GRAPHIONALE: How Graph Visualizations of LLM Rationales Affect Human Decision Making

Xinru Wang*
Singapore-MIT Alliance for Research and Technology
Singapore
xinru.wang@smart.mit.edu

Zhexuan Ma*
National University of Singapore
Singapore
e1499160@u.nus.edu

Ming Yin
Purdue University
West Lafayette, Indiana, USA
mingyin@purdue.edu

Shuai Ma†
Institute of Software, Chinese Academy of Sciences
Beijing, China
mashuai@iscas.ac.cn

Thomas W Malone
Massachusetts Institute of Technology
Cambridge, Massachusetts, USA
malone@mit.edu

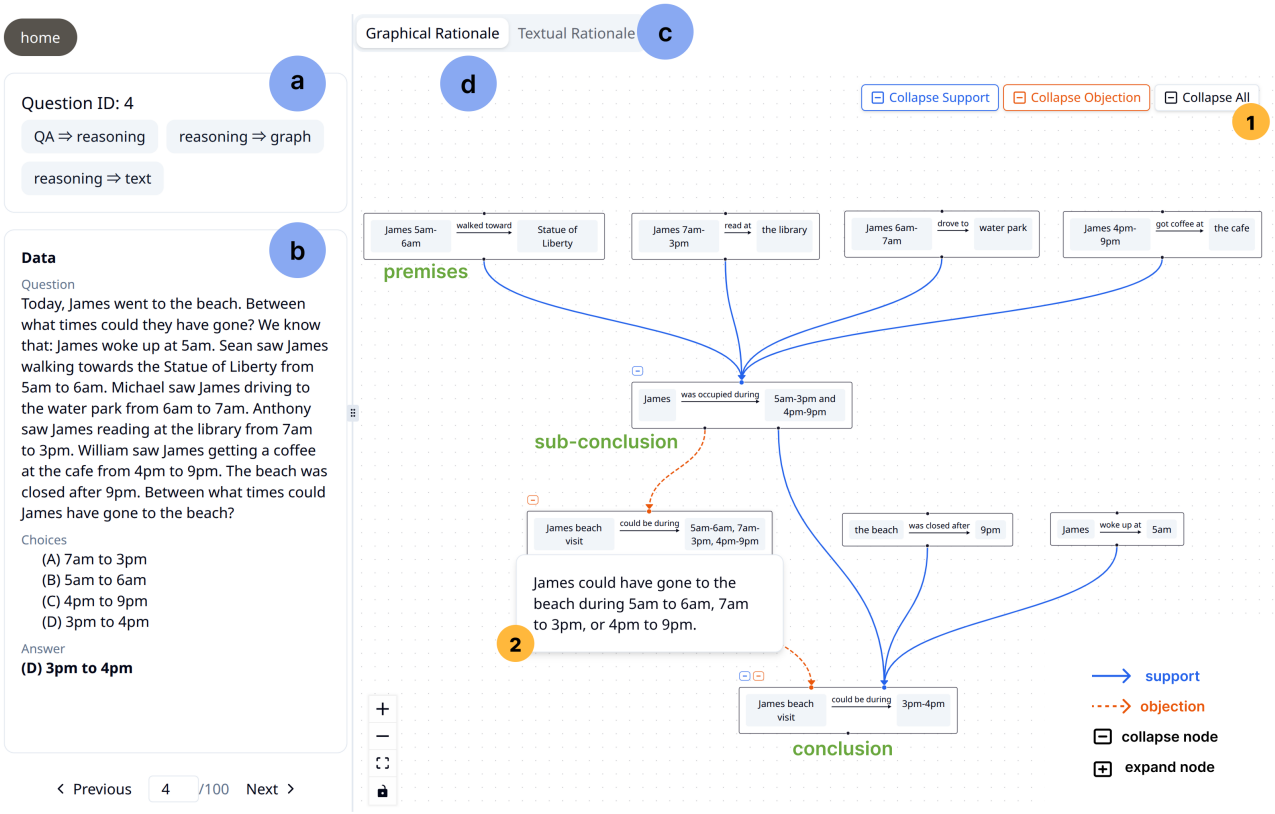


Figure 1: GRAPHIONALE, a co-design-informed prototype that transforms LLM rationales into structured argument graphs with compressed node representations. The interface supports textual and graphical rationale generation (a), task question display (b), switching between views (c), and an interactive graph panel (d). Users can selectively expand or collapse reasoning branches through global controls (1) and inspect detailed explanations by hovering over nodes (2).

*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/10.1145/3830398.3830638

Abstract

Large Language Models (LLMs) are increasingly equipped with augmented reasoning capabilities to generate rationales that support human decision-making. Yet these text-dense rationales often impose substantial cognitive burdens. Building on a formative co-design study that identified user preferences for non-linear reasoning representations, we developed Graphionale as a testbed for empirically studying argument-map-style rationale visualization. This system transforms linear LLM rationales into interactive, multi-level graphs. It explicitly structures logical relationships (e.g., conclusions, premises, support, and objections), while further extracting entities and relations within each statement to construct condensed node-link representations. We conduct a large-scale online user study ($N = 204$) to examine when graphical rationales are more effective than textual ones, across varying task modality (verbal vs. visual reasoning), rationale format (textual vs. graphical), and question difficulty (easy vs. hard). Our results show that graphical rationales do not help uniformly: they improve trust calibration for verbal reasoning yet feel more cognitively demanding and less satisfying; for visual reasoning, they impair calibration yet feel more engaging and helpful. In each modality, the format that better supports calibrated decisions is not the one users prefer, highlighting that matching rationale format to task modality is key to effective AI explanation design. Our findings contribute empirical design knowledge about when and how graphical rationales support human decision making, and inform the next-generation reasoning-aware AI interfaces.

CCS Concepts

• **Human-centered computing** → **Empirical studies in HCI**; **Interactive systems and tools**.

Keywords

LLM rationale, argument map, user study, explainable AI, interaction design

ACM Reference Format:

Xinru Wang, Zhexuan Ma, Ming Yin, Shuai Ma, and Thomas W Malone. 2026. Graphionale: How Graph Visualizations of LLM Rationales Affect Human Decision Making. 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, 21 pages. https://doi.org/10.1145/3830398.3830638

1 Introduction

Recent advances in Large Language Models (LLMs) have significantly improved their reasoning capabilities. With these developments, LLMs are increasingly used as conversational assistants for complex, knowledge-intensive questions. To improve transparency and support human decision making, LLMs provide detailed rationales describing how an answer is derived [4, 34, 62, 65, 66]. However, these rationales are traditionally presented as long blocks of text. Even when organized into sections, text remains inherently linear and cannot intuitively represent relational structures among reasoning steps, such as branching, aggregation, or elimination. As a result, users must parse sequential sentences to reconstruct the underlying logic, making it difficult to quickly grasp the overall reasoning or evaluate the model's conclusions [23, 37].

In contrast, humans often externalize complex explanations into diagrams to support understanding. Students draw concept maps to visualize relationships between ideas [35, 44]. In deliberation settings, discussions are represented as argument maps to help people navigate competing viewpoints [10, 24]. Inspired by these practices, representing LLM rationales graphically—by making logical relationships explicit and visually structured—may better align with human reasoning processes and support users in understanding and critically evaluating model outputs. However, graphical rationales may also make flawed or unfaithful LLM reasoning appear more credible than it is [32]. Whether they help users identify such problems or instead inflate trust remains an empirical question.

Several recent systems have begun exploring graphical interfaces for LLM outputs, including tools for prompt engineering [2], interactive diagrams of LLM responses [23, 47], and reasoning trace visualizations [27, 37]. However, the design principles for graphical LLM rationales remain underexplored, and little empirical evidence exists on whether, when, and how such representations benefit users.

To address this gap, we investigate two research questions:

- **RQ1:** How should graphical representations of LLM rationales be designed to align with user needs?
- **RQ2:** Compared with textual rationales, how do graphical LLM rationales influence users' interaction experience and decision-making outcomes?

To answer RQ1, we conducted a formative co-design study in which participants interacted with a low-fidelity interface. Through sketching and discussion, we examined how users expect LLM rationales to be graphically represented. Guided by argument mapping theory [1] and insights from this study, we developed Graphionale[1], an LLM-specific instantiation of argument-map-style rationale visualization that generates glanceable graphical representations of LLM rationales by (1) structuring reasoning as argument maps and (2) compressing node-level text to highlight essential information [23, 60]. Graphionale serves as a co-design-informed prototype for the empirical comparison that follows.

To answer RQ2, we conducted a controlled online user study ($N = 204$) comparing graphical and textual LLM rationales across different decision-making contexts. We varied task modality (verbal vs. visual reasoning) and question difficulty (easy vs. hard) to examine how these factors influence the effectiveness of graphical rationales. Our results show that graphical rationales do not help uniformly: contrary to intuitive expectations, they improve trust calibration for verbal reasoning yet are rated as more cognitively demanding and less satisfying; for visual reasoning, graphical rationales impair calibration yet are rated as more engaging and helpful. In each modality, the format that supports better-calibrated decisions is thus not the one users prefer. This objective–subjective dissociation suggests that perceived fluency does not reliably predict calibration quality. Furthermore, the calibration benefit in verbal reasoning is amplified on harder questions and among users higher in analytical thinking.

In summary, this paper makes the following contributions:

[1] https://graphionale.vercel.app/

- A formative co-design study grounded in argument mapping that identifies key design principles for presenting LLM rationales graphically.
- Graphionale, a co-design-informed prototype that instantiates argument-map-style visualization for LLM rationales by combining structured argument mapping with text condensation, serving as the vehicle for our empirical comparison.
- A large-sample controlled user study that systematically evaluates, compared with traditional textual rationales, how graphical LLM rationales influence user perception and behavior across task modalities, contributing empirical design knowledge about when and how graphical rationales support human decision making.

2 Related Work

2.1 Graph-driven Explainable AI and Reasoning in LLMs

Graph structures—particularly Knowledge Graphs (KGs)—have long been used to provide transparency in AI systems. Research has shown that inferring preference paths over user–item–entity graphs can generate human-readable reasoning chains [14, 46], and KGs have been applied for causal and counterfactual reasoning [21, 41].

In the context of LLMs, the focus has evolved toward multi-hop explanation graphs. Datasets such as WorldTree [22] formalize rationales as lexically connected facts. To augment LLM reasoning, researchers have moved beyond linear Chain-of-Thought prompting [62] toward structured approaches such as Tree-of-Thought [65] and Graph-of-Thought [4], which allow models to explore multiple reasoning branches. Recent test-time scaling methods [34] further enhance reasoning performance.

Despite these advances, a critical gap persists at the human–AI *interface*: rationales are typically presented as dense text, which obscures their multi-hop structure and limits users' ability to efficiently inspect and evaluate the reasoning process.

2.2 Graph Structures in Human Mental Models

Human cognition is inherently graph-structured. Research suggests that humans represent knowledge as associative networks in which sensory, spatial, and semantic nodes are multiply connected. The ability to infer and navigate these relationships is fundamental to forward planning and logical reasoning [42].

Humans naturally organize semantic knowledge using graph-like structures [9]. Representations such as mind maps and concept maps leverage perceptual strengths to manage cognitive load, for example by exploiting Gestalt principles such as proximity and continuity [38]. Visual working memory is typically limited to roughly 4–7 chunks [61], motivating design strategies that prioritize task-relevant nodes while compacting less relevant ones. Effective conceptual mapping also relies on hierarchical layering to structure complex information [35, 44].

If knowledge provides the cognitive space, reasoning can be understood as tracing paths through it. Deductive, inductive, and abductive reasoning correspond to different ways of selecting and chaining relations. Argument maps provide a structural representation of these inference processes by explicitly linking premises, evidence, and conclusions [1]. By organizing arguments into explicit structures, they help clarify complex reasoning, identify logical weaknesses, and support critical evaluation and collaborative discussion.

Early systems such as gIBIS [10] and platforms like the Deliberatorium [24] externalize reasoning as navigable graphs, enabling users to synthesize complex information more effectively than through text-based discourse alone and supporting large-scale deliberation and collective intelligence.

2.3 Design and Evaluation of Graphical XAI and LLM Interfaces

Researchers have explored graph-based visualizations to support AI explainability in ways that align with human mental models, typically representing explanations as structured relationships among concepts or decision provenance. For example, ConceptExplainer [19] uses concept graphs to visualize relationships between learned concepts and model predictions. Other work [20, 29] leverages provenance graphs to capture security-aware meta-information for explaining AI systems.

