electronics

MDPI

Article

Advancing Generative Intelligent Tutoring Systems with GPT-4: Design, Evaluation, and a Modular Framework for Future Learning Platforms

Siyang Liu [1], Xiaorong Guo [1], Xiangen Hu [2,*] and Xin Zhao [3,*]

[1] Key Laboratory of Adolescent Cyberpsychology and Behavior (CCNU), Ministry of Education, Wuhan 430079, China; liusiyang@mails.ccnu.edu.cn (S.L.); gxr@mails.ccnu.edu.cn (X.G.)
[2] Department of Applied Social Sciences, Hong Kong Polytechnic University, Hong Kong 100872, China
[3] Manchester Institute of Education, The University of Manchester, Manchester M13 9PL, UK
* Correspondence: xiangen.hu@polyu.edu.hk (X.H.); skye.zhao@manchester.ac.uk (X.Z.)

Abstract: Generative Intelligent Tutoring Systems (ITSs), powered by advanced language models like GPT-4, represent a transformative approach to personalized education through real-time adaptability, dynamic content generation, and interactive learning. This study presents a modular framework for designing and evaluating such systems, leveraging GPT-4's capabilities to enable Socratic-style interactions and personalized feedback. A pilot implementation, the Socratic Playground for Learning (SPL), was tested with 30 undergraduate students, focusing on foundational English skills. The results showed significant improvements in vocabulary, grammar, and sentence construction, alongside high levels of engagement, adaptivity, and satisfaction. The framework employs lightweight JSON structures to ensure scalability and versatility across diverse educational contexts. Despite its promise, challenges such as computational demands and content validation highlight the main areas for future refinement. This research establishes a foundational approach for advancing Generative ITSs, offering key insights into personalized learning and the broader potential of Generative AI in education.

Keywords: GPT-4; generative AI; intelligent tutoring system (ITS); Socratic Playground for Learning (SPL); personalized learning (PL)

Citation: Liu, S.; Guo, X.; Hu, X.; Zhao, X. Advancing Generative Intelligent Tutoring Systems with GPT-4: Design, Evaluation, and a Modular Framework for Future Learning Platforms. *Electronics* **2024**, *13*, 4876. https://doi.org/10.3390/electronics13244876

Academic Editor: Ajay Bandi

Received: 25 November 2024
Revised: 7 December 2024
Accepted: 10 December 2024
Published: 11 December 2024

1. Introduction

Intelligent Tutoring Systems (ITSs) have emerged as a transformative tool within educational technology, offering adaptive and personalized instruction that meets the unique needs of individual learners [1]. Traditionally, ITSs have relied on rule-based algorithms and static content. While these systems can be effective in structured environments, they often lack the flexibility needed to adapt to diverse educational contexts and the evolving needs of students [1,2].

The advent of Generative Artificial Intelligence (AI), particularly large language models (LLMs) such as GPT-4, has added a new dimension to ITSs by enabling real-time content generation and highly personalized learner interactions. These advancements reshape the educational experience, offering adaptable, interactive support that more closely resembles human tutoring [3]. Unlike traditional ITSs, which are often constrained by predefined rules, generative models like GPT-4 engage learners dynamically—facilitating interactive dialogue, generating questions in real time, and providing tailored feedback that responds immediately to each learner's progress [4]. Moreover, Generative AI holds the potential to democratize access to high-quality individualized education, offering scalable support in contexts where traditional ITS solutions have limited reach [5,6]. For example, virtual tutors like ChatGPT help to address gaps in conventional classroom instruction, which frequently targets the collective rather than the individual. As a result, students at varying

ability levels may become either overburdened or disengaged, highlighting the need for more personalized forms of support [7].

Despite the transformative potential of Generative AI in ITSs, several critical challenges remain unresolved. There is a clear need for structured frameworks to effectively integrate GPT-4 into the ITSs and robust evaluation methods to measure its impact on personalized learning experiences. Addressing these gaps is essential to unlocking the full educational potential of Generative AI and ensuring its practical application across diverse learning settings. Therefore, this study's objectives are twofold: first, to design a Generative ITS framework that incorporates GPT-4's language-generation capabilities through a modular JSON-based structure, enabling real-time personalized educational support; and, second, to evaluate this system through a case study, the Socratic Playground for Learning (SPL), thus assessing its feasibility and effectiveness. The specific research questions guiding this study are as follows:

RQ1: How can a Generative ITS framework be designed using GPT-4's language-generation capabilities to enable real-time personalized educational support through a modular JSON-based structure?

RQ2: How can the performance of the Generative ITS be comprehensively evaluated in terms of feasibility, effectiveness, and learner engagement through case study methodologies like the Socratic Playground for Learning (SPL)?

2. Literature Review

2.1. GPT-4'S Application Features

GPT-4 represents a notable advancement in Generative AI, characterized by enhanced contextual understanding, dynamic feedback generation, and multimodal integration [8,9]. These capabilities distinguish GPT-4 from its predecessors and other generative models, making it a versatile tool in tasks requiring real-time, context-aware interactions [10]. Specifically, GPT-4 excels in providing tailored feedback and generating adaptive content dynamically. For instance, its ability to analyze inputs and produce detailed context-specific outputs has made it particularly effective in areas such as personalized feedback and programming support. Unlike earlier iterations, such as GPT-3, GPT-4 demonstrates greater coherence across multi-turn interactions and offers more nuanced responses that align closely with users' needs [9].

A key innovation of GPT-4 is its multimodal integration, enabling it to process and generate outputs from both text and image inputs. This capability broadens its application scope to include visual question answering, document analysis, and complex problem-solving tasks. By incorporating diverse data types, GPT-4 achieves a level of adaptability that surpasses text-only generative models like GPT-3.5. These multimodal capabilities make it particularly effective in contexts where textual and visual information must be synthesized simultaneously [9,11]. In addition to its multimodal functionality, GPT-4 shows significant improvements over its predecessors in multilingual and domain-specific tasks. For instance, it achieves superior performance on benchmarks such as the MMLU, excelling in over 24 languages and outperforming models like PaLM in reasoning and contextual adaptation. These enhancements highlight its capacity to handle diverse, complex tasks with higher accuracy and flexibility than prior models [9]. Despite these advancements, GPT-4 is not without limitations. It continues to face challenges such as generating inaccurate or biased outputs, and its high computational requirements limit its accessibility. Moreover, its finite context window can constrain performance in tasks requiring long-term dependency tracking. Addressing these issues remains a critical area for future development [8–10].

2.2. Intelligent Tutoring Systems

Intelligent Tutoring Systems have evolved substantially, from early rule-based frameworks to sophisticated, personalized learning environments. These systems are designed to simulate human tutoring by providing personalized instruction and support tailored to individual learners' needs [1]. A typical and efficient ITS architecture consists of four key interacting components that work together to facilitate this personalization [12]: the Expert Module, the Student Module, the Tutor Module, and the User Interface Module (see Figure 1).

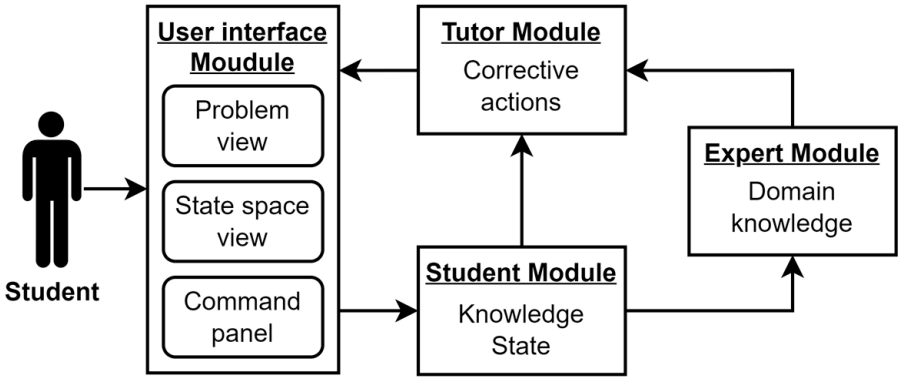


Figure 1. A traditional architecture of an Intelligent Tutoring System (ITS).

