SPL: A Socratic Playground for Learning Powered by Large Language Model*

Liang Zhang[1,2], Jionghao Lin[3], Ziyi Kuang[4,*], Sheng Xu[5] and Xiangen Hu[6]

[1]Institute for Intelligent Systems, University of Memphis, Memphis, TN 38152, USA
[2]Department of Electrical and Computer Engineering, University of Memphis, Memphis, TN 38152, USA
[3]School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, 15213, USA
[4]School of Psychology, Shaanxi Normal University, Xi'an, 710062, PR China
[5]School of Psychology, Central China Normal University, Wuhan, 430079, PR China
[6]Department of Applied Social Sciences, Hong Kong Polytechnic University, Hong Kong, PR China

Abstract

Dialogue-based Intelligent Tutoring Systems (ITSs) have significantly advanced adaptive and personalized learning by automating sophisticated human tutoring strategies within interactive dialogues. However, replicating the nuanced patterns of expert human communication remains a challenge in Natural Language Processing (NLP). Recent advancements in NLP, particularly Large Language Models (LLMs) such as OpenAI's GPT-4, offer promising solutions by providing human-like and context-aware responses based on extensive pre-trained knowledge. Motivated by the effectiveness of LLMs in various educational tasks (e.g., content creation and summarization, problem-solving, and automated feedback provision), our study introduces the Socratic Playground for Learning (SPL), a dialogue-based ITS powered by the GPT-4 model, which employs the Socratic teaching method to foster critical thinking among learners. Through extensive prompt engineering, SPL can generate specific learning scenarios and facilitates efficient multi-turn tutoring dialogues. The SPL system aims to enhance personalized and adaptive learning experiences tailored to individual needs, specifically focusing on improving critical thinking skills. Our pilot experimental results from essay writing tasks demonstrate SPL has the potential to improve tutoring interactions and further enhance dialogue-based ITS functionalities. Our study, exemplified by SPL, demonstrates how LLMs enhance dialogue-based ITSs and expand the accessibility and efficacy of educational technologies.

Keywords

Large Language Model, Socratic Teaching Method, Dialogue-based Intelligent Tutoring System, Prompt Engineering

1. Introduction

Dialogue-based Intelligent Tutoring Systems (ITSs) leverage artificial intelligence to simulate human-like tutoring through interactive dialogues [1, 2]. These systems aim to provide personalized and adaptive learning experiences by engaging learners in conversation, such as asking questions and providing feedback, and guiding them towards the expected learning goals. Over the past three decades, dialogue-based ITSs have demonstrated effectiveness in supporting learning, particularly in STEM subjects [3] as well as in reading and language learning [4, 1, 2]. However, dialogue-based ITSs can still be improved by incorporating more human-like guidance (e.g., effective tutoring strategies and polite language [2, 5, 6]), which underscores the importance of fully replicating the nuanced patterns of expert human tutoring communication within ITSs. In this context, advancements in large language models (LLMs) offer promising solutions, such as in-context learning and detailed feedback, for enhancing the quality of instruction [7, 8, 9].

LLMs, such as OpenAI's GPT-4 [10], are pre-trained on extensive datasets and can generate human-like dialogue when properly prompted. These models leverage their vast knowledge base to exhibit human-like reasoning and deliver insightful responses in natural language [11]. A critical technique for maximizing the capabilities of LLMs is prompt engineering, which includes methods like chain-of-thought (CoT) prompting [12] and few-shot prompting [11]. These methods enhance the models' ability to replicate human interaction and provide more adaptive text generation. Previous research has highlighted the promise of LLMs in improving various educational tasks, including providing better feedback [13, 9], enhancing learning guidance and interaction strategies [14], understanding student behaviors [15], and stimulating tutoring dialogues through answer evaluation and content generation [16].

Inspired by the potential of LLMs in education, our study introduces a dialogue-based Intelligent Tutoring System (ITS) named the Socratic Playground for Learning (SPL)[1], which simulates the Socratic teaching method in specific learning scenarios. SPL guides learners to solve questions by fostering self-reflection, critical thinking, and the development of independent thinking skills through interactive dialogue [17, 18]. Leveraging the capability of GPT models with advanced prompt engineering, SPL aims to deliver adaptive and flexible learning experience that can adjust to various educational contexts and learner profiles. Figure 1 illustrates an example of user interface for SPL dialogue, designed to enhance English proficiency for learners by applying second language learning principles. The interface features a menu with multiple selectable learning principles (e.g., Zone of Proximal Development) in the left-side column and five types of wh-questions (What?, Why?, How?, Who?, When?) at the top. This design enables learners to engage in interactive dialogues that promote critical thinking and language acquisition. The initial dialogue developed by the SPL system begins with the question "*How do you think this method could be applied to exchange students enhancing their English proficiency within the context of second language learning theories such as The Input Hypothesis?*". This is followed by a multi-turn dialogue with additional prompt wh-questions to further stimulate the learner's crit-

Educational Data Mining 2024 Workshop: Leveraging Large Language Models for Next Generation Educational Technologies, July 14, 2024, Atlanta, Georgia, USA
*Corresponding author.
lzhang13@memphis.edu (L. Zhang); jionghao@cmu.edu (J. Lin); ziyikuangzyk@gmail.com (Z. Kuang); willxusheng@gmail.com (S. Xu); xiangen.hu@polyu.edu.hk (X. Hu)
0009-0002-0017-2569 (L. Zhang); 0000-0003-3320-3907 (J. Lin)

