석사학위논문
Master's Thesis

VIVID: 강의 동영상으로부터 인간-인공지능 협업을 통한 대리 대화 저작

VIVID: Human-AI Collaborative Authoring of Vicarious Dialogues from Lecture Videos

2024

최 슬 기 (最 슬 기 Choi, Seulgi)

한국과학기술원

Korea Advanced Institute of Science and Technology

석사학위논문

VIVID: 강의 동영상으로부터 인간-인공지능 협업을 통한 대리 대화 저작

2024

최슬기

한국과학기술원

전산학부

VIVID: 강의 동영상으로부터 인간-인공지능 협업을 통한 대리 대화 저작

최슬기

위 논문은 한국과학기술원 석사학위논문으로
학위논문 심사위원회의 심사를 통과하였음

2024년 06월 14일

심사위원장　김주호　(인)

심사위원　이의진　(인)

심사위원　안소연　(인)

VIVID: Human-AI Collaborative Authoring of Vicarious Dialogues from Lecture Videos

Seulgi Choi

Advisor: Juho Kim

A dissertation submitted to the faculty of
Korea Advanced Institute of Science and Technology in
partial fulfillment of the requirements for the degree of
Master of Science in Computer Science

Daejeon, Korea
June 14, 2024

Approved by

Juho Kim
Professor of Computer Science

The study was conducted in accordance with Code of Research Ethics[1].

[1] Declaration of Ethical Conduct in Research: I, as a graduate student of Korea Advanced Institute of Science and Technology, hereby declare that I have not committed any act that may damage the credibility of my research. This includes, but is not limited to, falsification, thesis written by someone else, distortion of research findings, and plagiarism. I confirm that my thesis contains honest conclusions based on my own careful research under the guidance of my advisor.

MCS

최슬기. VIVID: 강의 동영상으로부터 인간-인공지능 협업을 통한 대리 대화 저작. 전산학부 . 2024년. 47+v 쪽. 지도교수: 김주호. (영문 논문)
Seulgi Choi. VIVID: Human-AI Collaborative Authoring of Vicarious Dialogues from Lecture Videos. School of Computing . 2024. 47+v pages. Advisor: Juho Kim. (Text in English)

초 록

독백 형식의 온라인 강의는 학습자들이 쉽게 흥미를 잃게 만듭니다. ”대리적 대화” 형식으로 강의를 디자인하면 독백 형식보다 학습자의 인지 활동을 더 촉진할 수 있습니다. 그러나 학습자의 다양한 요구에 맞춘 대화 형식의 온라인 강의를 디자인하는 것은 교사에게 많은 시간과 노력을 요합니다. 그래서 여덟 명의 교육 전문가와 일곱 명의 강사와 함께 디자인 워크숍을 진행했고, 본 학위 논문에서 독백 형식의 강의 대본을 교육적으로 의미 있는 대화로 변환하는 데 필요한 주요 가이드라인과 대형 언어 모델(LLM)이 어떻게 사용될 수 있을지에 대해서 제시했습니다. 또한, 이 디자인 가이드라인을 적용하여 교사들이 LLM과 협력하여 교육 대화를 설계, 평가, 수정할 수 있도록 하는 시스템인 VIVID를 만들었습니다. 12명의 교사들을 대상으로 한 동일 집단 내 연구에서 VIVID가 강사들이 대화를 효율적으로 선택하고 수정할 수 있도록 도와 양질의 대화를 작성하는 데 도움이 된다는 것을 확인했습니다. 본 연구 결과는 다양한 학습 단계에 적용 가능한 고품질의 교육 대화를 만드는 데 LLM이 교사들에 의해 어떻게 사용될 수 있을지에 대한 방향을 제시하고 있습니다.

핵 심 낱 말 대화형 강의, 대리적 학습, LLM 기반 저작 도구, 교사 지원 도구, 비디오 기반 학습

Abstract

The lengthy monologue-style online lectures cause learners to lose engagement easily. Designing lectures in a “vicarious dialogue” for- mat can foster learners’ cognitive activities more than monologue- style. However, designing online lectures in a dialogue style catered to the diverse needs of learners is laborious for instructors. We conducted a design workshop with eight educational experts and seven instructors to present key guidelines and the potential use of large language models (LLM) to transform a monologue lec- ture script into pedagogically meaningful dialogue. Applying these design guidelines, we created VIVID which allows instructors to collaborate with LLMs to design, evaluate, and modify pedagogical dialogues. In a within-subjects study with instructors (N=12), we show that VIVID helped instructors select and revise dialogues effi- ciently, thereby supporting the authoring of quality dialogues. Our findings demonstrate the potential of LLMs to assist instructors with creating high-quality educational dialogues across various learning stages.

Keywords Dialogic lecture, Vicarious learning, LLM-based authoring tool, Instructor assist tool, Video-based learning

Contents

List of Tables

List of Figures

Chapter 1. Introduction

Online lectures are widely used for conveying knowledge in various learning contexts. Notably, instructors often use them as educational resources in various teaching contexts like flipped learning [36] or supplementary materials [8, 24], which usually take the format of knowledge-transfer-oriented online lectures. However, they usually take the form of a lengthy monologue. This format can cause learners to feel disengaged or quickly lose interest [4], potentially resulting in persistent negative emotions that detrimentally affect learning outcomes [28]. To address the limitations of this lecture format, studies have explored the application of conversational agents (CA) to video-based learning [55, 67, 68, 76]. Many studies used CAs to mimic human tutoring behaviors such as scaffolding [30], and these direct interactions with CA have improved the learning experience of online learners.

Although these studies imply the importance of CAs' scaffolding mechanisms in online video lecture settings, they have supported mostly the learners who prefer to interact directly with an instructor and peer learners [65, 66]. Yet, for *vicarious learners*, who prefer to learn from others and actively process the interactions of others, interactions that can be vicariously processed are more beneficial to their learning [65, 66]. To enhance *vicarious learners*' experience, systems with multiple CAs that simulate interactions between an instructor and a *direct learner* [68] have been introduced based on *vicarious learning theory* [49]. *Vicarious learning theory* explains the benefits of learning when *vicarious learners* observe tutoring between an instructor and *a direct learner* who interact directly with the instructor in a video lecture [11, 12, 49, 2]. The studies found that *vicarious learners* preferred dialogic lecture videos that incorporate CAs over monologue-style lecture videos, and it resulted in a positive effect on students' engagement [68]. Therefore, introducing vicarious dialogue into monologue-style lectures can serve as a promising solution to address the limitations of conventional online lectures and satisfy *vicarious learners.*

However, the current approach has not yet addressed how to create high-quality dialogues that cater to vicarious learners through adaptation or expansion of original lecture contents. It is important to consider the quality of the learning content, such as the level of detail provided by the lecturer, because it can significantly affect a vicarious learner's cognitive load and engagement. Thus, rather than simply enhancing lectures, we converted the original lecture script into a format that can reduce the cognitive load for vicarious learners and found a pedagogically meaningful format for high-quality dialogue. To do this, we integrated LLM in this conversion process since LLM has been discussed as a feasible way to design vicarious dialogues while reducing the extra effort for instructors [68].

In short, this work aims to alleviate the manual effort of instructors authoring vicarious dialogue and establish a scalable pipeline for designing educationally high-quality dialogues from lectures.

To achieve this goal, we have developed five guidelines for transforming monologic lectures into a vicarious dialogue that can benefit online learners: *Dynamic, Academically Productive, Cognitive Adaptable, Purposeful*, and *Immersive.* As an initial step in crafting these guidelines, we conducted an iterative inductive literature analysis to define what constitutes a pedagogically meaningful dialogue. However, most existing literature focused on insights derived from classrooms or intelligent tutoring systems, not video lectures. Furthermore, there is limited research on transforming the content in video lectures into high-quality educational dialogues. Therefore, we conducted a design workshop with eight educational experts and seven secondary school teachers to develop the guidelines to be tailored for a STEM video learning setting.

To facilitate the efficient authoring of video-based vicarious dialogues based on our guidelines, we propose a system, VIVID (**VI**deo to **VI**carious **D**ialogue), which allows instructors to design, evaluate, and modify vicarious interactions with video lectures. To empower this system, we propose a collaborative design process between LLM and instructors to generate high-quality vicarious dialogues efficiently. This process consists of three stages, guided by the developed guidelines in the workshop: (1) **Initial Generation**: After an instructor chooses where to convert in a lecture, LLM configures a direct learner's understanding level for each concept in the selected section of the lecture and generates initial dialogues. (2) **Compare and Selection**: Instructors compare and select from multiple generated dialogues, and (3) **Refinement**: Instructors collaborate with LLM to refine the final dialogue, which will replace a section of the video lecture.

To determine whether VIVID is helpful for instructors to transform monologue lectures into high-quality dialogue lectures, we conducted a within-subjects study with 12 instructors. VIVID helped instructors simulate a direct learner effectively through co-designing with VIVID. Furthermore, instructors found that VIVID is significantly better in monitoring essential considerations ($p = 0.04$) with an effect size (Cohen's d) of 0.8 than the Baseline when designing dialogue. To evaluate the pedagogical quality of the authored dialogues designed through VIVID, we also conducted a human evaluation with six secondary instructors in four criteria which is if the dialogue is *Dynamic*, *Academically productive, Immersive*, and *Correct*. We found that the dialogues made by VIVID were significantly better quality in most criteria than the dialogues generated by Baseline.

The contributions of this work are as follows:

- Design guidelines through design workshop for making vicarious educational dialogues from lecture videos.
- VIVID, a system that collaborates with LLM to assist instructors in authoring vicarious dialogues from the monologue-styled lecture videos.
- Findings from a user study with 12 instructors showing how VIVID can assist instructors in dialogue authoring (Section 6.2), and a technical evaluation with six instructors that demonstrates the higher quality of dialogues created by instructors using VIVID compared to the Baseline (Section 6.4).

Chapter 2. Related Work

We reviewed previous research on simulating vicarious learning in online learning contexts and approaches for generating diverse educational dialogues at scale.

2.1 Simulating Vicarious Learning in an Online Learning Environment.

Vicarious learning [11, 19] in an online environment typically occurs when observing the interaction between other learners and an instructor on platforms like Zoom or when witnessing peer discussions on QA platforms. Such situations of vicarious learning can stimulate learners' cognitive activity and enhance their level of engagement.

Thus, research has employed a Conversational Agent (CA) [60, 25, 31, 72, 37] to simulate interactions between a virtual tutor and tutee for supporting vicarious learners in video-based learning. For instance, Nugraha et al. [55] explored how a CA in the role of a tutee to Massive Open Online Course (MOOC) videos could enhance the vicarious learners' learning experiences. Similarly, Tanprasert et al. [67] implemented vicarious interaction in MOOCs as if participating in a Zoom class. To do this, they added scripted vicarious dialogues between virtual learners and an instructor to a lecture video in a chat format. These studies showed learners preferred dialogue-like lecture videos with CAs that mimic vicarious interactions over monologue lecture videos. This type of interpersonal interaction positively impacted vicarious learners' engagement.

However, it's important to note that these studies employed manually crafted dialogues of assumed equal quality even though the quality of dialogues can significantly influence learner engagement and outcomes [57]. Furthermore, there is limited research on designing high-quality educational dialogues to facilitate vicarious learning in video-based learning contexts. Consequently, we aim to fill this research gap by developing guidelines for creating high-quality educational dialogues that can promote effective vicarious learning experiences [68].

2.2 Generating Diverse Educational Dialogues for Vicarious Learners at Scale

Large Language Models (LLMs) are becoming increasingly useful for educators [48, 73]. One promising area of research involves utilizing them to create a wide range of educational materials [18, 61]. For example, Wang et al. [74] found that large pretrained language models (PLMs) can automate the generation of educational assessment questions. Other approaches introduce question generation models that automatically produce questions from educational content such as textbooks [74, 75].

However, they primarily focus on addressing the challenge of scaling the generation of specific question types and provide solutions primarily at the model level without considering the needs of instructors and learners. In contrast, Promptiverse [38] proposes a novel approach aimed at reducing the workload for instructors while delivering useful and diverse prompts to learners. Furthermore, ReadingQuizMaker [45] introduces a system to enable instructors to conveniently generate high-quality questions. Both systems allow instructors to create prompts or questions at scale, but neither considers the learners' level when generating them. Furthermore, they mainly focus on enhancing the diversity of single prompts or quizzes. Consequently, applying these approaches to generating diverse educational dialogues, which involve dynamic interactions between tutees and a tutor, may present challenges.

To evaluate various uses of LLM in generating learning materials, such as code explanations [40], learning objectives [62], they have been evaluated based on general criteria, such as "easy to understand" or "accuracy" without thoroughly considering the quality for specific tasks. However, to ensure quality, it is essential to establish specific and measurable criteria tailored to each task. Moreover, integrating LLM into education practice requires balancing the use of LLM with the role of instructors since relying solely on an automatic pipeline with LLM may result in low quality. Thus, we aim to establish criteria for assessing education dialogues and propose an LLM-based pipeline that can generate high-quality dialogues scalable while considering vicarious learners. Further, based on this pipeline, we aim to design an interactive system that allows collaboration between LLM and instructors in authoring dialogues.

Chapter 3. Design Workshop

To develop a guideline for designing high-quality vicarious dialogues, we employed the two-step approach. In the first step, we conducted an iterative inductive literature analysis to define what a pedagogically meaningful dialogue should look like. Despite the increasing amount of research on video learning, there has been little research on how to design beneficial vicarious dialogues based on lecture videos and how to support instructors in doing this. To address these issues, we conducted a design workshop to develop new guidelines for designing vicarious dialogues in the context of video-based learning.

3.1 Utterance Patterns and Teaching Strategies

Two of the authors conducted an iterative inductive analysis of literature to define what constitutes a pedagogically meaningful dialogue in literature. To identify relevant literature, we conducted a query-based search with the PRISMA process [51] on Google Scholar and the ACM Digital Library, and the 50 final papers were selected for meta-analysis. Our review was based on three search queries related to main keywords (Detailed analysis method is in the Supplemental Material):

- **Vicarious Learning:** ”vicarious learning” + (”learning gain” OR ”tutorial dialogue” OR ”monologue”)
- **Classroom Interaction:** ”classroom interaction” + ”science” + ”dialogic” + ”teacher questioning” + (”secondary school” OR ”undergraduate”)
- **Human Tutoring:** ”human tutoring” + ”tutorial dialogue” + (”strategy” OR ”move”) +(”scaffolding” OR ”feedback”)

Based on our literature analysis, we created initial guidelines for designing vicarious dialogue in video lectures. The vicarious dialogue should be perceived as a natural conversation occurring during a lecture and should be effective for vicarious learners. Thus, we derived two main factors for designing vicarious dialogues: (1) the most commonly observed utterance categories in real tutoring (Table 3.1, Table 3.2) and (2) effective teaching strategies for vicarious learners.

3.1.1 Key utterance categories that are commonly observed in 1-to-1 tutoring and classroom.

Several studies have collected data from actual one-on-one tutoring or classroom session recordings and performed qualitative coding at the statement level to classify representative types of utterances made by tutors and tutees. We categorized the tutor's utterances into nine types (Table 3.1) and the learner's utterances into five types (Table 3.2) to utilize for designing vicarious dialogues that simulate a natural tutoring scenario.

3.1.2 Three teaching strategies that can positively affect vicarious learners.

Research indicates that vicarious learners are notably affected by the direct learner's discourse following the instructor's statements as vicarious learners tend to mimic direct learner's actions [11].

Table 3.1: This table displays nine categories of tutor utterances and their corresponding definitions. The table consists of two columns, the first containing the tutor's utterance categories, and the second containing their respective definitions. Categories include Self-monitoring, Lecturing, Demonstrating, Questioning, Off-topic, Summarizing, Answering, Scaffolding, and Diagnosing.

Tutor's utterance	**Definition**
Self-monitoring	Utterance related to self-monitoring of one's teaching style [13].
Summarizing	Utterance that summarizes what has bee done so far or restates student's questions or statements comments [10, 44, 3, 46, 56, 34, 42]
Lecturing	Utterance explaining declarative knowledge, which includes facts and conceptual principles. [13, 10, 44, 6, 34, 9]
Answering	Utterance in response to student questions [13, 44, 3, 69, 42].
Demonstrating	Utterance related to solving specific problems in a way that allows student to model the instructor's problem-solving approach [10, 44].
Scaffolding	Utterance involving assistance or hints to help students reach answers on their own. [13, 10, 44, 6, 53, 14, 3, 46, 56, 52, 34, 9, 32, 42, 50, 20, 7, 5]
Questioning	Utterance containing questions to encourage students to recall knowledge or think productively (e.g., deep-level reasoning/short-answer questions [13, 71, 14, 3, 46, 56, 52, 69, 34, 41, 9, 32, 5, 42, 6, 59].
Diagnosing	Utterance used to diagnose student understanding or progress [13, 44, 6].
Off-topic	Introduction or unrelated utterances, such as small talk not related to learning [10, 44, 69, 42].

The three most influential dialogue patterns in vicarious learning include:

Integrate a direct learner's cognitive conflict. A tutoring video that contains a *cognitive conflict* situation, where the instructor corrects errors made by the direct learner, can improve the attention and interest of vicarious learners [19, 54, 11]. Thus, we propose designing dialogues as if the instructor encourages the direct learner to reach confusion and addresses misconceptions productively [39, 70, 29].

Integrate a direct learner's deep-level reasoning questions. We suggest incorporating deep-level reasoning questions [21, 17, 27, 16, 22] that address comparisons, inferences, and causal relationships among concepts into the direct learner's utterances. According to previous research in Intelligent Tutoring System (ITS) [21, 17, 27, 16, 22], the vicarious learners' learning was significantly improved when the direct learner posed deep questions. Therefore, if a direct learner asks a deep-level reasoning question during a lecture, it can encourage vicarious learners to engage in critical thinking.

Integrate a direct learner's substantial and relevant follow-up responses. A direct learner should provide answers or self-explanations based on the learning contents followed by an instructor's scaffolding or lecturing [16, 12, 11, 23].

3.2 Workshop Overview

To verify and improve the literature-based guidelines for vicarious interactions in video learning environments, we conducted design workshops with eight educational experts (7 female, 1 male) and seven

Table 3.2: Five categories of tutee utterances and their corresponding definitions.

