Penquiry: A Pen-based Interactive In-situ Q&A System Leveraging LLMs

Jeongmin Rhee*
Seoul National University
Seoul, Republic of Korea
jmrhee@hcil.snu.ac.kr

Changhee Lee*
Pohang University
of Science and Technology
Pohang, Republic of Korea
chlee0811@postech.ac.kr

Hyunwoo Kim
Seoul National University
Seoul, Republic of Korea
hyunkim@hcil.snu.ac.kr

Kiroong Choe
Seoul National University
Seoul, Republic of Korea
krchoe@hcil.snu.ac.kr

Bohyoung Kim†
Hankuk University
of Foreign Studies
Yongin-si, Republic of Korea
bkim@hufs.ac.kr

Sungahn Ko†
Pohang University
of Science and Technology
Pohang, Republic of Korea
sungahn@postech.ac.kr

Jinwook Seo†
Seoul National University
Seoul, Republic of Korea
jseo@snu.ac.kr

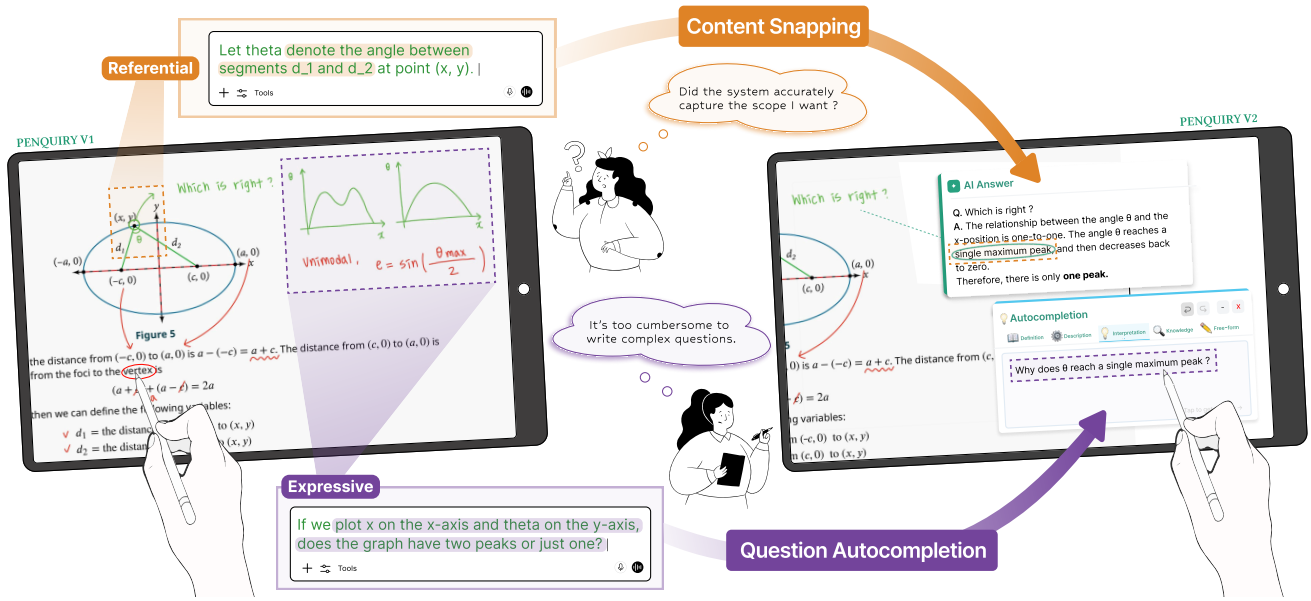


Figure 1: Penquiry is an in-situ Q&A system that enables learners to query directly on digital study materials using a pen. (Left) Penquiry V1 leverages in-situ pen input to mitigate the referential and expressive barriers inherent in a separate keyboard interface for LLM queries. However, it introduces new challenges, such as uncertainty in referential grounding and the physical burden of handwriting lengthy questions. (Right) Penquiry V2 resolves these emergent barriers by introducing Content Snapping, which aligns user sketches with nearby document text elements, and Question Autocompletion, which generates full queries from minimal ink keywords. Together, these mechanisms mediate free-form pen input into LLM-interpretable questions while sustaining inquiry within the learner's pen-centered workflow.

*Both authors contributed equally to this research.
†Corresponding Author.

UIST '26, Detroit, MI, USA

ACM ISBN 979-8-4007-2856-3/2026/11
https://doi.org/10.1145/3830398.3830637

Abstract

Pen-based digital devices remain a preferred medium for active, cognitively engaging study. Concurrently, Large Language Models (LLMs) have become indispensable for self-directed learning, enabling students to clarify concepts. However, a fundamental interaction gap exists between the fluid, spatial nature of pen-based workflows and the discrete, keyboard-heavy requirements of LLMs. We present Penquiry, an in-situ question-and-answer system that bridges this gap by enabling learners to pose questions directly on digital study materials via a pen. We characterize two primary interaction challenges in this multimodal transition: a Referential Barrier, which hinders grounding fine-grained visual elements into the query context, and an Expressive Barrier, which forces learners to translate diverse, non-textual intents—such as equations and diagrams—into rigid, typed sentences. To resolve these, Penquiry introduces a mediation layer featuring Content Snapping for unambiguous referencing and Question Autocompletion to expand sparse ink keywords into rich semantic queries. Through two iterative user studies ($N = 16$ per study), we demonstrate that Penquiry significantly reduces the cognitive and physical overhead of inquiry compared to traditional interfaces, providing a new blueprint for pen-based, in-situ AI interaction.

CCS Concepts

• **Human-centered computing** → **User interface programming**; *Interaction techniques*; Touch screens.

Keywords

In-situ Interaction, Pen-based Interaction, Large Language Models, Pen-based Learning, Human-AI Interaction

ACM Reference Format:

Jeongmin Rhee, Changhee Lee, Hyunwoo Kim, Kiroong Choe, Bohyoung Kim, Sungahn Ko, and Jinwook Seo. 2026. Penquiry: A Pen-based Interactive In-situ Q&A System Leveraging LLMs. In *The 39th Annual ACM Symposium on User Interface Software and Technology (UIST '26), November 02–05, 2026, Detroit, MI, USA.* ACM, New York, NY, USA, 19 pages. https://doi.org/10.1145/3830398.3830637

1 Introduction

Pen-based digital environments remain a preferred method for active learning due to their cognitive benefits [22, 34, 40, 47]. Simultaneously, learners increasingly rely on Large Language Models (LLMs) as indispensable tools for real-time concept clarification [46, 49] during self-directed study. However, integrating these two paradigms reveals a fundamental interaction gap: the fluid and spatial nature of stylus workflows is at odds with the discrete text-heavy requirements of current LLM interfaces. Because querying an LLM typically requires abandoning the document canvas for a separate, keyboard-driven chat window, learners face a severe modality mismatch. This constant context-switching not only fractures the learner's focus but also imposes a high interaction cost, often discouraging inquiry and disrupting the flow of learning [20].

While in-situ multimodal approaches combining pen and speech [21, 39] can reduce such constant context-switching, speech input remains limited by situational constraints, such as quiet libraries or noisy public spaces, and lacks the fine-grained spatial precision of stylus input. Therefore, we deliberately scope this work to pen-only interaction, seeking to understand how LLM-based inquiry can be fully realized through direct, ink-based modalities without disrupting existing learning workflows.

To empirically understand the friction caused by this modality mismatch, we conducted a formative study ($N = 6$) observing learners navigating standard, out-of-situ LLM workflows. Our findings reveal that learners struggle to map their spatial and visual intent into the discrete text required by the system. We characterize this friction as two distinct interaction challenges: (1) a Referential Barrier, the difficulty of precisely grounding inquiries to specific visual regions; and (2) an Expressive Barrier, the high physical and cognitive overhead of articulating non-textual concepts via standard keyboard input.

These barriers manifest as bottlenecks in the user's workflow. The Referential Barrier arises because learners must provide detailed textual descriptions for visual elements. For example, to simply reference an angle θ in a diagram (Figure 1, Top Left), a learner must construct a lengthy, unambiguous prompt (e.g., "Let theta denote the angle between segments d_1 and d_2 at point (x, y)"), making the mere act of referencing inefficient. Similarly, the Expressive Barrier stems from the low bandwidth of text-only input when dealing with symbolic or structural information. When questioning the shape of a mathematical graph (Figure 1, Bottom Left), learners are forced to translate 2D visual properties into cumbersome 1D text strings (e.g., "If we plot x on the horizontal axis and θ on the vertical axis, does the graph have two peaks or just one?").

To address these barriers, we designed and implemented **Penquiry Version 1** (V1), an in-situ system enabling learners to query directly on study materials via digital ink. In our first user study ($N = 16$), a comparison against a baseline ex-situ keyboard environment demonstrated that V1 mitigated initial barriers by allowing users to directly point to visual elements and freely sketch formulas. However, this study revealed that simply shifting the interface to the pen introduced secondary interaction challenges stemming from raw, unmediated ink input. Regarding the Referential Barrier, high degrees of freedom inherent to free-form sketches created interpretive ambiguity; without explicit visual grounding, learners could not verify whether the system accurately bounded their intended referential scope. Regarding the Expressive Barrier, while sketching bypassed the need for textual workarounds, the high physical effort of handwriting lengthy, semantically complex prompts severely hampered learners' inquiry bandwidth.

To resolve the friction of unmediated ink, we developed **Penquiry Version 2** (V2), which introduces an LLM-mediated layer consisting of two core interaction techniques. First, to eliminate the interpretive ambiguity associated with the Referential Barrier, we implemented Content Snapping, which automatically aligns free-form sketches with document elements, building on prior work on imprecise pen selection [29], to provide immediate visual confirmation of the system's spatial understanding. Second, to overcome the bandwidth constraints of the Expressive Barrier, we integrated Question Autocompletion, an LLM-driven feature that synthesizes complete, context-aware candidate questions from sparse ink keywords, reducing the physical overhead of handwriting lengthy queries. In our second user study comparing V1 and

V2 ($N = 16$), we found that Content Snapping successfully disambiguated the learner's referential intent. Furthermore, Question Autocompletion did more than just reduce physical effort; it actively augmented learners' inquiry behavior, scaffolding their thought processes and prompting them to explore complex concepts beyond their initial intent.

Our core contributions are:

(1) We formalize the Referential Barrier (spatial grounding) and the Expressive Barrier (input bandwidth) that characterize the modality mismatch between fluid pen-based workflows and discrete LLM interfaces in pen-based digital learning contexts.
(2) We present the iterative design and implementation of Penquiry. While Version 1 establishes the foundational in-situ questioning framework, Version 2 introduces an intelligent mediation layer featuring Content Snapping and Question Autocompletion (to synthesize sparse ink into rich prompts).
(3) Through two user studies ($N = 16$ per study), we demonstrate that Penquiry not only significantly reduces the physical and cognitive overhead of in-situ inquiry, but also actively scaffolds and augments learners' complex questioning behaviors, providing empirical insights for pen-centric AI interaction.

2 Related Work

2.1 LLMs in Education

The advent of LLMs has fundamentally invigorated research on conversational learning agents. Recent LLM-based systems facilitate more flexible, natural tutoring dialogues [11, 30] and have been extensively studied in domains like mathematics to explore both their pedagogical potential and inherent risks [28]. Beyond merely providing direct answers, LLM-powered interfaces are increasingly designed to scaffold strategic thinking and complex problem-solving processes [15, 17, 37]. Furthermore, survey-based research indicates that while generative AI offers transformative opportunities in the classroom, it also necessitates significant pedagogical and institutional adaptation [7, 60].

