LanPIP: A Pedagogically-Grounded AI Pipeline for Customizable Language Learning

Abstract

The integration of Large Language Models (LLMs) in language education has the potential to revolutionize teaching and learning practices. However, current AI-driven learning tools often lack grounding in well-established language acquisition theories and fail to consider real-world lesson planning. Additionally, most existing conversational agents are limited to handling only text-based interactions, while effective teaching practices require the integration of multiple modalities. Consequently, there is a scarcity of educational practices that successfully integrate AI with teacher guidance. To address these challenges, this study proposes a set of criteria for selecting language models and designing prompts to effectively incorporate AI in language education. Based on these criteria, we develop a framework that integrates autonomous teaching, teaching assistance, and self-learning facilitation functions. We also design a prototype, LanPIP, to demonstrate the practical application of the proposed criteria and framework. Through medium-scale user studies on each function of the prototype, we demonstrate its efficiency in enhancing language learning experiences. By considering educational theory, real classroom settings, and multimodal interactions, our study aims to bridge the gap between AI technology and effective educational practices, ultimately facilitating the successful integration of AI in language education with proper teacher guidance. This research contributes to the advancement of AI-assisted language education and provides valuable insights for developers, educators, and researchers in the field.

CCS Concepts

• **Human-centered computing → Human computer interaction (HCI)**.

Keywords

Human-Computer Interaction, Computer-Assisted Language Learning (CALL), Software Design, User Study

ACM Reference Format:

. 2024. LanPIP: A Pedagogically-Grounded AI Pipeline for Customizable Language Learning. In . ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Large Language Models (LLMs) have emerged as transformative tools in natural language processing (NLP) and language education. Their ability to understand and generate human-like text offers extended capacity for learners with various backgrounds.

The integration of AI in language education has revolutionized traditional teaching methods, offering personalized and adaptive learning experiences through various applications such as Intelligent Tutoring Systems (ITS), Natural Language Processing (NLP) tools, and conversational agents. ITS provides customized instruction and feedback based on individual needs and proficiency levels [56]. NLP technologies enable real-time feedback on grammar, vocabulary, and pronunciation [12, 27]. Conversational agents simulate real-life conversations and enhance learner engagement and motivation [15, 25, 51]. Adaptive learning platforms tailor content to each learner's specific needs [32]. AI can also foster collaborative learning environments by facilitating group interactions and providing feedback [22, 37]. While AI can be a valuable tool in education, simply implementing it to satisfy learners' immediate needs does not constitute an effective tutoring system. Developing an ITS that can effectively conduct tutoring or assist human teachers requires a more comprehensive approach that goes beyond merely satisfying learners' requests. Language acquisition is a complex process that involves fostering critical thinking, promoting authentic language practices, and providing timely constructive corrective feedback [34, 35]. To create a successful AI-driven ITS, one must consider the pedagogical implications and design the system to support meaningful learning experiences rather than focusing solely on content delivery or question answering [33].

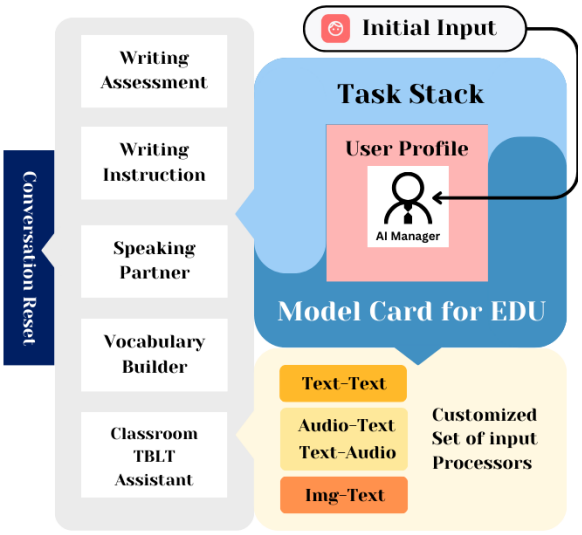

Figure 1: LanPIP Prototype Overview

Currently, there is a lack of commercially available or publicly accessible learning assisting tools that have been developed based on well-tested language acquisition theories and have taken actual lesson planning into consideration [18]. Furthermore, most existing conversational agents are limited to handling only one modality, typically text, while actual teaching practices often require the integration of multiple modalities tailored to the needs of learners in different situations. As a result, there have been few observed

Conference'30, Aug 2024, Barcelona, Spain

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

educational practices that effectively integrate AI with teacher guidance. The development of AI-driven educational tools should be grounded in pedagogical principles and designed to support real-world classroom dynamics. Incorporating multiple modalities, such as speech and visual aids, could enhance the effectiveness of these tools.

As a solution to the mentioned research gaps, our study makes the following contributions:

(1) We propose criteria for selecting language models for developers and designing prompts for educators and developers to effectively integrate AI in language education.
(2) We develop a framework that integrates autonomous teaching, teaching assistance, and self-learning facilitation five tasks based on the proposed criteria.
(3) We design a prototype LanPIP, as shown in Figure 1, with five different applications integrated for in and outside-classroom learning based on the proposed criteria and framework for user studies to demonstrate the practical application of these aspects.
(4) Medium-scale user studies were conducted on each function of the prototype and demonstrated the efficiency of the prototype.

