Sensemaking in Multi-Agent LLM Interfaces: How Users Interpret Transparency and Trustworthiness Cues

Saumya Pareek
School of Computing and Information Systems
University of Melbourne
Melbourne, Victoria, Australia
saumya.pareek@student.unimelb.edu.au

Jarod Govers
School of Computing and Information Systems
University of Melbourne
Melbourne, Victoria, Australia
jarod.govers@unimelb.edu.au

Naja Kathrine Kollerup
Department of Computer Science
Aalborg University, Aalborg
Denmark
nkka@cs.aau.dk

Emily Wong
School of Computing and Information Systems
University of Melbourne
Melbourne, Victoria, Australia
emily.wong.1@unimelb.edu.au

Eduardo Velloso
School of Computer Science
The University of Sydney
Sydney, New South Wales, Australia
eduardo.velloso@sydney.edu.au

Jorge Goncalves
School of Computing and Information Systems
University of Melbourne
Melbourne, Victoria, Australia
jorge.goncalves@unimelb.edu.au

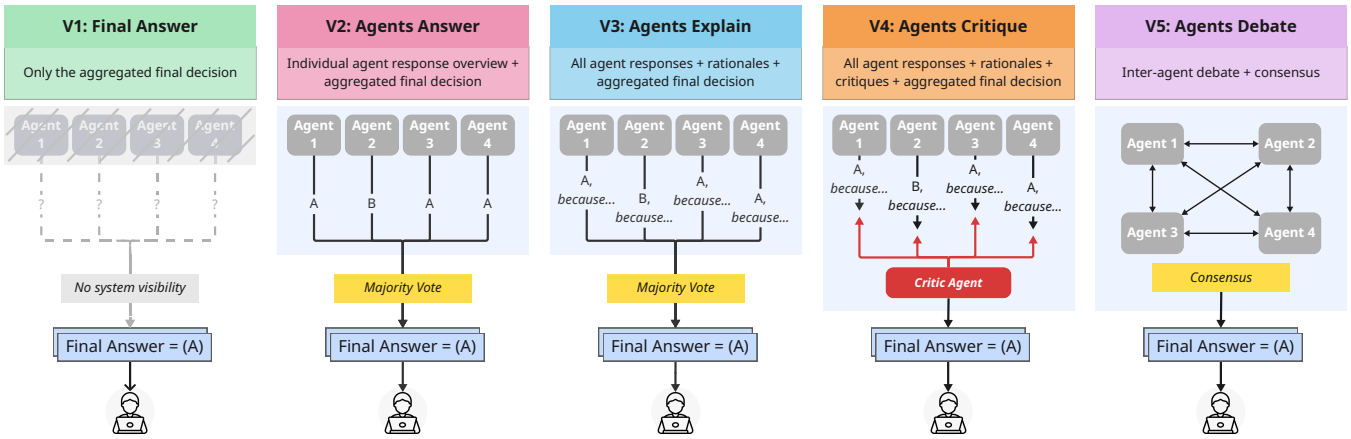

Figure 1: Diagrammatic overview of the five multi-agent Large Language Model (LLM) interface variants (V1–V5) used as exploratory probes in our study. Each variant instantiates a distinct combination of transparency-related design dimensions to surface multi-agent reasoning to the user. *V1 (Final Answer)* shows only the system's final decision, with no agent-level visibility. *V2 (Agents Answer)* displays each agent's individual answer alongside the majority outcome. *V3 (Agents Explain)* extends this by surfacing a brief rationale from each agent, before presenting the majority decision. *V4 (Agents Critique)* introduces a dedicated critic agent who evaluates each agent's response, followed by a summarised final answer. *V5 (Agents Debate)* presents a multi-turn inter-agent discussion, where initially disagreeing agents deliberate and converge on a consensus answer. Together, these variants span a range of process visibility and aggregation mechanisms found in emerging multi-agent LLM systems. We examine how these differences shape user perceptions of transparency and trustworthiness, and the tensions that arise between increased system visibility and its cognitive costs, across different task types.

Abstract

As multi-agent Large Language Models (LLMs) gain traction, designers must consider how to surface their internal reasoning in ways that foster appropriate trust. We present a design-led, qualitative, comparative structured observation study, exploring how users interpret and evaluate transparency in multi-agent LLMs. Participants interacted with five interface variants, each instantiating different combinations of transparency-related design dimensions, across two task types: information-seeking and logical reasoning. We surface participants' mental models, the cues they interpret as signals of transparency and trustworthiness, and how they weigh the costs and benefits of increasing process visibility. Transparency needs were dynamic and context-sensitive, with the ideal "Goldilocks" (i.e., "just right" transparency) level shaped jointly by task demands,

CHI '26, Barcelona, Spain

ACM ISBN 979-8-4007-2278-3/26/04
https://doi.org/10.1145/3772318.3791157

interface affordances, and user characteristics such as task expertise and dispositional AI trust. We highlight tensions between process visibility, information sufficiency, and cognitive effort, and synthesise these insights into design considerations for aligning transparency with user needs in future multi-agent LLM interfaces.

CCS Concepts

• **Human-centered computing → HCI theory, concepts and models**; **Collaborative interaction**.

Keywords

transparency, trust, multi-agent chatbots, multi-agent LLM, reliance, sensemaking, mental models, information seeking, human-AI interaction, human-AI decision-making

ACM Reference Format:

Saumya Pareek, Jarod Govers, Naja Kathrine Kollerup, Emily Wong, Eduardo Velloso, and Jorge Goncalves. 2026. Sensemaking in Multi-Agent LLM Interfaces: How Users Interpret Transparency and Trustworthiness Cues. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 20 pages. https://doi.org/10.1145/3772318.3791157

1 Introduction

An effective collaboration between humans and Large Language Models (LLMs) depends upon both the model capability and its end-users' ability to accurately calibrate their trust in the model [84]. In classical AI systems, trust calibration has been explored through methods aiming to enhance AI transparency, typically via explanations [67, 69], confidence scores [107], or uncertainty indicators [41, 81]. However, these strategies often fall short when applied to LLMs, which generate fluent, open-ended, and seemingly confident outputs that can appear coherent even when incorrect [38, 42, 97]. As a result, transparency-based trust calibration techniques originally developed for classical AI systems may not align with how users interpret or engage with the outputs of generative LLMs, leaving Human-Computer Interaction (HCI) and trust calibration efforts to play catch-up.