These advancements represent a paradigm shift toward more adaptive, learner-centered experiences. However, while most existing systems still rely on disjointed, chat-style interfaces, the emergence of multimodal interaction highlights opportunities to embed questioning and feedback more seamlessly into the learning environment [2, 19, 26, 27]. Among these, pen-based interaction stands out as a particularly promising modality, enabling learners to engage in flexible, expressive, and contextually grounded inquiry without the cognitive disruption inherent in traditional text-based inputs.

2.2 Pen-based Learning

The cognitive benefits of handwriting and pen-based interaction have been consistently validated in educational and cognitive science research. Compared to typing, handwriting more effectively reinforces memory and conceptual understanding by engaging specific motor and spatial processing mechanisms [22, 31, 34]. In digital contexts, pen-based systems have similarly been shown to promote metacognitive strategies and improve the overall quality of learning [1, 12, 13, 52, 55].

Despite these advantages, most contemporary digital note-taking tools treat pen annotations as static marks rather than interactive, functional elements [53]. To address this, prior research has introduced techniques that dynamically adjust document layouts to accommodate ink [44, 59] or treat handwriting as an active input that interacts with underlying content structures [42, 43]. Building on these advancements, our work aims to bridge the disconnect between the act of inquiry and the process of information retrieval [56]. By transforming handwritten annotations into dynamic query inputs, our approach more effectively supports generative learning processes, allowing learners to remain immersed in their primary study workflow.

2.3 Interaction Paradigms beyond Chat and Keyboard Input

While conversational interfaces have lowered the barrier to LLM adoption, they inherently decouple the user's active workspace from the interaction interface. To bridge this gap, recent Human-Computer Interaction (HCI) research has applied Direct Manipulation principles [48] to AI interactions. By acting directly on document content, these systems use text selections, cursor positions, or predefined elements as implicit prompts, reducing the cognitive effort required to articulate context [5, 33, 50, 51]. However, such approaches are less suited to spatially diffuse regions that lack explicit structural boundaries but remain semantically meaningful to users.

To move beyond keyboard input and structure-dependent selection, researchers have explored multimodal interactions, particularly through pen and sketching tools. These approaches demonstrate that drawing or scribbling can convey spatial intent through free-form marks. Yet, existing LLM-integrated sketching systems predominantly treat sketches as a medium for instruction or generation [21, 58]—functioning as spatial commands for structural edits or visual output rather than facilitators of open-ended exploration.

Compared with these editing- and generation-oriented systems, fewer studies have examined how partial pen input can support inquiry formulation during active reading. Our work addresses this gap by introducing an interaction technique dedicated to the act of questioning. By transforming free-form ink into contextually meaningful targets, our approach enables learners to designate and refine complex inquiries without the structural rigidity of traditional digital selection.

3 Formative Study: Understanding Barriers in Standard Workflow

We conducted a formative study to investigate learners' standard workflow for asking questions to LLMs during pen-based learning, identify the fundamental barriers inherent in these workflows, and explore the potential of an in-situ pen-based Q&A system.

3.1 Study Design

Participants. We recruited university students who had used LLMs during pen-based learning within the previous six months

(N = 6; five women and one man; mean age = 24.2 years, range = 23–25; four undergraduate and two graduate students). All participants had used LLMs for at least one year during learning activities such as concept clarification, summarization, question generation, and mathematical or coding assistance. They typically took notes with a stylus while attending classes or studying on a tablet and used an LLM on a laptop when asking questions.

Materials. To ensure that participants would naturally generate a diverse range of pen-based questions, we selected study materials based on two key criteria: (1) coverage of a wide range of knowledge domains and (2) inclusion of varied content layouts (e.g., figures, tables, graphs, code snippets, and formulas) designed to elicit different pen-based questioning strategies. Based on these criteria, we searched across a broad collection of OpenStax[1] textbooks. Following this process, we selected four subjects—history, data science, algebra, and biology. During the study, participants read sections from the four subjects sequentially, spending approximately 10 minutes on each. To help them quickly immerse themselves in the content, we also provided a brief supporting document for each subject outlining the learning context (e.g., chapter introduction, summary of preceding content), learning objectives, and a few example questions.

Procedure. We designed the study procedure to observe how learners naturally formulate questions through pen input. Each study session lasted approximately 60 minutes and was divided into three parts:

(1) **Pre-Study Interview (10 min):** Participants discussed their typical pen usage patterns during pen-based learning, the types of systems they use when asking questions to LLMs, and any inconveniences they had experienced.
(2) **Task Performance (40 min):** Participants engaged in a learning task within a pen-based environment, where they read the provided materials and asked questions.
(3) **Post-Study Interview (10 min):** Participants reflected on their overall experience and discussed how it differed from their typical questioning workflow.

Protocol. During the task, we employed a Wizard-of-Oz [24] protocol for the pen-based Q&A task. Participants used an iPad and an Apple Pencil to read and annotate PDF documents, with the screen displaying both the document and a web-based chat interface. To enable real-time observation and response simulation, their screen was shared with the researcher via Zoom. As illustrated in Appendix A.1, participants asked questions by sketching (e.g., underlines, circles, arrows, or keywords) with a green pen or highlighter and appending a question mark to signal submission. The Wizard captured the sketched region from the live stream, forwarded it to GPT-4o [38], and returned the model's response via the chat interface. Participants were able to retrieve and view the answer by refreshing the page. A predefined system prompt was used to ensure that the LLM accurately interprets the intentions of the participants from the shared images and pen annotations.

Data Collection and Analysis. We collected screen and audio recordings, interview transcripts, observational notes, and participants' question sketches.

[1]https://openstax.org/

3.2 Findings

Pre-study interviews showed that participants commonly annotated study materials on a tablet but switched to a separate keyboard-equipped device for LLM queries, often transferring context through text descriptions, copy-and-paste, or screenshots. This formative finding later informed the baseline condition in Study 1.

Through thematic analysis of the interview transcripts and observational data, we identified two barriers that hinder effective question-asking when learners rely on standard workflow where they use a keyboard on a separate device during pen-based learning: Referential Barrier and Expressive Barrier.

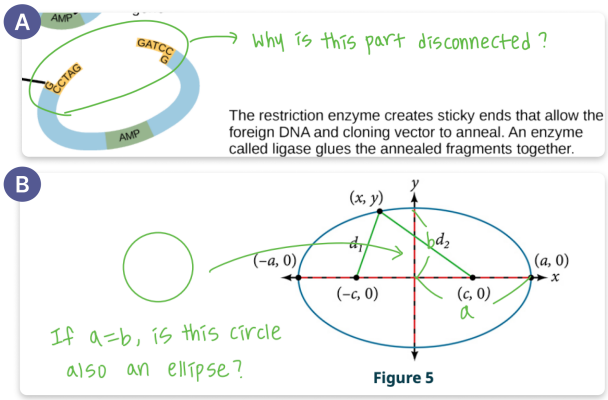


Figure 2: Pen-based questioning strategies observed in the formative study. (A) Using circles to reference specific areas. (B) Drawing sketches to express questions.

Finding 1: Referential Barrier—Difficulty in Clearly Referencing Specific Regions. In standard workflow, participants found it challenging to precisely indicate which part of the study material their question referred to. As a result, they often resorted to inconvenient workarounds, such as repeatedly capturing screenshots or copying and pasting portions of the document. For instance, P2 noted that he typically has to "*capture specific parts and insert the image or copy and paste*" whenever he needs to reference particular content within the study material. In contrast, with pen-based interaction, participants could naturally sketch to reference specific elements. As shown in Figure 2(A), we observed that participants used circles or underlines to reference specific areas. They also used familiar annotation practices such as arrows, highlights, and handwritten keywords to link questions to their targets. These observed practices informed Penquiry's interactions rather than defining a fixed one-to-one gesture vocabulary.

Finding 2: Expressive Barrier—Limitations of Textual Expression for Visual, Symbolic, and Structural Information. Participants also noted that text alone was insufficient to fully express visual, symbolic, and structural aspects of their questions. For instance, P5 described the difficulty of inputting mathematical formulas, stating that "*it is difficult to type formulas with a keyboard when studying mathematics.*" As a result, he often had to "*write the formula with a pen on an iPad, capture it, and send the image to GPT*" whenever he needed to ask a formula-related question. Conversely, participants using pen input not only naturally wrote formula but also expressed

their questions by drawing sketches, as shown in Figure 2(B). These observations indicate that the barrier involved translating spatial and symbolic intent into linear textual descriptions, rather than merely composing the question itself.

Design Requirements. These findings led to three design goals: (1) support flexible pen marks for direct reference and visual or symbolic expression, (2) capture the surrounding semantic scope and document context of concise sketches, and (3) preserve spatial and conversational continuity between questions and answers. In particular, participants' preference for locating responses in the closest available page margin informed the sticky-note-style Answer Shape, while their requests for consecutive questioning motivated the in-place Re-Ask interaction.

4 Penquiry V1: Addressing Barriers of Standard Workflow

To address Referential Barrier and Expressive Barrier identified in our formative study, we designed and implemented Penquiry V1, an in-situ pen-based Q&A system.

4.1 User Interface and Interactions

4.1.1 Sketch Interface and Interactions. The Penquiry V1 interface is designed to support pen-based learning activities directly on top of PDF study materials. It consists of two layers: a Document Layer that renders the original PDF content and a transparent Canvas Layer above it that captures pen input. Using basic sketching tools, learners can freely take notes, highlight passages, and add annotations (Figure 3(B)). They can also navigate the document by performing gestures such as zooming, or by using a thumbnail view (Figure 3(A)). The basic sketching tools are implemented using the open-source tldraw library [54].

Beyond these basic interactions, the core feature of Penquiry V1 is its ability to transform pen-based sketches into queries directed to an LLM. This process is enabled through a dedicated Question Mode, where all sketches are interpreted as question intents (Figure 3(C)).

4.1.2 Question Creation. In Question Mode, learners can naturally express their questions by sketching directly on the document. To overcome Referential Barrier, learners can precisely indicate their target references using visual markers such as underlines, circles, highlights, and arrows. Furthermore, to address Expressive Barrier, learners can freely incorporate visual and symbolic expressions—such as mathematical equations, diagrams, or symbols—that are difficult to articulate through text. Importantly, sketches created for questioning are explicitly distinguished from regular annotations through this mode; these sketches are temporary and automatically disappear once the Question Mode is deactivated.

Once one or more question sketches are created, the "AI Question" button becomes active (Figure 3(1, 2)). When tapped, the system generates a query to GPT-4o by combining the sketches with the relevant document content. The LLM's response is displayed as an Answer Shape—a sticky note-like object containing the interpreted question text and the answer. To fulfill the design requirement derived from our formative study that the act of inquiry and learning flow must coincide in the same spatial context, the Answer Shape is positioned by default in the document margin closest to the original sketch. This placement preserves spatial association while avoiding interference with the study material and other elements. Furthermore, learners can manually reposition the Answer Shapes using the selection tool, while a dotted arrow persistently connects the shape to the location where the sketch was created to maintain a clear visual link.

4.1.3 Re-Ask. Learners often develop additional questions based on the answers they receive. To support iterative inquiry, Penquiry V1 enables follow-up questioning. When learners wish to ask further questions related to a previous answer, they can create new sketches directly on top of the Answer Shape (Figure 3(3)). By then tapping the "AI Question" button, the system recognizes this interaction as a follow-up query and automatically incorporates the context of the prior exchange. The LLM's response to the follow-up query is appended as a new page within the existing Answer Shape (Figure 3(4)).

4.2 LLM Input Generation

4.2.1 Input Pipeline Implementation. To ensure the LLM can accurately interpret sketch-based questions, Penquiry V1 utilizes a multimodal input pipeline. In the formative study, simple visual differentiation of question intents proved problematic due to conflicts with existing content, sketches obscuring underlying text, and ambiguous semantics from insufficient local context.