2 Related Work

2.1 Large Language Models

Large Language Models (LLMs) have garnered significant attention in recent years due to their impressive capabilities in various natural language processing (NLP) tasks. Notable LLMs with billions of parameters include OpenAI's GPT series [10], Anthropic's Claude [4], Cohere's Command [19], Mixtral [2], Meta's LLaMA [54], and Google's Gemini [21]. Large Language Models (LLMs) exhibit several advantages that make them valuable tools for various natural language processing tasks. These models have demonstrated strong performance across a wide range of NLP tasks without task-specific fine-tuning, showcasing their ability to generalize [10]. Additionally, LLMs can be pre-trained on large corpora and then fine-tuned on specific tasks, leveraging transfer learning to significantly improve performance in various applications [31]. Moreover, models with billions of parameters utilize large-scale pre-training to understand and generate contextually relevant text, demonstrating their contextual understanding capabilities on various benchmarks [16, 59, 61], such as SuperGLUE [57], MMLU[29], HumanEval [17], etc. However, despite their advantages, LLMs also have several notable disadvantages that should be considered. Training and deploying LLMs require substantial computational resources and memory, which can limit their accessibility [50]. Furthermore, LLMs inherit biases present in their training data, which can lead to biased or harmful outputs [7], raising concerns about fairness and potential negative impacts. Additionally, LLMs sometimes generate hallucinated content [60], which can be misleading or inaccurate. Lastly, the decision-making processes of LLMs are often opaque, making it difficult to interpret how they arrive at specific outputs [44]. All these disadvantages need to be taken into account when designing a system for learners who lack the ability to recognize the errors made in their target learning language.

2.2 Language Pedagogical Approaches

Second language acquisition (SLA) has been extensively studied, leading to the development of various theories and methods. Language teaching methods have evolved significantly over time, reflecting changes in the understanding of how languages are best learned. The Grammar Translation Method, originating in the 19th century, focuses on translating sentences between the target and native languages, emphasizing grammar rules and vocabulary [42]. However, it has been criticized for neglecting speaking and listening skills [9]. The Direct Method, developed as a reaction to the Grammar Translation Method, emphasizes immersion in the target language without using the native language, focusing on oral skills and inductive grammar learning [53]. While effective in improving speaking abilities, it requires highly skilled teachers and may be challenging for learners without a solid grammatical foundation [52].

The Presentation, Practice, and Production (PPP) method, widely used today, involves presenting new language items, practicing them in a controlled context, and producing them in a more communicative setting [40]. Although effective for structured learning, it has been criticized for being too rigid and not reflecting the dynamic nature of language use in real-life communication [46]. Communicative-based approaches, such as Communicative Language Teaching (CLT) and Task-Based Language Teaching (TBLT), focus on using language as a tool for communication. CLT emphasizes interaction as both the means and goal of learning [41, 49], fostering learners' ability to communicate in diverse situations. However, it may lack the grammatical rigor needed for accurate language use [52]. TBLT, a well-tested and efficient approach that draws student engagement while teaching with context, uses meaningful tasks as the central unit of planning and instruction [43]. It promotes natural language use and real-world relevance but can be difficult to implement effectively due to the need for careful task design and classroom management [11, 13, 30].

TBLT is a prominent approach in second language acquisition, emphasizing authentic language use through meaningful tasks. It comprises three stages: Pre-task, During-task, and Post-task. The Pre-task stage prepares learners for the main task by activating prior knowledge and providing essential language input. [58] highlights its importance, while [38] elaborates on pre-task activities. Studies show that effective pre-task activities enhance learners' confidence and reduce anxiety [39, 47]. The During-task stage is the core of TBLT, involving problem-solving, information-gap activities, or decision-making tasks. [26] emphasizes authentic language use and negotiation of meaning. [14] underscores its importance in providing contextually rich language practice, improving fluency, accuracy, and complexity [38, 43]. The Post-task stage involves reflection and consolidation through feedback, error correction, and task repetition. [58] suggests this phase allows learners to analyze performance and focus on emergent language forms. [48] demonstrates that post-task activities solidify learning and promote self-awareness, while [38] highlights their role in enhancing metacognitive skills and fostering long-term language development.

Table 1: Factors Included in our Model Card Construction and an Example with Claude 3 Opus

Factors	Example Model Card for Claude 3 Opus
Non-Optimal Uses	Fetching new resources, Generate short and defined responses, *etc.*
Intended Uses	Graduate-level writing feedback, Producing long and creative responses, *etc.*
Performance (Metrics)	**GPQA, Diamond (Graduate level reasoning)**: 50.4% (Accuracy). **DROP (Reasoning over Text)**: 83.1 (F1); *etc* [6].
Training Data	Publicly available information via the Internet, Third party licensed datasets, *etc* [5].

3 Model Card for Educational Applications

Inspired by the work on model cards [36], we have developed specialized criteria for selecting models for educational purposes. Our educational model cards provide essential information about pre-trained models used in education, focusing on their capabilities and characteristics most relevant to educational applications. As shown in Table 1, our criteria emphasize specific aspects that distinguish one pre-trained model from another, facilitating the selection of the most appropriate models for educational use. To enhance efficiency in model searching, this new format omits less immediately relevant—though still important—aspects such as licenses and ethical considerations. This streamlined approach allows for quicker identification of suitable models that meet specific educational needs. Our educational model cards include all functions in order of recommendation, categorized as either "Intended Uses" or "Non-Optimal Uses" for each model. This comprehensive yet focused presentation enables users to rapidly assess a model's strengths and limitations in educational contexts. Table 1 presents an example model card for Claude 3 Opus, a model pre-trained by Anthropic and accessible via API. It's important to note that this example does not exhaustively list all characteristics of the model, but rather highlights some key aspects relevant to language education applications.

3.1 Non-Optimal Uses

We introduce "Non-Optimal Uses" as a key factor in our educational model card. This approach differs from traditional model cards that typically list intended uses. Given the versatility of major Large Language Models (LLMs) and their capacity for numerous tasks, exhaustively listing all intended uses can be inefficient and computationally demanding. Instead, by highlighting uses that are less suitable or potentially problematic, we provide AI managers with a more streamlined and efficient method to quickly determine if a model aligns with specific educational requirements. Non-optimal uses are identified through two primary methods: comparative performance on different benchmarks and observations from actual practice. For instance, benchmark comparisons might reveal that while a model performs well in certain areas, it lags in others. Additionally, practical application often uncovers nuanced limitations. As an example, some Claude models, despite showing comparable overall performance to certain GPT models on the same benchmarks, excel at generating long, natural contexts but struggle with producing concise responses that directly address prompts. Such insights from real-world usage are invaluable in defining non-optimal use cases.

