

Difficulties in diagram interpretation in physics: An eye-tracking study of visual attention and diagram interpretation strategies

M. Klassen^{*} and G. Friege[†]

Institute for Mathematics and Physics Education, Physics Education Research Group, Leibniz University Hannover, Welfengarten 1, 30167 Hannover, Germany



(Received 12 May 2025; accepted 15 January 2026; published 23 February 2026)

Analyzing and understanding the differences between novices and experts in domains such as physics has been proven to be important for enhancing competence acquisition in that field. This study investigates these differences through an expert-novice comparison by analyzing visual attention, cognitive strategies, and behavior during scientific diagram interpretation. A total of 23 participants (11 novices and 12 experts) completed 10 diagram-based tasks, while their eye movements and verbalizations were recorded. Experts outperformed novices across several dimensions: higher task success (86% vs 70%), significantly greater visual efficiency (23% vs 17%), and longer dwell times in task-relevant areas. Their shorter, more structured scan paths reflected goal-directed, conceptually driven strategies. In contrast, novices showed disoriented gaze behavior, frequent shifts between the diagram and task text, and relied on surface-level reverse-matching strategies, indicating higher extraneous cognitive load. Think-aloud data revealed qualitative reasoning differences: experts used precise, technical language and referenced diagrams accurately, while novices were vague, hesitant, and uncertain. Emotional expressions in novice protocols aligned with signs of cognitive overload. These findings support cognitive load theory, suggesting experts draw on schemas to manage intrinsic load, while novices struggle with element interactivity. Unexpectedly, some experts in the study showed novicelike behavior, whereas some novices exhibited patterns typically associated with expert performance. Novicelike experts often lacked recent exposure to diagram-related coursework. Conversely, some novices demonstrated efficient gaze patterns and structured scanning but still made conceptual errors. To contextualize human performance, two large language models (ChatGPT-4 and DeepSeek-R1) were tested on the same tasks. ChatGPT-4 reached an overall success rate of 50%, handling simple extractions but failing with visual relationships. DeepSeek-R1 performed poorly (10%), revealing current AI limitations in visual interpretation. These results underscore the uniquely human capacity for visual reasoning, schema activation, and metacognitive monitoring, which AI still lacks. The findings emphasize the critical role of graph literacy, representational competence, and cognitive strategy use in diagram interpretation. They highlight the need for explicit instructional interventions to help novices develop expertlike visual attention patterns and reasoning skills.

DOI: [10.1103/58hn-493c](https://doi.org/10.1103/58hn-493c)

I. INTRODUCTION

Graphs, charts, and visual representations are indispensable tools in science and everyday life. They allow complex relationships and large datasets to be communicated clearly and efficiently. In physics, in particular, graphical representations not only facilitate faster data processing but also support deeper conceptual understanding.

Acquiring graph literacy begins early in education. At the school level, students are expected to create, interpret, and critique graphical data. Competence in dealing with diagrams is essential for developing scientific reasoning. At the university level, the systematic teaching of graphical representation is often neglected. Students in physics and other STEM disciplines are frequently expected to intuitively understand complex diagrams without explicit instruction.

Beyond academia, diagram literacy is a key skill for informed decision making in everyday life. Individuals are routinely exposed to graphs in news articles, political debates, and financial analyses. Public understanding of health statistics, climate trends, or economic developments increasingly depends on the ability to interpret visual data correctly.

^{*}Contact author: klassen@idmp.uni-hannover.de

[†]Contact author: friege@idmp.uni-hannover.de

Published by the American Physical Society under the terms of the [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/). Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

At the same time, advances in artificial intelligence, especially large language models (LLMs), are reshaping the landscape of knowledge work and education. While these systems demonstrate impressive performance in text-based reasoning, their ability to interpret diagrams remains limited. Including AI performance as a point of comparison therefore provides a useful contrast to human diagram interpretation and helps contextualize which aspects of visual reasoning, such as schema activation and strategic attention allocation, are uniquely human.

Prior research in physics education has typically focused on relatively narrow domains, such as specific content areas or single diagram types. While valuable, such restrictions limit the generalizability of findings across the wide range of diagram types used in physics. Moreover, much of the existing literature emphasizes task outcomes rather than the cognitive processes that underlie diagram interpretation. Understanding these processes, however, is crucial for identifying the mechanisms that distinguish expert from novice performance.

The present study seeks to address these gaps. By combining eye tracking with think-aloud protocols, it analyzes both the visual attention patterns and the reasoning strategies that participants employ while solving diagram-based tasks.

II. THEORETICAL BACKGROUND

In science, technology, engineering, and mathematics (STEM) subjects, research shows that experts process visual information more efficiently than novices [1]. This efficiency is reflected in chunking and a more selective focus on relevant areas [2]. Empirical studies support these findings: experts tend to focus more on relevant features and apply more effective visual strategies [3]. Participants with higher graph literacy spent significantly more time fixating on key elements relative to the total viewing time [4].

Graph literacy refers to the ability to understand and work with graphically presented information, including identifying relevant features and interpreting relationships between variables [5,6]. It encompasses not only general knowledge of graphical formats but also internalized schemas that guide interpretation processes. Individuals with higher levels of graph literacy tend to have more advanced educational backgrounds [7], suggesting that this skill develops through formal instruction and repeated exposure to graphical information. Research further suggests that individuals with high graph literacy are better equipped to extract meaningful patterns and avoid misinterpretations [8,9].

Building on this, expertise research emphasizes the importance of first identifying the differences between novices and experts in order to design effective learning interventions [10], particularly in their abilities and skills related to handling scientific content.

Complex scientific diagrams that require learners to simultaneously integrate multiple pieces of information often exhibit high element interactivity and place considerable demands on working memory, making them ideal for investigating the relationship between visual processing, expertise, and cognitive load [11,12]. In eye-tracking research, such processing is often examined using areas of interest (AOIs), defined as specific regions of a stimulus (e.g., axes, legends, and data points) that are selected for detailed analysis of visual attention. This methodological approach makes it possible to determine how participants allocate their attention to critical features during diagram interpretation.

A. Diagram interpretation

Developing diagram interpretation skills is crucial for physics but also essential in fields like engineering, medicine, data analysis, and daily life. In STEM disciplines, graphical representations are central to instruction and support the processing and comparison of data [13]. The ability to interpret graphs is a key ability for understanding scientific information [14]. According to Shah and Hoeffner [15], effective graph comprehension relies on several essential components, including the ability to identify relevant visual features, understand the conceptual relationships they represent, and integrate this information with prior domain knowledge.

Research on multiple representations highlights that diagrams are only one of several representational forms that learners must coordinate. According to Ainsworth [16], multiple external representations serve complementary, constraining, and constructing functions, meaning that deep understanding often requires learners to integrate and translate between diagrams, text, and symbolic notation. This perspective helps explain why diagram interpretation in physics is demanding, as students must often shift flexibly between representations while avoiding misinterpretations.

This perspective aligns with the foundational work of Larkin and Simon, who demonstrated that diagrams can facilitate understanding by making underlying structures more explicit and reducing the cognitive load associated with abstract reasoning [17].

Despite their importance, interpreting graphs presents significant challenges for learners, including difficulties in translating between graphs and equations, as well as misinterpreting context-specific data in fields like physics [18]. These challenges highlight the need to foster representational competence, which encompasses the ability to extract meaning from visual data, identify variables and relationships, and connect them to underlying concepts.

Building on Rau's framework [19], representational competence involves not only interpreting single diagrams but also flexibly coordinating multiple representations, selecting the appropriate representation for a given

problem, and translating between them. This broader competence plays a crucial role in supporting schema construction and reducing unnecessary cognitive load when learners engage with complex scientific content. Students also must learn to transfer previously acquired strategies to new contexts [20].

The presentation format of diagram-based tasks can modulate cognitive processing and visual attention. When multiple-choice options are displayed next to diagrams, learners may adopt trial-and-error strategies, shifting attention between answer options and the diagram. From a theoretical perspective, this introduces additional extraneous cognitive load [21] and may influence the deployment of representational competence. Empirical evidence suggests that multiple-choice formats affect both visual attention patterns and problem-solving strategies in physics tasks [22,23]. Accounting for these effects is therefore essential for understanding how learners construct meaning and apply visual strategies when interpreting complex scientific diagrams.

The need for strong graph literacy skills in physics education is also evident in the way physics assessments are structured. Wang and Wei [24] found that many physics exams, particularly in China, heavily rely on students' ability to interpret given diagrams. They cautioned that this emphasis on interpretation without requiring students to construct or critically analyze these diagrams might limit the development of deeper graph literacy. To foster a more comprehensive understanding of physics, they recommended that educational systems focus more on developing students' ability to create and evaluate diagrams, not just interpret them.

In addition to these content-specific challenges, learners' self-perceptions also play a crucial role in diagram interpretation. Concepts such as self-efficacy [25] and metacognition [26] help explain how students monitor, regulate, and evaluate their own abilities. Low self-efficacy may lead novices to underestimate their competencies, while limited metacognitive awareness can prevent them from effectively reflecting on their strategies. These psychological factors are therefore highly relevant when analyzing self-assessment data in the context of diagram-based tasks.