[1]SPL Platform: https://polyu.skoonline.org/

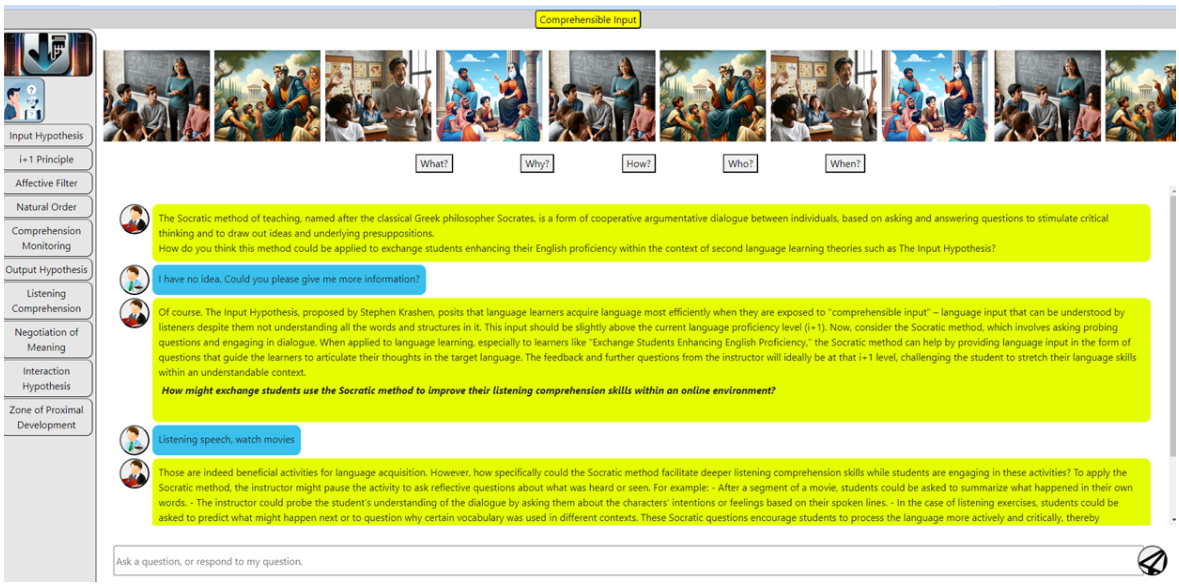

Figure 1: An Example Interactive Dialogue Interface of SPL System.

ical thinking.

Our study introduces the Socratic Playground for Learning (SPL) system, which employs GPT-4-based prompt strategies to create personalized learning scenarios grounded in the Socratic teaching method, enhancing dialogue-driven educational interactions. SPL demonstrates a significant enhancement over traditional dialogue-based ITSs by automating lesson design for specific learning scenarios and utilizing sophisticated NLP capabilities for multi-turn dialogue tutoring, thereby reducing reliance on human effort and predefined rules. Our preliminary evaluation of the SPL system's capabilities was conducted using essay writing tasks with college students. The results demonstrate the positive impact of the system's effective use of LLM in facilitating learning through the Socratic teaching method, promoting both critical thinking and deeper comprehension. Additionally, SPL provides adaptive and flexible learning experiences, increasing scalability and enabling the system to adjust to various educational contexts and learner profiles, thus broadening its potential for widespread adoptions in AI-based education.

2. Related Work

2.1. Dialogue-based Intelligent Tutoring Systems

Dialogue-based ITSs have proven to be effective in fostering cognitive engagement and improving learning outcomes by utilizing conversational interactions modeled on the best practices of human tutors [19, 20]. Since the development of the early SCHOLAR tutor by Carbonell in 1970, which offered Socratic tutoring through natural language text input and output [21], dialogue-based ITSs have employed mixed-initiative dialogues, semantic networks, question-answering, and tailored feedback to enhance learning. The SCHOLAR system encouraged learners to both ask and answer questions, providing feedback based on their responses to guide them toward the correct answers. Despite these advances, fully replicating all the capabilities of a human tutor remains a distant goal due to persistent challenges in natural language processing techniques [2].

The development of AutoTutor marked a significant advancement in dialogue-based ITSs by incorporating tutoring strategies derived from human tutoring protocols [22]. AutoTutor poses questions and problems from a curriculum script, understands learner inputs entered via keyboard, generates tutoring strategies in response (such as brief feedback, prompts, elaborations, corrections, and hints), and presents these strategies through a talking head [22, 4]. The dialogue structure in AutoTutor is guided by the expectation-and-misconception-tailored (EMT) dialogue rule, a pedagogical method for scaffolding student answers. Later on, many dialogue-based ITSs have been developed for diverse subjects. For example, Why2-Atlas is a natural language-based ITS for qualitative physics that uses deep syntactic analysis and abductive theorem proving to identify and address misconceptions in students' explanatory essays through dialogue-based feedback [23]. The Geometry Explanation Tutor engages students in dialogue-based self-explanation to improve their understanding and articulation of geometry rules [24]. DeepTutor is a conversational ITS that aligns assessment, learning progressions, and instructional tasks to guide students through conceptual physics problems with personalized instruction and feedback [25, 26].

2.2. LLMs for Enhancing ITSs

Large language models, such as ChatGPT, have brought opportunities to the ITS community in areas such as lesson design, feedback generation, and assessment of learner knowledge mastery. Ahmed [27] explored the potential of ChatGPT for conversation design and assessment in a course, facilitating the Generalized Intelligent Framework for Tutoring (GIFT) and reducing the effort required to design EMT conversation scripts. The conversational tutoring system Ruffle & Riley, developed by Schmucker et al., automatically generates tutoring scripts from lesson texts using GPT-4 to accelerate content authoring and employs the EMT-based rules to facilitate free-form conversational tutoring [28, 29]. Abu-Rasheed et al. [30] proposed an LLM-

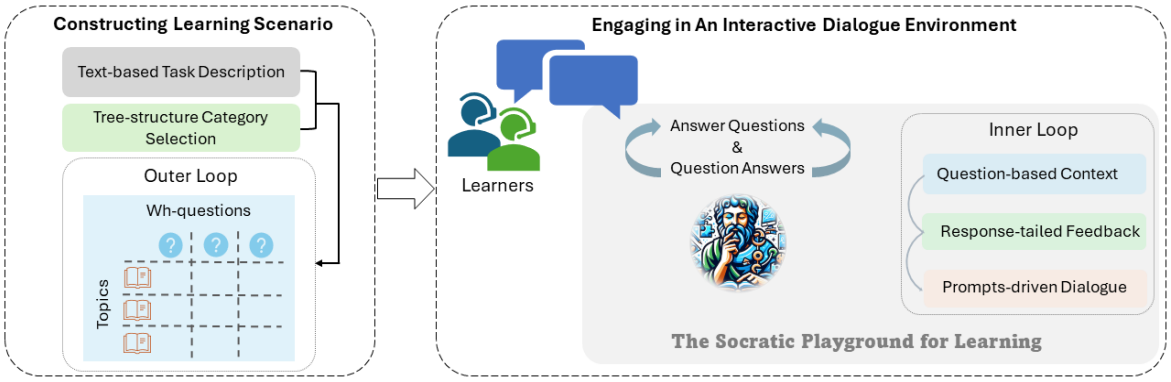

Figure 2: The Socratic Playground for Learning System Architecture.

based chatbot that engages students in conversation, similar to a discussion with a peer or mentor, augmented with knowledge graphs and human mentorship, enhancing conversational explainability (e.g., clarifying the reasons behind specific content suggestions) and mentoring in educational recommendations. Dan et al. [31] developed EduChat, a large-scale language model-based chatbot system for intelligent education that provides personalized, comprehensive and timely support for teachers, students, and parents by integrating retrieval-augmented question-answering, essay assessment, Socratic teaching, and emotional support to facilitate personalized and compassionate learning, leveraging pre-trained knowledge from educational and psychological domains. Nye et al. [16] highlighted opportunities for enhancing educational experiences through content generation with LLMs, while also addressing concerns around inaccuracies and equitable access.