This approach aids in preventing the misallocation of models in educational contexts where their performance may be suboptimal or inappropriate. By focusing on non-optimal uses, the AI manager can process and filter model capabilities more efficiently for faster and more accurate model selection for educational applications. This nuanced understanding of a model's limitations ensures that AI managers can make informed decisions, optimizing the deployment of LLMs in various educational scenarios.

3.2 Intended Uses

The "Intended Uses" section of our educational model card provides a concise overview of the primary applications for which the model is optimized. This information is crucial for AI managers to quickly identify the model's strengths and ideal implementation scenarios. We derive these intended uses from both comparative benchmark performances and real-world applications. For instance, a model might consistently outperform others in specific tasks like in-depth analysis or creative writing, as evidenced by benchmark scores. Additionally, practical implementations might reveal unique strengths, such as a model's exceptional ability to maintain context over long conversations or to generate culturally nuanced content. Rather than an exhaustive list, we focus on these empirically proven core competencies and key educational tasks where the model excels. By clearly defining the model's intended uses based on concrete data and experience, we aim to enable AI managers to efficiently match models to appropriate educational contexts, ensuring optimal performance and resource allocation. This data-driven approach streamlines the decision-making process and allows for rapid and informed deployment of the most suitable models in various educational settings.

3.3 Performance with Metric

Performance with metric offers a comprehensive overview of the model's functionality across diverse educational tasks and scenarios. This section goes beyond raw metrics to provide context-specific insights into the model's strengths and limitations. It includes detailed performance data across various academic disciplines, demonstrating the model's subject-specific accuracy. The card also outlines the model's proficiency in processing different educational content formats, such as text, mathematical equations, and code. Additionally, it presents performance metrics across varying levels of task complexity, illustrating how the model scales with difficulty. For instance, the card might include specific benchmark results such as GPQA, Diamond for graduate-level reasoning: 50.4% (Accuracy); DROP for Reasoning over Text: 83.1 (F1) [6]. This detailed performance breakdown enables AI managers to precisely match

the model's capabilities with specific educational requirements, facilitating more accurate and efficient model selection for diverse educational applications.

3.4 Training Data

Training Data is vital for understanding the foundation of the model's knowledge and potential biases. For large, pre-trained Language Models, detailed information about training data may not be publicly available due to proprietary concerns, such as Claude 3 opus as shown in Table 1. However, for smaller models or open-source pre-trained models, this section can provide valuable insights to guide model selection. When available, this factor offers information about the nature, scope, and sources of the data used to train the model. It may include details on the diversity of educational materials, the range of topics covered, the languages represented, and the time frame of the data. Understanding the training data helps AI managers anticipate the model's areas of expertise, potential knowledge gaps, and possible biases. For models where such information is accessible, this data is crucial for ensuring that the model's outputs align with educational standards and for identifying areas where supplementary resources might be necessary. In cases where detailed training data information is not available, AI managers should consider this limitation when evaluating the model's suitability for specific educational tasks.

4 Prototype Design

Prototype Overview. This prototype is designed to enhance English language learning through a comprehensive suite of functions. At its core, it offers writing instruction that provides feedback and flexible guidance based on user input, complemented by a writing assessment feature that evaluates submissions using ETS TOEFL writing rubrics [24]. To support vocabulary development, the prototype incorporates a Vocabulary Builder that leverages real-world corpus for exemplification and offers in-depth morphological instructions. For spoken language skills, it includes a Speaking Partner function that employs a Communicative Language Teaching (CLT) approach, utilizing ACTFL speaking rubrics for assessment [1]. Additionally, the prototype features a Task-Based Language Teaching (TBLT) Classroom Assistant to facilitate the in-class writing process. This multifaceted tool aims to provide a holistic approach to English language learning, addressing various aspects of language acquisition in an integrated manner. To prioritize user privacy, the prototype only retains data for the duration of a single conversation, disposing of it once the user ends the session. Because the prototype is implemented on a server with limited resources (8GB RAM, 16GB CPU), some functions that could be handled by smaller, less resource-intensive models are instead processed by LLMs through an API. This approach, while more resource-intensive, is chosen to enhance the overall user experience, especially when multiple users are simultaneously engaging with the system. By leveraging LLMs, the prototype is able to process multiple languages and maintains high performance and responsiveness across all features of the language learning tool, even under concurrent user load.

Prompt Design. The CO-STAR prompt engineering framework[3, 20, 28] is a general approach for meta-prompt construction widely used in the field. It is designed to optimize interactions with Large Language Models (LLMs) and consists of six key elements: Context (providing relevant background information), Objective (defining the specific task), Style (specifying the desired writing approach), Tone (setting the response's attitude or mood), Audience (identifying the intended recipients), and Response (outlining the desired output format). By systematically addressing these components, users can craft more precise prompts, leading to more accurate and tailored responses from LLMs, thus enhancing the efficiency and effectiveness of AI-assisted tasks.

In our prototype, we primarily use the CO-STAR framework to design prompts for generators that generate content for direct user interaction. However, we also employ LLMs for tasks that don't directly output to users, such as information extraction and classification. For these simpler processing tasks, we adopt a streamlined approach, focusing mainly on the Objective and Response elements of the CO-STAR framework, with Context included when necessary. The modified approach of COR allows us to maintain efficiency for processors while still leveraging the strengths of the CO-STAR framework where appropriate. In addition, to further enhance the performance of LLMs, we adhere to the principle of assigning only one task per prompt to ensure focused and optimized responses for each specific function [45].

4.1 Pre-task Preparation

The prototype employs a sophisticated two-stage processing approach. Initially, GPT-3.5-turbo, accessed via the OpenAI API, serves as the primary processor, extracting relevant personal information from the user's input. This includes details about their English language background, learning goals, and interests, which are used to construct a one-time user profile. Building upon this profile, GPT-4 then takes over to perform two crucial functions as AI manager: it selects an appropriate task from a predefined stack of five tasks (writing instruction, writing assessment, vocabulary builder, speaking partner, and TBLT classroom assistant) and chooses a suitable model based on the modality requirements of the selected task. The model selection is based on a diverse set of AI models, each specialized for different modalities:

- Text processing: GPT-4, GPT-3.5-turbo, Claude 3 Opus, and Llama2-70b-chat
- Audio processing: Whisper-1
- Image processing: GPT-4 and Claude 3 Opus

This dynamic process ensures that each user receives a tailored learning experience that aligns with their individual needs and preferences while optimizing the use of appropriate AI models for different language learning tasks and modalities.

