Criticality: Scaffolding Decision-Making with Interactive Critical Thinking and Evidence-Based Reasoning Traces

Minsuk Chang*
School of Interactive Computing
Georgia Institute of Technology
Atlanta, Georgia, USA
minsuk@gatech.edu

Arjun Srinivasan
Tableau Research
Salesforce
Seattle, Washington, USA
arjunsrinivasan@salesforce.com

Srishti Palani
Tableau Research
Salesforce
Palo Alto, California, USA
srishti.palani@salesforce.com

Abstract

Decision-making requires examining underlying assumptions and concepts, considering diverse perspectives, and weighing potential consequences with clear, accurate reasoning. Recent large language models (LLMs) show promise for assisting decision-makers by combining reasoning capabilities with the ability to retrieve relevant information from large documents. However, our formative study with five professional decision-makers revealed key limitations of using LLM in workflow: time-consuming alignment of user goals, lack of evidence-based grounding, overwhelmingly long outputs, and unsurfaced assumptions undermined user trust in the LLM output and the validity of the final decision. We introduce Criticality, a system that operationalizes the Paul-Elder Critical Thinking framework to structure reasoning into interactive Elements of Thought (e.g., purpose, assumptions, perspectives, implications), and evaluates and guides reasoning using Intellectual Standards (e.g., clarity, fairness, logic). It also retrieves evidence for each claim, classifies it as supporting, neutral, or contradictory, and explains the claim-evidence link. A within-subjects study (n=13) comparing Criticality to ChatGPT 5 Pro, a state-of-the-art reasoning model in conversational interface, found that Criticality improved user interaction of steering and repairing through the decision-making process, producing better decision rationales compared to the baseline.

CCS Concepts

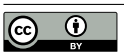
• **Human-centered computing → Interactive systems and tools**.

Keywords

Critical Thinking, Decision-Making, Large Language Models, Systems Thinking, Human-AI Interaction

ACM Reference Format:
Minsuk Chang, Arjun Srinivasan, and Srishti Palani. 2026. Criticality: Scaffolding Decision-Making with Interactive Critical Thinking and Evidence-Based Reasoning Traces. In *31st International Conference on Intelligent User Interfaces (IUI '26), March 23–26, 2026, Paphos, Cyprus.* ACM, New York, NY, USA, 21 pages. https://doi.org/10.1145/3742413.3789077

*This work was done during an internship at Salesforce Tableau Research.



IUI '26, Paphos, Cyprus

ACM ISBN 979-8-4007-1984-4/26/03
https://doi.org/10.1145/3742413.3789077

1 Introduction

Decision-making is the process of identifying and selecting the best course of action among alternatives to achieve a desired outcome. This ranges from everyday choices like comparing which product to buy or route to take to work, to more high-stakes ones such as selecting between competing investment strategies, career choices, or policy interventions. Sound decision-making requires more than just intuition; it demands *critical thinking*, i.e., the ability to reason logically, examine assumptions, explore different perspectives, and ground conclusions in evidence [14, 20, 77, 78]. In today's information-rich world, decision-makers also increasingly rely on evidence-based reasoning, using documents, often filled with visualizations, tables, textual insights, and analyses to ensure decisions rest on credible evidence and transparent, defensible logic [3–5].

Large Language models (LLMs) with advanced mechanisms (e.g., extended context windows [21], data retrieval-drieven methods [34], and deep research modes [68, 114]) have shown promise in processing large volumes of text documents, but present critical limitations for decision-making. LLMs frequently generate hallucinated or biased responses [28, 36, 84, 96], and traditional conversational interfaces, where users type a prompt and receive a single lengthy textual output, trap them in inefficient multi-turn repair cycles to clarify intent [92]. Recent reasoning models (e.g., GPT o-series [67], Claude Sonnet [2], or Gemini Pro [21]) are trained to generate intermediate steps in a *reasoning trace* before arriving at the final answer. While these approaches improve performance on benchmarked reasoning tasks, they often fail to generalize to unfamiliar or open-ended problems [88]. For users, the interaction remains passive: they must wait as the model "thinks," with latency that grows alongside task complexity [79, 117]. This limits opportunities for users to inspect, steer, or repair reasoning, leading to frustration and reduced trust [80, 109]. Even with the reasoning trace, users face challenges in evaluating whether the model has engaged in critical, evidence-based exploration examining multiple perspectives, potential biases, and implications [13, 118]. Evaluation that is especially important given that these reasoning traces do not faithfully represent the models' actual reasoning, reducing their transparency value and their reliability for decision-making [18].

To investigate needs and challenges in document-driven decision-making workflows, we conducted a formative study with five professionals who regularly make or recommend business decisions. Through interviews, we identified key challenges, including aligning on intents, unclear evidence referencing, and unexpressed assumptions or alternatives. Through a participatory design exercise, we validated and extended the Paul-Elder Critical Thinking Framework [77], demonstrating that it can be a useful way to structure a

reasoning trace and get alignment between users and LLMs. Specifically, the framework's Elements of Thought (purpose, question at issue, assumptions, point of views, concepts, information, inferences, consequences) can scaffold the reasoning process into manageable components that users can inspect and modify, while the Intellectual Standards (clarity, accuracy, precision, relevance, depth, breadth, logic, significance, and fairness) provide quality control mechanisms to ensure rigorous critical thinking (Table 1).

Based on these insights and design considerations, we built Criticality, an interactive system that structures LLM reasoning into editable elements of thought, assesses and guides thinking quality with intellectual standards, and bases each reasoning on relevant evidence from the report with explanatory links of whether the evidence supports or contradicts the claim. We conducted a within-subjects study asking 13 participants to engage in two document-based decision-making tasks and produce a rationale for each using Criticality in one condition and using a Baseline of ChatGPT 5 Pro, a state-of-the-art reasoning model embedded in a popular conversational interface, in another. Our findings demonstrate that when using Criticality, participants were able to interact with the reasoning trace to steer and repair it during the process, often preferring to use Criticality over Baseline. Also, blind-to-condition raters' evaluations of participants' decision-rationales indicated that participants produced higher-quality decisions when using Criticality.

In summary, this work makes the following contributions:

- Insights from formative interviews and participatory design exercises with five decision-makers that identified common workflows, challenges, and design considerations for supporting users during document-driven decision-making.
- The design of Criticality, a system that operationalizes the Paul-Elder Critical Thinking framework into an interactive human-AI decision-making interface. This includes interactive techniques for: i) structuring, assessing, and guiding LLM outputs into reasoning traces, ii) retrieving evidence with explanatory links to show how evidence supports or contradicts reasoning steps, and iii) providing affordances for users to engage with reasoning traces.
- Qualitative insights from a within-subjects study (n=13) suggesting that scaffolding reasoning traces with interactive affordances, critical-thinking-based structure and guidance, and evidence links enhance the human-AI decision-making experience.

2 Related Work

Our work builds on prior research studying how people make decisions, especially using critical thinking and evidence-based reasoning, and LLMs and systems built to support these workflows.

2.1 Supporting Decision-Making Workflows

Research in decision-support systems has long examined how people use information to make decisions, revealing common workflows and needs. Simon [89] divides decision-making into phases: (i) identifying issues and collecting information, (ii) developing alternative options, and (iii) evaluating these options [69]. Many decision-supporting systems try to automate the process and present final decisions that nudge users to comply (e.g., employment hiring [49], loan approval [86], investment advice [55]). However, automation deprives users of the opportunity to develop decision-making strategies [33], while fostering inappropriate user reliance due to a disconnect between AI output and user reasoning [81]. External factors such as time pressure increase AI dependence and decrease deep cognitive engagement [94], underscoring the need for tools that foster more agency and cognitive engagement. Previous decision-support systems have assisted parts of these stages. To support identifying issues and collecting information, systems have been designed to suggest concepts [72], problem frames [73], and data [38, 90] to mitigate the challenge of under-specified, ambiguous, or biased user requests. For later stages, research has focused on externalizing decision options and criteria into decision matrices that help users compare alternatives more effectively [32, 51, 52]. Criticality is designed to support users throughout their decision-making process.

2.1.1 With Evidence-Based Reasoning. Providing evidence and explanations that reveal a model's reasoning has been shown to strengthen decision-making by improving understanding, uncertainty awareness, and trust calibration [29, 66, 103]. Further research demonstrates that the presentation of evidence in various forms, as visual layouts [112], logical structures [14], deliberation [56], and interactive affordances [101], strongly shapes users' trust and reliance on AI. Building on this work, Criticality advances evidence-based reasoning by focusing on the presentation of evidence. Instead of merely displaying an explanation, it systematically links each reasoning claim to specific supporting or contradicting passages in data reports. It then uses strong affordances, such as visual indicators and direct links, to encourage users to examine this evidence. This is designed to make evidence evaluation a seamless part of the workflow, moving beyond fragmented verification towards evidence-based reasoning.

2.1.2 With Critical Thinking. Critical thinking is a disciplined, reflective process of rationally analyzing and evaluating information to guide belief or action [27, 45], which involves questioning assumptions, considering multiple viewpoints, using logic and evidence, and reflecting on the reasoning process [1, 12]. Among various critical thinking models [7, 30, 77, 98], the Paul-Elder framework [77] is particularly well-suited for operationalizing in computational systems because it provides a clear, structured mapping between the Elements of Thought (the components of reasoning Table 1 (left)) and evaluating its quality using Intellectual Standards (criteria for evaluating reasoning quality Table 1 (right)). Other models only introduce limited components (three for Ennis [30] and six for Bloom [7]) or lack the quality assessment (Toulmin [98]).

Recent HCI research has also begun to shed light on the importance of fostering critical thinking during knowledge work. Textual critiques and provocations can enhance critical thinking in AI-supported knowledge tasks [25], while structured group discussions that challenge assumptions improve decision quality [20, 48]. Similarly, AI chat-bots employing Socratic questioning techniques cultivate deeper reflection and critical thinking [26, 31, 57, 71]. Building on these insights, Criticality structures reasoning traces into interactive components based on the Paul-Elder framework's

Table 1: Paul-Elder Critical Thinking Framework [77]: Elements and Standards for Data-Driven Decision-Making

Element	Definition	Standard	Assessment Criteria
Purpose	The overarching goal directing the reasoning process.	Clarity	Evaluates comprehensibility and precision of reasoning elements.
Question at Issue	The specific problem and sub-problems requiring resolution.	Accuracy	Assesses factual correctness and empirical validity of claims.
Assumptions	Underlying beliefs and presuppositions supporting the reasoning process.	Precision	Evaluates the specificity and detail of reasoning components.
Point of View	The perspective from which the reasoning is made.	Relevance	Assesses the degree to which reasoning elements contribute to addressing the central question.
Concepts	Theoretical constructs, definitions, and principles shaping analytical frameworks.	Depth	Evaluates whether reasoning adequately addresses inherent problem complexity.
Data / Evidence	Factual foundation supporting reasoning processes.	Breadth	Assesses comprehensiveness of perspective and consideration of alternative approaches.
Inferences	Logical conclusions derived from available evidence.	Logic	Evaluates the internal consistency and validity of inferential processes.
Implications	Both intended consequences and potential unintended effects of proposed decisions.	Significance	Assesses whether reasoning focuses on the most consequential aspects.
		Fairness	Evaluates reasoning for bias, self-interest, and adequate stakeholder consideration.

Elements of Thought, and offers interactive guidance to improve the quality of thinking in each element.

2.2 LLMs, Conversational Interfaces and Limits of Human-AI Decision-Making

LLMs started as next-token predictors but are now evolving into systems exhibiting human-like reasoning capabilities [53, 105], enhanced by stepwise prompting strategies like Chain-of-Thought [106] and Tree of Thought [113]. Recent reasoning models (e.g., OpenAI's GPT o-series [67], Anthropic Claude [2], and Google Gemini Pro [21]) generate intermediate steps in a reasoning trace before producing a final answer. While this improves performance on benchmarked tasks, LLMs remain fundamentally optimized for fluency over factual accuracy [39], often hallucinate [24], fail at complex or unfamiliar reasoning tasks [64, 88], and amplify cognitive and data biases [28]. Users' role is also limited as passive readers: they wait as the model processes the reasoning chain, which grows with task complexity [79, 117]. This limits users' opportunities to inspect, steer, or repair reasoning [80, 109], where reasoning traces frequently fail to reflect the model's actual logic or critical evaluation, reducing transparency and reliability [13, 18, 118].

Most users interact with LLMs through conversational interfaces, which are intuitive but limited for decision-making tasks [16]. Initial prompts are often underspecified, typically requiring multi-turn repair to align on intent [92, 116], and users tend to trust chat responses more than traditional search results even when quality is comparable [93, 115]. HCI theory emphasizes that high automation requires high human control [87], yet conversational interfaces provide minimal agency. Recent interfaces reveal intermediate reasoning: AI Chains [108] and Stepwise/Phasewise [43] enable interactive task decomposition, DirectGPT [58] allows direct manipulation of generated text, and many systems expose step-by-step reasoning traces via bullet points, tree structures, or interface components [46, 75, 99].

Criticality builds on these efforts by structuring human-AI reasoning into explicit steps that users can steer and edit during the process. It replaces generic reasoning traces with a rigorous pedagogical framework that assesses and guides specific cognitive behaviors of both the model and the user. It further evolves the "human-in-the-loop" paradigm by not only spotting errors, but auditing each claim and reasoning step in retrievable evidence, and intellectual standards. Unlike reasoning traces in current conversational interfaces and reasoning models, it enables users to interactively steer, inspect, and verify each step, fostering critical, evidence-based reasoning and collaborative human-AI decision-making rather than black-box automation.

3 Formative Study

We conducted a formative study with professionals who routinely reason with reports often combining text, tables, and visualizations. Our aims were to (i) understand current workflows and challenges, (ii) elicit design considerations for LLM-assisted decision-support systems, and (iii) characterize how people evaluate the AI reasoning quality. See supplemental material for detailed examples of interview materials.

3.1 Participants

Using purposeful sampling [74], we recruited five participants (P1–P5) by posting on three Slack workspaces frequented by business executives and analysts, professionals who often reason with and make decisions based on data and reports. Participants represented diverse backgrounds: gender (two women and three men), from five domains (banking, marketing, consulting, construction, and finance) who routinely make or influence decisions, with an

average of 14 years (8-20 years) of experience. Roles included VP of Business Intelligence, Analyst, Consultant, Enablement Lead, and Marketing Partner. All had previous experience using conversational interfaces of LLMs over the past year.

3.2 Procedure & Analysis

This study was conducted in accordance with the internal research policies of the authors' affiliated organization, an anonymized company. All participants provided informed consent, and data handling practices adhered to the company's ethics, privacy, and confidentiality standards.