In parallel, recent work in machine learning and natural language processing has introduced a new class of LLMs — *multi-agent* LLMs — which comprise multiple AI models that collaborate to reach a more accurate final answer, through mechanisms such as sampling answers and voting, inter-agent deliberation, and casting agents into roles (e.g., a 'critic') to encourage divergent reasoning [53, 58, 95]. While primarily developed to boost model performance, multi-agent systems inherently embody a variety of *epistemic signals*, such as agent disagreement, diverse reasoning paths, critique, and consensus, features that could *also* function as user-facing transparency and trustworthiness signals in such systems.

A small but growing body of HCI research has begun exploring this space, for example, introducing a *'devil's advocate'* agent into group decision-support reduced overreliance in some cases [18], while assigning agents with specialised roles such as *'critics'* and *'summarisers'* helped users recognise diverse opinions [50]. Despite such promising initial signals, the literature remains scattered and exploratory. We still lack a principled understanding of how users interpret, engage with, and desire multi-agent reasoning to be presented in end-user interfaces. A key part of this interpretive process involves users' *mental models* — their internal representations of how a system works, which guide their expectations, understanding, and trust [29, 39, 63]. In systems as complex as multi-agent LLMs, mismatches between user mental models and actual system behaviour could lead to a range of issues, such as miscalibrated trust, misplaced expectations, or inaccurate perceptions of system capability. This raises questions: *how do users make sense of multiple agent responses? What epistemic cues (e.g., disagreement, critique, consensus) do they attend to when evaluating transparency and trustworthiness? And how should multi-agent transparency be designed to support trust calibration?* As multi-agent architectures grow more common, these questions become increasingly central for their thoughtful, human-centred design.

In this work, we treat multi-agent reasoning not just as a performance enhancing back-end technique, but as a novel design material for transparency in multi-agent LLM interfaces. We pose the following research questions (RQs):

- **RQ-1 - Mental Models:** How do users conceptualise a multi-agent LLM system, and how do these mental models shape their interpretation of its outputs?
- – **RQ-2a - Affordances as Signals:** What affordances (e.g., agent visibility, reasoning diversity, consensus, disagreement) of multi-agent LLM systems do users interpret as cues for transparency and trustworthiness?
  - **RQ-2b - Operationalisation Preferences:** What are users' preferences for how multi-agent reasoning should be structured, surfaced, and summarised to gauge system trustworthiness?
- **RQ-3 - Role of Task Type:** How do users' transparency needs vary across different tasks, such as information-seeking versus reasoning-based tasks?

To answer these RQs, we conducted an in-person, design-led qualitative lab study. Our approach was inspired by Mackay and McGrenere [60]'s Comparative Structured Observation (CSO), and card-sorting in design research [25]. We first reviewed existing literature on multi-agent LLMs to identify patterns in how multi-agent transparency is surfaced and presented. This yielded seven design dimensions (e.g., *Number of Agents*, *Role Specialisation*, *Reasoning Visibility*, *Disagreement & Critique*), from which we formalised a multi-agent transparency design space. We then instantiated this space by designing five mock multi-agent interfaces, each operationalising a distinct combination of the identified dimensions. Our participants engaged with all five interfaces in both an information-seeking and reasoning task, and completed comparative sorting and reflection activities. The interfaces served as epistemic probes during our user study, used to prompt participant reflection and elicit rich, comparative insights into how they interpret and evaluate multi-agent transparency (e.g., *why* one variant appears more trustworthy than another).

Our findings reveal that users actively interpret the epistemic signals embedded in multi-agent systems — such as disagreement, critique, and consensus — as cues for deciding when and how much to trust the system. These cues often gave rise to heuristics: for instance, participants frequently interpreted the number of agents as a

proxy for system reliability, or took visible agent consensus as a sign of accuracy. However, these heuristics also proved double-edged: while they offered intuitive shortcuts for assessing trustworthiness, they could also create a false sense of reliability when these signals do not reflect the true quality of the multi-agent reasoning. We highlight the trade-offs these cues introduce and discuss challenges they pose for designing multi-agent transparency.

Furthermore, participants did not desire complete transparency into all agent reasoning. Instead, they sought *contextually sufficient* transparency: just enough information to support their current decision-making without incurring undue mental workload. Across tasks, users consistently gravitated towards the interface that offered a "Goldilocks" (i.e., "just right") level of transparency — where the informational value of the interface was balanced against their cognitive effort. Interestingly, this choice was dynamic, shaped by task type, interface affordances, and individual dispositions. For instance, what felt sufficient in a simple fact-checking task often felt inadequate in a complex reasoning task. We highlight this central cost–benefit tension in designing multi-agent transparency: *participants appreciate signals of internal system deliberation (e.g., disagreement, diverse rationales) to gauge reliability, but do not want to sift through all the deliberation.* They also expressed a desire for progressive disclosure: being able to start with a simple output and request deeper transparency (such as through collapsible interface toggles) *when needed.* Our findings raise design-oriented hypotheses and questions that future work can test more systematically. We make the following theoretical contributions:

- **A foundation for theorising the determinants of user trust in multi-agent LLMs.** We provide a rich account of how users interpret, form mental models around, and desire multi-agent reasoning. We identify key epistemic cues that users instinctively rely on to assess trustworthiness, along with the risks and miscalibrations these cues can produce, and provide suggestions on how to de-risk them. By surfacing which design elements users attend to and why, our findings lay the conceptual groundwork for future hypothesis-driven studies to test how specific multi-agent configurations influence trust calibration and perceived transparency.
- **A reconceptualisation of transparency as *sufficiency*, not volume.** We rethink transparency in multi-agent LLMs not as a binary (transparent vs. opaque), nor as a linear "more-is-better" dial, but rather as *sufficiency judgement*. We show that both too much and too little transparency can undermine trust, challenging the assumption that more system visibility is always better.
- **Design tensions between visibility, interpretability, and cognitive effort.** We make explicit the core tensions between users' desire for process visibility, interpretability, and cognitive effort, arguing that multi-agent transparency must be designed in a manner that is task-sensitive, cognitively manageable, and responsive to users' information needs *in the moment*.
- **Methodologically**, we demonstrate the use of Comparative Structured Observation [60], with interfaces used as epistemic probes, as a promising approach for early-stage, design-led exploration of user needs in novel AI system designs, prior to formal hypothesis testing.