4.2 Inner Structures of Different Functionalities

As illustrated in Figure 2, the **Writing Assessment module** comprises two processors and one generator, with an optional image transcription step. The process begins with an optional image transcription, where any user-provided images are converted to text using a model selected based on the user's initial input. Following this, the writing classification processor categorizes the user's input as either integrated writing or academic discussion, adhering to TOEFL writing standards [23]. Finally, the assessment generator

evaluates the writing based on this classification. It incorporates the user's profile and applies the appropriate TOEFL rubric to produce a comprehensive assessment. This modular approach allows for flexible handling of different input types and ensures a tailored assessment process that considers both the specific writing task and the individual user's characteristics, all while aligning with established TOEFL criteria [24].

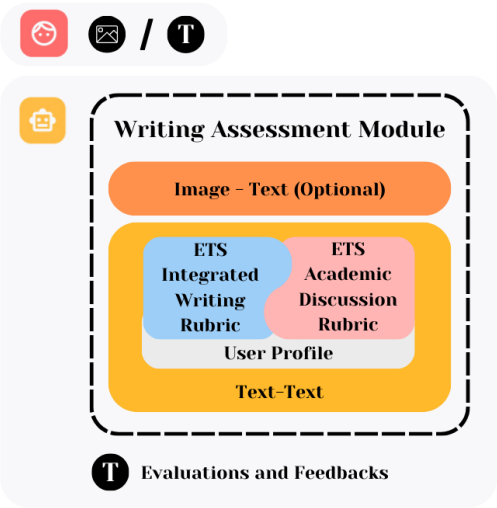

Figure 2: Inner Structure of Writing Assessment. Writing Assessment module will take images or text and output text.

The **vocabulary builder module** consists of three components: two text processors and one generator, as illustrated in Figure 3. By default, the text processors utilize GPT-3.5-turbo, though the AI manager may opt for a different model if deemed necessary. The first text processor identifies and extracts relevant vocabularies from the input text, while the second further refines and processes these extracted vocabularies. The generator then creates output based on a usage-based approach, leveraging real language data [55]. It sources its corpus from WordNet, accessed through the NLTK package [8]. The generator's process involves analyzing the roots and affixes of each vocabulary item, retrieving relevant information from WordNet, incorporating the user's profile data, and applying embedded morphological rules. This comprehensive approach allows the generator to produce contextually appropriate and personalized vocabulary output.

The **speaking module** accommodates users who prefer entering text for expressions, comprising three components in Figure 4: an audio-to-text processor, a generator, and a text-to-audio processor. Once the audio is transcribed into text, the generator analyzes it based on ACTFL speaking rubrics [1]. It then produces text output that reflects on the user's input and facilitates further communication, adhering to the Communicative Language Teaching (CLT) approach [49]. This process allows for a flexible interaction method while maintaining a structured, rubric-based assessment and a communicative teaching style.

We present two different approaches for implementing task-based language learning experiences: the writing instruction model

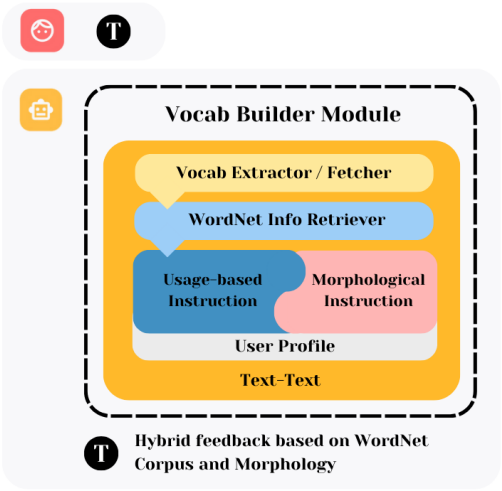

Figure 3: Inner Structure of Vocabulary Builder. Vocabulary Builder module will take text and output text.

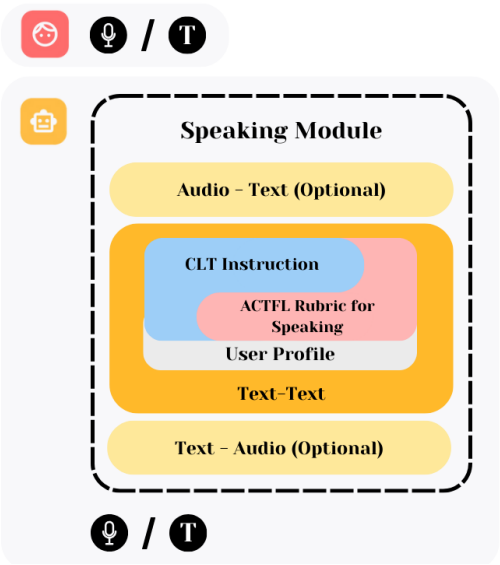

Figure 4: Inner Structure of Speaking Module. This module will take text or audio as input and output text or audio corresponding to the type of input.

and the TBLT Classroom assistant module. The **writing instruction module** in Figure 5 caters to users with diverse learning purposes and writing assignments, making it impractical to use predefined prompts for the three stages of task-based learning. This model requires two key components: (1) a stage identifier to determine the user's current phase in the learning process and (2) a generator that automatically creates appropriate prompts based on user input and profile. Once a suitable TBLT prompt is generated, a final generator produces customized responses for users. To

accommodate scenarios where users provide no personal information, we've also implemented a general prompt for delivering user feedback.