3.2.1 Semi-Structured Interview. (20 mins) Each interview followed a semi-structured guide. Participants described their background in decision-making, walked through a recent example, and discussed how they ensured reasoning quality and trust. We also explored their prior experiences with LLMs, desired AI roles, and where AI support would be most beneficial.

3.2.2 Participatory Design Exploration. (15 mins) We then facilitated a participatory design exercise around interfaces for human-AI decision-making. We initially presented participants with a user scenario of making an investment strategy and walked them through low-fidelity Figma prototypes showing variations in reasoning traces, guidance, and layout. They annotated designs, suggested alternatives, and discussed how each supported their workflow. Interviewers also posed user scenario questions to prompt deeper reflection and design feedback.

3.2.3 Evaluation of AI Reasoning Traces. (10 mins) Participants rated a sample of LLM reasoning traces using the Paul–Elder framework [77]'s intellectual standards (Table 1(right)) on a 5-point Likert scale. They thought aloud about their evaluations and what improvements could raise each rating.

3.2.4 Analysis. Each session was transcribed and thematically analyzed. We iteratively clustered themes along workflow strategies, user challenges, and co-design feedback.

3.3 Findings

3.3.1 Workflows & Challenges. We observed four recurring parts to the workflow with corresponding pain points:

[C1] Scoping & alignment (3/5; P2,P3,P4). When collaborating as a team, people iteratively clarify goals, stakeholders, and constraints; with LLMs, this becomes slow prompt–response repair. *"The most difficult part is to understand their requirement... that's the key."* (P3) *"I ask a lot of questions... to really understand them."* (P1) Participants wanted proactive clarification *"I want it [LLM] to act like a concierge. Ask clarifying questions with options. Solving the wrong problem faster is still the wrong problem."* (P5).

[C2] Referencing behavior (2/5; P1,P5). Decisions rely on inspecting original data sources (dashboards, sheets, slides). For example, P2 noticed that for a conference *"VP registration is down 46% ... here are the accounts with VPs that haven't registered yet [to reach out to with the next marketing message]." "KPI signals show this [construction] job's got issues... that drives what resources we need to allocate and change today."* (P4) However, linking claims to precise anchors (cell, filter state, figure) is tedious, and conversational interfaces cite reports but not exact locations, undermining trust in data and stakeholder persuasion. As P5 said, *"I want the reference. If I uploaded a PDF, show me exactly where, which table and which data point."*

[C3] Interacting with long outputs (2/5; P1,P2). Traditional workflows, as well as those with LLMs, can have multiple artifacts generated over the course of multiple back-and-forth messages. This can be verbose and meandering in terms of context carryover as well as task focus. Users struggle to navigate and make sense of this: *"This feels like scribbled notes, not thoughtfully written and clear."* (P2) *"I don't know what I'm reading... it feels disconnected from what I need"* (P3)

[C4] Considering alternatives & assumptions (3/5; P1,P2,P5). Participants need visible assumptions, uncertainty, and explicit option comparisons across criteria to judge quality and defend choices. As P5 said, *"How do I know it's really 34–42%... how did that number come to exist? Is it hallucinating?" Provide caveats and what's beyond scope; suggest who to talk to."*

3.3.2 Preferences and Challenges when using the Critical Thinking Framework for Structuring, Assessing, and Guiding Reasoning Traces. When interacting with low-fidelity prototypes where the reasoning trace was represented as the elements of thought in Paul-Elder's critical thinking framework, all five participants preferred it to the reasoning traces they had previously experienced in conversational interfaces to LLMs. Participants described the framework as "a clear map of how the model is thinking" (P2) and "something I can follow, not just a wall of text" (P1). *"It's covering all the important aspects, making sure we don't have any blindspots or biases."* (P2)

All participants also favored integrating intellectual standards–based assessments and guidance directly within each reasoning element, rather than displaying feedback in a separate panel. *"If it's embedded right there, I can fix it on the spot instead of hunting for what it's talking about."* (P4) In terms of guidance style, four of the five preferred receiving guidance as *direct action recommendations. "I like when it gives me a concrete next step instead of a question. It saves time and feels like it's collaborating.* (P4) *"Actionable feedback is way more helpful than asking me vague why questions."*(P1) One of them preferred Socratic questioning. *"Sometimes I want it to prompt me to think deeper on why I made that assumption or choice, and I can address it how I want."* (P3) while none preferred receiving provocations, describing these styles as *"confrontational"* and *"unhelpfully vague"*.

However, participants also identified limitations. Some participants desired evidence-based reasoning support to know whether or not each reasoning step was grounded in the source report. *"I love the structure, but how do I know it's not making this up? Show me where that's from in the report."* (P1) Similarly, P5 added, *"If I upload a report, I want to see exactly which table or paragraph supports that claim."* Another challenge was wanting increased interactivity to steer and repair each element: *"This is great. I love how it breaks down the reasoning, but I want a way that I can step in and correct it."* (P3) Participants also wanted support for an executive summary

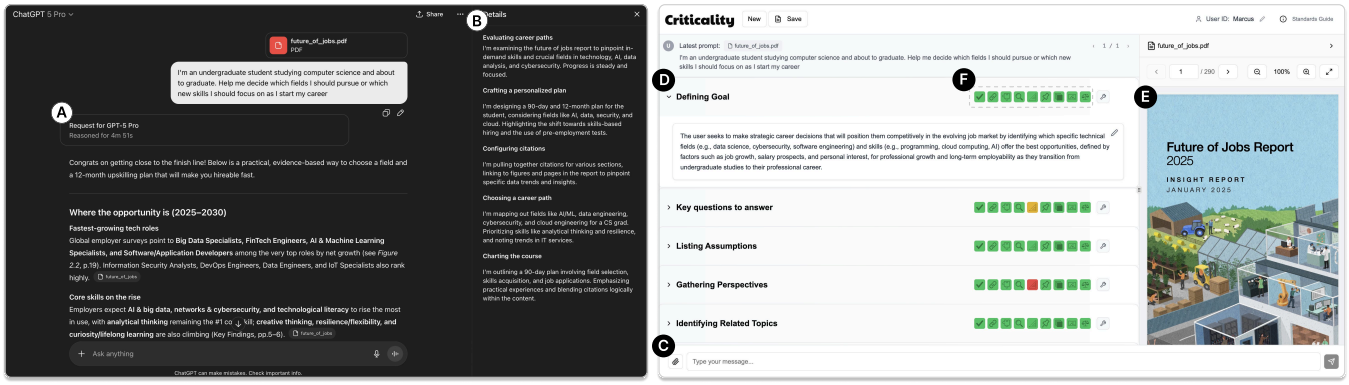


Figure 1: (left) ChatGPT 5 Pro, where the reasoning process took about 5 minutes (A) and produced a reasoning trace with limited interactivity (B). (right) Criticality with the same prompt and report (C), incrementally produces an interactive reasoning trace (D). Reasoning is based on evidence retrieved from the report, where clicking on an evidence link navigates to the exact passage in the report panel (E). The intellectual standards' heatmap strip shows the quality of reasoning in that element (F).

that distilled the reasoning trace into an output they could present to their stakeholders and collaborators. *"Give me a summary that keeps the evidence links. I need to defend my decision later to my team."* (P4) *"At the end of the day, I need something I can show my VP—a clear, defendable takeaway."* (P3)

3.3.3 Evaluation of Reasoning Traces. We also collected the ratings and the underlying reasoning from the participants for the LLM reasoning traces and converted them into a few-shot prompts used to score the decision-criterion pair. The detailed formulation of the scores is described in Section 4.3.4 (see Supplementary Materials for few-shot prompts).

3.4 Design Considerations for Human–AI Decision-Making Systems

Grounded in these findings and in related work, we list our core design considerations:

[D1] Structure reasoning into transparent, interactive steps that support early alignment and iterative steering. [C1, C3] [16, 79, 87] LLM-assisted decision-making systems should proactively clarify user goals through early and continuous alignment loops. Reasoning should be externalized into structured, inspectable elements that users can iteratively clarify, correct, or refine. Systems should also support non-linear navigation and preserve traceability of edits so users can understand how earlier steps influence final outcomes.

[D2] Ground reasoning in verifiable, linked evidence, maintaining transparent claim–evidence traceability. [C2] [14, 29, 101, 103, 112] To strengthen trust and accountability, each claim should be linked specific evidence and indicate whether that evidence supports, contradicts, or remains neutral.

[D3] Embed ongoing, adaptive guidance to improve reasoning quality and reflection. [C4] [13, 25, 40, 48, 118] Guidance grounded in intellectual standards (e.g., clarity, logic, fairness) should appear within each reasoning step, combining actionable recommendations with reflective prompts. To help users monitor and improve reasoning quality throughout, systems should visualize this at each step.

[D4] Summarize decision and rationale for communication [C1, C4] [44, 63, 70, 85]. The decision-making goal, explored alternatives, underlying rationale, supporting evidence, and any trade-offs considered should be summarized in a report. To support audit-ability during stakeholder communication, the report should preserve interactive evidence links and surface key assumptions, trade-offs, and implications, enabling downstream users to quickly inspect and verify reasoning.

4 Criticality

Following the design guidelines derived from the formative study, we developed the Criticality system.

4.1 User Scenario

Consider Marcus, a senior undergraduate student studying computer science, exploring which specific field to pursue and what skills to develop for his future career. Wanting to base his decision on credible insights, he consults the World Economic Forum's Future of Jobs Report [107], a 290-page report aggregating the perspectives of over 1,000 global employers, across 22 industry clusters and 55 economies, outlining workforce transformations, emerging job categories, in-demand skills, and industry disruptions through 2030, and presenting these insights as visualizations, infographics, data tables, and text.

Wanting to make a sound career decision, Marcus attempts to use ChatGPT-5 Pro (Figure 1(left)), which offers multi-step reasoning and promises "research-grade intelligence". After prompting, he has to passively wait and watch as the model reasons step-by-step for many minutes. Upon receiving a lengthy output, he feels overwhelmed [C3]. He is unsure if it considered all alternatives, diverse perspectives, and implications of different options, but the information overload forces him to focus on making sense of the

output instead of critically engaging with the content [C4]. Wanting to trace how the output is based on the report content, he clicks on the in-line citations, but they just reference the entire report, leaving him unsure about the model's evidence base [C2]. Reading the output, he realizes that he needs more specific job descriptions, so he revises his prompt to include his major as criteria [C1].

Feeling frustrated with his limited agency, and not wanting to wait longer, Marcus switches to Criticality (Figure 1(right)). He enters the same prompt and report. Within seconds, Criticality processes the prompt and report, and displays its interpretation of his *Goal*, allowing him to verify and edit for alignment before proceeding, saving time and tokens on irrelevant subsequent steps. A red icon in the *intellectual standards' heatmap strip* indicates a lack of *Precision* in his Goal. Hovering reveals a suggestion to specify the roles, careers, or industries he is targeting (Figure 2). Marcus clicks *auto-fix*, reviews the proposed edit that adds 'technology roles emphasizing leadership, continual learning', and finding that it aligns with his context, approves it. The icon turns green. Satisfied with the refined goal, he clicks the check mark to continue.

Following Criticality's reasoning trace, Marcus reaches the *Listing Assumptions* step, which lists the LLM's assumptions (Figure 3). Some are initially toggled off due to contradictory evidence in the report. Curious why the assumption 'Technical skills can be acquired and mastered within...' is toggled off, Marcus clicks on the retrieved evidence classified as weakly-contradicting, which takes him directly to that visualization in the report. Thinking that the claim should reflect his current situation rather than the report, he activates the assumption so that it can be included in his and the LLM's reasoning process.

Further along Criticality's reasoning trace, the *Comparing Options* element presents a *decision matrix* (Figure 4) of career options and criteria derived from previous elements in the reasoning trace. Marcus adjusts criteria weights via sliders, and the system dynamically re-ranks options to highlight those best aligned with his preferences. Wondering why 'AI/ML Specialization' scores low on 'Skill Acquisition Speed', he inspects its scoring breakdown and sees that contradictory data reduced the score's confidence (Figure 4D).

Subsequent elements in the reasoning trace, *Insights & Takeaways and Implications* (Figure 5), help Marcus reason through the consequences of each choice. The former distills key insights from the matrix, flagging those with quality issues that Marcus can toggle off; the latter outlines short-, mid-, and long-term implications, helping him understand each option's broader outcomes.

After aligning on every element in the reasoning trace, Criticality presents an *Executive Summary* (Figure 6), condensing the reasoning process into a concise recommendation, a rationale with trade-offs with alternatives and evidence. Clicking on highlighted keywords lets Marcus trace each suggested decision back to its evidence from the report. Reviewing the summary, Marcus feels confident in his decision and the reasoning behind it and begins planning his next steps toward achieving his career goals.

4.2 System Features

Incorporating the design considerations from the formative study, Criticality provides four core features that support critical reasoning and the data-driven aspect of decision-making tasks.

*4.2.1 Elements of Thought **[D1]**.* According to Paul-Elder's framework, all reasoning can be decomposed into *elements of thought*, which form Criticality's reasoning trace. As there is a natural order in which some elements are clearly defined before others (e.g., purpose precedes consequences), we sequentially generate the elements of thought following the logical order in [77]. Users can edit each element to ensure alignment and quality of reasoning. For instance, users can toggle one of the *Assumption* element by clicking on the toggle button, or edit using the edit pen (Figure 3C). In contrast to the typically inaccessible, immutable, and non-aligned internal reasoning traces of LLMs [18], Criticality generates reasoning traces through conversation, making them accessible, steerable, and well-aligned with the overall context.

*4.2.2 Heatmap and Guidance System Based on Intellectual Standards **[D3]**.* On every element, a heatmap strip of nine icons providing an overview of the element's quality of thought assessment based on nine intellectual standards (Figure 1F). Red indicates major areas needing improvement, yellow signals minor issues, and green shows all content meets the standard.

Hovering over any heatmap indicator reveals targeted feedback specific to that intellectual standard and element's content(see Figure 2). Notification pills above the element's content blocks highlight the most critical areas and hovering over them displays actionable guidance to support critical thinking.

*4.2.3 Evidence-based Reasoning **[D2]**.* To ensure reliable and trustworthy chains of critical reasoning, we implement direct linking between evidence and elements of thought. Criticality retrieves relevant report snippets and systematically evaluates whether they support or contradict each assumption, point of view, inference, and implication within the reasoning elements. The result is presented as a thin stacked bar chart, which represents the distribution of evidence, notifying users about the data document's level of support for the element (Figure 3A). Evidence is shown on a five-level scale: dark green for explicit support, light green for implicit support, gray for neutral, light red for implicit contradiction, and dark red for explicit contradiction. In the interface, "explicit" and "implicit" were simplified to "strong" and "weak" for improved clarity.