Tutee's utterance	**Definition**
Questioning	Utterance related to posing cognitive deep questions or simple questions to the instructor [44, 13].
Answering	Utterance related to providing responses or completing scaffolding in response to a instructor's question [44, 13, 42].
Reflecting	Utterance related to assessing one's understanding level in response to an instructor's question or voluntarily [44, 13].
Explanation	Utterance related to speaking spontaneously, as if articulating one's thoughts simultaneously, without necessarily being prompted by the instructor's scaffolding [44, 13].
Off-topic	Introduction or unrelated utterances, such as small talk not related to learning [44, 42].

secondary school teachers (5 female, 2 male). We aimed to (1) derive design guidelines for effective conversion of monologue-style lecture videos into dialogue-style videos and (2) discover design opportunities for a system that can facilitate easy authoring of dialogue-style lectures with LLM. We mainly target STEM lectures in our workshop. STEM lectures can cause more intrinsic cognitive load, require more critical thinking than other subjects, and be prone to disengagement while watching lectures because they mostly consist of abstract concepts and complex formulas [63, 64]. Thus, we decided to present STEM lectures in a dialogue format as it might help with processing the dense knowledge of STEM lectures.

Chapter 4. Findings from Design Workshop

We identified the two most commonly mentioned issues by participants and formulated five design recommendations for creating high-quality vicarious dialogues. Additionally, we propose how LLM can be integrated into the educational dialogue authoring process.

4.1 Challenges in Converting Video Lectures to Dialogue

Two challenges were observed when instructors converted video lectures to dialogue.

Challenge 1: Designing the overall structure of dialogues. We observed that the participants faced difficulties in designing the overall structure of the dialogue when creating from scratch. Participants mostly first struggled with which part of the lecture should be converted to dialogue. P3 mentioned that it was *"difficult to figure out which parts of a monologue should be transformed into direct learner's questions"* and P4 said it was *"hard to decide when and how much dialogue to create"*. It poses the cold start problem when designing dialogues by considering the improvement of the vicarious learner's learning. Furthermore, participants struggled to determine the appropriate format for the dialogue as they were unsure how the dialogue format would affect learning outcomes. P5 said that *"while it was easy to convert the lecture into a simple question-and-answer format, I'm not sure if these would be meaningful dialogues for vicarious learners"*. P15 also mentioned, *"If it ends up looking too similar to the original lecture format, converting the material to a dialogue format might not be necessary"*, asserting the need to define what kind of dialogue format would be helpful for vicarious learners in an online learning environment.

Challenge 2: Anticipating direct learner's utterances based on their level of understanding. Both instructors and experts needed help with designing a direct learner's utterances. This is evident from comments: *"It is hard to add direct learners' misconceptions to dialogues effectively"* (P15) and *"It was difficult to consider individual responses of the direct learners"* (P7).

4.2 Design recommendations that should be considered while designing dialogue for vicarious learners.

Based on the challenges above, we propose five dialogue design recommendations. Furthermore, we suggest four teaching strategies (Table 9.2 in Appendix) validated by workshop participants as likely effective even in a video-based learning context among pre-defined guidelines based on literature (Section 3.1.2).

DR1. Dynamic: Include various interaction patterns to reflect the dialogic dynamics between the tutor and tutee. A vicarious dialogue should be structured with fast turn-taking and various utterance patterns (Table 3.1, Table 3.2) that capture the dynamism of an actual tutoring scenario. Moreover, P14 mentioned that *"fast turn-taking is required to hold the attention of vicarious learners in online education, as it is more difficult to retain focus on digital learning platforms than in physical classrooms"*. Furthermore, instructors and experts often divided the tutor's lengthy utterances into smaller sub-dialogues between the tutor and the direct learner, highlighting the quick turn-taking in vicarious dialogues.

DR2. Academically productive: Encourage the metacognitive and constructive utterances of the direct learner to make a dialogue academically productive. Direct learners' utterances should be pedagogically meaningful to enhance vicarious learners' learning and engagement. Most workshop participants consistently emphasized the influence of direct learners on vicarious learners throughout the dialogue design process. Notably, they stressed the importance of direct learners displaying "interactive engagement" in dialogues, as vicarious learners are highly likely to empathize with the direct learner's learning process. The term "interactive engagement" refers to the active engagement of direct learners both cognitively and metacognitively.

Direct learner's cognitive engagement: P15 highlighted the importance of a tutor in a vicarious dialogue who should encourage active engagement by facilitating connections between direct learners' existing knowledge and the new material, citing *Ausubel's meaningful learning theory* [33]. In addition, P14 mentioned that *"When the instructor links the learning contents with the learner's personal experiences, the transfer learning occurs more easily"*.

Direct learner's metacognitive engagement: P15 and P9 proposed incorporating self-assessment and explanations of understanding from the direct learner into vicarious dialogues: *"When a direct learner self-assesses their level of understanding or performs self-summarization, a vicarious learner could potentially check their comprehension"*. In addition, P15 suggested that a tutor continuously promotes the direct learner's metacognition. This guide aligns with the findings that in an ITS [1, 47], the constructive actions of a direct learner, such as answering based on what they learned from the instructor's scaffolding and asking deep-level reasoning questions [47, 35], significantly influenced the learning outcomes and participation of vicarious learners.

DR3. Cognitively adaptive: Adapt the teaching strategies to the level of understanding of the vicarious learner, learning objectives, and lecture contents. Previous literature suggests that strategies requiring higher cognitive engagement, like inducing cognitive conflicts and posing deep-level reasoning questions, benefit vicarious learners [21, 17, 27, 16]. However, applying cognitively demanding strategies, like *cognitive conflict* in Table 9.2 in Appendix, may not always suit all learning materials or learners when converting lecture videos into dialogues. P15 noted that the choice of cognitive strategy may vary depending on the granularity of the learning content being transformed into a dialogue. In addition, he emphasized the importance of aligning cognitive strategies with learning objectives and the level of vicarious learners, stating that *"Frequent placement of lighter, easily answerable questions and minimal use of cognitive strategies on important content could lower the cognitive load on vicarious learners"*.

DR4. Purposeful: Define a learning objective for the vicarious learner and ensure that the learning objective is achieved through that dialogue. To create meaningful dialogue for vicarious learners, we recommend aligning the dialogue's goal with the vicarious learner's learning objective and illustrating the achievement of this objective through interactions between a direct learner and a tutor. P15 and P8 emphasized the importance of defining clear learning objectives for vicarious learners as an initial step in dialogue creation. Additionally, P8 highlighted that learning objectives should be intimately tied to the difficulties vicarious learners face.

DR5. Immersive: Utilize realistic teaching scenarios and match the direct learner's cognitive level with the vicarious learner's level. We suggest considering two factors that can immerse vicarious learners in their vicarious interaction.

- *Incorporate common teaching scenarios*: Some participants suggested using real classroom scenarios

Table 4.1: Criteria when instructors evaluated the pedagogical quality of LLM-generated dialogues in our design workshop.

Criteria	Key questions
Dynamic	Are various interaction patterns (Table 3.1, Table 3.2) incorporated to reflect the dynamics of real classroom dialogue?
Academic Productivity	Is the teacher effectively eliciting the learner's metacognitive and constructive utterances to ensure the discourse is academically productive?
Cognitive Adaptability	Are the cognitive strategies used in the dialogue adaptively applied based on the vicarious learner's level, learning objectives, and the lecture contents?
Purposefulness	Is the learning objective of vicarious learners achieved through the dialogue between the direct learner and teacher?
Immersion	Does the dialogue represent realistic teaching scenarios and establish a direct learner's level comparable to that of a vicarious learner, thereby improving vicarious learner engagement?
Usefulness	Is the dialogue satisfactory and useful, considering personal experience with students, what an instructor wants to emphasize, and the instructor's usage context, such as the level of vicarious learners being targeted?
Correctness	Are domain-specific words used accurately, and is the conversation content based on facts?

for vicarious learner engagement. For example, P11 proposed scenarios in which the direct learner is given an incorrect problem and asked to explain what is wrong and a situation where another learner responds correctly to the tutor's question when a student gives wrong answers. P14 also suggested a scenario where a tutor makes the direct learner apply what they have learned in different examples.

- *Match cognitive levels*: Instructors and experts highlighted aligning the cognitive levels of direct and vicarious learners in lecture videos to benefit the vicarious learners.— *"Vicarious learners often lose interest when confronted with familiar material but are more likely to engage when unfamiliar or essential information is presented."* (P12). Therefore, addressing vicarious learners' unfamiliar or challenging parts through direct learners' dialogue could be an effective way to design meaningful and high-quality dialogue.

4.3 Enhancing the Educational Dialogue Design Process with LLMs

After establishing guidelines, we explored how instructors and experts used LLM-generated dialogues and developed evaluation criteria (Table 6.2) for evaluating their pedagogical quality based on how workshop participants assess the dialogues (Table 4.1). We also explored strategies for integrating LLMs into the educational dialogue design process.

4.3.1 Utilization of LLM-Generated Dialogues

We propose two ways in which the LLM could enhance the dialogue design process for vicarious learners. Firstly, it can provide pre-generated dialogues, stimulating instructors' ideation. P2 commented that using the LLM felt like it provided helpful guidelines, making it more effective than starting from scratch. Secondly, it can assist in modifying dialogues at different levels, refining sub-dialogues and crafting direct learners' responses. Participants proposed presenting expected responses at different levels (P12) and automating the process of generating questions from the direct learner's perspective (P2).

Despite the LLM's advantages, the dialogue authoring process still requires active instructor involvement. In our observation, we have noted that instructors have their own set of criteria when designing high-quality dialogues. These criteria are based on their teaching experiences and can vary depending on the instructor's emphasis on specific aspects where they believe vicarious learners may face challenges. Guided by these personalized criteria, instructors designed and revised their dialogues.

Some instructors found the generated dialogues satisfactory because they aligned with their intended teaching points or teaching style. P13 chose the dialogue, stating *"When teaching math, using fewer variables is better. So, I initially emphasized reducing the number of characters and utilizing known information. The dialogue aligns well with my problem-solving approach that focuses on minimizing variables"*. Some instructors didn't use the dialogues because the content didn't meet their quality criteria. For example, P11 made revisions to emphasize a specific point, stating, *"The tutee's question: 'So, is x-2 the square root of 6?' is crucial in the problem-solving process. It would be helpful if the tutor followed up with a question like, 'What is the number that becomes 6 when squared?' to elaborate on this point"*.

4.3.2 Criteria for Evaluating the Educational Dialogues

Instructors evaluated the quality of LLM-generated dialogue based on seven criteria (Table 4.1). Five of these criteria aligned with the key factors to consider when designing educational dialogues (Section 4.2), while the other two criteria, *Usefulness* and *Correctness*, pertain to evaluating dialogues generated by the LLM.

4.4 Design Goals

Based on LLM's strengths and limitations in designing educational dialogue and criteria that instructors emphasized the most when evaluating the quality of dialogues (Table 4.1), we propose four design goals (DG):

DG1. Enable instructors to easily simulate direct learners easily.

DG2. Assist instructors in designing dialogues by referencing utterances generated at various levels of granularity.

DG3. Assist instructors in creating dialogues that reflect the user's dialogue usage context and personal experience with students.

DG4. Ensure that instructors consistently monitor important considerations when designing vicarious dialogues.

Chapter 5. VIVID: A System for Authoring Vicarious Dialogues from Monologue-styled Lecture Videos with LLM Assistance

Based on our design goals from the workshop, we developed VIVID, an LLM-based system to assist instructors in crafting vicarious dialogues from their monologue-styled lecture videos. While LLM holds potential benefits for the dialogue design process, as detailed in Section 4.3.1, they may not be practically utilized in real educational settings if *Correctness* and *Usefulness* (Table 4.1) are not ensured. Thus, VIVID provides a collaborative authoring process between LLM and instructors, facilitating the generation of high-quality and correct vicarious dialogues. Based on our four design goals and observed dialogue design process in the workshop, this collaborative authoring process consists of three stages: (1) *Initial Generation*, (2) *Comparison and Selection*, and (3) *Refinement.*

To motivate VIVID's design, we describe a usage scenario where an instructor collaborates with LLM to author dialog through VIVID. A high school biology teacher, Sophia requires her students to watch recorded lectures before class. Sophia wants to make sure that students easily understand parts of the lectures with the most common misconceptions. In this context, she uses VIVID to transform the sections in her recorded lecture where misconceptions frequently occur into dialogues so that her students gain a better understanding. Thus, she uploads her lecture video to VIVID **(A1, Figure 5.1)** and selects the sections she wants to transform into dialogues **(A2)**.

Initial Generation. She then highlights areas where her students might develop misconceptions or key examples she wants to emphasize in the dialogue **(B1, Figure 5.1)**. Sophia aims to design the dialogue scenario as if it is occurring in a high school biology class, where a teacher addresses the direct learner's misconceptions in the dialogue **(B2)**. Upon highlighting, VIVID generates four dialogues reflecting the dialogue scenario.

Comparison and Selection. VIVID shows generated dialogues with an 'understanding level rubric' **(C1, Figure 5.1)** that shows four levels of learners' understanding for each key concept in the selected part and the 'dialogue cards **(C2)**' that contains key information of each dialogue. Sophia compares each dialogue, considering the knowledge levels of the direct learner for each concept illustrated in the dialogue cards **(C2)**. She then chooses to modify 'Dialogue 2' because it highlights the misconceptions she wants to include.

Refinement. Sophia modifies 'Dialogue 2' by adding questions in the tutor's utterance to address direct learner's misconceptions. She clicks the *Generate* button **(D1-A, Figure 5.2)** to add a new utterance. However, she is unsure what answers the direct learner could provide for these newly added questions. To view different examples of how the learner might respond, she first selects the learner's utterance that she wants to see more variations of clicked sub-dialogue **(D2)**. Afterward, she clicks the *Laboratory* button **(D4-1)**, and VIVID generates four variations of the chosen utterances.

After reviewing the results, she wants to replace the existing utterances with new ones that better represent the learner's misconceptions. She clicks the *Apply* button **(D4-2)** to replace the previous utterances with new ones. This allows Sophia to create a dialogue where misconceptions are effectively addressed in the final dialogue.

5.1 Initial Generation

VIVID initially creates various dialogues for instructors to choose the one that aligns best with their intention for converting monologue to dialogue as we found that the LLM-generated dialogues can be utilized as prototypes in the process of educational dialog design (Section 4.3.1). Notably, our LLM-based pipeline of the Initial Generation stage is designed to generate dialogues that satisfy the most emphasized characteristics by workshop participants, which are *Dynamic, Academically Productive*, and *Immersive* (DR1, DR2, and DR5 in Section 4.2). Furthermore, when generating dialogues, VIVID reflects instructors' needs in our pipeline, making instructors easily simulate direct learners with knowledge levels similar to their target vicarious learner (DG1 in Section 4.4). Thus, the Initial Generation stage consists of four steps to generate dialogues that finely adjust the direct learner's knowledge state based on the instructor's needs.

We determined our final prompts (further details are in the Supplemental Material) by evaluating the quality of various dialogues based on our evaluation criteria (Table 6.2).

5.1.1 Step 1. Create a rubric for highlighted areas, indicating the learner's understanding level for each concept.

DR5 (*Immersive*) in Section 4.2 suggests that the dialogue should align the cognitive level of direct learners with vicarious learners. *Highlighting* feature allows instructors to highlight sections in the script that vicarious learners might find challenging. It reflects the intention of instructors to convert the dialogue for a specific level of vicarious learners. Therefore, VIVID leverages the highlighted sections to make assumptions about the level of vicarious learners and uses it to model the direct learner (DR5 in Section 4.2).

Before configuring the direct learners' understanding state, we extract the core concepts of the selected area in the transcript and divide the direct learners' possible understanding state of each concept into four levels. These levels are based on the cognitive domain of Bloom's taxonomy [26] as it has been used by instructors to design, assess, and evaluate student's learning [43]. VIVID then generates four *understanding levels* for each key concept with LLM and presents them in a rubric format **(B1)** (Figure 5.1). The *understanding level* here refers to the understanding state expected of direct learners when they learn new concepts from the instructor during the dialogue.

5.1.2 Step 2. Determine the direct learner's understanding level using the highlighted parts and the rubric.

The highlighted parts present the concepts that the direct learner may not fully comprehend after the tutor's explanation in the dialogue. We set the direct learner's understanding level based on the highlighted concepts, using 'level 1', 'level 2', or 'level 3' in the generated rubric to indicate the direct learner's knowledge deficits. The direct learner is prompted at the highest understanding level, 'level 4' for unhighlighted areas.

The process of determining a direct learner's understanding level didn't consider prerequisite relationships between concepts to generate a dialogue that reflects varied levels of comprehension of each concept, as shown in Figure 5.4. For example, consider a case where Concept A is a prerequisite for Concept B. Even if the LLM model sets Concept A at 'level 1' and Concept B at 'level 4', a scenario can be designed where the learner studies Concept A with the teacher to fill the knowledge gap (level 1) and then responds well to Concept B (level 4).

5.1.3 Step 3. Create an answer sheet consisting of the learner's expected answers to the tutor's questions and questions showing where the learner struggles.

We designed our prompt to create expected questions and responses to the instructor's questions when the direct learner is in a specific knowledge deficit state. The expected answer sheet was designed in a descriptive format to reflect the learner's nuanced understanding. We prompted an LLM to manipulate the expected answers to the instructor's questions concerning the learner's knowledge level for each concept. We also designed a prompt to generate questions that direct learners might struggle with the concepts set to a low level.

5.1.4 Step 4. Generate dialogues.

The final dialogues are generated through prompts based on the following three elements as shown in Figure 5.3: (1) Adjusted direct learner's knowledge state information through Step 1 to Step 3 to achieve **Immersive (DR5)**, (2) Key utterance categories of a tutor and a tutee in Table 3.2 and Table 3.1 to achieve **Dynamic (DR1)**, and (3) Key teaching strategies described in Table 9.2 to achieve **Academic Productive (DR2)**.

5.2 Comparison and Selection

In *Comparison and Selection* stage, VIVID provides the instructors with an *Understanding level rubric* **(B1)** and *Dialogue cards* **(B2)** (Figure 5.1) to enable monitoring and selecting based on the criteria that were important during *Initial Generation* stage (DG4 in Section 4.4). Each *dialog card* **(B2)** contains the primary information of the dialogue, such as the direct learner's understanding level of each concept, key teaching strategies, and key dialogue patterns. Besides, *Understanding level rubric* represents a four-level understanding state for each key concept appearing in the selected part in the transcript.