Prior work has also examined how such visualizations affect human understanding. Delarue et al. [11] found that graph-based explanations were perceived as more usable and intuitive than feature-importance explanations in recommender systems, although textual explanations sometimes led to higher objective understanding. This highlights the importance of carefully designing graph representations. Montagna et al. [33] further emphasize the need for rigorous, metric-driven evaluation of graph-based explanation quality (e.g., [13, 28, 52, 55–58]), rather than relying solely on qualitative case studies.

Looking at graph representations of LLM-generated text, one line of work focuses on extracting causal structures. Yang et al. [64] review methods for identifying cause–effect relationships, while recent work explores using LLMs to discover causal structures [53] and generate concept maps [39]. Systems such as PaperTrail [30] extract claim–evidence structures from research papers and LLM responses, representing them as provenance graphs for scholarly question answering.

Another line of research explores graphical interfaces for interacting with LLM outputs. Graphologue [23] converts responses into interactive diagrams for non-linear exploration, while Sensecape [47] supports multilevel organization for navigating complex topics. Tools such as ChainForge [2] provide visual environments for prompt engineering and hypothesis testing. These systems primarily facilitate LLM output exploration and content organization.

In contrast, visualizing LLM rationales requires representing implicit reasoning structures—such as dependencies among steps and how conclusions are derived—in a way that is explicit and interpretable. Only recently have efforts begun to address this. HaLLMark [17] visualizes authorship provenance, while Wang et al. [60] improve glanceability by structuring text into taxonomy components without modeling reasoning as graphs. ReasonGraph [27] and HIPPO [37] visualize reasoning traces as graphs or trees with interactive features. However, these systems lack clear design principles for making graphical rationales glanceable and interpretable, and their evaluations are small-scale usability studies that rarely

consider factors such as user reliance on AI or decision-making task characteristics.

Research Gap. While prior work demonstrates the potential of graph representations for explaining AI systems and exploring LLM outputs, there is limited empirical evidence on how graphical representations of LLM *rationales* can be effectively designed, and how they affect human understanding, decision-making, and human–AI collaboration.

3 Formative Study: A Co-Design Exploration

3.1 Participants and Study Procedure

To explore what forms of graphical rationales users prefer, and to inform the design of a strong argument-map-style representation for our empirical comparison, we conducted a formative study with co-design activities. We recruited 10 participants from our institution (5 female, 5 male; mean age 30), including 3 graduate students, 6 research scientists, and 1 software engineer. All were familiar with LLM applications and used them daily.

We built a low-fidelity interface that allowed participants to interact with both textual and graphical LLM rationales. We selected XplainLLM [8], in which each instance includes explanatory text describing LLM reasoning on CommonsenseQA [50]. Using GPT-5.2 with a prompt adapted from [15] (see Appendix A.1), we converted each explanation into a graphical representation where nodes represent statements (premises, intermediate conclusions, final conclusion) and directed edges represent inference relationships. A screenshot of the interface is provided in Appendix A.2.

During the study, participants completed two randomly selected CommonsenseQA questions. They first reviewed the textual rationale and described any difficulties, then switched to the graphical rationale and compared the two formats. Participants proposed improvements and sketched how they would like the rationale represented graphically. Each session lasted around 25 minutes ($6 compensation). All sessions were recorded for subsequent analysis.

3.2 Analysis and Results

Recorded sessions were transcribed and analyzed using thematic analysis [5]. Two authors iteratively discussed and refined a shared codebook. The analysis revealed four themes describing how users believe graphical representations of LLM rationales should be improved.

Theme 1: Glanceable nodes require aggressive text compression. **Challenge (C1).** Participants reported that many nodes felt like *"reading the text again,"* as sentence-level content prevented quick scanning. As P1 noted, *"If this block of text could be split into segments and listed as items, it would be easier."* When diagrams became overloaded, participants preferred falling back to structured text. As P10 commented, *"It didn't really help me solve the problem. I still had to read a lot of text."* Participants suggested replacing full sentences with keywords or short phrases and adjusting node granularity.

Design Goal (DG1). Enable fast scanning through compressed keyword-based nodes, while retaining structured text as a fallback when diagrams do not improve clarity.

Theme 2: Explanation maps should align with human decision logic. **Challenge (C2).** Participants struggled when the map's reasoning direction conflicted with common reading habits. They expected explanations to support elimination and comparison across answer options. As P2 described, *"I would first eliminate obviously wrong options, then compare the remaining ones."* Some rationales contained irrelevant terms not present in the answer options. P8 reacted: *"This looks like a hallucination… these things are not even in the answer options."* Participants proposed restructuring explanations around a decision template: criteria → comparison → elimination → summary (example sketches are provided in Appendix A.3), and adapting depth to task difficulty.

Design Goal (DG2). Make explanation maps function as decision-support artifacts that mirror human reasoning and option elimination, without exposing unnecessary model internals.

Theme 3: Visual semantics are essential for fast interpretation. **Challenge (C3).** Participants had difficulty interpreting relationships when visual semantics were unclear. P2 noted, *"This graph is messy… the lines are confusing, and I can't tell what the relationships are."* Participants wanted to quickly distinguish supporting and rejecting evidence and understand reasoning strength without reading every node. They proposed adding colors, icons, or symbols and clarifying edge semantics with labels and grouping structures.

Design Goal (DG3). Enable fast relational scanning so that support, rejection, grouping, and reasoning strength are visually interpretable at a glance.

Theme 4: Complexity must be bounded through constraints and progressive disclosure. **Challenge (C4).** Participants found diagrams overwhelming when too many nodes were displayed. P4 reacted: *"This is way too much… how did it generate so many nodes?"* Navigation costs were high for large diagrams. P3 noted, *"It's so long that I have to keep dragging around, which is exhausting."* Participants suggested merging redundant nodes, imposing limits on depth, and using progressive disclosure so that details appear only on demand. P1 proposed keeping the structure within *"about three to four layers, with three to five nodes per layer."*

Design Goal (DG4). Keep explanations bounded, navigable, and user-controllable so that diagrams remain glanceable rather than overwhelming.

4 Graphionale Interface

4.1 System Walkthrough

We instantiate these design goals in Graphionale, as illustrated in the following scenario and in Figure 1.

Alex is solving a logic puzzle with the help of a chatbot. The task involves determining when a person could have gone to the park based on a series of time-stamped activities (e.g., walking toward the Statue of Liberty from 5am to 6am, driving to the water park from 6am to 7am). In a traditional interface, the rationale appears as multiple paragraphs that Alex must read through and mentally reconstruct.

With Graphionale, the rationale is presented as a graphical structure that explicitly organizes statements and their relationships. The graph aligns with human decision logic: constraints

are first aggregated, then used to eliminate candidate time ranges, leading to a single valid interval (**DG2**). The graph adapts its complexity: high-level constraints are shown first, while details can be expanded on demand (**DG4**).

Each node represents a compressed, keyword-based statement for quick scanning (**DG1**). Within each node, the statement is structured using subject–predicate–object triplets, e.g., "5am–6am" $\xrightarrow{\text{walking toward}}$ "Statue of Liberty". Edges visually encode relationships between statements such as support and objection (**DG3**).

4.2 Graphical Rationale Generation

Given a reasoning question and an answer, we decompose graphical rationale generation into two steps: (1) constructing an argument map, and (2) compressing statement text within each node.

4.2.1 Argument Map Construction. Our formative study suggests that people tend to organize premises into intermediate sub-conclusions, progressively derive a final conclusion, and eliminate conflicting alternatives. This reasoning pattern closely matches argument maps [1], which have long been used in education, deliberation, and critical thinking.

Therefore, we adopt argument map principles to guide LLMs in generating structured graphical rationales (**DG2**). Specifically, we represent premises, sub-conclusions, and the final conclusion as nodes, and logical relationships as edges, including support (reasons in favor) and objection (reasons against). The depth of sub-conclusions scales with reasoning complexity (**DG2**). Edge color and line style distinguish support and objection relations (**DG3**). We constrain the number of nodes and incorporate progressive disclosure for users to expand the graph to reveal deeper reasoning layers or inspect supporting and opposing arguments on demand (**DG4**).

4.2.2 Statement Text Compression. To further improve glanceability, we compress the textual content within each node. Specifically, we distill each statement into an entity–relation–entity representation (**DG1**). The compressed representation captures essential semantic components while reducing verbosity, and is presented in a structured node-link format within each node (**DG3**). We impose length constraints for conciseness. As a fallback, users can view the full text on demand. Prompts for argument map construction and statement compression are provided in Appendix B.1.

4.3 Implementation

Graphionale is a full stack web application implemented with Next.js. The front-end renders questions and LLM rationales, while the backend manages task data retrieval, generation of graphical rationales, derivation of textual and graphical rationales, and persistence of task data and generated rationales in PostgreSQL.

Rationale generation uses GPT-5.4 with high reasoning effort. The back-end first constructs a structured argument map, then derives both textual and graphical rationales from it. This keeps the two views aligned while supporting different presentation forms.

The graphical view is built on XYflow (React Flow) [51], which renders graphical rationales as interactive diagrams with zooming and viewport navigation. We implemented controls for selectively expanding and collapsing support and objection branches, either for individual nodes or globally. By default, each node displays a compact subject–predicate–object triplet when suitable or a short label otherwise. Hovering over a node reveals the full text explanation. An overview of the system architecture is provided in Appendix B.2.

5 Technical Evaluation

To assess the quality of LLM-generated graphical rationales, we conducted a technical evaluation focusing on the correctness of both graph structure and reasoning content.

5.1 Setup

5.1.1 Dataset Selection. We varied task modality (verbal vs. visual reasoning) and question difficulty to improve generalizability, as further explained in Section 6.1.

For verbal reasoning, we selected BIG-Bench Hard (BBH) [48], a suite of 23 challenging tasks [3] that requires multi-step reasoning. The benchmark reports average human-rater performance, allowing us to select tasks with different difficulty levels. We focused on tasks involving verbal reasoning that remain accessible to lay participants: one easy task with high human performance ("Temporal Sequences") and two hard tasks with low human performance ("Logical Deduction (7 Objects)" and "Tracking Shuffled Objects (7 Objects)").

For visual reasoning, we selected I-RAVEN [18, 67], which improves upon RAVEN [68] for fairer evaluation on Raven's Progressive Matrices. Among its seven figure configurations, we selected "L-R" (Left-Right) as the easy task, and "3×3Grid" and "Out-InGrid" as the hard tasks, since configurations with more components require more complex reasoning. Because I-RAVEN is visual, we prompted the LLM to generate step-by-step graphical logic patterns in SVG-based representations. Example questions of BBH and I-RAVEN are provided in Appendix C.

For each selected question category, we randomly sampled 30 questions. For 20 questions, we generated graphical LLM rationales for the correct answer; for the remaining 10, we generated rationales for a randomly selected incorrect answer.

5.1.2 Evaluation Criteria. Two coders evaluated the generated graphical rationales on three aspects:

- **Graph generation quality:** Whether generated graphs followed structural constraints [27, 35, 38, 44], including depth scaling with difficulty, node count limits, text length per node, and consistency between compressed nodes and full textual rationales.
- **Semantic reasoning quality:** Whether reasoning paths (supporting or rejecting relationships) were logically correct and coherent, and whether node content was valid and interpretable [27].
- **Functional correctness for error detection:** Whether graphical rationales could help users detect reasoning errors [16]—correct answers should produce reasonable rationales, while incorrect answers should reflect flawed or inconsistent reasoning structures.

5.2 Findings

Coders recorded issues or sub-optimal patterns in generated rationales. Based on these observations, we conducted an iterative prompt refinement process. All findings below are based on the final refined generation. Inter-rater reliability was high (Cohen's $\kappa = 0.94$).

Graph generation quality was high. Across 180 rationales (30 questions × 6 task categories), only 5 structural issues (2.8%) were identified, all involving excessive edges in BBH Logical Deduction questions. No violations were observed in node scaling, node count, text length, or consistency between compressed nodes and full text.

Correct-answer rationales had two recurring sub-optimal patterns. (1) *Redundant reasoning steps* (15/120, 13%): unnecessary intermediate nodes that added clutter, particularly in I-RAVEN tasks. (2) *Unverifiable reasoning* (10/120, 8%): nodes containing claims about subtle visual attributes that could not be reliably verified from the image.

Three error patterns characterized incorrect-answer rationales. (1) *Missing evidence* (10/60, 17%): reasoning chains omitting critical constraints. (2) *Problematic evidence* (26/60, 43%): hallucinated or factually incorrect content, especially in I-RAVEN. (3) *Incorrect (sub-)conclusions* (30/60, 50%): valid evidence with incorrect inferences, most prominent in BBH. All three patterns were almost entirely absent in correct-answer rationales, confirming their diagnostic value as indicators of LLM reasoning failure.