These difficulties can also be understood from a cognitive perspective. Sweller's cognitive load theory (CLT) provides a theoretical framework to explain the mental demands placed on learners when engaging with complex visual materials [21]. According to this theory, working memory has a limited capacity, and effective learning depends on how cognitive load is managed. CLT distinguishes between intrinsic load (inherent complexity of the content), extraneous load (load imposed by poor design or irrelevant information), and germane load (mental effort devoted to constructing and automating schemas). In particular, complex diagrams often require learners to process multiple visual elements simultaneously, such as

variable relationships, axes labels, and contextual annotations, which increases element interactivity and contributes to intrinsic cognitive load [11,27].

Chi *et al.* [28] further elucidated that the development of expertise involves the acquisition of structured knowledge schemas, enabling experts to chunk information effectively and reduce cognitive load during tasks. Prior knowledge helps manage this load by enabling chunking [29] and selective attention [30,31], allowing experts to reduce irrelevant processing and focus on meaningful features.

B. Expert-novice comparison

While experts are skilled at interpreting and utilizing these visual representations, novices often struggle to extract relevant information or accurately interpret complex diagrams [32].

This distinction is well captured by the cognitive theory of visual expertise [33], which describes how expertise develops through perceptual learning, attentional control, and the progressive automatization of domain-specific visual strategies. CTVE provides a valuable framework for interpreting our findings on diagram use, as it emphasizes both the perceptual and cognitive mechanisms that differentiate novices from experts.

When interpreting scientific diagrams, especially in physics, novices may experience a high intrinsic load due to a lack of prior knowledge and struggle with unnecessary extraneous elements that impede comprehension. This can hinder not only understanding but also the ability to transfer learning to new contexts.

This is supported by eye tracking findings indicating that experts and novices differ significantly in their gaze behavior: experts show more task-specific fixations and structured scan paths, while novices exhibit more exploratory patterns that suggest uncertainty and inefficient processing [34–36]. Recent studies have shown that correct graph solvers (across disciplines) exhibit similar, effective scanning strategies, whereas incorrect solvers focus on less informative features. This underscores that expertlike reasoning in physics involves using visual cues efficiently [37].

Research in cognitive science has consistently shown two main differences between experts and novices: how they structure their knowledge and the strategies they use during their process [38–40]. Experts tend to organize their knowledge hierarchically, clustering it around a few core principles, which allows them to access and apply relevant concepts quickly [41,42]. Their decision-making processes are highly structured and efficient, often proceeding automatically without conscious awareness [28]. Novices, in contrast, possess fragmented knowledge and rely on surface-level cues or rote procedures that are heavily context-specific.

Experts typically begin by qualitatively analyzing a situation, identifying the fundamental principles that apply, and constraining the problem accordingly [43]. Their

subsequent approach is organized, principle-driven, and includes self-monitoring mechanisms to verify progress and correctness [44]. Metacognitive control is therefore a hallmark of expertise. Novices, however, tend to jump directly to manipulating equations based on identified quantities, often neglecting a deeper qualitative analysis unless explicitly instructed [45].

The path to developing expertise in physics involves building strong conceptual frameworks and integrating them with effective strategies [46]. Knowledge organization are mutually reinforcing: improvements in one domain lead to advances in the other. Research has shown that enhancing students' general abilities also contributes to their conceptual understanding [47–49].

C. Eye tracking and think-aloud protocols

Combining think-aloud protocols with eye tracking provides a richer, more complete picture of student cognition in physics tasks than either method alone. Eye tracking reveals where students look (visual attention and information gathering), while think-aloud protocols reveal why they look there (underlying reasoning, strategies, and activated knowledge). In physics education research, this multimodal approach has begun to yield insights into diagram interpretation. For example, a pilot study found that tracking students' eye gaze while they verbalized their reasoning revealed phenomena such as “foretelling” eye movements and pauses in speech that aligned with shifts in gaze [50].

Think-aloud protocols ask learners to verbalize their thoughts while examining a diagram. This method, long used in cognitive and educational research, yields qualitative data about learners' reasoning, interpretations, and strategy use [51], providing a window into the germane cognitive load involved in diagram interpretation. The combination of eye-movement data and concurrent verbalizations enables a more comprehensive analysis of the cognitive demands associated with diagram interpretation.

Eye-tracking technology has emerged as a powerful tool in educational research. By tracking eye movements, researchers can gain a deeper understanding of learners' cognitive processes, such as how they process information and which areas of diagrams capture their attention. Eye tracking in physics education research underscores its growing importance in providing valuable insights into the processes involved in students' learning [1]. It especially enables the exploration of how multiple representations are used in diagram interpretations by analyzing the areas of interest (AOIs).

Recent studies using eye tracking have explored students' visual attention during diagram-based interpretation tasks [13,34,52–54]. Many prior studies on graph literacy and diagram interpretation have focused narrowly on specific domains, such as graphs in kinematics, which

may limit the generalizability of their findings to other types of scientific diagrams [22].

Rarely have studies investigated multiple types of diagrams across various physics topics, even though such an approach would enable a more comprehensive understanding of visual strategies and representational competence.

For example, Klein *et al.* [35,36] found that vertical and horizontal saccades were linked to better performance in tasks involving Cartesian diagrams and vector fields. Further, Klein *et al.* [54] showed that physics students follow line graphs more closely with their eyes, indicating better understanding of slope.

Overall, these studies confirm that eye tracking effectively reveals cognitive strategies in working with diagrams and offer insights into analyzing visual interaction. Additionally, more expertlike scan paths, such as those following the gradient of line graphs, have been interpreted as indicators of “theoretical” perceptual actions and correct conceptual processing [35,36,54].

Eye-tracking studies have shown that unknown or unfamiliar axis labels attract more visual attention, suggesting that prior domain knowledge shapes not only interpretation but also visual attention distribution [34,54]. Such findings further support the role of graph literacy and expertise in reducing cognitive effort during interpretation.

From a cognitive load perspective, eye-movement measures—such as dwell times, fixation distributions, and gaze transitions—can be interpreted as indirect indicators of cognitive processing demands. Longer dwell times in task-relevant areas are often associated with deeper processing and schema-driven attention, whereas frequent gaze shifts, scattered scan paths, or prolonged attention to irrelevant areas have been linked to increased extraneous cognitive load and uncertainty [21,34,36]. In this sense, eye tracking does not measure cognitive load directly, but rather provides behavioral correlates of how learners allocate limited working memory resources during complex visual tasks. When combined with verbal data, such as think-aloud protocols, eye-movement patterns can therefore offer converging evidence for differences in intrinsic and extraneous load between novices and experts.

D. AI comparison to human performance

Recent research highlights significant limitations in AI tools like ChatGPT-4, particularly when it comes to interpreting visual information such as diagrams (e.g., “Our quantitative analysis found that ChatGPT's performance is far from expertlike” [55]). While AI models excel in providing logical reasoning and formulating problem-solving strategies in text-based contexts, their performance tends to falter when tasked with interpreting diagrams that are central to physics problems.

For instance, studies have shown that ChatGPT-4, even with image input capabilities, struggles with interpreting graphs and diagrams. Polverini and Gregorcic [55] found

that ChatGPT-4 was able to provide correct problem-solving strategies for graphs in kinematics but failed to accurately interpret key visual elements, reflecting a broader issue in AI's ability to engage with physics content visually. Similarly, Aldazharova *et al.* [56] observed that ChatGPT-4 struggled significantly with problems involving the interpretation of figures, often leading to incorrect conclusions despite sound reasoning. These findings suggest that current AI tools, while advanced in language processing, still lack reliable visual comprehension skills, especially in the context of complex physics problems.

Further research by Polverini *et al.* [57] also emphasized the persistent difficulties of AI in understanding spatial information and diagrams. Even with the enhanced visual capabilities of ChatGPT-4o, errors in interpreting diagrams such as vector fields and circuit diagrams were common, particularly in tasks requiring spatial reasoning, like applying the right-hand rule [57].

E. Framing the research questions

The difficulty of AI models in accurately interpreting complex diagrams, particularly with respect to spatial relationships and multiple interacting elements, aligns with findings from physics education research showing that even human learners require domain knowledge and representational competence to correctly interpret similar visual information. Küchemann *et al.* [58] emphasize that the ability to fluently move between visual and mathematical representations is essential for achieving deep comprehension of physics concepts. Without this skill, students are often unable to effectively understand or solve problems, regardless of their textual understanding.

Studies on student behavior further support the importance of graph interpretation skills. For example, Madsen *et al.* [59] found that students who successfully solved physics problems tended to focus their attention on the most relevant parts of a diagram, using expertlike visual scanning strategies.

Given these insights, it is clear that graph literacy is not just a supplementary skill but a fundamental aspect of physics expertise. In contrast, AI tools such as ChatGPT-4 or Deepseek-R1, despite their advancements in text-based problem solving, remain limited in their ability to engage with and interpret complex visual representations. This highlights a gap that currently exists between human expert performance in physics and the capabilities of AI in educational contexts.

While previous research has highlighted the relevance of graph literacy and differences between experts and novices in diagram interpretation, most studies have focused primarily on final task performance rather than the underlying cognitive processes, leaving a gap regarding real-time visual attention and strategy use. Many existing studies focus narrowly on specific subdomains such as kinematics,

often limiting the variety of diagram types under investigation.