5.3 Refinement

Basic tools for instructor's direct refinement.

In the workshop, we observed that instructors were proficient in using existing dialogue content, like breaking down lengthy tutor utterances into smaller segments or incorporating script contents into dialogue. To facilitate this kind of authoring, VIVID provides four basic functions: *add* **(D1-a)**, *duplicate* **(D1-b)**, *delete utterance* **(D1-c)**, and *change speaker* **(D1-d)**. As visible in **(D1)**, each utterance box in the final dialogue is clickable and can be moved with drag-and-drop (Figure 5.2). Additionally, we aimed to enhance the *Correctness* of the dialogue through direct refinement.

5.3.1 LLM-based refinement tool: *Laboratory*

In addition to basic functions, VIVID offers the *Laboratory* tool **(D4-1)** that provides alternatives **(D3)** for the selected sub-dialogues **(D2)** through LLM (Figure 5.2). It is designed to address the instructor's challenges in developing direct learners' utterances while considering their understanding level (*Challenge 2* in Section 4.1) and achieve DG3 (Section 4.4). To do this, we designed the prompt used in *Laboratory* tool while maintaining four key elements except for the original dialog patterns (in Supplemental Material): (1) learner's level of the selected dialogue in the *Comparison and Selection* phase, (2) dialogue context, (3) main learning contents, and (4) the number of turns. On the other hand, we diversified the dialogue patterns, reflecting utterance categories in Table 3.2 and Table 3.1 in our prompt. When the instructor clicks the *Apply* button **(D4-2)**, the selected sub-dialogue **(D2)** is replaced with the new sub-dialog **(D4-2)**.

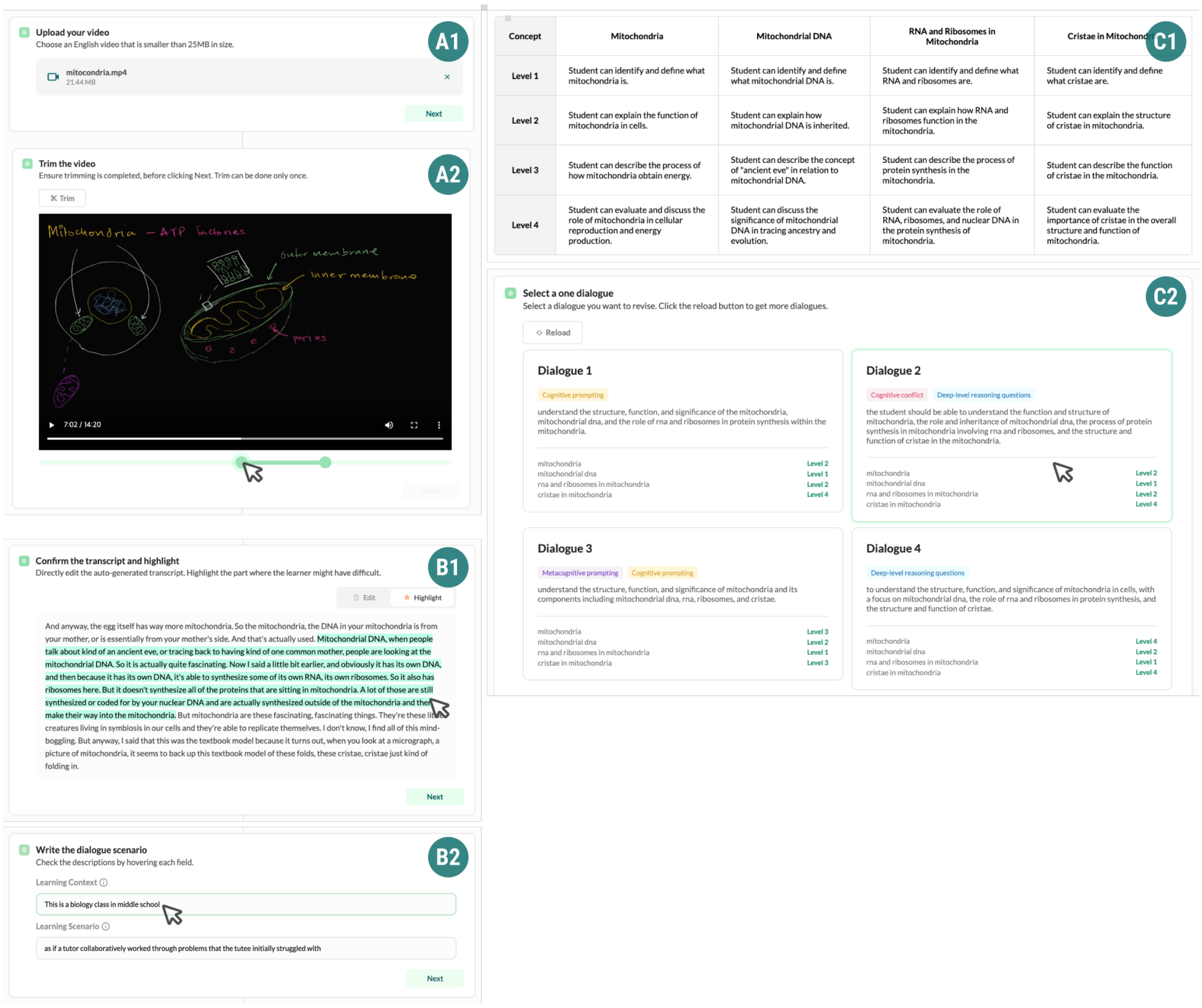
5.4 Implementation

VIVID is implemented using React [1], connected to a Flask [2]-based back-end server that utilizes GPT API. Whisper [58], an automatic speech recognition model by OpenAI, auto-generated the script of the section that the instructor chose from the lecture video (**B1** in Figure 5.1). To address limitations in text-to-speech (TTS) models like noise or language and get more precise dialogue conversion, VIVID allows instructors to modify the TTS output directly during the *Initial Generation* stage.

Subsequently, the system harnessed the API of the latest trained GPT-4, OpenAI's advanced language model, to generate the rubric, learner's knowledge level, predicted answer sheet, and the final dialogue. Considering the importance of model accuracy in an educational context, we conducted prompt engineering experiments using GPT-3.5 and GPT-4. We chose to use GPT-4 due to its superior generation quality. We set a temperature of 0.65 for the rubric generation, which was empirically determined through trial to maintain consistency, and used the default temperature for other features.

[1] https://react.dev/

[2] https://flask.palletsprojects.com/



Upload your video

Choose an English video that is smaller than 25MB in size.

mitocondria.mp4
21.44 MB

Trim the video

Ensure trimming is completed, before clicking Next. Trim can be done only once.

Confirm the transcript and highlight

Directly edit the auto-generated transcript. Highlight the part where the learner might have difficult.

And anyway, the egg itself has way more mitochondria. So the mitochondria, the DNA in your mitochondria is from your mother, or is essentially from your mother's side. And that's actually used. Mitochondrial DNA, when people talk about kind of an ancient eve, or tracing back to having kind of one common mother, people are looking at the mitochondrial DNA. So it is actually quite fascinating. Now I said a little bit earlier, and obviously it has its own DNA, and then because it has its own DNA, it's able to synthesize some of its own RNA, its own ribosomes. So it also has ribosomes here. But it doesn't synthesize all of the proteins that are sitting in mitochondria. A lot of those are still synthesized or coded for by your nuclear DNA and are actually synthesized outside of the mitochondria and then make their way into the mitochondria. But mitochondria are these fascinating, fascinating things. They're these little creatures living in symbiosis in our cells and they're able to replicate themselves. I don't know, I find all of this mind-boggling. But anyway, I said that this was the textbook model because it turns out, when you look at a micrograph, a picture of mitochondria, it seems to back up this textbook model of these folds, these cristae, cristae just kind of folding in.

Write the dialogue scenario

Check the descriptions by hovering each field.

Learning Context: This is a biology class in middle school

Learning Scenario: as if a tutor collaboratively worked through problems that the tutee initially struggled with

Concept	Mitochondria	Mitochondrial DNA	RNA and Ribosomes in Mitochondria	Cristae in Mitochondria
Level 1	Student can identify and define what mitochondria is.	Student can identify and define what mitochondrial DNA is.	Student can identify and define what RNA and ribosomes are.	Student can identify and define what cristae are.
Level 2	Student can explain the function of mitochondria in cells.	Student can explain how mitochondrial DNA is inherited.	Student can explain how RNA and ribosomes function in the mitochondria.	Student can explain the structure of cristae in mitochondria.
Level 3	Student can describe the process of how mitochondria obtain energy.	Student can describe the concept of "ancient eve" in relation to mitochondrial DNA.	Student can describe the process of protein synthesis in the mitochondria.	Student can describe the function of cristae in the mitochondria.
Level 4	Student can evaluate and discuss the role of mitochondria in cellular reproduction and energy production.	Student can discuss the significance of mitochondrial DNA in tracing ancestry and evolution.	Student can evaluate the role of RNA, ribosomes, and nuclear DNA in the protein synthesis of mitochondria.	Student can evaluate the importance of cristae in the overall structure and function of mitochondria.

Select a one dialogue

Select a dialogue you want to revise. Click the reload button to get more dialogues.

Dialogue 1

Cognitive prompting

understand the structure, function, and significance of the mitochondria, mitochondrial dna, and the role of rna and ribosomes in protein synthesis within the mitochondria.

mitochondria	Level 2
mitochondrial dna	Level 1
rna and ribosomes in mitochondria	Level 2
cristae in mitochondria	Level 4

Dialogue 2

Cognitive conflict · Deep-level reasoning questions

the student should be able to understand the function and structure of mitochondria, the role and inheritance of mitochondrial dna, the process of protein synthesis in mitochondria involving rna and ribosomes, and the structure and function of cristae in the mitochondria.

mitochondria	Level 2
mitochondrial dna	Level 1
rna and ribosomes in mitochondria	Level 2
cristae in mitochondria	Level 4

Dialogue 3

Metacognitive prompting · Cognitive prompting

understand the structure, function, and significance of mitochondria and its components including mitochondrial dna, rna, ribosomes, and cristae.

mitochondria	Level 3
mitochondrial dna	Level 2
rna and ribosomes in mitochondria	Level 1
cristae in mitochondria	Level 3

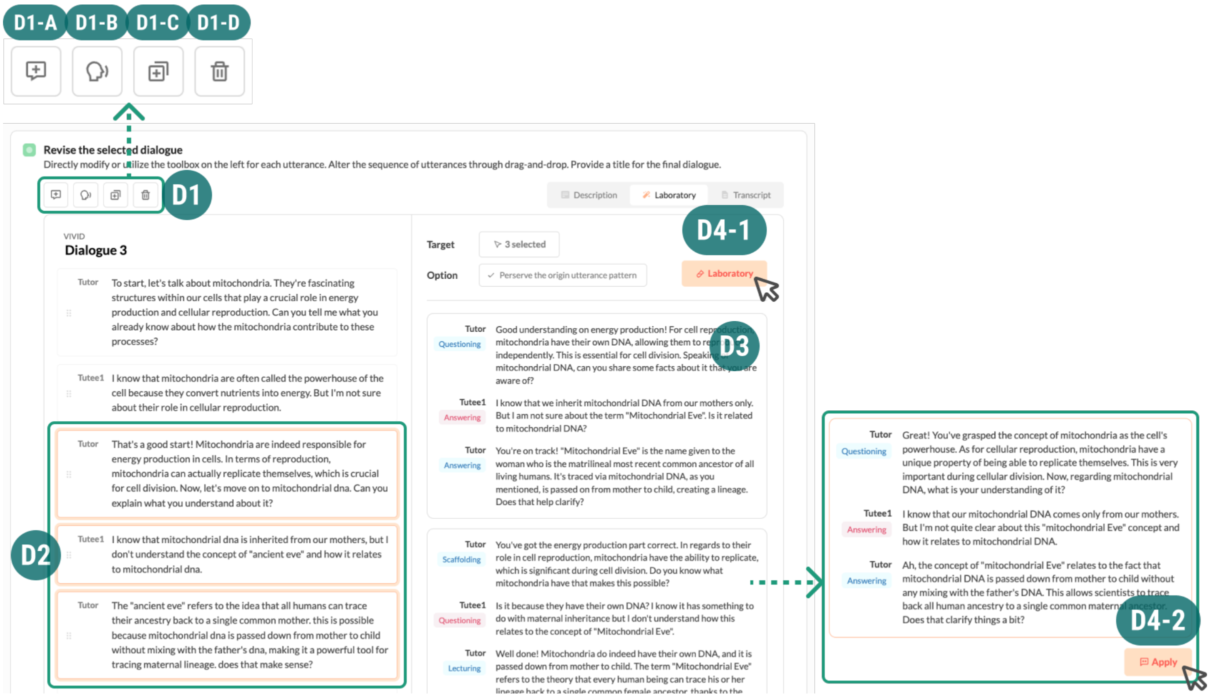
Dialogue 4

Deep-level reasoning questions

to understand the structure, function, and significance of mitochondria in cells, with a focus on mitochondrial dna, the role of rna and ribosomes in protein synthesis, and the structure and function of cristae.

mitochondria	Level 4
mitochondrial dna	Level 2
rna and ribosomes in mitochondria	Level 1
cristae in mitochondria	Level 4

Figure 5.1: VIVID's key components of *Initial Generation* : (A1) User uploads lecture video; (A2) User trims a video section to convert ; (B1) User uses the *highlighting* feature by selecting a part of the video transcript, where vicarious learners may face difficulty understanding ; (B2) User writes down the learning context and the scenario of dialogue that they want to depict in final dialogue, and *Comparison and Selection* phase : (C1) VIVID shows a rubric table of learners' *understanding level* regarding key concepts stated in the transcript; (C2) VIVID presents generated dialogues in the form of *dialogue cards* comprising of core information from each dialogue.



Revise the selected dialogue

Directly modify or utilize the toolbox on the left for each utterance. Alter the sequence of utterances through drag-and-drop. Provide a title for the final dialogue.

Dialogue 3

Tutor: To start, let's talk about mitochondria. They're fascinating structures within our cells that play a crucial role in energy production and cellular reproduction. Can you tell me what you already know about how the mitochondria contribute to these processes?

Tutee1: I know that mitochondria are often called the powerhouse of the cell because they convert nutrients into energy. But I'm not sure about their role in cellular reproduction.

Tutor: That's a good start! Mitochondria are indeed responsible for energy production in cells. In terms of reproduction, mitochondria can actually replicate themselves, which is crucial for cell division. Now, let's move on to mitochondrial dna. Can you explain what you understand about it?

Tutee1: I know that mitochondrial dna is inherited from our mothers, but I don't understand the concept of "ancient eve" and how it relates to mitochondrial dna.

Tutor: The "ancient eve" refers to the idea that all humans can trace their ancestry back to a single common mother. this is possible because mitochondrial dna is passed down from mother to child without mixing with the father's dna, making it a powerful tool for tracing maternal lineage. does that make sense?

Tutor (Questioning): Good understanding on energy production! For cell reproduction, mitochondria have their own DNA, allowing them to reproduce independently. This is essential for cell division. Speaking of mitochondrial DNA, can you share some facts about it that you are aware of?

Tutee1 (Answering): I know that we inherit mitochondrial DNA from our mothers only. But I am not sure about the term "Mitochondrial Eve". Is it related to mitochondrial DNA?

Tutor (Answering): You're on track! "Mitochondrial Eve" is the name given to the woman who is the matrilineal most recent common ancestor of all living humans. It's traced via mitochondrial DNA, as you mentioned, is passed on from mother to child, creating a lineage. Does that help clarify?

Tutor (Scaffolding): You've got the energy production part correct. In regards to their role in cell reproduction, mitochondria have the ability to replicate, which is significant during cell division. Do you know what mitochondria have that makes this possible?

Tutee1 (Questioning): Is it because they have their own DNA? I know it has something to do with maternal inheritance but I don't understand how this relates to the concept of "Mitochondrial Eve".

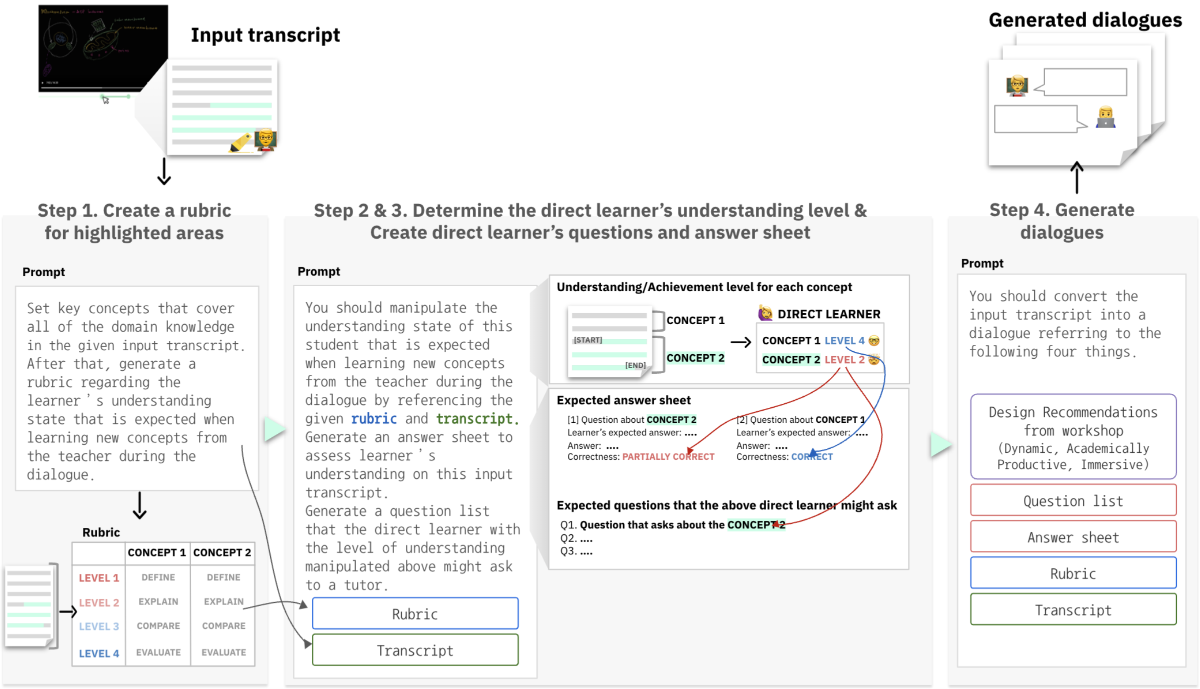
Tutor (Lecturing): Well done! Mitochondria do indeed have their own DNA, and it is passed down from mother to child. The term "Mitochondrial Eve" refers to the theory that every human being can trace his or her lineage back to a single common female ancestor, thanks to the

Tutor (Questioning): Great! You've grasped the concept of mitochondria as the cell's powerhouse. As for cellular reproduction, mitochondria have a unique property of being able to replicate themselves. This is very important during cellular division. Now, regarding mitochondrial DNA, what is your understanding of it?