Recent advancements in large language models (LLMs) have driven further innovation in education. Dai et al. [13, 32] demonstrated that GPT models could automate students' performance assessment and feedback generation in a manner more readable than that of human tutors. Lin et al. [33] developed a GPT-4-powered feedback system that that provides explanatory feedback by identifying trainees' responses as desired or undesired and automatically generating template-based feedback, with the GPT-4 model rephrasing incorrect responses to ensure clarity and understanding. Zhang et al. [34] explored the potential of LLMs in predicting learning performance, finding that they outperform traditional knowledge tracing methods in predictive accuracy in the context of adult literacy.

3. Socratic Playground for Learning (SPL)

The SPL is an LLM-powered, dialogue-based ITS designed to facilitate in-context learning through the Socratic teaching method [35]. It uses standard prompt strategies for lesson creation and Socratic dialogue to stimulate critical thinking and uncover underlying ideas and assumptions. The SPL offers personalized, adaptive, and flexible learning experiences that promote self-reflection and the development of critical thinking skills in learners.

3.1. System Architecture

As illustrated in Figure 2, the SPL system architecture supports usage in two main stages: (1) constructing learning scenarios and (2) engaging in an interactive dialogue environment.

Constructing Learning Scenarios. This system allows both educators and learners to easily and automatically construct personalized learning scenarios. Users can create scenarios through text-based task descriptions or by selecting options from a tree-structured format, which includes hierarchical categories in up-down relations such as domain, subdomain, objective, context, concepts, target learners, environments, and tutoring pedagogies. For example, a user might describe their learning request as: "*I am John, struggling with time management affecting grades. Any tips on effective time management would be welcome.*" Alternatively, they can as select from a tree structure: choosing "Psychology" as the domain, "Educational Psychology" as the subdomain, and setting the goal as "To understand the impact of motivation on student learning.". Based on this input, the system, powered by LLM (GPT-4), generates a list of learning contexts, such as "Explore the role of extrinsic rewards in student motivation.". The system then generates a corresponding list of concepts, such as the "Behavior Reinforcement", to be chosen as main focus. "College Students" is chosen from the target learner list, identifying the primary audience, and "Online Discussions" is selected from the learning environments list, indicating the mode of interaction. Finally, users can select from the list of pedagogical strategies, such as the "Socratic Method", though other methods like BLOOM (tutoring concepts/skills at all 6 levels of Bloom), TIMSS (Trends in International Mathematics and Science Study that tutoring based on different cognitive domains), Game-based learning (e.g., Who wants to be Millionaire), and Teachable Agents (OpenAI needs your help to understand the concepts) are also available. The up-down tree structure category selection process provides the necessary information, knowledge, and background to facilitate the automatic construction of learning scenarios. The established scenarios displayed on the SPL user interface showcase a matrix format of knowledge components or topics derived from input information and wh-questions (e.g., "what?", "why?", "how?", "who?", "when?") [36, 37]. By selecting and integrating these two dimensions, use can refine creation process, initiate the defined learning scenario, and start with specific questions to provoke dialogue, targeting their preferred knowledge areas (as illustrated in

the outer loop in Figure 2). For example the questions like "*What* effect do you think earning badges for your contributions might have on your motivation to participate in the online discussions?" and "*How* do you think earning badges for your contributions might impact your motivation to participate in the online discussions?".

Engaging in An Interactive Dialogue Environment. Within the dialogue interface, users engage in interactive dialogues driven by the Socratic method (as illustrated in the inner loop of the SPL architecture in Figure 2). The process includes: 1) *Question-based Context:* Initially, the SPL system kicks off the dialogue by presenting an example scenario or context related to the specific task or knowledge and posing a wh-question about that context. 2) *Learner Response-tailed Feedback:* After the user responds, the system captures the learners' responses or historical records, assesses their understanding, and provides timely feedback concluding with an another prompt question to encourage deeper thinking. 3) *Iterative Prompts-driven Dialogue:* The system persistently guides users through iterative prompts, deepening their thinking, correcting errors, and leading to correct solutions, providing a dynamic and interactive experience as a key pedagogical strategy within the dialogue-based environment. For example, considering the established learning scenario: "*Imagine a student named Taylor who has set a goal to improve their grades this semester. Taylor is exploring different motivational strategies to stay on track and achieve this goal. Chart your path to success by mastering the art of motivation. Let's embark on this journey together!*". The question "*What motivational strategies do you think Taylor could use to achieve their goal?*" kicks off the multiple-turn dialogue. If the user responds with "*I believe it requires hard work*", the system might reply, "*Absolutely, hard work is essential. But let's dive deeper into specific strategies that can help Taylor stay motivated.*". Here, the word "*Absolutely*" asserts agreement by assessing the user's response and provides positive feedback. This is further followed by the prompt question, "*What types of positive reinforcement could Taylor use to maintain their motivation and improve their grades?*". If the user's response is, "*I think some verbal praise and goal setting.*", the system would follow up with, "*Great start! Verbal praise and goal setting can be powerful motivators.*", and then prompt again, "*How do you think verbal praise can impact Taylor's motivation and academic performance?*". This approach both validates the user's response again and encourages deeper thinking and elaboration on "*how*" aspect. The iterative loop continues with diverse wh-questions, fostering critical thinking and deeper engagement for learners.