To overcome these issues, Penquiry V1 captures a multimodal input set: a Cropped Annotated Image (containing sketches), a Cropped Original Image (without sketches), and the Context Text explicitly parsed from the PDF document. By providing both the image with the sketch and the image without the sketch, the system effectively separates the user's strokes from the surrounding local context. The two cropped inputs resolve color-conflict and content-obscuration issues, ensuring the user's intent is visible while the original content remains accessible. The Context Text ensures a more precise semantic baseline, addressing ambiguities caused by limited local views. Dedicated system prompts guide the LLM to synthesize these inputs and interpret minimal gestures as meaningful queries. For follow-up actions (Re-Ask), the system provides the relevant cropped region alongside the prior conversation history. The prompts used in the Penquiry V1 pipeline are provided in Appendix A.5.

4.2.2 Capturing Semantic Scope. A key challenge for the LLM is inferring the semantic scope of minimal sketches (e.g., a simple line or circle), which often refer to broader content like an entire paragraph or equation. To accurately capture this intended scope, the crops generated for the LLM input encompass the direct annotation as well as the local context expanded by a fixed padding area around the sketch. For general questions, the cropping region is defined relative to the learner's annotation: vertically, it includes the annotation area plus fixed padding; horizontally, it spans the full width of the document. If the annotation extends beyond the document boundaries, the crop automatically expands to encompass the strokes. Furthermore, if this cropping region includes a part of a table or figure, it automatically expands to encompass the entire boundaries of that element. For follow-up actions, the crop defaults to the dimensions of the target Answer Shape and expands if the new annotation exceeds its boundaries.

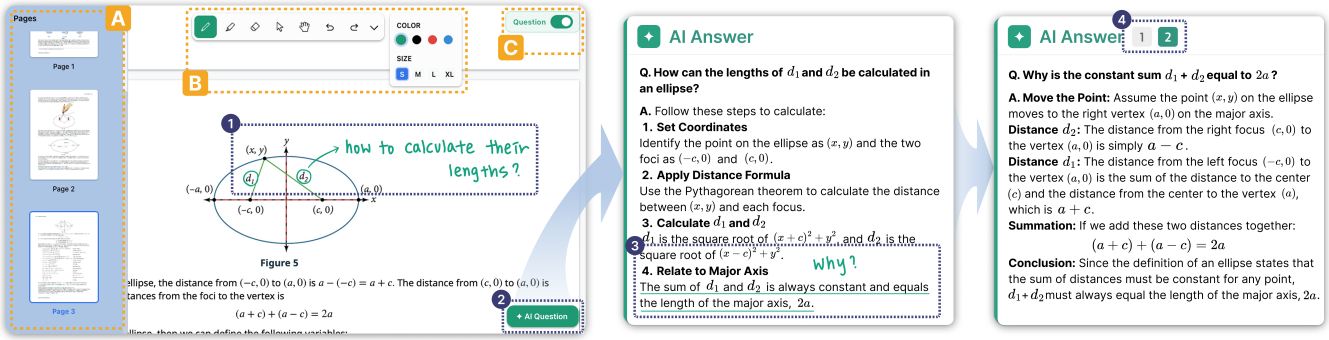


Figure 3: Penquiry V1 user interface and workflow. The interface features (A) a Thumbnail View, (B) Sketching Tools, and (C) Question Mode. The sequence illustrates the integrated inquiry workflow: (1) sketching an in-situ question, (2) receiving a context-aware in-place answer, (3) sketching a follow-up question on the previous response, and (4) navigating iterative answers via tabs.

5 Study 1: Standard Workflow vs. Penquiry V1

We conducted User Study 1 to compare Penquiry V1 against standard workflows, guided by three research questions: (RQ1) How effectively does V1 resolve the barriers inherent in standard workflows? (RQ2) What types of questions do learners ask in an in-situ pen environment? and (RQ3) What new, pen-specific barriers emerge when using pen input?

5.1 Study Design

Participants. This study was conducted with 16 participants (9 male, 7 female; mean age = 24.8 years). All participants were undergraduate or graduate students with prior experience in both tablet-based pen interactions and using LLMs for academic purposes. To minimize the influence of domain-specific knowledge, we excluded applicants majoring in related fields. Each participant received USD 14 for completing the 75 minute study. All study procedures were approved by our Institutional Review Board (IRB).

Materials. We selected two chapters from a geotechnical engineering textbook in the LibreTexts[2] library: "1.5 Standard Penetration Test" and "1.7 Cone Penetration Test." These chapters were chosen to ensure comparable difficulty, roughly equivalent to a first-year undergraduate elective while minimizing the likelihood of prior familiarity. Both chapters were of similar length and designed to be fully read and understood within a 25 minute session. Each chapter featured a balanced mix of text, figures, tables, graphs, and formulas, ensuring that at least one instance of each element was included to maintain consistency across conditions.

Conditions. The study was designed to compare two distinct question–answering environments. To ensure ecological validity, we designed the baseline based on our formative study, instantiating the standard workflow where learners study with a pen on a tablet but use a keyboard-equipped device for LLM queries. In the baseline condition, participants read and annotated the document on a tablet, which was implemented by retaining only the basic learning tools from the Penquiry V1 system. When a question arose, they switched to a separate web browser and typed their query to an LLM (GPT-4o) using a keyboard. To enable a fair comparison of question-asking capabilities, the baseline also provided access to the same source materials on the LLM. Specifically, participants could copy–paste text from an adjacent browser tab into the LLM input, and were allowed to attach the screenshots as image references. In the Penquiry V1 condition, participants used the system for all tasks—including reading, note-taking, and asking questions to the LLM—exclusively through pen-based interactions.

Apparatus. The user study was conducted using an Apple iPad Air running iPadOS 18, paired with an Apple Pencil. For the Baseline condition, the keyboard-based environment was provided using a MacBook Air.

Design. We employed a within-subjects design [10] in which each participant experienced both experimental conditions. To mitigate potential order effects, conditions were counterbalanced [9].

Procedure. The experiment lasted approximately 75 minutes and followed this sequence:

(1) **Briefing and Consent (10 min):** We explained the purpose and procedure and obtained informed consent.
(2) **First Learning Session (25 min):** Participants studied the first material under their assigned initial condition (Baseline or Penquiry V1), asking questions as needed.
(3) **Break (5 min)**
(4) **Second Learning Session (25 min):** Participants studied the second material under the remaining condition.
(5) **Post-study Activities (10 min):** Participants completed in-depth interviews and questionnaires evaluating their experiences and the usability of both systems.

Each learning session was limited to 25 minutes, following the principles of the Pomodoro Technique [16], to maintain participants' concentration and minimize fatigue. All sessions were screen- and audio-recorded with participants' consent for subsequent analysis.

5.2 Result

A total of 429 questions were generated throughout the user study, with 225 (52.45%) were generated in the baseline condition, and 204 (47.55%) in the Penquiry V1 condition. Regarding follow-up questions, 49 (21.78%) were recorded in the baseline condition, compared to 43 (21.08%) in the Penquiry V1 condition.

5.2.1 System Usability. To compare the user experience across the two conditions, we analyzed participants' responses using three

[2]https://eng.libretexts.org/

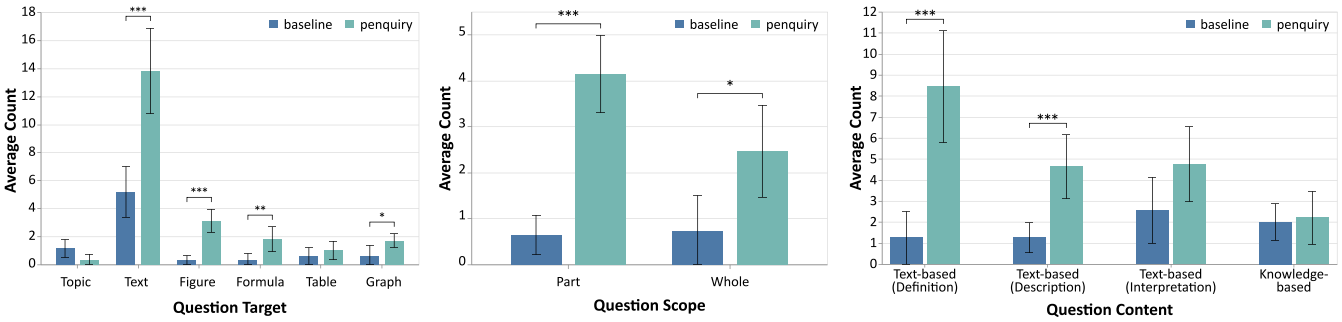


Figure 4: Comprehensive preference analysis across three dimensions. (A) Target-based: Penquiry V1 was favored for all specific targets, with the exception of general `Topic` queries. (B) Scope-based: A strong preference for Penquiry V1 was observed in `Part`-scoped questions, where spatial referencing is crucial. (C) Content-based: Preference shifted with question complexity; Penquiry V1 was most effective for short, direct question types such as `Definition` and `Description`.

measures: the System Usability Scale (SUS) [8], the Technology Acceptance Model (TAM) [14], and the NASA-Task Load Index (NASA-TLX) [18] to evaluate the usability of Penquiry V1. All questionnaire items were rated on a 7-point Likert scale, with statistical significance denoted as $^{*}p < .05$, $^{**}p < .01$, and $^{***}p < .001$.

SUS. The average SUS score for the Penquiry V1 condition was **80.78**, which corresponds to an "Excellent" usability rating according to standard SUS benchmarks.

Table 1: Results from the TAM

Question	Baseline	Penquiry V1	p-value
1. This system is useful for learning.	5.98	6.63	0.029*
2. This system improves learning efficiency.	5.28	6.25	0.011*
3. The functions of this system are intuitive.	5.88	6.38	n.s.
4. I intend to continue using this system.	5.44	6.31	0.008**
5. I feel positive about using this system.	5.69	6.50	0.007**

TAM. The evaluation of TAM showed that Penquiry V1 received significantly higher ratings than the baseline condition in most categories, including perceived usefulness and intention to use (Table 1). This suggests that participants rated Penquiry V1 as a more helpful, efficient, and easier-to-use tool for learning, and also expressed a sustained intention to use the system.

Table 2: Results from the NASA-TLX

Question	Baseline	Penquiry V1	p-value
1. Mental Demand	4.00	3.13	n.s.
2. Physical Demand	4.06	2.31	<0.001***
3. Temporal Demand	3.81	3.69	n.s.
4. Performance	5.00	5.50	n.s.
5. Maintained Learning Flow	4.06	5.63	0.046*

NASA-TLX. Compared to the baseline, the results of the NASA-Task Load Index (NASA-TLX) for Cognitive Load and Learning Flow indicated that the Penquiry V1 condition overall reduced the cognitive burden of performing tasks, particularly by imposing significantly lower physical demand. Through this reduction in cognitive burden, Penquiry V1 showed a statistically significant advantage in Maintained Learning Flow as well (Table 2).

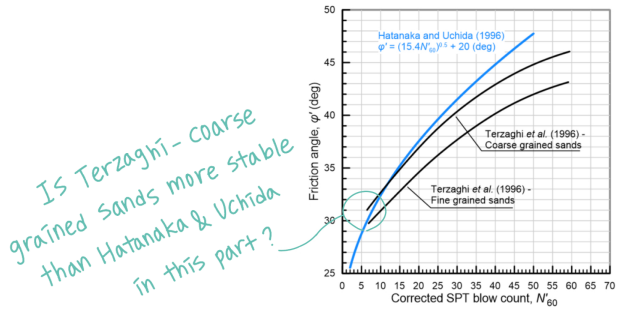


Figure 5: Directly referencing an area difficult to describe textually.

5.2.2 *Addressing Referential Barrier.*

Question Target-based Analysis. We classified the targets of participants' questions into `topic`, `text`, `figure`, `formula`, `table`, and `graph` (Appendix A.2(A)). As shown in the preference data (Figure 4(A)), participants generally preferred using Penquiry V1 when asking questions with specific, concrete referents. However, for questions targeting a general `topic` that lack a clear deictic target, participants favored the baseline condition.