6 User Study Design

We conducted a user study to compare the effects of graphical and textual LLM rationales on human decision making across tasks. Specifically, we investigate:

- **Q1:** How does representing LLM rationales as graphs affect users' interaction experience?
- **Q2:** How does graphical representation affect users' decision-making outcomes?
- **Q3:** Which design features of graphical rationales are helpful, and how can they be improved?

6.1 Conditions

We compared graphical LLM rationales against a textual baseline. Both conditions were derived from the same underlying argument map. The textual rationale contained the same information as the full-text content in the graphical rationale's nodes, so the comparison isolates presentation format rather than reasoning content. Headings and bullet points were added so that the textual baseline presented information in a similarly organized form and resembled common LLM chat responses, which are now often organized into subsections. Prompts for textual rationale generation are in Appendix D.1.

Beyond rationale format (graphical vs. textual), we varied task modality and question difficulty to account for different decision-making contexts [37]. We conjectured that the effectiveness of graphical rationales may depend on whether the underlying reasoning is primarily verbal or visual. In verbal reasoning tasks, graphical representations can externalize relational structures that would otherwise require parsing long text. In visual reasoning tasks, which already involve visual patterns, graphical explanations may provide less additional benefit. We also varied question difficulty, as prior work [52] suggests users rely more on AI explanations when tasks are sufficiently challenging. We selected BIG-Bench Hard (BBH) [48] as the verbal reasoning task and I-RAVEN [18, 67] as the visual reasoning task. Both datasets report human performance across subcategories, allowing us to select both easy and hard questions. Datasets are described in Section 5.1.1.

Overall, we employed a 2 × 2 between-subjects design with (1) rationale format (graphical vs. textual) and (2) task modality (verbal (BBH) vs. visual reasoning (I-RAVEN)). Within each condition, participants completed both easy and hard questions. Screenshots of the four conditions are shown in Figure 2.

6.2 Participants

We conducted a power analysis using G*Power [12] ($f = 0.25$, $\alpha = 0.05$, power = 0.9), resulting in a required sample size of 171 participants across four conditions. After IRB approval, we recruited participants from Prolific using standard quality filters (U.S.-based, ⩾99% approval rate with ⩾1,000 submissions, native English speakers, desktop users). Participants received $4.50 base compensation, with a performance-based bonus of up to $4.00 if they achieved an overall accuracy above 70% ($0.20 per correct answer; $0.70 for two designated trials in which the AI provided incorrect answers).

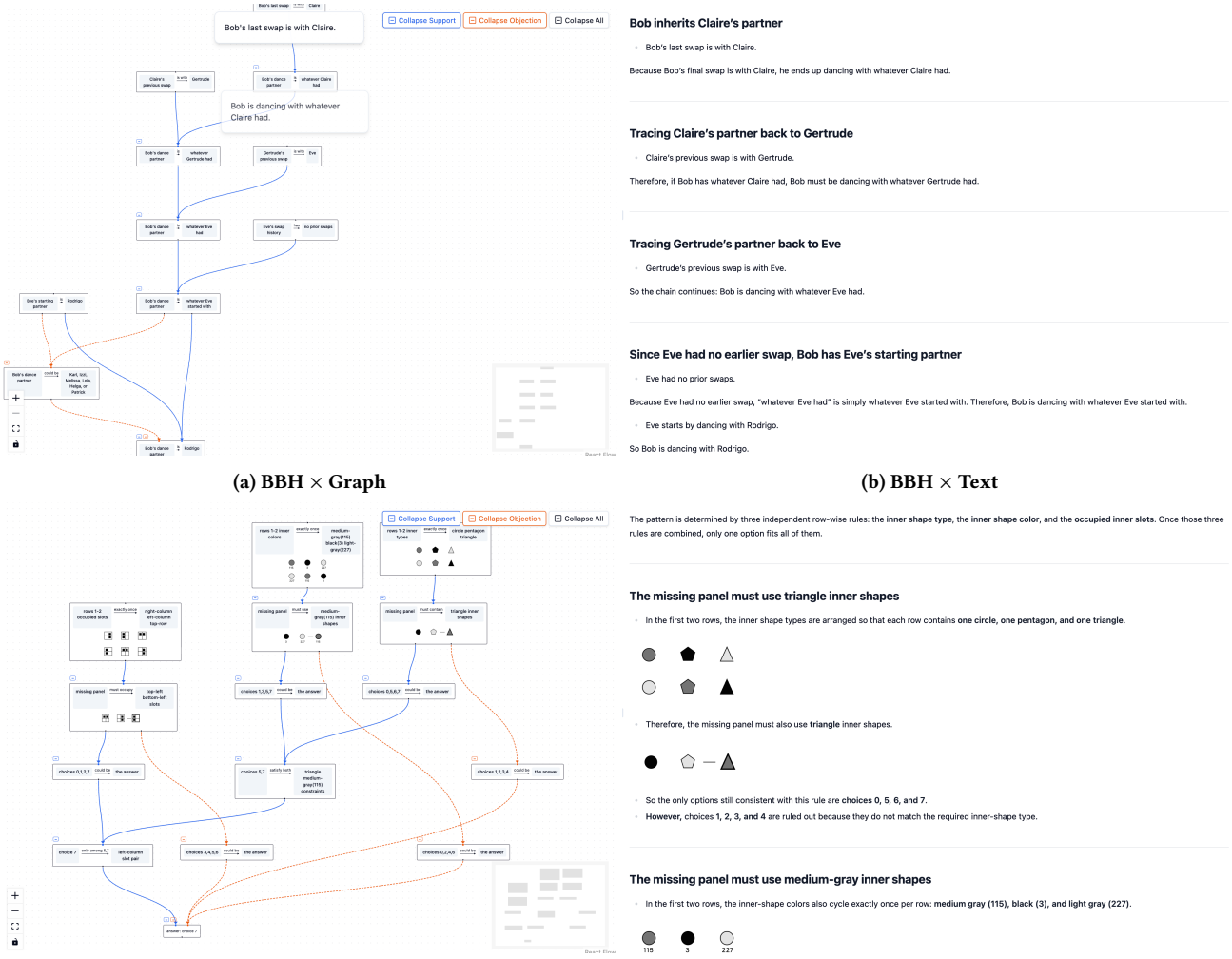
Out of 223 participants recruited, 204 valid participants passed the attention checks (BBH × Graph: 57, BBH × Text: 50, I-RAVEN × Graph: 51, I-RAVEN × Text: 46). The final sample was 52% female ($M_{\text{age}} = 41$), with 27.0% holding a graduate degree or higher, 45.6% a bachelor's degree, 17.2% some college education, 9.3% a high school diploma, and 1.0% some high school education or lower. Regarding AI experience, 8.8% had professional or research experience, 54.4% were regular users, 35.8% had basic familiarity, and 1.0% had no prior experience. The study lasted 47 minutes on average, yielding an average hourly compensation of ~$8.85.

6.3 Study Procedure

Participants were randomly assigned to one of the four conditions. The study consisted of four stages: pre-study survey, tutorial, task phase, and post-study survey.

In the pre-study survey, participants reported demographics, AI experience [28], AI reliance [27, 45], and need for cognition [6]. A tutorial then walked them through example questions step by step, introduced the interface, and explained how to interpret the assigned LLM rationales. During the task phase, participants completed 15 trials. Each trial followed the sequence: viewing question, providing initial answer, viewing LLM rationale, and making revised decision. The system provided rationales only, without explicit answer recommendations, as prior work shows users engage in more analytical reasoning without recommendations [6, 13]. Rationales were pre-generated to ensure consistency across participants.

The 15 trials included 5 easy and 10 hard questions. To examine whether rationales help users identify incorrect reasoning, 5 questions (1 easy, 4 hard) were paired with incorrect LLM answers and selected to reflect the error pattern distribution identified from our technical evaluation (Section 5.2).



(a) BBH × Graph

(b) BBH × Text

(c) I-RAVEN × Graph

(d) I-RAVEN × Text

Figure 2: Screenshots of the four user study conditions, organized by task modality (verbal vs. visual reasoning) and rationale format (graphical vs. textual).

After completing all trials, participants completed a post-study survey. Attention-check questions were included in both the main study and the post-study survey.

6.4 Measurements

6.4.1 Independent Variables. Our study included three independent variables: **rationale format** (graphical vs. textual; between-subjects), **task modality** (verbal vs. visual reasoning; between-subjects), and **question difficulty** (easy vs. hard; within-subjects).

We also measured **Need for Cognition (NFC)** as an individual-difference variable using a 4-item scale [6, 7], which assesses dispositional tendency to engage in effortful thinking. NFC was included as a moderator in secondary analyses.

6.4.2 Dependent Variables.

Objective Behavioral Measures. Our behavioral analysis examines whether participants' final decisions appropriately track the correctness of AI recommendations. We operationalize trust calibration using four trial-level metrics:

- **Appropriate Trust:** $\frac{\#(\text{accept correct} \cup \text{reject incorrect})}{\#(\text{all trials})}$
- **Over-Trust:** $\frac{\#(\text{follow incorrect})}{\#(\text{AI incorrect trials})}$
- **Under-Trust:** $\frac{\#(\text{reject correct})}{\#(\text{AI correct trials})}$
- **Error Correction:** $\frac{\#(\text{override incorrect and choose correct})}{\#(\text{AI incorrect trials})}$

where $\#(\cdot)$ denotes the number of trials satisfying the condition.

Subjective Self-Report Measures. After completing all trials, participants rated their experience on nine dimensions: Usability [26], Task Load [28, 57], Satisfaction [28], Engagement[36], Helpfulness [28, 60], Understanding [28, 57, 60], Trust [28, 57, 60], Critical Thinking, Learning Gain [54]. Participants in the graphical rationale

condition additionally rated four design features (Alignment with Human Decision Logic, Semantic Color-Coding, Text Compression, Progressive Disclosure) on a 7-point helpfulness scale [59]. Two open-ended questions were included in the post-study survey to capture participants' perceptions and improvement suggestions. All instruments are in Appendix D.2 and D.3.

6.5 Analysis Method

For trial-level objective outcomes, we fitted linear mixed-effects models with participant as a random intercept. All categorical predictors were sum-coded (−1, +1), and models were estimated using maximum likelihood. We computed estimated marginal means (EMMs) and conducted pairwise comparisons.

Subjective outcomes were analyzed at the participant level using 2×2 between-subjects ANOVA, with within-dataset simple effects examined using Welch's t-tests.

We conducted an exploratory moderation analysis on NFC. We treated standardized NFC as a continuous moderator by extending the models to include NFC and its interactions. For consistency with prior work that stratifies participants by NFC [6], we additionally report a descriptive median-split analysis. NFC was median-split within each dataset ($Mdn = 4.00$ for both BBH and I-RAVEN, which was also the scale midpoint) to form Low-NFC and High-NFC subgroups, and models were re-estimated within each subgroup.

Open-ended responses were analyzed using thematic analysis [5], separately by task modality and rationale format.

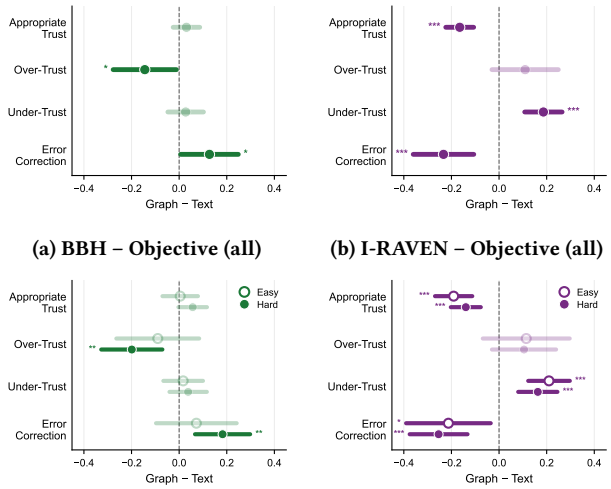
7 Results

7.1 Objective Outcomes

The rationale format effect reverses between verbal and visual reasoning. The rationale format × task modality interaction was significant across all four trust-calibration metrics (Appropriate Trust: $b = 0.049$, $p < .001$; Over-Trust: $b = -0.064$, $p = .005$; Under-Trust: $b = -0.040$, $p = .004$; Error Correction: $b = 0.090$, $p < .001$). The direction of the format effect reverses between verbal (BBH) and visual (I-RAVEN) reasoning tasks. Point estimates and 95% CIs are shown in Figure 3.