The Expert Module contains the essential domain knowledge, forming the base for instructional content and assessments. The Student Module represents the learner's current knowledge state, tracking their progress and performance through data on responses, exercise completion, and learning behaviors. The Tutor Module leverages these data to deliver personalized instruction, adjusting the content difficulty and approach based on the student's needs. Finally, the User Interface Module facilitates interactions between the ITS and the student, enabling question responses, feedback, and interactive exercises, thus creating an integrated and responsive learning experience.

In practice, the process works as follows: the student interacts with the ITS via the User Interface Module, providing responses and engaging in exercises. The Student Module continuously tracks these interactions, updating the learner's knowledge state. The Expert Module then references this state to guide the Tutor Module in adjusting the instructional approach, offering feedback, modifying the difficulty, and selecting content that best supports the learner's progress [13].

This four-component structure enables an ITS to provide a dynamic and responsive learning experience, which can be more effective than traditional static instructional methods. As a result, ITSs are now widely applied across various educational domains, enhancing outcomes in subjects like STEM, languages, and professional training by offering personalized guidance and targeted practice based on each learner's unique progression [14].

2.3. Integrating GPT-4 into ITS

The integration of LLMs like GPT-4 has opened up new opportunities for ITSs across various educational domains. LLMs enhance ITS capabilities by facilitating lesson design, generating feedback, and assessing learner knowledge. Studies have explored the utility of GPT-4 and related models in automating tasks traditionally handled by human educators. For instance, Ahmed [15] investigated ChatGPT's potential for conversation design and assessment within the Generalized Intelligent Framework for Tutoring (GIFT), demonstrating its ability to streamline the creation of educational scripts with reduced effort. Similarly, Schmucker et al. developed "Ruffle & Riley", a conversational tutoring system that utilizes GPT-4 to generate tutoring scripts from lesson texts, speeding up content authoring through EMT-based rules for free-form conversation [16].

Beyond script generation, GPT-4 has been employed in interactive conversational support. Abu-Rasheed et al. proposed an LLM-based chatbot that emulates peer-like mentorship by incorporating knowledge graphs and human supervision. This model enhances explainability, enabling the chatbot to clarify suggestions and provide educational recommendations in a mentoring role [17]. Another application, EduChat, developed by Dan et al., integrates retrieval-augmented question-answering, essay assessment, and Socratic teaching to offer personalized and compassionate support to students, teachers, and parents alike. By combining educational and psychological knowledge, EduChat provides both academic and emotional support, facilitating a more holistic approach to learning [18].

These Generative AI tools also contribute to automated performance assessment. Dai et al. demonstrated that GPT models can evaluate student submissions and generate feedback that is both readable and accessible, often surpassing the readability of feedback from human tutors [19]. Lin et al. further explored a GPT-4-powered feedback system that classifies trainee responses as correct or incorrect and provides explanatory feedback. The system automatically generates template-based responses, with GPT-4 refining incorrect responses for clarity and understanding [20]. Additionally, Zhang et al. found that LLMs can outperform traditional knowledge-tracing methods in predicting learning outcomes, highlighting their potential to enhance adaptive learning models, especially in adult literacy education [21].

3. Our Proposed Framework of Generative ITS

3.1. Framework and Functionality of Generative ITS

The Generative ITS framework is designed to provide adaptive, interactive, and highly personalized educational experiences. As shown in Figure 2, it comprises four primary components—Content Retriever, Data Analyzer, Instructional Advisor, and Feedback Assessor—alongside a User Interface Module, which facilitates seamless student interaction. Each component works in synergy to deliver a coherent and responsive learning environment.

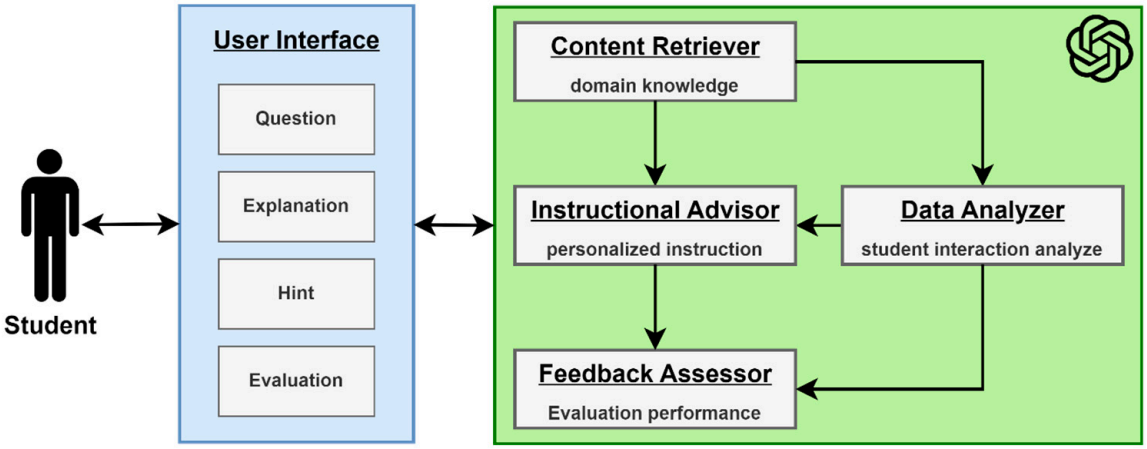


Figure 2. A demonstration of Generative Intelligent Tutoring System.

User Interface Module: This module acts as the central interaction platform for students. It features tools such as Question, Explanation, Hint, and Evaluation, enabling students to access tailored guidance and real-time feedback. For example, a student struggling with a math problem might receive targeted hints or explanatory visuals directly through the interface, fostering deeper understanding. The Feedback Assessor also integrates results into this module, helping students to track progress and identify areas for improvement.

Content Retriever: This module holds the system's domain knowledge, structured by topics and subtopics (e.g., mathematics, physics, and history). It dynamically adjusts content based on the student's preferences and progress. For instance, a student focusing on physics could receive progressively challenging content tailored to their learning goals.

Data Analyzer: Acting as the system's "brain", the Data Analyzer continuously monitors the student's knowledge state and learning trajectory. By analyzing responses, completion rates, feedback ratings, and time-on-task metrics, it generates a nuanced learner profile. This profile guides subsequent system decisions, such as recommending targeted practice exercises.

Instructional Advisor: Using data from the Data Analyzer and resources from the Content Retriever, this module delivers customized learning materials. It adapts the difficulty and type of content based on the student's performance. For example, if a learner demonstrates proficiency in fundamental algebra, the Instructional Advisor might introduce more advanced concepts or problem-solving strategies.

Feedback Assessor: This module evaluates overall learning outcomes, synthesizing insights from the Data Analyzer and Instructional Advisor. By generating detailed feedback reports, it supports both students and teachers in refining future learning strategies. For instance, it may highlight recurring errors in problem-solving or recommend additional resources.

Advanced Personalization Features: The Generative ITS employs AI-driven techniques to personalize learning further. It generates questions spanning diverse cognitive demands—recall, application, and analysis—and presents them in formats like multiple-choice or essay questions. The system also provides contextual explanations and dynamic hints, ensuring that the guidance aligns precisely with individual learner needs. For example, hints are adjusted for specificity, offering more detailed guidance when students struggle significantly while remaining concise for confident learners.