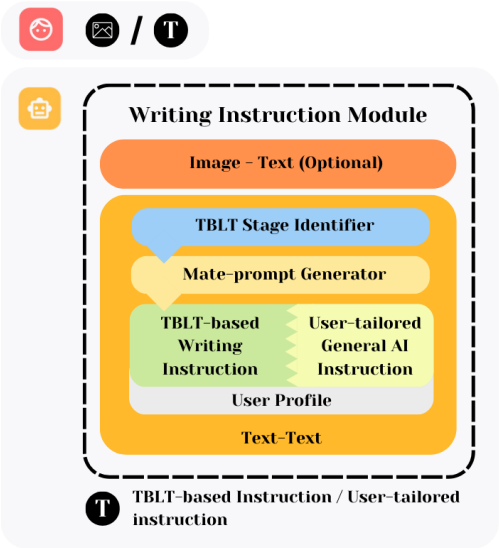

Figure 5: Inner Structure of Writing Instruction. Writing Instruction module will take images or text and output text.

The **TBLT Classroom Assistant module**, in contrast, is specifically designed for a literary writing lesson. This focused approach allows for clear expectations of students across the pre-task, during-task, and post-task stages. Consequently, we've implemented a processor to categorize users into their appropriate stages and designed defined prompts for each stage to generate tailored output. This structure enables a more directed and stage-specific learning experience within the context of task-based language teaching for literary writing.

5 User Study

We conducted user studies for each functionality, surveying two aspects: system understanding of users' input and response relevance. The number of completed surveys varied across functionalities, with writing assessment receiving 33 responses, vocabulary builder 36, writing instruction 35, speaking partner 41, and TBLT classroom assistant 15. The TBLT classroom assistant received fewer responses due to its specific use case for in-class activities. For each aspect, multiple questions were asked, and the overall satisfaction level was calculated as the average of all scores for that aspect. We categorized the average scores into five levels: Level 5 (4.5 to 5.0), Level 4 (3.5 to 4.49), Level 3 (2.5 to 3.49), Level 2 (1.5 to 2.49), and Level 1 (below 1.5). This categorization allows for a more nuanced evaluation of user satisfaction across different functionalities and aspects.

5.1 Input Comprehension

To assess how well the system comprehends and processes user inputs, we designed questions to evaluate the system's clarity and

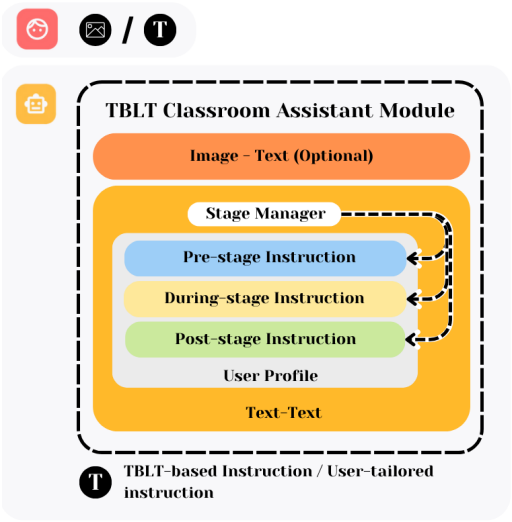

Figure 6: Inner Structure of TBLT Classroom Assistant. TBLT Classroom Assistant module will take images or text and output text.

precision in understanding inputs, contextual awareness, handling of complex queries, consistency, adaptability to variations in input style, and error handling. These questions help identify the system's strengths and areas needing improvement in understanding user interactions, ensuring that user inputs are correctly interpreted and acted upon. User satisfaction with input accuracy varied across functions. As shown in Figure 7, peaking Partner (41 respondents) and Writing Instruction (35 respondents) received high ratings, with over 90% giving level 4 or 5. Vocabulary Builder (36 respondents) and Writing Assessment (33 respondents) also performed well, with about 90% rating them 4 or 5, though Writing Assessment had the highest level 2 ratings (3%). The TBLT Classroom Assistant (15 respondents) had 86.7% level 4 or 5 ratings. Overall, users found input accuracy reliable across functions, with minor improvements needed, particularly for Writing Assessment. User feedback highlighted the need for multiple rubrics to assess language proficiency accurately and suggested that integrating instruction based on assessment would be more helpful for system to understand their needs.

5.2 Response Evaluation

Relevance and Accuracy of System Responses evaluate the system's ability to provide appropriate responses based on user inputs. We assessed understanding of user intent, relevance, accuracy, consistency, helpfulness, clarity, and handling of follow-up questions. This helps determine how well the system meets user expectations and enhances overall effectiveness. User satisfaction with response relevance varied across functions. In Figure 8, Speaking Partner (41 respondents) received 90.2% level 4 or 5 ratings (70.7% at level 5). Writing Instruction (35 respondents) had 82.8% at levels 4 or 5, with 17.1% at level 3. Vocabulary Builder (36 respondents) garnered 83.3% level 4 or 5 ratings. Writing Assessment (33 respondents) achieved 93.9% at levels 4 or 5, with 3% at level 2. TBLT Classroom Assistant

(15 respondents) received 93.3% level 4 or 5 ratings. Overall, users found the approaches effective across functions. Users Mentioned they expect writing instruction to have more features such as academic paraphrasing but the system concentrate more on providing feedbacks. This address the importance of diversify the general feedback prompt when users do not input their information.

5.3 Evaluation on how Language Learning Approach

This section assesses the system's impact on users' language learning progress. We evaluated learning progress, engagement, practical application of skills, personalization, feedback quality, activity variety, motivation, and ease of use. These factors help understand how effectively the system supports and enhances the language learning journey. User satisfaction with the language learning approach varied across functions. Speaking Partner (41 respondents) received 90.2% level 4 or 5 ratings, with 51.2% at level 5, though 2.4% rated it at level 2. Writing Instruction (35 respondents) had 91.4% at levels 4 or 5, with 60% at level 5. Vocabulary Builder (36 respondents) garnered 83.3% level 4 or 5 ratings, with 50% at level 5. Writing Assessment (33 respondents) achieved 93.8% at levels 4 or 5, with 56.3% at level 5, but 3% rated it at level 2. TBLT Classroom Assistant (15 respondents) received 86.7% level 4 or 5 ratings, with 60% at level 4 and 13.3% at level 3. Overall, users found the system's language learning approach effective across functions. However, there's room for improvement, particularly in Writing Assessment. Assessing rubrics is basically how users consider the approach in assessment. Therefore, their decision is also affected by the satisfaction of input comprehension.