Question Scope-based Analysis. We classified the scope of the targeted elements into `Whole` (e.g., an entire figure or table) and `Part` (e.g., a specific axis of a graph or a row in a table) (Appendix A.2(B)). According to the preference data (Figure 4(B)), participants showed an overwhelming preference for using Penquiry V1 when asking these `Part`-scoped questions.

Pen-Mark Type. Based on Marshall's annotation classification [32], we analyzed 204 queries in the Penquiry V1 dataset. Participants predominantly used four pen-mark types: `underlining` (45%), `circles` (29%), `boxes` (9%), and `braces` (2%) (Appendix A.3). Notably, 85% of queries contained at least one visible mark. This overwhelming reliance on visual deictic cues demonstrates that pen input effectively supports precise referencing behaviors.

Ease of Referencing and Qualitative Insights. Regarding the "Ease of referencing specific elements," Penquiry V1 showed a statistically significant improvement over the baseline (6.00 vs. 4.38, p = 0.015*), allowing participants to easily designate specific targets. Interview responses further supported this efficiency. For instance,

Figure 5 demonstrates a case where pointing with a pen is essential. When asking about a specific part of a graph, P16 noted: "*If I had tried to write this question with text, it would have become much longer, like 'between 5 and 10 on the x-axis...' With the pen, I could just point and ask directly.*" Collectively, these results demonstrate that Penquiry V1 successfully introduces the principle of deictic reference [6] into the Q&A process. The ability to "just point and ask directly" provides unambiguous, in-situ interaction, effectively resolving the referential barrier.

5.2.3 *Addressing Expressive Barrier.*

Question Content-based Analysis. We conducted a question content-based analysis to categorize the underlying semantic intent of the inquiries. To assess the cognitive depth of participants' questions, we adopted the classification framework proposed by Scardamalia and Bereiter [45]. This framework categorizes questions into *Text-based* questions, which seek to understand explicitly presented information, and *Knowledge-based* questions, which involve higher-level cognitive processes such as extending, applying, or comparing beyond the text. We further sub-divided *Text-based* questions into `Definition` (seeking the meaning of a short term), `Description` (requesting a simple explanation), and `Interpretation` (asking about reasons, implications, or principles) (Appendix A.2(C)). According to the preference data (Figure 4(C)), participants showed a significant preference for Penquiry V1 when asking shorter questions, such as `Text-based (Definition)` and `Text-based (Description)`. In contrast, as questions became more interpretive or required longer, more elaborate descriptions, a trend toward preferring the baseline emerged. Interviews further clarified this shift; several participants (e.g., P1, P6, P11, P13) reported that typing longer questions was faster and more convenient with a keyboard, leading them to favor the baseline in such cases.

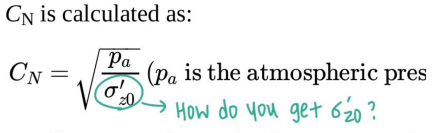


Figure 6: Example of free-form pen-based inquiries. Participants comfortably articulated mathematical symbols and notations using pen input.

Free-form Pen Input. Our analysis examined specific inquiry types where free-form pen input offered a distinct expressive advantage over keyboard modalities. As illustrated in the representative examples in Figure 6, participants leveraged the pen to effortlessly articulate mathematical symbols that would otherwise be cumbersome to express via keyboard. These cases confirm that free-form pen input significantly reduces the barrier to entry for visual, structural, or symbolic inquiries, enabling participants to express these complex intents directly.

Ease of Expressing Intent and Qualitative Insights. Regarding the "Ease of expressing question intent," the Penquiry V1 score was slightly lower than the baseline (4.88 vs. 5.50, p = 0.374), primarily due to difficulties in handwriting long questions or occasional recognition issues. However, this difference was not statistically significant. In fact, interview responses revealed that Penquiry V1 provided distinct advantages for queries that are difficult to articulate in text. As shown in Figure 6, several participants (e.g., P4, P9) expressed a clear preference for the pen modality when handling symbolic and notational content, noting that "*entering mathematical formulas via a keyboard presents many difficulties, making the pen an easier alternative.*" They also appreciated the system's accurate recognition of handwritten formulas. Similarly, P10 remarked, "*With Penquiry V1, I could easily express my intent by directly denoting elements related to the x-axis or y-axis.*"

5.3 Finding

While Penquiry V1 mitigated the inherent barriers of the standard workflow, our analysis revealed that pen-based interaction introduces its own set of challenges. We redefined Referential Barrier and Expressive Barrier within the context of pen-based inquiry as follows.

Finding 1: Referential Barrier—Lack of Visual Feedback in Verifying Referential Intent. Although pen input enables participants to reference materials with high degrees of freedom, this flexibility often introduces uncertainty regarding system recognition. Participants noted a critical lack of visual feedback to confirm whether their intended targets were accurately captured, occasionally resulting in irrelevant system responses. For instance, P14 remarked, "*I underlined a sentence to ask a question, but the system provided an irrelevant answer. The fact that my pen marks could be misinterpreted made it slightly uncomfortable.*" Similar frustrations were reported by P1, P15, and P16 when their referential intent was misaligned with the system's internal processing. These instances underscore that providing explicit visual grounding is essential for users to verify that their intended targets are accurately reflected in the query.

Finding 2: Expressive Barrier—High Interaction Cost in Formulating Complex Queries. Although pen input successfully supports visually and symbolically rich queries, a significant limitation emerged when participants attempted to formulate `Knowledge-based` questions. Those who engaged in complex inquiry reported that handwriting lengthy, structured queries was both physically burdensome and cognitively demanding. Furthermore, participants expressed recognition anxiety, fearing that the system might fail to accurately digitize their text even after the effort of writing it out in full. These findings underscore the necessity for an interaction mechanism that can assist learners in articulating lengthy and complex questions, thereby reducing the interaction cost associated with manual handwriting in pen-based environments.

6 Penquiry V2: Resolving Pen-Specific Barriers

Building on the findings from Study 1, we identified residual challenges in pen-based questioning: a lack of immediate confirmation regarding referential accuracy and the significant physical effort required for handwriting extended inquiries. To address these issues and further mitigate Referential Barrier and Expressive Barrier, we refined the system into Penquiry V2, which retains the core functionality of V1 while introducing targeted enhancements to the interface interactions and the LLM input generation pipeline.

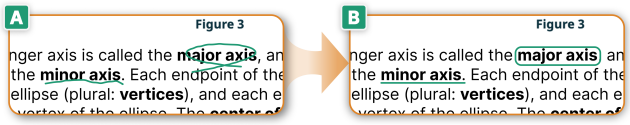


Figure 7: Content Snapping. (A) A learner draws a rough circle around "major axis" and underlines "minor axis". (B) The system automatically snaps these sketches into precise geometric shapes (rounded rectangle and line) to enclose the text, providing immediate visual confirmation of the referential target.

6.1 User Interface Enhancements

6.1.1 Content Snapping. To address the uncertainty learners experienced regarding their referenced regions, V2 introduces Content Snapping. When a learner draws a circle or underline over the document (Figure 7(A)), the system maps the sketch to nearby character-level PDF text elements and redraws it to enclose or underline those elements (Figure 7(B)). This immediate visual feedback confirms that the system has accurately captured their referential target, eliminating ambiguity before the question is submitted to the LLM. This approach to snapping builds on imprecise pen-selection work such as Sloppy Selection [29].

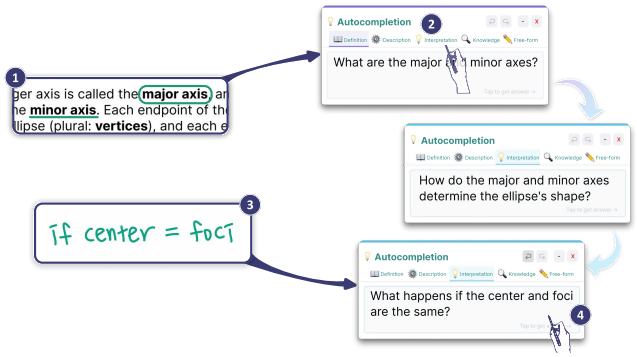


Figure 8: Question Autocompletion. (1) Referencing "major axis" and "minor axis" generates an autocompleted question under the Definition tab (highest confidence). (2) The learner switches to the Interpretation tab to align with their specific intent. (3) Adding a supplementary sketch dynamically refines the suggested question to incorporate the new context. (4) Tapping the updated question submits the request.

6.1.2 Question Autocompletion. To alleviate the physical and cognitive burden of handwriting entire questions, V2 features Question Autocompletion. When a learner provides a minimal sketch—such as a brief underline or a circled term, occasionally accompanied by a keyword—the system analyzes the action within its local context to dynamically generate four autocomplete question candidates (Figure 8(1)). These candidates are categorized based on the question-content analysis detailed in Section 5.2.3.

The LLM evaluates the user's sketch to calculate a confidence score for each category, displaying the autocompleted question with the highest confidence. Learners can navigate through alternative categories to find a question that aligns with their specific intent (Figure 8(2)). If the initially autocompleted questions do not fully capture their meaning, learners can continue adding sketches to refine them (Figure 8(3)), or revert their actions using the undo button. Once satisfied, the learner can tap the autocompleted question to submit it (Figure 8(4)).

Furthermore, to preserve user autonomy, V2 maintains the *Free Question* feature from V1. If no autocompleted questions align with their intent, learners can bypass the autocompletion entirely and formulate inquiries through free-form sketching.

6.2 LLM Input Generation

6.2.1 Question Autocompletion and Refinement Pipeline. To support Question Autocompletion, V2 employs a specialized multi-stage pipeline using GPT-4o for all stages. For the initial Question Autocompletion, the system transmits a multimodal input set to the LLM: the Context Text, a Cropped Original Image, and a Cropped Annotated Image containing the user's minimal sketch. Using a dedicated prompt, the LLM processes these spatial and semantic inputs to generate four candidate questions, each accompanied by a confidence score.

If a learner refines a suggestion by adding supplementary sketches or keywords, a refinement request is triggered. In this stage, the LLM receives the Selected Category, the Seed Question, the Context Text, and the updated Cropped Original Image and Cropped Annotated Image. The LLM synthesizes these continuous strokes with the existing semantic intent to produce a refined, contextually accurate question. The prompts used across the Penquiry V2 pipeline are provided in Appendix A.6.

6.2.2 Question Selection and Answering Pipeline. Once the inquiry is finalized the system proceeds to the final answering stage. The input set for this phase includes the Selected Question, the Context Text, and the corresponding Cropped Original/Annotated Images. The LLM generates a precise, context-aware response, maintaining strict grounding within the surrounding textual and visual boundaries of the document.

7 Study 2: Penquiry V1 vs. Penquiry V2

We conducted Study 2 to evaluate the iterative improvements of Penquiry V2 relative to V1, focusing on two research questions: (RQ1) How effectively does V2 resolve the pen-specific barriers identified in Study 1? and (RQ2) How do learners practically utilize the newly added features during their study?

7.1 Study Design

User Study 2 followed the same design and procedure as User Study 1 (Section 5.1). We recruited a new cohort of 16 participants (7 male, 9 female; mean age = 22.69 years) and employed a within-subjects design to compare two system iterations, Penquiry V1 and V2. For the study materials, we selected two new chapters ("4. Theory" and "5. Carbon") from a climate science textbook[3].

7.2 Result

A total of 458 questions were generated across both conditions: 219 (47.8%) in Penquiry V1 and 239 (52.2%) in Penquiry V2. Follow-up questions accounted for 51 (23.3%) in V1 and 41 (17.2%) in V2.