By supporting natural language conversations, the system simulates human-like tutoring. These interactive dialogues enable the ITS to respond dynamically to diverse topics, fostering engagement and critical thinking. For instance, a student studying history might engage in a multi-turn Socratic dialogue exploring the causes of a historical event, guided by the system's contextual prompts.

3.2. Module Development with GPT-4 and JSON

The development of the Generative ITS leverages GPT-4 and the lightweight JSON format to streamline modular system design and enhance scalability. Traditional ITS development involves phases such as needs assessment, cognitive task analysis, initial implementation, and evaluation [22]. These phases often require significant human expertise, particularly during task analysis and module creation. However, GPT-4's advanced programming capabilities provide a unique opportunity to automate and optimize these processes [23].

Why JSON? JSON (JavaScript Object Notation) is a widely adopted data format due to its simplicity, readability, and efficiency. It has become the preferred choice for data interchange in ITS development, replacing XML for its lightweight nature [24]. JSON's modular structure facilitates the design of system components, enabling seamless integration and scalability [25–27].

Module Design with GPT-4: Using GPT-4, the ITS modules are constructed as JSON objects with 13 pre-defined keys: unique identifier, name, type, role, data requirements, responsibilities, input sources, output targets, recommended models/modules, initial prompts, target prompts, output keys, and notes. This standardized design ensures modularity and compatibility across components. For instance, the Content Retriever module might include fields specifying input from the Data Analyzer and outputs to the Instructional Advisor, maintaining alignment between modules. The initial prompt of the ITS is shown in Table 1.

Table 1. The initial prompt structure for Generative Intelligent Tutoring System (ITS) modules.

Section	Description
Role Definition	You are an expert in Intelligent Tutoring Systems (ITSs).
Task Description	Create 4 JSON objects, each representing a key model or module within an ITS. These objects will collaborate to build a fully functional ITS for teaching purposes.
JSON Object Requirements	Each JSON object must contain the following 13 keys:
BotID	Unique identifier, typically an abbreviation of the module's name plus a random number.
name	The name of the module.
type	The model or module type within the ITS.
role	The module's job function.
data	The type of data required.
responsibility	The module's specific tasks.
input_source	List of BotIDs from which data are retrieved. Use "missing" if none.
output_target	List of BotIDs to which data are sent. Use "missing" if none.
model_and_module_recommended	List of recommended models/modules if "missing" is found in "input_source" or "output_target".
initial_prompt	Initial instructions for the module to clarify its role and tasks, summarize interactions with "input_source" and "output_target" modules, and ensure the output is in JSON format.
prompt_to_target	Additional instructions when sending output to target modules, mentioning the information being sent.
output_keys	List of possible keys in the JSON output.
notes	Additional explanations about the module.
Execution Steps	Steps to perform before generating the final output:
(a) Create Recommended Modules	For any module with a non-empty "model_and_module_recommended", create additional modules accordingly and include them in the list.
(b) Validate JSON Objects	Ensure each JSON object contains all 13 keys. If not, recreate it to include all keys.
(c) Update Module Relationships	If new modules are added, adjust "input_source" and "output_target" values to reflect the correct relationships. Ensure consistency across all modules.
(d) Ensure Pure JSON Format	Present each module as a pure JSON object containing all 13 keys.
(e) Organize Modules	Combine all modules into a cohesive list, maintaining the correct order and relationships.
(f) Repeat Validation	Repeat steps (a) to (e) until all modules have an empty "model_and_module_recommended" field.
Note	Always provide the final list of modules in JSON format, each containing all 13 keys.

Automated Module Generation: GPT-4 facilitates automated creation and validation of these modules. For example, when a module requires additional submodules (e.g., for specialized feedback), GPT-4 generates these autonomously while ensuring adherence to the JSON structure. Each module undergoes iterative validation to confirm completeness and alignment with system objectives. This process reduces development time and ensures a robust framework.

3.3. Evaluation Method for Generative ITS

To assess the effectiveness of the Generative ITS, this study adopts a comprehensive multi-dimensional evaluation framework based on the work of Chrysafiadi et al. [28].

The evaluation encompasses five key dimensions: effectiveness, engagement, adaptivity, satisfaction, and recommendation accuracy.

Effectiveness: Measured through pre- and post-tests to determine knowledge gains. For instance, students complete assessments before and after using the ITS, with improvements in scores indicating the system's impact on learning outcomes. Additional feedback from students provides qualitative insights into perceived learning effectiveness [28–30].

Engagement: Evaluated using interaction data, such as time-on-task and feature usage frequency, complemented by questionnaires assessing students' interest and attention levels. For example, sustained engagement during multi-turn dialogues suggests the system's ability to maintain learner focus [28,30].

Adaptivity: Assessed through system log tracking regarding how frequently and effectively the ITS adjusts content based on individual needs. Feedback surveys capture the perceived relevance of these adaptations, providing a balanced view of the system's responsiveness [28,30].

Satisfaction: Measured using metrics such as the Net Promoter Score (NPS) and questionnaire responses regarding interface usability, clarity, and overall experience. For example, students who rate the system highly in these areas are likely to recommend it to peers, reflecting user acceptance [28,30].

Recommendation Accuracy: Evaluated by analyzing the alignment between system recommendations and student needs. Corrective actions resulting from system feedback are tracked alongside student perceptions of the guidance provided [28,30].

By adopting this rigorous evaluation methodology, the study demonstrates the Generative ITS's ability to deliver personalized, adaptive, and engaging learning experiences while providing actionable insights for further refinement.

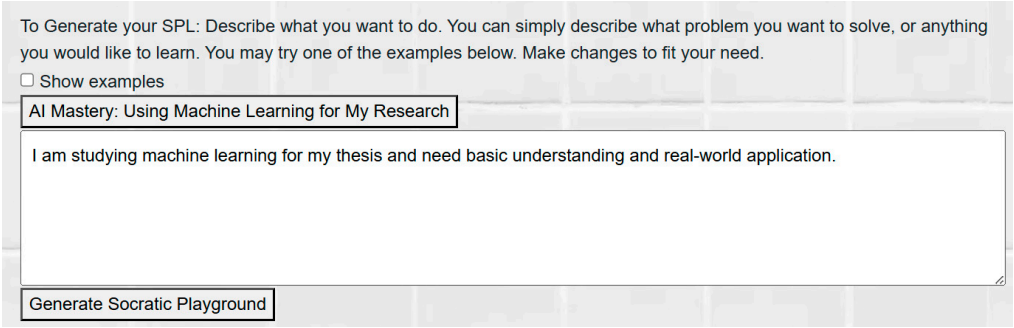
4. A Case Study: Socratic Playground for Learning (SPL)

The Socratic Playground for Learning (SPL) is structured around a sophisticated modular architecture that aligns with the four key components from the Generative ITS framework: Content Retriever, Data Analyzer, Instructional Advisor, Feedback Assessor, and User Interface. These modules collectively enable the SPL to deliver an intuitive Socratic-method-based educational experience that fosters critical thinking, adaptability, and personalized learning.