2 Related Work

To design for transparency and (appropriate) trustworthiness in LLMs, it is essential to examine both how users perceive, interpret, and act on AI outputs, and the technical paradigms that shape how AI agents can think, deliberate, and generate responses. In the sections that follow, we first review prior work on trust, reliance, and transparency in traditional (non-LLM) human-AI collaboration, and then outline the unique transparency and trustworthiness challenges posed by the more fluent and conversational LLMs. We then discuss multi-agent reasoning techniques, originally developed as a strategy to boost raw LLM performance, and highlight their untapped potential as user-facing transparency mechanisms in human-LLM interaction contexts.

2.1 Trust and Transparency in AI-Assisted Decision-Making

Trust is a foundational concept in human-AI collaboration. It shapes whether, when, and how users engage with AI-generated outputs, especially in contexts with risk and uncertainty [49, 73]. Following Lee and See [49], we define trust as *"an attitude that an agent will help achieve an individual's goals in a situation characterised by uncertainty and vulnerability."*

In AI-assisted decision-making, the effectiveness of collaboration hinges not only on the AI's capabilities, but also on whether users know when and how much to trust the AI system [84]. This process, known as trust calibration, refers to regulating human trust to align with AI competence, i.e., trusting it when warranted and withholding trust when it is likely to err. Miscalibrated trust, whether as over-trust (trusting incorrect AI outputs) or under-trust (not trusting correct AI outputs), can lead to poor decision quality and negate the very benefits of AI-assisted decision-making [5, 67, 92]. Lee and Moray [48] identify three core determinants of trust: (1) *Process*: pertains to conveying the AI's rationale behind its decisions in order to foster trust, often facilitated through explanations or rationales; (2) *Performance*: tied to the perceived accuracy of the AI, which may differ from actual system accuracy; and (3) *Purpose*: deals with the perceived intentions behind designing the system and the intent of the system, whether aligned with users' goals or aiming to deceive them.

2.1.1 Transparency as a Mechanism for Trust Calibration. Building on these determinants, AI transparency has emerged as a central design strategy to support trust calibration. Transparency, in this context, refers to the extent to which users can inspect, understand, or contextualise an AI's behaviour to help them make informed trust judgements [89, 106]. Following prior work on *"just-enough"* explanation design, we operationalise *perceived transparency* as the felt sufficiency of process insight — i.e., whether the information provided about how an AI answer was generated feels right for the user's goal and context [35, 43, 45]. This framing aligns with research showing that users respond best to explanations that balance completeness with cognitive usability [43, 45]. Without adequate transparency into the operations of an AI system, users lack the necessary cues to gauge whether AI is acting accurately,

benevolently, and in alignment with their goals — key elements that underpin trust [49].

Transparency mechanisms typically target two aspects of AI systems: their *performance* and their *process*. Performance transparency focusses on surfacing cues about how accurate or reliable the AI is, through accuracy scores [32, 102, 104], confidence estimates [107], or uncertainty markers [7, 41, 81], while process transparency focuses on making the AI's decision-making more interpretable, often via explanations [3, 67, 69] and insights into model internals [66, 70].

However, these approaches are not without limitations. The presence of explanations can lend (unwarranted) credibility to (incorrect) AI outputs, or anchor users to the AI output and discourage critical thinking, highlighting "pitfalls of explainability" [5, 23, 67]. Similarly, system performance indicators are not always interpreted correctly and can be overshadowed by subjective perceptions of AI reliability during a task, which may diverge from reality [102, 103]. Even uncertainty information, which is often intended to serve as a proxy for (expected) AI performance, can have unintended effects: for instance, being interpreted as a signal of transparency and honesty by the AI designers, thereby boosting trust (even when unwarranted) [7, 36, 65].

2.1.2 Transparency Challenges in LLMs. The limitations of conventional AI transparency strategies become especially pronounced in the context of Large Language Models (LLMs). Generative AI systems like LLMs represent a paradigm shift in human-AI interaction — from terse predictions by traditional classification models to fluent and seemingly confident natural language responses. While this shift offers new possibilities, it also hinders users' ability to assess trustworthiness. Prior research suggests that LLM fluency can mask its response inaccuracies, making it harder for users to determine when to trust or question the model's output [38, 42, 97]. This emerging interaction paradigm challenges traditional approaches to AI transparency and trustworthiness. Mechanisms such as confidence scores, uncertainty indicators, or post-hoc explanations, originally developed for classification or retrieval-based models, may no longer align with how users interpret or engage with open-ended LLM responses. Moreover, the longstanding assumption that *more transparency improves reliance* may not even hold in LLM contexts: users' transparency needs may differ depending on the nature of the task. For instance, in some cases, users may simply want a definitive answer from the LLM and ignore auxiliary transparency information; in others, they may value deeper insight into how the response was generated and why.

Although recent HCI studies have begun exploring trust and reliance in LLM-based systems [30, 31, 42, 91], much of this work has borrowed transparency boosting strategies and design cues from traditional human-AI interaction contexts, such as natural language or visual signals for uncertainty [41, 90], explanations, source citations, and internal contradictions [40, 42]. However, little is known about the cues users naturally notice, seek, or interpret as signals of transparency and trustworthiness in LLM outputs, highlighting a pressing need to understand users' transparency needs and preferences in such novel decision-making contexts. This study adopts a discovery-oriented approach to examine how users conceptualise and engage with transparency signals in LLM interfaces. Taking a bottom-up approach, we focus on the cues, features, and reasoning strategies that shape users' interpretations, with particular attention to multi-agent settings where epistemic signals such as debates, critiques, and consensus may emerge.