[3]https://open.oregonstate.education/

7.2.1 System Usability. We evaluated the usability of PENQUIRY V2 using the same three instruments: SUS, TAM, and NASA-TLX.

SUS. PENQUIRY V2 achieved an average SUS score of **84.9**, reflecting an "Excellent" usability rating and surpassing the score from the initial study.

TAM. Both V1 and V2 received consistently high scores across all metrics. While PENQUIRY V2 scored slightly higher on items such as perceived usefulness (5.69 vs. 6.19), learning efficiency (5.38 vs. 5.81), intuitiveness (6.25 vs. 6.38), intention to use (5.06 vs. 5.50), and overall positive affect (5.50 vs. 6.06), these differences were not statistically significant. These results indicate that participants found both pen-based systems to be highly intuitive, engaging, and valuable for their learning process.

NASA-TLX. Both systems minimized user burden while maintaining high learning quality. Participants reported low levels of mental (3.19 vs. 3.13), physical (3.63 vs. 2.75), and temporal demand (2.69 vs. 1.94) for V1 and V2, respectively. Conversely, positive metrics such as perceived performance (5.63 vs. 5.75) and learning flow (5.38 vs. 5.25) remained consistently high. Although PENQUIRY V2 showed a noticeable trend toward lower physical and temporal demands, the differences between the two versions did not reach statistical significance.

7.2.2 System Performance and Robustness.

System Error Rate. The error rates were 4.6% (10/219) in PENQUIRY V1 and 3.8% (9/239) in PENQUIRY V2. Among the 19 total errors, eight involved handwriting misrecognition and nine involved intent misinterpretation. The remaining two were caused by referential misalignment in V1 and context over-restriction in V2, with the latter resulting from constraining answers strictly to the provided document without external search.

System Latency. The average end-to-end response time was 9.88 seconds for the baseline, 8.17 seconds for PENQUIRY V1, and 9.04 seconds for PENQUIRY V2. For V2, the latency comprised 4.34 seconds for Question Autocompletion and 4.70 seconds for answer generation, remaining comparable to the baseline despite the additional processing stage.

7.2.3 Addressing Referential Barrier.

Content Snapping. Participants reported that Content Snapping provided immediate visual confirmation of their intended referents, bridging the gap between action and system perception. While PENQUIRY V1 users often felt uncertainty—noting it was "*not clear if the system accurately captured my range*" (P5)—the instant alignment in V2 explicitly visualizes the recognition state. From an HCI perspective, this immediate feedback mitigates the Gulf of Execution [35, 36, 41] by confirming that pointing actions are successfully translated into captured intent. As P4 noted, "*I could clearly see that what I wanted to ask was being well recognized*," which allowed participants to formulate questions "*with trust*" (P6) and resolved the anxiety (P16) previously caused by ambiguous system feedback.

Qualitative Insights on Ease of Referencing. Both versions received high scores for "Ease of referencing specific elements" (5.19 vs. 5.75, $p = 0.112$), confirming that pen-based systems are inherently effective for target designation. Beyond simple bounding, the free-form nature of pen input enables users to specify abstract spatial concepts intuitively. For example, P5 inquired about the area under a curve by simply drawing hatching marks: "*I wanted to ask about the area under the graph, so I drew hatching to indicate the space, and V2 recommended exactly what I wanted.*" This illustrates how PENQUIRY V2 leverages flexible input to convey complex spatial references that are otherwise cumbersome to describe textually.

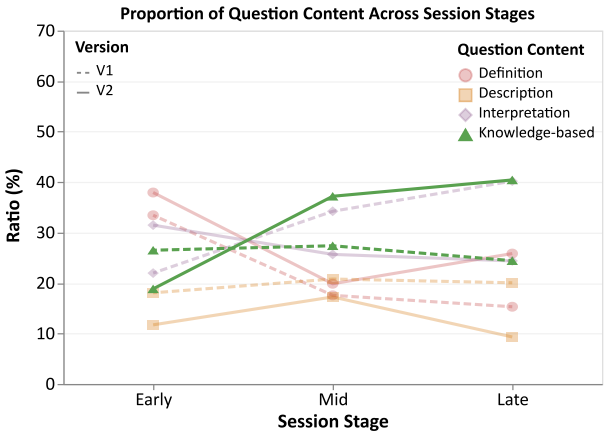


Figure 9: Temporal trends in average proportions of question contents across session phases by system version.

7.2.4 Addressing Expressive Barrier.

Question Autocompletion. The effectiveness of the Question Autocompletion is evidenced by its high adoption rate, with **74.9%** of all submitted queries utilizing autocompletion. While V1 users often found handwriting entire questions physically tiring and time-consuming, V2 resolved this by enabling learners to generate complete inquiries from minimal sketches or keywords. P3 noted that the ability to select autocompleted questions meant they "*didn't have to write the sentence to the end, which reduced the burden.*" Furthermore, participants reported that the suggestions reflected their intentions across predefined categories (P5, P12). At the same time, retaining the Free Question option allowed them to bypass suggestions when they did not match their intent.

Beyond reducing physical effort, we investigated whether Question Autocompletion enabled learners to formulate more complex, high-level inquiries. Since knowledge-based questions typically emerge in later stages as learners gain deeper comprehension, we analyzed the questioning sessions across three chronological phases: early, mid, and late. Results revealed a significant progressive increase in the proportion of knowledge-based questions: 18.85% (early), 37.19% (mid), and 40.43% (late) (Figure 9; Friedman test, $p = 0.015$; Wilcoxon signed-rank test for early vs. late, $p = 0.019$).

These findings suggests that the Question Autocompletion may have supported an effective cognitive scaffold [57]; rather than merely completing sentences, it progressively guided learners toward higher levels of exploration as the session advanced. Interview data corroborated this effect; participants noted that the system prompted inquiries they had not initially considered. We also observed cases in which participants first asked their intended question and later returned to a suggestion-inspired follow-up. Ultimately, the feature scaffolded deeper engagement (P1, P3) by lowering the threshold for complex inquiry without imposing additional cognitive load.

Qualitative Insights on Ease of Expressing Intent. Regarding the "Ease of expressing question intent," Penquiry V2 scored slightly higher than V1 (5.88 vs. 5.50), indicating that both pen-based systems effectively support inquiry conveyance. However, qualitative feedback revealed that V2's Question Autocompletion specifically mitigated the cognitive burden of question formulation—a well-documented challenge for learners [25]. Participants emphasized that generating full question from minimal keywords significantly lowered this barrier. As P5 noted, the system translated "*abstract thoughts into well-formed questions, reducing the cognitive burden.*" P2 echoed this, explaining that even when they held a vague curiosity without a clear explanation, simply writing a keyword like "VS" allowed the system to generate an appropriate structured question. Ultimately, this keyword-driven approach alleviated both "*the burden of making question and the burden of handwriting*" (P8).

8 Discussion

8.1 Unique Affordances of Pen Input Modality

In this section, we examine the unique affordances of pen-based interaction, focusing on its advantages in referential precision and direct contextual articulation. Unlike direct manipulation approaches that operate on predefined document structures, such as text spans or SVG elements [33], free-form ink allows learners to indicate structurally undefined yet semantically meaningful regions through deictic marks and free trajectories (Figure 5 and Figure 6). This in-situ referential expression captures intent seamlessly, without forcing users to translate visual attention into cumbersome textual descriptions. Furthermore, pen input also supports combining references, symbols, equations, and relational marks within the same canvas, allowing learners to express their inquiry intent in forms closely aligned with the source material. These affordances preserve critical spatial relationships among user marks, document content, and generated responses.

8.2 Material-Dependent Adaptive Scaffolding

While the Question Autocompletion feature mitigated the Expressive Barrier, qualitative feedback revealed that its utility scaled dynamically with the learner's domain expertise. Participants relied heavily on the LLM's generative scaffolding when encountering novel concepts; P12 found the feature ideal when "*studying a topic for the first time*" while P9 explained that formulating questions from scratch is especially challenging with unfamiliar material. Conversely, the proactive nature of the autocompletion could sometimes hinder learners, with P4 reporting that the automated suggestions occasionally "*interfered*" when they already had a very specific question in mind for familiar content. Furthermore, generated candidate questions may anchor learners to system-framed perspectives or narrow the range of alternatives they consider, highlighting a potential risk of over-steering. These insights emphasize that scaffolding should not be a static intervention. Consequently, future pen-based Q&A interfaces must evolve beyond static features to implement adaptive fading [3, 4, 23]. By intelligently shifting the locus of control—providing aggressive query synthesis during initial conceptual shifts, then fading to subtle, user-driven inquiry as proficiency increases—systems can prevent AI over-reliance while preserving learner agency.

8.3 Limitation and Future Work

Despite the promising results, this study has several limitations. First, the limited session duration may have prevented participants from fully adapting to the pen-based questioning environment. The controlled experimental design—requiring the study of unfamiliar materials within a strict timeframe—may also have constrained the diversity of query types. Long-term learning outcomes remain to be explored, highlighting the need for extended longitudinal studies to assess the sustained impact of Penquiry on learning habits.

Second, while the baseline in Study 1 was chosen for ecological validity based on prevailing workflows at the time, the rapid evolution of LLM-integrated tools continues to shift standard practices. Future research should incorporate contemporary benchmarks to evaluate Penquiry's distinct advantages within increasingly AI-saturated learning environments.

Furthermore, building on our findings regarding cognitive scaffolding, future iterations should explore temporally adaptive Question Autocompletion mechanisms. As learners' comprehension deepens over a session, the system could dynamically adjust the cognitive complexity of its autocompletions based on elapsed time or interaction history. By progressively transitioning from foundational fact-checking in early phases to high-level, knowledge-based prompts in later stages, Penquiry could provide a truly tailored, adaptive scaffolding experience.

9 Conclusion

We identified the Referential Barrier and Expressive Barrier that learners encounter when integrating LLMs into pen-based digital learning. To bridge this interaction gap, we presented Penquiry, an in-situ Q&A system to facilitate inquiry directly within active study workflows. Through an iterative design process, we introduced Content Snapping to resolve referential ambiguity and Question Autocompletion to alleviate the physical and cognitive overhead of handwriting queries. Our evaluations demonstrated that these features not only mitigate the limitations of standard keyboard-centric interfaces but also function as effective cognitive scaffolds, guiding learners toward deeper, knowledge-based exploration. Ultimately, Penquiry establishes a new blueprint for intuitive and effective human–AI interaction, empowering learners to harness the full potential of LLMs without disrupting the cognitive flow of pen-based learning.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grants funded by the Korean government (MSIT) (No. NRF-2023R1A2C2005209; No. RS-2024-00456247; No. RS-2023-00218913; No. RS-2025-00520170), the Institute of Information & Communications Technology Planning & Evaluation (IITP) grants funded by the Korean government (MSIT) [No. RS-2021-II211343, Artificial Intelligence Graduate School Program (Seoul National University); No. RS-2019-II191906, Artificial Intelligence Graduate School Program (POSTECH); No. RS-2026-25546560, the Leading Generative AI Human Resources Development], and the Hankuk University of Foreign Studies Research Fund. The ICT at Seoul National University provided research facilities for this study.