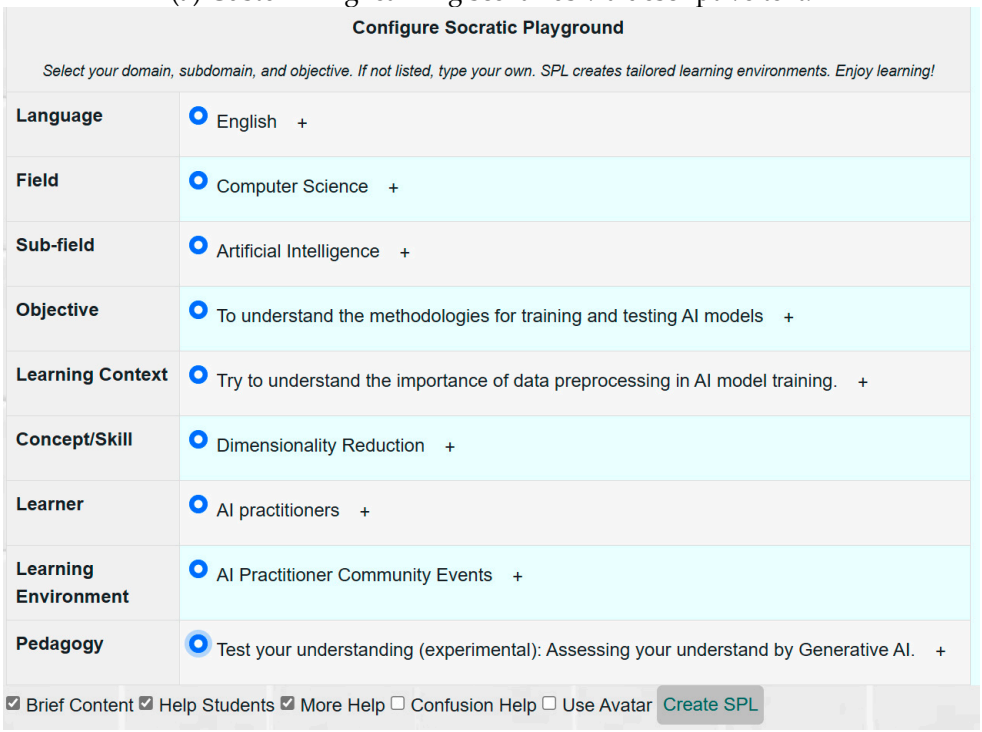
4.1. System Architecture of SPL

The SPL adopts a two-tiered architecture that operationalizes the Generative ITS framework through scenario construction and interactive dialogue stages, enhancing both critical thinking and independent learning [31]. Each stage is supported by specific modules from the framework.

Constructing Learning Scenarios: The Content Retriever module dynamically selects domain-relevant content and organizes it based on the learner's objectives and preferences. This enables users—both learners and educators—to create customized learning scenarios either through descriptive text or by selecting options from a hierarchical structure (see Figure 3). The structure incorporates domains, objectives, context, and concepts tailored to target learner groups. For example, a user interested in psychology might select "Educational Psychology" as the domain and set the objective as understanding the "Impact of Motivation on Student Learning". The Content Retriever then generates a structured learning scenario with associated contexts, such as "Role of Extrinsic Rewards in Motivation", and concepts like "Behavioral Reinforcement". The Data Analyzer ensures that the learner's current knowledge state is considered during this process. It analyzes the past learning behavior and achievements to recommend contextually relevant scenarios, thereby increasing the alignment between the learning content and the learner's needs.



(**a**) Customizing learning scenarios via descriptive text.



(**b**) Customizing learning scenarios via hierarchical selection.

Figure 3. Methods for customizing learning scenarios in the SPL system. (**a**) Customizing learning scenarios via descriptive text, enabling users to specify their learning objectives and areas of interest. (**b**) Customizing learning scenarios via hierarchical selection, enabling users to define parameters such as field, sub-field, objectives, and context to create tailored learning environments.

Interactive Dialogue Environment: Once a scenario is established, the Instructional Advisor and Feedback Assessor modules collaboratively support a multi-turn Socratic dialogue driven by a sequence of wh-questions (What, Why, How, etc.) (see Figure 4). The dialogue starts with the Instructional Advisor presenting questions tailored to the learner's context based on input from the Data Analyzer. The learners respond, and the Feedback Assessor evaluates their responses in real time, providing iterative feedback to prompt deeper thinking, correct misconceptions, and expand understanding. The Instructional Advisor dynamically adjusts the difficulty and type of follow-up questions based on learner progress. For instance, if a learner consistently demonstrates proficiency in foundational concepts, subsequent questions target higher-order thinking skills. The feedback loop remains responsive, ensuring that each interaction mirrors human-like tutoring with nuanced adjustments to questions and explanations.

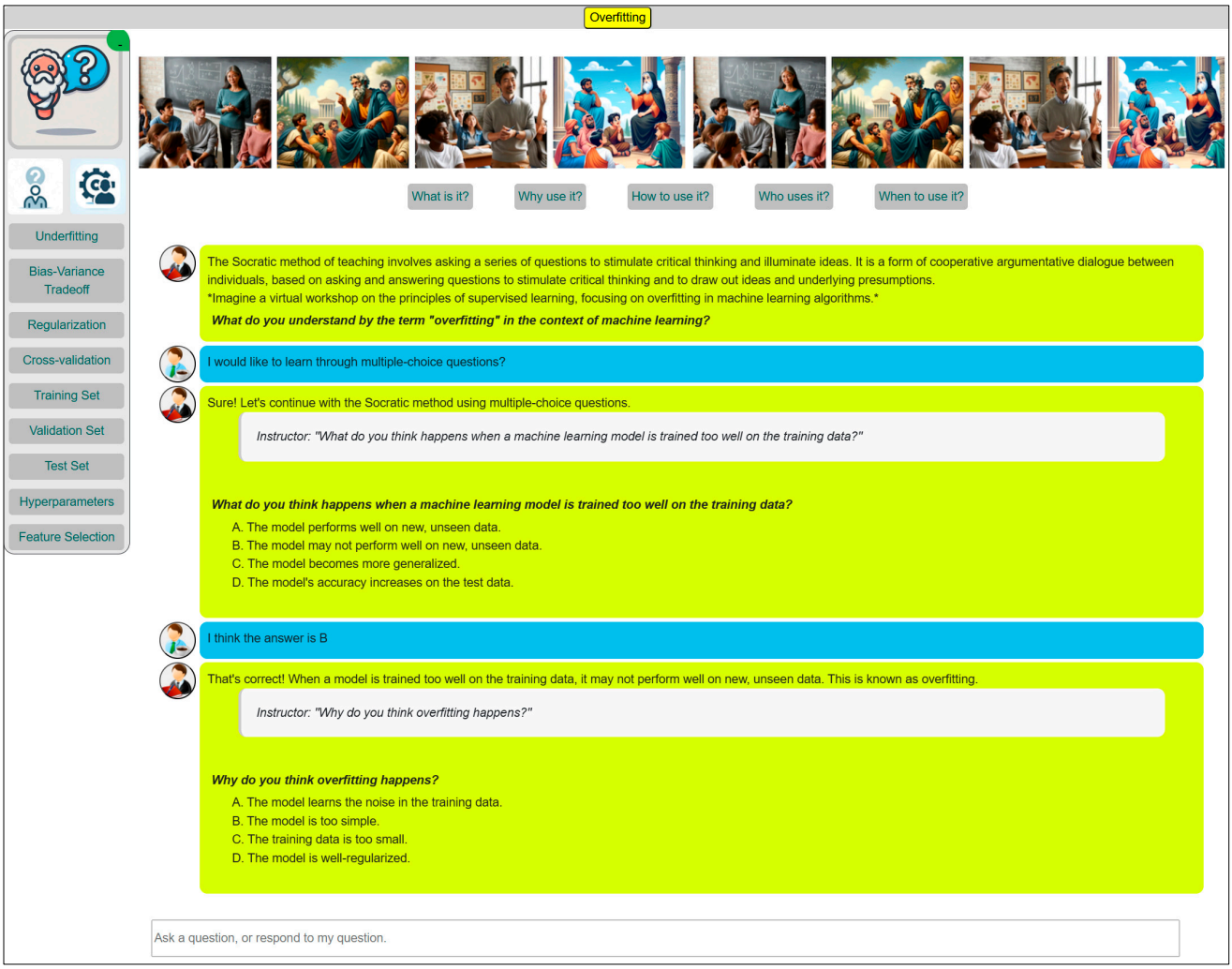


Figure 4. Interactive dialogue interface for the SPL system.