Tutee1 (Answering): I know that our mitochondrial DNA comes only from our mothers. But I'm not quite clear about this "mitochondrial Eve" concept and how it relates to mitochondrial DNA.

Tutor (Answering): Ah, the concept of "mitochondrial Eve" relates to the fact that mitochondrial DNA is passed down from mother to child without any mixing with the father's DNA. This allows scientists to trace back all human ancestry to a single common maternal ancestor. Does that clarify things a bit?

Figure 5.2: VIVID's key components of *Refinement* phase : (D1) User can edit each utterance content directly or using basic editing tools; (D2) User can use *laboratory* feature by selecting consecutive utterances and clicking (D4-1) *laboratory* button ; (D3) VIVID suggests four variations of sub-dialogues as a result; (D4-2) *apply* button ; User can replace the original utterances with a variation by clicking button.


Input transcript
Generated dialogues
Step 1. Create a rubric for highlighted areas
Step 2 & 3. Determine the direct learner's understanding level & Create direct learner's questions and answer sheet
Step 4. Generate dialogues
Prompt
Rubric
Understanding/Achievement level for each concept
[START]
[END]
CONCEPT 1
CONCEPT 2
DIRECT LEARNER
CONCEPT 1 LEVEL 4
CONCEPT 2 LEVEL 2
Expected answer sheet
Expected questions that the above direct learner might ask

Prompt

Set key concepts that cover all of the domain knowledge in the given input transcript. After that, generate a rubric regarding the learner ' s understanding state that is expected when learning new concepts from the teacher during the dialogue.

Rubric

	CONCEPT 1	CONCEPT 2
LEVEL 1	DEFINE	DEFINE
LEVEL 2	EXPLAIN	EXPLAIN
LEVEL 3	COMPARE	COMPARE
LEVEL 4	EVALUATE	EVALUATE

Prompt

You should manipulate the understanding state of this student that is expected when learning new concepts from the teacher during the dialogue by referencing the given **rubric** and **transcript**. Generate an answer sheet to assess learner ' s understanding on this input transcript. Generate a question list that the direct learner with the level of understanding manipulated above might ask to a tutor.

Rubric

Transcript

Expected answer sheet

[1] Question about **CONCEPT 2**
Learner's expected answer: ....
Answer: ....
Correctness: PARTIALLY CORRECT

[2] Question about **CONCEPT 1**
Learner's expected answer: ....
Answer: ....
Correctness: CORRECT

Expected questions that the above direct learner might ask

Q1. **Question that asks about the CONCEPT 2**
Q2. ....
Q3. ....

Prompt

You should convert the input transcript into a dialogue referring to the following four things.

Design Recommendations from workshop (Dynamic, Academically Productive, Immersive)

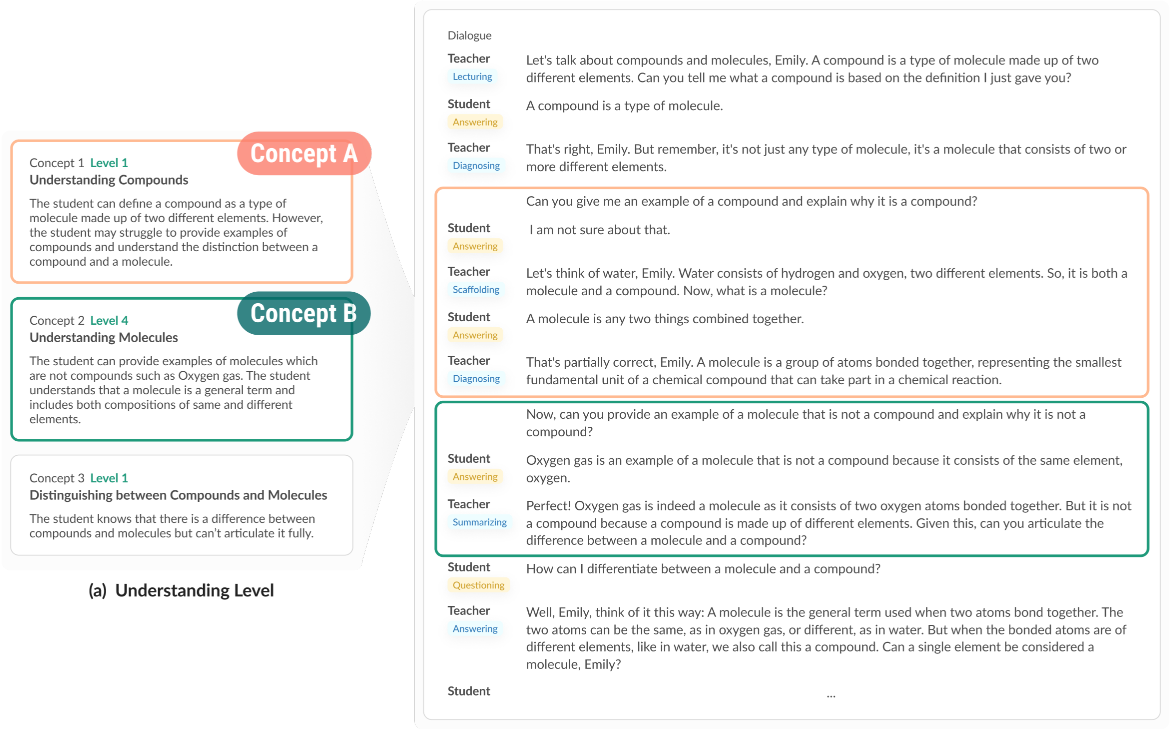
Question list

Answer sheet

Rubric

Transcript

Figure 5.3: Overview of prompting pipeline for *Initial Generation* phase. Each step corresponds to following subsections: (1) Create a rubric for highlighted areas, indicating the learner's understanding level for each concept ; (2) Determine the direct learner's understanding level using the highlighted parts and the rubric ; (3) Create an answer sheet consisting of the learner's expected answers to the tutor's questions and questions showing where the learner struggles ; (4) Generate dialogues based on the guidelines.



Concept 1 Level 1
Understanding Compounds

The student can define a compound as a type of molecule made up of two different elements. However, the student may struggle to provide examples of compounds and understand the distinction between a compound and a molecule.

Concept 2 Level 4
Understanding Molecules

The student can provide examples of molecules which are not compounds such as Oxygen gas. The student understands that a molecule is a general term and includes both compositions of same and different elements.

Concept 3 Level 1
Distinguishing between Compounds and Molecules

The student knows that there is a difference between compounds and molecules but can't articulate it fully.

(a) Understanding Level

Dialogue

Teacher (Lecturing): Let's talk about compounds and molecules, Emily. A compound is a type of molecule made up of two different elements. Can you tell me what a compound is based on the definition I just gave you?

Student (Answering): A compound is a type of molecule.

Teacher (Diagnosing): That's right, Emily. But remember, it's not just any type of molecule, it's a molecule that consists of two or more different elements.

Can you give me an example of a compound and explain why it is a compound?

Student (Answering): I am not sure about that.

Teacher (Scaffolding): Let's think of water, Emily. Water consists of hydrogen and oxygen, two different elements. So, it is both a molecule and a compound. Now, what is a molecule?

Student (Answering): A molecule is any two things combined together.

Teacher (Diagnosing): That's partially correct, Emily. A molecule is a group of atoms bonded together, representing the smallest fundamental unit of a chemical compound that can take part in a chemical reaction.

Now, can you provide an example of a molecule that is not a compound and explain why it is not a compound?

Student (Answering): Oxygen gas is an example of a molecule that is not a compound because it consists of the same element, oxygen.

Teacher (Summarizing): Perfect! Oxygen gas is indeed a molecule as it consists of two oxygen atoms bonded together. But it is not a compound because a compound is made up of different elements. Given this, can you articulate the difference between a molecule and a compound?

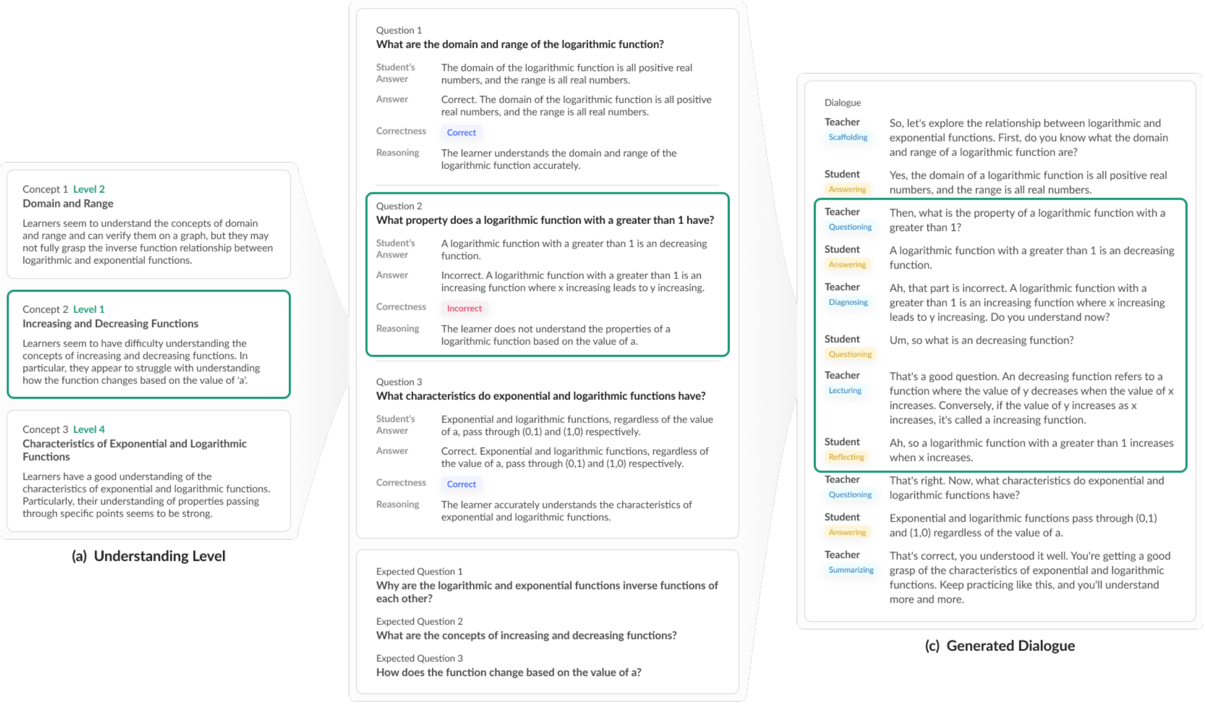
Student (Questioning): How can I differentiate between a molecule and a compound?

Teacher (Answering): Well, Emily, think of it this way: A molecule is the general term used when two atoms bond together. The two atoms can be the same, as in oxygen gas, or different, as in water. But when the bonded atoms are of different elements, like in water, we also call this a compound. Can a single element be considered a molecule, Emily?

Student ...

(b) Generated Dialogue

Figure 5.4: Example of generated dialogue regardless of the prerequisite relationships between key concepts. Concept A is a prerequisite for Concept B. During the conversation, the direct learner didn't understand the Concept A initially, but grasped it through question-and-answer, and answered Concept B correctly later.



Concept 1 Level 2
Domain and Range
Learners seem to understand the concepts of domain and range and can verify them on a graph, but they may not fully grasp the inverse function relationship between logarithmic and exponential functions.

Concept 2 Level 1
Increasing and Decreasing Functions
Learners seem to have difficulty understanding the concepts of increasing and decreasing functions. In particular, they appear to struggle with understanding how the function changes based on the value of 'a'.

Concept 3 Level 4
Characteristics of Exponential and Logarithmic Functions
Learners have a good understanding of the characteristics of exponential and logarithmic functions. Particularly, their understanding of properties passing through specific points seems to be strong.

(a) Understanding Level

Question 1
What are the domain and range of the logarithmic function?

Student's Answer	The domain of the logarithmic function is all positive real numbers, and the range is all real numbers.
Answer	Correct. The domain of the logarithmic function is all positive real numbers, and the range is all real numbers.
Correctness	Correct
Reasoning	The learner understands the domain and range of the logarithmic function accurately.

Question 2
What property does a logarithmic function with a greater than 1 have?

Student's Answer	A logarithmic function with a greater than 1 is an decreasing function.
Answer	Incorrect. A logarithmic function with a greater than 1 is an increasing function where x increasing leads to y increasing.
Correctness	Incorrect
Reasoning	The learner does not understand the properties of a logarithmic function based on the value of a.

Question 3
What characteristics do exponential and logarithmic functions have?

Student's Answer	Exponential and logarithmic functions, regardless of the value of a, pass through (0,1) and (1,0) respectively.
Answer	Correct. Exponential and logarithmic functions, regardless of the value of a, pass through (0,1) and (1,0) respectively.
Correctness	Correct
Reasoning	The learner accurately understands the characteristics of exponential and logarithmic functions.

Expected Question 1
Why are the logarithmic and exponential functions inverse functions of each other?

Expected Question 2
What are the concepts of increasing and decreasing functions?

Expected Question 3
How does the function change based on the value of a?

(b) Answer Sheet and Questions

Dialogue

Teacher Scaffolding	So, let's explore the relationship between logarithmic and exponential functions. First, do you know what the domain and range of a logarithmic function are?
Student Answering	Yes, the domain of a logarithmic function is all positive real numbers, and the range is all real numbers.
Teacher Questioning	Then, what is the property of a logarithmic function with a greater than 1?
Student Answering	A logarithmic function with a greater than 1 is an decreasing function.
Teacher Diagnosing	Ah, that part is incorrect. A logarithmic function with a greater than 1 is an increasing function where x increasing leads to y increasing. Do you understand now?
Student Questioning	Um, so what is an decreasing function?
Teacher Lecturing	That's a good question. An decreasing function refers to a function where the value of y decreases when the value of x increases. Conversely, if the value of y increases as x increases, it's called a increasing function.
Student Reflecting	Ah, so a logarithmic function with a greater than 1 increases when x increases.
Teacher Questioning	That's right. Now, what characteristics do exponential and logarithmic functions have?
Student Answering	Exponential and logarithmic functions pass through (0,1) and (1,0) regardless of the value of a.
Teacher Summarizing	That's correct, you understood it well. You're getting a good grasp of the characteristics of exponential and logarithmic functions. Keep practicing like this, and you'll understand more and more.

(c) Generated Dialogue

Figure 5.5: *Initial Generation* pipeline. (a) **Understanding level**: Example of the direct learner's understanding level using the highlighted parts and the rubric, (b) **Answer Sheet and Questions**: Example of the answer sheet consisting of learner's expected answers to the tutor's questions and expected questions of direct learner, (3) **Generated Dialogue**: Example of final dialogue based our guideline-based prompt. The green box shows how the concept that set in *level 1* reflects on the final dialogue.

Chapter 6. Evaluation

To evaluate the performance of VIVID in designing high-quality educational dialogues, we conducted a two-fold evaluation — user study and technical evaluation. In this section, we provide the details of each evaluation and results, respectively.

6.1 User Study

VIVID is designed to autonomously generate *Dynamic, Academically Productive,* and *Immersive* dialogues between a tutor and a direct learner and support instructors in efficiently modifying them. To validate the efficacy of VIVID, we conducted a within-subjects experiment with 12 participants, comparing it with the baseline system that lacks VIVID's core features.

6.1.1 *Study Setup*

Participants were asked to transform a part of the lecture video chosen by the authors, into a dialogue using the systems under each condition. Participants experienced both conditions with different videos in a counterbalanced order to prevent bias and ensure validity. We analyzed user behavior logs, post-survey, and interview data to understand how our system supported the authoring process.

Baseline Condition The following text describes how the Baseline system differs from the VIVID system regarding the four design goals. In the *Initial Generation* phase of the Baseline, it utilized a simple prompt (the detailed prompt is in the Supplemental) to create a dialogue that did not reflect the learner's understanding. Thus, the entire process of adjusting direct learner's knowledge through *Highlighting* feature (in Section 5.1.1) was excluded. During the *Compare and Selection* stage of the Baseline, the *summarized card function* and *understanding level rubric* were excluded from VIVID, enabling compare and selection of one out of four dialogues for revision without any background information about the generated dialogues. In the *Refinement* phase of the Baseline, the *laboratory function*, which offers multiple contextual alternatives for the sub-dialogue selected by the instructor, was removed.

Lecture Selection The clarity of the lecture video can have an impact on the quality of the resulting dialogue. Other factors, such as the length of the video, the difficulty of the content, and the subject matter, can also influence the dialogue creation process. Therefore, when selecting lecture videos, we carefully considered the lecturer's explanation style and balanced the educational content and level of difficulty across all conditions. All videos were aimed at secondary school students, and we chose lecture content with similar prerequisite levels and granularity. Each video was in Korean and was approximately 10 minutes in length.

As we targeted STEM subjects, we selected two science and two mathematics lectures to use: topics for the science lecture were *generation of waves* [1] and the *refraction of waves* [2], and topics for mathematics were *exponential function* [3] and *logarithmic function* [4]. Mathematics lectures are presented in the format of writing board screencasts with voice-over [15]. Science lectures are presented in the same

[1] https://www.youtube.com/watch?v=u0KO1rm8neI

[2] https://www.youtube.com/watch?v=64dZGBCELBc

[3] https://www.youtube.com/watch?v=FBAgxbQ931Y

[4] https://www.youtube.com/watch?v=I_H04p9HHcI

Table 6.1: User study participants' demographic, career, and their subject taught in the classroom.

ID	Gender	Age	Career	Subject taught
P1	F	50s	30 years	Math
P2	M	20s	2 years	Science
P3	F	20s	1 year	Science
P4	M	40s	15 years	Math
P5	M	30s	7 years	Engineering
P6	M	20s	2 years	Math
P7	M	20s	4 years	Engineering
P8	M	20s	5 years	Math
P9	F	20s	4 years	Math
P10	F	20s	2 years	Science
P11	M	20s	Graduated teacher's college	Math
P12	F	20s	1 year	Math & Science

format but based on slides. Each video follows a monologue-style lecture, where the instructor teaches without direct learners. The audio recording quality of all videos is at a level where the instructor can watch the lectures without any issues.