Question difficulty does not explain this reversal: none of the rationale format × task modality × question difficulty three-way interactions reached significance. However, difficulty modulates the *magnitude* of format effects: a marginal format × difficulty interaction ($b = -0.013$, $p = .068$ for Appropriate Trust) and a significant task modality × difficulty interaction ($p < .001$ for all metrics) indicate that the pattern strengthens on harder items within verbal reasoning.

Verbal reasoning (BBH): graph advantage concentrates on hard questions. Collapsed across difficulty, graphical rationales significantly reduced Over-Trust ($\Delta = -0.144$, $p = .033$) and increased Error Correction ($\Delta = +0.127$, $p = .040$). Difficulty-stratified contrasts (Figure 3c) show this benefit is absent on easy questions and emerges on hard ones: Over-Trust ($\Delta = -0.199$, $p = .002$) and Error Correction ($\Delta = +0.182$, $p = .002$) both reached significance, with Appropriate Trust showing a marginal positive trend ($\Delta = +0.057$, $p = .064$). This suggests that explicit argument structure helps only when reasoning demands are sufficiently high.



(a) BBH – Objective (all) **(b) I-RAVEN – Objective (all)**

(c) BBH – Objective (by difficulty) **(d) I-RAVEN – Objective (by difficulty)**

Figure 3: Objective trust-calibration outcomes (Graph − Text) across task modalities. Left column: Estimated Marginal Means collapsed across difficulty. Right column: Difficulty-stratified contrasts (open = Easy; filled = Hard). In verbal reasoning, the graph advantage is absent on easy questions and emerges on hard ones. In visual reasoning, the graph disadvantage is present at both difficulty levels. Faded items are non-significant ($p \geq .05$); * $p < .05$, ** $p < .01$, *** $p < .001$.

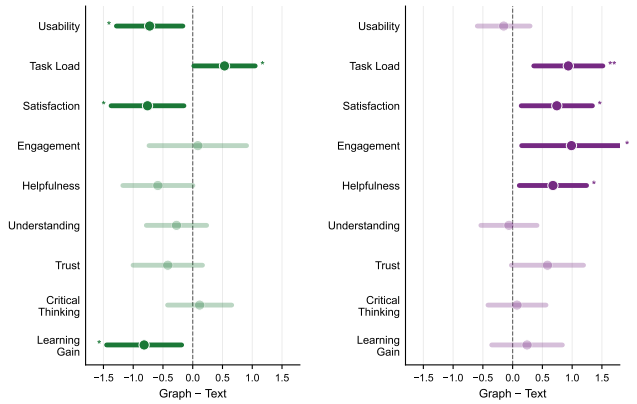
Visual reasoning (I-RAVEN): graph disadvantage is robust across difficulty. Graphical rationales impaired trust calibration in visual reasoning at both difficulty levels. Collapsed, Appropriate Trust ($\Delta = -0.165$, $p < .001$), Under-Trust ($\Delta = +0.186$, $p < .001$), and Error Correction ($\Delta = -0.233$, $p < .001$) all differed significantly. This pattern held at both difficulty levels (Figure 3d). The fact that graphical rationales hurt performance even on simple visual items rules out a question-difficulty explanation and instead points to a modality mismatch: the graph structure does not provide the perceptual grounding required for visual reasoning verification, regardless of task demand.

Task modality thus determines the *direction* of the format effect; question difficulty shapes its *magnitude* within BBH (verbal reasoning) but does not alter the cross-modality reversal.

7.2 Subjective Experience Diverges from Objective Outcomes

Despite the split in objective outcomes, subjective ratings show the opposite pattern (Figure 4), revealing a dissociation between how participants *experienced* graphical rationales and how they *performed* with them.

Verbal reasoning (BBH): graph works better but feels worse. Where graphical rationales objectively reduced over-trust and improved error correction, participants rated the experience lower on Satisfaction ($\Delta = -0.76$, $p = .017$), Usability ($\Delta = -0.72$, $p = .013$),



(a) BBH – Subjective (b) IRAVEN – Subjective

Figure 4: Subjective self-report outcomes (Graph − Text) within each task modality. Welch's *t*-test simple effects. Faded = non-significant ($p \geq .05$); * $p < .05$, ** $p < .01$, *** $p < .001$.

and Learning Gain ($\Delta = -0.82$, $p = .013$), and reported higher Task Load ($\Delta = +0.53$, $p = .045$).

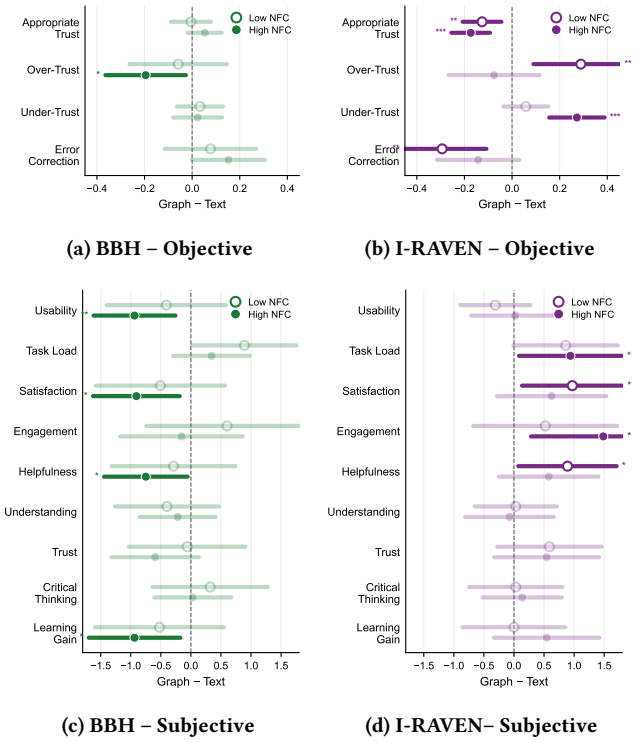
Visual reasoning (I-RAVEN): graph feels better but works worse. Where graphical rationales objectively impaired trust calibration, participants reported greater Satisfaction ($\Delta = +0.74$, $p = .017$), Engagement ($\Delta = +0.99$, $p = .023$), and Helpfulness ($\Delta = +0.68$, $p = .021$), alongside elevated Task Load ($\Delta = +0.94$, $p = .002$).

This pattern reveals that graphical rationales feel engaging in visual tasks—where they most likely produce miscalibrated trust, while imposing friction in verbal tasks—where they deliver the greatest objective benefit.

7.3 Need for Cognition Moderates the Trust-Calibration Effect

We examined whether Need for Cognition moderated the format effect. The rationale format × NFC interaction was significant for Over-Trust ($b = -0.072$, $p = .009$) and Under-Trust ($b = +0.035$, $p = .032$), but not for Appropriate Trust or Error Correction. Higher-NFC participants showed a stronger reduction in Over-Trust but also a stronger increase in Under-Trust with graphical rationales. No significant rationale format × NFC interactions were found for any subjective outcome.

To descriptively interpret this moderation, we additionally stratified participants by a median split. NFC-stratified contrasts are shown in Figure 5. In verbal reasoning (BBH), low-NFC participants showed no reliable difference between rationale formats. High-NFC participants, by contrast, showed reduced Over-Trust ($\Delta = -0.195$, $p = .024$) and a marginal improvement in Error Correction ($\Delta = +0.152$, $p = .052$), together with lower subjective ratings on Usability ($\Delta = -0.938$, $p = .009$), Satisfaction ($\Delta = -0.906$, $p = .017$), Learning Gain ($\Delta = -0.938$, $p = .019$), and Helpfulness ($\Delta = -0.750$, $p = .039$). In visual reasoning, the graph disadvantage



(a) BBH – Objective (b) I-RAVEN – Objective

(c) BBH – Subjective (d) I-RAVEN– Subjective

Figure 5: NFC-stratified format effects (Graph − Text), median-split ($Mdn = 4.00$). Open = Low NFC; filled = High NFC. Top: Objective outcomes. Bottom: Subjective outcomes. In verbal reasoning, the objective benefit concentrates in High-NFC participants; in visual reasoning, calibration impairment is present across both NFC groups. Faded = non-significant; * $p < .05$, ** $p < .01$, *** $p < .001$.

on trust calibration was present across both NFC subgroups. Both groups also rated graphical rationales more positively.

Taken together, the continuous analysis supports NFC moderation of objective trust calibration, and the median-split suggests that the calibration benefit in verbal reasoning concentrates among High-NFC participants. In contrast, subjective effects showed no reliable NFC moderation in the continuous analysis, so the subgroup differences in ratings should be interpreted as exploratory.

7.4 Design Features and Qualitative Themes

All graphical features were rated as helpful. Participants in graphical rationale conditions rated all four design features above the scale midpoint ($p < .001$): Progressive Disclosure ($M = 5.66$), Semantic Color-Coding ($M = 5.58$), Text Compression ($M = 5.45$), and Alignment with Human Decision Logic ($M = 5.39$).

Qualitative themes mirror the task modality split. Thematic analysis of open-ended responses showed that participants in the two task modalities used the same design in different ways, despite sharing a common appreciation of its stepwise structure (BBH: 22/57 = 38.6%; I-RAVEN: 21/51 = 41.2%).

In verbal reasoning (BBH), the graph served as a logic-tracing tool whose structural transparency made AI errors visible and prompted correction, but at the cost of perceived trustworthiness when errors occurred. Complaints centered on branching complexity (17.5% of BBH × Graph vs. 7.8% of I-RAVEN × Graph participants), and a disproportionate share of BBH × Graph participants (22.8% vs. 5.9%) tied their confidence in the design directly to the AI's reasoning accuracy: *"Once it gets one thing wrong it's hard to follow the logic. Once I encountered that I didn't trust the AI to give me a correct answer"* (P75). Improvement requests focused on simplification including fewer nodes and shorter reasoning chains.

In visual reasoning (I-RAVEN), the graph was valued for decomposing complex visual patterns into trackable attributes: *"It was very helpful as it breaks down complex visual patterns into simple ones"* (P48). However, the gap between the graph's abstract structure and the perceptual task limited its usefulness. Participants asked for color-coded attribute tracks, visual legends, and direct highlighting in the source images. Even participants in I-RAVEN × Text condition frequently requested visual overlays (12/44 = 27.3%), indicating that the need for perceptual grounding is broadly felt and not met by either format.

7.5 Summary

In summary, graphical rationales improve trust calibration in verbal reasoning (BBH)—particularly for analytically engaged participants and on harder questions—albeit at the cost of perceived usability. In contrast, the same design feels more engaging and helpful in visual reasoning (I-RAVEN), yet consistently impairs calibrated trust. Qualitative evidence points to a shared underlying mechanism: in verbal reasoning, the graph acts as a traceable logical structure that exposes AI errors, whereas in visual reasoning, it functions as a decomposition aid that is intuitive but lacks sufficient perceptual support for reliable verification.

8 Discussion

Our results show that graphical rationales did not help uniformly: they improved trust calibration for verbal reasoning yet impaired it for visual reasoning, while subjective ratings showed the opposite pattern in each domain. We discuss the implications of these findings in this section.

8.1 Modality Mismatch

The reversal in effectiveness across task modalities can be explained by the complementarity of representations [25, 49]. A diagram helps when its structure matches the structure of the problem. It allows problem-related operations to be done perceptually rather than symbolically, and reduces the working memory load [25].

In verbal reasoning, the argument graph achieves this match—each node maps to a logical inference and edges encode dependencies that participants can inspect sequentially. But visual reasoning operates under different representational demands. I-RAVEN tasks require integrating spatial and figural information across multiple pattern regularities. The graphical rationale, however, represents this process through an abstract argument graph disconnected from the source images. Rather than grounding participants in the perceptual features, it asks them to re-encode visual information into propositional nodes, imposing extraneous cognitive load [31, 49]. Participants must maintain the spatial matrix in working memory while also parsing the abstract graph—a dual-channel demand that, under multiple-resource theory [63], degrades rather than supports verification. Qualitative evidence supports this: I-RAVEN participants specifically requested visual overlays and color-coded attribute tracks. Importantly, this mismatch is not explained by task difficulty. The reversal held at both difficulty levels and the three-way interaction was non-significant. This points to a structural incompatibility between the graph's abstract vocabulary and the verification demands of visual reasoning.