3.2. System Prompt Engineering

The entire process, including the scenario construction and multiple-turn interactive dialogue, is driven by GPT-4 based prompt engineering, which supports the Socratic teaching method for learning in the SPL. Several important nodes are described below:

Standard Prompt for Lesson Creation. This approach structures the creation of educational scenarios by starting with broad knowledge areas and refining them into specific sub-components. Leveraging GPT-4's reasoning, knowledge, prediction, and generative abilities, it transitions from general concepts to detailed elements essential for generating specific scenarios. This method effectively navigates complex information, facilitating the construction of learning scenarios, including role definitions, task clarifications, context setting, content specification, question generation, instructional resource preparation, pedagogical approach selection, and detailed scenario development. See the Table 1 as an example structure of a standard prompt template for lesson creation, demonstrating how broad concepts are refined into specific, actionable components. This systematic approach ensures clarity and precision in generating learning scenarios, fostering an effective and engaging learning environment. The prompt defines some variables, which are detailed below:

- `%[theLang]%` refers to the language (e.g., English, Chinese Mandarin, etc.) that will be displayed in the SPL learning scenario. This ensures that the content is accessible to learners in their preferred language.
- `%[theKC]%` refers to the knowledge components required to constitute the knowledge space for the domain-specific scenario. These components are essential elements or concepts that form the foundation of the subject matter.
- `%[theNumber]%` refers to the number of concepts needed for the creation of the learning scenario. This helps in defining the scope and depth of the learning material.
- `%[theDomain]%` refers to the domains (e.g., computer science, business, psychology, etc.) used in creating the specific learning scenario. This specifies the academic or professional field to which the learning scenario belongs.
- `%[theTarget]%` refers to the target learner group (e.g., college students, graduate students, online learners, etc.). This identifies the primary audience for the learning scenario, ensuring that the content is tailored to their needs and level of understanding.
- `%[theAvatar]%` refers to the avatar displayed in the user interface of the SPL dialogue. This personalized character can enhance engagement and provide a more interactive learning experience.
- `%[theTutorName]%` refers to the name that users prefer for the virtual tutor, adding a personalized touch to the tutoring experience.
- `%[theContext]%` refers to the context by topics for learning scenarios. This specifies the thematic areas or situations that the learning material will address.
- `%[theEnvironment]%` refers to the learning environment (e.g., online learning) used for learning engagement. This defines the setting in which the learning activities will take place, influencing the methods and tools used for instruction.
- `%[theUserName]%` refers to the user name for designing the SPL learning scenario. This personalizes the experience and can be used for tracking progress and providing feedback.
- `%[theType]%` refers to the style of pedagogical strategies for the learning scenario, e.g., Socratic method. This defines the instructional approach used to facilitate learning and ensure the material is effectively delivered.
- `%[theObjective]%` refers to the goal set for the learning, e.g., understanding the principles of working memory and understanding problem-solving strategies, in a specific learning subject.

Standard Prompt for Interactive Socratic Dialogue. The SPL employs a carefully designed prompt architecture powered by GPT-4, integrating the Socratic method to foster interactive and engaging tutoring. The design rules are implemented through prompt templates, which are dynamically updated based on the dialogue interactions. For

Table 1
Standard Prompt Template for Lesson Creation.

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).

more details, please refer to Table 2. The table outlines various prompt types involved the interaction process, their descriptions, and example wh-questions, showcasing how the system guides learners through context, feedback, and iterative questioning. This structured approach ensures a personalized and adaptive learning experience, encouraging critical thinking and reflection.

The SPL features a carefully crafted prompt architecture, incorporating both the standard for constructing learning scenarios and the interactive Socratic dialogue for fostering engaging and interactive tutoring. The system dynamically refines guidance based on user input and feedback, ensuring responses are aligned with the learner's needs and learning status. This exemplifies the innovative application of dialogue-based ITSs in education.

3.3. System Highlights

This system aims to provide learners with personalized, adaptive and flexible learning experiences. The main features of SPL include:

- **Personalization:** SPL creates personalized learning paths for learners, allowing them to explore different learning domains based on their interests. For example, a learner interested in psychology can choose specific topics like cognitive behavioral therapy or developmental psychology. The system traces the learner's responses and provides adaptive feedback with tailored prompt wh-questions.
- **Socratic Teaching:** The system employs the Socratic teaching method, encouraging learners to think critically, reflect, and explore concepts deeply by asking thought-provoking questions instead of directly providing answers. For instance, instead of explaining the principles of cognitive dissonance directly, SPL might ask, "What do you think happens when someone's actions contradict their beliefs?".
- **Interactivity:** SPL offers a dynamic and engaging learning experience through interactive dialogues, emulating the interactions that occur with human tutors. An example is a dialogue where the system asks, "How would you apply the concept of reinforcement in a classroom setting?" and provides feedback based on the learner's response.
- **Context-Sensitivity:** SPL generates rich problem scenarios (e.g., understanding the mechanisms of attention in psychology, understanding developmental milestones in early childhood education, understanding the architecture and functioning of computer processors in computer engineering) around key concepts or knowledge components (e.g., cognitive processes, developmental stages, computer architecture) and provides guidance and feedback based on the learners' responses.
- **Adaptability:** It adjusts tutoring strategies and content in response to the learner's progress and understanding, ensuring that the learning experience is continuously optimized. For instance, if a learner struggles with a particular psychology problem, SPL might provide additional questions and multi-turn dialogues to engage the learner further.
- **Cross-Domain Coverage:** The system supports learning across various domains, overcoming the limitations of many ITS that are restricted to specific fields. Examples include providing tailored content for subjects as diverse as computer science, business, engineering, psychology, nursing, mathematics, physics, and economics, etc.

These features foster advanced critical thinking, dynamic interactive learning, collaborative questioning, personalized learning journeys, comprehensive analytical skills, reflective metacognition, cross-disciplinary integration, and engaging motivational strategies.

4. System Evaluation Through Pilot Testing

To evaluate the SPL system's effectiveness in enhancing learner engagement, understanding, and satisfaction, we use one pilot study using the example task on essay writing.

4.1. Experimental Design

This pilot testing experiment involved 10 graduate-level participants recruited from the campus. Upon entering the laboratory, participants filled out demographic information and then engaged with the SPL system for dialogue-based communication on the topic of essay writing. Learners described their essay writing needs based on their field of study, for example: "*I am John, my major is Psychology, and I want to learn how to write empirical research papers. Please help me!*". The 10 survey questions focused on various aspects, including the effectiveness and fluency of dialogue (**Q1**), perception of human-like interaction (**Q2**), user enjoyment (**Q3**), attractiveness of learning methods (**Q4**), happiness with learning (**Q5**),understanding enhancement (**Q6**), learning motivation (**Q7**), improvement in learning outcomes (**Q8**), satisfaction of learning needs (**Q9**), willingness to recommend the system (**Q10**), along with two open-ended questions to gather feedback. Responses were collected using a 7-point Likert scale, ranging from 1 (strongly disagree)

Table 2
Standard Prompt for Interactive Socratic Dialogue.