Participants We recruited 12 participants via social media platforms, including the local community for instructors. The participants were required to 1) teach STEM subjects, 2) have experience in designing online lectures or using them in their classes, and 3) be either school teachers or part-time instructors. We recruited participants for VIVID without considering teachers' experience levels, as VIVID is designed to support teachers regardless of their experience. All sessions were carried out via Zoom, and participants were compensated at a rate of 45,000 won per hour (equivalent to 34 USD).

6.1.2 *Study Procedure*

The study consisted of three tasks, followed by a post-task survey and interview.

Task 1. Eliciting ambiguous intent for the direct learner design. The participants were asked to convert a challenging section of a lecture into a dialogue that would help vicarious learners better understand the topic. In VIVID condition, the instructors had to select the specific contents that might be difficult for vicarious learners and convert them into dialogue using the *highlighting* feature. The specific guidelines on how the *highlighting* feature would affect the dialogue generation pipeline were not provided. On the other hand, instructors were only asked to choose where to convert without the *highlighting* feature in the Baseline condition. They then wrote about the teaching scenarios they wanted to depict in a dialogue.

Task 2. Comparing and selecting a dialogue to revise. Participants in the VIVID condition referred to dialogue cards and rubric to select one dialogue from four generated in *Initial generation* stage for revision. However, in the Baseline condition, instructors had to choose a dialogue that was designed without considering the direct learner, and they could not consider *rubrics* and information regarding the direct learner in choosing a dialogue.

Task 3. Revising a chosen dialogue. Participants in both conditions could refine the selected dialogue, employing the system's basic refinement functions. In the VIVID condition, participants could use the *laboratory* feature (Section 5.3.2) to refine their dialogue.

Post-task survey and interview After completing the tasks with both conditions, participants were asked to fill out a 7-point Likert Scale questionnaire that consists of nine questions to evaluate whether each feature of the system under each condition well reflected the design goals in Section 4.4 for creating quality educational dialogue and whether it produced quality dialogue (Figure 6.1). We conducted a semi-structured interview to understand participants' experiences with each system, the generated conversation, and the dialogue authoring experiences.

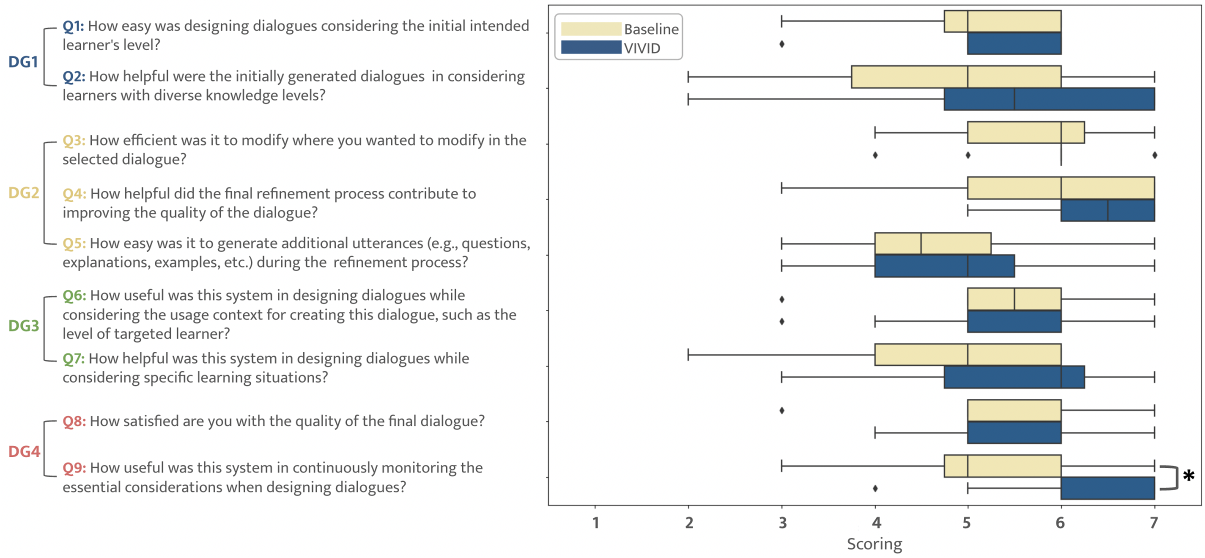


Figure 6.1: Post-task survey results on nine questions regarding task experiences. Each question was evaluated on a 7-point Likert scale. Treatment is corresponding to the VIVID.

6.2 *User Study Results*

Despite the overall high utility of the Baseline (Figure 6.1), nine out of 12 participants found VIVID to be better for designing vicarious dialogues due to its unique features such as *rubric*, *dialogue card*, and *laboratory* features. Notably, instructors considered VIVID to be significantly more helpful than the Baseline in monitoring important factors in dialogue design, as shown in Q9 of Figure 6.1. However, apart from this, no other significant differences in usefulness were observed.

6.2.1 VIVID helped participants monitor essential considerations when designing conversations.

Participants rated VIVID (M = 6.1, SD = 0.9) as significantly more useful in assisting them in monitoring key considerations persistently in dialogue design (Q9 in Figure 6.1) compared to the Baseline (M = 5.2, SD = 1.3, p = 0.04, Wilcoxon signed-rank test). Furthermore, while instructors felt that VIVID (M = 5.5, SD = 1.31) was more useful than the Baseline (M = 4.75, SD = 1.13) in considering specific teaching scenarios when designing dialogues (Q7), the difference was not statistically significant (p = 0.07, Wilcoxon signed-rank test). In terms of satisfaction with dialogue quality (Q8), there was a minimal difference between VIVID (M = 5.7, SD = 0.94) and the Baseline (M = 5.6, SD = 0.95). Although VIVID played a significant role in managing the educational dialogue design process, both conditions resulted in similar satisfaction levels due to manual refinement.

Table 6.2: Measuring questions used in our expert evaluation of the *Initial Generation* pipeline and statements used in our human evaluation of the end-to-end pipeline of VIVID to measure the educational quality of designed dialogue.

Criteria	Statement (7-point Likert Scale)	Measuring questions (Pairwise Comparison)
Dynamic	SD1. The dialogue demonstrates clear and fast turn-taking. SD2. The dialogue utilizes diverse interaction patterns between a tutor and tutees.	QD1. Which one demonstrates clearer and faster turn-taking? QD2. Which one utilizes more diverse interaction patterns between a tutor and a tutee?
Academic Productivity	SAP1. The dialogue encourages the learner's cognitive engagement (e.g., asking about what they've learned, asking various types of questions, and inquiring about a student's experiences) SAP2. The dialogue prompts a student's metacognitive thinking.	QAP1. Which one encourages the learner's cognitive engagement more? (e.g., asking about what they've learned, asking various types of questions, and inquiring about a student's experiences) QAP2. Which one prompts a learner's metacognitive thinking more?
Immersion	SI1. The dialogue appears to describe a specific and natural learning situation. SI2. The dialogue reveals and addresses a learner's knowledge deficits more clearly.	QI1. Which one describes a more specific and natural learning situation? QI2. Which one reveals and addresses a learner's knowledge deficits more clearly?

6.2.2 VIVID helped instructors simulate a direct learner with diverse levels of understanding.

Although the difference in Q2 (Figure 6.1), which evaluates how helpful the initially generated dialogue was in considering learners of various knowledge levels, was not significant, VIVID (M = 5.3, SD = 1.6) had a higher average than the Baseline (M = 4.6, SD = 1.56). In addition, some instructors highlighted VIVID was better at selecting a suitable dialogue by considering the direct learner's knowledge level for each dialogue than Baseline. P1 mentioned, *"VIVID was more conducive to constructing a lesson script optimized for the target learner as it clearly indicates the learning stage compared to the Baseline."*. Furthermore, P4 stated, *"VIVID was preferable as it allows selection and refinement according to the learner's level by showing rubric, so it was helpful for selecting dialogues with an appropriate difficulty level."*. Notably, P5 and P11 mentioned that the understanding level rubric provided with the dialogue cards allowed them to consider the direct learner's level more specifically when choosing a dialogue.

6.2.3 VIVID's *laboratory* feature helped instructors better predict the direct learner's responses and improve the dialogue's pedagogical quality.

Eight of eleven instructors who used the *laboratory* feature were satisfied with this feature. One instructor did not use this feature. Some instructors highlighted how this feature positively impacted the dialogue quality. We observed that the *laboratory* feature helped instructors explore the design space of dialogues while considering possible responses from direct learners. P1 said, *"Especially regarding the utterances of direct learners, it was difficult for the participants to imagine what questions the learner would ask, but through this feature, I was able to consider various learning situations and learner's responses that I hadn't thought of before."*. P5 also mentioned, *"I could consider answers and questions that direct learner might have from a wider range of perspectives"*.

6.3 Technical Evaluation

To evaluate whether VIVID supports authoring dialogues that meet the design requirements for educational dialogues (Section 4.2), we conducted a technical evaluation focusing on three primary parts: (1) *Initial generation* prompting pipeline, (2) our end-to-end pipeline designed for dialogue authoring, and (3) *Correctness* of the final dialogue. For the human evaluation of two pipeline outputs, we invited four instructors who participated in our user study to evaluate the pedagogical quality of the dialogues using the metrics shown in Table 6.2.

6.3.1 ***Initial generation*** prompting pipeline evaluation

We created a test dataset to explore how dialogues are generated through the Initial Generation prompting pipeline because the pipeline is designed to play the most crucial role in generating quality dialogue. Notably, we aim to investigate whether the language and the subject factors affect the quality of the pipeline to test the generalizability of the system for different subjects and languages.

To do this, we construct our test dataset on two lecture videos. We selected science and mathematics lectures to use: topics for the science lecture (Properties of periodic waves)[5] and topics for mathematics (Linear equation)[6], as our target domain is STEM subjects. In addition, to compare across different languages, we selected Khan Academy videos with transcripts available in both Korean and English. For each subject, we selected one segment of approximately 2-3 minutes for dialogue generation. We generated 32 dialogues that consist of 16 Baseline evaluation dialogues (8 in Korean and 8 in English) and 16 VIVID evaluation dialogues (8 in Korean and 8 in English). Detailed test dataset generation process and dialogue examples are in Appendix.

Two evaluators evaluated the Korean dialogues, while the other two who are proficient in reading and listening in English assessed the English dialogues, utilizing the given evaluation metrics (*Measuring questions* column of Table 6.2). Each evaluator conducted a pairwise comparison on a set of 32 pairs of dialogues and was asked to choose the dialogue generated by VIVID or the Baseline condition. Then, we calculated the preference percentage of selecting each condition to provide a comprehensive view of the system comparison.

6.3.2 ***End-to-end*** dialogue authoring pipeline evaluation

We assessed the final dialogues in two ways. Firstly, we compared the Likert scores to determine which one produced more *Dynamic*, *Academically Productive,* and *Immersive* dialogue. Secondly, we compared the percentage of incorrect responses for each dialogue to evaluate the variations in correctness before and after the instructor's refinement and between the different conditions.

Dynamic, Academically Productivity, and Immersion *evaluation of authored dialogues by VIVID.* We collected expert evaluations on 20 dialogues designed during the user study, ten from Baseline and ten from VIVID. Each dialogue was evaluated by three or four evaluators, as evaluators did not evaluate the dialogues designed by themselves. The evaluators used the evaluation metrics shown in *Statement* column of Table 6.2, which consisted of six 7-point Likert-scale questions.

Correctness evaluation of authored dialogues in both conditions. During the *Refinement* stage, instructors were allowed to make direct modifications. We conducted an evaluation study with four instructors to validate our approach using 48 dialogues from our user study. 24 dialogues were generated

[5] https://www.youtube.com/watch?v=tJW_a6JeXD8&t=345s

[6] https://www.youtube.com/watch?v=9IUEk9fn2Vs

before direct modifications by instructors, and 24 were created after modifications. Two instructors each evaluated the same dataset and their respective teaching subjects. We compared the two sets of dialogues to identify how our approach improved *Correctness*.

Based on the definition of typical hallucination [78, 77], we classified three types of incorrectness that may occur in educational dialogues on a turn-by-turn basis: (1) incorrect case where original numbers or explanations in the transcript were transformed incorrectly, (2) inconsistency observed when the answer deviates from the question from the student (e.g., when a student asks a question about logarithm function but the teacher provides an answer about the exponential function), and (3) inconsistencies observed across multiple turns (e.g., inconsistency in the student's knowledge level).

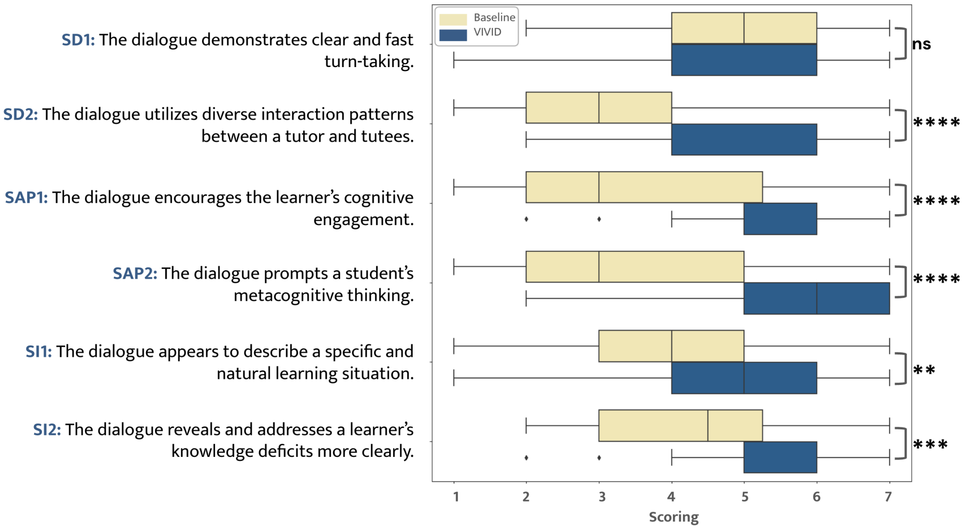


Figure 6.2: Results of authored dialogues' quality. ****, ***, **, *, and ns indicate significance of p $<= 0.0001$, $0.0001 < p < 0.001$, $p < 0.01$, $p < 0.05$, and $p > 0.05$, respectively.

6.4 Technical Evaluation Results

The technical evaluation showed that instructors designed significantly higher quality educational dialogues using VIVID compared to Baseline in all criteria except SD1 (Figure 6.2). Our study also found that the *Initial Generation* stage produces better educational dialogues than the Baseline, with the exception of QD1 (Figure 6.4). However, the overall usefulness of each system feature was not significantly high among the instructors, as we reported in Section 6.2, so we discussed possible reasons for the gap between the usefulness and quality of dialogues in Section 7.1.

6.4.1 Dynamic, Academically Productivity, and Immersion evaluation of authored dialogues by VIVID

Technical evaluation results showed that the instructors created significantly better educational dialogues with VIVID than the baseline. As shown in Figure 6.2, the dialogues designed through the entire pipeline of VIVID were rated significantly higher in quality in all aspects, except for SD1, compared to the baseline. The most significant findings were shown in SD2 (VIVID: M= 5.11, SD= 1.45, Baseline: M= 3.4, SD= 1.73), SAP1 (VIVID: M= 5.47, SD= 1.13, Baseline: M= 3.7, SD= 1.9), SAP2 (VIVID:

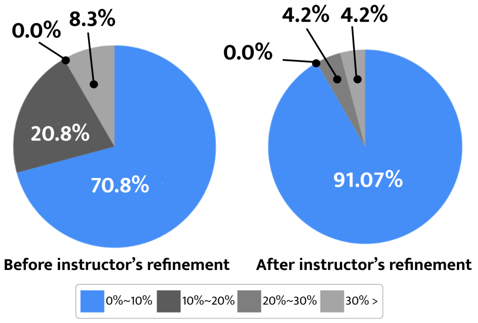


Figure 6.3: Human evaluation results on *Correctness*. The figure illustrates how the instructor's refinement has affected the correctness.

M= 5.64, SD= 1.4, Baseline: M= 3.3, SD= 2, p <= 1.00e-04, Wilcoxon signed-rank test). In other words, dialogues authored by the end-to-end pipeline of VIVID better described the metacognitive and cognitive activities of direct learners and consisted of more diverse patterns than the baseline. Difference of SI2 (p <= 0.001) and SI1 (p <= 0.01) also showed significance. The result implies that the dialogues authored with VIVID described a more natural learning situation and the direct learner's knowledge deficit better than the baseline.

6.4.2 Correctness evaluation of authored dialogues in both conditions.

We analyzed the percentage of turns with errors in each dialogue. As shown in Figure 6.3, after the modification, the total incorrectness rate of 0%-10% increased from 71% (17) to 92% (22). Before modification, VIVID generated more incorrect dialogues than Baseline because VIVID had to consider more details about the direct learner's understanding states when generating dialogue. After modification, the percentage of VIVID's incorrect dialogues in the 0-10% range increased from 67% (8) to 92% (11), while Baseline increased from 75% (9) to 92% (11) (Figure 6.5). These results indicate that VIVID improved correctness better than Baseline, and the instructor's refinement produced high-correctness final dialogues in both conditions. Additionally, we calculated the percentage of exceptional incorrect dialogues due to transcript errors (e.g., absence of essential conditions like 'x < 0', incorrect concept definition). Before the instructor made corrections, 25% of the entire dialogue resulted from incorrect transcripts. Even after the instructor's refinement, around 17% persisted, especially due to the absence of essential conditions in math dialogues.

6.4.3 *Initial Generation* pipeline evaluation.

The dialogues generated by VIVID's *Initial Generation* pipeline were rated higher in quality compared to the corresponding baseline in terms of all metrics listed in Table 6.2, except QD1. The most significant difference (Baseline: 5.5%, VIVID: 86.7%) was on QAP2 (Table 6.2) as in the end-to-end pipeline (Section 6.4.1), indicating that VIVID generates initial dialogues that effectively reflect a direct learner's metacognitive activity (Figure 6.4). Figure 6.4 shows that QD2, QAP1, QI1, and QI2 had over a 50% difference between VIVID and Baseline except for QD1 (Baseline: 71.01%, VIVID: 14.1%). We discussed the issue of poor quality for QD1 in Section 7.1.