However, I-RAVEN tests a specific instantiation of graphical rationales, i.e., argument-map-style graphs, rather than graphical rationales for visual reasoning in general. The poor calibration outcomes should not be interpreted as evidence that graphical rationales are universally harmful for visual tasks. Instead, they reveal an important *boundary condition of our design*: abstract argument graphs are poorly matched to tasks requiring perceptual information from source images. Argument-map-style graphs appear better suited to verbal reasoning, where nodes and edges can represent logical dependencies. In contrast, visual reasoning likely requires explanations perceptually grounded in source stimuli.

8.2 The Objective–Subjective Dissociation

In verbal reasoning, graphical rationales improved calibration but were rated as more demanding and less satisfying. In visual reasoning, they impaired calibration but were rated as more engaging and helpful. This shows that participants' judgments of their own reasoning quality are unreliable when rationale format varies.

Response time offers complementary behavioral evidence. We compared per-trial response time and found no format effect in verbal reasoning, whereas in visual reasoning participants spent less time with graphical than with textual rationales (28.2s vs. 36.7s, $p = .013$). This pattern further reinforces the objective–subjective dissociation. In BBH, graphical rationales increased perceived task load without increasing time spent; in I-RAVEN, they were rated as more demanding yet processed more quickly.

Therefore, for visual reasoning, the structured graph created a sense of comprehension that substituted for genuine verification—an effect similar to the illusion of explanatory depth [43]. The risk that graphical presentation gives flawed reasoning unwarranted credibility was realized here. For verbal reasoning, the graph made AI errors highly visible, improving calibration but reducing perceived trustworthiness. This is consistent with prior work showing that increased model transparency can reduce user confidence even when it improves decision quality [6, 40]. This tension reflects a core challenge for explainable AI design: formats that support calibrated trust may undermine the subjective experience that drives continued engagement. Graph structure thus can support calibration when its representation matches the task, but may increase the perceived credibility of flawed rationales when it does not.

8.3 Analytic Engagement as a Gating Condition

In verbal reasoning, only high-NFC participants showed reduced Over-Trust and improved Error Correction with graphical rationales. This suggests graphical rationales function as argument scaffolds whose benefit requires effortful engagement. This is consistent with the Elaboration Likelihood Model [7], which posits that central-route processing—careful evaluation of argument quality—requires both ability and motivation. While the argument graph provides structure that facilitates elaboration, effective evaluation still depends on users' willingness to engage with it.

The graph advantage also emerged exclusively on hard questions, where explicit structure provides scaffolding for deeper inference. Prior work on AI-assisted decision-making has similarly shown that AI explanation benefits are largest when task difficulty is high enough [52].

Together, these findings indicate that graphical rationales impose an onboarding barrier. They require a threshold of effortful engagement to deliver their benefit.

8.4 Design Implications

Graph Complexity as an Inverted-U Function. Verbal-reasoning participants complained about branching complexity, suggesting a non-linear relationship between node count and calibration benefit. Rationale graphs should be pruned to retain only decision-critical nodes. As the highest-rated design feature, progressive disclosure is a promising mechanism for managing this tradeoff by collapsing branches on demand.

Canvas-Based Presentation for Visual Tasks. The visual-reasoning findings motivate a different design: rather than imposing abstract argument graphs, rationale interfaces should anchor explanations directly in the source visual representation, preserve perceptual grounding, and incorporate argument structure.

Hybrid Representations. Aligning with the redundancy principle in multimedia learning [31], the modality-split findings motivate research on hybrid rationales that adapt their representational format based on task type, e.g., graph-structured arguments for verbal reasoning and spatial annotations for visual reasoning. We preliminarily explored augmenting verbal reasoning with additional visual cues (e.g., SVG-based spatial diagrams for BBH logical deduction tasks; see Appendix B.3). A subset of participants from our formative interviews indicated that embedding such visual cues within graphical rationales largely improved perceived clarity, suggesting a promising direction for future work.

8.5 Limitations and Future Work

Several limitations constrain the generalizability of our findings. First, our study focuses on well-structured QA tasks, whereas real-world reasoning often involves ambiguity and value-laden tradeoffs that do not decompose into discrete steps. Second, we adopt a decision-making paradigm and do not capture more open-ended, subjective aspects of human–AI collaboration, such as whether AI introduces novel perspectives or supports co-creative thinking. Third, our findings are tied to one instantiation of graphical rationales, i.e., argument-map-style graphs with condensed node text. The visual-reasoning results in particular may not generalize to other visual tasks or to other forms of graphical rationale representation.

A key direction for future work is moving from read-only to *editable* rationale graphs, where users can modify reasoning steps and prompt the model to re-evaluate from intermediate states.

9 Conclusion

We explored when and how graphical LLM rationales support human decision making, using Graphionale, a prototype that transforms linear LLM rationales into interactive argument graph visualizations. Grounded in a formative co-design study, Graphionale combines argument mapping with text condensation and adapts graph complexity through progressive disclosure. A controlled user study ($N = 204$) showed that graphical rationales did not help uniformly: they improved trust calibration for verbal reasoning—reducing over-trust and increasing error correction, especially on hard questions and among analytically engaged users—yet were perceived as more cognitively demanding and less satisfying. For visual reasoning, graphical rationales impaired trust calibration but were rated as more engaging, helpful, and satisfying, which points to a mismatch between the graph's abstract structure and the perceptual demands of the task. In each modality, the format that better supported calibrated decisions was thus not the one users preferred. Our findings provide empirical design knowledge for next-generation reasoning-aware AI interfaces that adapt rationale presentation to task modality and user needs.

Acknowledgments

We would like to thank our colleagues Vasiliki Charisi, Shuo Sun, and Alok Prakash at Singapore–MIT Alliance for Research and Technology (SMART) center for their thoughtful feedback and support throughout this project. We are also grateful to all participants who contributed their time to our formative study and user study. Finally, we thank all anonymous reviewers for their insightful and constructive comments. This research has been supported by the National Research Foundation (NRF), Prime Minister's Office, Singapore, under its Campus for Research Excellence and Technological Enterprise (CREATE) program, through the Mens, Manus and Machina (M3S) interdisciplinary research group (IRG) of the Singapore–MIT Alliance for Research and Technology (SMART) center.

References

[1] 2026. Argument map. https://en.wikipedia.org/w/index.php?title=Argument_map&oldid=1334700013 Page Version ID: 1334700013.

[2] Ian Arawjo, Chelse Swoopes, Priyan Vaithilingam, Martin Wattenberg, and Elena L. Glassman. 2024. ChainForge: A Visual Toolkit for Prompt Engineering and LLM Hypothesis Testing. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–18. doi:10.1145/3613904.3642016

[3] BIG-bench authors. 2023. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research* (2023). https://openreview.net/forum?id=uyTL5Bvosj

[4] Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. 2024. Graph of Thoughts: Solving Elaborate Problems with Large Language Models. *Proceedings of the AAAI Conference on Artificial Intelligence* 38, 16 (March 2024), 17682–17690. doi:10.1609/aaai.v38i16.29720

[5] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (Jan. 2006), 77–101. doi:10.1191/1478088706qp063oa _eprint: https://doi.org/10.1191/1478088706qp063oa.

[6] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1 (April 2021), 188:1–188:21. doi:10.1145/3449287

[7] John T. Cacioppo, Richard E. Petty, and Chuan Feng Kao. 1984. The Efficient Assessment of Need for Cognition. *Journal of Personality Assessment* 48, 3 (June 1984), 306–307. doi:10.1207/s15327752jpa4803_13 _eprint: https://doi.org/10.1207/s15327752jpa4803_13.

[8] Zichen Chen, Jianda Chen, Ambuj Singh, and Misha Sra. 2024. XplainLLM: A Knowledge-Augmented Dataset for Reliable Grounded Explanations in LLMs. 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, 7578–7596. doi:10.18653/v1/2024.emnlp-main.432

[9] Allan M. Collins and M. Ross Quillian. 1969. Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior* 8, 2 (April 1969), 240–247. doi:10.1016/S0022-5371(69)80069-1

[10] Jeff Conklin and Michael L. Begeman. 1988. gIBIS: a hypertext tool for exploratory policy discussion. In *Proceedings of the 1988 ACM conference on Computer-supported cooperative work (CSCW '88)*. Association for Computing Machinery, New York, NY, USA, 140–152. doi:10.1145/62266.62278

[11] Simon Delarue, Astrid Bertrand, and Tiphaine Viard. 2024. Evaluating graph-based explanations for AI-based recommender systems. doi:10.48550/arXiv.2407.12357 arXiv:2407.12357 [cs].

[12] Franz Faul, Edgar Erdfelder, Axel Buchner, and Albert-Georg Lang. 2009. Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods* 41, 4 (Nov. 2009), 1149–1160. doi:10.3758/BRM.41.4.1149

[13] Krzysztof Z. Gajos and Lena Mamykina. 2022. Do People Engage Cognitively with AI? Impact of AI Assistance on Incidental Learning. In *Proceedings of the 27th International Conference on Intelligent User Interfaces (IUI '22)*. Association for Computing Machinery, New York, NY, USA, 794–806. doi:10.1145/3490099.3511138

[14] Shijie Geng, Zuohui Fu, Juntao Tan, Yingqiang Ge, Gerard de Melo, and Yongfeng Zhang. 2022. Path Language Modeling over Knowledge Graphsfor Explainable Recommendation. In *Proceedings of the ACM Web Conference 2022 (WWW '22)*. Association for Computing Machinery, New York, NY, USA, 946–955. doi:10.1145/3485447.3511937

[15] Maralee Harrell. 2010. Creating Argument Diagrams. https://www.academia.edu/772321/Creating_Argument_Diagrams

[16] Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. 2019. Metrics for Explainable AI: Challenges and Prospects. doi:10.48550/arXiv.1812.04608 arXiv:1812.04608 [cs].

[17] Md Naimul Hoque, Tasfia Mashiat, Bhavya Ghai, Cecilia D. Shelton, Fanny Chevalier, Kari Kraus, and Niklas Elmqvist. 2024. The HaLLMark Effect: Supporting Provenance and Transparent Use of Large Language Models in Writing with Interactive Visualization. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–15. doi:10.1145/3613904.3641895

[18] Sheng Hu, Yuqing Ma, Xianglong Liu, Yanlu Wei, and Shihao Bai. 2021. Stratified Rule-Aware Network for Abstract Visual Reasoning. *Proceedings of the AAAI Conference on Artificial Intelligence* 35, 2 (May 2021), 1567–1574. doi:10.1609/aaai.v35i2.16248

[19] Jinbin Huang, Aditi Mishra, Bum Chul Kwon, and Chris Bryan. 2023. ConceptExplainer: Interactive Explanation for Deep Neural Networks from a Concept Perspective. *IEEE Transactions on Visualization & Computer Graphics* 29, 01 (Jan. 2023), 831–841. doi:10.1109/TVCG.2022.3209384

[20] Fariha Tasmin Jaigirdar, Carsten Rudolph, Gillian Oliver, David Watts, and Chris Bain. 2020. What Information is Required for Explainable AI? : A Provenance-based Research Agenda and Future Challenges. In *2020 IEEE 6th International Conference on Collaboration and Internet Computing (CIC)*. 177–183. doi:10.1109/CIC50333.2020.00030

[21] Utkarshani Jaimini and Amit Sheth. 2022. CausalKG: Causal Knowledge Graph Explainability Using Interventional and Counterfactual Reasoning. *IEEE Internet Computing* 26, 1 (Jan. 2022), 43–50. doi:10.1109/MIC.2021.3133551

[22] Peter Jansen, Elizabeth Wainwright, Steven Marmorstein, and Clayton Morrison. 2018. WorldTree: A Corpus of Explanation Graphs for Elementary Science Questions supporting Multi-hop Inference. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga (Eds.). European Language Resources Association (ELRA), Miyazaki, Japan. https://aclanthology.org/L18-1433/

[23] Peiling Jiang, Jude Rayan, Steven P. Dow, and Haijun Xia. 2023. Graphologue: Exploring Large Language Model Responses with Interactive Diagrams. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)*. Association for Computing Machinery, New York, NY, USA, 1–20. doi:10.1145/3586183.3606737

[24] Mark Klein. 2007. *How to Harvest Collective Wisdom for Complex Problems: An Introduction to the MIT Deliberatorium.* doi:10.13140/RG.2.2.32743.24489

[25] Jill H. Larkin and Herbert A. Simon. 1987. Why a Diagram Is (Sometimes) Worth Ten Thousand Words. *Cognitive Science* 11, 1 (1987), 65–100. doi:10.1111/j.1551-6708.1987.tb00863.x

[26] James R. Lewis, Brian S. Utesch, and Deborah E. Maher. 2013. UMUX-LITE: when there's no time for the SUS. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. Association for Computing Machinery, New York, NY, USA, 2099–2102. doi:10.1145/2470654.2481287

[27] Zongqian Li, Ehsan Shareghi, and Nigel Collier. 2025. ReasonGraph: Visualization of Reasoning Methods and Extended Inference Paths. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Pushkar Mishra, Smaranda Muresan, and Tao Yu (Eds.). Association for Computational Linguistics, Vienna, Austria, 140–147. doi:10.18653/v1/2025.acl-demo.14

[28] Shuai Ma, Qiaoyi Chen, Xinru Wang, Chengbo Zheng, Zhenhui Peng, Ming Yin, and Xiaojuan Ma. 2025. Towards Human-AI Deliberation: Design and Evaluation of LLM-Empowered Deliberative AI for AI-Assisted Decision-Making. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, 1–23. doi:10.1145/3706598.3713423

[29] Karthic Madanagopal, Eric D. Ragan, and Perakath Benjamin. 2019. Analytic Provenance in Practice: The Role of Provenance in Real-World Visualization and Data Analysis Environments. *IEEE Computer Graphics and Applications* 39, 6 (Nov. 2019), 30–45. doi:10.1109/MCG.2019.2933419

[30] Anna Martin-Boyle, Cara A. C. Leckey, Martha C. Brown, and Harmanpreet Kaur. 2026. PaperTrail: A Claim-Evidence Interface for Grounding Provenance in LLM-based Scholarly Q&A. doi:10.1145/3772318.3791101 arXiv:2602.21045 [cs].