Prompt Engineering: The SPL architecture leverages GPT-4's generative capabilities, with prompt engineering playing a central role in aligning the learner's input with system-generated responses. The prompts are carefully designed to ensure that the interaction is both contextually appropriate and pedagogically effective, enabling the SPL to deliver a coherent and dynamic learning experience. Each prompt follows a standardized format, as shown in Figure 5, which specifies the key components required for effective system functionality. These components include the knowledge component (KC), learning context, target user, and instructional style. For instance, the Content Retriever uses the KC and context fields to generate domain-specific materials, while the Instructional Advisor leverages the target and the objective to adjust the difficulty and type of questions. Additionally, the Feedback Assessor iteratively refines prompts based on learner interactions, ensuring personalized and contextually relevant feedback. This design enables the SPL to deliver dynamic, scalable, and pedagogically sound learning experiences.

Your answers, both for now and for future interactions, will be presented in %[theLang]%.

You are producing some basic concepts, called knowledge components relevant to %[theKC]%, in %[theDomain]% for a group of %[theTarget]%.

Please give me %[theNumber]% concepts relevant to %[theDomain]%. output each separately, in pure json, following this format:

```
{
  "theAvatar":"%[theAvatar]%",
  "theLang":"%[theLang]%",
  "theKC":%[the_concept]%,
  "theType":"%[theType]%",
  "theTarget":"%[theTarget]%",
  "theTutorName":"%[theTutorName]%",
  "theContext":"%[theContext]%",
  "theEnvironment":"%[theEnvironment]%",
  "theUserName":"%[theUserName]%",
  "theStyle":"%[theType]%",
  "theObjective":"%[theObjective]%"
}
```

Making sure each of the entry in its own, pure, json.

Do not put all in one array. one json for each of %[theNumber]% concepts.

And the last but not least, making sure the value of the json objects are in %[theLang]%. in English only if you are not sure.

Make theKC short (less than 3 words if the language is English).

Figure 5. Structured prompt template for lesson creation regarding SPL.

4.2. A Pilot Evaluation Method

To evaluate the SPL system's effectiveness, engagement, adaptivity, satisfaction, and recommendation accuracy, a pilot study was conducted using both questionnaire-based and experimental methods. This multi-dimensional evaluation aimed to capture both objective learning outcomes and subjective user perceptions across key areas of system performance.

4.2.1. Experimental Design

This study involved 30 first-year students from the Faculty of Education at Central China Normal University, all majoring in Early Childhood Education. None of the participants had passed the CET-4 (College English Test Band 4), a standardized test widely used in China to assess undergraduate English proficiency, indicating limited English skills.

The study aimed to develop foundational English skills in vocabulary, grammar, and sentence construction. The experimental design followed a single-group pre-test/post-test format over one week. The participants completed a pre-test at the start to assess their baseline English skills. Each day, they used the SPL system for 30–40 min in a quiet computer lab equipped with stable internet connectivity. All the exercises were completed on desktop computers provided by the research team. Instructors were present throughout the sessions to ensure smooth system operation and to provide technical support when needed. To maintain a focused learning environment, non-study-related activities were prohibited during the sessions. The learners' characteristics are summarized in Table 2.

Table 2. Age and gender data of participants.

Age (Years)	Female (*n*)	Male (*n*)	Total (*n*)
18	4	5	9
19	5	5	10
20	5	4	9
21	2	0	2
Total	16	14	30

The evaluation consisted of two parts: objective learning gains and subjective user experience. The objective learning gains were measured by comparing the pre-test and post-test scores to evaluate the knowledge improvement. The subjective user experience was assessed through a questionnaire (see Table 3) covering perceived learning effectiveness, engagement, adaptivity, satisfaction, and recommendation accuracy, which were assessed using standardized questionnaire items rated on a 5-point Likert scale (1 = Strongly Disagree to 5 = Strongly Agree).

Table 3. The User Experience Questionnaire for the SPL system.

Dimension	Question	
Effectiveness	Q01	The learning activities in SPL helped me achieve my learning objectives.
	Q02	I feel that my understanding of the subject has improved through SPL.
	Q03	The feedback provided in SPL effectively facilitated my understanding.
Engagement	Q04	I was fully engaged and focused during my interactions with SPL.
	Q05	I felt motivated to continue using SPL during the learning process.
	Q06	I found the learning activities in SPL interesting and stimulating.
Adaptivity	Q07	SPL adjusted the learning content according to my responses and progress.
	Q08	The feedback provided by SPL was relevant to my needs and learning state.
	Q09	I felt that the learning experience in SPL was personalized to my needs.
Satisfaction	Q10	I am satisfied with my overall learning experience using SPL.
	Q11	The interface and design of SPL met my expectations for ease of use.
	Q12	I would recommend using SPL for learning to others.
Recommendation Accuracy	Q13	The system's recommendations in SPL helped me effectively learn the content.
	Q14	The content recommended by SPL met my learning needs and expectations.
	Q15	I trust the system's recommendations to guide my learning process.

4.2.2. Data Analysis

The data from the pre- and post-tests and questionnaires were analyzed to evaluate the effectiveness of the SPL system in enhancing the English proficiency and user experiences. Paired *t*-tests were conducted separately for the three components of the pre-test and post-test—vocabulary (scored out of 40), grammar (scored out of 30), and sentence construction (scored out of 30)—to assess the knowledge improvement in each area. Additionally, the questionnaire responses were analyzed to examine the participants' perceptions across five dimensions: learning effectiveness, engagement, adaptivity, satisfaction, and recommendation accuracy. Descriptive statistics, including mean scores and response distributions, were calculated to summarize the participants' evaluations of the system.

4.3. Results and Discussion

4.3.1. The Pre-Test and Post-Test

As shown in Table 4, the pre-test and post-test results demonstrate significant improvements in all the measured areas after using the SPL system. The paired *t*-tests revealed that the participants achieved higher scores in vocabulary, grammar, and sentence construction following the intervention. Specifically, the mean vocabulary score increased from 26.4 (SD = 9.2) to 30.7 (SD = 9.0), with a t-value of -9.8 ($p < 0.05$). The grammar scores improved

from 18.2 (*SD* = 7.3) to 23.1 (*SD* = 6.5), with a *t*-value of −8.1 ($p < 0.05$). Similarly, the sentence construction scores increased from 19.3 (*SD* = 8.3) to 23.2 (*SD* = 6.5), with a *t*-value of −5.4 ($p < 0.05$). The sum of all the scores showed the most substantial improvement, rising from 63.9 (*SD* = 11.8) to 77.0 (*SD* = 11.1), with a *t*-value of −12.5 ($p < 0.05$).

Table 4. Paired samples *t*-test results for pre-test and post-test scores.

Measure	Pre-Test Mean (*SD*)	Post-Test Mean (*SD*)	*t*	*df*	*p*
Vocabulary	26.4 (9.2)	30.7 (9.0)	−9.8	29	<0.05
Grammar	18.2 (7.3)	23.1 (6.5)	−8.1	29	<0.05
Sentence Construction	19.3 (8.3)	23.2 (6.5)	−5.4	29	<0.05
Sums	63.9 (11.8)	77.0 (11.1)	−12.5	29	<0.05

The significant improvements suggest that the personalized and adaptive exercises provided by the SPL addressed participants' learning gaps effectively. The enhanced vocabulary scores indicate improved word retention and understanding, which are critical for language proficiency. These results are consistent with prior research demonstrating that personalized learning systems effectively tailor instruction to individual needs, leading to improved learning efficiency and outcomes [32,33].