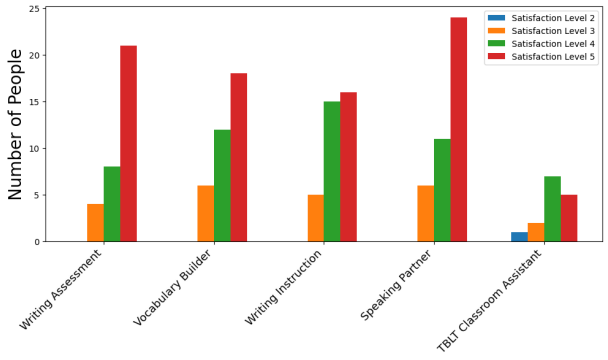

Figure 7: Overall Satisfaction on How the system understands the input.

6 Conclusion

The LanPIP prototype represents a significant advancement in integrating Large Language Models into language education. Grounded in established acquisition theories and real-world lesson planning, LanPIP offers a versatile framework for autonomous teaching, teaching assistance, and self-learning. Medium-scale user studies validated its efficiency and effectiveness, revealing several key insights. Participants consistently reported high satisfaction with LanPIP's

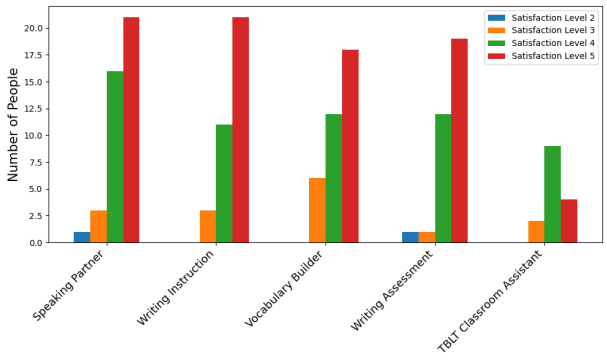

Figure 8: Overall Satisfaction on Relevance and Accuracy of System Responses

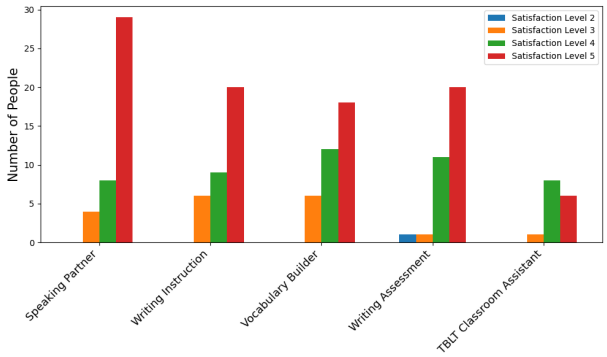

Figure 9: Overall Satisfaction on Effectiveness of the Adopted Language Learning Approach

ability to understand diverse inputs and provide relevant, accurate responses. The system's personalized feedback mechanism was particularly praised, with users noting improved confidence in their language skills. Interestingly, the multimodal interactions, combining text, audio, and visual elements, proved especially effective for vocabulary retention and speaking practice. The studies also highlighted LanPIP's adaptability across different proficiency levels, from beginners to advanced learners. The prototype's key features include writing assessment, vocabulary building, speaking practice, and task-based learning. By developing criteria for selecting appropriate language models and designing educational prompts, LanPIP ensures pedagogically sound AI integration. This successful implementation highlights the potential of AI-enhanced language learning when combined with teacher guidance, addressing key challenges in the field and supporting meaningful learning experiences.

References

[1] ACTFL. 2024. ACTFL Performance Descriptors for Language Learners. https://www.actfl.org/uploads/files/general/ACTFLPerformance_Descriptors.pdf.
[2] Mixtral AI. 2024. Au Large. *Mixtral News* (2024). https://mistral.ai/news/mistral-large/
[3] AI Advisory Boards. 2024. CO-STAR Framework. https://aiadvisoryboards.wordpress.com/2024/01/30/co-star-framework/.

[4] Anthropic. 2023. Introducing Claude. *Anthropic News* (2023). https://www.anthropic.com/news/introducing-claude

[5] Anthropic. 2024. How do you use personal data in model training? *Anthropic Privacy Legal* (2024). https://support.anthropic.com/en/articles/7996885-how-do-you-use-personal-data-in-model-training

[6] Anthropic. 2024. Introducing the next generation of Claude. *Anthropic Annoucements* (2024). https://www.anthropic.com/news/claude-3-family

[7] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Margaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (2021), 610–623.

[8] Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.".

[9] H Douglas Brown. 2014. *Principles of language learning and teaching: A course in second language acquisition.* Pearson.

[10] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[11] Lara Bryfonski and Todd H McKay. 2019. TBLT implementation and evaluation: A meta-analysis. *Language Teaching Research* 23, 5 (2019), 603–632.

[12] Jill Burstein, Martin Chodorow, and Claudia Leacock. 2004. Automated essay evaluation: The Criterion online writing service. *Ai magazine* 25, 3 (2004), 27–27.

[13] Yuko Goto Butler. 2011. The implementation of communicative and task-based language teaching in the Asia-Pacific region. *Annual review of applied linguistics* 31 (2011), 36–57.

[14] Martin Bygate, Peter Skehan, and Merrill Swain. 2013. *Researching pedagogic tasks: Second language learning, teaching, and testing.* Routledge.

[15] S. Caballé and Jordi Conesa. 2018. Conversational Agents in Support for Collaborative Learning in MOOCs: An Analytical Review. In *International Workshop on Intelligent Networking and Collaborative Systems.* https://api.semanticscholar.org/CorpusID:58005767

[16] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2023. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology* (2023).

[17] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374* (2021).

[18] Kee-Man Chuah and Muhammad Kamarul Kabilan. 2022. The development of mobile applications for language learning: A systematic review of theoretical frameworks. *International Journal of Learning, Teaching and Educational Research* 21, 8 (2022), 253–270.