2.2 Multi-Agent Reasoning Structures

Recent work in Large Language Models (LLMs) has introduced systems composed of multiple agents — either instantiated as distinct models or as role-specialised versions of the same model — that collaborate to reason over a shared problem. These multi-agent approaches aim to improve performance and robustness by introducing diversity in how reasoning is generated, evaluated, and synthesised. Traditionally, multi-agent systems are defined as *"a collection of, possibly heterogeneous, computational entities, having their own problem-solving capabilities and which are able to interact in order to reach an overall goal"* [64]. In the context of human-centred AI, this concept has been adapted to include systems where agents exhibit attributes such as inter-agent collaboration, communication, and coordination [15, 20, 51]. In our work, we follow this framing to study multi-agent LLM systems designed for decision-support, where agents collectively reason, critique, or consolidate their outputs to assist users.

In NLP and ML research, multi-agent reasoning methods have shown promise in boosting AI performance on tasks such as maths word problems and multi-step inference. For instance, Du et al. [22] found that multiple agents debating the final answer to arithmetic problems outperformed single-agent baselines. Similarly, Li et al. [53] demonstrated that simply scaling up the number of sampled agents and choosing the final answer through voting improved accuracy. Recent evaluations have also explored multi-persona prompting, casting AI agents into roles such as *"angel vs. devil"* or an overseeing *"judge"* to encourage more divergent reasoning and accurate solutions [55, 82]. While these approaches aim to improve LLM reasoning performance, the very mechanisms they employ, such as simulated deliberation, agent disagreement, or jury-like consensus, may shape users' perceptions of trustworthiness, *perhaps even in ways that are not aligned with actual model accuracy*. This underscores the need to examine how such multi-agent structures could function as user-facing transparency mechanisms.

2.2.1 Multi-Agent Reasoning in HCI. A handful of recent HCI studies have begun exploring how AI systems involving multiple rather than a single agent might impact human-AI interaction, though this work remains scattered and exploratory. For example, Chiang et al. [18] introduced a *"devil's advocate"* agent into an AI-assisted group decision-making process, finding that adversarial questioning by this agent can reduce overreliance in some cases. Similarly, Swoopes et al. [87] presented users with ten different AI responses to the same query and found that users appreciated the ability to cross-verify responses, though how this impacted users' trust or decision performance was not studied. Other relevant works have implemented multi-agent systems where they give agents distinct roles, such as a *'summariser'*, *'critic'*, and *'redundancy-checker'* [50], and found how inconsistencies within multiple AI responses shape perceived AI competence [51].

Despite these promising signals, there is currently no unified framework for understanding how multi-agent reasoning should

be operationalised, structured, or surfaced in human-AI interfaces. Multi-agent systems inherently generate rich epistemic signals, such as disagreement, consensus, and visibility into diverse agent rationales, that likely shape user perceptions of transparency and trustworthiness in understudied ways. Existing work tends to investigate isolated interaction patterns or specific agent configurations without systematically connecting them to user goals, transparency needs, and perceived trustworthiness. To date, no work has systematically examined how different structures of multi-agent reasoning affect perceived transparency and trustworthiness, or what users' mental models of multi-agent LLMs are like — internal representations of how a system works that can guide expectations, understanding, and trust [29, 39, 63]. We address this gap by treating multi-agent reasoning structures as a novel design material for transparency. We ask: *how do users interpret these structures? What epistemic signals do they rely on to gauge trustworthiness, and what additional information do they seek? Is one level of transparency more desirable than others — for instance, should interfaces surface every agent's reasoning path, or present only a distilled consensus?*

2.3 Task Characteristics and Their Influence on Transparency Needs

Task-level characteristics such as complexity and uncertainty can strongly influence how users engage with AI systems. Prior work has distinguished between fact-based, information-seeking tasks and more open-ended, reasoning-based tasks, inherently varying along dimensions such as complexity, ambiguity, uncertainty, and verifiability [4, 16, 76, 92]. For example, Salimzadeh et al. [76] classified tasks as diagnostic (low uncertainty, more factual) and prognostic (high uncertainty, more inferential), showing how these differences impacted reliance on AI. In other HCI studies, tasks along this spectrum have ranged from fact-based general knowledge questions, for example, *"Has Paris hosted the Summer Olympics more times than Tokyo?"* [42] or *"Which country in Europe has the most Nobel Laureates in science?"* [51] to more cognitively demanding, LSAT-style logical reasoning tasks [5, 8].

Building on this, we posit that users' transparency needs may also vary across task types: for information-seeking tasks, surfacing agent consensus might be most desirable, whereas exposing agent debate and critique could help users navigate the ambiguity in reasoning-based tasks. Therefore, in this work, we explore both information-seeking and reasoning-based tasks as a lens through which to understand user expectations and transparency preferences in multi-agent systems.

3 Deriving A Design Space for Surfacing Multi-Agent Reasoning

Multi-agent LLM systems vary widely in how they surface internal reasoning — exposing multiple agents, disagreement, critique, or collaborative deliberation. While these features offer rich epistemic signals, it remains unclear how their presentation shapes users' perceptions of transparency and trust. To address this, we systematically mapped patterns in existing multi-agent LLM interfaces in the literature, and distilled a set of design dimensions that characterise how reasoning is exposed to end users. This section outlines (1) our scoping and screening process for relevant literature, (2) the core dimensions that emerged, and (3) how these dimensions informed the construction of interface used as our study stimuli.