The targeted focus on individual weaknesses may have facilitated efficient knowledge acquisition, as supported by evidence from personalized tutoring studies [34]. However, the lack of a control group in the study limits causal inferences. The improvements could partially stem from repeated testing or greater familiarity with the assessment format. Additionally, while statistically significant, it remains uncertain to what extent these improvements translate into practical language use or long-term retention, a concern also raised in the personalized learning literature [35]. Future studies should address these limitations by incorporating diverse performance-based assessments and establishing control groups to validate the effects further. Despite these limitations, the SPL system demonstrates strong potential as a personalized learning tool, emphasizing the importance of personalization in achieving meaningful learning outcomes.

4.3.2. The User Experience Questionnaire

The questionnaire results, visualized in Figure 6, provide insights into participants' perceptions of the SPL system across five dimensions: effectiveness, engagement, adaptivity, satisfaction, and recommendation accuracy. Among these dimensions, effectiveness received the highest mean score (9.70), indicating that the participants strongly perceived the system as helpful in achieving their learning goals. Engagement (9.53) and adaptivity (9.43) also scored highly, reflecting the system's ability to maintain attention and provide personalized learning experiences. Recommendation accuracy (9.07) was well-rated, suggesting that the system effectively aligned its recommendations with participants' needs. Satisfaction received a slightly lower but still positive score (8.70), indicating room for improvement in the user experience design. The high ratings for effectiveness and engagement underscore the SPL system's ability to deliver an impactful and engaging learning experience. These results highlight the potential of adaptive systems to create tailored educational experiences that address individual learner needs [36]. The positive ratings for adaptivity and recommendation accuracy suggest that the participants appreciated the system's ability to customize the learning content and offer relevant suggestions, which aligns with research emphasizing the importance of personalization in educational technologies [34].

Figure 7 illustrates the percentage distribution of the scores across these dimensions, further reinforcing the system's strong performance in effectiveness, engagement, adaptivity, and recommendation accuracy. However, the lower satisfaction score points to areas for improvement, such as enhancing the interface design, resolving usability issues, or providing more comprehensive user guidance. The research on personalized learn-

ing technologies underscores the importance of a user-centered design to optimize both functionality and learner experience [37].

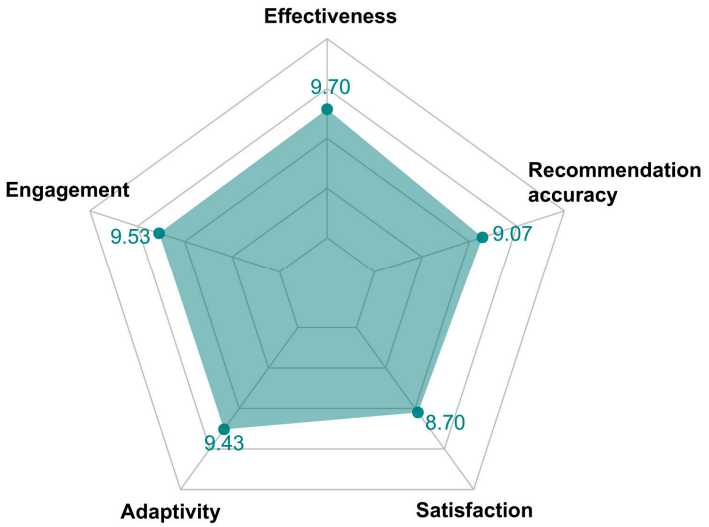


Figure 6. SPL system evaluation across five user experience dimensions.

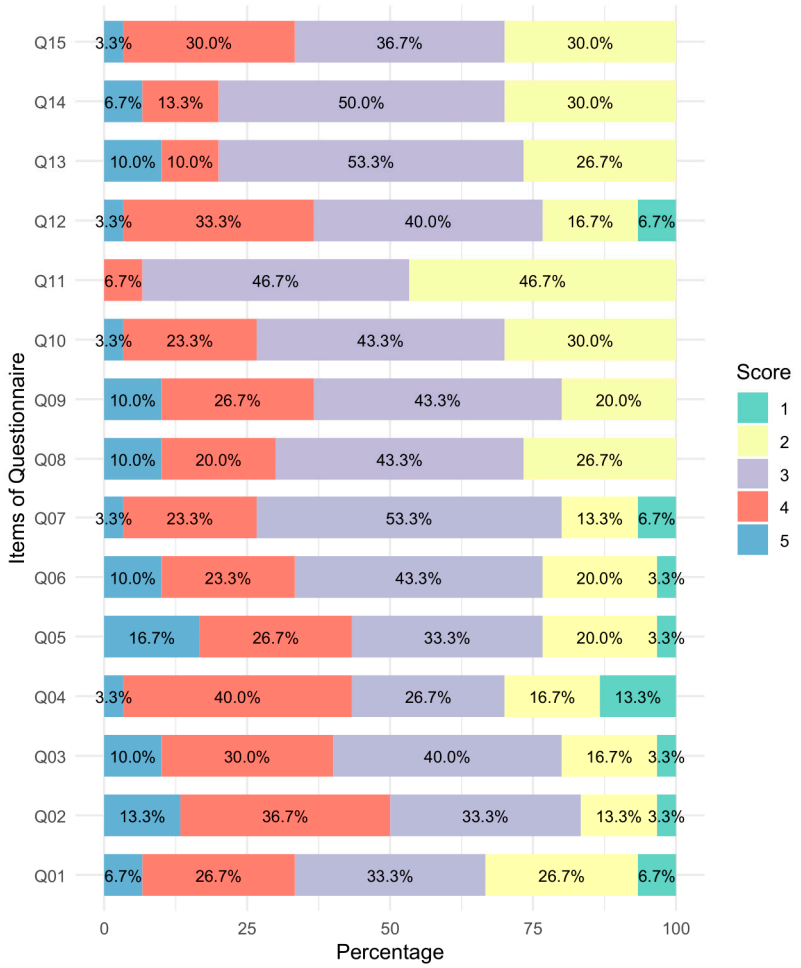


Figure 7. Percentage distribution of scores for User Experience Questionnaire.

5. Discussion

5.1. Contributions and Implications

This study makes significant contributions to the integration of GPT-4 into Intelligent Tutoring Systems (ITSs), demonstrating their potential to revolutionize personalized learning. The modular framework developed in this research, along with its implementation through the Socratic Playground for Learning (SPL), showcases measurable improvements in learner outcomes, particularly in vocabulary acquisition, grammatical accuracy, and sentence construction. By dynamically adapting content and feedback to individual learner profiles, the Generative ITS framework effectively addresses personalized learning needs, outperforming traditional ITSs in flexibility and adaptability.

The practical implications are substantial. The use of lightweight data formats like JSON and the scalability of GPT-4 significantly enhance the accessibility and efficiency of ITS development. This modular approach provides a clear, replicable pathway for deploying Generative AI in diverse educational settings, from formal classrooms to self-directed online learning platforms. Additionally, the integration of Socratic methodologies promotes critical engagement with content, leading to deeper understanding and improved long-term retention. These contributions offer a transformative model for advancing ITS development and adapting it to a wide range of educational contexts.

5.2. Limitations and Future Work

The integration of GPT-4 into an ITS provides significant advantages, particularly in personalization and scalability. By dynamically tailoring content and feedback to individual learners, GPT-4 facilitates personalized learning, enhancing engagement and outcomes [38]. Its natural language capabilities further enable human-like, interactive learning environments, making educational experiences more engaging [29]. Additionally, GPT-4 supports scalability by automating the generation of instructional materials and feedback, reducing educators' workload and broadening accessibility across diverse educational settings. These advantages align with studies that demonstrate the superior effectiveness of ITSs compared to traditional instruction methods [39].