In addition to the stacked bar chart, Criticality also offers guidance through pre-filtered selections. When contradicting evidence is detected in the data or when the number of supporting evidence is insufficient, that element is automatically deselected by default and shown with low opacity (Figure 3A). Also, all elements are reordered based on the evidence supportiveness, where those strongly supported by the data rise to the top. This process guides the user, ensuring that only evidence-backed results are passed to generate the next element, thereby maintaining the integrity of the reasoning chain.

*4.2.4 Decision Matrix **[D1, D2]**.* The decision matrix (Figure 4) is presented as the sixth element, allowing users to compare and evaluate possible decision options, following the design of prior

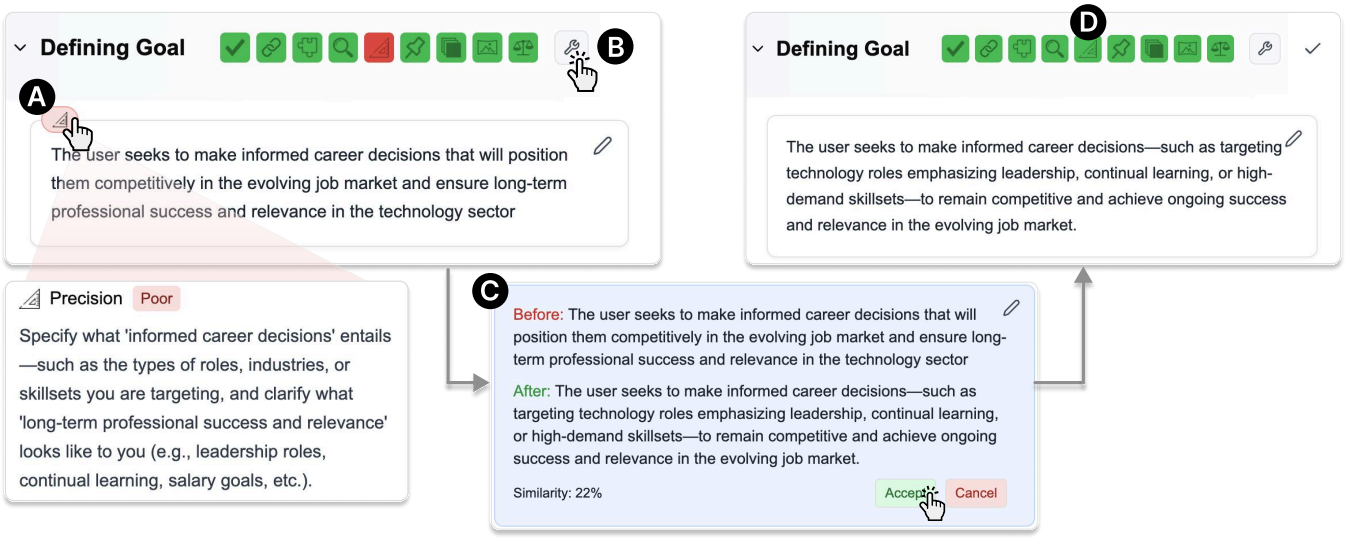


Figure 2: The guidance-based content auto-fix interaction. Hovering over the red icon, a guidance snippet appears as a tooltip (A). On pressing the auto-fix button (B), Criticality suggests the fixed content and contrasts it to the existing content (C). When accepted, the system re-evaluates the quality of the fixed content (D).

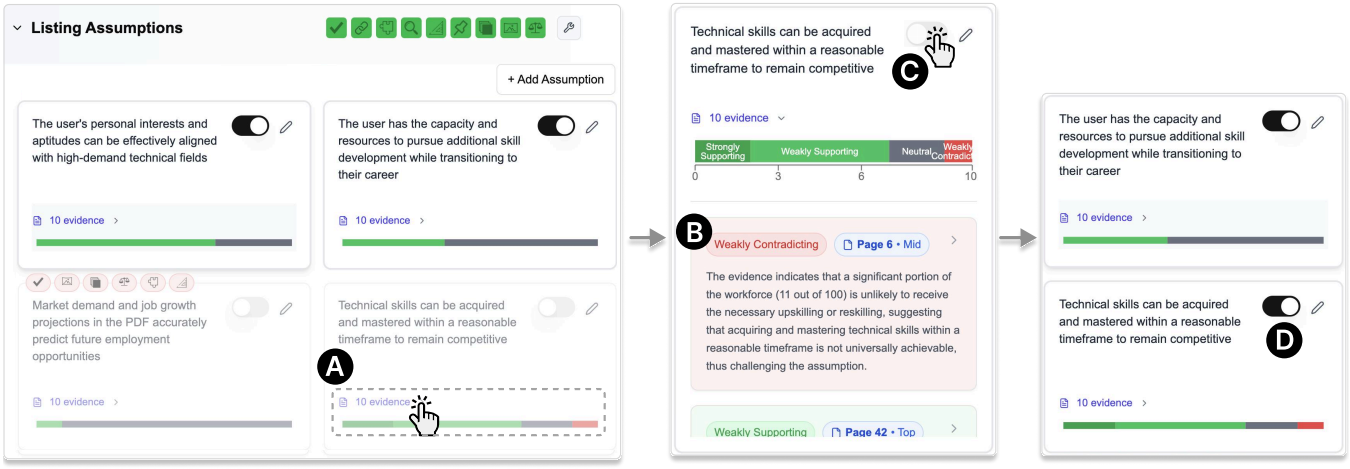


Figure 3: Assumptions having insufficient supporting evidence from the data are toggled off by default shown as low opacity (A). The user can inspect the individual evidence (B) and override the decision based on the assumption and decision-making context (C). Only the contents that are toggled on are used to generate the next elements (D)

works [50, 69] that provided interactivity and affordance. Each cell of the matrix corresponding to the decision option and criterion pair is colored according to a heatmap, representing the score of how advantageous the decision option is in terms of the criterion. When the user clicks on a specific cell, the score distribution view (Figure 4D) appears, allowing the user to observe how the score is calculated based on the evidence and intellectual standards of critical thinking. The detailed composition and formulation of the score are described subsequently in Section 4.3.4.

Each criterion has a relative weight value [1-10], reflecting the user's context and preference for prioritizing options. As the user modifies the weight with a slider, the rows are dynamically reordered so that the most relevant and significant criteria appear at the top. At the same time, the overall score for each decision option is calculated as a weighted sum of all scores on each criterion, where the options with higher scores are sorted to the left.

*4.2.5 Executive Summary **[D4]**.* After going through the reasoning trace elements, Criticality generates an *Executive Summary* that synthesizes the reasoning trace into a concise argumentative narrative. It integrates key insights from each element of thought and highlights the highest-scoring decision options from the decision matrix. The summary presents structured reasoning in a format suitable for stakeholder communication while maintaining

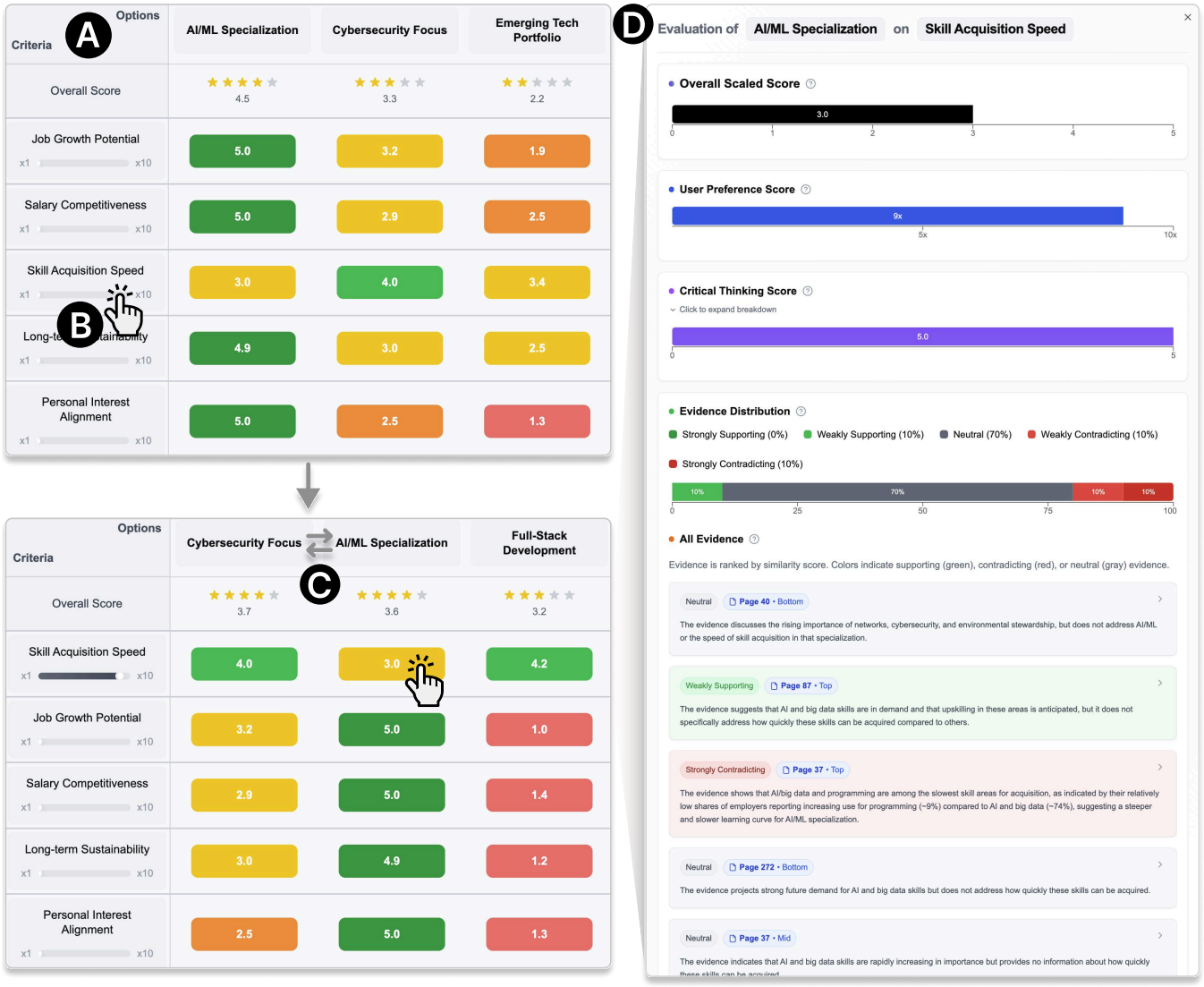


Figure 4: The Decision Matrix (A) shows the decision options and criteria to compare those. Users can adjust the criteria importance with the slider (B), which automatically reorders the matrix layout (C). Selecting a score cell opens the score breakdown modal (D), which shows the decomposition of the overall score into scores of user preference, critical thinking, and evidence distribution.

evidence links for transparency. Users can click references within the summary to navigate back to specific reasoning elements or report sections, maintaining full transparency and auditability of the decision-making process (Figure 6A). The executive summary thus serves as both an entry point for a deeper exploration of the reasoning trace and a deliverable for shared understanding among stakeholders.

4.3 System Architecture

CRITICALITY operationalizes the Paul-Elder Critical Thinking Framework [77] to scaffold AI-assisted decision-making through a sequential processing architecture described in Figure 7. The system accepts a user prompt and a report as input. The *Document Indexer* processes the input report to enable evidence retrieval throughout the reasoning process. The *Reasoning Trace Generator* then produces Elements of Thought in logical sequence (purpose, key questions, assumptions, perspectives, concepts, decision matrix, inferences, and implications, allowing users to intervene and steer the reasoning before subsequent elements are generated. As each element is produced, two parallel processes evaluate its quality: the *Intellectual Standard Evaluator* assesses reasoning against nine dimensions (clarity, accuracy, precision, relevance, depth, breadth, logic, significance, and fairness) and displays visual indicators as heatmap strips for potential issues (Figure 1F), while the *Evidence Retriever* searches indexed reports for related content and classifies passages as supporting, neutral, or contradictory to generated

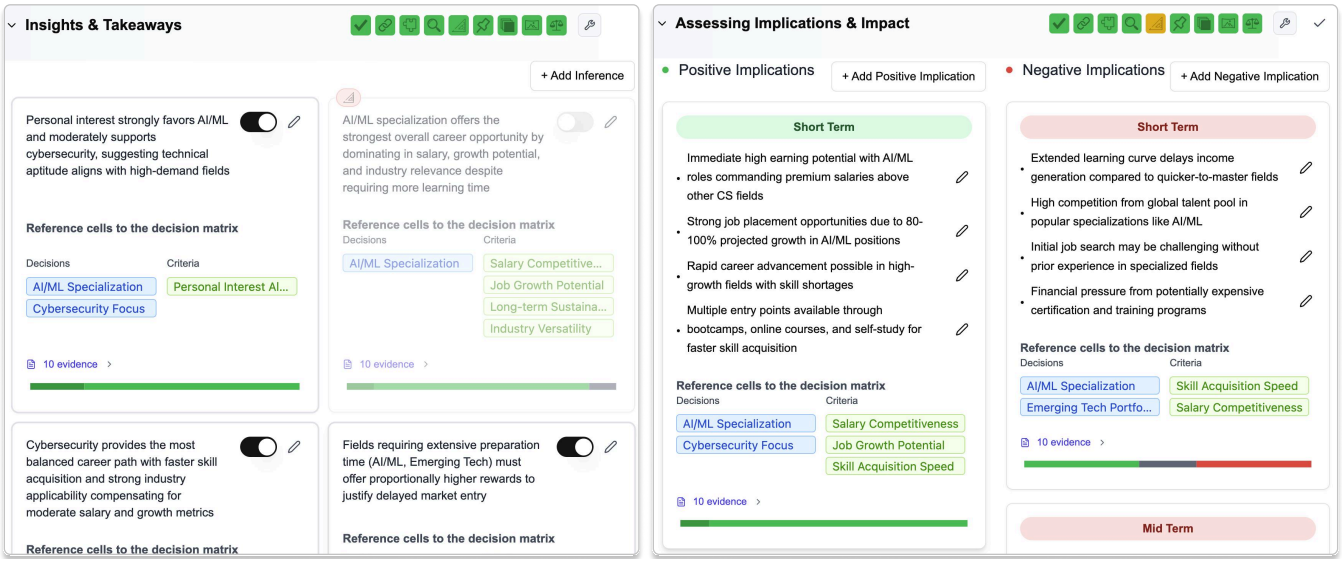


Figure 5: (left) Insights and Takeaways show possible inferences derived from the decision matrix. (right) Implications outline potential consequences organized by the temporal scope (short, mid, and long-term)