3.1 Gathering Design Requirements for Multi-Agent Interfaces

Following prior HCI work that derives design considerations through targeted literature analyses [9, 19, 71], we conducted a design-oriented scoping review of multi-agent LLM systems in literature. Our goal was to uncover recurring interface patterns, ensemble configurations, and reasoning structures that shape how such systems expose their inner workings to end users. We began with a targeted keyword search in the ACM Digital Library as it indexes key HCI venues (e.g., CHI, CSCW, IUI, FAccT), using the following query: `("multi-agent" OR "multiple agents") AND ("LLM" OR "large language model") AND ("deliberation" OR "critique" OR "debate" OR "rationale" OR "collaboration")`. This returned 691 results (687 published between 2022-2025). After restricting to full research articles and excluding extended abstracts, magazine articles, and workshop papers, 235 remained. We manually screened titles and abstracts for relevance, yielding 92 papers. We supplemented this pool with six arXiv preprints and four additional relevant papers known to us, bringing our initial pool to 102.

We then applied the following inclusion criteria: (1) involves *multiple* LLM agents; (2) described agent behaviours or interactions (e.g., response form, voting, disagreement, summarisation, collaboration); and (3) reports or illustrates system-level or user-facing design details (e.g., interface layout, process flows), and not just back-end architectures or benchmarking results. This screening reduced our set to 33 papers (including two surveys). We then performed a full-text, open (bottom-up) coding of these 33 manuscripts, tagging observed interface patterns such as the ensemble size, until no new codes emerged. Through affinity mapping and team discussions, we iteratively clustered our codes into mid-level categories (e.g., reasoning visibility, aggregation), and then consolidated into seven interface-level design dimensions (D1–D7). The complete codebook and list of included papers are available in our supplementary materials.

3.2 Core Design Dimensions of Multi-Agent Interfaces

From analysing the 33 papers, we distilled seven interface-level design dimensions that govern how multi-agent LLM systems are structured, and how they surface their internal reasoning to end-users:

Agent Composition: *How many agents are involved, and what roles do they play?*

- ***D1 — Number of Agents:*** Ensemble size (typically 2–6).
- ***D2 — Role Specialisation:*** Whether the system involves *homogeneous* agents with identical roles, or *heterogeneous* ones with functionally distinct roles, such as *critics* who challenge others' responses or *summarisers*.

Reasoning Visibility: *How much of each agent's internal reasoning is revealed, and in what form?*

- ***D3 — Response Format:*** Output granularity, ranging from opaque, aggregated final answers only to individual agent

answers with or without explanations, or visible chain-of-thought reasoning.
- ***D4 — Disagreement & Critique:*** Whether and how systems surface internal dissent, with some systems not revealing disagreement at all, while others making it salient implicitly via response diversity, or explicitly via peer critiques/debates or dedicated critic agents.

Interaction Topology: *How do agents interact, and how is this shown?*
- ***D5 — Interaction Paradigm:*** How agents interact with one another: parallel (independent) responses, sequential responses, or multi-turn deliberations where agents build upon one another's responses.
- ***D6 — Information Flow Architecture:*** How information flows between agents: *flat* configurations (all agents contribute equally), *pipeline* flows (e.g., agent A proposes → B critiques → C revises), or *hierarchical* flows where a lead agent (e.g., *"judge"*, *"coordinator"*) synthesises inputs from others.

Output Aggregation: *How is the final system recommendation derived?*
- ***D7 — Aggregation Mechanism:*** How individual agent outputs are consolidated into a final response. Some systems surface all agent responses, while others employ a *meta-agent* to produce an aggregated output: via voting, summarisation, or post-debate consensus.

3.3 Operationalising the Design Space as Study Stimuli

The design space above illustrates the diverse ways in which multi-agent systems can be instantiated — varying in the number of agents, their roles, interaction patterns, reasoning visibility, how disagreement is surfaced, and more. *But how do end-users interpret these epistemic signals and cues? What do they seek when assessing a system's transparency and trustworthiness? And how can multi-agent interfaces be designed to align with these user needs?*

To investigate these questions, we instantiated the design space through a set of multi-agent interface variants, each constructed using distinct combinations of the identified design dimensions (D1-D7). Rather than exhaustively testing all possible design space configurations, we curated a tractable set of interface variants grounded in the surveyed multi-agent literature, designed to expose a range of cognitive and epistemic strategies that participants may employ when assessing trustworthiness in such contexts. We sought to capture key patterns and salient configurations observed across prior work (e.g., opaque responses, majority voting, agent-level explanations, dedicated critics, peer debate), while ensuring each variant differed meaningfully along the dimensions so participants could compare them and articulate trade-offs.

These interfaces served as epistemic probes, evaluative design artefacts embedded with some theoretical intent, designed to provoke reflection, surface mental models, and elicit rich insight into users' beliefs and attitudes [28, 93]. Such probes are widely used in design-led research to explore user perceptions of emerging technologies [59, 77]. We designed the following five representative interface variants, V1-V5, each embodying a distinct approach to multi-agent transparency and foregrounding particular epistemic cues:

- **V1 (*Final Answer*)** represents a fully opaque baseline, presenting only the final system output with no visible multi-agent structure, individual agent responses, rationales, or aggregation logic. It seeks to expose how users evaluate trust in a multi-agent system in the absence of process visibility.
- **V2 (*Agents Answer*)** reveals each individual agent's decision (without explanations), with the final system output decided through the majority outcome. This interface highlights epistemic cues such as response diversity and majority voting, reflecting LLM architectures that employ such sampling and voting strategies [96, 101].
- **V3 (*Agents Explain*)** extends V2 by including brief agent-level explanations alongside decisions. This makes agents' reasoning diversity and any (dis)agreement amongst them visible, helping users contrast and compare divergent justifications, representing multi-agent systems where each agent explains itself [17, 86, 100].
- **V4 (*Critique*)** presents individual responses and explanations as V3, and adds a dedicated critic agent that evaluates and critiques its peers' responses. This variant makes disagreement and dissent explicit, and reflects multi-agent architectures that incorporate specialised critic or verifier agents [6, 24, 94, 109].
- **V5 (*Debate*)** presents a multi-turn peer-to-peer debate in which agents engage in dialogue and converge on a consensus answer. This highlights deliberation, evolving reasoning, and consensus formation, reflecting systems that employ multi-agent debates and discussions [52, 54, 82, 108].