However, graph literacy requires the ability to interpret a broad range of diagrammatic representations. There is a lack of empirical studies systematically comparing how individuals with different levels of expertise engage with a diverse set of diagrams. Also, prior work has focused primarily on end products of diagram interpretation (e.g., correctness of answers), offering limited insight into the underlying cognitive processes. Yet, the interpretation of complex visual information involves a dynamic interplay between attention, reasoning strategies, and prior knowledge. Investigating these processes in real time is essential for understanding how representational competence develops.

Beyond the focus on cognitive processes, participants in prior studies on expert-novice differences in physics diagram interpretation were often recruited from a single academic setting, thereby limiting the generalizability of the findings [60].

This study seeks to address these gaps by adopting an expert-novice comparison, applying eye tracking and think-aloud methods to investigate diagram interpretation processes across multiple diagram types in physics. Indeed, individual differences observed in previous research suggest that graph literacy is better conceptualized as a continuum rather than a strict binary classification [7]. This perspective underpins the current study's design, which aims to identify differences in strategy use and performance between participants with different levels of expertise and to examine how these strategies are applied during the interpretation of various diagram types.

This will be guided by the following research questions:

RQ1: How do the success rates of experts and novices differ for simple and complex diagrams?

RQ2: How do visual attention patterns (e.g., dwell times in AOIs and scan paths) differ between experts and novices when interpreting diagrams?

RQ3: What differences can be observed in the reasoning strategies and verbal expressions of experts and novices during diagram interpretation?

RQ4: How do the performances of AI models (ChatGPT-4 and DeepSeek-R1) compare to those of experts and novices on the same diagram-based tasks?

III. METHODS

A. Participants

The present study was conducted with a sample of 23 students, divided into two groups: 11 novices (age: $M = 25.55$; $SD = 5.01$; $Min = 22$; $Max = 39$; gender: 3 male, 8 female; semester: $M = 7.5$) studying special education and 12 experts (age: $M = 24.08$; $SD = 1.61$; $Min = 21$; $Max = 27$; gender: 8 male, 4 female; semester: $M = 10$) studying physics.

TABLE I. Overview of the ten physics diagram-based tasks used in the study.

Task no.	Diagram type	Complexity level	Notes on task demand
1	Diode characteristic curve	Simple	Direct reading of voltage and current values
2	Velocity-time graph	Simple	Straightforward linear relationship
3	Force-displacement plot (points only)	Moderately higher	No connecting lines; approximate value reading
4	Refraction diagram (y given, read x)	Moderately higher	Consultation of axis labels; basic conceptual mapping
5	Distance-time (dotted, exponential)	Intermediate	Exponential relationship; requires interpretation of trend
6	Radioactive decay (two functions)	Intermediate	Increased conceptual difficulty
7	Harmonic oscillator (amplitude, distractors)	High	Determine amplitude; ignore misleading zero-crossing options
8	Current-time graph	High	Apply conceptual knowledge
9	Logarithmic spiral (radius at $360^\circ/2\pi$)	High	Integrates angle/radius understanding
10	Phase-space diagram (multiple trajectories)	High	Analyze maximum velocities; consult legend for multiple oscillators

The novice group consisted of bachelor students with minimal physics knowledge beyond upper-secondary school, attending an introductory seminar in physics and technology at our institute. This provided a relatively homogeneous and accessible sample comparable to the beginning of undergraduate physics studies. The expert group included advanced master's students in secondary physics education nearing completion of their degree. The choice of these groups ensured that they represented learners with limited domain-specific knowledge, while still allowing for the observation of individual differences.

Before the study began, participants were informed about data protection and received a short briefing on the eye-tracking methodology. The testing room was darkened to minimize external visual noise and reflections, which can especially affect data quality for participants wearing non-anti-reflective glasses.

B. Diagram complexity classification

Participants completed all tasks in a fixed order, which was designed to gradually increase in complexity. The sequence began with simple diagrams featuring few interacting elements and familiar physics contexts and progressed to more complex or abstract diagrams. This ordering aimed to reduce initial cognitive overload, allowing participants to become familiar with the task format before confronting higher element interactivity and less familiar physics scenarios.

Diagram complexity was defined *a priori* based on theoretical considerations grounded in cognitive load theory (CLT) [21]. Complexity was conceptualized along two dimensions: the number of interacting components within a diagram (element interactivity) and the conceptual familiarity or difficulty of the underlying physics topic.

Simple diagrams contained fewer interrelated components and straightforward relationships between variables (e.g., linear decay and velocity-time graphs), while complex diagrams involved multiple interacting variables,

nonlinear relationships, or abstract representations such as logarithmic spirals or phase-space diagrams of harmonic oscillators. Task difficulty was also influenced by factors such as the requirement to integrate information from axes, legends, or multiple trajectories, as well as the presence of distractors or less familiar physics contexts.

For clarity and transparency, Table I provides an overview of all ten diagram-based tasks, including the diagram type, and assigned complexity level. Performance data (success rates) were subsequently analyzed to validate the difficulty categorization, but they were not used to determine which diagrams were classified as simple or complex.

All diagram-based tasks were self-constructed by the authors and designed to reflect common diagram types used in secondary and early university-level physics education. To enhance transparency and reproducibility, representative examples of the diagrams and corresponding task prompts are provided as Supplemental Material [61].

C. Expertise classification

Participants were classified based on their field of study, with one group consisting of master's students in physics and the other of bachelor students in special education (no STEM subject). This selection enabled a direct comparison between two distinct groups in terms of their academic background and assumed levels of representational competence. In cases where individuals deviated strongly from their group's performance, their self-assessments and previous university coursework involving graph interpretation were used to provide explanatory context.

D. Procedure

The study includes questionnaires that were administered before and after the study. The prestudy questionnaire included demographic items, as well as questions on participants' prior experience with physics or science-related diagrams and their self-assessed competence in

interpreting such diagrams. The poststudy questionnaire asked participants to rate the difficulty of the tasks and how well they were able to manage them.

Participants had to solve ten diagram-based tasks, where the solution was directly extractable from the respective diagram, by answering single-choice questions with five answer options, each within a 60-s time limit. The 60-s limit was chosen based on pilot testing, which indicated that most participants could complete an initial scan and begin interpretation of each diagram within this time frame. This duration balances the need to capture meaningful engagement while maintaining experimental feasibility across multiple items. All tasks were presented as single-choice items, with the answer options displayed immediately next to each diagram. This setup allowed us to analyze participants' transitions between the diagrams and the answer options to capture potential trial-and-error strategies and the interaction between visual attention and decision making. The ten diagrams used in the study were self-constructed and initially based on common content from physics education, drawing on topics familiar mostly from school and the first and second years at the university.

To increase task complexity, diagrams with a stronger mathematical orientation were included. These consisted of a diagram of a logarithmic spiral and a phase-space diagram of the undamped harmonic oscillator, showing three trajectories with different initial energy levels. The diagrams spanned a wide thematic range across physics, from classical dynamics to radioactive decay. All tasks were presented in the same fixed sequence for every participant.

In this study, diagram complexity does not refer to empirically observed task difficulty, but was defined *a priori* based on theoretical considerations grounded in cognitive load theory. Specifically, challenge was operationalized through the number of interacting elements within a diagram (element interactivity) and the conceptual familiarity of the underlying physics topic.

Areas of Interest were defined *a priori* to capture participants' attention allocation (see Fig. 1). Relevance was determined based on the information necessary to solve each task. Accordingly, AOIs included the axes and axis labels, reference points on the function, the legend, the title and task description, as well as the correct answer option.

For broader gaze-shift analyses, two larger AOIs were defined to capture transitions between major display regions: the diagram and the task prompt (see Fig. 2). This allowed us to measure both detailed visual attention within diagrams and broader scanning behavior between the task description and diagram.

Two complementary types of Areas of Interest (AOIs) were defined to address different aspects of visual processing. Fine-grained AOIs within the diagram (e.g., axes, curves, legends, reference points, correct answer option) were used to analyze dwell times and to compute visual

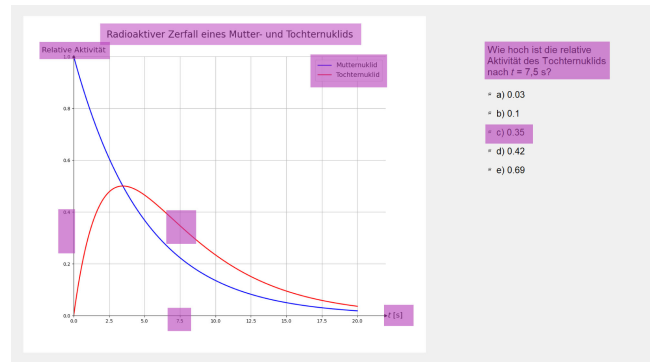


FIG. 1. Item design with task representing the structure of all ten items, areas of interest marked as colored rectangles.