Prompt Type	Description	Example Prompt WH-Question
Initial Context and Questioning	The system starts by presenting a scenario context and posing a wh-question to stimulate the learner's thinking.	"*What aspect of the context do you find most challenging to understand?*"
Response Evaluation and Feedback	The system evaluates responses and provides hints and feedback to guide learners toward correct understanding without directly giving answers.	"*How does this part of the context relate to the overall scenario?*"
Iterative Prompting	Through iterative prompts, the system deepens the learner's reasoning, encouraging detailed exploration and articulation.	"*Can you explain why this particular detail is significant in the scenario?*"
Feedback and Exploration	Feedback highlights correct elements and offers hints for further exploration.	"*What other factors might influence this outcome?*"
Maintaining Engagement	This approach maintains engagement through continuous, thought-provoking questions that connect new concepts to prior knowledge.	"*How would you connect this concept to what you have learned previously?*"
Fostering Critical Thinking	The system prompts learners to evaluate and critique their own responses, fostering critical thinking skills.	"*What could be a potential limitation of your current understanding?*"
Encouraging Reflection	It encourages learners to reflect on their learning process and outcomes.	"*How has your understanding changed after considering this question?*"
Providing Incremental Hints	The system offers incremental hints that build upon each other to guide the learner progressively towards deeper understanding.	"*What is a simpler way to think about this problem before tackling the more complex aspects?*"
Adaptive Feedback	The feedback adapts to the learner's responses, becoming more specific as the learner's understanding develops.	"*Given your explanation, what would be the next logical step to explore?*"
Encouraging Synthesis	It encourages learners to synthesize information from different parts of the scenario to form a comprehensive understanding.	"*How can you combine these different pieces of information to solve the problem?*"

to 7 (strongly agree). For more detailed survey questions, please refer to Appendix A.

The distribution of survey scores, along with their frequency percentages, was analyzed to assess various dimensions of user experience. Open-ended feedback was semantically annotated using ChatGPT (GPT-4) [38], with a network-based visualization highlighting similar semantic themes (using the NetworkX python package) [39]. ChatGPT facilitated the precise annotation of each feedback entry, capturing the essence and complexity of responses, and identifying shared themes across different answers. This comprehensive approach allowed for a nuanced understanding of participant feedback, enhancing the evaluation of the SPL system's performance and user satisfaction.

5. Results

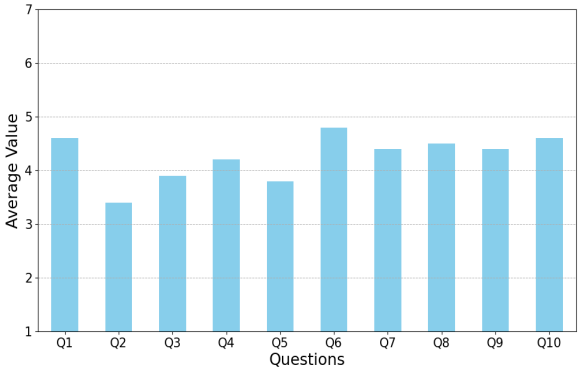

Figure 3: Average Scores for **Q1** to **Q10**.

Figure 3 shows the average scores for **Q1** to **Q10** from the 7-point Likert scale survey conducted during the pilot experiment. The results indicate positive responses for most questions, with average scores greater than 4 for **Q1**, **Q4**, **Q6**, **Q7**, **Q8**, **Q9**, and **Q10**. These high scores suggest that participants found the system effective and engaging, particularly in terms of effectiveness and fluency of dialogue, attractiveness of learning methods, and enhancement of understanding. Additionally, the questions related to learning motivation, improvement in learning outcomes, satisfaction of learning needs, and willingness to recommend the system also received positive feedback, indicating overall satisfaction with the system's performance in these areas. Conversely, the relatively lower scores for **Q2**, **Q3**, and **Q5**, which are below 4, highlight potential areas for improvement. These questions pertain to the perception of human-like interaction, user enjoyment, and happiness with learning. The lower scores in these areas suggest that while the system is effective in delivering content and enhancing understanding, there may be a need to enhance the interactive and enjoyable aspects of the system to better engage users and make the learning experience more pleasurable. For a detailed breakdown of the scores from **Q1** to **Q10**, please refer to Figure 5 in the Appendix B.

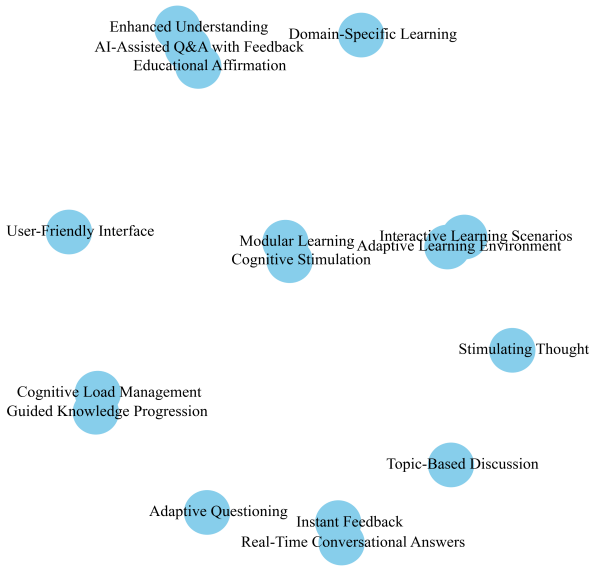

Figure 4: The Semantics Visualization for Q11 about Favourite Features of SPL.

Figure 4 presents the semantic annotation results and network visualization for the open question **Q11**, which investigates users' favorite features of the SPL system. Results reveal that most participants expressed positive sentiments about the system's features. Notably, the "AI-Assisted Q&A with Feedback", "Enhanced Understanding", and "Educational Affirmation" were highlighted as key advantages, underscoring the favored aspects of AI integration in learning. The other feedback also showcases the system's various strengths and benefits, with "Interactive Learning Scenarios" and "Adaptive Learning Environment" emphasize the system's flexibility and engagement, and "Guided Knowledge Progression" and "Cognitive Load Management" highlight its efficiency in organizing the learning process. Additionally, "Real-Time Conversational Answers" and "Instant Feedback" are highlighted for enhancing the interactive experience. The "User-Friendly Interface" and "Topic-Based Discussion" further improve the learning environment's usability. Overall, the feedback highlights the system's ability to deliver adaptive learning experiences, with particular appreciation for its AI-driven features and user-friendly design.