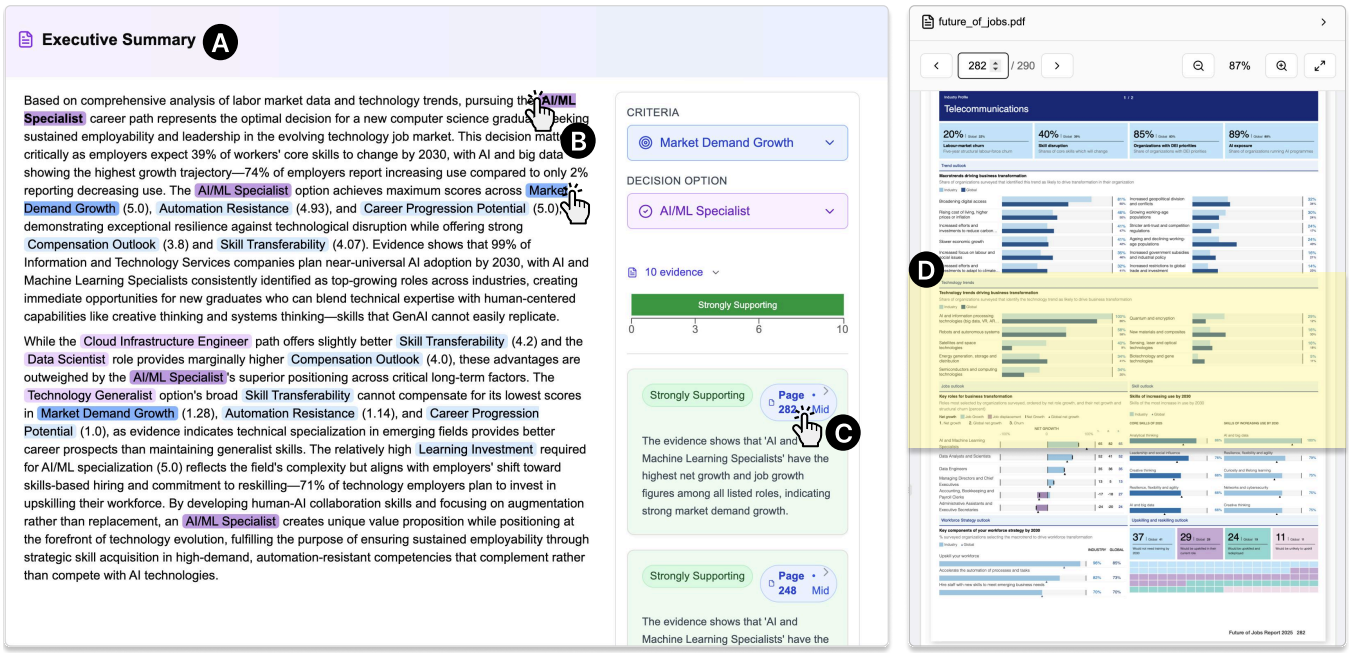


Figure 6: The Executive Summary (A) consolidates reasoning across all generated elements. Users can click keywords or use the right panel dropdown to view detailed evidence supporting or contradicting each decision option against the selected criterion (B). Clicking any evidence in the list (C) opens the Evidence Viewer (D), which highlights the relevant section in the data report.

claims. These evaluations inform a Guidance and Recommendation Loop that provides actionable suggestions for improving reasoning quality and presents evidence distributions for user validation. Users maintain full control to accept, modify, or reject any generated content throughout this iterative process. After passing through all elements of thought, the system concludes by generating an Executive Summary that synthesizes the critical thinking elements into a cohesive decision recommendation supported by evidence and acknowledged trade-offs. Please refer to the supplemental materials for detailed prompts, including a few-shot example derived from the formative study.

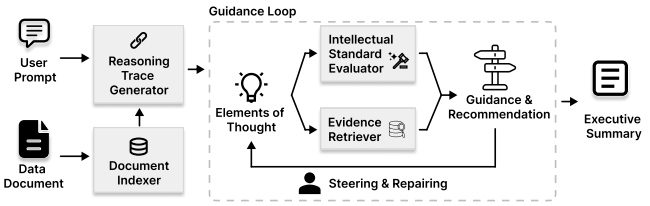


Figure 7: The system architecture overview. The system accepts user prompts and reports, indexes the data, and generates Elements of Thought sequentially through a Reasoning Trace Generator. Then, it evaluates each element via parallel quality assessment (Intellectual Standards) and evidence retrieval processes. Users can steer and repair the reasoning at each step through an interactive guidance loop before the system produces a final Executive Summary.

4.3.1 Document Indexer. When a user submits a report, it undergoes automated processing through a document parser (LlamaParse [54]) that converts the file into structured markdown format. The parser preserves page numbers from the original report to enable evidence-linking while maintaining inter-page context for multi-page visualizations or tables. Visualization elements are converted into data tables to help with indexing and retrieval. The parsed content is then chunked into manageable segments based on its semantic structure and the constraints of the embedding model. These chunks are converted into embeddings and indexed for efficient retrieval during the reasoning process.

4.3.2 Evidence Retrieval & Evaluation. To ensure reliable critical reasoning, each generated assumption, point of view, inference, and implication must be grounded in verifiable evidence from reports **[D2]**. While the LLM processes document context to generate reasoning elements, we implement a pipeline to verify that supporting evidence exists and assess its strength. When the LLM generates a reasoning statement, Criticality uses it as a semantic query to retrieve relevant evidence from the document corpus. The statement is converted into a text embedding using OpenAI's text-embedding-3-large model, then matched against indexed document embeddings using cosine similarity to retrieve the top-k most relevant snippets (k=10). Each retrieved evidence snippet is evaluated by a separate LLM that classifies the relationship between evidence and claim into five categories: [explicitly supports, implicitly supports, neutral, implicitly contradicts, or explicitly contradicts]. This assessment captures nuanced relationships, enabling users to gauge the strength of evidence that supports their reasoning and access relevant report sections directly for details.

4.3.3 Guidance Generation & Automated Fix. Each element of thought content (shown as gray blocks in Figure 8 undergoes quality assessment against the nine intellectual standards of the Paul-Elder Framework. We adopted an LLM-as-a-judge approach where the assessment criteria are clearly embedded in the prompt to increase reliability [35, 111]. When the evaluator LLM processes an element's content block, it returns a binary evaluation ('poor' or 'acceptable') for each standard along with specific guidance on content improvement. This creates a direct feedback loop, allowing users to identify weaknesses and receive actionable recommendations **[D3]**.

As users modify content based on guidance, the revised text is automatically re-evaluated by the LLM to verify that changes address the identified issues. This iterative process ensures that user edits align with critical thinking principles and that quality improvements are measurable. When multiple guidance items require attention simultaneously, users may experience cognitive overload from managing numerous improvement suggestions. To address this, we provide an auto-fix feature (Figure 2) that consolidates all guidance recommendations and automatically generates revised content addressing the identified issues. Users retain control by reviewing the proposed changes and choosing to accept or reject the modifications, maintaining human agency while reducing cognitive burden in complex reasoning scenarios.

4.3.4 Automatic Scoring of Decision Matrix. Once the decision criteria and options are established, Criticality initially generates cell scores (S_C) for each matrix cell by combining two complementary assessments: critical thinking quality and evidence strength (see Figure 4E).

The **critical thinking score** $S_{CT}(i, j)$ evaluates how well decision option i can be justified against criterion j using the nine intellectual standards. We employ an LLM-as-a-judge approach, where the evaluator produces scores from 1 to 5 for each standard given the rubric, including the examples crafted from our formative study. The individual standard scores are averaged to produce the critical thinking score.

The **evidence score** $S_E(i, j)$ quantifies how strongly the reports support each option-criterion pairing. Our system retrieves the top-k relevant evidence snippets and classifies their relationship to the claim into five levels: explicitly supporting (+2), implicitly supporting (+1), neutral (0), implicitly contradicting (-1), or explicitly contradicting (-2). These classifications are averaged across retrieved evidence.

Both scores are normalized to a 1-5 scale using criterion-level min-max scaling, then averaged to produce the cell score (S_C)

$$S_C(i, j) = \frac{1}{2}\left(\text{norm}(S_{CT,}(i, j)) + \text{norm}(S_E(i, j))\right) \quad (1)$$

The **overall weighted cell score (S)** for each option-criteria pair is calculated as a weighted average across cell scores:

$$S(i, j) = \frac{\sum_j w_j \cdot S_C(i, j)}{\sum_j w_j} \quad (2)$$

where w_j is the user-adjusted weight for criterion j ranging from 1 to 10. This approach ensures decision recommendations are both logically sound and empirically grounded in the available evidence.

4.4 Implementation Details

Criticality is built as a full-stack web application composed of a Next.js front-end [102] and a Flask-based Python backend [82]. The frontend is built with React 18 [62], utilizing Recharts for bar chart components [97]. The backend uses both OpenAI's GPT models and Anthropic's Claude APIs for LLM tasks, including structured text extraction and Elements of Thought reasoning through carefully crafted prompt templates. We used the Anthropic Claude 4 Sonnet

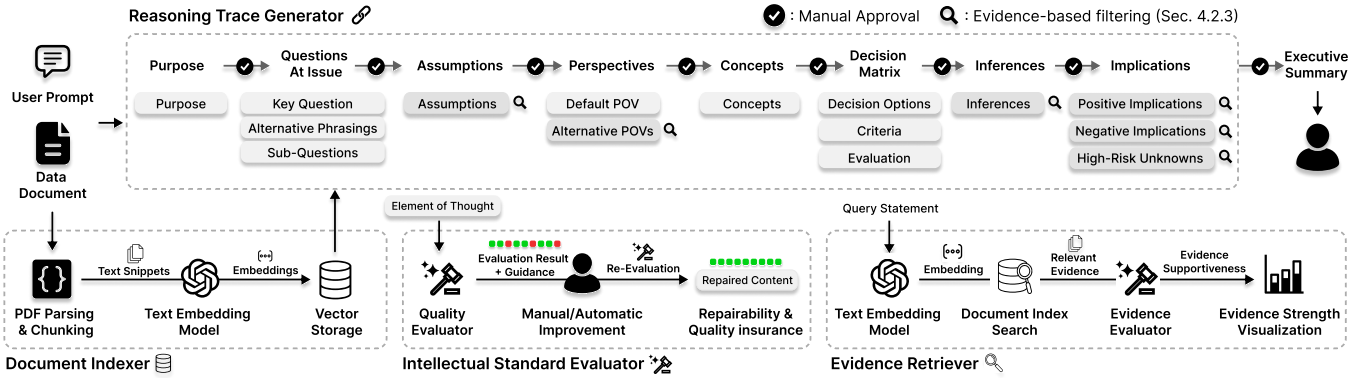


Figure 8: Detailed architecture diagram that expands Figure 7, illustrating how the prompt text and user-provided data are processed through our system. The Document Indexer parses, chunks, and stores the data document in a vector storage for evidence retrieval. The reasoning trace generator sequentially generates the elements of thought in a predefined order. The user iteratively reads and improves each block's content by referring to the guidance of the intellectual standard evaluator. At the end of the critical reasoning, an executive summary is generated that summarizes the elements and the content of the decision matrix.

model [2] without extended reasoning for generating elements of thought and the GPT-4.1 model from OpenAI [67] for the LLM-as-Judge evaluation phases to prevent self-preference bias [76, 104]. The Python FAISS library [41] is used for vector similarity search with OpenAI embeddings, enabling efficient evidence retrieval from the uploaded report. The backend implements asynchronous processing with Pydantic [23] to lower latency and ensure data validation. See supplemental materials for Criticality's source code, including prompts and rubrics for assessment and scoring.

5 User Evaluation Study

To investigate the effects of Criticality's scaffolds for human-AI decision-making, we conducted a within-subjects study with 13 participants making a decision using Criticality and making another decision using a state-of-the-art reasoning model embedded in a popular conversational interface. Specifically, we wanted to examine each system's effects on (i) user experience, (ii) interaction patterns and workflows, (iii) cognitive processes, (iv) trust and confidence in the decision-making process, as well as the blind-to-condition expert evaluations of decision rationale quality.

5.1 Baseline Condition

We used ChatGPT 5 Pro as our baseline condition because the model offered the strongest reasoning performance and was embedded in the popular LLM conversational interface ChatGPT. Also, it provides the model's internal reasoning trace, like Criticality, but without logical structure, user interaction, and explicit evidence linkage. Our formative findings and prior research indicate that flexible conversational LLM interfaces are already widely adopted by decision-makers, making this a practical and representative baseline. To ensure a controlled comparison, we did not enable web search or Deep Research modes, as these rely on external information retrieval rather than reasoning over the provided materials—potentially introducing variability unrelated to participants' decision-making processes.

5.2 Tasks

Participants completed two decision-making tasks—one using ChatGPT 5 Pro and another using Criticality. Each participant selected two of five available scenarios, adapted from recent industry and government reports, completing one scenario per condition to mitigate carryover effects. Each task took approximately 25 minutes. Participants were instructed to act as consultants preparing recommendations for stakeholders based on the provided reports. Their deliverable were a decision, their rationale, supporting evidence from the report, and consideration of trade-offs and alternatives.

Participants were allowed to freely choose their report below to encourage engagement and reflect personal interest. All reports were published within a year to ensure that the scenario is realistic.

(1) **HR Strategy**: Automation implementation decisions based on Bond Capital *AI trends* report [9]
(2) **Educational Strategy**: Curriculum development for emerging skills based on the Bond Capital *AI & Universities* report [8] and *AI and future of teaching and learning* report from the US Department of Education [100].
(3) **Career Planning**: Professional field selection informed by World Economic Forum *Future of Jobs* report [107]
(4) **Marketing Strategy**: Channel optimization decisions utilizing McKinsey *Sporting Goods 2025* report [60]
(5) **Investment Strategy**: Portfolio focus area determination based on McKinsey *Competition Arenas* report [61]

5.3 Participants

We recruited 13 participants (4 females) from three Slack workspaces frequented by business executives, analysts, and consultants, and university mailing lists. Participants represent diverse domains, including technology, finance, consulting, academia, and education. The majority (11/13) had over one year of experience using LLMs. In their professional work, most participants (10/13) used LLM-based tools daily or weekly, with the remaining three using them less than once a month. Regarding their use of data in making decisions,

(11/13) participants reported daily engagement with data or reports, while one reported monthly and one reported yearly engagement.

5.4 Procedure

This study was conducted in accordance with the internal research policies of the authors' affiliated organization, an anonymized company. All participants provided informed consent, and data handling practices adhered to the company's ethics, privacy, and confidentiality standards.