Dialogue quality difference by subject (Science, Math) As shown in Figure 6.6, the most significant difference between Baseline (2%) and VIVID (86%) was observed in the QD2 criterion, which

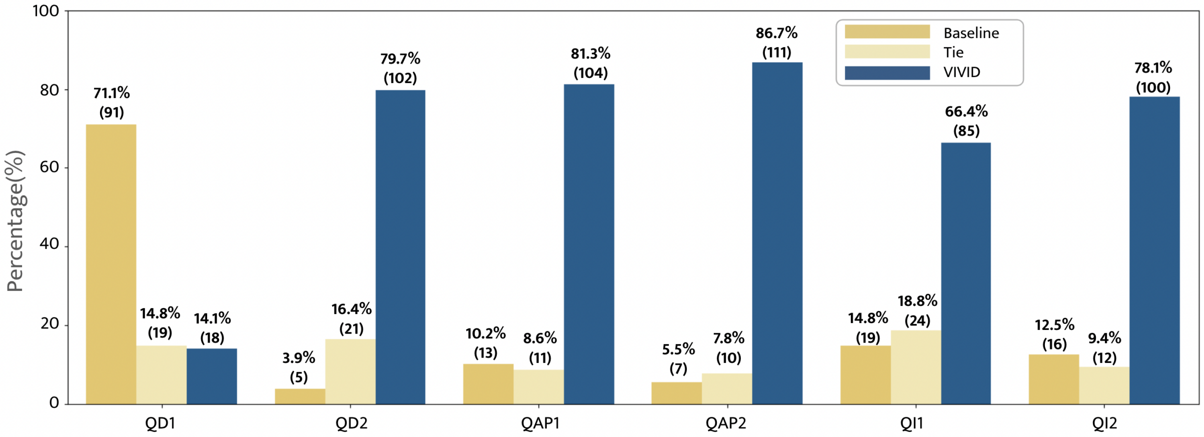


Figure 6.4: Human evaluation results on the six questions listed in Table 6.2. Four instructors evaluated dialogue sets, and each instructor conducted a pairwise comparison on a set of 32 pairs of dialogues. Except for QD1, VIVID outperformed the Baseline in the other five metrics.

is about diverse interaction patterns, in science dialogues. This suggests that VIVID effectively utilizes diverse dialogue patterns between the tutor and the direct learner, regardless of the language used.

Regarding math videos, the evaluation metric with the largest difference between Baseline (5%) and VIVID (84%) was QAP2 (Table 6.2). This indicates that VIVID is particularly effective in designing a dialogue that encourages metacognitive speech from a direct learner, regardless of the language used. On the other hand, QAP1 (Baseline: 13%, VIVID: 84%) and QI1 (Baseline: 8%, VIVID: 80%) showed smaller differences, but were still significant. In both subjects, evaluators preferred the dialogues generated by the Baseline in terms of QD1 (Table 6.2) which is about verbosity.

Dialogue quality difference by language (English, Korean) As shown in Figure 6.7, QAP2 (Table 6.2) had the most significant difference between Baseline (3%) and VIVID (91%) in English. This suggests that despite the subject, VIVID effectively created dialogues that elicited metacognitive activities from the instructor to the learner when the dialogues were converted from English to English. Yet, the Baseline performed better in terms of QD1 better than VIVID, both in English (Baseline: 41%, VIVID: 14%) and Korean (Baseline: 50%, VIVID: 4%) as in the results of dialogue quality difference by subject.

When converting Korean lecture into Korean dialogue, VIVID showed a significant contrast between Baseline (6%) and VIVID (92%) in the QI2 criterion while QAP1 had the least difference between Baseline (31%) and VIVID (51%). This indicates that VIVID effectively represented the learner's knowledge gaps directly and clearly, and depicted the process of addressing these difficulties in the dialogue, regardless of the subject.

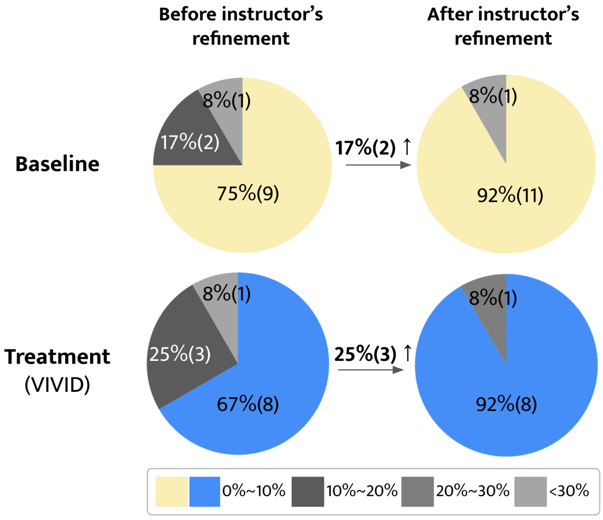


Figure 6.5: Human evaluation results on *Correctness* according to condition. This figure shows bigger changes in VIVID and high correctness of final dialogues in both conditions. The questionnaire used in this evaluation is from Table 6.2.

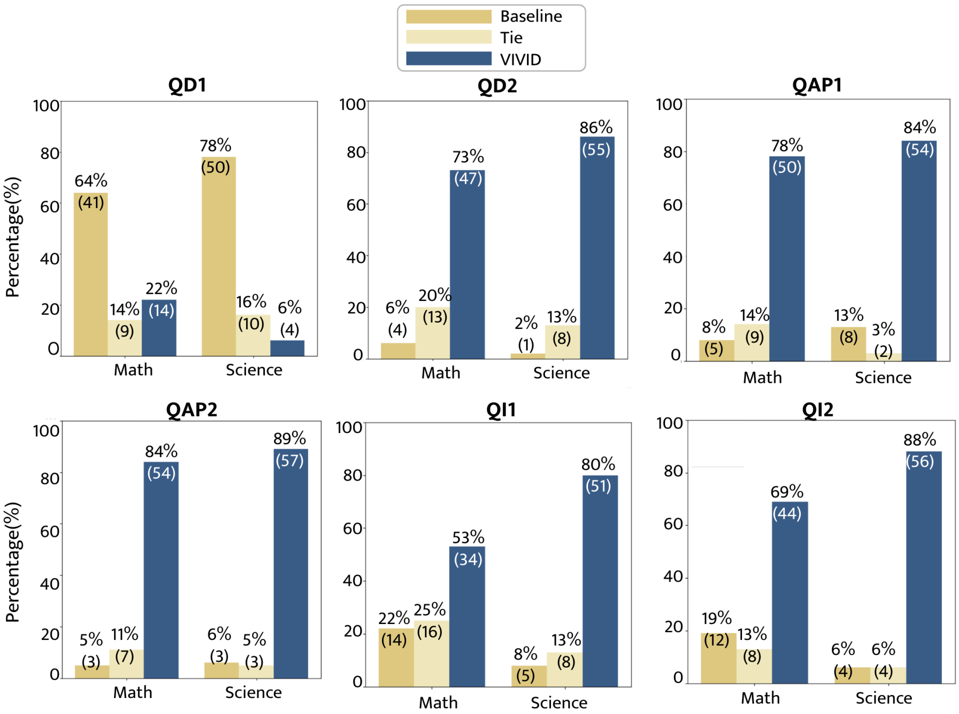


Figure 6.6: Results of dialogue quality by subject (Science, Math). The questions used are listed in Table 6.2. Except for QD1, VIVID outperformed the Baseline in the other five metrics.

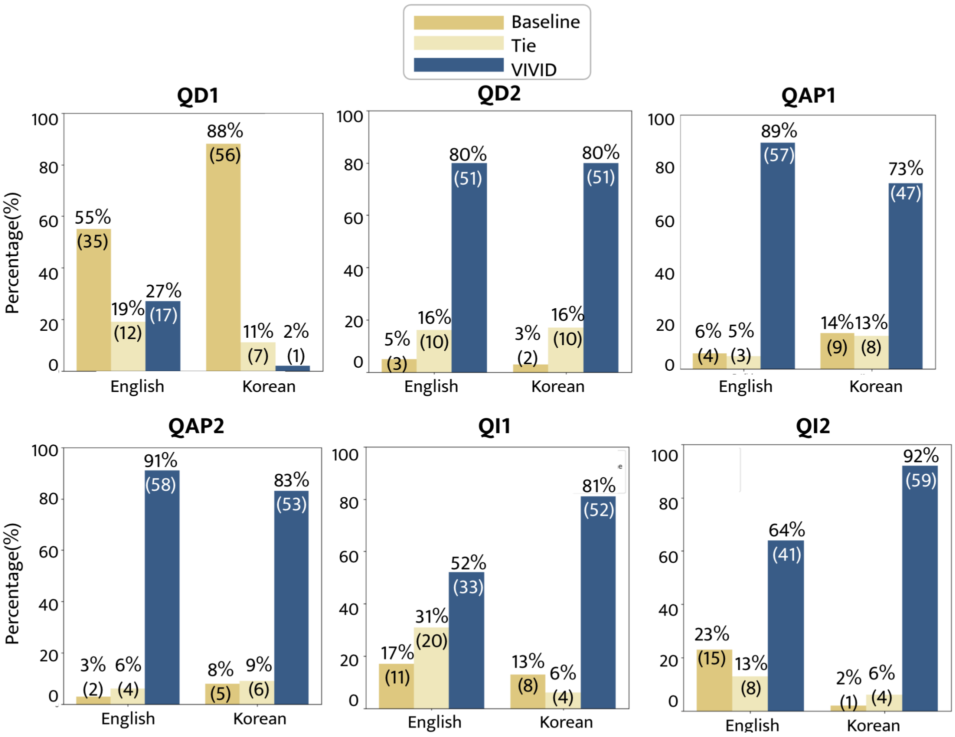


Figure 6.7: Results of generated dialogue quality by language (Korean, English). The questions used are listed in Table 6.2. Except for QD1, VIVID outperformed the Baseline in the other five metrics.

Chapter 7. Discussion

In this section, we discussed how to improve explainability, controllability, and verbosity for better utility, VIVID's potential beyond lecture videos, customizability for learners, and its generalizability.

Despite the positive results for VIVID, the overall usefulness of each system feature was relatively low among the instructors (Figure 6.1). We attribute this to two reasons: (1) low explainability and informativeness of dialogue design described in the *highlighting* feature and *dialogue cards* and (2) low controllability of the *laboratory* feature. Thus, we suggest three improvements:

- **Enhancing Explainability**: The *highlighting* feature and *dialogue cards* in VIVID need to offer greater explainability to the instructors. P7 highlighted that having prior knowledge of each feature's exact functionality could have led to more frequent and appropriate usage, potentially resulting in a higher level of satisfaction with the system. Notably, there is a need to investigate the types of information instructors require to effectively discern the diversity among learners and pedagogical dialogue patterns. We observed substantial differences between instructors in their ability to recognize differences in direct learners' understanding levels and how these differences are reflected in dialogue structure.

- **Providing Fine-grained Controllability**: Enhancing controllability and providing granular modifications on the *laboratory* feature can improve instructors' workflow. In our user study, we observed that instructors exhibited varying expectations for the modified versions offered by the *laboratory*, and they tended to rate usability lower when their expectations were not met. The improved version of *laboratory* feature could support the instructors in determining and expressing what features they would expect in the revised versions of the dialogue. For instance, enabling instructors to select elements, such as diverse versions of examples, questions, or versions with added prior knowledge with interactive guidance, could increase the perceived usefulness of the feature.

- **Improving Verbosity**: One unexpected downside was that the generated dialogues were perceived to be verbose, which is likely due to LLM's tendency to produce long text. This issue could be addressed by revising the prompting pipeline to limit the length of utterances and dialogues generated, which we leave as future work.

7.1 Potential Applications beyond the Video Lecture Context

In the educational context, dialogues can serve multiple roles, extending beyond the mere transmission of factual knowledge. In our user study, several instructors highlighted the adaptability of our dialogue design pipeline, suggesting its potential application in diverse learning contexts, instructional materials, and various learning stages. For instance, P7 proposed utilizing our dialogue design pipeline to use dialogues for the learner's review, or to use dialogue as a means to diagnose the learner's misconceptions by providing a dialogue in which the direct learner presents misconceptions.

Furthermore, VIVID and its process of transforming lectures into a dyadic format may take the role of a valuable active learning tool. Our dialogue design pipeline can be utilized in formulating questions in a dialogue format for learners and provide interactive guidance for the students' self-learning process using digital textbooks or in flipped learning settings. Learners can gain a better understanding of complex concepts by analyzing educational content and exploring effective teaching strategies.

7.2 Customizable VIVID for learners

VIVID is a system that supports instructors in transforming their lecture videos into educational dialogues in text format. Yet, it is important to consider how these dialogues can be seamlessly incorporated into the video learning environment (VLE) to enrich learners' experiences and optimize learning outcomes. We can integrate text-format dialogue into the VLE by delivering it in voice and text modes together, utilizing the VLE's multi-modality. For instance, the dialogue can be converted into human-like speech and played alongside the corresponding lecture clip by replacing the original explanations with dialogue. Furthermore, vicarious learners can simultaneously explore multimodal dialogue incorporating formulas in the lecture within a chat-like interface.

While VIVID, designed for instructors, utilizes data pertinent to the levels of a vicarious learner group considered by instructors, it is limited when incorporating teaching strategies, like transfer learning (DR2 in Section 4.2) and personalization of dialogue, which demand each vicarious learner's data, such as prior knowledge, personal background, and current understanding state. We believe that VIVID can be extended to collect data from vicarious learners by using a multi-modal representation of vicarious dialogue. This would enable customized modeling of direct learners, effective transfer learning, and personalization to vicarious learning. For instance, we can collect the data for generating personalized dialogue by requiring learners to click on challenging elements such as formulas or explanations within a lecture as they watch it. Therefore, future work should expand VIVID to include learners and evaluate dialogues from the learner-centered criteria, such as engagement and learning gain.

7.3 Generalizability of VIVID

Even when different lectures cover the same concept, variables such as material modalities, style of delivery, and language affect how a learner perceives and understands new knowledge. We found that instructors tend to adjust dialogue to fit their teaching style when the teaching style in the lecture differs from their preference. To enable instructors to use lecture videos of any teaching style and match them with their intended outcomes, it is necessary to explore a solution for converting dialogue, which includes a preprocessing step for scripting before the *Initial Generation* phase. To design dialogues based on lectures with varying teaching styles, VIVID needs to preprocess the lecture material to isolate core concepts, understand the instructor's intention, and transform the knowledge into a personalized format that matches the user's preferred teaching style.

Moreover, it is important to determine which lecture segments and lengths are suitable for a dialogue style. As P3 said, certain contents or subjects may be more suitable for dialogue formats to help learners better understand relatively complex concepts or examples. Further, we observed in our technical evaluation that the dialogue generation had varying degrees of improvement depending on the subject matter. Thus, enhancing the advantages of dialogue format can be achieved by understanding and reflecting on the differing effects of dialogue format between subjects in dialogue design.

7.4 Limitations and Future Work

We acknowledge several limitations in our current study. Firstly, the knowledge progression of the direct learner in the dialogue was not one-sided in VIVID. VIVID didn't consider prerequisite relationships to create diverse dialogues (Section 5.1.2). Yet, some dialogues depicted direct learners initially understanding a concept but later appearing to lack understanding. Thus, redesigning the knowledge state setting pipeline is needed to maintain consistent knowledge levels and prevent reverse progression. Secondly, our experiments involved instructors designing dialogues for only a single segment within a lecture. However, the generated dialogues are influenced by factors such as the length of the selected segment, the type of content, and the subject. To explore VIVID's use cases more deeply, it is necessary to conduct experiments under a more diverse set of conditions.

Chapter 8. Conclusion

We present design recommendations from an extensive literature review and insights gathered during a design workshop. These recommendations are aimed at facilitating the creation of high-quality educational dialogues. To put these guidelines into practice, we have developed VIVID, a web application designed to assist instructors in authoring pedagogical dialogues from their monologue-style lecture videos. Through our technical evaluation and user study, we found that instructors can consider important factors in dialogue design effectively, generating *Dynamic, Pedagogically productive, Immersive*, and *Correct* dialogues. We hope VIVID helps create more engaging lecture videos, providing a personalized learning experience for online students.

Chapter 9. Appendix

9.1 Workshop Details

9.1.1 Subjects and Lesson Content Used In Workshop

Table 9.1 indicates the subjects and lesson contents used in our workshop.

9.1.2 Teaching Strategies for Designing Pedagogically Effective Dialogue

Table 9.2 indicates the teaching strategies for designing pedagogically effective dialogue.

9.2 Dialogue Generation Examples

9.2.1 Evaluation Dataset Generation Process

Figure 9.1 indicates the evaluation dataset generation process.

9.2.2 Transcript Example

The green section indicates the area highlighted by the authors as potentially challenging for vicarious learners to understand. VIVID creates a direct learner who lacks knowledge of this green area.

9.2.3 Example 1

Figure 9.3 and Figure 9.4 are two dialogue examples generated based on the English script of the physics lecture.

9.2.4 Example 2

Figure 9.5 and Figure 9.6 are two dialogue examples generated from the Korean script of the physics lecture. We translated them into English as the output was Korean.

Table 9.1: Subjects and lesson contents of the lecture that were addressed in the workshop by each group.

Group	Subject	Main lecture content
1 (P1, P2)	Math	Composite functions and inverse functions
2 (P3, P4)	Science	Phases of the moon and the reasons behind these lunar phases
3 (P5, P6)	Math	Concept of unit vectors and their alignment with a given vector
4 (P7, P8)	Science	Einstein's General Theory of Relativity, covering concepts such as the warping of spacetime but massive celestial bodies, gravitational lensing, time dilation due to gravity, and phenomena associated with black holes
5 (P9, P10)	Math	Concepts of radical (square root) functions and rational (fractional) functions
6 (P11, P12)	Math	Process of transforming a quadratic equation into a perfect square trinomial and then using square roots to find the solutions
7 (P13, P14)	Math	Method of expressing a third line passing through the intersection of two given lines and determining the equation of a line, even with an unknown slope, passing through a specified point
8 (P15)	Math	Classification of integers based on the remainders when divided by a positive integer

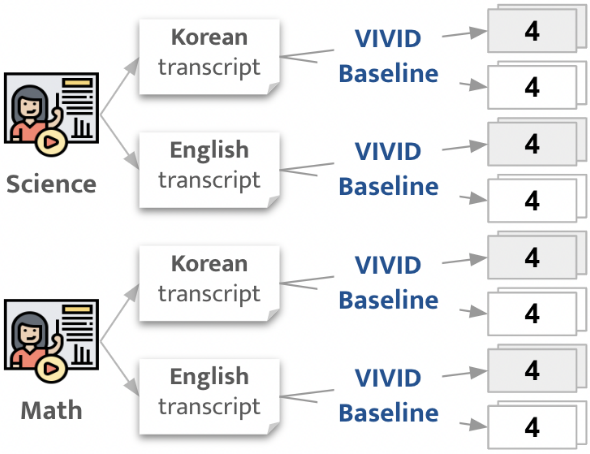


Figure 9.1: Evaluation Dataset Generation Process.