References

[1] Afaf H Al-Nadaf. 2024. The Effect of Smart Pen Technology on Distance Learning Teaching for Medicinal Chemistry Courses: An Exploratory Study. *Research Journal of Pharmacy and Technology* 17, 9 (2024), 4329–4336. doi:10.52711/0974-360X.2024.00669

[2] Fady Alnajjar, Alreem R Alneyadi, Habiba Almetnawy, Khawla S Alneyadi, Fatima Alhemeiri, Amna R Alneyadi, Ahed Orabi, and Muhammed S Palliyalil. 2024. Exploring the Pedagogical Potential of Large Language Models: A Multimodal Study on Student Learning Outcomes. In *2024 6th International Workshop on Artificial Intelligence and Education (WAIE)*. IEEE, 93–96. doi:10.1109/WAIE63876.2024.00024

[3] Kenneth C Arnold, Krysta Chauncey, and Krzysztof Z Gajos. 2020. Predictive text encourages predictable writing. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 128–138. doi:10.1145/3377325.3377523

[4] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–16. doi:10.1145/3411764.3445717

[5] Shraddha Barke, Michael B. James, and Nadia Polikarpova. 2023. Grounded Copilot: How Programmers Interact with Code-Generating Models. *Proc. ACM Program. Lang.* 7, OOPSLA1, Article 78 (April 2023), 27 pages. doi:10.1145/3586030

[6] Richard A Bolt. 1980. "Put-that-there" Voice and gesture at the graphics interface. In *Proceedings of the 7th annual conference on Computer graphics and interactive techniques*. 262–270. doi:10.1145/800250.807503

[7] Matt Bower, Jodie Torrington, Jennifer WM Lai, Peter Petocz, and Mark Alfano. 2024. How should we change teaching and assessment in response to increasingly powerful generative Artificial Intelligence? Outcomes of the ChatGPT teacher survey. *Education and Information Technologies* 29, 12 (2024), 15403–15439. doi:10.1007/s10639-023-12405-0

[8] John Brooke. 1995. SUS: A quick and dirty usability scale. *Usability Eval. Ind.* 189 (11 1995).

[9] Joseph L Brooks. 2012. Counterbalancing for serial order carryover effects in experimental condition orders. *Psychological methods* 17, 4 (2012), 600. doi:10.1037/a0029310

[10] Gary Charness, Uri Gneezy, and Michael A Kuhn. 2012. Experimental methods: Between-subject and within-subject design. *Journal of economic behavior & organization* 81, 1 (2012), 1–8. doi:10.1016/j.jebo.2011.08.009

[11] Yulin Chen, Ning Ding, Hai-Tao Zheng, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2024. Empowering private tutoring by chaining large language models. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. 354–364. doi:10.1145/3627673.3679665

[12] Michelene T. H. Chi and Ruth Wylie. 2014. The ICAP Framework: Linking Cognitive Engagement to Active Learning Outcomes. *Educational Psychologist* 49, 4 (2014), 219–243. doi:10.1080/00461520.2014.965823

[13] Lilian L Chimuma and Iris DeLoach Johnson. 2016. Assessing Students' Use of Metacognition during Mathematical Problem Solving Using Smartpens. *Educational Research: Theory and Practice* 28, 1 (2016), 22–36.

[14] Fred D. Davis. 1989. Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *Management Information Systems Quarterly* 13, 3 (09 1989), 319–340. doi:10.2307/249008

[15] David Dinucu-Jianu, Jakub Macina, Nico Daheim, Ido Hakimi, Iryna Gurevych, and Mrinmaya Sachan. 2025. From Problem-Solving to Teaching Problem-Solving: Aligning LLMs with Pedagogy using Reinforcement Learning. *arXiv preprint arXiv:2505.15607* (2025). doi:10.48550/arXiv.2505.15607

[16] Federico Gobbo and Matteo Vaccari. 2008. The pomodoro technique for sustainable pace in extreme programming teams. In *International conference on agile processes and extreme programming in software engineering*. Springer, 180–184. doi:10.1007/978-3-540-68255-4_18

[17] Eleonora Grassucci, Gualtiero Grassucci, Aurelio Uncini, and Danilo Comminiello. 2025. Beyond answers: How llms can pursue strategic thinking in education. *arXiv preprint arXiv:2504.04815* (2025). doi:10.48550/arXiv.2504.04815

[18] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*. Vol. 52. Elsevier, 139–183. doi:10.1016/S0166-4115(08)62386-9

[19] Ken Hinckley, Koji Yatani, Michel Pahud, Nicole Coddington, Jenny Rodenhouse, Andy Wilson, Hrvoje Benko, and Bill Buxton. 2010. Pen+ touch= new tools. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*. 27–36. doi:10.1145/1866029.1866036

[20] Kenneth Holstein, Bruce M McLaren, and Vincent Aleven. 2019. Designing for complementarity: Teacher and student needs for orchestration support in AI-enhanced classrooms. In *International conference on artificial intelligence in education*. Springer, 157–171. doi:10.1007/978-3-030-23204-7_14

[21] Zeyuan Huang, Cangjun Gao, Yaxian Shan, Haoxiang Hu, Qingkun Li, Xiaoming Deng, Cuixia Ma, Yu-Kun Lai, Yong-Jin Liu, Feng Tian, Guozhong Dai, and Hongan Wang. 2025. SketchGPT: A Sketch-based Multimodal Interface for Application-Agnostic LLM Interaction *(UIST '25)*. Association for Computing Machinery, New York, NY, USA, Article 157, 18 pages. doi:10.1145/3746059.3747598

[22] Aya S Ihara, Kae Nakajima, Akiyuki Kake, Kizuku Ishimaru, Kiyoyuki Osugi, and Yasushi Naruse. 2021. Advantage of handwriting over typing on learning words: Evidence from an N400 event-related potential index. *Frontiers in human neuroscience* 15 (2021), 679191. doi:10.3389/fnhum.2021.679191

[23] Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. 2023. Co-writing with opinionated language models affects users' views. In *Proceedings of the 2023 CHI conference on human factors in computing systems*. 1–15. doi:10.1145/3544548.3581196

[24] John F Kelley. 1984. An iterative design methodology for user-friendly natural language office information applications. *ACM Transactions on Information Systems (TOIS)* 2, 1 (1984), 26–41. doi:10.1145/357417.357420

[25] Alison King. 1994. Guiding knowledge construction in the classroom: Effects of teaching children how to question and how to explain. *American educational research journal* 31, 2 (1994), 338–368.

[26] Gerd Kortemeyer. 2023. Toward AI grading of student problem solutions in introductory physics: A feasibility study. *Physical Review Physics Education Research* 19, 2 (2023), 020163. doi:10.1103/PhysRevPhysEducRes.19.020163

[27] Stefan Küchemann, Karina E. Avila, Yavuz Dinc, Chiara Hortmann, Natalia Revenga, Verena Ruf, Niklas Stausberg, Steffen Steinert, Frank Fischer, Martin Fischer, Enkelejda Kasneci, Gjergji Kasneci, Thomas Kuhr, Gitta Kutyniok, Sarah Malone, Michael Sailer, Albrecht Schmidt, Matthias Stadler, Jochen Weller, and Jochen Kuhn. 2025. On opportunities and challenges of large multimodal foundation models in education. *npj Science of Learning* 10, 1 (2025), 11. doi:10.1038/s41539-025-00301-w

[28] Harsh Kumar, David M Rothschild, Daniel G Goldstein, and Jake M Hofman. 2025. Math education with large language models: Peril or promise?. In *International Conference on Artificial Intelligence in Education*. Springer, 60–75. doi:10.1007/978-3-031-98459-4_5

[29] Edward Lank and Eric Saund. 2005. Sloppy selection: Providing an accurate interpretation of imprecise selection gestures. *Computers & Graphics* 29, 4 (2005), 490–500. doi:10.1016/j.cag.2005.05.003

[30] Zhengyuan Liu, Stella Xin Yin, Geyu Lin, and Nancy F. Chen. 2024. Personality-aware Student Simulation for Conversational Intelligent Tutoring Systems. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 626–642. doi:10.18653/v1/2024.emnlp-main.37

[31] Giuseppe Marano, Georgios D. Kotzalidis, Francesco Maria Lisci, Maria Benedetta Anesini, Sara Rossi, Sara Barbonetti, Andrea Cangini, Alice Ronsisvalle, Laura Artuso, Cecilia Falsini, Romina Caso, Giuseppe Mandracchia, Caterina Brisi, Gianandrea Traversi, Osvaldo Mazza, Roberto Pola, Gabriele Sani, Eugenio Maria Mercuri, Eleonora Gaetani, and Marianna Mazza. 2025. The Neuroscience Behind Writing: Handwriting vs. Typing—Who Wins the Battle? *Life* 15, 3 (2025). doi:10.3390/life15030345

[32] Catherine C. Marshall. 1997. Annotation: from paper books to the digital library. In *Proceedings of the Second ACM International Conference on Digital Libraries* (Philadelphia, Pennsylvania, USA) *(DL '97)*. Association for Computing Machinery, New York, NY, USA, 131–140. doi:10.1145/263690.263806

[33] Damien Masson, Sylvain Malacria, Géry Casiez, and Daniel Vogel. 2024. DirectGPT: A Direct Manipulation Interface to Interact with Large Language Models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–16. doi:10.1145/3613904.3642462

[34] Pam A Mueller and Daniel M Oppenheimer. 2014. The pen is mightier than the keyboard: Advantages of longhand over laptop note taking. *Psychological science* 25, 6 (2014), 1159–1168. doi:10.1177/0956797614524581

[35] Jakob Nielsen. 1994. Enhancing the explanatory power of usability heuristics. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. 152–158. doi:10.1145/191666.191729

[36] Donald A Norman. 1986. Cognitive engineering. In *User centered system design: New perspectives on human-computer interaction*. CRC Press, 31–61.

[37] Hiroaki Ogata, Brendan Flanagan, Kyosuke Takami, Yiling Dai, Ryosuke Nakamoto, and Kensuke Takii. 2024. EXAIT: Educational eXplainable Artificial Intelligent Tools for personalized learning. *Research and Practice in Technology Enhanced Learning* 19 (Jan. 2024), 019. doi:10.58459/rptel.2024.19019

[38] OpenAI. 2024. OpenAI GPT-4o API Documentation. https://platform.openai.com/docs/models/gpt-4o.

[39] Sharon Oviatt, Antonella DeAngeli, and Karen Kuhn. 1997. Integration and synchronization of input modes during multimodal human-computer interaction. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems* (Atlanta, Georgia, USA) *(CHI '97)*. Association for Computing Machinery, New York, NY, USA, 415–422. doi:10.1145/258549.258821

[40] Sharon L Oviatt and Adrienne O Cohen. 2010. Toward high-performance communications interfaces for science problem solving. *Journal of science education and technology* 19, 6 (2010), 515–531. doi:10.1007/s10956-010-9218-7

[41] Seokhyeon Park, Soohyun Lee, Eugene Choi, Hyunwoo Kim, Minkyu Kweon, Yumin Song, and Jinwook Seo. 2026. Bridging Gulfs in UI Generation through

Semantic Guidance. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26)*. Association for Computing Machinery, New York, NY, USA, Article 880, 16 pages. doi:10.1145/3772318.3791966

[42] Yann Riche, Nathalie Henry Riche, Ken Hinckley, Sheri Panabaker, Sarah Fuelling, and Sarah Williams. 2017. As We May Ink? Learning from Everyday Analog Pen Use to Improve Digital Ink Experiences. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 3241–3253. doi:10.1145/3025453.3025716

[43] Hugo Romat, Nathalie Henry Riche, Ken Hinckley, Bongshin Lee, Caroline Appert, Emmanuel Pietriga, and Christopher Collins. 2019. ActiveInk: (Th)Inking with Data. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3290605.3300272

[44] Hugo Romat, Emmanuel Pietriga, Nathalie Henry-Riche, Ken Hinckley, and Caroline Appert. 2019. SpaceInk: Making Space for In-Context Annotations. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology* (New Orleans, LA, USA) *(UIST '19)*. Association for Computing Machinery, New York, NY, USA, 871–882. doi:10.1145/3332165.3347934

[45] Marlene Scardamalia and Carl Bereiter. 1992. Text-based and knowledge based questioning by children. *Cognition and instruction* 9, 3 (1992), 177–199. doi:10.1207/s1532690xci0903_1

[46] Sahil Sharma, Puneet Mittal, Mukesh Kumar, and Vivek Bhardwaj. 2025. The role of large language models in personalized learning: a systematic review of educational impact. *Discover Sustainability* 6, 1 (2025), 243. doi:10.1007/s43621-025-01094-z

[47] Madelynn D Shell, Maranda Strouth, and Alexandria M Reynolds. 2021. Make a Note of It: Comparison in Longhand, Keyboard, and Stylus Note-Taking Techniques. *Learning Assistance Review* 26, 2 (2021), 1–21.

