ethics of AI in education. In *Handbook of Artificial Intelligence in Education*. Edward Elgar Publishing, 571–604.

[70] Ryan Rafiola, Punaji Setyosari, Carolina Radjah, and M Ramli. 2020. The effect of learning motivation, self-efficacy, and blended learning on students' achievement in the industrial revolution 4.0. *International Journal of Emerging Technologies in Learning (iJET)* 15, 8 (2020), 71–82.

[71] Arvind Satyanarayan and Graham M. Jones. 2024. Intelligence as Agency: Evaluating the Capacity of Generative AI to Empower or Constrain Human Action. *An MIT Exploration of Generative AI* (mar 27 2024). https://mit-genai.pubpub.org/pub/94y6e0f8.

[72] Jordan Richard Schoenherr, Roba Abbas, Katina Michael, Pablo Rivas, and Theresa Dirndorfer Anderson. 2023. Designing AI using a human-centered approach: Explainability and accuracy toward trustworthiness. *IEEE Transactions on Technology and Society* 4, 1 (2023), 9–23.

[73] Donggil Song and Curtis J Bonk. 2016. Motivational factors in self-directed informal learning from online learning resources. *Cogent Education* 3, 1 (2016), 1205838.

[74] Eric J Topol. 2019. High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine* 25, 1 (2019), 44–56.

[75] Zhijie Wang, Yuheng Huang, Da Song, Lei Ma, and Tianyi Zhang. 2024. PromptCharm: Text-to-Image Generation through Multi-modal Prompting and Refinement. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–21.

[76] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* 35 (2022), 24824–24837.

[77] Tongshuang Wu, Michael Terry, and Carrie Jun Cai. 2022. Ai chains: Transparent and controllable human-ai interaction by chaining large language model prompts. In *Proceedings of the 2022 CHI conference on human factors in computing systems*. 1–22.

[78] Meng Xia, Mingfei Sun, Huan Wei, Qing Chen, Yong Wang, Lei Shi, Huamin Qu, and Xiaojuan Ma. 2019. Peerlens: Peer-inspired interactive learning path planning in online question pool. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–12.

[79] Ling Xu. 2020. The dilemma and countermeasures of AI in educational application. In *Proceedings of the 2020 4th international conference on computer science and artificial intelligence*. 289–294.

[80] Ling Zhang, James D Basham, and Sohyun Yang. 2020. Understanding the implementation of personalized learning: A research synthesis. *Educational research review* 31 (2020), 100339.

[81] Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A Smith. 2023. How language model hallucinations can snowball. *arXiv preprint arXiv:2305.13534* (2023).

[82] Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2024. Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology* 15, 2 (2024), 1–38.