These design choices enabled us to examine how the distinct epistemic cues embedded and surfaced in these interfaces shape users' mental models and their perceptions of transparency and trust, across tasks. In particular, the five variants enabled us to surface a wide range of user strategies: evaluating trust in opaque system outputs *(V1)*, making sense of response diversity and majority heuristics *(V2)*, comparing multi(ple)-agent explanations and their reasoning diversity *(V3)*, evaluating critiques and disagreement from a dedicated critic agent *(V4)*, and interpreting an evolving debate that reaches consensus *(V5)*.

Table 1 summarises how the variants (V1-V5) map onto the seven design dimensions (D1–D7). While we varied dimensions D2–D7 across interfaces, D1 (Number of Agents) was intentionally held constant at four to support meaningful comparisons between interfaces. Ensemble size shapes nearly all other visible aspects of the interface, such as the number of agent responses, explanations, or critiques, so varying it alongside D2–D7 would have compromised our ability to use each interface variant as a focused probe, and required an impractically large number of interface variants. Instead, we explored D1 in a focused comparison task, where participants viewed side-by-side mockups of otherwise identical interfaces with *four* agents (as used throughout V1–V5) versus *eight*, chosen to ensure sufficient contrast between the probes, and reflected on how ensemble size shaped their perceptions of the system (see subsection 4.3, Stage 4).

We emphasise that our aim was not to identify a "best" interface, but to use representative, theoretically grounded variants as stimuli, probing how users interpret and evaluate transparency and trustworthiness in multi-agent LLMs, and uncovering a range of their sensemaking strategies in the process.

4 Method

To examine how users interpret, evaluate, and form preferences around transparency in multi-agent LLMs, we conducted an in-person, qualitative, design-led lab study. In this emerging and under-theorised design space, where user needs and interpretive strategies remain to be fully understood, such an exploratory qualitative approach was suitable for our aims. Our aim was to surface rich accounts of users' mental models of multi-agent LLMs *(RQ-1)*, identify the affordances they treat as signals of transparency and trustworthiness *(RQ-2a)*, elicit preferences for how those affordances should be designed and presented *(RQ-2b)*, and understand how these needs vary by task *(RQ-3)*. Participants interacted with five multi-agent LLM interfaces (V1-V5), each instantiating a distinct combination of the design dimensions identified in our literature review (see Table 1).

We drew methodological inspiration from prior work in human–AI collaboration [75, 83, 98] and card-sorting methods in design research [25], and followed the Comparative Structured Observation (CSO) approach proposed by Mackay and McGrenere [60]. They define CSO as *"an interventionist, qualitative method for assessing and advancing a design concept where researchers observe participants as they compare and reflect deeply upon their experiences with selected design variants, exposure to which is structured [...]."* CSO is explicitly interpretivist and design-oriented: it treats qualitative, comparative reflections as the *primary* data, and any quantitative measures as *secondary*, with the goal of advancing a design concept rather than testing formal hypotheses or estimating effect sizes. CSO was thus well-suited to our goal of building theory around how multi-agent reasoning should be surfaced and what preferences, tensions, and trade-offs emerge, rather than hypothesis-testing interfaces to find the "best" one. Further, CSO enabled us to use interfaces that operationalised the design dimensions as *epistemic probes* [59, 77, 93] to prompt participant discussion and reflection, eliciting comparative judgements and trade-offs (e.g., *why* one variant appears more trustworthy than another).

4.1 Task Selection and Design

Building on prior HCI research that differentiates tasks by complexity, ambiguity, and uncertainty [5, 16, 76, 92], we selected two contrasting question-answering tasks that reflect common real-world use cases for LLMs: information-seeking and logical-reasoning. Across both tasks, we sought questions that (1) did not require specialised domain knowledge, allowing the general population to reason about them, yet (2) were non-trivial and sufficiently challenging to warrant AI assistance. Information-seeking tasks were binary (yes/no) general-knowledge questions with known answers (e.g., *"Has Paris hosted the Summer Olympics more times than Tokyo?"*), adapted from prior work on AI trust and uncertainty [42, 76]. Reasoning tasks, by contrast, involved more uncertainty and subjective interpretation. For this, we used LSAT [1] logical reasoning questions (multiple-choice questions with five options, A–E), which assess general aptitude and deductive reasoning rather than legal knowledge. These questions have been widely used in prior AI-assisted decision-making research [5, 8], and we selected questions marked 'difficult' in LSAT preparation materials [1, 2].

Together, these two tasks enhance the ecological validity of our study by reflecting common real-world contexts where people query LLMs for both knowledge lookup and reasoning, and let us examine how users' mental models and transparency preferences shift with (the cognitive and informational demands of) the different tasks.

4.2 Interface Design

To instantiate the design space defined in §3.2, we developed five interactive mock interface variants (V1-V5), implemented using HTML/JavaScript and designed to resemble a contemporary chat-based LLM platform. For visual consistency, we standardised typography, layout, and structural elements across all variants, and for added realism, included animations commonly observed in real LLM systems, such as staggered "typing" reveals.

All interfaces presented pre-generated outputs held constant across participants. V1 (*Final Answer*) and V2 (*Agents Answer*) did not present agent rationales, while V3 (*Agents Explain*), V4 (*Critique*), and V5 (*Debate*) required text generation: agent explanations in V3, explanations plus critique in V4, and a multi-turn convergent debate in V5. For information-seeking tasks, we manually crafted agent outputs for both correct and incorrect responses. For reasoning tasks, we sourced correct answers and explanations from official LSAT preparation materials [2], which were then provided to OpenAI's GPT-4o [2] to condense into concise rationales. We further prompted GPT-4o to generate stylistically similar but plausible rationales for agents selecting incorrect task responses, as well as critiques for all rationales.