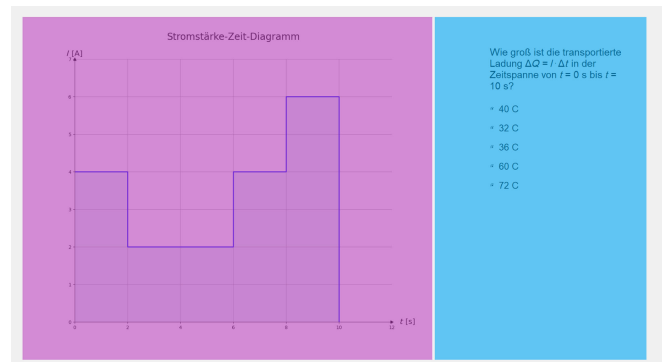


FIG. 2. Defined areas of interest (AOIs) used in the gaze shift analysis: task description (right section) and diagram (left section).

efficiency as the proportion of time spent in task-relevant regions. In addition, two larger AOIs were defined at the level of the entire diagram and the task description to capture gaze transitions between these regions. This distinction allowed us to separately analyze detailed information extraction processes and broader strategic scanning behavior.

Each item was implemented as an individual digital stimulus to allow for targeted analysis and interpretation. The interactive items were created using Python, combining the Tkinter library for the graphical interface with Matplotlib for diagram visualization. This setup enabled the precise placement and formatting of complex scientific graphs alongside contextually tailored questions and response options. Numerical values were calculated dynamically within the code to ensure internal consistency between the graphical data and the multiple-choice alternatives. All participants received identical diagrams and identical answer options. No randomization or variation across participants occurred.

In addition to the eye-tracking measures, think-aloud protocols were collected. Participants were asked to articulate their thought processes aloud while solving each diagram task. Audio recordings were captured using the Zoom H2n Handy Recorder. The verbal data were

transcribed using the F4TRANSCRIPT software and later coded into iterative categories.

Eye-tracking data were collected using the EyeFollower device from Interactive Minds. The system uses binocular tracking with separate cameras for each eye, automatic face and eye detection, and real-time pupil tracking. Calibration was performed prior to the diagram task phase using a standard nine-point dot-following procedure on screen and repeated as necessary until gaze data fell within the accepted tolerance range.

Participants were seated approximately 40 cm from the eye tracker. The EyeFollower operates at a constant sampling rate of 120 Hz and offers a tracking accuracy of $< 0.4^\circ$ visual angle across its full head-motion range [62]. The Nyan 3.0 XT software was used for data acquisition and subsequent analysis. Also, dwell times in predefined, task-relevant AOIs and overall scan paths of eye movements were recorded. Performance was operationalized by calculating an efficiency metric, dividing the dwell time in AOIs by the total time per diagram. Transitions between the task description and the diagram were specifically analyzed.

Additionally, the task descriptions and diagrams were processed by the AI models ChatGPT-4 and DeepSeek-R1 to evaluate their success rates and compare them with the performance of human participants.

To address each research question, we mapped specific data sources and analyses accordingly. RQ1 was analyzed using the multiple-choice task performance, with accuracy per diagram and participant compared across groups. RQ2 was analyzed using dwell times and scan-path measures from eye-tracking data, focusing on AOIs and gaze transitions. RQ3 was explored through inductive coding of think-aloud transcripts, with individual utterances as the unit of analysis. Given the limited amount of verbal data, no formal coding scheme or interrater reliability was applied. Instead, transcripts were examined iteratively to identify patterns related to reasoning strategies, verbal precision, and indicators of uncertainty or cognitive overload. This qualitative analysis was conducted in conjunction with the corresponding eye-tracking data to better interpret participants' visual attention and strategy use. RQ4 was addressed by applying the same tasks to ChatGPT-4 and DeepSeek-R1 and comparing their success rates and error patterns to human performance.

IV. FINDINGS

A. Questionnaire

1. Prestudy

Participants' self-perceived competence in interpreting scientific diagrams was measured using a five-point Likert scale. The results revealed a significant difference between the two groups. Experts reported a higher average confidence ($M = 3.75$, $SD = 0.72$) compared to novices, who

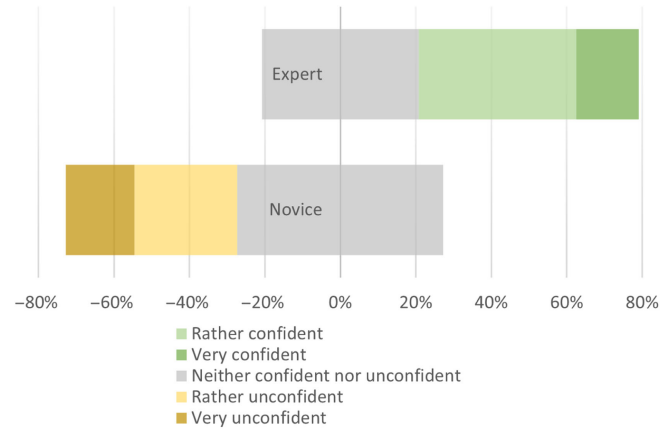


FIG. 3. Comparison of self-perception of diagram interpretation between experts and novices.

reported lower self-assessed competence ($M = 2.36$, $SD = 0.77$) (t test, $p < 0.01$, $d = 1.86$).

Figure 3 illustrates the comparison of self-perception of diagram interpretation between groups. A strong disparity was found: 92% of experts reported having attended courses that included diagram interpretation, compared to only 27% of novices. The same trend was visible in current coursework, with 58% of experts but none of the novices enrolled in classes involving diagram interpretation.

Notably, mismatches appeared between self-assessments and actual performance. One novice achieved 80% accuracy in relatively short task durations, yet rated their competence very low. Experts, in contrast, generally aligned their self-assessment more closely with actual performance.

2. Post-study

Both groups agreed that they had sufficient time to complete the tasks, with no strong disagreement recorded. Agreement was particularly high among participants who rated themselves as “very confident” or “rather confident.”

Figure 4 shows perceptions of time, task difficulty, and clarity. Novices were more likely to disagree with the statement “The diagram-related tasks matched my level of knowledge,” especially those with lower self-assessed confidence. Most participants found the multiple-choice options clear, with experts giving the highest ratings.

B. Diagram tasks

1. Overall average time

The total time summed across all diagrams by novices and experts was compared. On average, experts spent less time per diagram (based on the total time divided by the number of diagrams) compared to novices, suggesting a trend toward more efficient interpretation. The overall average total time for novices was 26.53 s ($SD = 7.85$), while experts spent 24.14 s ($SD = 8.67$). Although the difference is not statistically significant (t test, $p = 0.547$,

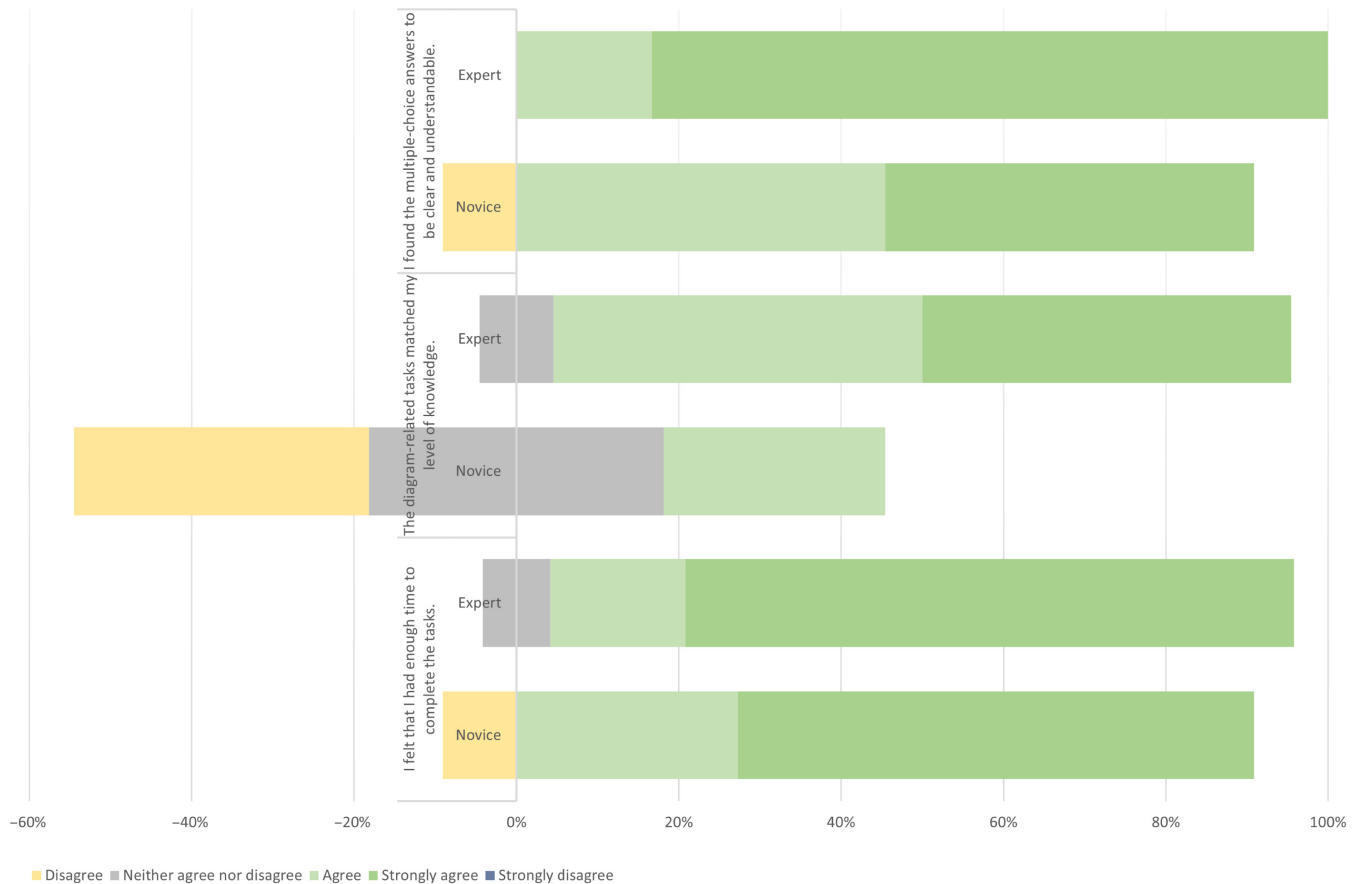


FIG. 4. Comparison of novices' and experts' perceptions of time, task difficulty, and clarity.