Before the main session, participants completed a pre-screening survey covering professional background, decision-making experience, frequency of LLM usage, and typical decision-making approaches in professional contexts. Among 71 submissions, we initially selected 20 participants based on various factors, including occupation, decision-making frequency, and LLM usage patterns. We prioritized business professionals over students and those who frequently use LLMs in data-driven decision-making. However, only 13 made it to their study appointments.

At the beginning of the study session, participants verbally consented to recording audio and screen; all study sessions were conducted over a video-conferencing call. Each session consisted of two decision-making task sessions, with condition order counterbalanced across participants to control for order effects. Each task session followed a standardized structure: participants first received a tutorial for the system (5 mins). Then they worked on the decision-making task (25 min) using the assigned system. Immediately following task completion, participants completed a post-task questionnaire (5-10 mins) assessing system usability, decision confidence, and perceived quality of critical thinking support.

Upon completion of both conditions, a semi-structured interview (5 mins) was conducted to elicit detailed reflections on user experience, strategic approaches to system utilization, and rationales underlying stated preferences. Throughout the experimental session, participants were instructed to verbalize their thought processes using a think-aloud protocol. All sessions were recorded using screen capture software for subsequent analysis and review. The complete session required approximately 75 minutes.

5.5 Measures

5.5.1 Interaction Patterns. We collected comprehensive interaction logs with millisecond-precision timestamps. For Criticality, we logged all interactions and analyzed them categorically as: (i) **Workflows** i.e., sequence and time allocation across elements of the reasoning trace; (ii) **Evidence actions**, such as examining the evidence stacked bar charts or navigating to parts of the source report from the reasoning trace; and (iii) **Steering actions**, such as guidance-based content edit or content selection toggle. For ChatGPT 5 Pro, we recorded all conversational interactions, such as a prompt formulation and refinement, clicking 'Details' to examine reasoning trace, and starts and stops to reasoning trace.

5.5.2 Self-Reported Scales. Post-task questionnaires assessed user experience using five-point Likert scales across three sections: (i) **Tool performance** measuring usability, learnability, interactivity, alignment, transparency, steerability, and repairability; (ii) **Cognitive effects** measuring perceived critical thinking support, confidence, and trust in system usage; and (iii) **Intellectual standards** based on Paul-Elder's framework [77], excluding Accuracy (covered in the confidence measures). Usability items were adapted from the System Usability Scale [6]. Five-point Likert formats are widely used in HCI user studies for self-reported usability, workload, and trust measures (e.g., SUS, NASA-TLX derivatives), as they reduce cognitive load, which can occur when non-expert respondents are asked to discriminate across finer scales, while maintaining consistency with established evaluation instruments [6, 11]. See supplemental materials for surveys.

5.5.3 Thematic Analysis. Our qualitative analysis included open coding, focused coding, and thematic clustering [15]. The two authors independently coded two randomly chosen study session videos through open coding. They discussed emerging themes and agreed on a common vocabulary. Once they identified similar codes and themes with no significant discrepancies, they finalized the coding scheme and shifted to a focused coding approach.

5.5.4 Quality of Decision Rationales. Two evaluators, decision-makers from the formative study, blind to condition, assessed the quality of participants' decision options and rationales using validated rubrics derived from the Paul-Elder framework. Responses were rated on 7-point Likert scales across: (i) Overall Critical Thinking, (ii) Evidence-based Reasoning, and (iii) Quality assessed with Intellectual Standards. The inter-rater reliability was 0.86 (Cohen's Kappa). A seven-point scale was used to enable finer distinctions among essays that may differ subtly across multiple criteria, particularly when raters are calibrating across multiple responses. Previous HCI studies suggest that increasing the number of scale points to seven can improve inter-rater discrimination without decreasing reliability when raters are guided by rubrics [11, 59].

6 Findings

We wanted to qualitatively understand the effect of interactive reasoning traces and the critical thinking framework applied in Criticality. We also introduce exploratory statistical results to suggest further work on generalizable results with sufficient sample sizes.

6.1 User Preference and Overall Experience

After using both systems, participants were asked about their preferences. 11/13 participants said they preferred Criticality's approach to scaffolding reasoning with evidence-based reasoning and critical thinking elements and standards, compared to the Baseline. The core feature the participants praised was transparency and control during the process and in the output. Although participants recognized that the Baseline's slow responses stemmed from its lengthy reasoning chain, they still desired more interactive feedback, clearer guidance on how to improve, and precise links to supporting evidence. As P5 put it, *"Criticality feels like AI that thinks with you, not for you."*

> *"Chat's [ChatGPT's] answer is not even close. And, the user isn't experienced enough to realize that. And for me this approach [Criticality] is much much better. If the user is inexperienced, he's still going to get a bad answer. But this [Criticality] is going to say, well I took these suggestions to say, "Hey, did you think about*

> *this? Is this important to you?" Even though it didn't say it in those words, but it's asking those questions back to me, and it's making me think, yeah, that's important or yeah not so important. So I think that's a good thing. I think that's a very good thing."* – P3

> *"Right now [ChatGPT] feels like a Burger King cashier. Where Burger King is have it your way. I get what I'm asking for but there's no back and forth. I want a concierge-type experience, where you go to the concierge and they say "Interesting, what's your budget? Are you walking or driving? Is it just you or is it a group? What kind of food do you like? Do you want local or chain?" It's that back and forth where you feel like you are part of that decision-making process, because you might not even think about some of the things that you should be thinking about."* – P4

> *"I feel like the industry is coming up with 40 different UIs for AI. And I think this one could be used in a lot of those, not just for decision-making. This flow is great. I want a standardized UI, even if I'm not working with data reports or making a decision, but say analyzing data, and I think it should be this one."* – P1

6.2 Interactive Steering and Repair

Participants' ratings (Figure 9) suggest that Criticality serves as a more *repairable*, *steerable*, *interactive*, *aligned*, and *transparent* system than the Baseline. Although the ratings themselves can not prove strict superiority over the baseline, they suggest that Criticality's core features successfully gave users more direct control during the decision-making process. This was achieved while maintaining comparable foundational usability. No significant differences were found between conditions for *Usability* and *Learning Curve*. This suggests that participants did not perceive higher effort, even though Criticality consisted of significantly more features and the Baseline's interface is popular, which indicates that the learning barrier for Criticality is minor and that the system's advanced features were added without sacrificing ease of use.

Participants were generally unsatisfied with the lack of interactivity, steerability, and repairability in the Baseline. (12/13) participants examined the reasoning trace by clicking on the 'Details' button. They could only steer and repair the conversation through prompts. They noted the process's lack of steerability and repairability, particularly when it involves the highest level of reasoning, which can take a long time. *"I'm left with a blank sheet of paper. If you want a better answer, ask a better question."* (P3) *"Wish I could change during the six minutes."* (P1) The only affordances to steer and repair were the stop button, which (1/13) participants used due to the long execution time, and none of the participants repaired their original prompt. (8/13) participants issued follow-up prompts to better understand the LLM's reasoning or restructure the output.

When using Criticality, all participants generally followed a sequential pattern from the purpose element to the decision matrix, as guided by the system (see Figure 10). However, some participants (e.g., P2, P8) frequently revisited previously generated elements after reaching the decision matrix. This illustrates the reflection centered on the decision matrix, which plays a key role in decision-making, by listing and comparing multiple decision options and criteria. Also, a few participants (e.g., P10, P11) showed a similar pattern at an earlier stage when the purpose and question at issue were generated. Reflection in the earlier stage could be a core strategy to better understand the goal and sub-questions that the user is currently targeting, which are often lost during long conversations with LLMs [22].

When using Criticality, participants attempted to edit most elements of thought, with the highest number of edits being made in the Assumptions and Purpose (Figure 11). A few participants attempted to add (2/13) or remove (1/13) content by pressing the 'Add' or 'Remove' button. The steering actions were done based on participants' preference or experience, where P4 edited the 'purpose' element directly to sharpen the goal *"Let's be bold… target salary over 200k... confirm."*, and used the add button within assumptions to capture a real-world constraint *"Add the assumption... the person is not willing to relocate... confirm it."* when working on deciding what job to pursue. This also appeared in the decision matrix, where (12/13) participants adjusted the sliders multiple times (M=7.46, SD=6.17) to match their preference.

Each element of thought is assessed for quality of critical thinking using Paul-Elder's intellectual standards (Table 1 (right)). While interacting on each element, participants frequently examined the intellectual standards strip (M=8.94, SD=13.41) while reading the provided suggestions (M=4.65, SD=7.64). They also repaired the element either manually (M=2.77, SD=2.78) or through the auto-fix feature (M=2.46, SD=3.0), while steering the context by toggling the contents (M=1.29, SD=1.00). Figure 11 shows the average number of manual modifications, automatic corrections, and the normalized selection toggle actions across the elements.

Also, individual differences were observed in the interaction pattern. When guided by the intellectual standards, some participants chose to fix the content manually. *"It's telling me it's, Things are too vague out there. I gotta define things."* (P3 fixing assumptions) However, some participants relied on auto-fix in a similar situation to reflect multiple guidances at the same time. *"Maybe I'll try auto-fixing it... Okay, Okay, this looks good, so I'll accept it."* (P12 fixing purpose)

6.3 Evidence-Based Reasoning

Surfacing evidence underlying a claim plays a crucial role in shaping how people trust the reasoning of a system. Among the nine elements of thought displayed in Criticality's reasoning trace, six displayed retrieved evidence from the data report. On average, participants opened and analyzed the report's evidence for each element of Criticality (M=6.1, SD=6.46) times, the majority in the Decision Matrix (M=5.80, SD=8.83), followed by Assumptions (M=4.3, SD=5.06). However, this varied considerably by user preference: some frequently checked the report to manually verify whether the extracted evidence existed and aligned with the supporting claim's context, whereas others trusted the evidence retrieval algorithm and relied solely on the aggregated stacked bar chart, which showed the distribution of evidence types (supportive, neutral, contradicting). Criticality supports per-claim verification in the flow of work, which was often used for navigating the

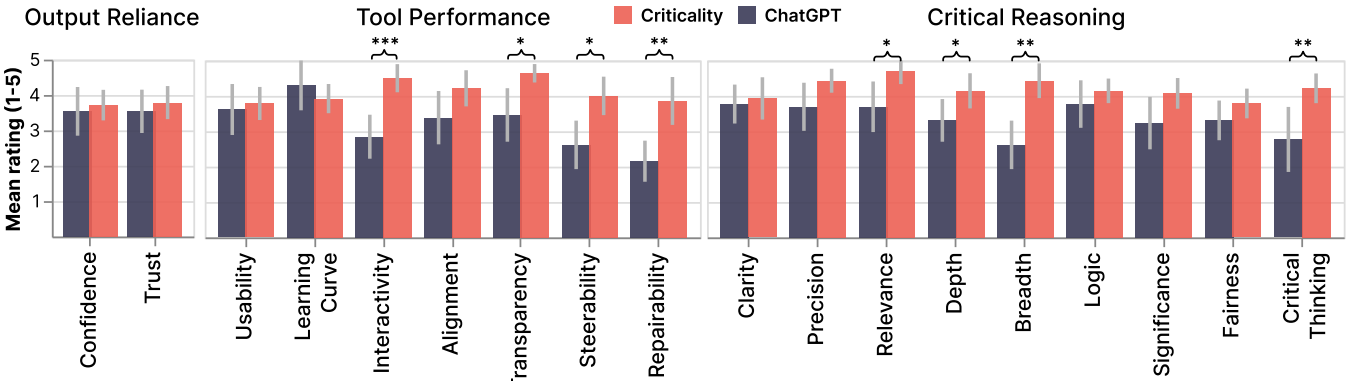


Figure 9: Average ratings (1–5) with 95% confidence interval comparing ChatGPT and Criticality. Each chart illustrates the users' reliance on output (e.g., critical thinking, trust), tool performance (e.g., usability, transparency), and the quality of critical reasoning (e.g., clarity, depth, fairness). Criticality generally scored higher across all sets of measures, and asterisks show a statistically significant advantage using the Wilcoxon signed-rank test (*: p<0.05, **: p<0.01, ***:p<0.001).

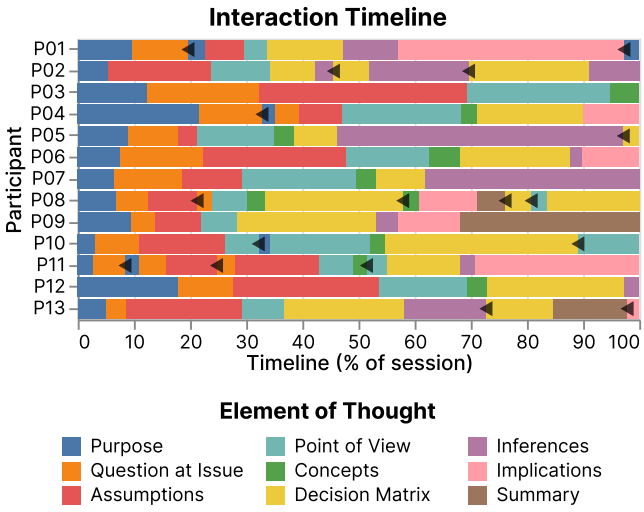


Figure 10: Interaction timelines of participants. While some participants (e.g., P6, P9, P12) followed a linear reasoning path, others (e.g., P1, P2, P8, P11) engaged in non-linear iteration, revisiting earlier components. The triangle marks indicate back-tracking moments, where participants reviewed and refined previously generated elements. The first three elements (Purpose, Questions, and Assumptions) and the Decision Matrix mark pivotal reflection points, fostering users to revisit and iterate on previous reasoning in an otherwise sequential workflow.