As for the collected feedback on question **Q12**, participants emphasized the need for enhanced system performance, improved guidance, and greater clarity in the user interface. Specific issues identified included slow response times and high latency. Recommendations for improvement included the incorporation of explanatory videos and more intuitive navigation tips. Users also stressed the importance of clearer icons, such as labeled buttons, to enhance usability. They suggested that AI responses should more effectively align with previous interactions and the system should be more accessible to beginners by clearly presenting essential instructional guidance.

6. Discussion

Our preliminary evaluation of the SPL system has yielded promising results, demonstrating positive aspects of learner engagement and leaning experience. Participants particularly valued the system's effective use of AI to facilitate learning through the Socratic method, which could promote critical thinking and deeper comprehension.

The system engages learners through interactive conversational process that deepens understanding, corrects misconceptions, and guides them towards their learning goals. This process is well-aligned with the expectation-misconception tailoring (EMT) principles [40, 41, 27], which are designed to address and rectify learners' misconceptions effectively. By persistently guiding users with targeted prompts and feedback, the SPL system has the potential to reinforce learners' knowledge, enhance problem-solving skills, and boost their confidence.

The SPL system enhances its educational interactions by leveraging GPT-4's capabilities. The system employs standard prompts for leasson creation and Socratic dialogue. The prompt for lesson creation organize educational scenarios by starting with broad topics and narrowing them down into specific details. This method utilizes GPT-4's strengths in reasoning, prediction, and generation to transition from general ideas to detailed learning scenarios, creating comprehensive and coherent learning experiences. Meanwhile, the prompt for interactive Socratic dialogue are carefully crafted to facilitate engaging tutoring sessions. These prompts are dynamically updated based on the flow of dialogue, ensuring that the system's responses are tailored to the learner's current level of understanding and learning needs. Thus, the SPL system has the potential to deliver personalized and adaptive learning experiences while ensuring that the educational content is contextually appropriate.

6.1. Limitations

The SPL system faces several limitations that impact its overall performance and user experience. A primary concern is the time latency associated with the ChatGPT API, which can hinder the responsiveness of the system. Additionally, the implementation of learning pathways guided by the EMT approach is still limited, affecting the system's ability to fully support learners in achieving their learning goals. Another significant challenge is the hallucination issue, where the system may produce responses that are seemingly plausible but incorrect or nonsensical. Moreover, enhancing domain-specific teaching through retrieval-augmented generation remains an area for further development. Achieving truly human-like dialogue also remains difficult, with ongoing issues related to the smoothness of conversational turn-taking and latency in real-time feedback [42, 43]. Further refinement of GPT-4 based prompt templates is needed to better assess learner states, capture responses, and compile specific knowledge source. Additionally, there is a need for generative AI models to trace and predict learner performance while exploring individual differences, building on the progress made in our previous work. [34, 44, 45].

6.2. Future Works

To enhance the SPL system, several future work directions are proposed to improve domain-specific learning design and overall functionality. Firstly, pre-training will be utilized to incorporate expertise from educational practices and professional criteria, guiding the system towards more formal and specialized teaching methodologies. We are also exploring a scalable learning model to extend SPL as a general learning framework, capable of encompassing diverse educational domains and scenarios. This also involves integrating multimedia elements to support multimodal learning through generative AI, enhancing the system's ability to deliver rich, interactive educational experiences.

Additionally, the development of multiple roles and agents using large language models (LLMs) will be pursued to create a more dynamic and versatile dialogue-based Intelligent Tutoring System (ITS). This will enable the system to simulate various educational roles and perspectives, providing a comprehensive learning environment.

For the evaluation of essay submissions, future work will focus on three key areas:

- Robustness of Essay Evaluation: We will assess the consistency of the evaluation process by repeatedly evaluating the same essay against a standardized rubric to ensure reliability.
- Sensitivity of the Evaluation: By systematically altering a well-written essay, we will evaluate how sensitive the rating system is to changes in the document, ensuring it can accurately reflect variations in quality.

- Psychometric Analysis of Evaluation Standards: Each evaluation standard will be treated as an individual "person" allowing us to analyze the effectiveness and consistency of each criterion.
- Evaluation of Standards: We will examine how different factors, such as the nature, type, and length of documents, influence the evaluation standards.

Following these evaluations, we aim to develop recommendations and potentially introduce a "grading wizard" as a user-friendly product, streamlining the grading process and enhancing user experience. This comprehensive approach aims to refine the SPL system's educational capabilities, making it more robust, sensitive, and adaptable to a wide range of learning and assessment scenarios.

7. Conclusion

In this study, we introduce the SPL system, powered by large language models (GPT-4), designed to enhance dialogue-based ITS through the Socratic method. The SPL system aims to provide personalized, adaptive, and flexible learning experiences that foster self-reflection, critical thinking, and independent thinking skills in learners. Leveraging GPT-4's prompt engineering capabilities, we employ a standard prompt for lesson creation and interactive Socratic dialogue to facilitate engaging and interactive tutoring. Preliminary pilot testing demonstrates the positive impact of SPL on learners, including increased engagement, enjoyment, and learning gains. The SPL system marks a significant improvement over traditional dialogue-based ITSs like SCHOLAR and AutoTutor, which depended on human effort for lesson design and predefined rules with limited NLP capabilities for multi-turn dialogue. Although this work is still in progress, it represents a promising step towards the next generation of dialogue-based ITS, encompassing lesson design, pedagogical strategy formulation, and the assessment of learner responses and feedback generation through generative AI.

8. Acknowledgments

We would like to thank Dr. Xiangen Hu from the Department of Applied Social Sciences at Hong Kong Polytechnic University for leading this collaborative project. Additionally, we express our gratitude to Arthur C. Graesser from the Institute for Intelligent Systems at the University of Memphis for providing theoretical support and valuable insights on multi-turn dialogues, which significantly contributed to this study.

References

[1] B. D. Nye, A. C. Graesser, X. Hu, Autotutor and family: A review of 17 years of natural language tutoring, International Journal of Artificial Intelligence in Education 24 (2014) 427–469.

[2] J. Paladines, J. Ramirez, A systematic literature review of intelligent tutoring systems with dialogue in natural language, IEEE Access 8 (2020) 164246–164267.

[3] A. C. Graesser, H. Li, Intelligent tutoring systems and conversational agents (2023).