$d = 0.14$), it is noteworthy that for certain more challenging diagrams, both novices and experts exhibited significantly higher times. Some novices also skipped certain more challenging diagrams, resulting in outliers with significantly shorter aggregated total times. When outliers related to skipped items were excluded, the difference reached statistical significance ($p < 0.04$, $d = 0.92$).

Experts, while also spending more time on difficult diagrams, showed greater consistency in their approach and ultimately higher success rates: based on the proportion of correctly answered items, summed across all diagrams, experts achieved an aggregated average success rate of 86% correct responses ($SD = 23.3\%$), significantly outperforming novices, who averaged 70% correct responses ($SD = 36.4\%$) (t test, $p < 0.01$, $d = 0.54$). Here, a correct response was defined as the selection of the correct answer option in multiple-choice format.

2. Expertise reversal

It is important to note that participants categorized as experts and novices were not domain experts *per se* but were nominally classified based on their academic background and level of study. However, individual performance revealed a more nuanced picture. Within the expert

group ($n = 12$), three participants displayed novicelike behavior, while in the novice group ($n = 11$), two participants exhibited expertlike gaze patterns. All participants remained in their original groups; these individual deviations were analyzed qualitatively to highlight variability in reasoning and performance. Specifically, expert participants whose visual efficiency dropped below 21% were considered novicelike, while novice participants exceeding an efficiency of 19% were classified as expertlike.

Interestingly, despite the similarity of gaze patterns, there were notable differences in task performance between these subgroups (outliers). The novicelike experts spent a relatively high amount of time in white space (empty or unmarked areas of a visual layout that do not contain visual content or stimuli at all) and significantly less time in task-relevant areas. Their average success rate remained high at 90%. In contrast, the expertlike novices, who demonstrated structured scanning and efficient attention allocation, achieved only a 70% success rate. This discrepancy indicates that while their gaze behavior resembled that of experts, they likely lacked the necessary domain-specific knowledge. For example, typical errors among this group included misjudging the amplitude of a sine wave or misinterpreting the area under

a curve as charge, revealing conceptual misunderstandings in basic physics. Interestingly, experts who were not currently engaged in diagram-related courses displayed novicelike gaze behavior, while novices with expertlike gaze patterns were attending courses where they frequently worked with diagrams.

3. Think-aloud

The analysis of the think-aloud protocols reinforces the observed differences in overall task duration and strategy use between experts and novices. Novice frequently used verbal fillers such as “um,” “hm,” or “ah,” especially during moments of hesitation or when articulating spontaneous decisions. Statements like “Okay, um, voltage 0.6. Here we are.” or “Oh, I have no idea.” illustrate this tendency. Statements like “I don’t know how this symbol is pronounced” or “What do I know?” show that participants lacked knowledge of certain symbols or terms, often used in science literature.

A particularly revealing difference emerged in terms of linguistic precision. Experts demonstrated structured and precise language, making targeted use of technical terminology and offering clear references to diagrams. Examples include statements like “I’ll look at where 0.6 Volt is on the X-axis,” or “So, I look at where the red curve is at 7.5 s.” In contrast, novice responses were often vague and unstructured. A typical example might be: “There, at 7.5 ... it goes up ...” These statements lacked not only precise localization but also crucial prepositions, which contributed significantly to the clarity in expert explanations.

Emotional expressions such as “Oh God” or “I have no idea” appeared predominantly in novice protocols, suggesting signs of cognitive overload.

The structured focus of experts in the quantitative data was echoed in their verbalizations. Experts often referred specifically to diagram features with technical precision (“the red curve at 7.5 s”), which corresponds with their longer dwell times in relevant AOIs. Novices, by contrast, frequently expressed confusion or asked aimless questions aloud (“How big is the transported charge? No idea.”), which aligns with their scattered gaze patterns and shorter dwell times in task-relevant areas.

C. Attention patterns

1. Efficiency

When examining the dwell time in predefined areas of interest (AOIs), a clear distinction emerges between the two groups. Experts spent significantly more time in task-relevant areas of the diagrams compared to novices. On average, novices spent 3.32 s (SD = 1.77) in the AOIs, while experts spent 5.98 s (SD = 2.79). The difference is significant (t test, $p < 0.03$, $d = 1.17$). This focused attention results in an efficiency rate of 23% (SD = 4.5%) for experts, compared to 17% (SD = 4.1%) for novices.

2. Scan-paths

The scan paths, representing the total distance traveled by the gaze during the diagram interpretation process, were also analyzed. Novices exhibited an average scan path length of 92.68 (SD = 22.19) units, while experts had a shorter average scan path of 85.54 (SD = 32.87) units. A t test showed no statistically significant difference ($p = 0.597$, $d = 0.14$). Excluding the same outliers as in Sec. IV B 2. resulted in a significant difference ($p < 0.02$, $d = 1.07$). Also, novices exhibited more gaze shifts between the two AOIs (the diagram and the task description [see Fig. 5]) (10.04 shifts, SD = 2.96), while experts demonstrated purposeful transitions (9.23 shifts, SD = 2.8). A t test after excluding outliers showed statistically significant difference in transitions between diagram and task ($p < 0.01$, $d = 0.28$).

D. AI comparison

The performance of ChatGPT-4 and DeepSeek-R1 was evaluated for comparison with human groups. ChatGPT-4 achieved an average success rate of 50%, performing well on diagrams that required straightforward value extraction. However, it struggled significantly with diagrams involving multiple graphs or requiring interpretation of relationships between variables, leading to a success rate of 0% for these particular diagrams. DeepSeek-R1, on the other hand, achieved a notably lower success rate of 10%. Its responses often indicated that it could only process task descriptions from image files and lacked the ability to interpret diagrams.

V. DISCUSSION

RQ1: The findings indicate that experts not only possess more experience but also exhibit a more positive self-concept regarding their ability to interpret diagrams. Experts who performed at a novicelike level tended to rate their confidence as neither confident nor unconfident, suggesting a realistic self-perception that corresponded more closely with their actual performance.

The poststudy questionnaire further showed that novices frequently disagreed with the statement that the tasks matched their level of knowledge, while most experts agreed. This perceived mismatch between prior knowledge and task requirements suggests that novices faced greater cognitive demands when processing the diagrams.

This supports the assumption from cognitive load theory (CLT) that high element interactivity in complex diagrams, meaning that many elements are closely connected and must be understood together rather than separately [11], increases intrinsic cognitive load, especially for novices who lack automated schemas to guide interpretation [11,21].

Novices often exhibited marked difficulties in managing the increased complexity. As CLT posits, this may be due to