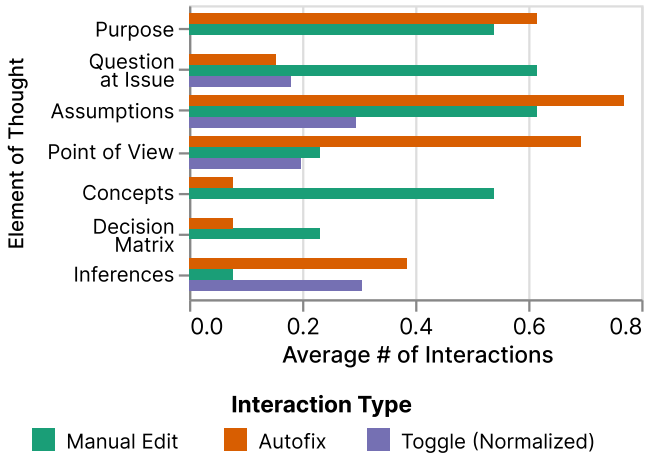


Figure 11: Average number of steering/repairing interactions observed in each element. Among elements of thought, *Purpose* and *Assumptions* had the highest number of interactions, suggesting that participants prioritized aligning the high-level goal and validating the foundational beliefs. Toggle selection action was normalized based on the total number of toggles present in each element. The Purpose and the Decision Matrix did not have a toggle button.

hundreds-of-pages-long reports in a contextual per-claim manner *"I'm going from 50 pages to get this focus here."* P12 said, as he used the blue link in the element to jump straight to the referenced spot in the report. The evidence classification labels of supporting, contradictory, or neutral were used to triage whether to exclude or scrutinize claims *"Wherever there is a red, decide whether it should be used or not."*(P12) *"These are bad assumptions. It's reasonable to turn them off."* (P8)

On the other hand, in the Baseline condition, (5/13) people clicked on references. But clicking a reference here downloaded their entire report, undermining traceability and trust and overwhelming the user. As P3 said, *"I don't see the link between a conclusion and where they got that information...[clicks reference] Ah, it just downloaded the whole PDF."* P11 said *"Inline citation... allow me to download to see the detail... maybe I could get some high-level details to make sure that it is not hallucinating."*

6.4 Perceived cognitive value in the critical thinking process

Participants rated Criticality higher than Baseline for supporting critical thinking, confirming its effectiveness in promoting analytical reasoning (see Figure 9). Criticality scaffolds critical thinking with explicit, inspectable checks (edits, intellectual standards' assessment and guidance, toggles). On the other hand, Baseline often ensures clarity and relevance in the output, while process-level precision, logic, and fairness are hard to interrogate or control. It was up to the user to specify the significance, breadth, and depth through systematic prompting; otherwise, these elements are ignored. This opens a future research direction to investigate whether deliberative prompting alone can match the benefits of Criticality's scaffolding, clearly distinguishing the value of the interface from the capability of the underlying model.

6.4.1 Clarity. Criticality ensured clarity by presenting an interactive reasoning trace with detail-on-demand (e.g., related concepts could be expanded to get definitions, and each step could be collapsed), presenting evidence linked to claims, and assessing each element while providing guidance on how to improve it. Participants reflected on the transparency of the reasoning trace, *"It's beyond my expectations...very transparent."* (P6) *"It really externalizes all details like what decision implications mean."*(P7) When using Baseline, there is less clarity in the reasoning process and how it generated the outcome. *"Trace helps visualize what it's doing, but it's also very lengthy to just sit there and read."* (P11) There was a lack of clarity even about the output, as P2 describes, *"I'm trying to understand: is this based on only the information I gave or [on] its own information?"*

6.4.2 Precision and Relevance. Criticality's suggestions, based on intellectual standards, frequently flagged a lack of precision and relevance, guiding participants to improve these areas through targeted suggestions. However, users questioned whether LLM's idea of improving precision by adding details and examples aligned with their own. *"I can edit or auto-fix the goal to be more precise."* (P12) *"I'm not sure whether providing detail means more careful reasoning."* (P8) When using Baseline, it was hard for participants to get the system to match their expectations for precision or assess relevance. *"I'm not super happy with the feedback because I think these are very out of nowhere."* (P11) For example, participants noticed that process steps sometimes diverged from the user's immediate prompt without explaining why. *"It's doing things I didn't ask"* (P8) And about the output *"I feel like this this part is definitely a bit repetitive in a way because you know they already talked about it earlier."*(P1)

6.4.3 Depth and Breadth. Criticality is scaffolded through structured elements, such as questions, implications, inferences, and guidance, at each level, suggesting consideration of additional factors at multiple levels. Additionally, Criticality ensured breadth through elements of thought, such as questions, perspectives, and concepts. On the other hand, Baseline left breadth to user initiative and required iterative and skilled prompting for depth. This was a clear differentiator between Criticality and Baseline. For example, *"It includes perspectives I hadn't thought of."* (P6 using Criticality) vs *"It's giving me the obvious ones"* (P2 using Baseline), or *"This is the sort of stuff I could get off the nightly news… not real deep."* (P3 using Baseline)

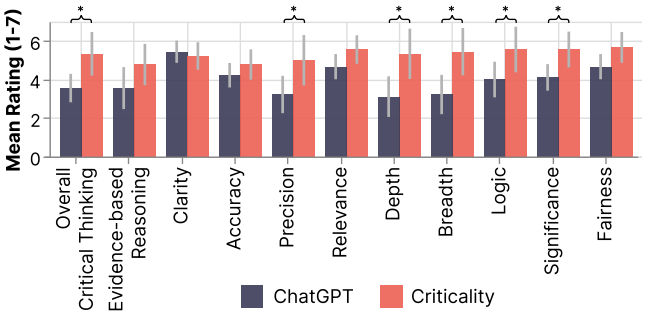


Figure 12: Mean ratings (1-7) with 95% confidence interval where two expert judges rated ChatGPT and Criticality blind to condition (IIR: 0.86). Criticality-generated decision rationales were rated significantly higher than Baseline across overall critical thinking and four intellectual standards. Asterisks show a statistically significant advantage of Criticality using the Mann-Whitney U test (*: p<0.05).

6.4.4 Logic and Fairness. Criticality's explicit logic and fairness checks made flaws visible. Links between evidence and claims, and the automatic exclusion of claims with contradictory evidence, helped follow the logic. Including multiple stakeholder perspectives, concepts, rephrased questions, and providing guidance to improve this helped account for fairness. As P6 mentioned, *"Pretty much follows logically from the evidence ... Presents information in an unbiased way."* P5: *"I feel more confident when it pointed out potentially bias and I could correct it"* On the other hand, Baseline's long outputs with hidden logic made assessment difficult, and required intentional and skilled prompting to ensure these standards. As P3 said, *"I don't see the link between a conclusion and where they got that information."* P1 reflects on how Baseline could be interacted with to ensure logic and fairness in the Baseline, *"Unless if you explicitly tell it to be as objective as possible to only follow the PDF and ask what other perspectives are there what are the other potential pros and cons and things like that."*

6.4.5 Significance. When using Criticality, participants valued the ability to interactively prioritize what's important to them at each step, by editing to add or remove details, adjusting criteria weights, or guidance suggestions to specify this. *"From the matrix, I could decide now."* (P7) when interacting with the decision matrix surfaced what mattered most given her weights. On the other hand, when using Baseline, participants had to intentionally prompt for what mattered to them when making the decision.

6.5 Decision Quality and User Confidence

Figure 12 shows that blind-to-condition experts rated decision rationales produced using Criticality as significantly higher in overall critical thinking and evidence-based reasoning compared to those produced using the Baseline condition. Across all Paul–Elder intellectual standards (clarity, accuracy, precision, relevance, depth, breadth, logic, significance, and fairness), Criticality-supported rationales received consistently higher average scores, except for

Clarity, where both showed approximately the same score ranges. As some participants (e.g., P1, P2, P15) only came to a decision and rationale on only one of the conditions, we applied Mann-Whitney U test to compare scores between conditions, treating them as independent samples. Still, we consider these statistical results as exploratory rather than conclusive, where further research using large-scale controlled studies is required to verify their generalizability.

Participants perceived different levels of trust and confidence in the final decisions made with Baseline and Criticality. When using the Baseline, participants often trusted the final decision tentatively. *"I can't really know whether or not I trust it."* (P1) *"don't know that I trust it though...Transparency is not there. It's doing all this stuff before it comes back to me."* (P4) *"But still you have to explain really careful what exactly you want and you have to come back and just guide it multiple times back, you know."* (P3)

On the other hand, Criticality built confidence and trust. As P6 said, *"I think I can trust the results. Looks pretty reasonable based on the data." "Helped me think ... helped me feel like I didn't need to ask as many follow-up questions. It really externalizes, like, what the decision implications mean."* (P7) and *"It helped me list out the assumptions... and reflect on different perspectives."* (P8).

7 Discussion

Our evaluation of Criticality reveals important insights about scaffolding human-AI decision-making workflows. This section discusses our findings and their implications for human-AI decision-making systems, as well as current limitations of the system and study.

7.1 From Post-Hoc Repair to In-Process Steering

Overall, participants preferred using Criticality over the Baseline for decision-making, valuing the transparency and agency during the process (Section 6.1). LLMs are fundamentally trained and optimized for fluency over factual accuracy [39] when responding to a user prompt [37], fail at complex or unfamiliar reasoning tasks [64, 88], and amplify cognitive and data biases [28]. This traps users in inefficient, multi-turn repair cycles, which our formative study identified as requiring significant time and effort (Section 3.3.3). Furthermore, vague and imprecise initial prompts [92, 116] introduce a tedious refinement going over lengthy responses, which frustrated our participants while using the Baseline. Criticality replaces this post-hoc repair with in-process steering, allowing users to guide the AI's reasoning one step at a time (Section 6.2). This approach reduces cognitive load while tracing a long conversation [22], and encourages reflection, which we observed as participants iterated around the decision matrix. This shifts the locus of control: users no longer merely prompt but truly *align* using this reasoning trace. Experts rated the decisions made with Criticality as significantly higher in overall critical thinking and evidence-based reasoning (Section 6.5)

This aligns with a broader trend in human-AI interaction towards structured alignment, helping human-AI decision-making systematically rather than through prompt trial-and-error [43, 108].

Implications for Future Human-AI Interaction Design:

- Support in-process steering rather than post-hoc repair and reformulation. Interfaces should expose intermediate reasoning states, assumptions, and evidence, allowing users to adjust direction early and maintain shared understanding throughout the decision process.
- Replace ad-hoc prompt refinement with structured scaffolds for alignment.

7.2 Designing for Positive Friction To Balance Cognitive Engagement and Efficiency

Participants consistently leveraged the heatmap and guidance system grounded in intellectual standards to repair or steer reasoning traces (Section 6.2). While effective, this process introduced cognitive overhead, leading to two distinct workarounds: manual edits and auto-fixes (Section 6.2). These differed in the depth of reflection and effort required.

This tension underscores the challenge of balancing cognitive engagement with efficiency. Showcased by Bucinca et al. [10], reduced friction can obscure biases or logical gaps, thereby degrading performance; conversely, excessive friction may frustrate users, discouraging both engagement and trust [96]. In our case, the *auto-fix* feature in Criticality enhances efficiency by aggregating multiple guidance points, yet it can also serve as a cognitive shortcut, limiting opportunities for deep reasoning.

Positive friction [19, 40] provides a productive middle ground. In Criticality, the heatmap and the guidance system with the auto-fix feature introduced this form of *positive friction*, allowing users to understand issues before deciding whether or not to delegate cognitively demanding tasks to AI. This approach transforms friction from a usability or efficiency barrier into a learning scaffold, where users can leverage critical reflection to calibrate their trust in AI [17]. However, extending Naiseh et al.'s [65] findings, participants occasionally chose to accept the AI guidance without scrutiny, with limited cognitive engagement.

Building on these findings, future work could investigate how varying degrees of friction shape the trade-off between engagement and efficiency. Such insights could inform adaptive mechanisms that dynamically tune friction based on user expertise, task complexity, and cognitive state, thereby mitigating over-reliance on automation [47].

Furthermore, Criticality's heatmap strip, recommended guidance, and auto-fix interaction reconfigure the human–AI relationship into a continuous, collaborative quality-control loop. Both, users *and* AI, can identify violations and iteratively improve each others' reasoning quality. This design embodies an implicit principle: *steerability is quality assurance*. Participants' higher ratings of steerability and clarity reflect these reciprocal interactions, where each detected violation becomes an opportunity for learning and repair (Section 6.2, Section 6.4). Also, by embedding guidance within context, Criticality transforms abstract intellectual standards into interactive evaluation criteria.

Implications for Future Human-AI Interaction Design:

- Embed interactions that prompt users to cognitively engage and reflect before accepting AI fixes, turning verification into an active learning process rather than a passive confirmation step.

- Design interactions that make the AI's limitations visible and interpretable, so users can calibrate trust and retain agency, even when automation assists with repair.
- Future systems should adjust the level and timing of positive friction based on user proficiency, task complexity, and cognitive load.

7.3 Evidence-Based Reasoning Calibrates User Reliance and Quality Control in AI

The claim–evidence linking mechanism and support/contradiction classification anchored every reasoning segment to concrete sources. This helped not only improve factual precision but also gave users agency and affordances to question and adjust the model's stance, whether to trust, challenge, or qualify a statement. This evidence-based interaction re-frames transparency: instead of global model explainability, Criticality offers claim-level auditability. Moreover, by embedding evidence links directly into the reasoning trace, users could navigate to the relevant passage, verify claims without leaving the task context, fluidly within the flow of work. This novel evidence-based reasoning and interaction mechanism adds to prior research in information retrieval and referencing in human-AI interaction [42, 78, 103].

Implication for Future Human-AI Interaction Design: Interactive claim–evidence linking with support/contradiction classification and explanations can foster engagement and build trust, efficiently in the flow of work.

7.4 Decision Matrix and Executive Summary as Bridges Between Process and Outcome

Both the Decision Matrix and the Executive Summary function as boundary objects in human–AI collaborative decision-making: they provide a shared representation that connects the AI's internal reasoning process with the human's interpretive understanding, while remaining flexible enough to serve different purposes for each party [91]. This design goes beyond merely recording or displaying results, it actively coordinates participation and enables translation between both the model and the user's reasoning.

Participants frequently moved back-and-forth between the matrix and the preceding Elements of Thought, reflecting a non-linear workflow of refining criteria and options rather than a fixed sequential generation of outputs. This observation aligns with prior research in intelligent interfaces that externalize decision criteria into matrix or tabular forms to support user-driven comparison and sensemaking. Each cell in our Decision Matrix, with its color-coded score breakdown, made the underpinnings of the AI's recommendation explicit: users could click on a cell to see how a given option's score was derived from evidence and critical-thinking standards. This transparent decomposition prompted users to revisit assumptions and re-evaluate evidence whenever something looked unexpected. This mirrors the goal of human–AI deliberation frameworks, which seek to expose and reconcile conflicting human–AI opinions through structured dialogue [56]. Conventional conversational AI interfaces typically present a single recommendation that the user can only accept or reject (with no granularity for partial agreement).

The Executive Summary complements this by linking process and outcome through evidence-grounded traceability. Participants used it not as a static report but as an interactive gateway into the reasoning trace. Instead of merely accepting the model's conclusion, they could trace each claim to the exact passage in the source report, enabling targeted verification and rapid error correction. This fine-grained traceability differentiates Criticality's output from conventional LLM outputs, which often collapse justification into long narratives.