However, challenges remain. High computational requirements may restrict the use of GPT-4 in resource-constrained environments, such as rural schools. Its potential to produce inaccurate or biased content underscores the need for rigorous validation and oversight [38]. Furthermore, effective system design and prompt engineering demand significant expertise to align the system with specific pedagogical goals [2]. Overcoming these challenges is critical to ensuring equitable access and the reliability of educational content.

While these technical and operational challenges highlight areas for improvement, the limitations of this study also warrant attention. This research provides a conceptual framework for a GPT-4-powered Generative ITS but lacks empirical evaluation. Future research should address these gaps by focusing on real-world implementation and testing in diverse educational contexts, involving both students and educators. Comparative analyses of Generative ITSs, the existing ITS systems, and the traditional learning approaches would offer valuable insights into their practical impact and effectiveness [40].

Additionally, future studies should explore the versatility of the proposed framework by extending its application to diverse academic disciplines, including STEM fields, the arts, humanities, and social sciences. Such interdisciplinary applications would demonstrate the framework's adaptability and utility across varied educational contexts. Addressing students' emotional factors, such as learning anxiety and motivation, is essential to fully understand the system's influence on user experience. Incorporating standardized questionnaires and advanced AI methods to assess and respond to these factors could significantly enhance the system's adaptability and relevance. Expanding the application of Generative ITSs across diverse academic disciplines—including the arts, humanities, and social sciences—would further demonstrate their versatility and utility [41]. Enhancements in AI techniques, such as machine learning, computer vision, and advanced natural language processing, could improve the system's accuracy, error detection, and responsive-

ness to learners' needs [19]. Integrating pedagogical strategies like collaborative learning, game-based learning, and personalized instruction would foster deeper engagement and sustained motivation, improving the overall learning experience [42].

Finally, the growing use of AI in education necessitates addressing ethical and social concerns. Issues such as algorithmic bias, data privacy, accountability, and transparency require careful attention. Establishing comprehensive guidelines and policies for responsible AI use will ensure equitable, secure, and ethically sound learning environments. Addressing these concerns proactively is essential to support the sustainable integration of AI in education [43].

6. Conclusions

The results demonstrated significant improvements regarding the learner outcomes, confirming the value of Generative ITSs in providing personalized and adaptive educational experiences. The primary contributions of this research include a modular framework designed for scalability and adaptability, a replicable evaluation methodology, and empirical evidence supporting the potential of Generative ITSs to expand access to high-quality education, particularly in underserved settings. As Generative AI continues to advance, this framework is poised to evolve alongside emerging technologies, paving the way for increasingly sophisticated applications and unlocking new opportunities for scalable, personalized learning.

Author Contributions: Conceptualization, S.L., X.H. and X.Z.; Software, S.L. and X.H.; Validation, S.L. and X.G.; Investigation, S.L. and X.G.; Data curation, S.L. and X.G.; Writing—original draft, S.L. and X.G.; Writing—review and editing, S.L., X.G. and X.Z.; Supervision, X.H. and X.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Informed Consent Statement: Informed consent was obtained from all participants involved in the study.

Data Availability Statement: The SPL Platform is available at https://polyu.skoonline.org (accessed on 10 November 2024). The datasets are available in the Figshare repository via the following link: https://figshare.com/s/e6ed0bc1104a45a4d93b (accessed on 18 November 2024).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Nwana, H.S. Intelligent Tutoring Systems: An Overview. *Artif. Intell. Rev.* **1990**, *4*, 251–277. [CrossRef]
2. VanLehn, K. The Relative Effectiveness of Human Tutoring, Intelligent Tutoring Systems, and Other Tutoring Systems. *Educ. Psychol.* **2011**, *46*, 197–221. [CrossRef]
3. Ahmad, K.; Iqbal, W.; El-Hassan, A.; Qadir, J.; Benhaddou, D.; Ayyash, M.; Al-Fuqaha, A. Data-Driven Artificial Intelligence in Education: A Comprehensive Review. *IEEE Trans. Learn. Technol.* **2024**, *17*, 12–31. [CrossRef]
4. Yan, L.; Greiff, S.; Teuber, Z.; Gašević, D. Promises and Challenges of Generative Artificial Intelligence for Human Learning. *Nat. Hum. Behav.* **2024**, *8*, 1839–1850. [CrossRef] [PubMed]
5. Bandi, A.; Adapa, P.V.S.R.; Kuchi, Y.E.V.P.K. The Power of Generative AI: A Review of Requirements, Models, Input–Output Formats, Evaluation Metrics, and Challenges. *Future Internet* **2023**, *15*, 260. [CrossRef]
6. Su, J.; Yang, W. Unlocking the Power of ChatGPT: A Framework for Applying Generative AI in Education. *ECNU Rev. Educ.* **2023**, *6*, 355–366. [CrossRef]
7. Qadir, J. Engineering Education in the Era of ChatGPT: Promise and Pitfalls of Generative AI for Education. In Proceedings of the 2023 IEEE Global Engineering Education Conference (EDUCON), Salmiya, Kuwait, 1–4 May 2023; pp. 1–9.
8. Sedkaoui, S.; Benaichouba, R. Generative AI as a Transformative Force for Innovation: A Review of Opportunities, Applications and Challenges. *Eur. J. Innov. Manag.* **2024**. *ahead-of-print*. [CrossRef]
9. OpenAI; Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F.L.; Almeida, D.; Altenschmidt, J.; Altman, S.; et al. GPT-4 Technical Report. *arXiv* **2024**, arXiv:2303.08774.
10. Sanderson, K. GPT-4 Is Here: What Scientists Think. *Nature* **2023**, *615*, 773. [CrossRef] [PubMed]
11. Jandhyala, V.S.V. GPT-4 and Beyond: Advancements in AI Language Models. *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol.* **2024**, *10*, 274–285. [CrossRef]
12. Akyuz, Y. Effects of Intelligent Tutoring Systems (ITS) on Personalized Learning (PL). *Creat. Educ.* **2020**, *11*, 953–978. [CrossRef]