[31] Richard E. Mayer and Roxana Moreno. 2003. Nine Ways to Reduce Cognitive Load in Multimedia Learning. *Educational Psychologist* 38, 1 (2003), 43–52. doi:10.1207/S15326985EP3801_6

[32] Aditi Mishra, Sajjadur Rahman, Kushan Mitra, Hannah Kim, and Estevam Hruschka. 2024. Characterizing Large Language Models as Rationalizers of Knowledge-intensive Tasks. In *Findings of the Association for Computational Linguistics: ACL 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 8117–8139. doi:10.18653/v1/2024.findings-acl.484

[33] Andrea Montagna, Alvise De Biasio, Nicolo' Navarin, Fabio Aiolli, and others. 2023. Graph-based Explainable Recommendation Systems: Are We Rigorously Evaluating Explanations? A Position Paper. In *CEUR WORKSHOP PROCEEDINGS*, Vol. 3502. CEUR-WS.

[34] Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candes, and Tatsunori Hashimoto. 2025. s1: Simple test-time scaling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (Eds.). Association for Computational Linguistics, Suzhou, China, 20275–20321. doi:10.18653/v1/2025.emnlp-main.1025

[35] Joseph D Novak and Alberto J Cañas. 2008. The theory underlying concept maps and how to construct and use them. (2008).

[36] Heather L. O'Brien and Elaine G. Toms. 2010. The development and evaluation of a survey to measure user engagement. *Journal of the American Society for Information Science and Technology* 61, 1 (2010), 50–69. doi:10.1002/asi.21229 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/asi.21229.

[37] Rock Yuren Pang, K. J. Kevin Feng, Shangbin Feng, Chu Li, Weijia Shi, Yulia Tsvetkov, Jeffrey Heer, and Katharina Reinecke. 2025. Interactive Reasoning: Visualizing and Controlling Chain-of-Thought Reasoning in Large Language Models. doi:10.48550/arXiv.2506.23678 arXiv:2506.23678 [cs].

[38] Paul Parsons. 2022. Understanding Data Visualization Design Practice. *IEEE Transactions on Visualization and Computer Graphics* 28, 1 (Jan. 2022), 665–675. doi:10.1109/TVCG.2021.3114959

[39] Wagner de A. Perin, Davidsom Cury, Camila Zacché de Aguiar, and Crediné S. de Menezes. 2023. From Text to Maps: Automated Concept Map Generation Using Fine-tuned Large Language Model. In *Simpósio Brasileiro de Informática na Educação (SBIE)*. SBC, 1317–1328. doi:10.5753/sbie.2023.234749

[40] Forough Poursabzi-Sangdeh, Daniel G. Goldstein, Jake M. Hofman, Jennifer W. E. Wortman Vaughan, and Hanna M. Wallach. 2021. Manipulating and Measuring Model Interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. ACM, New York, NY, USA. doi:10.1145/3411764.3445315

[41] Enayat Rajabi and Kobra Etminani. 2024. Knowledge-graph-based explainable AI: A systematic review. *Journal of Information Science* 50, 4 (Aug. 2024), 1019–1029. doi:10.1177/01655515221112844

[42] Milena Rmus, Harrison Ritz, Lindsay E. Hunter, Aaron M. Bornstein, and Amitai Shenhav. 2022. Humans can navigate complex graph structures acquired during latent learning. *Cognition* 225 (Aug. 2022), 105103. doi:10.1016/j.cognition.2022.105103

[43] Leonid Rozenblit and Frank Keil. 2002. The Misunderstood Limits of Folk Science: An Illusion of Explanatory Depth. *Cognitive Science* 26, 5 (2002), 521–562. doi:10.1207/s15516709cog2605_1

[44] Maria Araceli Ruiz-Primo and Richard J. Shavelson. 1996. Problems and issues in the use of concept maps in science assessment. *Journal of Research in Science Teaching* 33, 6 (1996), 569–600. doi:10.1002/(SICI)1098-2736(199608)33:6<569::AID-TEA1>3.0.CO;2-M _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/%28SICI%291098-2736%28199608%2933%3A6%3C569%3A%3AAID-TEA1%3E3.0.CO%3B2-M.

[45] David D. Scholz, Johannes Kraus, and Linda Miller. 2025. Measuring the Propensity to Trust in Automated Technology: Examining Similarities to Dispositional Trust in Other Humans and Validation of the PTT-A Scale. *International Journal of Human–Computer Interaction* 41, 2 (Jan. 2025), 970–993. doi:10.1080/10447318.2024.2307691 _eprint: https://doi.org/10.1080/10447318.2024.2307691.

[46] Weiping Song, Zhijian Duan, Ziqing Yang, Hao Zhu, Ming Zhang, and Jian Tang. 2022. Ekar: An Explainable Method for Knowledge Aware Recommendation. doi:10.48550/arXiv.1906.09506 arXiv:1906.09506 [cs].

[47] 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 (UIST '23)*. Association for Computing Machinery, New York, NY, USA, 1–18. doi:10.1145/3586183.3606756

[48] Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, and Jason Wei. 2023. Challenging BIG-Bench Tasks and Whether Chain-of-Thought Can Solve Them. In *Findings of the Association for Computational Linguistics: ACL 2023*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 13003–13051. doi:10.18653/v1/2023.findings-acl.824

[49] John Sweller, Jeroen J. G. van Merriënboer, and Fred G. W. C. Paas. 1998. Cognitive Architecture and Instructional Design. *Educational Psychology Review* 10, 3 (1998), 251–296. doi:10.1023/A:1022193728205

[50] Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, Minneapolis, Minnesota, 4149–4158. doi:10.18653/v1/N19-1421

[51] XYflow Team. 2026. React Flow. https://reactflow.dev/

[52] Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael S. Bernstein, and Ranjay Krishna. 2023. Explanations Can Reduce Overreliance on AI Systems During Decision-Making. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1 (April 2023), 129:1–129:38. doi:10.1145/3579605

[53] Aniket Vashishtha, Abbavaram Gowtham Reddy, Abhinav Kumar, Saketh Bachu, Vineeth N. Balasubramanian, and Amit Sharma. 2024. Causal Order: The Key to Leveraging Imperfect Experts in Causal Inference. https://openreview.net/forum?id=9juyeCqL0u

[54] Gayle Vogt, Catherine Atwong, and Jean Fuller. 2005. Student Assessment of Learning Gains (SALGains): An Online Instrument. *Business Communication Quarterly* 68, 1 (March 2005), 36–43. doi:10.1177/1080569904273332

[55] Xinru Wang, Chen Liang, and Ming Yin. 2023. The Effects of AI Biases and Explanations on Human Decision Fairness: A Case Study of Bidding in Rental Housing Markets. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization, Macau, SAR China, 3076–3084. doi:10.24963/ijcai.2023/343

[56] Xinru Wang and Ming Yin. 2021. Are Explanations Helpful? A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making. In *Proceedings of the 26th International Conference on Intelligent User Interfaces (IUI '21)*. Association for Computing Machinery, New York, NY, USA, 318–328. doi:10.1145/3397481.3450650

[57] Xinru Wang and Ming Yin. 2022. Effects of Explanations in AI-Assisted Decision Making: Principles and Comparisons. *ACM Trans. Interact. Intell. Syst.* 12, 4 (Nov. 2022), 27:1–27:36. doi:10.1145/3519266

[58] Xinru Wang and Ming Yin. 2023. Watch Out for Updates: Understanding the Effects of Model Explanation Updates in AI-Assisted Decision Making. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–19. doi:10.1145/3544548.3581366

[59] Xinru Wang, Ming Yin, Eunyee Koh, and Mustafa Doga Dogan. 2026. XAgen: An Explainability Tool for Identifying and Correcting Failures in Multi-Agent Workflows. doi:10.48550/arXiv.2512.17896 arXiv:2512.17896 [cs].

[60] Xinru Wang, Mengjie Yu, Hannah Nguyen, Michael Iuzzolino, Tianyi Wang, Peiqi Tang, Natasha Lynova, Co Tran, Ting Zhang, Naveen Sendhilnathan, Hrvoje Benko, Haijun Xia, and Tanya R. Jonker. 2025. Less or More: Towards Glanceable Explanations for LLM Recommendations Using Ultra-Small Devices. In *Proceedings of the 30th International Conference on Intelligent User Interfaces (IUI '25)*. Association for Computing Machinery, New York, NY, USA, 938–951. doi:10.1145/3708359.3712074

[61] Colin Ware. 2019. *Information visualization: perception for design*. Morgan Kaufmann.

[62] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Advances in Neural Information Processing Systems* 35 (Dec. 2022), 24824–24837. https://proceedings.neurips.cc/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html

[63] Christopher D. Wickens. 2002. Multiple Resources and Performance Prediction. *Theoretical Issues in Ergonomics Science* 3, 2 (2002), 159–177. doi:10.1080/14639220210123806

[64] Jie Yang, Soyeon Caren Han, and Josiah Poon. 2022. A survey on extraction of causal relations from natural language text. *Knowledge and Information Systems* 64, 5 (May 2022), 1161–1186. doi:10.1007/s10115-022-01665-w

[65] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. In *Advances in Neural Information Processing Systems*, Vol. 36. Curran Associates, Inc., 11809–11822. https://proceedings.neurips.cc/paper/2023/hash/271db9922b8d1f4dd7aaef84ed5ac703-Abstract.html

[66] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. 2022. ReAct: Synergizing Reasoning and Acting in Language Models. https://openreview.net/forum?id=WE_vluYUL-X

[67] Chen Yu. 2025. cwhy/i-raven. https://github.com/cwhy/i-raven original-date: 2021-10-02T03:54:11Z.