Implication for Future Human-AI Interaction Design: It is important to maintain *shared accountability* in human–AI interaction. In Criticality, this takes the form of a continuous loop linking process and outcome: the AI provides structured analysis and retrieval; the human contributes judgment and context; and the system ensures both inputs remain visible and are integrated into the final decision. This shared accountability contrasts with common human-AI interactions, where reasoning is hidden, fixed or presented as immutable explanations.

7.5 Limitations & Future Work

We also identified limitations in our system design and study as opportunities for future work:

Criticality generates the reasoning trace's elements in a sequential order. During the study, some participants wished to explore the decision options first and then move on to the other elements (P8), while others appreciated the earlier steps that shape the decision space (P5). Based on these observations, Criticality could be more adaptive to user preferences or context, providing flexible exploration around critical thinking. Adaptive content generation could also benefit users in domain-specific cases where decision options are nuanced. For instance, one participant, a learning scientist, claimed that the decision options from Criticality were quite generic when they expected them to be more specific.

The LLM-as-a-judge approach offers the advantage of providing consistent, scalable, and fast evaluations of text or model outputs, but it also carries risks such as bias reinforcement, lack of true understanding, and overreliance on automated judgment without human oversight. To mitigate these issues, we follow best practices from prior literature [83, 95, 110], and establish a clear rubric with few-shot examples to approximate the evaluation problem as a pattern-matching task and produce more reliable judgments. However, as LLMs are fundamentally next-token predictors, the result is not guaranteed to be reasonable based on human knowledge. Future work could complement the guidance generation and scoring mechanism with a more reliable approach (e.g., carefully crafted heuristics) or include a technical evaluation of the generated guidance and scoring to prove its validity.

Participant Pool While we successfully recruited 13 participants who frequently make decisions based on data. While this helps identify qualitative insights into interaction and cognitive behavior, to make more quantitative claims, larger-scale longitudinal deployments are required.

Study Design Trade-Offs We had to balance ecological validity with experimental control in our evaluation study, which informed specific design choices. For example, our baseline was motivated by ecological validity: professional decision-makers commonly

use conversational LLMs (e.g., ChatGPT). While an ablation study would be valuable for disentangling the effects of individual component contributions, this was beyond the scope of this paper, which focuses on emergent interaction behaviors arising from the use of each system.

Furthermore, the tasks assigned to participants during the study were artificial scenarios having a single report file per task, which may have reduced the criticality of their decisions. However, web search is commonly used to collect data in real world, where P13 instinctively recognized this when trying ChatGPT's agentic mode, noting, *"I mean some of the stuff that might not be in the report."* Future versions of Criticality could leverage web search while maintaining the same evidence-based reasoning mechanisms to identify knowledge gaps, verify claims across sources, and consider alternative perspectives outside user-uploaded reports.

Additionally, having less than 30 minutes to both fully learn how to use the tool and make the decision could have prevented the participants from being fully immersed in the scenario, as time-pressure changes the user behavior while interacting with AI [94]. Future work could deploy Criticality to real-world decision makers with a larger pool size and observe how they use it to make high-stakes decisions to seek more generalizable results.

8 Conclusion

This paper presents Criticality, a system that scaffolds human-AI decision-making through interactive critical thinking structures and evidence-based reasoning traces. The operationalized Paul-Elder Critical Thinking Framework decomposes reasoning into editable Elements of Thought, evaluates them against Intellectual Standards, and grounds every claim in verifiable evidence. A within-subjects study (n=13) showed that this approach improved steerability, repairability, and interactivity compared to the baseline of a state-of-the-art reasoning model in a conversational interface. Expert assessments further confirmed that participants produced higher-quality decisions and exhibited stronger critical thinking across multiple intellectual standards.

Our findings highlight that effective human-AI collaboration requires more than strong reasoning capabilities; it requires systems that can show the reasoning process, ensure user agency when the context should be steered, and preserve the link between the claims and evidence. Criticality embodies these principles, demonstrating how AI can function not merely as an answer sheet but as a *tool for thinking*, which guides and strengthens users' reasoning toward more reliable, transparent, and inclusive decisions.

9 GenAI Usage Disclosure

The system interface was developed with assistance from Claude Sonnet 4 and ChatGPT 5 via Cursor IDE. All code was reviewed and modified by the authors.

Acknowledgments

This study was conducted in accordance with the internal research policies of the authors' affiliated organization. All participants provided informed consent, and data handling practices adhered to the company's ethics, privacy, and confidentiality standards. Supplemental materials are available at https://osf.io/23ya7, and a live demonstration of the system can be accessed at https://iui26-criticality.up.railway.app.

References

[1] Narayan Prasad Adhikari, Jhupa Kumari Budhathoki, and Sharad Adhikari. 2025. Use of Critical Thinking in Decision-Making Process. *Perspectives on Higher Education* 15, 01 (2025), 155–168.

[2] Anthropic. 2025. *System Card: Claude Opus 4 & Claude Sonnet 4.* Technical Report. Anthropic. https://www-cdn.anthropic.com/4263b940cabb546aa0e3283f35b686f4f3b2ff47.pdf

[3] Zinat Ara, Hossein Salemi, Sungsoo Ray Hong, Yasas Senarath, Steve Peterson, Amanda Lee Hughes, and Hemant Purohit. 2024. Closing the Knowledge Gap in Designing Data Annotation Interfaces for AI-powered Disaster Management Analytic Systems. In *Proceedings of the 29th International Conference on Intelligent User Interfaces* (Greenville, SC, USA) *(IUI '24)*. Association for Computing Machinery, New York, NY, USA, 405–418. doi:10.1145/3640543.3645214

[4] Susan Gardner Archambault, Joanne Helouvry, Bonnie Strohl, and Ginger Williams. 2015. Data visualization as a communication tool. *Library Hi Tech News* 32, 2 (2015), 1–9.

[5] Melanie Bancilhon, R Jordan Crouser, and Alvitta Ottley. 2023. Communicating Intel to Decision-Makers: Toward the Integration Text and Charts in Reports. *arXiv preprint arXiv:2302.10885* (2023).

[6] Aaron Bangor, Philip T Kortum, and James T Miller. 2008. An empirical evaluation of the system usability scale. *Intl. Journal of Human–Computer Interaction* 24, 6 (2008), 574–594.

[7] Benjamin S Bloom, Max D Engelhart, Edward J Furst, Walker H Hill, and David R Krathwohl. 1956. *Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain.* David McKay.

[8] Bond Capital. 2023. AI & Universities, Expanded. https://www.bondcap.com/report/aiu-e/. Accessed: 2025-10-10.

[9] Bond Capital. 2023. Thoughts on AI. https://www.bondcap.com/reports/tai. Accessed: 2025-10-10.

[10] 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. *Proceedings of the ACM on Human-computer Interaction* 5, CSCW1 (2021), 1–21.

[11] M. Y. Cai, Y. Lin, and W. J. Zhang. 2016. Study of the optimal number of rating bars in the likert scale. In *Proceedings of the 18th International Conference on Information Integration and Web-Based Applications and Services* (Singapore, Singapore) *(iiWAS '16)*. Association for Computing Machinery, New York, NY, USA, 193–198. doi:10.1145/3011141.3011213

[12] Bette Case. 1994. Walking around the elephant: A critical-thinking strategy for decision making. 101–109 pages.

[13] Federico Castagna, Isabel Sassoon, and Simon Parsons. 2024. Critical-Questions-of-Thought: Steering LLM reasoning with Argumentative Querying. *arXiv preprint arXiv:2412.15177* (2024).

[14] Federico Maria Cau, Hanna Hauptmann, Lucio Davide Spano, and Nava Tintarev. 2023. Supporting High-Uncertainty Decisions through AI and Logic-Style Explanations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces (IUI '23)*. ACM, Sydney, NSW, Australia, 251–263. doi:10.1145/3581641.3584080

[15] Kathy Charmaz. 2014. Constructing grounded theory. (2014).

[16] Jiaqi Chen, Yanzhe Zhang, Yutong Zhang, Yijia Shao, and Diyi Yang. 2025. Generative Interfaces for Language Models. *arXiv preprint arXiv:2508.19227* (2025).

[17] Valerie Chen, Q Vera Liao, Jennifer Wortman Vaughan, and Gagan Bansal. 2023. Understanding the role of human intuition on reliance in human-AI decision-making with explanations. *Proceedings of the ACM on Human-computer Interaction* 7, CSCW2 (2023), 1–32.

[18] Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato, Carson Denison, John Schulman, Arushi Somani, Peter Hase, Misha Wagner, Fabien Roger, et al. 2025. Reasoning Models Don't Always Say What They Think. *arXiv preprint arXiv:2505.05410* (2025).

[19] Zeya Chen and Ruth Schmidt. 2024. Exploring a behavioral model of "positive friction" in human-AI interaction. In *International Conference on Human-Computer Interaction*. Springer, 3–22.

[20] Chun-Wei Chiang, Zhuoran Lu, Zhuoyan Li, and Ming Yin. 2024. Enhancing AI-Assisted Group Decision Making through LLM-Powered Devil's Advocate. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*. ACM, 103–119. doi:10.1145/3640543.3645199

[21] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261* (2025).

[22] Adam Coscia, Shunan Guo, Eunyee Koh, and Alex Endert. 2025. OnGoal: Tracking and Visualizing Conversational Goals in Multi-Turn Dialogue with Large Language Models. *arXiv preprint arXiv:2508.21061* (2025).

[23] Samuel Crosariol et al. 2024. *Pydantic: Data validation and settings management using Python type hints.* https://pydantic.dev/

[24] Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2024. Chain-of-Verification Reduces Hallucination in Large Language Models. 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, 3563–3578. doi:10.18653/v1/2024.findings-acl.212

[25] Ian Drosos, Advait Sarkar, Neil Toronto, et al. 2025. " It makes you think": Provocations Help Restore Critical Thinking to AI-Assisted Knowledge Work. *arXiv preprint arXiv:2501.17247* (2025).

[26] Aline Duelen, Iris Jennes, and Wendy Van den Broeck. 2024. Socratic AI against disinformation: Improving critical thinking to recognize disinformation using Socratic AI. In *Proceedings of the 2024 ACM International Conference on Interactive Media Experiences*. 375–381.

[27] Robert Duron, Barbara Limbach, and Wendy Waugh. 2006. Critical thinking framework for any discipline. *International Journal of teaching and learning in higher education* 17, 2 (2006), 160–166.

[28] Jessica Maria Echterhoff, Yao Liu, Abeer Alessa, Julian McAuley, and Zexue He. 2024. Cognitive Bias in Decision-Making with LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. Association for Computational Linguistics, 12640–12653.

[29] Upol Ehsan, Pradyumna Tambwekar, Larry Chan, Brent Harrison, and Mark O. Riedl. 2019. Automated Rationale Generation: A Technique for Explainable AI and Its Effects on Human Perceptions. In *Proceedings of the 24th International Conference on Intelligent User Interfaces (IUI '19)*. ACM, Marina del Ray, CA, USA, 263–274. doi:10.1145/3301275.3302316

[30] Robert H Ennis. 1962. A concept of critical thinking. *Harvard educational review* 32, 1 (1962), 81–111.

[31] Lucile Favero, Juan Antonio Pérez-Ortiz, Tanja Käser, and Nuria Oliver. 2024. Enhancing critical thinking in education by means of a Socratic chatbot. In *International Workshop on AI in Education and Educational Research*. Springer, 17–32.

[32] Raymond Fok, Nedim Lipka, Tong Sun, and Alexa F Siu. 2024. Marco: Supporting Business Document Workflows via Collection-Centric Information Foraging with Large Language Models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–20.

[33] 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*. 794–806.

[34] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-Augmented Generation for Large Language Models: A Survey. arXiv:2312.10997 [cs.CL] https://arxiv.org/abs/2312.10997

[35] Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. 2024. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594* (2024).

[36] Paul Hager, Friederike Jungmann, Robbie Holland, Kunal Bhagat, Inga Hubrecht, Manuel Knauer, Jakob Vielhauer, Marcus Makowski, Rickmer Braren, Georgios Kaissis, and Daniel Rueckert. 2024. Evaluation and Mitigation of the Limitations of Large Language Models in Clinical Decision-making. *Nature Medicine* 30, 9 (2024), 2613–2622.

[37] Hasan Abed Al Kader Hammoud, Hani Itani, and Bernard Ghanem. 2025. Beyond the last answer: Your reasoning trace uncovers more than you think. *arXiv preprint arXiv:2504.20708* (2025).

[38] Jeffrey Heer, Fernanda B. Viégas, and Martin Wattenberg. 2007. Voyagers and voyeurs: supporting asynchronous collaborative information visualization. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '07)*. Association for Computing Machinery, New York, NY, USA, 1029–1038. doi:10.1145/1240624.1240781

[39] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Trans. Inf. Syst.* 43, 2, Article 42 (Jan. 2025), 55 pages. doi:10.1145/3703155

[40] Mert Inan, Anthony Sicilia, Suvodip Dey, Vardhan Dongre, Tejas Srinivasan, Jesse Thomason, Gökhan Tür, Dilek Hakkani-Tür, and Malihe Alikhani. 2025. Better slow than sorry: Introducing positive friction for reliable dialogue systems. *arXiv preprint arXiv:2501.17348* (2025).

[41] Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2017. *Billion-scale similarity search with GPUs*. Technical Report. Facebook AI Research. https://arxiv.org/abs/1702.08734

[42] Hita Kambhamettu, Alyssa Hwang, Philippe Laban, and Andrew Head. 2025. Attribution Gradients: Incrementally Unfolding Citations for Critical Examination of Attributed AI Answers. *arXiv preprint arXiv:2510.00361* (2025).

[43] Majeed Kazemitabaar, Jack Williams, Ian Drosos, Tovi Grossman, Austin Zachary Henley, Carina Negreanu, and Advait Sarkar. 2024. Improving Steering and Verification in AI-Assisted Data Analysis with Interactive Task Decomposition *(UIST '24)*. Association for Computing Machinery, New York, NY, USA, Article 92, 19 pages. doi:10.1145/3654777.3676345

[44] Predrag Klasnja, Eric B Hekler, Elizabeth V Korinek, John Harlow, and Sonali R Mishra. 2017. Toward usable evidence: optimizing knowledge accumulation in HCI research on health behavior change. In *Proceedings of the 2017 CHI conference on human factors in computing systems*. 3071–3082.

[45] Emily R Lai. 2011. Critical thinking: A literature review. *Pearson's research reports* 6, 1 (2011), 40–41.

[46] Kin-Ho Lam, Zhengxian Lin, Jed Irvine, Jonathan Dodge, Zeyad T Shureih, Roli Khanna, Minsuk Kahng, and Alan Fern. 2021. Identifying reasoning flaws in planning-based rl using tree explanations. *arXiv preprint arXiv:2109.13978* (2021).