13. Marouf, A.; Al-Dahdooh, R.; Ghali, M.J.A.; Mahdi, A.O.; Abunasser, B.S.; Abu-Naser, S.S. Enhancing Education with Artificial Intelligence: The Role of Intelligent Tutoring Systems. *Int. J. Eng. Inf. Syst. IJEAIS* **2024**, *8*, 10–16.
14. Guo, S.; Zheng, Y.; Zhai, X. Artificial Intelligence in Education Research during 2013–2023: A Review Based on Bibliometric Analysis. *Educ. Inf. Technol.* **2024**, *29*, 16387–16409. [CrossRef]
15. Ahmed, F.; Shubeck, K.; Hu, X. Chatgpt in the Generalized Intelligent Framework for Tutoring. In Proceedings of the 11th Annual Generalized Intelligent Framework for Tutoring (GIFT) Users Symposium (GIFTSym11), Orlando, FL, USA, 1 July 2023; p. 109.
16. Schmucker, R.; Xia, M.; Azaria, A.; Mitchell, T. Ruffle &Riley: Insights from Designing and Evaluating a Large Language Model-Based Conversational Tutoring System. In Proceedings of the Artificial Intelligence in Education, Recife, Brazil, 1 July 2024; Olney, A.M., Chounta, I.-A., Liu, Z., Santos, O.C., Bittencourt, I.I., Eds.; Springer Nature Switzerland: Cham, Switzerland, 2024; pp. 75–90.
17. Abu-Rasheed, H.; Abdulsalam, M.H.; Weber, C.; Fathi, M. Supporting Student Decisions on Learning Recommendations: An LLM-Based Chatbot with Knowledge Graph Contextualization for Conversational Explainability and Mentoring. *arXiv* **2024**, arXiv:2401.08517.
18. Dan, Y.; Lei, Z.; Gu, Y.; Li, Y.; Yin, J.; Lin, J.; Ye, L.; Tie, Z.; Zhou, Y.; Wang, Y.; et al. EduChat: A Large-Scale Language Model-Based Chatbot System for Intelligent Education. *arXiv* **2023**, arXiv:2308.02773.
19. Dai, W.; Lin, J.; Jin, H.; Li, T.; Tsai, Y.-S.; Gašević, D.; Chen, G. Can Large Language Models Provide Feedback to Students? A Case Study on ChatGPT. In Proceedings of the 2023 IEEE International Conference on Advanced Learning Technologies (ICALT), Orem, UT, USA, 10–13 July 2023; pp. 323–325.
20. Lin, J.; Han, Z.; Thomas, D.R.; Gurung, A.; Gupta, S.; Aleven, V.; Koedinger, K.R. How Can I Get It Right? Using GPT to Rephrase Incorrect Trainee Responses. *Int. J. Artif. Intell. Educ.* **2024**. [CrossRef]
21. Zhang, L.; Lin, J.; Borchers, C.; Sabatini, J.; Hollander, J.; Cao, M.; Hu, X. Predicting Learning Performance with Large Language Models: A Study in Adult Literacy. In Proceedings of the Adaptive Instructional Systems, Washington, DC, USA, 29 June–4 July 2024; Sottilare, R.A., Schwarz, J., Eds.; Springer Nature Switzerland: Cham, Switzerland, 2024; pp. 333–353.
22. Kurni, M.; Mohammed, M.S.; Srinivasa, K.G. Intelligent Tutoring Systems. In *A Beginner's Guide to Introduce Artificial Intelligence in Teaching and Learning*; Kurni, M., Mohammed, M.S., Srinivasa, K.G., Eds.; Springer International Publishing: Cham, Switzerland, 2023; pp. 29–44. ISBN 978-3-031-32653-0.
23. Pham, T.; Nguyen, T.B.; Ha, S.; Ngoc, N.T.N. Digital Transformation in Engineering Education: Exploring the Potential of AI-Assisted Learning. *Australas. J. Educ. Technol.* **2023**, *39*, 1–19. [CrossRef]
24. Kumar, P. Development of Intelligent Tutoring System Framework For Game-Based Learning. Master's Thesis, IIT Bombay, Mumbai, India, 2012.
25. Cabada, R.Z.; Barrón Estrada, M.L.; González Hernández, F.; Oramas Bustillos, R. Intelligent Tutoring System with Affective Learning for Mathematics. In Proceedings of the Human-Inspired Computing and Its Applications, Tuxtla, Mexico, 5 November 2014; Gelbukh, A., Espinoza, F.C., Galicia-Haro, S.N., Eds.; Springer International Publishing: Cham, Switzerland, 2014; pp. 483–493.
26. Schez-Sobrino, S.; Gmez-Portes, C.; Vallejo, D.; Glez-Morcillo, C.; Redondo, M.Á. An Intelligent Tutoring System to Facilitate the Learning of Programming through the Usage of Dynamic Graphic Visualizations. *Appl. Sci.* **2020**, *10*, 1518. [CrossRef]
27. Zografos, G.; Moussiades, L. A GPT-Based Vocabulary Tutor. In Proceedings of the Augmented Intelligence and Intelligent Tutoring Systems, Corfu, Greece, 2–5 June 2023; Frasson, C., Mylonas, P., Troussas, C., Eds.; Springer Nature Switzerland: Cham, Switzerland, 2023; pp. 270–280.
28. Chrysafiadi, K.; Virvou, M.; Tsihrintzis, G.A.; Hatzilygeroudis, I. Evaluating the User's Experience, Adaptivity and Learning Outcomes of a Fuzzy-Based Intelligent Tutoring System for Computer Programming for Academic Students in Greece. *Educ. Inf. Technol.* **2023**, *28*, 6453–6483. [CrossRef]
29. Kulik, J.A.; Fletcher, J.D. Effectiveness of Intelligent Tutoring Systems: A Meta-Analytic Review. *Rev. Educ. Res.* **2016**, *86*, 42–78. [CrossRef]
30. Mousavinasab, E.; Zarifsanaiey, N.; Niakan Kalhori, S.R.; Rakhshan, M.; Keikha, L.; Ghazi Saeedi, M. Intelligent Tutoring Systems: A Systematic Review of Characteristics, Applications, and Evaluation Methods. *Interact. Learn. Environ.* **2021**, *29*, 142–163. [CrossRef]
31. Zhang, L.; Lin, J.; Kuang, Z.; Xu, S.; Hu, X. SPL: A Socratic Playground for Learning Powered by Large Language Model. *arXiv* **2024**, arXiv:2406.13919.
32. Walkington, C.A. Using Adaptive Learning Technologies to Personalize Instruction to Student Interests: The Impact of Relevant Contexts on Performance and Learning Outcomes. *J. Educ. Psychol.* **2013**, *105*, 932–945. [CrossRef]
33. Bernacki, M.L.; Greene, M.J.; Lobczowski, N.G. A Systematic Review of Research on Personalized Learning: Personalized by Whom, to What, How, and for What Purpose(s)? *Educ. Psychol. Rev.* **2021**, *33*, 1675–1715. [CrossRef]
34. Xie, H.; Chu, H.-C.; Hwang, G.-J.; Wang, C.-C. Trends and Development in Technology-Enhanced Adaptive/Personalized Learning: A Systematic Review of Journal Publications from 2007 to 2017. *Comput. Educ.* **2019**, *140*, 103599. [CrossRef]
35. Gliner, J.A.; Morgan, G.A.; Leech, N.L. *Research Methods in Applied Settings: An Integrated Approach to Design and Analysis, Second Edition*, 2nd ed.; Routledge: New York, NY, USA, 2009; ISBN 978-0-203-84310-9.
36. Sajja, R.; Sermet, Y.; Cikmaz, M.; Cwiertny, D.; Demir, I. Artificial Intelligence-Enabled Intelligent Assistant for Personalized and Adaptive Learning in Higher Education. *Information* **2024**, *15*, 596. [CrossRef]

37. Woolf, B.P. *Building Intelligent Interactive Tutors: Student-Centered Strategies for Revolutionizing e-Learning*; Morgan Kaufmann: Burlington, MA, USA, 2010.
38. Chen, S.; Xu, X.; Zhang, H.; Zhang, Y. Roles of ChatGPT in Virtual Teaching Assistant and Intelligent Tutoring System: Opportunities and Challenges. In Proceedings of the 2023 5th World Symposium on Software Engineering, Tokyo, Japan, 22–24 September 2023; Association for Computing Machinery: New York, NY, USA, 26 December 2023; pp. 201–206.
39. Steenbergen-Hu, S.; Cooper, H. A Meta-Analysis of the Effectiveness of Intelligent Tutoring Systems on College Students' Academic Learning. *J. Educ. Psychol.* **2014**, *106*, 331–347. [CrossRef]
40. Alshahrani, A. The Impact of ChatGPT on Blended Learning: Current Trends and Future Research Directions. *Int. J. Data Netw. Sci.* **2023**, *7*, 2029–2040. [CrossRef]
41. Wang, J. The Research about the Innovative Application in Education Field Based on ChatGPT Foundation Model. *Adult High. Educ.* **2023**, *5*, 127–132. [CrossRef]
42. Bekeš, E.R.; Galzina, V. Exploring the Pedagogical Use of AI-Powered Chatbots Educational Perceptions and Practices. In Proceedings of the 2023 46th MIPRO ICT and Electronics Convention (MIPRO), Opatija, Croatia, 22–26 May 2023; pp. 636–641.
43. Stahl, B.C.; Eke, D. The Ethics of ChatGPT—Exploring the Ethical Issues of an Emerging Technology. *Int. J. Inf. Manag.* **2024**, *74*, 102700. [CrossRef]