To support accurate stimuli generation, we iteratively refined our prompt to specify the desired structure, tone, length, and logical characteristics of each rationale and critique, and included additional contextual details about the experiment. The full final prompt is presented in Appendix A. All generated outputs were then manually reviewed and refined by three authors in a two-step process. In the first pass, we checked for plausibility and intended (mis)alignment, ensuring that correct rationales were faithful to the task while incorrect rationales were believable yet misaligned with the task logic (V3), including critiques that surfaced plausible limitations of each agent's reasoning (V4). In the second pass, we ensured stylistic consistency across responses and tasks (e.g., length, tone). Lastly, the V5 debate sequences were manually authored by the research team using the vetted V3/V4 rationales as inputs, ensuring a coherent multi-turn dialogue without introducing new facts. Final stimuli are available in our supplementary materials.

[1] The Law School Admission Test (LSAT) is a standardised test administered by the Law School Admission Council (LSAC).
[2] https://openai.com/index/hello-gpt-4o/

Variant	What the Multi-Agent System displays	D1	D2	D3	D4	D5	D6	D7
		Number of Agents	Role Spec.	Response Format	Disagreement & Critique	Interaction Paradigm	Information Flow	Aggregation Mechanism
V1	Only the aggregated response	4	No	Answer only	None	Hidden	Hidden	Hidden
V2	Individual agent response overview + aggregated response	4	No	Answer + vote count	Through response diversity (summary only)	Parallel	Flat	Majority vote
V3	All agent responses + explanations + aggregated response	4	No	Individual agent responses + explanation	Through response diversity (visible)	Parallel	Flat	Majority vote
V4	All agent responses + critique + aggregated response	4 + Critic	Yes	Individual agent responses + explanation + critique	Through a dedicated critic agent	Sequential	Hierarchical	Summarisation
V5	Individual agent responses + explanations + inter-agent debate + summarised consensus	4	No	Individual agent responses + explanation + reasoning evolution	Through peer debate & convergence	Multi-turn	Hierarchical	Consensus

Table 1: Interface variants (V1—V5) and their operationalisation using the seven identified design dimensions (D1—D7). Shorthand legend: V1 = Final Answer; V2 = Agents Answer; V3 = Agents Explain; V4 = Critique; V5 = Debate.

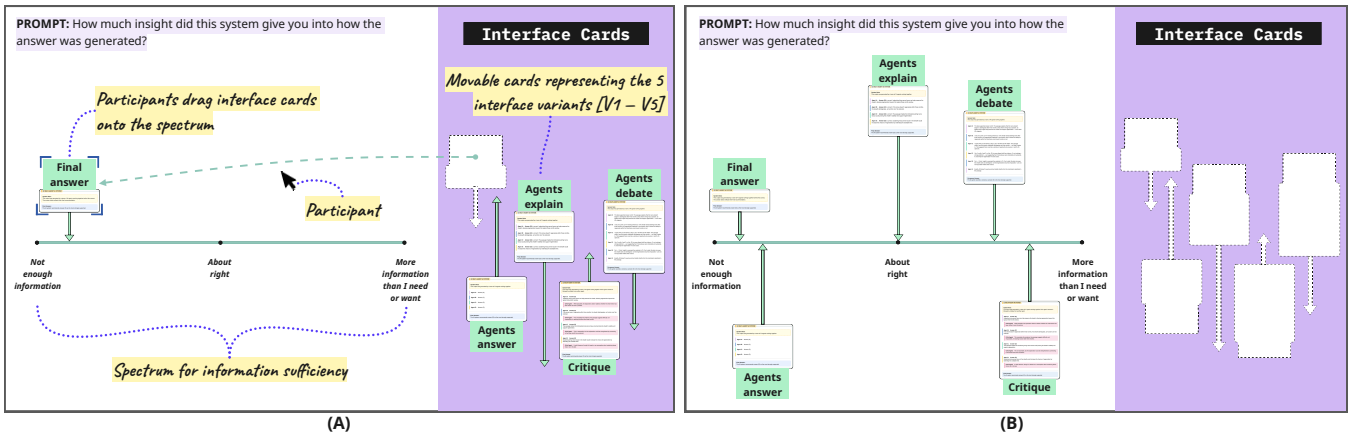

Figure 2: A before (A) and after (B) snapshot of the digital card-sorting activity participants completed for the Transparency scale, informed by Comparative Structured Observation [60] and inspired by card-sorting design-led research [25]. Participants arranged cards representing the five interface variants (V1–V5) along a spectrum of perceived transparency (*"How much insight did this system give you into how the answer was generated?"*), anchored from *"not enough information"* to *"more than I need or want."* This activity formed the basis of our implementation of the Comparative Structured Observation method, with the interface cards serving as elicitation probes, giving participants a concrete way to externalise judgements, compare variants (and their design dimensions) side by side, and articulate trade-offs. The same activity was repeated for two additional scales capturing *perceived helpfulness* and *perceived reliability.*

4.3 Study Procedure

We ran face-to-face study sessions in a quiet lab setting using a laptop. Participants interacted with the interfaces and engaged in comparative and reflective activities, outlined below. Each session was recorded with permission. The study unfolded as follows:

Stage 1. Participants were introduced to the study context and provided a short pre-task questionnaire, capturing demographic data (age, gender), AI usage frequency, propensity to trust automation (TiA-PtT), and AI literacy (see Appendix B for scales and measures). As a warm up for the think-aloud, participants interacted with ChatGPT by asking it a light question (e.g., a historical fact) and then describing what they believed happened "behind the scenes."

Stage 2. Next, participants engaged with the tasks and the five interface variants. The task type was a within-subjects factor: each participant completed both tasks (information-seeking and logical

reasoning), presented in a counterbalanced order. When introducing each task, we asked the participant to imagine a usage context tailored to the task type (e.g., *'looking up a fact to settle a debate during dinner with friends'* for the *information-seeking task*, and *'coming across a tricky question when preparing for a standardised exam'* for the *logical reasoning task*). Each task began with participants answering the question independently and recording their confidence (0–100). They were then shown all five interface variants, one by one, in randomised order. While interacting with each variant, participants were prompted to think aloud, followed by a brief semi-structured interview probing their impressions, what elements stood out, how they believed the system produced its answer, whether they would follow its advice, and what they would change about the interface. This stage helped us surface mental models and interpretations of the design dimensions at play.