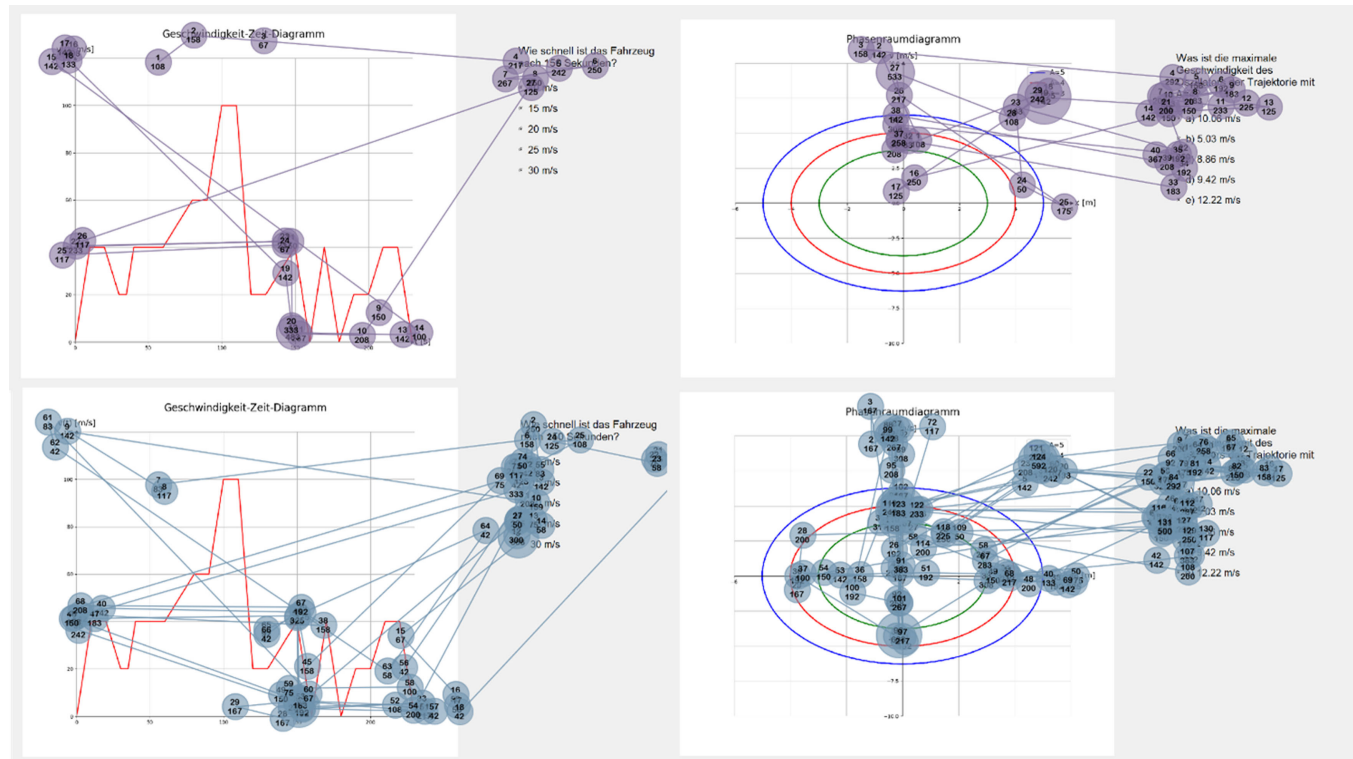


FIG. 5. Comparison of diagram 2 and 10 interpretation process; expert no. 3 (top) and novice no. 12 (bottom).

a lack of chunking and schema-based processing, which typically characterizes expert performance [29,30].

RQ2: Eye-tracking data indicated that, when confronted with more challenging diagrams, novices frequently began to shift their gaze in an unstructured manner, often oscillating between the task description and the diagram.

It should be noted that tasks were generally constrained to a maximum of 60 s per diagram. In practice, some participants completed simpler diagrams more quickly, and particularly complex diagrams were sometimes terminated before the time limit. While this approach allowed for largely standardized data collection, it may have limited novices' opportunity to fully engage with more complex diagrams. Consequently, aspects of their gaze behavior or verbalized reasoning may not have been fully captured, potentially underestimating the cognitive effort required. Future studies could consider longer or adaptive time limits to better accommodate differences in processing speed and strategy deployment.

It is important to note that all tasks were designed as multiple-choice items with answer options displayed adjacent to the diagrams. This format inherently allows for trial-and-error strategies, particularly for novices who may scan the answer options first and then seek corresponding elements in the diagram rather than systematically interpreting the diagram itself. Prior research indicates that the multiple-choice layout can significantly influence visual attention allocation and cognitive processing [22,23].

The longer dwell times of experts indicate a more deliberate and targeted visual search strategy, while novices exhibit more exploratory or inefficient gaze patterns. This supports prior findings that novices lack selective attention mechanisms and instead distribute their attention inefficiently across irrelevant areas, resulting in higher cognitive load [2,31].

This visual search behavior reflects the theoretical concept of “perceptual chunking” and structured scanning in experts [29,54]. The scan path of experts suggests that they are better able to identify and focus on the most relevant areas of the diagram more quickly. This pattern of visual attention aligns with the cognitive strategies employed by experts, who are able to extract relevant information more efficiently and avoid unnecessary exploration of irrelevant areas.

This aligns with prior studies that associate longer and more scattered scan paths with exploratory behavior and cognitive uncertainty in novices [35,36]. Figure 5 reveals that novices indeed exhibit more scattered eye movements.

These scan path irregularities can be interpreted as signs of disorientation and inefficient strategy use, confirming findings from eye-tracking research that associates such patterns with lower levels of graph literacy and increased extraneous load [34,36]. These patterns were consistently observed when reviewing all participants' scan paths for each diagram, indicating that novices tend to follow less structured and more random paths through

the diagram, while experts demonstrate a more focused and efficient approach.

Incorporating this perspective helps contextualize why some observed gaze patterns and transitions may reflect strategic compensation for task structure rather than purely content-driven reasoning. By explicitly acknowledging the role of task format, we recognize that cognitive load, attention allocation, and representational competence interact not only with diagram complexity but also with the constraints and affordances of the assessment design.

RQ3: Additionally, rather than extracting information directly from the diagram, many novices adopted a compensatory approach by scanning the multiple-choice options first and attempting to locate visually matching elements within the diagram. This reverse-matching strategy suggests a surface-level approach to diagram interpretation, in contrast to the deeper, content-driven engagement typically observed in experts. Such behavior reflects the absence of internalized interpretive schemas and a reliance on superficial cues, as theorized in graph literacy literature [5,9].

The observed reverse-matching strategy among novices can be understood as a compensatory behavior resulting from insufficient prior knowledge and weakly developed interpretive schemas. Instead of constructing meaning from the diagram itself, novices appeared to rely on surface-level visual correspondences between answer options and diagram features. From a cognitive load perspective, this strategy likely increases extraneous load, as attention is repeatedly shifted between diagram and answer options without reducing task complexity. The frequent speech pauses and verbal fillers observed in novice think-aloud protocols can be interpreted as manifestations of this increased processing demand, reflecting moments in which working memory resources are overloaded or strategic uncertainty arises. Similar associations between gaze shifts, speech pauses, and cognitive effort have been reported in prior multimodal studies [50].

This discrepancy indicates that while their gaze behavior resembled that of experts, they likely lacked the necessary domain-specific knowledge. However, these individual differences underscore that graph literacy is better understood as a continuum rather than a binary classification [7], and they suggest that while interpretive strategies can be acquired through experience, deep conceptual understanding requires domain-specific expertise.

This supports the assumption that novices experience higher germane and extraneous load due to insufficient internalization of concepts, as they struggle to formulate coherent reasoning under pressure [21].

These differences in verbalization reflect varying degrees of representational competence [15,20], suggesting that experts not only perceive information more efficiently but also express it more conceptually. These expressions, combined with eye tracking and summarized performance

data, offer behavioral evidence of working memory overload as described by CLT [21]. Altogether, these patterns support the interpretation that longer overall processing times and irregular behaviors among novices are linked to uncertainty and deficits in strategic approach.

Many novices also used a reverse-matching strategy, described in protocols as “Let’s see if that fits.” suggesting they relied more on superficial visual matching than on conceptual understanding. This confirms the importance of prior knowledge and representational competence in reducing intrinsic and extraneous cognitive load [8,15].

This observation is consistent with findings by Oh *et al.* [50], whose results suggest that speech pauses often coincide with shifts in visual attention.

These findings also confirm earlier research that individuals with high graph literacy spend more time focusing on key features of a diagram, facilitating deeper conceptual integration [4]. Our findings extend previous work by Harsh *et al.* [60], who documented structured gaze patterns and precise verbalization among experts relative to novices.

RQ4: While recent AI models are increasingly capable of producing explicit reasoning steps in text-based contexts, this should be distinguished from internally grounded, perceptually driven reasoning. In the present study, AI-generated explanations were not reliably linked to the visual structure of the diagrams and often failed when relational or spatial interpretation was required.

By including AI models in this comparison, we highlight that, while AI can mimic certain expertlike reasoning patterns in text-based tasks, it lacks the schema-based visual processing and domain-specific strategies that characterize human expertise, particularly in complex diagram interpretation.

These findings align with prior research indicating that while AI models like ChatGPT-4 can mimic expertlike reasoning patterns in text-based tasks, they still lack the visual perception capabilities and schema-based interpretation strategies required for diagram analysis [55].

These results also highlight the limitations of both AI models in handling complex visual diagram interpretation tasks, with ChatGPT-4 demonstrating relatively better capabilities due to its advanced processing power. This underscores the gap in diagram interpretation abilities between the AI models and human groups, particularly in interpreting complex relationships within visual data. This contrast further emphasizes the uniquely human capacity for strategy use, emotional monitoring, and schema activation in visual diagram interpretation [15,54].

This strategy of focused visual attention helps explain why experts are able to interpret complex diagrams more quickly and accurately, while novices often struggle to identify relevant information. Additionally, they require less time overall to solve tasks compared to novices [59,63].

While the AI models themselves did not produce think-aloud data, the contrast between their performance and that

of human participants further underscores the role of structured reasoning and cognitive engagement. The AI models' failure, particularly with complex diagrams, reflects an absence of interpretive strategy, something clearly observable in the human experts' verbal protocols. The absence of emotional expressions in AI outputs should not be interpreted as a deficiency *per se* but rather as an indicator of missing metacognitive and affective monitoring processes. In human participants, emotional utterances such as expressions of uncertainty or frustration often coincided with speech pauses and gaze shifts, serving as observable cues of cognitive overload and strategy regulation. Current AI models do not exhibit such self-monitoring signals, which limits their ability to adapt reasoning strategies in response to task difficulty.