[68] Chi Zhang, Feng Gao, Baoxiong Jia, Yixin Zhu, and Song-Chun Zhu. 2019. RAVEN: A Dataset for Relational and Analogical Visual REasoNing. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Long Beach, CA, USA, 5312–5322. doi:10.1109/CVPR.2019.00546

A Formative Study Details

A.1 Prompt for Converting Text to Argument Map

Instruction Prompt.

```
You are an expert logic analyzer and you are good at making argument map.
```

System Prompt.

```
You are to transform written text into an **argument map** following these steps:
1. **Identify all claims** made by the author.
2. **Rewrite each claim** as an independent, concise statement.
- Keep it **minimal but readable**, just enough for humans to understand.
- Avoid redundancy and filler words.
3. **Classify** each statement as a *premise*, *sub-conclusion*, or *main conclusion*.
4. **Optionally** include implied premises or conclusions if they are necessary for logical completeness.
5. **Represent** each statement as a node (box) in JSON format.
6. **Connect** nodes with directed edges (arrows) that indicate *support* relationships — from premise(s) to sub-conclusion(s)
↪ or conclusion(s).

**Edge rules**
- Use arrows only to show *support* (never opposition).
- If multiple premises support the same conclusion, each should have its own edge to that conclusion.
- Sub-conclusions can serve as intermediate nodes supported by earlier premises and supporting later conclusions.
- **Do not** connect sibling premises to each other unless one explicitly depends on the other.
- **Do not** create edges between independent reasons; they should each point directly to the conclusion they support.
```

User Input Template.

```
text: <rationale text>
```

Example Input.

```
text: The model selected "airport" because the phrase "checked baggage" strongly suggests air travel. Checking luggage is a
↪ common airport activity.
Therefore, the most plausible answer is "airport." Other options such as "garbage can," "military," "jewelry store," and "safe"
↪ are less consistent with the context.
```

Structured Output Schema (Using Zod).

```
{
"nodes": [
  {
    "id": "...",
    "data": {
      "label": "..."
    },
    "type": "premise" | "sub-conclusion" | "conclusion"
  }
],
"edges": [
  {
    "id": "...",
    "source": "...",
    "target": "..."
  }
]
}
```

A.2 Low-Fidelity Interface Screenshot

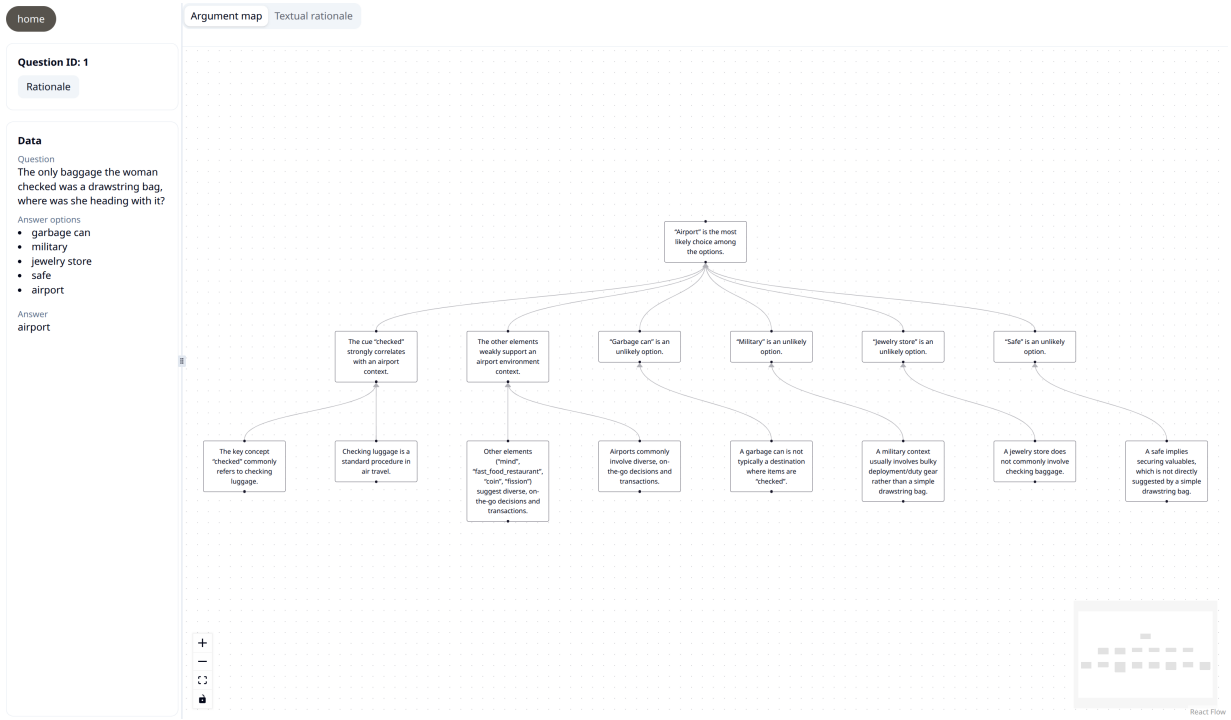
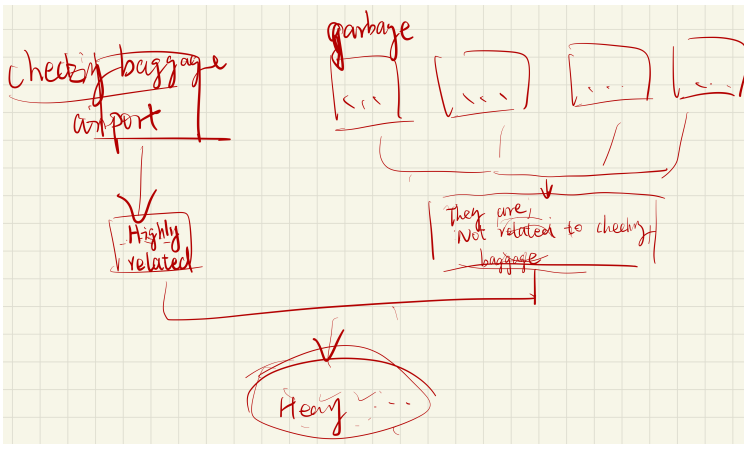
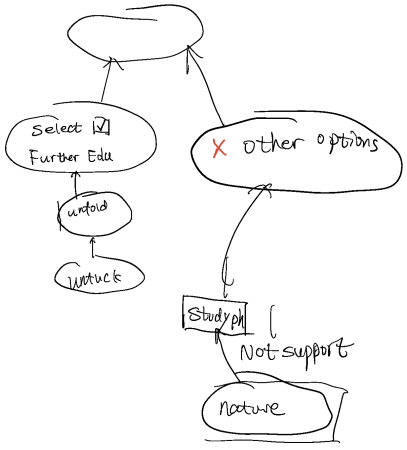


Figure A1: Screenshot of the low-fidelity prototype used in the formative study. The left panel shows a CommonsenseQA question. Participants first reviewed the textual rationale on the right panel before switching to the graphical rationale for comparison.

A.3 Example Participant Sketches



(a)



(b)

Figure A2: Example participant sketches from the formative co-design study. (a) P1 sketched a graph that groups the correct answer ("checking baggage / airport") and incorrect options (e.g., "garbage") into separate branches, with a shared node summarizing why the distractors are unrelated, then leading to a final conclusion node. (b) P10 sketched a graph distinguishing the selected answer ("Further Edu," marked with a checkmark) from rejected options (marked with a cross), using hierarchical nodes with progressive unfolding and explicit "not support" edges for elimination.

B System Implementation Details

B.1 Prompts for Graphical Rationale Generation

B.1.1 Question–Answer Pair to Reasoning Structure.

Developer Prompt.

```
transform a question-answer pair into an **argument map** — a structured reasoning graph.

**node rules**
- each node must express exactly **one atomic claim**. if a sentence covers multiple independent aspects (e.g.,
different attributes, different entities) or chains multiple reasoning steps (e.g., "A, so B, therefore C"),
split it into separate nodes connected by an edge.
- similarly, if one reasoning step resolves multiple independent entities (e.g., "mango → position 5 and kiwi →
position 6"), split them into separate nodes — one per entity.
- each node label must be a **complete sentence** clearly stating the claim or observation.
- classify each node as *premise*, *sub-conclusion*, or *conclusion*.
- the final conclusion node states the chosen answer.
- use the fewest nodes necessary to convey the full reasoning. every node must earn its place — if removing a
node would not lose any logical step, remove it.

**edge rules**
- each edge must have a *relation*: **support** (reason in favor) or **objection** (reason against).
- **support**: use when a premise provides evidence FOR a claim. e.g., premise "inner shape is triangle"
--support--> "choice 1 could be the answer".
- **objection**: use when a premise provides evidence AGAINST a candidate. e.g., premise "these choices do not
show four triangles" --objection--> "choices 2-7 could be the answer".
- **elimination flow**: a candidate sub-conclusion whose set does NOT include the final answer is an eliminated
candidate — it contradicts the conclusion and connects via **objection**. full chain: evidence --objection-->
eliminated candidate --objection--> conclusion.
- **surviving candidate flow**: a candidate sub-conclusion whose set INCLUDES the final answer is narrowing
toward the answer — it connects to the conclusion (or to a more specific surviving candidate) via **support**.
e.g., "choices 2 and 6 fit the size constraint" --support--> "answer: choice 2".
- if multiple premises support or object to the same node, each should have its own edge.
- sub-conclusions may support other sub-conclusions to form deeper reasoning chains when the logic requires
intermediate steps.
- do not connect sibling premises unless one explicitly depends on the other.
```

Output Schema.

```
{
  "nodes": [
    {
      "id": "...",
      "data": {
        "label": "...",
        "svg": "..." | null
      },
      "type": "premise" | "sub-conclusion" | "conclusion"
    }
  ],
  "edges": [
    {
      "id": "...",
      "source": "...",
      "target": "...",
      "relation": "support" | "objection"
    }
  ]
}
```

B.1.2 Reasoning Structure to Graphical Rationale.

Developer Prompt.

```
you are given a list of reasoning nodes with full-sentence labels.
for each node, produce a compressed representation:
```

```
1. try to decompose the label into a **triplet**:
   - subject: the main entity or concept (2-4 keyword tokens)
   - predicate: the relationship or action (1-3 words)
   - object: the related entity, value, or result (2-4 keyword tokens)
   - also provide a short fallback label (max 6 words)
2. if the sentence does not fit a natural subject-predicate-object structure, set subject/predicate/object to
null and provide only the label.

**important**:
- labels must contain concrete specifics (attribute names, values, choice numbers) — never vague summaries like
"missing panel specification" or "rule analysis". a good label lets the reader understand the claim without
hovering for details.
- when the original label describes a rule or pattern (e.g. progression, constant, distribute three), the
compressed label must preserve the rule name and key values. e.g. "color progression: 31→59→87", not just "color
fixed as grayscale 87".
- preserve human-readable formatting from the original: e.g. keep "grayscale 199" not "199", keep "the largest"
not "0.9x" or "level 5".

keep the same node id.

examples:
  "the left-shape sizes follow a distribute-three cycle of medium, small, and large"
  → subject: "left-shape sizes", predicate: "distribute-three", object: "M, S, L", label: "left size: M-S-L
cycle"

  "choices 0, 3, and 7 are eliminated because their size does not match"
  → subject: "choices 0,3,7", predicate: "eliminated by", object: "size mismatch", label: "eliminate 0,3,7: size"

  "the answer is choice 5"
  → subject: null, predicate: null, object: null, label: "answer: choice 5"
```

Output Schema (organized by Zod).

```
{
  "nodes": [
    {
      "id": "...",
      "label": "...",
      "subject": "..." | null,
      "predicate": "..." | null,
      "object": "..." | null
    }
  ]
}
```

B.1.3 Dataset-Specific Prompt Prefixes. In the first stage (B.1.1: question-answer pair to reasoning structure), we prepend dataset-specific prompt prefixes before the general developer prompt. For I-RAVEN, these prefixes describe the matrix layout, valid rule types, configuration-specific structure, and SVG-related constraints. For Big-Bench Hard, these prefixes define task-specific reasoning templates for temporal reasoning, logical deduction, and tracking shuffled objects.

B.2 System Architecture

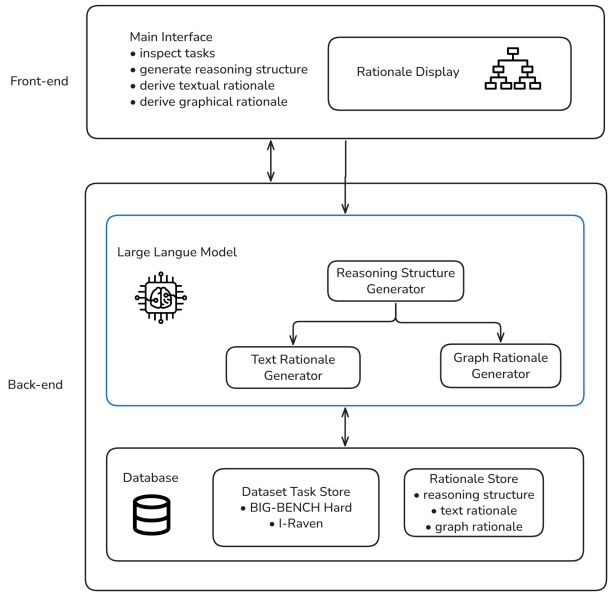


Figure B3: System architecture of GRAPHIONALE.