[4] A. C. Graesser, K. VanLehn, C. P. Rosé, P. W. Jordan, D. Harter, Intelligent tutoring systems with conversational dialogue, AI magazine 22 (2001) 39–39.

[5] J. Lin, S. Singh, L. Sha, W. Tan, D. Lang, D. Gašević, G. Chen, Is it a good move? mining effective tutoring strategies from human–human tutorial dialogues, Future Generation Computer Systems 127 (2022) 194–207.

[6] J. Lin, M. Rakovic, D. Lang, D. Gasevic, G. Chen, Exploring the politeness of instructional strategies from human-human online tutoring dialogues, in: LAK22: 12th International Learning Analytics and Knowledge Conference, 2022, pp. 282–293.

[7] X. Xu, Y. Liu, P. Pasupat, M. Kazemi, et al., In-context learning with retrieved demonstrations for language models: A survey, arXiv preprint arXiv:2401.11624 (2024).

[8] J. Lin, Z. Han, D. R. Thomas, A. Gurung, S. Gupta, V. Aleven, K. R. Koedinger, How can i get it right? using gpt to rephrase incorrect trainee responses, arXiv preprint arXiv:2405.00970 (2024).

[9] J. Stamper, R. Xiao, X. Hou, Enhancing llm-based feedback: Insights from intelligent tutoring systems and the learning sciences, arXiv preprint arXiv:2405.04645 (2024).

[10] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al., Gpt-4 technical report, arXiv preprint arXiv:2303.08774 (2023).

[11] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Advances in neural information processing systems 33 (2020) 1877–1901.

[12] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al., Chain-of-thought prompting elicits reasoning in large language models, Advances in neural information processing systems 35 (2022) 24824–24837.

[13] W. Dai, J. Lin, H. Jin, T. Li, Y.-S. Tsai, D. Gašević, G. Chen, Can large language models provide feedback to students? a case study on chatgpt, in: 2023 IEEE International Conference on Advanced Learning Technologies (ICALT), IEEE, 2023, pp. 323–325.

[14] H. Kumar, I. Musabirov, M. Reza, J. Shi, A. Kuzminykh, J. J. Williams, M. Liut, Impact of guidance and interaction strategies for llm use on learner performance and perception, arXiv preprint arXiv:2310.13712 (2023).

[15] M. Park, S. Kim, S. Lee, S. Kwon, K. Kim, Empowering personalized learning through a conversation-based tutoring system with student modeling, arXiv preprint arXiv:2403.14071 (2024).

[16] B. D. Nye, D. Mee, M. G. Core, Generative large language models for dialog-based tutoring: An early consideration of opportunities and concerns, 2023.

[17] R. Yang, K. Narasimhan, The socratic method for self-discovery in large language models, Technical Report, tech. rep., Princeton NLP, 2023.

[18] C. Kong, F. Yaxin, X. Wan, F. Jiang, B. Wang, Platolm: Teaching llms via a socratic questioning user simulator (2023).

[19] S. Feng, A. J. Magana, D. Kao, A systematic review of literature on the effectiveness of intelligent tutoring systems in stem, in: 2021 IEEE frontiers in education conference (fie), IEEE, 2021, pp. 1–9.

[20] S. K. D'Mello, A. Graesser, Intelligent tutoring systems: How computers achieve learning gains that rival human tutors, in: Handbook of educational psychology, Routledge, 2023, pp. 603–629.
[21] J. R. Carbonell, Ai in cai: An artificial-intelligence approach to computer-assisted instruction, IEEE transactions on man-machine systems 11 (1970) 190–202.
[22] A. C. Graesser, K. Wiemer-Hastings, P. Wiemer-Hastings, R. Kreuz, T. R. Group, et al., Autotutor: A simulation of a human tutor, Cognitive Systems Research 1 (1999) 35–51.
[23] K. VanLehn, P. W. Jordan, C. P. Rosé, D. Bhembe, M. Böttner, A. Gaydos, M. Makatchev, U. Pappuswamy, M. Ringenberg, A. Roque, et al., The architecture of why2-atlas: A coach for qualitative physics essay writing, in: Intelligent Tutoring Systems: 6th International Conference, ITS 2002 Biarritz, France and San Sebastian, Spain, June 2–7, 2002 Proceedings 6, Springer, 2002, pp. 158–167.
[24] V. Aleven, A. Ogan, O. Popescu, C. Torrey, K. Koedinger, Evaluating the effectiveness of a tutorial dialogue system for self-explanation, in: Intelligent Tutoring Systems: 7th International Conference, ITS 2004, Maceió, Alagoas, Brazil, August 30-September 3, 2004. Proceedings 7, Springer, 2004, pp. 443–454.
[25] V. Rus, D. Stefanescu, N. Niraula, A. C. Graesser, Deeptutor: towards macro-and micro-adaptive conversational intelligent tutoring at scale, in: Proceedings of the first ACM conference on Learning@ scale conference, 2014, pp. 209–210.
[26] V. Rus, N. Niraula, R. Banjade, Deeptutor: An effective, online intelligent tutoring system that promotes deep learning, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 29, 2015.
[27] F. Ahmed, K. Shubeck, X. Hu, Chatgpt in the generalized intelligent framework for tutoring, in: Proceedings of the 11th Annual Generalized Intelligent Framework for Tutoring (GIFT) Users Symposium (GIFTSym11), US Army Combat Capabilities Development Command–Soldier Center, 2023, p. 109.
[28] R. Schmucker, M. Xia, A. Azaria, T. Mitchell, Ruffle&riley: Towards the automated induction of conversational tutoring systems, arXiv preprint arXiv:2310.01420 (2023).
[29] R. Schmucker, M. Xia, A. Azaria, T. Mitchell, Ruffle&riley: Insights from designing and evaluating a large language model-based conversational tutoring system, arXiv preprint arXiv:2404.17460 (2024).
[30] H. Abu-Rasheed, M. H. Abdulsalam, C. Weber, M. Fathi, 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).
[31] Y. Dan, Z. Lei, Y. Gu, Y. Li, J. Yin, J. Lin, L. Ye, Z. Tie, Y. Zhou, Y. Wang, A. Zhou, Z. Zhou, Q. Chen, J. Zhou, L. He, X. Qiu, Educhat: A large-scale language model-based chatbot system for intelligent education, 2023. arXiv:2308.02773.
[32] W. Dai, Y.-S. Tsai, J. Lin, A. Aldino, F. Jin, T. Li, D. Gasevic, et al., Assessing the proficiency of large language models in automatic feedback generation: An evaluation study (????).
[33] J. Lin, Z. Han, D. R. Thomas, A. Gurung, S. Gupta, V. Aleven, K. R. Koedinger, How can i get it right? using gpt to rephrase incorrect trainee responses, 2024. arXiv:2405.00970.
[34] L. Zhang, J. Lin, C. Borchers, J. Sabatini, J. Hollander, M. Cao, X. Hu, Predicting learning performance with large language models: A study in adult literacy, arXiv preprint arXiv:2403.14668 (2024).
[35] Xiangen Hu, FAQ About SPL, https://spl.skoonline.org/FAQ/index.html?lang=en, 2024.
[36] I. Koshik, Wh-questions used as challenges, Discourse Studies 5 (2003) 51–77.
[37] B. A. Fox, S. A. Thompson, Responses to wh-questions in english conversation, Research on Language and Social Interaction 43 (2010) 133–156.
[38] F. Gilardi, M. Alizadeh, M. Kubli, Chatgpt outperforms crowd workers for text-annotation tasks, Proceedings of the National Academy of Sciences 120 (2023) e2305016120.
[39] A. Hagberg, D. Conway, Networkx: Network analysis with python, URL: https://networkx. github. io (2020).
[40] A. C. Graesser, S. Lu, G. T. Jackson, H. H. Mitchell, M. Ventura, A. Olney, M. M. Louwerse, Autotutor: A tutor with dialogue in natural language, Behavior Research Methods, Instruments, & Computers 36 (2004) 180–192.
[41] A. C. Graesser, X. Hu, D. S. McNamara, Computerized learning environments that incorporate research in discourse psychology, cognitive science, and computational linguistics. (2005).
[42] K. Mitsui, Y. Hono, K. Sawada, Towards human-like spoken dialogue generation between ai agents from written dialogue, arXiv preprint arXiv:2310.01088 (2023).
[43] D. Adiwardana, M.-T. Luong, D. R. So, J. Hall, N. Fiedel, R. Thoppilan, Z. Yang, A. Kulshreshtha, G. Nemade, Y. Lu, et al., Towards a human-like open-domain chatbot, arXiv preprint arXiv:2001.09977 (2020).
[44] L. Zhang, J. Lin, C. Borchers, M. Cao, X. Hu, 3dg: A framework for using generative ai for handling sparse learner performance data from intelligent tutoring systems, arXiv preprint arXiv:2402.01746 (2024).
[45] L. Zhang, P. I. Pavlik Jr, X. Hu, J. L. Cockroft, L. Wang, G. Shi, Exploring the individual differences in multidimensional evolution of knowledge states of learners, in: International Conference on Human-Computer Interaction, Springer, 2023, pp. 265–284.