Table 9.2: Four teaching strategies for pedagogically effective vicarious dialogues: Cognitive conflict, Metacognitive prompting, Cognitive prompting, and Spontaneous deep-level reasoning question.

By initiative	Key strategies that can be effective to vicarious learners	Description	Example Dialogues between a tutor and a tutee	
Tutor	Cognitive conflict	It is a teaching strategy that examines the learner's prior knowledge, creates a mismatch situation that causes conflict, and then helps the learners to see that his or her understanding is incorrect.	**Tutor** **Tutee** **Tutor**	Absolutely, that's the usual method. But let me throw a curveball. What if I told you that solving them using a different approach might lead to a different solution? Really? I thought there was only one way to solve equations. That's what we're here to explore! Let's try this. Instead of isolating x right away, . . .
	Metacognitive prompting	It orients learners towards higher-level strategies (e.g., goal-setting, planning, monitoring, evaluation, reflection). It includes an instructor's utterances that encourage the learner to express their current level of understanding or articulate their thought process.	**Tutor** **Tutee** **Tutor**	Got it. How about we take a slightly different approach this time? Before you jump into solving, let's start by identifying what the problem is asking. Can you read the question and tell me what this question is requesting? Sure. It's asking me to solve for the sum of 'x' and 'y' in the equation. That's a nice interpretation, but let's take a closer look.
	Cognitive prompting	It engages learners in lower-level strategies (e.g., organization, rehearsal, elaboration). It includes the instructor's utterances that prompt the learner to talk about what they are learning or to draw out the learner's prior knowledge and personal experiences.	**Tutor** **Tutee**	As you work through an equation, think about the basic operations you've learned. Can you explain how these operations are helping you manipulate this equation? Sure. When there's addition on one side, I subtract to balance it out. And if it's multiplication, I divide to get 'x' by itself.
Tutee	Spontaneous deep-level reasoning question	It refers to starting a conversation where the learner spontaneously asks deep-level reasoning questions that help them better understand and engage in critical thinking.	**Tutee** **Tutor**	How can a manufacturer increase the speed of the computer? What can they do to make it faster? Well, one thing manufacturers do is increase the clock speed of the computer.

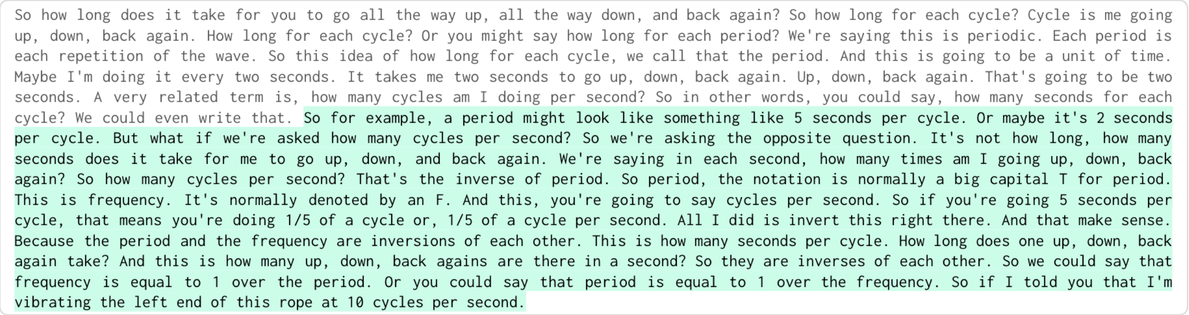
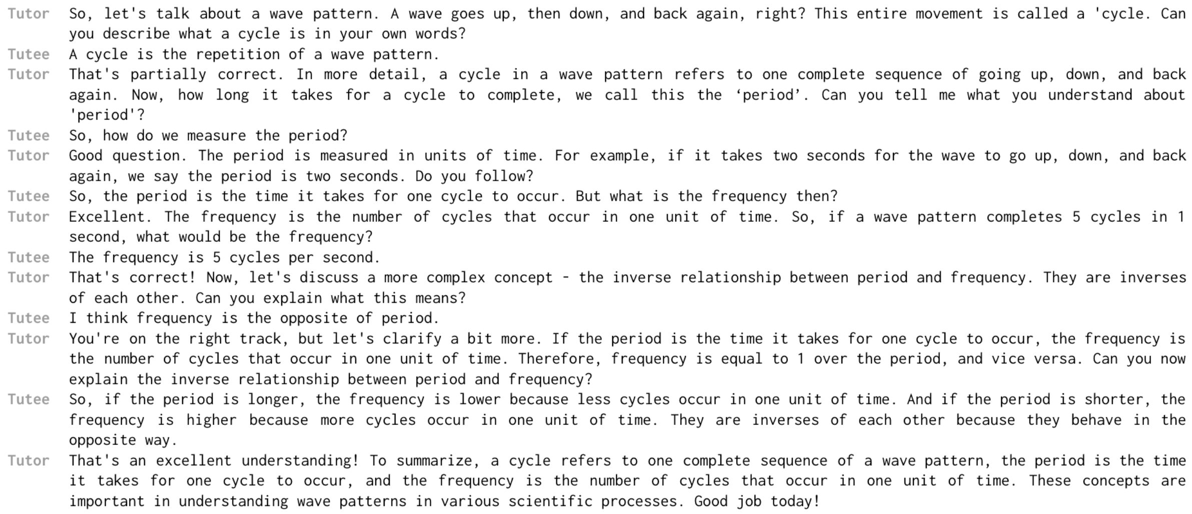

So how long does it take for you to go all the way up, all the way down, and back again? So how long for each cycle? Cycle is me going up, down, back again. How long for each cycle? Or you might say how long for each period? We're saying this is periodic. Each period is each repetition of the wave. So this idea of how long for each cycle, we call that the period. And this is going to be a unit of time. Maybe I'm doing it every two seconds. It takes me two seconds to go up, down, back again. Up, down, back again. That's going to be two seconds. A very related term is, how many cycles am I doing per second? So in other words, you could say, how many seconds for each cycle? We could even write that. So for example, a period might look like something like 5 seconds per cycle. Or maybe it's 2 seconds per cycle. But what if we're asked how many cycles per second? So we're asking the opposite question. It's not how long, how many seconds does it take for me to go up, down, and back again. We're saying in each second, how many times am I going up, down, back again? So how many cycles per second? That's the inverse of period. So period, the notation is normally a big capital T for period. This is frequency. It's normally denoted by an F. And this, you're going to say cycles per second. So if you're going 5 seconds per cycle, that means you're doing 1/5 of a cycle or, 1/5 of a cycle per second. All I did is invert this right there. And that make sense. Because the period and the frequency are inversions of each other. This is how many seconds per cycle. How long does one up, down, back again take? And this is how many up, down, back agains are there in a second? So they are inverses of each other. So we could say that frequency is equal to 1 over the period. Or you could say that period is equal to 1 over the frequency. So if I told you that I'm vibrating the left end of this rope at 10 cycles per second.

Figure 9.2: A transcript that is used to generate dialogue data for technical evaluation study (Figure 9.3, 9.4, 9.5, 9.6). This transcript is from a physics lecture used in the technical evaluation study. The physics lecture from which this script was derived supports Korean and English scripts, enabling the generation of dialogues in both languages, and this is an English transcript. The authors highlighted the *green part* as potentially difficult for vicarious learners to understand when generating dialogue data.



Tutor So, let's talk about a wave pattern. A wave goes up, then down, and back again, right? This entire movement is called a 'cycle. Can you describe what a cycle is in your own words?

Tutee A cycle is the repetition of a wave pattern.

Tutor That's partially correct. In more detail, a cycle in a wave pattern refers to one complete sequence of going up, down, and back again. Now, how long it takes for a cycle to complete, we call this the ‘period’. Can you tell me what you understand about 'period'?

Tutee So, how do we measure the period?

Tutor Good question. The period is measured in units of time. For example, if it takes two seconds for the wave to go up, down, and back again, we say the period is two seconds. Do you follow?

Tutee So, the period is the time it takes for one cycle to occur. But what is the frequency then?

Tutor Excellent. The frequency is the number of cycles that occur in one unit of time. So, if a wave pattern completes 5 cycles in 1 second, what would be the frequency?

Tutee The frequency is 5 cycles per second.

Tutor That's correct! Now, let's discuss a more complex concept - the inverse relationship between period and frequency. They are inverses of each other. Can you explain what this means?

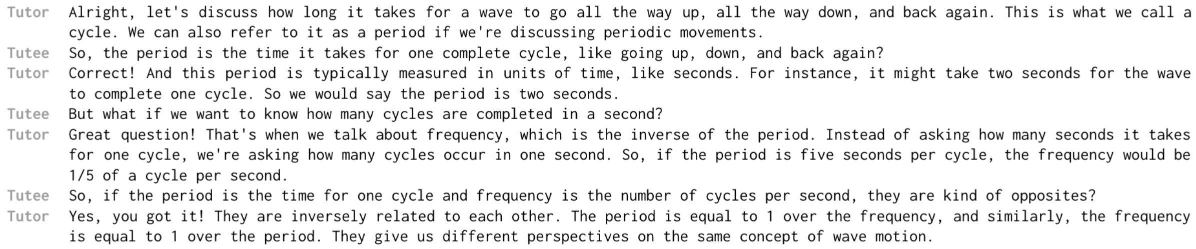
Tutee I think frequency is the opposite of period.

Tutor You're on the right track, but let's clarify a bit more. If the period is the time it takes for one cycle to occur, the frequency is the number of cycles that occur in one unit of time. Therefore, frequency is equal to 1 over the period, and vice versa. Can you now explain the inverse relationship between period and frequency?

Tutee So, if the period is longer, the frequency is lower because less cycles occur in one unit of time. And if the period is shorter, the frequency is higher because more cycles occur in one unit of time. They are inverses of each other because they behave in the opposite way.

Tutor That's an excellent understanding! To summarize, a cycle refers to one complete sequence of a wave pattern, the period is the time it takes for one cycle to occur, and the frequency is the number of cycles that occur in one unit of time. These concepts are important in understanding wave patterns in various scientific processes. Good job today!

Figure 9.3: An example of technical evaluation dataset. This dialogue is one of the 16 examples generated by **VIVID** to evaluate the *Initial Generation* phase (in Section 5.1) during the Technical Evaluation study (in Section 6.3). It is based on the **English** script of the first physics lecture.



Tutor Alright, let's discuss how long it takes for a wave to go all the way up, all the way down, and back again. This is what we call a cycle. We can also refer to it as a period if we're discussing periodic movements.

Tutee So, the period is the time it takes for one complete cycle, like going up, down, and back again?

Tutor Correct! And this period is typically measured in units of time, like seconds. For instance, it might take two seconds for the wave to complete one cycle. So we would say the period is two seconds.

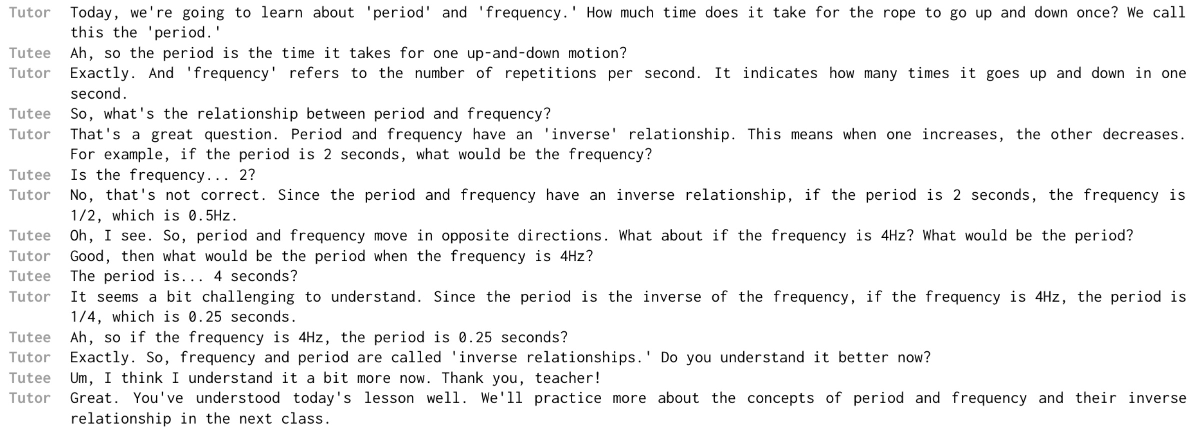
Tutee But what if we want to know how many cycles are completed in a second?

Tutor Great question! That's when we talk about frequency, which is the inverse of the period. Instead of asking how many seconds it takes for one cycle, we're asking how many cycles occur in one second. So, if the period is five seconds per cycle, the frequency would be 1/5 of a cycle per second.

Tutee So, if the period is the time for one cycle and frequency is the number of cycles per second, they are kind of opposites?

Tutor Yes, you got it! They are inversely related to each other. The period is equal to 1 over the frequency, and similarly, the frequency is equal to 1 over the period. They give us different perspectives on the same concept of wave motion.

Figure 9.4: An example of technical evaluation dataset. This dialogue is one of the 16 examples generated by **Baseline** to evaluate the *Initial Generation* phase (in Section 5.1) during the Technical Evaluation study (in Section 6.3). It is based on the **English** script of the first physics lecture.



Tutor Today, we're going to learn about 'period' and 'frequency.' How much time does it take for the rope to go up and down once? We call this the 'period.'

Tutee Ah, so the period is the time it takes for one up-and-down motion?

Tutor Exactly. And 'frequency' refers to the number of repetitions per second. It indicates how many times it goes up and down in one second.

Tutee So, what's the relationship between period and frequency?

Tutor That's a great question. Period and frequency have an 'inverse' relationship. This means when one increases, the other decreases. For example, if the period is 2 seconds, what would be the frequency?

Tutee Is the frequency... 2?

Tutor No, that's not correct. Since the period and frequency have an inverse relationship, if the period is 2 seconds, the frequency is 1/2, which is 0.5Hz.

Tutee Oh, I see. So, period and frequency move in opposite directions. What about if the frequency is 4Hz? What would be the period?

Tutor Good, then what would be the period when the frequency is 4Hz?

Tutee The period is... 4 seconds?

Tutor It seems a bit challenging to understand. Since the period is the inverse of the frequency, if the frequency is 4Hz, the period is 1/4, which is 0.25 seconds.

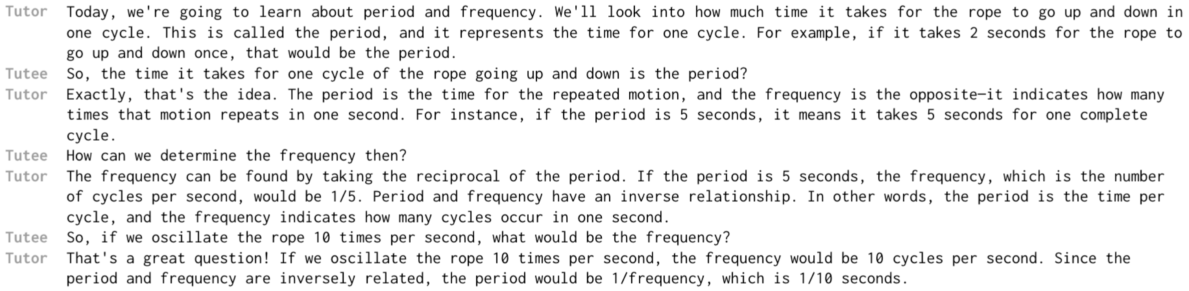
Tutee Ah, so if the frequency is 4Hz, the period is 0.25 seconds?

Tutor Exactly. So, frequency and period are called 'inverse relationships.' Do you understand it better now?

Tutee Um, I think I understand it a bit more now. Thank you, teacher!

Tutor Great. You've understood today's lesson well. We'll practice more about the concepts of period and frequency and their inverse relationship in the next class.

Figure 9.5: An example of technical evaluation dataset. This dialogue is one of the 16 examples generated by **VIVID** to evaluate the *Initial Generation* phase (in Section 5.1) during the Technical Evaluation study (in Section 6.3). It is based on the Korean script of the first physics lecture. Thus, this example is a translation of a dialogue generated in **Korean**.



Tutor Today, we're going to learn about period and frequency. We'll look into how much time it takes for the rope to go up and down in one cycle. This is called the period, and it represents the time for one cycle. For example, if it takes 2 seconds for the rope to go up and down once, that would be the period.

Tutee So, the time it takes for one cycle of the rope going up and down is the period?

Tutor Exactly, that's the idea. The period is the time for the repeated motion, and the frequency is the opposite—it indicates how many times that motion repeats in one second. For instance, if the period is 5 seconds, it means it takes 5 seconds for one complete cycle.

Tutee How can we determine the frequency then?

Tutor The frequency can be found by taking the reciprocal of the period. If the period is 5 seconds, the frequency, which is the number of cycles per second, would be 1/5. Period and frequency have an inverse relationship. In other words, the period is the time per cycle, and the frequency indicates how many cycles occur in one second.

Tutee So, if we oscillate the rope 10 times per second, what would be the frequency?

Tutor That's a great question! If we oscillate the rope 10 times per second, the frequency would be 10 cycles per second. Since the period and frequency are inversely related, the period would be 1/frequency, which is 1/10 seconds.

Figure 9.6: An example of technical evaluation dataset. This dialogue is one of the 16 examples generated by **Baseline** to evaluate the *Initial Generation* phase (in Section 5.1) during the Technical Evaluation study (in Section 6.3). It is based on the Korean script of the first physics lecture. Thus, this example is a translation of a dialogue generated in **Korean**.