VI. IMPLICATION

The findings hold important implications for teaching and learning in physics. Many of the novice errors observed are directly tied to core physics concepts and highlight specific conceptual hurdles that require targeted instructional support. While novice errors are ultimately rooted in conceptual difficulties in physics, our results show that these difficulties manifest systematically in characteristic visual strategies. Specifically, novices' tendency toward reverse-matching, frequent gaze shifts between diagram and answer options, and reduced dwell time in task-relevant regions indicate an absence of stable conceptual schemas guiding visual attention. In this sense, visual strategies do not merely accompany conceptual errors but provide observable markers of how these errors emerge during diagram interpretation. From an instructional perspective, this suggests that analyzing students' visual attention can help instructors identify underlying conceptual misunderstandings even before an incorrect answer is produced. The frequent reliance on reverse-matching strategies suggests that students often approach physics diagrams as pattern-matching tasks rather than as representations of underlying principles, particularly in multiple-choice formats commonly used in assessments. From a science education perspective, this underscores the importance of explicitly teaching representational competence in physics, including how to extract meaning from diagrams, connect visual features with physical concepts, and evaluate the adequacy of different representational forms. Moreover, the more structured and selective strategies employed by experts point to the potential benefits of instructional approaches that make expertlike visual processing explicit to students, for example through eye-tracking-based demonstrations, scaffolded diagram interpretation tasks, or guided reflection on problem-solving strategies. Taken together, these results suggest that strengthening representational and metacognitive skills should be considered a central aim of physics education, alongside the development of conceptual and mathematical understanding.

The findings of this study offer valuable insights into the cognitive and perceptual differences between novices and experts in interpreting complex scientific diagrams. Experts consistently demonstrated higher success rates, more efficient use of time, and more structured visual attention patterns. These advantages can be attributed to their cognitive strategies, greater schema use, and refined graph literacy, skills that allow them to reduce cognitive load by focusing on relevant features and chunking information more effectively.

In contrast, novices exhibited longer overall processing times, more scattered scan paths, and frequent signs of disorientation, particularly when confronted with more challenging diagrams. These patterns may suggest increased cognitive demands, as reflected in their visual attention and verbal behaviors, though cognitive load was not directly measured in this study. These difficulties were further reflected in their verbal protocols, which revealed uncertainty, strategy deficits, and superficial reasoning. Such behaviors are indicative of increased intrinsic and extraneous load and underscore the critical role of prior knowledge and representational competence in diagram interpretation. The ability to interpret diagrams is a continuous skill spectrum, not a binary one. While experts generally demonstrated higher success rates and more efficient strategies, the study also shows that even experts exhibit novicelike behavior under certain conditions, particularly when diagram interpretation is not regularly practiced. This highlights the importance of continual skill development. Interestingly, all expert participants who exhibited novicelike gaze behavior also reported irregular engagement with diagram interpretation tasks, as indicated in the prestudy questionnaire. This finding aligns with our observation that even nominally expert participants may display less structured visual attention patterns when they have limited recent practice with diagram-related tasks.

The comparison with AI models such as ChatGPT-4 and DeepSeek-R1 further highlights the uniquely human components of visual reasoning. While ChatGPT-4 performed well on simple retrieval tasks, it lacked the ability to interpret complex visual relationships. DeepSeek-R1 demonstrated even more limited capabilities. These findings confirm previous research indicating that current AI systems, despite impressive text-processing skills, still struggle with visual interpretation due to the absence of perceptual grounding and schema-guided reasoning. These models still lag behind human experts in interpreting complex visual data, raising new research questions about how AI could evolve to enhance personal diagram interpretation skills in education.

The ability to interpret diagrams extends beyond the natural sciences, impacting fields like mathematics, finance, and medicine. This suggests that diagram interpretation should be integrated across disciplines. Taken together, these results emphasize the growing importance

of graph literacy in STEM education and support the need for targeted instructional interventions. Eye-tracking data reveal clear markers of expert performance, which can serve as practical teaching tools. For instance, instructors can provide step-by-step visualizations of expert scanning strategies to guide students' focus and reasoning, thereby addressing common cognitive challenges and fostering expertlike diagram interpretation skills.

This study also addresses a research gap by examining a broad range of diagrams, rather than focusing on specific diagram types or disciplines, thus expanding the scope of previous research. In addition, this research also calls attention to the limitations of AI in the visual domain. As reliance on multimodal data increases in science education and assessment, it becomes crucial to explore how AI tools can evolve to interpret not just language,

but visual representations as well. Future research could also explore how different disciplines, or a second major, influence the development of expertise in diagram interpretation. Additionally, investigating whether interest in fields like financial charts or medical statistics fosters diagram interpretation skills.

ACKNOWLEDGMENTS

The publication of this article was funded by the Open Access Fund of Leibniz Universität Hannover.

DATA AVAILABILITY

The data are not publicly available. The data are available from the authors upon reasonable request.

-
- [1] L. Hahn and P. Klein, Eye tracking in physics education research: A systematic literature review, *Phys. Rev. Phys. Educ. Res.* **18**, 013102 (2022).
 - [2] A. Gegenfurtner, E. Lehtinen, and R. Säljö, Expertise differences in the comprehension of visualizations: A meta-analysis of eye-tracking research in professional domains, *Educ. Psychol. Rev.* **23**, 523 (2011).
 - [3] S. Brams, G. Ziv, O. Levin, J. Spitz, J. Wagemans, A. M. Williams, and W. F. Helsen, The relationship between gaze behavior, expertise, and performance: A systematic review, *Psychol. Bull.* **145**, 980 (2019).
 - [4] Y. Okan, M. Galesic, and R. Garcia-Retamero, How people with low and high graph literacy process health graphs: Evidence from eye-tracking, *Behav. Decis. Making* **29**, 271 (2016).
 - [5] E. G. Freedman and P. Shah, Toward a model of knowledge-based graph comprehension, in *Diagrammatic Representation and Inference* edited by G. Goos, J. Hartmanis, J. van Leeuwen, M. Hegarty, B. Meyer, and N. H. Narayanan (Springer, Berlin, Heidelberg, 2002), pp. 18–30.
 - [6] P. Shah and E. G. Freedman, Bar and line graph comprehension: An interaction of top-down and bottom-up processes, *Top. Cognit. Sci.* **3**, 560 (2011).
 - [7] M. Galesic and R. Garcia-Retamero, Graph literacy: A cross-cultural comparison, *Med. Decis. Making* **31**, 444 (2011).
 - [8] S. Pinker and R. Feedle, Theory of graph comprehension, in *Artificial Intelligence and the Future of Testing*, edited by R. Freedle (Erlbaum, Hillsdale, NJ, 1990), pp. 73–126.
 - [9] R. M. Ratwani and J. G. Trafton, Shedding light on the graph schema: Perceptual features versus invariant structure, *Psychon. Bull. Rev.* **15**, 757 (2008).
 - [10] D. Rosengrant, C. Thomson, T. Mzoughi, M. Sabella, C. Henderson, and C. Singh, Comparing experts and novices in solving electrical circuit problems with the help of eye-tracking, *AIP Conf. Proc.* **1179**, 249 (2009).
 - [11] J. Sweller, Element interactivity and intrinsic, extraneous, and germane cognitive load, *Educ. Psychol. Rev.* **22**, 123 (2010).
 - [12] L. C. McDermott, M. L. Rosenquist, and E. H. van Zee, Student difficulties in connecting graphs and physics: Examples from kinematics, *Am. J. Phys.* **55**, 503 (1987).
 - [13] A. Susac, A. Bubic, P. Martinjak, M. Planinic, and M. Palmovic, Graphical representations of data improve student understanding of measurement and uncertainty: An eye-tracking study, *Phys. Rev. Phys. Educ. Res.* **13**, 020125 (2017).
 - [14] N. Glazer, Challenges with graph interpretation: A review of the literature, *Stud. Sci. Educ.* **47**, 183 (2011).
 - [15] P. Shah and J. Hoeffner, Review of graph comprehension research: Implications for instruction, *Educ. Psychol. Rev.* **14**, 47 (2002).
 - [16] S. Ainsworth, The functions of multiple representations, *Comput. Educ.* **33**, 131 (1999).
 - [17] J. H. Larkin and H. A. Simon, Why a diagram is (sometimes) worth ten thousand words, *Cogn. Sci.* **11**, 65 (1987).
 - [18] C. F. J. Pols, P. J. J. M. Dekkers, and M. J. de Vries, What do they know? Investigating students' ability to analyse experimental data in secondary physics education, *Int. J. Sci. Educ.* **43**, 274 (2021).
 - [19] M. A. Rau, Conditions for the effectiveness of multiple visual representations in enhancing STEM learning, *Educ. Psychol. Rev.* **29**, 717 (2017).
 - [20] D. Carraher and A. Schliemann, The transfer dilemma, *J. Learn. Sci.* **11**, 1 (2002).
 - [21] J. Sweller, Cognitive load during problem solving: Effects on learning, *Cogn. Sci.* **12**, 257 (1988).
 - [22] P. Klein, A. Lichtenberger, S. Küchemann, S. Becker, M. Kekule, J. Viiri, C. Baadte, A. Vaterlaus, and J. Kuhn, Visual attention while solving the test of understanding graphs in kinematics: An eye-tracking analysis, *Eur. J. Phys.* **41**, 025701 (2020).