B.3 Example of Visually Augmented Verbal Reasoning Rationale

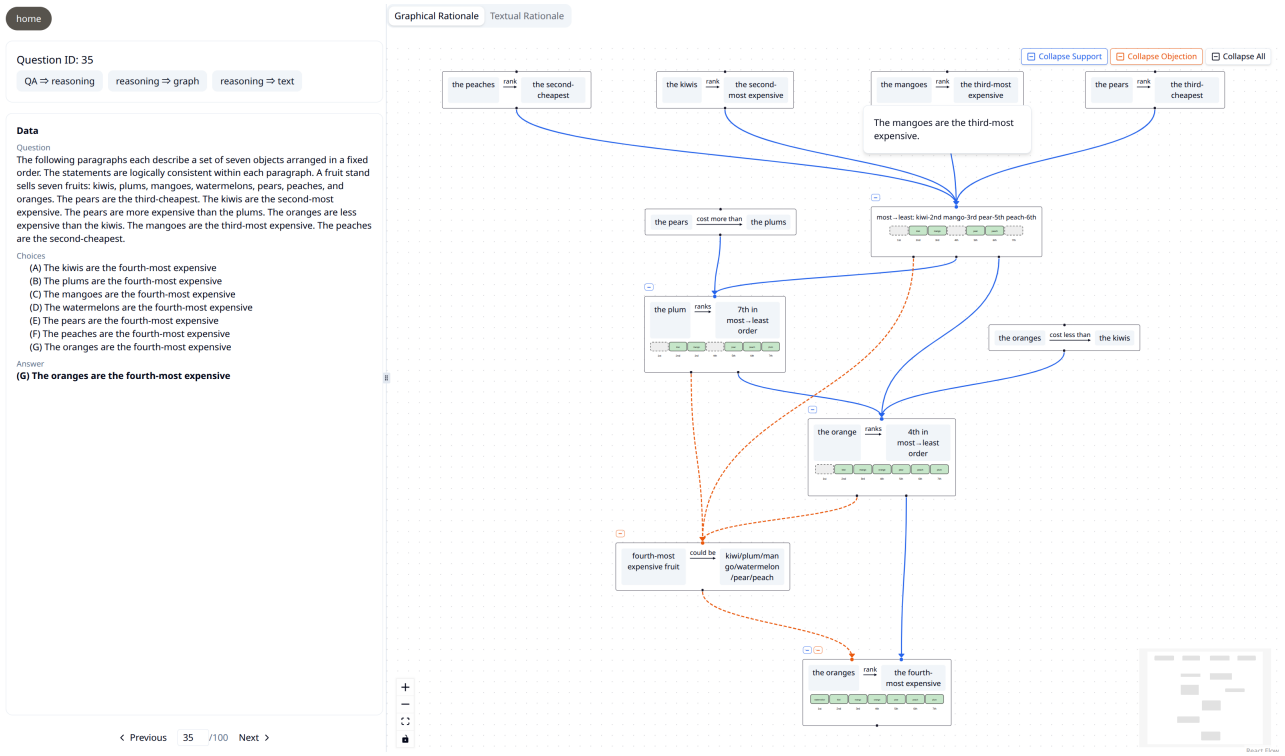


Figure B4: Example of a visually augmented rationale for a BBH Logical Deduction (7 Objects) task.

C Technical Evaluation Details

C.1 Example Tasks from BIG-Bench Hard

Data

Question

Today, Elizabeth went to the amusement park. Between what times could they have gone? We know that: Elizabeth woke up at 6am. Sarah saw Elizabeth taking photos near the Leaning Tower of Pisa from 8am to 9am. Jessica saw Elizabeth working out at the gym from 9am to 3pm. Betty saw Elizabeth watching a movie at the theater from 3pm to 7pm. James saw Elizabeth stretching at a yoga studio from 7pm to 8pm. The amusement park was closed after 8pm. Between what times could Elizabeth have gone to the amusement park?

Choices

(A) 8am to 9am
(B) 6am to 8am
(C) 3pm to 7pm
(D) 7pm to 8pm

Answer

(B) 6am to 8am

(a) Temporal Sequence

Data

Question

Alice, Bob, Claire, Dave, Eve, Fred, and Gertrude are dancers at a square dance. At the start of a song, they each have a partner: Alice is dancing with Melissa, Bob is dancing with Ophelia, Claire is dancing with Patrick, Dave is dancing with Jamie, Eve is dancing with Lola, Fred is dancing with Helga, and Gertrude is dancing with Izzi. Throughout the song, the dancers often trade partners. First, Alice and Fred switch partners. Then, Gertrude and Eve switch partners. Then, Gertrude and Alice switch partners. Then, Bob and Alice switch partners. Then, Dave and Fred switch partners. Then, Eve and Alice switch partners. Finally, Eve and Claire switch partners. At the end of the dance, Gertrude is dancing with

Choices

(A) Melissa
(B) Ophelia
(C) Patrick
(D) Jamie
(E) Lola
(F) Helga
(G) Izzi

Answer

(F) Helga

(b) Tracking Shuffled Objects (7 Objects)

Data

Question

The following paragraphs each describe a set of seven objects arranged in a fixed order. The statements are logically consistent within each paragraph. In a golf tournament, there were seven golfers: Eli, Eve, Mel, Joe, Mya, Rob, and Ada. Ada finished above Mya. Eli finished below Joe. Eli finished above Eve. Ada finished second-to-last. Mel finished third. Rob finished fourth.

Choices

(A) Eli finished third-to-last
(B) Eve finished third-to-last
(C) Mel finished third-to-last
(D) Joe finished third-to-last
(E) Mya finished third-to-last
(F) Rob finished third-to-last
(G) Ada finished third-to-last

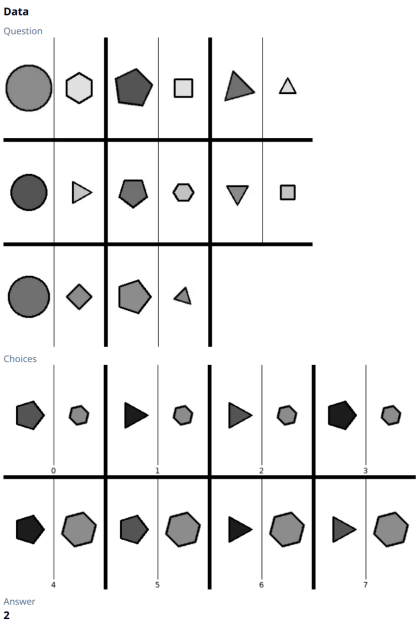
Answer

(B) Eve finished third-to-last

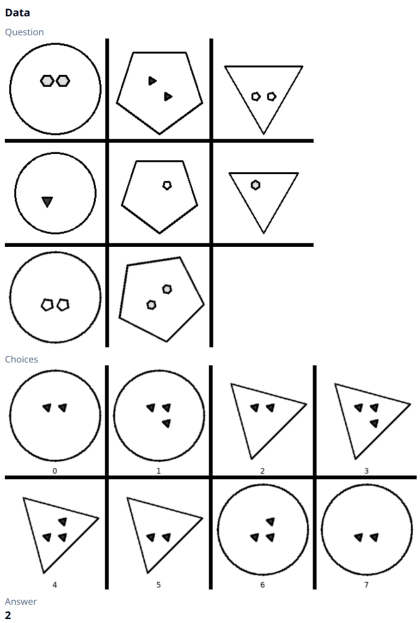
(c) Logical Deduction (7 Objects)

Figure C5: Three example tasks from Big-Bench-Hard.

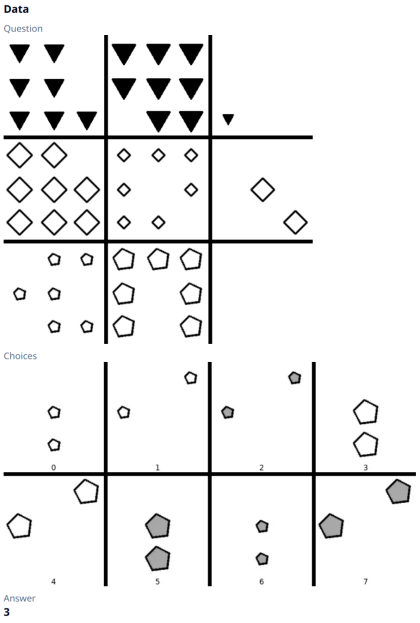
C.2 Example Tasks from I-RAVEN



(a) L-R



(b) Out-InGrid



(c) 3×3Grid

Figure C6: Three example tasks from I-RAVEN.

D User Study Details

D.1 Prompt for Generating Textual Rationale

Developer Prompt.

```
you are a reasoning analyst.
given an argument map (JSON with nodes and edges), convert it into a well structured markdown rationale.

**goals**
- the text must read as a natural, self contained explanation, not a list of claims.
- guide the reader through the reasoning step by step so the logic feels intuitive.
- use clear causal connectors (because, therefore, however) to link ideas.

**formatting**
- use markdown headings (##) for the conclusion and key sub-conclusions.
- use bullet points for supporting premises under each heading.
- use "---" dividers between major reasoning sections.
- when a premise is an objection (relation = "objection"), mark it clearly (e.g. "however, ...").
- if a node has an svg field, embed the svg directly in the markdown (not in a code block).
- preserve the logical flow: premises → sub-conclusions → conclusion.
- keep it concise and direct. do not reference node IDs or include JSON/metadata.
```

user input. This stage takes as input the argument map produced in the previous stage.

```
Argument map:
{JSON serialized argument map}
```

D.2 Pre-Study Survey

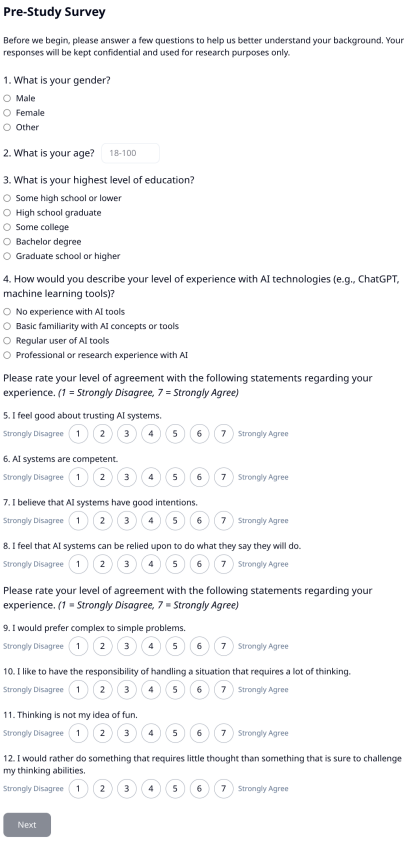


Figure D7: Screenshot of the pre-study survey.

D.3 Post-Study Survey

Post-Study Survey

Thank you! You have completed all 15 tasks. We would now like to ask you a few questions about your experience with the AI's rationale.

Your responses are an important part of this study, so we encourage you to answer as carefully and honestly as possible.

Some attention check questions were included throughout the study. Failing these checks may affect your eligibility for the performance bonus.

Please rate your level of agreement with the following statements regarding your experience. *(1 = Strongly Disagree, 7 = Strongly Agree)*

1. The way the AI organizes its rationale information meets my requirements for completing the task effectively.
Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

2. The AI rationale is easy to follow and navigate.
Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

3. Understanding the AI's rationale and using it for decision-making was mentally demanding.
Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

4. I had to work hard to process the AI's logic and integrate it into my own judgment.
Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

5. The structure of the AI's rationale and the interaction required to understand it were complex.
Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

6. I am satisfied with the assistance provided by the AI rationale.
Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

7. I was so involved in actively analyzing the AI's reasoning structure that I lost track of my immediate surroundings.
Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

8. I think the AI's rationale is helpful/useful for me to make good decisions.
Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

9. For this item, please select "Strongly Agree (7)'.
Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

10. I understand the internal reasoning process the AI used to support the final conclusion.
Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

11. I trust the AI's rationale to provide reliable decision assistance.
Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

12. I carefully evaluated the AI's reasoning by considering potential weaknesses or alternatives in the explanation.
Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

13. Interacting with the AI's rationale helped me gain a deeper understanding and learn new perspectives on the problem.
Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

Please rate how much you found these specific features helpful. *(1 = Not at all helpful, 7 = Extremely helpful)*

14. Organizing the graph in a way that follows how people naturally reason through the problem.
Not at all helpful 1 2 3 4 5 6 7 Extremely helpful

15. Using different colors to clearly distinguish supporting vs. opposing arguments.
Not at all helpful 1 2 3 4 5 6 7 Extremely helpful

16. Condensing long text within each node into a concise, easy-to-scan format.
Not at all helpful 1 2 3 4 5 6 7 Extremely helpful

17. Allowing users to expand or collapse supporting and opposing details as needed to manage complexity.
Not at all helpful 1 2 3 4 5 6 7 Extremely helpful

18. Do you think the way the AI presented its reasoning was helpful for your decision-making? Please explain why or why not.
Your answer...

19. To make the AI's reasoning more effective and easier to follow, which specific design elements do you think should be improved, and how?
Your answer...

Submit

Figure D8: Screenshot of the post-study survey.