[19] Cohere. 2024. Introducing Command R+: A Scalable LLM Built for Business. *Cohere Blog* (2024). https://cohere.com/blog/command-r-plus-microsoft-azure

[20] DataStax Docs. 2024. Generate Prompt | RAGStack. https://docs.datastax.com/en/ragstack/default-architecture/generation.html.

[21] Google DeepMind. 2024. Our next-generation model: Gemini 1.5. *Google Company News* (2024). https://blog.google/technology/ai/google-gemini-next-generation-model-february-2024/#sundar-note

[22] Pierre Dillenbourg. 2013. Design for classroom orchestration. *Computers & education* 69 (2013), 485–492.

[23] Mary K Enright and Thomas Quinlan. 2010. Complementing human judgment of essays written by English language learners with e-rater® scoring. *Language testing* 27, 3 (2010), 317–334.

[24] ETS. 2024. TOEFL® Essentials - Test Writing Scoring Guide. https://www.ets.org/pdfs/toefl/toefl-essentials-writing-scoring-guide-rubric.pdf.

[25] Luke K. Fryer, Mary D. Ainley, Andrew Thompson, Aaron Gibson, and Zelinda Sherlock. 2017. Stimulating and sustaining interest in a language course: An experimental comparison of Chatbot and Human task partners. *Comput. Hum. Behav.* 75 (2017), 461–468. https://api.semanticscholar.org/CorpusID:5404727

[26] Susan M Gass and Carolyn G Madden. 1985. Input in Second Language Acquisition. (1985).

[27] Víctor González-Calatayud, Paz Prendes-Espinosa, and Rosabel Roig-Vila. 2021. Artificial intelligence for student assessment: A systematic review. *Applied Sciences* 11, 12 (2021), 5467.

[28] GovTech Singapore. 2024. Prompt Engineering Playbook. https://go.gov.sg/promptengineering-playbook-file.

[29] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring Massive Multitask Language Understanding. *arXiv preprint arXiv:2009.03300* (2020).

[30] Siros Izadpanah. 2010. A study on task-based language teaching: From theory to practice. *US-China Foreign Language* 8, 3 (2010), 47–56.

[31] Israt Jahan, Md Tahmid Rahman Laskar, Chun Peng, and Jimmy Xiangji Huang. 2024. A comprehensive evaluation of large language models on benchmark biomedical text processing tasks. *Computers in Biology and Medicine* (2024), 108189.

[32] Deepak Kem. 2022. Personalised and Adaptive Learning: Emerging Learning Platforms in the Era of Digital and Smart Learning. *International Journal of Social Science and Human Research* (2022). https://api.semanticscholar.org/CorpusID:246506872

[33] Jinhee Kim, Hyunkyung Lee, and Young Hoan Cho. 2022. Learning design to support student-AI collaboration: Perspectives of leading teachers for AI in education. *Education and Information Technologies* 27, 5 (2022), 6069–6104.

[34] Shaofeng Li. 2010. The effectiveness of corrective feedback in SLA: A meta-analysis. *Language learning* 60, 2 (2010), 309–365.

[35] Alison Mackey. 2013. *Input, interaction and corrective feedback in L2 learning.* Oxford University Press.

[36] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19).* ACM. https://doi.org/10.1145/3287560.3287596

[37] Naomi Miyake. 2007. Computer supported collaborative learning. *The SAGE handbook of e-learning research* (2007), 248–265.

[38] David Nunan. 2004. *Task-based Language Teaching.* Cambridge University Press.

[39] David Nunan. 2004. *Task-based Language Teaching.* Cambridge university press.

[40] Fatma Oryza, Fitriah Asad, and Irma Soraya. 2022. Presentation-practice-production (PPP): elicitation technique used by the English teacher to teach grammar. *FOSTER: Journal of English Language Teaching* 3, 3 (2022), 149–159.

[41] Jack C Richards. 2006. *Communicative language teaching today.* Cambridge University Press.

[42] Jack C Richards and Theodore S Rodgers. 2001. *Approaches and methods in language teaching.* Cambridge university press.

[43] Peter Robinson. 2011. Task-based language learning: A review of issues. *Language learning* 61 (2011), 1–36.

[44] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence* 1, 5 (2019), 206–215.

[45] Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. 2021. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207* (2021).

[46] Peter Skehan. 1996. A framework for the implementation of task-based instruction. *Applied Linguistics* 17, 1 (1996), 38–62.

[47] Peter Skehan. 1998. *A cognitive approach to language learning.* Oxford University Press.

[48] Peter Skehan. 2003. Task-based instruction. *Language teaching* 36, 1 (2003), 1–14.

[49] Nina Spada. 2007. Communicative language teaching: Current status and future prospects. *International handbook of English language teaching* (2007), 271–288.

[50] Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics.* 3645–3650.

[51] Stergios Tegos and Stavros Demetriadis. 2017. Conversational agents improve peer learning through building on prior knowledge. *Journal of Educational Technology & Society* 20, 1 (2017), 99–111.

[52] Scott Thornbury. 1999. How to teach grammar. *Readings in Methodology* 129 (1999).

[53] Makhan Lal Tickoo. 1985. A History of English Language Teaching Howatt APR, Oxford University Press, 1984. *RELC Journal* 16, 2 (1985), 103–112.

[54] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint arXiv:2302.13971* (2023).

[55] Jose Tummers, Kris Heylen, and Dirk Geeraerts. 2005. Usage-based approaches in Cognitive Linguistics: A technical state of the art. (2005).

[56] Kurt VanLehn. 2011. The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational psychologist* 46, 4 (2011), 197–221.

[57] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems* 32 (2019).

[58] Jane Willis. 2021. *A Framework for Task-based Learning.* Intrinsic Books Ltd.

[59] Sean Wu, Michael Koo, Lesley Blum, Andy Black, Liyo Kao, Zhe Fei, Fabien Scalzo, and Ira Kurtz. 2024. Benchmarking Open-Source Large Language Models, GPT-4 and Claude 2 on Multiple-Choice Questions in Nephrology. *NEJM AI* 1, 2 (2024), AIdbp2300092.