- [23] L. Hahn and P. Klein, Clustering eye-movement data uncovers students' strategies for coordinating equations and diagrams of vector fields, *Educ. Stud. Math.* **118**, 359 (2025).
- [24] C. Wang and B. Wei, Analysis of visual-based physics questions of the senior high school entrance examination in China, *Phys. Rev. Phys. Educ. Res.* **20**, 010112 (2024).
- [25] A. Bandura, *Self-Efficacy: The Exercise of Control*, edited by S. F. Brennan and C. Hastings (W. H. Freeman and Company, New York, 1997).
- [26] J. H. Flavell, Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry, *Am. Psychol.* **34**, 906 (1979).
- [27] T. de Jong, Cognitive load theory, educational research, and instructional design: Some food for thought, *Instr. Sci.* **38**, 105 (2010).
- [28] M. Chi, R. Glaser, and E. Rees, Expertise in problem solving, in *Advances in the Psychology of Human Intelligence*, edited by R. J. Sternberg (Erlbaum, Hillsdale, NJ, 1982), Vol. 1, pp. 7–75.
- [29] F. Gobet, P. C. Lane, S. Croker, P. C.-H. Cheng, G. Jones, I. Oliver, and J. M. Pine, Chunking mechanisms in human learning, *Trends Cognit. Sci.* **5**, 236 (2001).
- [30] H. Haider and P. A. Frensch, The role of information reduction in skill acquisition, *Cogn. Psychol.* **30**, 304 (1996).
- [31] H. Haider and P. A. Frensch, Eye movement during skill acquisition: More evidence for the information-reduction hypothesis, *J. Exp. Psychol.* **25**, 172 (1999).
- [32] K. Stofer and X. Che, Comparing experts and novices on scaffolded data visualizations using eye-tracking, *J. Eye. Mov. Res.* **7**, 1 (2015).
- [33] A. Gegenfurtner, Cognitive theory of visual expertise, edited by A. Gegenfurtner and R. Stahnke, *Teacher Professional Vision: Theoretical and Methodological Advances* (Routledge, London, 2024), pp. 71–85.
- [34] A. Susac, A. Bubic, E. Kazotti, M. Planinic, and M. Palmovic, Student understanding of graph slope and area under a graph: A comparison of physics and nonphysics students, *Phys. Rev. Phys. Educ. Res.* **14**, 020109 (2018).
- [35] P. Klein, J. Viiri, S. Mozaffari, A. Dengel, and J. Kuhn, Instruction-based clinical eye-tracking study on the visual interpretation of divergence: How do students look at vector field plots?, *Phys. Rev. Phys. Educ. Res.* **14**, 010116 (2018).
- [36] P. Klein, J. Viiri, and J. Kuhn, Visual cues improve students' understanding of divergence and curl: Evidence from eye movements during reading and problem solving, *Phys. Rev. Phys. Educ. Res.* **15**, 010126 (2019).
- [37] V. Ruf, Y. Dinc, S. Küchemann, M. Berndt, S. Steinert, D. Kugelman, J. Bortfeldt, J. Schreiber, M. R. Fischer, and J. Kuhn, Comparison and AI-based prediction of graph comprehension skills based on the visual strategies of first-year physics and medicine students, *Phys. Rev. Phys. Educ. Res.* **20**, 020138 (2024).
- [38] D. Maloney, An overview of physics education research on problem solving, in *Getting Started in PER*, edited by C. Henderson and K. Harper (American Association of Physics Teachers, College Park, 2011), Vol. 2.
- [39] L. Hsu, E. Brewster, T. M. Foster, and K. A. Harper, Resource letter RPS-1: Research in problem solving, *Am. J. Phys.* **72**, 1147 (2004).
- [40] J. L. Docktor and J. P. Mestre, Synthesis of discipline-based education research in physics, *Phys. Rev. ST Phys. Educ. Res.* **10**, 020119 (2014).
- [41] B.-S. Eylon and F. Reif, Effects of knowledge organization on task performance, *Cognit. Instr.* **1**, 5 (1984).
- [42] J. H. Larkin and F. Reif, Understanding and teaching problem-solving in physics, *Eur. J. Sci. Educ.* **1**, 191 (1979).
- [43] J. Larkin, J. McDermott, D. P. Simon, and H. A. Simon, Expert and novice performance in solving physics problems, *Science (New York, N.Y.)* **208**, 1335 (1980).
- [44] F. Reif and J. I. Heller, Knowledge structure and problem solving in physics, *Educ. Psychol.* **17**, 102 (1982).
- [45] E. Bagno and B.-S. Eylon, From problem solving to a knowledge structure: An example from the domain of electromagnetism, *Am. J. Phys.* **65**, 726 (1997).
- [46] R. Elio and P. B. Scharf, Modeling novice-to-expert shifts in problem-solving strategy and knowledge organization, *Cogn. Sci.* **14**, 579 (1990).
- [47] K. Cummings, J. Marx, R. Thornton, and D. Kuhl, Evaluating innovation in studio physics, *Am. J. Phys.* **67**, S38 (1999).
- [48] J. L. Docktor, N. E. Strand, J. P. Mestre, and B. H. Ross, Conceptual problem solving in high school physics, *Phys. Rev. ST Phys. Educ. Res.* **11** (2015).
- [49] P. Heller and M. Hollabaugh, Teaching problem solving through cooperative grouping. Part 2: Designing problems and structuring groups, *Am. J. Phys.* **60**, 637 (1992).
- [50] K. Oh, J. T. Almarode, and R. H. Tai, An exploration of think-aloud protocols linked with eye-gaze tracking: Are they talking about what they are looking at, *Procedia* **93**, 184 (2013).
- [51] J. Halszka, K. Holmqvist, and H. Gruber, Eye tracking in educational science: Theoretical frameworks and research agendas, *J. Eye Mov. Res.* **10**, 1 (2017).
- [52] C. Hoyer and R. Girwidz, Animation and interactivity in computer-based physics experiments to support the documentation of measured vector quantities in diagrams: An eye tracking study, *Phys. Rev. Phys. Educ. Res.* **16**, 020124 (2020).
- [53] A. Susac, A. Bubic, M. Planinic, M. Movre, and M. Palmovic, Role of diagrams in problem solving: An evaluation of eye-tracking parameters as a measure of visual attention, *Phys. Rev. Phys. Educ. Res.* **15**, 013101 (2019).
- [54] P. Klein, S. Küchemann, S. Brückner, O. Zlatkin-Troitschanskaia, and J. Kuhn, Student understanding of graph slope and area under a curve: A replication study comparing first-year physics and economics students, *Phys. Rev. Phys. Educ. Res.* **15**, 020116 (2019).
- [55] G. Polverini and B. Gregoric, Performance of ChatGPT on the test of understanding graphs in kinematics, *Phys. Rev. Phys. Educ. Res.* **20**, 010109 (2024).
- [56] S. Aldazharova, G. Issayeva, S. Maxutov, and N. Balta, Assessing AI's problem solving in physics: Analyzing reasoning, false positives and negatives through the force concept inventory, *Contemp. Educ. Technol.* **16**, ep538 (2024).
- [57] G. Polverini, J. Melin, E. Önerud, and B. Gregoric, Performance of ChatGPT on tasks involving physics visual representations: The case of the brief electricity and

- magnetism assessment, *Phys. Rev. Phys. Educ. Res.* **21**, 010154 (2025).
- [58] S. Küchemann, S. Malone, P. Edelsbrunner, A. Lichtenberger, E. Stern, R. Schumacher, R. Brünken, A. Vaterlaus, and J. Kuhn, Inventory for the assessment of representational competence of vector fields, *Phys. Rev. Phys. Educ. Res.* **17**, 020126 (2021).
- [59] A. M. Madsen, A. M. Larson, L. C. Loschky, and N. S. Rebello, Differences in visual attention between those who correctly and incorrectly answer physics problems, *Phys. Rev. ST Phys. Educ. Res.* **8**, 010122 (2012).
- [60] J. A. Harsh, M. Campillo, C. Murray, C. Myers, J. Nguyen, and A. V. Maltese, “Seeing” data like an expert: An eye-tracking study using graphical data representations, *CBE Life Sci. Educ.* **18**, ar32 (2019).
- [61] See Supplemental Material at <http://link.aps.org/supplemental/10.1103/58hn-493c> for representative diagrams from the study.
- [62] A. R. Strohmaier, K. J. MacKay, A. Obersteiner, and K. M. Reiss, Eye-tracking methodology in mathematics education research: A systematic literature review, *Educ. Stud. Math.* **104**, 147 (2020).
- [63] P. Klein, A. Müller, and J. Kuhn, Assessment of representational competence in kinematics, *Phys. Rev. Phys. Educ. Res.* **13**, 010132 (2017).