[48] Shneiderman. 1983. Direct Manipulation: A Step Beyond Programming Languages. *Computer* 16, 8 (1983), 57–69. doi:10.1109/MC.1983.1654471

[49] Steffen Steinert, Karina E. Avila, Stefan Ruzika, Jochen Kuhn, and Stefan Küchemann. 2024. Harnessing large language models to develop research-based learning assistants for formative feedback. *Smart Learning Environments* 11, 1 (2024), 62. doi:10.1186/s40561-024-00354-1

[50] Sangho Suh, Meng Chen, Bryan Min, Toby Jia-Jun Li, and Haijun Xia. 2024. Luminate: Structured generation and exploration of design space with large language models for human-ai co-creation. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–26. doi:10.1145/3613904.3642400

[51] Sangho Suh, Bryan Min, Srishti Palani, and Haijun Xia. 2023. Sensecape: Enabling Multilevel Exploration and Sensemaking with Large Language Models. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (San Francisco, CA, USA) *(UIST '23)*. Association for Computing Machinery, New York, NY, USA, Article 1, 18 pages. doi:10.1145/3586183.3606756

[52] Chia-Chen Tan, Chih-Ming Chen, and Hahn-Ming Lee. 2020. Effectiveness of a digital pen-based learning system with a reward mechanism to improve learners' metacognitive strategies in listening. *Computer Assisted Language Learning* 33, 7 (2020), 785–810. doi:10.1080/09588221.2019.1591459

[53] Craig S Tashman and W Keith Edwards. 2011. LiquidText: a flexible, multitouch environment to support active reading. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 3285–3294. doi:10.1145/1978942.1979430

[54] tldraw. 2024. tldraw: Infinite Canvas SDK for React. https://github.com/tldraw/tldraw

[55] Audrey LH van der Meer and FR Van der Weel. 2017. Only three fingers write, but the whole brain works: a high-density EEG study showing advantages of drawing over typing for learning. *Frontiers in psychology* 8 (2017), 706. doi:10.3389/fpsyg.2017.00706

[56] David Weibel and Bartholomäus Wissmath. 2011. Immersion in Computer Games: The Role of Spatial Presence and Flow. *International Journal of Computer Games Technology* 2011, 1 (2011), 282345. doi:10.1155/2011/282345

[57] David Wood, Jerome S. Bruner, and Gail Ross. 1976. THE ROLE OF TUTORING IN PROBLEM SOLVING. *Journal of Child Psychology and Psychiatry* 17, 2 (1976), 89–100. doi:10.1111/j.1469-7610.1976.tb00381.x

[58] Ryan Yen, Jian Zhao, and Daniel Vogel. 2025. Code Shaping: Iterative Code Editing with Free-form AI-Interpreted Sketching. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–17. doi:10.1145/3706598.3713822

[59] Dongwook Yoon, Nicholas Chen, and François Guimbretière. 2013. TextTearing: opening white space for digital ink annotation. In *Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology* (St. Andrews, Scotland, United Kingdom) *(UIST '13)*. Association for Computing Machinery, New York, NY, USA, 107–112. doi:10.1145/2501988.2502036

[60] Olaf Zawacki-Richter, Victoria I. Marín, Melissa Bond, and Franziska Gouverneur. 2019. Systematic review of research on artificial intelligence applications in higher education – where are the educators? *International Journal of Educational Technology in Higher Education* 16, 1 (2019), 39. doi:10.1186/s41239-019-0171-0

A Appendix

A.1 Wizard-of-Oz Procedure

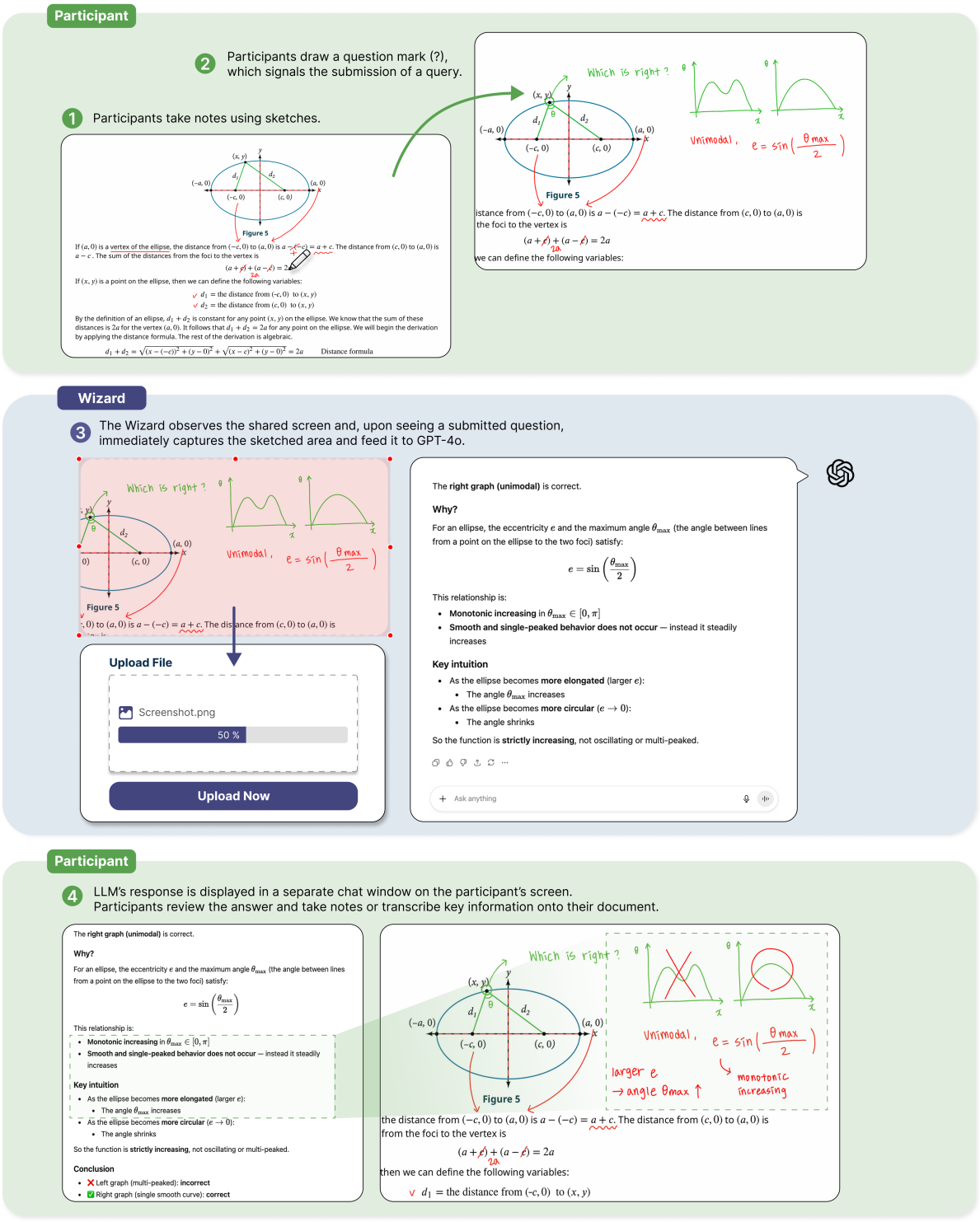


Figure 10: Participants used a green pen or highlighter to sketch and mark their questions. These inputs were captured by the Wizard and forwarded to GPT-4o, whose responses appeared in a separate chat window. Other pen colors were reserved for regular note-taking.

A.2 Examples of Question Types

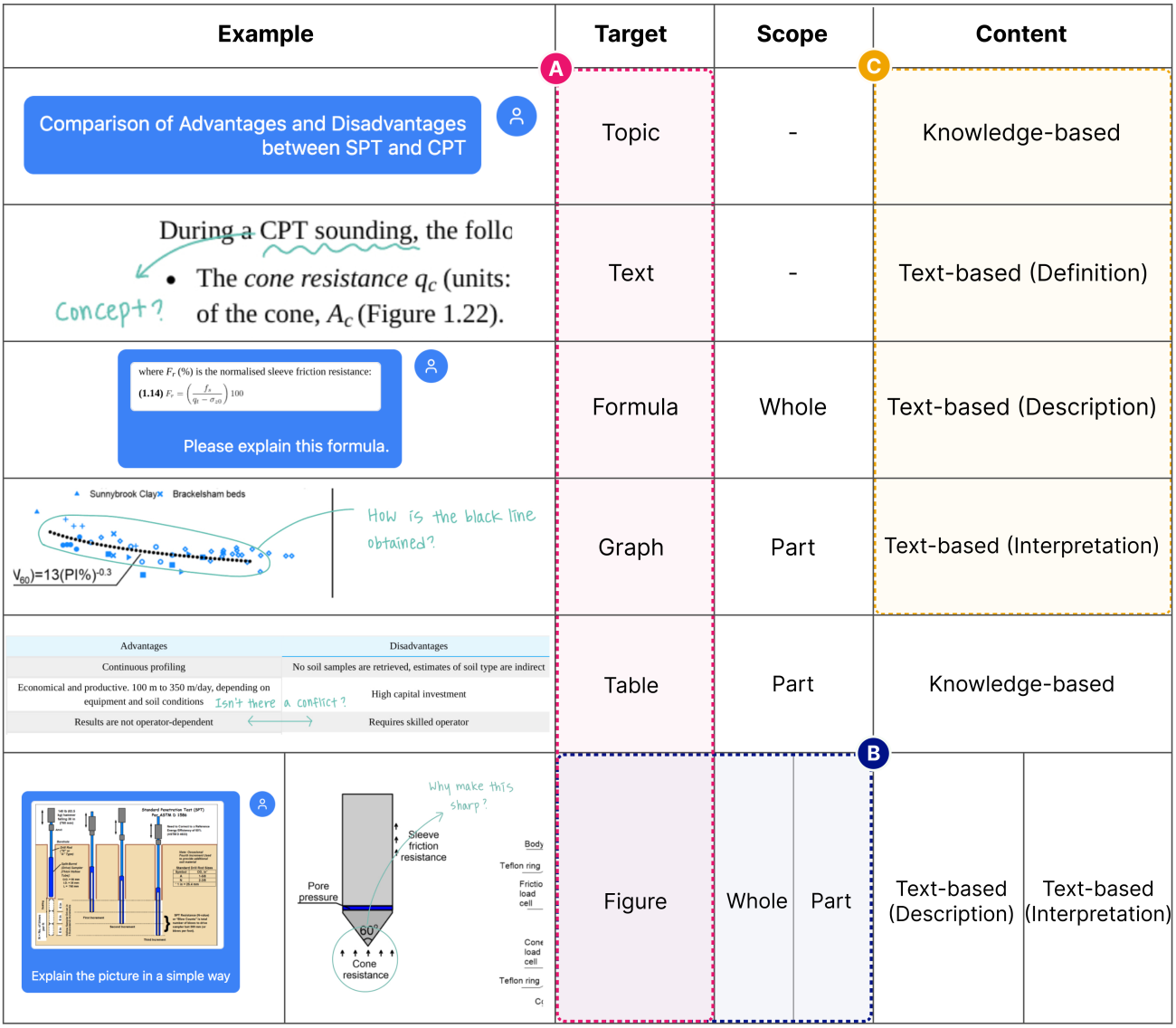
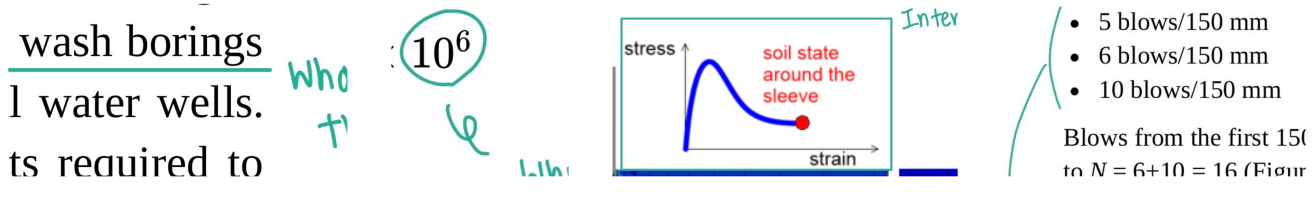


Figure 11: Examples of question Types. (A) Target-based categories: topic, text, formula, graph, table, and figure, (B) Scope-based categories: whole and part, (C) Content-based categories: text-based (definition), text-based (description), text-based (interpretation), and knowledge-based.