A. Survey Table

As shown in Table 3 , all the survey questions following the 7-point Likert scale are presented. These questions were designed to evaluate various aspects of the SPL system, including its effectiveness, user interaction, enjoyment, and overall satisfaction. By using a 7-point Likert scale, we aimed to capture a wide range of participant responses, from strong disagreement to strong agreement, providing a nuanced understanding of their experiences. Additionally, two open-ended questions were included to gather qualitative feedback, allowing participants to elaborate on their favorite features and provide suggestions for improvement. This comprehensive approach ensures a thorough evaluation of the SPL system from multiple perspectives.

Table 3
Survey Questions for SPL System Evaluation

Question No.	Survey Question
Q1	I believe the dialogue in the SPL is effective and smooth (1 = strongly disagree, 7 = strongly agree; same below).
Q2	I feel like I am interacting with a person in the SPL.
Q3	I enjoy learning in the SPL.
Q4	I find the learning methods provided by the SPL attractive.
Q5	I feel happy while learning in the SPL.
Q6	The SPL helps me understand the learning content.
Q7	I am motivated to learn in the SPL.
Q8	Learning in the SPL can improve my current knowledge performance.
Q9	I feel that the SPL meets my learning needs.
Q10	I am willing to recommend the SPL to others.
Q11	What is your favorite feature or function of the system?
Q12	What other feedback or suggestions do you have for the system?

B. Percentage Distribution of Scores for Q1 to Q10

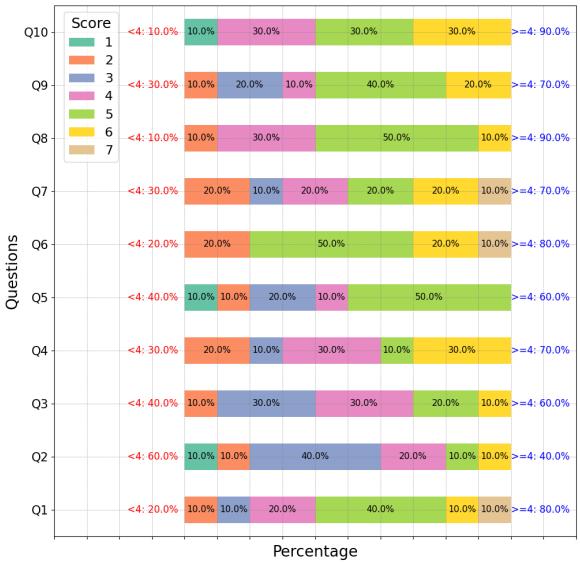

Figure 5: Percentage Distribution of Scores for Q1 to Q10.

Figure 5 shows the percentage distribution of scores for **Q1** to **Q10**, centered around a score of 4, the neutral point on the 7-point Likert scale. Generally, most participants indicated positive impacts of the SPL system, as the total percentage of scores above 4 exceeds those below 4, as shown by the labeled percentages on the left and right sides of the bars for each question. The results suggest that the system's performance is above average, with approximately 28% of scores below 4 and 72% of scores 4 or above. Additionally, the average score for each question is predominantly above 4, with the exception of **Q2**, which assesses the human-likeness of the SPL system, indicating it is perceived as less human-like.

Examining the individual questions, **Q8** (improvement in learning outcomes) and **Q10** (willingness to recommend the system) have the highest percentage (both are 90%) of scores in the 5-7 range, suggesting strong positive feedback in these areas. **Q6** (understanding enhancement) also shows 80% of responses in the positive range (50% at score 5, 20% at score 6, and 10% at score 7). On the other hand, **Q2** stands out with a significant portion of responses in the lower range (40% at score 3, 10% at score 2, and 10% at score 1), suggesting a critical area for improvement regarding the perceived human-likeness of the system.