[47] Hao-Ping Lee, Advait Sarkar, Lev Tankelevitch, Ian Drosos, Sean Rintel, Richard Banks, and Nicholas Wilson. 2025. The impact of generative AI on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–22.

[48] Soohwan Lee, Seoyeong Hwang, Dajung Kim, and Kyungho Lee. 2025. Conversational Agents as Catalysts for Critical Thinking: Challenging Social Influence in Group Decision-making. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–12.

[49] Lan Li, Tina Lassiter, Joohee Oh, and Min Kyung Lee. 2021. Algorithmic hiring in practice: Recruiter and HR Professional's perspectives on AI use in hiring. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 166–176.

[50] Michael Xieyang Liu, Jane Hsieh, Nathan Hahn, Angelina Zhou, Emily Deng, Shaun Burley, Cynthia Taylor, Aniket Kittur, and Brad A Myers. 2019. Unakite: Scaffolding developers' decision-making using the web. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*. 67–80.

[51] Michael Xieyang Liu, Aniket Kittur, and Brad A Myers. 2022. Crystalline: Lowering the cost for developers to collect and organize information for decision making. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–16.

[52] Michael Xieyang Liu, Tongshuang Wu, Tianying Chen, Franklin Mingzhe Li, Aniket Kittur, and Brad A Myers. 2024. Selenite: Scaffolding Online Sensemaking with Comprehensive Overviews Elicited from Large Language Models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–26.

[53] Xingyu Bruce Liu, Haijun Xia, and Xiang Anthony Chen. 2025. Interacting with Thoughtful AI. *arXiv preprint arXiv:2502.18676* (2025).

[54] LlamaIndex. 2024. *LlamaParse: GenAI-Native Document Parser.* https://docs.llamaindex.ai/en/stable/llama_cloud/llama_parse/ Accessed: 2025-09-28.

[55] Carlos JS Lourenço, Benedict GC Dellaert, and Bas Donkers. 2020. Whose algorithm says so: The relationships between type of firm, perceptions of trust and expertise, and the acceptance of financial robo-advice. *Journal of Interactive Marketing* 49, 1 (2020), 107–124.

[56] 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. arXiv:2403.16812 [cs.HC] https://arxiv.org/abs/2403.16812

[57] Pratyusha Maiti and Ashok Goel. 2025. Can an AI Partner Empower Learners to Ask Critical Questions?. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*. 314–324.

[58] Damien Masson, Sylvain Malacria, Géry Casiez, and Daniel Vogel. 2024. Directgpt: A direct manipulation interface to interact with large language models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–16.

[59] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. 2019. Reliability and Inter-rater Reliability in Qualitative Research: Norms and Guidelines for CSCW and HCI Practice. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 72 (Nov. 2019), 23 pages. doi:10.1145/3359174

[60] McKinsey & Company. 2025. *Sporting Goods 2025—The new balancing act: Turning uncertainty into opportunity.* Technical Report. McKinsey & Company. https://www.mckinsey.com/industries/retail/our-insights/sporting-goods-industry-trends

[61] McKinsey Global Institute. 2024. *The next big arenas of competition.* Technical Report. McKinsey & Company. https://www.mckinsey.com/mgi/our-research/the-next-big-arenas-of-competition

[62] Inc. Meta Platforms. 2024. *React: The library for web and native user interfaces.* https://react.dev/

[63] Yakira Mirabito, Megane Annaelle Tchatchouang Kayo, and Kosa Goucher-Lambert. 2024. Feature, specification and evidence framework for communicating design rationale. *Design Science* 10 (2024), e20.

[64] Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. 2024. GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models. *arXiv preprint*

arXiv:2410.05229 (2024).

[65] Mohammad Naiseh, Reem S Al-Mansoori, Dena Al-Thani, Nan Jiang, and Raian Ali. 2021. Nudging through friction: an approach for calibrating trust in explainable AI. In *2021 8th International Conference on Behavioral and Social Computing (BESC)*. IEEE, 1–5.

[66] An T Nguyen, Aditya Kharosekar, Saumyaa Krishnan, Siddhesh Krishnan, Elizabeth Tate, Byron C Wallace, and Matthew Lease. 2018. Believe it or not: designing a human-ai partnership for mixed-initiative fact-checking. In *Proceedings of the 31st annual ACM symposium on user interface software and technology*. 189–199.

[67] OpenAI. 2024. OpenAI o1 System Card. https://openai.com/index/openai-o1-system-card/

[68] OpenAI. 2025. Introducing Deep Research. https://openai.com/index/introducing-deep-research/ OpenAI Blog Post.

[69] Emre Oral, Ria Chawla, Michel Wijkstra, Narges Mahyar, and Evanthia Dimara. 2023. From information to choice: A critical inquiry into visualization tools for decision making. *IEEE Transactions on Visualization and Computer Graphics* 30, 1 (2023), 359–369.

[70] Rita Orji and Karyn Moffatt. 2018. Persuasive technology for health and wellness: State-of-the-art and emerging trends. *Health informatics journal* 24, 1 (2018), 66–91.

[71] Christine A Padesky. 1993. Socratic questioning: Changing minds or guiding discovery. In *A keynote address delivered at the European Congress of Behavioural and Cognitive Therapies, London*, Vol. 24. 44.

[72] Srishti Palani, Zijian Ding, Austin Nguyen, Andrew Chuang, Stephen MacNeil, and Steven P Dow. 2021. CoNotate: Suggesting queries based on notes promotes knowledge discovery. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–14.

[73] Srishti Palani, Yingyi Zhou, Sheldon Zhu, and Steven P Dow. 2022. InterWeave: Presenting Search Suggestions in Context Scaffolds Information Search and Synthesis. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*. 1–16.

[74] Lawrence A Palinkas, Sarah M Horwitz, Carla A Green, Jennifer P Wisdom, Naihua Duan, and Kimberly Hoagwood. 2015. Purposeful sampling for qualitative data collection and analysis in mixed method implementation research. *Administration and policy in mental health and mental health services research* 42, 5 (2015), 533–544.

[75] 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. arXiv:2506.23678 [cs.HC] https://arxiv.org/abs/2506.23678

[76] Arjun Panickssery, Samuel Bowman, and Shi Feng. 2024. Llm evaluators recognize and favor their own generations. *Advances in Neural Information Processing Systems* 37 (2024), 68772–68802.

[77] Richard Paul and Linda Elder. 2019. *The miniature guide to critical thinking concepts and tools*. Rowman & Littlefield.

[78] Snehal Prabhudesai, Leyao Yang, Sumit Asthana, Xun Huan, Q Vera Liao, and Nikola Banovic. 2023. Understanding uncertainty: how lay decision-makers perceive and interpret uncertainty in human-AI decision making. In *Proceedings of the 28th international conference on intelligent user interfaces*. 379–396.

[79] Xiaoye Qu, Yafu Li, Zhaochen Su, Weigao Sun, Jianhao Yan, Dongrui Liu, Ganqu Cui, Daizong Liu, Shuxian Liang, Junxian He, et al. 2025. A survey of efficient reasoning for large reasoning models: Language, multimodality, and beyond. *arXiv preprint arXiv:2503.21614* (2025).

[80] Hans Peter Lynsgøe Raaschou-jensen, Constanza Fierro, and Anders Søgaard. 2025. Predicting thinking time in Reasoning models. *arXiv preprint arXiv:2506.23274* (2025).

[81] Leon Reicherts, Zelun Tony Zhang, Elisabeth von Oswald, Yuanting Liu, Yvonne Rogers, and Mariam Hassib. 2025. AI, help me think—but for myself: Assisting people in complex decision-making by providing different kinds of cognitive support. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–19.

[82] Armin Ronacher et al. 2024. *Flask*. https://flask.palletsprojects.com/

[83] Swarnadeep Saha, Xian Li, Marjan Ghazvininejad, Jason Weston, and Tianlu Wang. 2025. Learning to plan & reason for evaluation with thinking-llm-as-a-judge. *arXiv preprint arXiv:2501.18099* (2025).

[84] A Saxena. 2024. Ethical Considerations and Best Practices for Using Large Language Models in Decision-Making. *International Journal of Science and Research (IJSR)* 13 (2024).

[85] Edward Segel and Jeffrey Heer. 2010. Narrative visualization: Telling stories with data. *IEEE transactions on visualization and computer graphics* 16, 6 (2010), 1139–1148.

[86] Mohammad Ahmad Sheikh, Amit Kumar Goel, and Tapas Kumar. 2020. An approach for prediction of loan approval using machine learning algorithm. In *2020 international conference on electronics and sustainable communication systems (ICESC)*. IEEE, 490–494.

[87] Ben Shneiderman. 2020. Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction* 36, 6 (2020), 495–504.

[88] Parshin Shojaee, Iman Mirzadeh, Keivan Alizadeh, Maxwell Horton, Samy Bengio, and Mehrdad Farajtabar. 2025. The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity. *arXiv preprint arXiv:2506.06941* (2025).

[89] Herbert A Simon. 1960. The new science of management decision. (1960).

[90] Arjun Srinivasan and Vidya Setlur. 2021. Snowy: Recommending Utterances for Conversational Visual Analysis. In *The 34th Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) *(UIST '21)*. Association for Computing Machinery, New York, NY, USA, 864–880. doi:10.1145/3472749.3474792

[91] Susan Leigh Star and James R. Griesemer. 1989. Institutional Ecology, "Translations" and Boundary Objects: Amateurs and Professionals in Berkeley's Museum of Vertebrate Zoology, 1907–39. *Social Studies of Science* 19, 3 (1989), 387–420. doi:10.1177/030631289019003001

[92] Hari Subramonyam, Roy Pea, Christopher Pondoc, Maneesh Agrawala, and Colleen Seifert. 2024. Bridging the gulf of envisioning: Cognitive challenges in prompt based interactions with llms. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–19.

[93] Xin Sun, Rongjun Ma, Xiaochang Zhao, Zhuying Li, Janne Lindqvist, Abdallah El Ali, and Jos A Bosch. 2024. Trusting the search: unraveling human trust in health information from Google and ChatGPT. *arXiv preprint arXiv:2403.09987* (2024).

[94] Siddarth Swaroop, Zana Bućinca, Krzysztof Z. Gajos, and Finale Doshi-Velez. 2024. Accuracy-Time Tradeoffs in AI-Assisted Decision Making under Time Pressure. In *Proceedings of the 29th International Conference on Intelligent User Interfaces (IUI '24)*. ACM, Greenville, SC, USA, 138–154. doi:10.1145/3640543.3645206

[95] Annalisa Szymanski, Noah Ziems, Heather A Eicher-Miller, Toby Jia-Jun Li, Meng Jiang, and Ronald A Metoyer. 2025. Limitations of the llm-as-a-judge approach for evaluating llm outputs in expert knowledge tasks. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*. 952–966.

[96] Alaina N. Talboy and Elizabeth Fuller. 2023. Challenging the Appearance of Machine Intelligence: Cognitive Bias in LLMs and Best Practices for Adoption. *arXiv preprint arXiv:2304.01358* (2023).

[97] Recharts Team. 2024. Recharts: A composable charting library built on React components. http://recharts.org

[98] Stephen E Toulmin. 1958. *The uses of argument*. Cambridge university press.

[99] Christoph Treude and Raula Gaikovina Kula. 2025. Interacting with ai reasoning models: Harnessing" thoughts" for ai-driven software engineering. *arXiv preprint arXiv:2503.00483* (2025).

[100] U.S. Department of Education, Office of Educational Technology. 2023. *Artificial Intelligence and the Future of Teaching and Learning: Insights and Recommendations*. Technical Report. U.S. Department of Education, Washington, D.C. https://www2.ed.gov/documents/ai-report/ai-report.pdf

[101] 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. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–38.

[102] Vercel. 2024. Next.js: The React Framework for the Web. https://nextjs.org

[103] 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)*. ACM, College Station, TX, USA, 318–328. doi:10.1145/3397481.3450650

[104] Koki Wataoka, Tsubasa Takahashi, and Ryokan Ri. 2024. Self-preference bias in llm-as-a-judge. *arXiv preprint arXiv:2410.21819* (2024).

[105] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682* (2022).

[106] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems* (New Orleans, LA, USA) *(NIPS '22)*. Curran Associates Inc., Red Hook, NY, USA, Article 1800, 14 pages.

[107] World Economic Forum. 2025. *The Future of Jobs Report 2025*. Technical Report. World Economic Forum, Geneva, Switzerland. https://www.weforum.org/publications/the-future-of-jobs-report-2025/

[108] Tongshuang Wu, Michael Terry, and Carrie Jun Cai. 2022. AI chains: Transparent and controllable human-AI interaction by chaining large language model prompts. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. ACM.

[109] Liwenhan Xie, Chengbo Zheng, Haijun Xia, Huamin Qu, and Chen Zhu-Tian. 2024. Waitgpt: Monitoring and steering conversational llm agent in data analysis with on-the-fly code visualization. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*. 1–14.

[110] Yi Xu, Laura Ruis, Tim Rocktäschel, and Robert Kirk. 2025. Investigating Non-Transitivity in LLM-as-a-Judge. *arXiv preprint arXiv:2502.14074* (2025).

[111] Yusuke Yamauchi, Taro Yano, and Masafumi Oyamada. 2025. An Empirical Study of LLM-as-a-Judge: How Design Choices Impact Evaluation Reliability. *arXiv preprint arXiv:2506.13639* (2025).

[112] Fumeng Yang, Zhuanyi Huang, Jean Scholtz, and Dustin L Arendt. 2020. How do visual explanations foster end users' appropriate trust in machine learning?. In *Proceedings of the 25th international conference on intelligent user interfaces*. 189–201.

[113] 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. *Advances in neural information processing systems* 36 (2023), 11809–11822.

[114] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing Reasoning and Acting in Language Models. In *International Conference on Learning Representations (ICLR)*.

[115] Hye Sun Yun and Timothy Bickmore. 2025. Framing Health Information: The Impact of Search Methods and Source Types on User Trust and Satisfaction in the Age of LLMs. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–7.

[116] J Diego Zamfirescu-Pereira, Richmond Y Wong, Bjoern Hartmann, and Qian Yang. 2023. Why Johnny can't prompt: how non-AI experts try (and fail) to design LLM prompts. In *Proceedings of the 2023 CHI conference on human factors in computing systems*. 1–21.

[117] Jason Zhu and Hongyu Li. 2025. Towards concise and adaptive thinking in large reasoning models: A survey. *arXiv preprint arXiv:2507.09662* (2025).

[118] Xuyang Zhu, Sejoon Chang, and Andrew Kuik. 2025. Enhancing Critical Thinking with AI: A Tailored Warning System for RAG Models. *arXiv preprint arXiv:2504.16883* (2025).