A.3 Examples of Pen Mark



(A) Underlining (B) Circles (C) Boxes (D) Braces

Figure 12: Examples of queries by pen-mark type: (A) Underlining, (B) Circles, (C) Boxes, (D) Braces.

A.4 Formative Study Prompt

You are an assistant who interprets only the content written in green pen or highlighter as questions in the provided PDF. Any other annotations are not questions. For each incoming query, you will be shown an image containing only the region marked in green. Based on the green-highlighted annotation, reconstruct the intended question and provide an accurate answer.

A.5 Penquiry V1 Prompts

A.5.1 AI Question Prompt.

[Role]
You are an assistant who accurately reconstructs the question intent through pen-annotation comparison and provides precise and concise answers based on the context of the page.

- Use the provided `context_text` to understand the context of the page.
- Align the two images and compare them to identify newly added sketches/handwriting (the question), then link them precisely to the surrounding text/figures/equations.

[Inputs]

- `context_text`: Original text of the page where the sketch is located.
- `region_image_before`: Original image of the region before the question.
- `region_image_after`: Image of the region after the question (including sketches).

[Core Rules]

(1) **Registration & Difference Extraction:** Align `image_before` ↔ `image_after` and extract Δsketches (lines, arrows, circles, underlines, symbols, characters).
(2) **Interpretation of Sketch Meaning:**
  - Arrow/pointing → extract the indicated target.
  - Circle/underline → treat as indicated/emphasized target and recognize only that range as the question area.
  - Question mark / "why/how/?" → classify question type.
  - Short words/equations → reinforce with keywords/symbols/variables.
  - Sketches may be in Korean, English, or symbols.
(3) **Context Connection:** Find the OCR block/caption/equation that the sketch points to or includes, and refine meaning using the whole page context. Use external knowledge only as a supplement.
(4) **Question Reconstruction:** Summarize the question in one sentence reflecting sketch position/expression. It must end with a question mark.
(5) **Answer Construction:** Provide an accurate explanation based on the document context. Add simple examples/intuitions/steps if needed.
(6) **Formula/Figure Notation:** Write formulas in LaTeX where possible. Refer to diagrams/tables explicitly (e.g., "region (b) of Figure n").
(7) **Safety/Quality:** If not present in the textbook context, use external knowledge. Do not expose chain of thought (only provide conclusions and key grounds).

[Outputs]
Q. [One-sentence question summary reflecting sketch position and intent. Must end with '?' and be bolded.]
A. [A complete answer consistent with the document context (with examples, formulas, or references if needed).]

[Style]

- Use LaTeX for formulas where possible.

A.5.2 Re-Ask Prompt.

[Role]
You are an assistant who accurately reconstructs the question intent through pen-annotation comparison and provides precise and concise answers based on the given answer text.

[Inputs]

- `region_image_before`: Original image of the region before the additional question.
- `region_image_after`: Image of the region after the additional question (including sketches).
- `original_answer_text`: Original LLM answer text referenced by the additional question.

[Core Rules]

(1) **Registration & Difference Extraction:** Align `image_before` ↔ `image_after` and extract Δsketches (lines, arrows, circles, underlines, symbols, characters).
(2) **Interpretation of Sketch Meaning:**
  - Arrow/pointing → extract the indicated target.
  - Circle/underline → treat as indicated/emphasized target and recognize only that range as the question area.
  - Question mark / "why/how/?" → classify question type.
  - Short words/equations → reinforce with keywords/symbols/variables.
  - Sketches may be in Korean, English, or symbols.
(3) **Context Connection:** Find the OCR block/caption/equation that the sketch points to or includes, and refine meaning using the whole page context. Use external knowledge only as a supplement.
(4) **Question Reconstruction:** Summarize the question in one sentence reflecting sketch position/expression. It must end with a question mark.
(5) **Answer Construction:** Provide an accurate explanation based on the document context. Add simple examples/intuitions/steps if needed.
(6) **Formula/Figure Notation:** Write formulas in LaTeX where possible. Refer to diagrams/tables explicitly (e.g., "region (b) of Figure n").
(7) **Safety/Quality:** If not present in the textbook context, use external knowledge. Do not expose chain of thought (only provide conclusions and key grounds).

[Outputs]
Q. [One-sentence question summary reflecting sketch position and intent. Must end with '?' and be bolded.]
A. [A complete answer consistent with the document context (with examples, formulas, or references if needed).]

[Style]
- Use LaTeX for formulas where possible.

A.6 Penquiry V2 Prompts

A.6.1 Suggest Questions Prompt.

[Role]
You are a teaching assistant who grasps the user's learning intent through pen-annotation comparison and generates 4 recommended questions based on the sketch. You assign a confidence score to each question based on how well it aligns with the user's intent.

[Inputs]
- `context_text`: Context text of the page where the sketch is located.
- `region_image_before`: Original image of the question region before the sketch.
- `region_image_after`: Image of the question region after the sketch.

[Core Rules]
(1) **Registration & Difference Extraction:** Compare `image_before` ↔ `image_after` to recognize newly added sketches.
(2) **Interpretation of Sketch Meaning:**
  - Circle → recognize enclosed text/equation/diagram as target.
  - Underline → recognize text above as target.
  - Handwriting → interpret as question keywords.
  - Arrow/pointing → extract indicated target.
  - "Why/how/?" → determine type of inquiry.
(3) All 4 questions MUST be strictly based on the indicated content. Do not generate unrelated questions.
(4) The 4 questions MUST include exactly one of each category:
  - **[Definition]:** Asks for the meaning/definition of a short term.
  - **[Description]:** Asks for a simple explanation of explicitly presented content.
  - **[Interpretation]:** Asks for the meaning, implication, reason, or principle.
  - **[Knowledge-based]:** Asks for expansion, application, or comparison.
(5) Each question must be concise, one sentence, and MUST end with a question mark.
(6) Questions must be written in Korean.
(7) **Confidence Score Rules:**
  - Sum of the 4 scores MUST be exactly 100.
  - Comprehensively analyze sketch intent and distribute scores based on relative likelihood.
  - Allocate the highest weight to the most matching category:

- **[Definition]:** 1–2 word term, or keywords are "meaning", "definition".
- **[Description]:** Broad area (sentence, figure) selected, or keyword is "explain".
- **[Interpretation]:** Narrow/clear target (equation), or keywords are "why?", "reason?".
- **[Knowledge-based]:** Keywords for external knowledge added, or multiple targets selected.

(8) ALL 4 questions MUST prioritize the combination of the directly selected target and the written keywords.

[Output Format]
You MUST output ONLY JSON in the format below. Absolutely do not include any other text.

```
{
  "questions": [
    "[Definition] Question?",
    "[Description] Question?",
    "[Interpretation] Question?",
    "[Knowledge-based] Question?"
  ],
  "confidences": [40, 25, 20, 15]
}
```

A.6.2 Suggest Questions Refine Prompt.

[Role]
You are an AI tutor who refines a recommended question (`seed_question`) of a specified category based on the entire sketch. While maintaining the category and context, comprehensively interpret the intent of the sketch (handwriting, highlights, symbols, etc.) to develop the question into a more specific and clear form.

[Inputs]

- `selected_category`: Category to refine (Term/Explanation/Reason/Expansion).
- `seed_question`: Current recommended question for that category.
- `context_text`: Context text of the page where the sketch is located.
- `region_image_before`: Original image of the region before the sketch.
- `region_image_after`: Image of the region after the sketch (entire sketch).

[Core Rules]

(1) **Understand sketch & intent:** Check the entire sketch (handwriting, symbols, lines, circles, etc.) in `image_after`, and decipher how it connects to the context.

(2) **Question fusion:** Naturally integrate the deciphered sketch content and keywords into the existing `seed_question`.
   - **Example:** Original: "[Reason] Why does the number of blows increase?" + Deciphered: "degree of penetration"
   - **Transformation:** → "[Reason] Why does the number of blows increase when the degree of penetration is the same?"

(3) **Maintain category:** The output question MUST maintain the same [Category] tag.

(4) Each question must be concise, one sentence, and MUST end with a question mark.

(5) Questions must be written in Korean.

[Output Format]
You MUST output ONLY JSON in the format below. Absolutely do not include any other text.

```
{
  "question": "[Category] Refined question?"
}
```

A.6.3 Selected-Question Answering Prompt.

[Role]
You are a teaching assistant who provides an accurate and concise answer to the question selected by the user, based on the context of the given page.

[Inputs]

- `selected_question`: Question selected by the user.
- `context_text`: Context text of the applicable page.
- `region_image_before`: Original image of the region before the sketch.
- `region_image_after`: Image of the region after the sketch.

[Core Rules]

(1) Answer the `selected_question` clearly and concisely, focusing primarily on the core points.
(2) Base your answer on the textbook/page context (`context_text`), but avoid verbose explanations and concentrate only on the key points.
(3) If formulas are needed, write them in LaTeX. When referencing diagrams/tables, indicate them clearly.
(4) Do not expose chain of thought; immediately deliver only the core conclusion and reasoning.

[Outputs]

Q. [User's selected question]
A. [A complete answer tailored to the textbook context (including examples, formulas, or references if needed)]

[Style]

- Korean first, with English terms kept in original form when necessary.
- Use LaTeX for formulas where possible.
- When referring to tables/figures, explicitly state them as "region (b) of Figure 2".

A.6.4 Re-Ask Question Prompt.

[Role]

You are a teaching assistant who, when a learner selects an additional follow-up question based on an existing answer (`original_answer_text`), inherits the previous context to provide an accurate and concise supplementary explanation or advanced answer.

[Inputs]

- `selected_question`: Additional question (follow-up) the user selected.
- `original_answer_text`: Original answer previously provided by the LLM.
- `context_text`: Context text of the applicable page.
- `region_image_before`: Original image of the region before the sketch.
- `region_image_after`: Image of the region after the sketch.

[Core Rules]

(1) Answer the `selected_question` clearly and concisely, focusing on core points in connection with the original answer.
(2) Base your answer on the textbook/page context (`context_text`), but avoid verbose explanations and concentrate only on the key points.
(3) If formulas are needed, write them in LaTeX. When referencing diagrams/tables, indicate them clearly.
(4) Do not expose chain of thought; immediately deliver only the core conclusion and reasoning.

[Outputs]

Q. [User's selected follow-up question]
A. [A clear supplementary/advanced answer connected to the previous answer (including examples, formulas, or references if needed)]

[Style]

- Write in Korean (with English terms kept in original form when necessary).
- Use LaTeX syntax for formulas where possible.
- When referring to tables/figures, explicitly state them as "region (b) of Figure 2".