[60] Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023. Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models. *ArXiv* abs/2309.01219 (2023). https://api.semanticscholar.org/CorpusID:261530162

[61] Kaijie Zhu, Jindong Wang, Qinlin Zhao, Ruochen Xu, and Xing Xie. 2024. DyVal 2: Dynamic Evaluation of Large Language Models by Meta Probing Agents. *arXiv preprint arXiv:2402.14865* (2024).

A User Experience Questionnaires

A.1 System Understanding of User Inputs

Please rate the following questions on a scale of 1 to 5:

1. **Clarity and Precision:**
   - On a scale of 1 to 5, how accurately do you feel the system understands the precise meaning of your inputs?
     1 - Very inaccurate
     2 - Inaccurate
     3 - Neutral
     4 - Accurate
     5 - Very accurate
2. **Contextual Awareness:**
   - On a scale of 1 to 5, how well does the system understand the context of your inputs (e.g., recognizing references to previous interactions or the broader topic)?
     1 - Very poor context understanding
     2 - Poor context understanding
     3 - Neutral
     4 - Good context understanding
     5 - Excellent context understanding
3. **Handling Complex Queries:**
   - On a scale of 1 to 5, how effectively does the system handle complex or multi-part queries?
     1 - Very ineffective
     2 - Ineffective
     3 - Neutral
     4 - Effective
     5 - Very effective
4. **Consistency:**
   - On a scale of 1 to 5, how consistent is the system in understanding your inputs across different interactions?
     1 - Very inconsistent
     2 - Inconsistent
     3 - Neutral
     4 - Consistent
     5 - Very consistent
5. **Adaptability:**
   - On a scale of 1 to 5, how well does the system adapt to variations in your input style (e.g., different wording, abbreviations, slang)?
     1 - Poor adaptation
     2 - Limited adaptation
     3 - Moderate adaptation
     4 - Good adaptation
     5 - Excellent adaptation
6. **Error Handling:**
   - On a scale of 1 to 5, how well does the system handle misunderstandings or errors in your inputs?
     1 - Very poor error handling
     2 - Poor error handling
     3 - Neutral
     4 - Good error handling
     5 - Excellent error handling

A.2 Effectiveness of Language Learning Approach

Please rate the following questions on a scale of 1 to 5:

1. **Learning Progress:**
   - On a scale of 1 to 5, how effectively does the system help you make progress in learning the language?
     1 - No progress
     2 - Minimal progress
     3 - Moderate progress
     4 - Significant progress
     5 - Excellent progress
2. **Engagement:**
   - On a scale of 1 to 5, how engaging do you find the language learning activities provided by the system?
     1 - Very unengaging
     2 - Unengaging
     3 - Neutral
     4 - Engaging
     5 - Very engaging
3. **Practical Application:**
   - On a scale of 1 to 5, how well do the learning activities prepare you for practical use of the language in real-life situations?
     1 - Not at all
     2 - Poorly
     3 - Moderately
     4 - Well
     5 - Very well
4. **Personalization:**
   - On a scale of 1 to 5, how well does the system personalize the learning experience to your individual needs and learning style?
     1 - Not at all personalized
     2 - Poorly personalized
     3 - Moderately personalized
     4 - Well personalized
     5 - Very well personalized
5. **Feedback Quality:**
   - On a scale of 1 to 5, how helpful is the feedback provided by the system in improving your language skills?
     1 - Very unhelpful
     2 - Unhelpful
     3 - Neutral
     4 - Helpful
     5 - Very helpful
6. **Variety of Activities:**
   - On a scale of 1 to 5, how diverse and varied are the language learning activities offered by the system?
     1 - Very limited
     2 - Limited
     3 - Neutral
     4 - Varied
     5 - Very varied

7. **Motivation:**
   - On a scale of 1 to 5, how motivated do you feel to continue using the system for language learning?
     1 - Not motivated at all
     2 - Slightly motivated
     3 - Moderately motivated
     4 - Highly motivated
     5 - Extremely motivated
8. **Ease of Use:**
   - On a scale of 1 to 5, how easy is it to navigate and use the language learning features of the system?
     1 - Very difficult
     2 - Difficult
     3 - Neutral
     4 - Easy
     5 - Very easy

A.3 Relevance and Accuracy of System Responses

Please rate the following questions on a scale of 1 to 5:

1. **Understanding User Intent:**
   - On a scale of 1 to 5, how well does the system understand your intent behind the input?
     1 - Very poorly
     2 - Poorly
     3 - Neutral
     4 - Well
     5 - Very well
2. **Relevance to Input:**
   - On a scale of 1 to 5, how relevant are the system's responses to your inputs?
     1 - Not relevant at all
     2 - Slightly relevant
     3 - Moderately relevant
     4 - Very relevant
     5 - Completely relevant
3. **Accuracy of Information:**
   - On a scale of 1 to 5, how accurate is the information provided by the system in its responses?
     1 - Very inaccurate
     2 - Inaccurate
     3 - Neutral
     4 - Accurate
     5 - Very accurate
4. **Consistency of Responses:**
   - On a scale of 1 to 5, how consistent are the system's responses over multiple interactions?
     1 - Very inconsistent
     2 - Inconsistent
     3 - Neutral
     4 - Consistent
     5 - Very consistent
5. **Helpfulness:**
   - On a scale of 1 to 5, how helpful are the system's responses in addressing your queries?
     1 - Not helpful at all
     2 - Slightly helpful
     3 - Moderately helpful
     4 - Very helpful
     5 - Extremely helpful
6. **Clarity of Responses:**
   - On a scale of 1 to 5, how clear are the system's responses?
     1 - Very unclear
     2 - Unclear
     3 - Neutral
     4 - Clear
     5 - Very clear
7. **Handling Follow-up Questions:**
   - On a scale of 1 to 5, how well does the system handle follow-up questions related to your initial input?
     1 - Very poorly
     2 - Poorly
     3 - Neutral
     4 - Well
     5 - Very well