Bibliography

[1] J. R. Anderson, C. F. Boyle, and B. J. Reiser. Intelligent tutoring systems. *Science*, 228(4698):456–462, 1985.

[2] A. Bandura and R. H. Walters. *Social learning theory*, volume 1. Englewood cliffs Prentice Hall, 1977.

[3] G. Bansal. Teacher discursive moves: Conceptualising a schema of dialogic discourse in science classrooms. *International Journal of Science Education*, 40(15):1891–1912, 2018.

[4] M. Beheshti, A. Taspolat, O. S. Kaya, and H. F. Sapanca. Characteristics of instructional videos. *World Journal on Educational Technology: Current Issues*, 10(1):61–69, 2018.

[5] K. Boyer, R. Phillips, M. Wallis, M. Vouk, and J. Lester. Learner characteristics and feedback in tutorial dialogue. In *Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications*, pages 53–61, 2008.

[6] K. E. Boyer, E. Ha, M. D. Wallis, R. Phillips, M. A. Vouk, and J. C. Lester. Discovering tutorial dialogue strategies with hidden markov models. In *AIED*, pages 141–148, 2009.

[7] K. E. Boyer, R. Phillips, M. Wallis, M. Vouk, and J. Lester. Balancing cognitive and motivational scaffolding in tutorial dialogue. In *Intelligent Tutoring Systems: 9th International Conference, ITS 2008, Montreal, Canada, June 23-27, 2008 Proceedings 9*, pages 239–249. Springer, 2008.

[8] H. D. Brecht. Learning from online video lectures. *Journal of Information Technology Education. Innovations in Practice*, 11:227, 2012.

[9] K. Brodie. Working with learners' mathematical thinking: Towards a language of description for changing pedagogy. *Teaching and Teacher Education*, 27(1):174–186, 2011.

[10] W. L. Cade, J. L. Copeland, N. K. Person, and S. K. D'Mello. Dialogue modes in expert tutoring. In *Intelligent Tutoring Systems: 9th International Conference, ITS 2008, Montreal, Canada, June 23-27, 2008 Proceedings 9*, pages 470–479. Springer, 2008.

[11] M. Chi, S. Kang, and D. Yaghmourian. Why students learn more from dialogue- than monologue-videos: Analyses of peer interactions. *Journal of the Learning Sciences*, 26, 06 2016.

[12] M. Chi, M. Roy, and R. Hausmann. Observing tutorial dialogues collaboratively: Insights about human tutoring effectiveness from vicarious learning. *Cognitive science*, 32:301–41, 03 2008.

[13] M. T. Chi, S. A. Siler, H. Jeong, T. Yamauchi, and R. G. Hausmann. Learning from human tutoring. *Cognitive science*, 25(4):471–533, 2001.

[14] C. Chin. Classroom interaction in science: Teacher questioning and feedback to students' responses. *International journal of science education*, 28(11):1315–1346, 2006.

[15] K. Chorianopoulos. A taxonomy of asynchronous instructional video styles. *The International Review of Research in Open and Distributed Learning*, 19(1), 2 2018.

[16] S. D. Craig, B. Gholson, J. K. Brittingham, J. L. Williams, and K. T. Shubeck. Promoting vicarious learning of physics using deep questions with explanations. *Computers & Education*, 58(4):1042–1048, 2012.

[17] S. D. Craig, J. Sullins, A. Witherspoon, and B. Gholson. The deep-level-reasoning-question effect: The role of dialogue and deep-level-reasoning questions during vicarious learning. *Cognition and Instruction*, 24(4):565–591, 2006.

[18] P. Denny, H. Khosravi, A. Hellas, J. Leinonen, and S. Sarsa. Human vs machine: Comparison of student-generated and ai-generated educational content. *arXiv preprint arXiv:2306.10509*, 2023.

[19] L. Ding, K. Cooper, M. Stephens, M. Chi, and S. Brownell. Learning from error episodes in dialogue-videos: The influence of prior knowledge. *Australasian Journal of Educational Technology*, pages 20–32, 03 2021.

[20] S. D'Mello, B. Lehman, and N. Person. Expert tutors feedback is immediate, direct, and discriminating. In *Twenty-Third International FLAIRS Conference*, 2010.

[21] D. M. Driscoll, S. D. Craig, B. Gholson, M. Ventura, X. Hu, and A. C. Graesser. Vicarious learning: Effects of overhearing dialog and monologue-like discourse in a virtual tutoring session. *Journal of Educational Computing Research*, 29(4):431–450, 2003.

[22] D. M. Driscoll, S. D. Craig, B. Gholson, M. Ventura, X. Hu, and A. C. Graesser. Vicarious learning: Effects of overhearing dialog and monologue-like discourse in a virtual tutoring session. *Journal of Educational Computing Research*, 29(4):431–450, 2003.

[23] I. Dubovi and V. R. Lee. Instructional support for learning with agent-based simulations: A tale of vicarious and guided exploration learning approaches. *Computers & Education*, 142:103644, 2019.

[24] M. S. Ellman and M. L. Schwartz. Article commentary: Online learning tools as supplements for basic and clinical science education. *Journal of medical education and curricular development*, 3:JMECD–S18933, 2016.

[25] E. Fast, B. Chen, J. Mendelsohn, J. Bassen, and M. S. Bernstein. Iris: A conversational agent for complex tasks. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–12, 2018.

[26] M. Forehand. Bloom's taxonomy. *Emerging perspectives on learning, teaching, and technology*, 41(4):47–56, 2010.

[27] B. Gholson, A. Witherspoon, B. Morgan, J. K. Brittingham, R. Coles, A. C. Graesser, J. Sullins, and S. D. Craig. Exploring the deep-level reasoning questions effect during vicarious learning among eighth to eleventh graders in the domains of computer literacy and newtonian physics. *Instructional Science*, 37:487–493, 2009.

[28] A. C. Graesser and S. D'Mello. Emotions during the learning of difficult material. In *Psychology of learning and motivation*, volume 57, pages 183–225. Elsevier, 2012.

[29] A. C. Graesser, S. Lu, G. T. Jackson, H. H. Mitchell, M. Ventura, A. Olney, and M. M. Louwerse. Autotutor: A tutor with dialogue in natural language. *Behavior Research Methods, Instruments, & Computers*, 36:180–192, 2004.

[30] A. C. Graesser, K. Wiemer-Hastings, P. Wiemer-Hastings, and R. Kreuz. Autotutor: A simulation of a human tutor. *Cognitive Systems Research*, 1(1):35–51, 1999.

[31] J. Grossman, Z. Lin, H. Sheng, J. T.-Z. Wei, J. J. Williams, and S. Goel. Mathbot: Transforming online resources for learning math into conversational interactions. *AAAI 2019 Story-Enabled Intelligence*, 2019.

[32] G. Hume, J. Michael, A. Rovick, and M. Evens. Hinting as a tactic in one-on-one tutoring. *The Journal of the Learning Sciences*, 5(1):23–47, 1996.

[33] S. D. Ivie. Ausubel's learning theory: An approach to teaching higher order thinking skills. *The High School Journal*, 82(1):35–42, 1998.

[34] P. Kranzfelder, J. L. Bankers-Fulbright, M. E. García-Ojeda, M. Melloy, S. Mohammed, and A.-R. M. Warfa. The classroom discourse observation protocol (cdop): A quantitative method for characterizing teacher discourse moves in undergraduate stem learning environments. *PloS one*, 14(7):e0219019, 2019.

[35] J. A. Kulik and J. Fletcher. Effectiveness of intelligent tutoring systems: a meta-analytic review. *Review of educational research*, 86(1):42–78, 2016.

[36] J. Lee, C. Lim, and H. Kim. Development of an instructional design model for flipped learning in higher education. *Educational Technology Research and Development*, 65:427–453, 2017.

[37] K. J. Lee, A. Chauhan, J. Goh, E. Nilsen, and E. Law. Curiosity notebook: the design of a research platform for learning by teaching. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–26, 2021.

[38] Y. Lee, J. J. Y. Chung, T. S. Kim, J. Y. Song, and J. Kim. Promptiverse: Scalable generation of scaffolding prompts through human-ai hybrid knowledge graph annotation. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–18, 2022.

[39] B. Lehman, S. D'Mello, and A. Graesser. Confusion and complex learning during interactions with computer learning environments. *The Internet and Higher Education*, 15(3):184–194, 2012.

[40] J. Leinonen, P. Denny, S. MacNeil, S. Sarsa, S. Bernstein, J. Kim, A. Tran, and A. Hellas. Comparing code explanations created by students and large language models. *arXiv preprint arXiv:2304.03938*, 2023.

[41] B. L. Levy, E. E. Thomas, K. Drago, and L. A. Rex. Examining studies of inquiry-based learning in three fields of education: Sparking generative conversation. *Journal of teacher education*, 64(5):387–408, 2013.

[42] D. Litman and K. Forbes-Riley. Correlations between dialogue acts and learning in spoken tutoring dialogues. *Natural Language Engineering*, 12(2):161–176, 2006.

[43] T. Lord and S. Baviskar. Moving students from information recitation to information understanding-exploiting bloom's taxonomy in creating science questions. *Journal of College Science Teaching*, 36(5):40, 2007.

[44] X. Lu, B. Di Eugenio, T. C. Kershaw, S. Ohlsson, and A. Corrigan-Halpern. Expert vs. non-expert tutoring: Dialogue moves, interaction patterns and multi-utterance turns. In *Computational Linguistics and Intelligent Text Processing: 8th International Conference, CICLing 2007, Mexico City, Mexico, February 18-24, 2007. Proceedings 8*, pages 456–467. Springer, 2007.

[45] X. Lu, S. Fan, J. Houghton, L. Wang, and X. Wang. Readingquizmaker: A human-nlp collaborative system that supports instructors to design high-quality reading quiz questions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–18, 2023.

[46] S. Lyle. Dialogic teaching: Discussing theoretical contexts and reviewing evidence from classroom practice. *Language and education*, 22(3):222–240, 2008.

[47] W. Ma, O. O. Adesope, J. C. Nesbit, and Q. Liu. Intelligent tutoring systems and learning outcomes: A meta-analysis. *Journal of educational psychology*, 106(4):901, 2014.

[48] R. Martinez-Maldonado, A. Clayphan, K. Yacef, and J. Kay. Mtfeedback: providing notifications to enhance teacher awareness of small group work in the classroom. *IEEE Transactions on Learning Technologies*, 8(2):187–200, 2014.

[49] J. T. Mayes. Still to learn from vicarious learning. *E-learning and digital media*, 12(3-4):361–371, 2015.

[50] A. Mitrovic, S. Ohlsson, and D. K. Barrow. The effect of positive feedback in a constraint-based intelligent tutoring system. *Computers & Education*, 60(1):264–272, 2013.

[51] D. Moher, A. Liberati, J. Tetzlaff, and D. Altman. Preferred reporting items for systematic reviews and meta-analyses: the prisma statement. *Br Med J*, 8:336–341, 07 2009.

[52] H. Muhonen, E. Pakarinen, A.-M. Poikkeus, M.-K. Lerkkanen, and H. Rasku-Puttonen. Quality of educational dialogue and association with students' academic performance. *Learning and Instruction*, 55:67–79, 2018.

[53] H. Muhonen, H. Rasku-Puttonen, E. Pakarinen, A.-M. Poikkeus, and M.-K. Lerkkanen. Scaffolding through dialogic teaching in early school classrooms. *Teaching and teacher education*, 55:143–154, 2016.

[54] K. Muldner, K. Dybvig, R. Lam, and M. Chi. Learning by observing tutorial dialogue versus monologue collaboratively or alone. 01 2011.

[55] A. Nugraha, T. Harada, I. A. Wahono, and T. Inoue. A tool to add a tutee agent in a monologue lecture video improves students' watching experience. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–7, 2020.

[56] K. K. A. Ong, C. E. Hart, and P. K. Chen. Promoting higher-order thinking through teacher questioning: a case study of a singapore science classroom. *New Waves-Educational Research and Development Journal*, 19(1):1–19, 2016.

[57] J. Paladines and J. Ramirez. A systematic literature review of intelligent tutoring systems with dialogue in natural language. *IEEE Access*, 8:164246–164267, 2020.

[58] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR, 2023.

[59] C. P. Rose, D. Bhembe, S. Siler, R. Srivastava, and K. VanLehn. The role of why questions in effective human tutoring. In *Proceedings of AIED*, volume 3, 2003.

[60] S. Ruan, L. Jiang, J. Xu, B. J.-K. Tham, Z. Qiu, Y. Zhu, E. L. Murnane, E. Brunskill, and J. A. Landay. Quizbot: A dialogue-based adaptive learning system for factual knowledge. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2019.

[61] S. Sarsa, P. Denny, A. Hellas, and J. Leinonen. Automatic generation of programming exercises and code explanations using large language models. In *Proceedings of the 2022 ACM Conference on International Computing Education Research-Volume 1*, pages 27–43, 2022.

[62] P. Sridhar, A. Doyle, A. Agarwal, C. Bogart, J. Savelka, and M. Sakr. Harnessing llms in curricular design: Using gpt-4 to support authoring of learning objectives. *arXiv preprint arXiv:2306.17459*, 2023.

[63] M. Stanislaw Paul. World class stem - benchmarking and delivering based on evidence based cognitive science. In *2021 IEEE International Conference on Engineering, Technology & Education (TALE)*, pages 1139–1144, 2021.

[64] S. Suhirman and S. Prayogi. Overcoming challenges in stem education: A literature review that leads to effective pedagogy in stem learning. *Jurnal Penelitian Pendidikan IPA*, 9(8):432–443, 2023.

[65] L. A. Sutton. The principle of vicarious interaction in computer-mediated communications. *International Journal of Educational Telecommunications*, 7(3):223–242, 2001.

[66] K. Swan. Learning effectiveness online: What the research tells us. *Elements of quality online education, practice and direction*, 4(1):13–47, 2003.

[67] T. Tanprasert, S. S. Fels, L. Sinnamon, and D. Yoon. Authoring virtual peer interactions for lecture videos. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, pages 1–7, 2022.

[68] T. Tanprasert, S. S. Fels, L. Sinnamon, and D. Yoon. Scripted vicarious dialogues: Educational video augmentation method for increasing isolated students' engagement. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–25, 2023.

[69] P. Teo. Exploring the dialogic space in teaching: A study of teacher talk in the pre-university classroom in singapore. *Teaching and Teacher Education*, 56:47–60, 2016.

[70] K. VanLehn, S. Siler, C. Murray, T. Yamauchi, and W. B. Baggett. Why do only some events cause learning during human tutoring? *Cognition and Instruction*, 21(3):209–249, 2003.

[71] M. Vrikki, L. Wheatley, C. Howe, S. Hennessy, and N. Mercer. Dialogic practices in primary school classrooms. *Language and Education*, 33(1):85–100, 2019.

[72] T. Wambsganss, T. Kueng, M. Soellner, and J. M. Leimeister. Arguetutor: An adaptive dialog-based learning system for argumentation skills. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–13, 2021.

[73] X. Wang, C. Rose, and K. Koedinger. Seeing beyond expert blind spots: Online learning design for scale and quality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2021.

[74] Z. Wang, A. S. Lan, W. Nie, A. E. Waters, P. J. Grimaldi, and R. G. Baraniuk. Qg-net: a data-driven question generation model for educational content. In *Proceedings of the fifth annual ACM conference on learning at scale*, pages 1–10, 2018.

[75] A. Willis, G. Davis, S. Ruan, L. Manoharan, J. Landay, and E. Brunskill. Key phrase extraction for generating educational question-answer pairs. In *Proceedings of the Sixth (2019) ACM Conference on Learning@ Scale*, pages 1–10, 2019.

[76] R. Winkler, S. Hobert, A. Salovaara, M. Söllner, and J. M. Leimeister. Sara, the lecturer: Improving learning in online education with a scaffolding-based conversational agent. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–14, 2020.

[77] H. Ye, T. Liu, A. Zhang, W. Hua, and W. Jia. Cognitive mirage: A review of hallucinations in large language models. *arXiv preprint arXiv:2309.06794*, 2023.

[78] Y. Zhang, Y. Li, L. Cui, D. Cai, L. Liu, T. Fu, X. Huang, E. Zhao, Y. Zhang, Y. Chen, et al. Siren's song in the ai ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*, 2023.

Acknowledgment

First and foremost, I want to express my gratitude to my advisor, Juho Kim, for allowing me to dream an HCI researcher. I could work in an exceptional research environment, meet wonderful and amazing people, and learn how to do research. I am deeply thankful for your unwavering belief in me, amazing guidance, and encouragement throughout this journey. I want to express many thanks to all the members of KIXLAB for their continuous support and encouragement. I would especially like to thank my fellow master's students (Yoonsu, Da**E**un, Jeongeon, and Juhoon) who made unforgettable memories for me. Lastly, I am profoundly grateful to my family for their steadfast belief in me and their support.

Curriculum Vitae in Korean

이름: 최슬기

생년월일: 1998년 06월 06일

전자주소: igules8925@kaist.ac.kr

학력

2014. 3. – 2017. 2.	포항여자고등학교
2017. 3. – 2022. 2.	한동대학교 컴퓨터공학과 (학사)
2022. 3. – 2024. 8.	한국과학기술원 전산학부 (석사)

연구업적

1. Lim, H., Cho, J., Kim, T., Park, J., Shin, H., **Choi, S.**, Park, S., Lee, K., Kim, J., Lee, M., & Hong, H. (2024). Co-Creating Question-and-Answer Style Articles with Large Language Models for Research Promotion. Proceedings of the 2024 ACM Designing Interactive Systems Conference (DIS). (To Appear)
2. Shin, H., **Choi, S.**, Cho, J., Admoni, S., Lim, H., Kim, T., Hong, H., Lee, M., & Kim, J. (2024). Towards an Evaluation of LLM-Generated Inspiration by Developing and Validating Inspiration Scale.
3. **Choi, S.**, Lee, H., Lee, Y., & Kim, J. (2024). VIVID: Human-AI Collaborative Authoring of Vicarious Dialogues from Lecture Videos. Proceedings of the CHI Conference on Human Factors in Computing Systems.
4. Kim, D., **Choi, S.**, Kim, J., Setlur, V., & Agrawala, M. (2023). EC: A Tool for Guiding Chart and Caption Emphasis. IEEE Transactions on Visualization and Computer Graphics, 30, 120-130.