Stage 3. After engaging with all five variants for a task, participants completed three comparative card-sorting activities on a digital canvas in Miro [3]. See Figure 2 for an illustration of one such activity. The canvas contained movable cards representing each interface variant (V1-V5), consisting of a screenshot of the interface and a brief title to anchor participants' memory and assist with recall. Following design-led approaches based on card sorts [25], for each sort, participants arranged cards along spectrums capturing *perceived transparency* (*"how much insight did this system give you into how the answer was generated?"*, anchored from *"not enough information"* to *"more information than I need or want"*), *perceived interface helpfulness* in the task's stated context (*"how helpful would this design be if you were using it in a real-world context – [preparing for a standardised exam (logical reasoning task)] or [looking up a fact during dinner with friends (info seeking task)?]"*), anchored from *"least helpful"* to *"most helpful"*, and *perceived interface reliability* in the task's stated context, anchored from *"least reliable – I would not follow its answer"* to *"most reliable – I would follow its answer"*. Following the Comparative Structured Observation method, we utilised each sort order as a structured prompt to probe participants and elicit comparative reasoning, both within each sort (e.g., *"What makes [V3] more helpful in this scenario than [V2]?"*) and across sorts, for example, when an interface ranked high on transparency but low on helpfulness, we probed this tension further by inviting participants to articulate why (e.g., *"What makes [V2] more informative yet less helpful in this scenario?"*).

The card-sorting stage served as the anchor for Comparative Structured Observation in this study. By asking participants to arrange and then explain the relative placement of designs, we were able to elicit fine-grained insights in how they interpreted transparency, usefulness, and reliability, and why and how they distinguished between the interfaces on the basis of these constructs. The structured comparisons gave our participants a concrete basis for reflection, while the facilitator's follow-up questions within and across sorts encouraged them to articulate the reasoning behind their choices, surface tensions between dimensions, and make explicit the trade-offs that might otherwise remain implicit. We repeated the same protocol sequence for the second task.

Stage 4. After the second task, participants completed a brief exploratory comparison of ensemble size (Design Dimension D1 - Number of Agents). We presented side-by-side mockups of otherwise identical interfaces with *four* agents (as used throughout) versus *eight* agents, and probed how the additional agents shaped participants' impressions of the system and perceived complexity. Since D1 was held constant across design variants V1-V5, this closing activity isolated how ensemble size can influence perceptions of multi-agent systems, helping us surface thresholds for *'how many is too many'* and whether preferences vary by task type.

The session concluded with a short, semi-structured wrap-up interview inviting participants to reflect on their experiences and preferences across tasks, giving us more contextualised insights into the tensions we were interested in and better understand how participants' information expectations shifted between reasoning and information-seeking tasks.

4.4 Participants and Recruitment

We recruited 12 participants (6 men and 6 women) through our university's notice board and snowball sampling, ensuring diversity in educational backgrounds and levels of AI familiarity. Our sample size aligns with typical sizes reported in similar within-subject Comparative Structured Observation (CSO) studies in HCI (e.g., [21, 88]), and follows established guidance in qualitative research to achieve thematic saturation [11, 27]. Eligibility criteria required participants to be fluent in English and to have at least some prior experience with Large Language Model (LLM) chatbots (e.g., ChatGPT, Gemini, etc). All participants provided informed consent before taking part and were compensated approximately US$16 for their time. The study received approval from our university's Human Ethics Committee, and sessions lasted approximately 65 minutes each.

4.5 Data Analysis

We adopted a *qualitative-first* approach, consistent with [60]'s Comparative Structured Observation (CSO). For qualitative analysis of our think-aloud and interview data, we followed Braun and Clarke's six-phase Reflexive Thematic Analysis [10, 12]. Our approach was both deductive and inductive: we began with a structured coding framework grounded in our research questions, study objectives, and identified design dimensions/affordances, while also remaining open to unanticipated patterns, tensions, and interpretations that emerged from the data itself. All interview recordings were transcribed using a research tool called Dovetail [4] and manually corrected for transcription errors. We used participants' screen recordings to clarify deictic references in audio recordings (e.g., "*this* interface is helpful") and ensure accurate tagging. The first author familiarised themselves with the full set of transcripts through multiple close readings, and iteratively tagged utterances using a dynamic codebook. The author team collaboratively reviewed, refined, and merged overlapping low-level codes, resolving ambiguities through discussion, before forming high-level themes. The resulting set of themes forms the basis of our findings, presented in the following section.

[3] https://miro.com/index/

[4] https://dovetail.com/

5 Findings

Our analysis focuses on how participants interpreted and evaluated the design dimensions operationalised by the multi-agent interfaces, and how these interpretations shaped their perceptions of transparency and trustworthiness in multi-agent systems. We first provide an overview of our participants and then present our findings.

5.1 Participant Overview

Table C.1 (Appendix C) summarises our participants' demographics and dispositional characteristics. Our 12 participants represented a mix of backgrounds and levels of AI familiarity. Most participants described themselves as frequent users of LLMs, often interacting with systems like ChatGPT multiple times a day, while a smaller group reported occasional use (a few times per week or month). Participants also varied in both their propensity to trust automation, and in their self-reported AI literacy. Together, these observations suggest that our sample included participants spanning the spectrum of AI familiarity and dispositional trust tendencies.

Before interacting with the AI interfaces, participants answered the information-seeking and logical reasoning tasks independently and rated their confidence (0–100). We find that their confidence was low in both domains: $M = 40.8$ ($SD = 22.3$) for the reasoning task and $M = 40.3$ ($SD = 17.1$) for the information-seeking task. These scores indicate that participants had limited prior knowledge in both domains, supporting the suitability of our tasks for studying AI-assisted decision-making.

5.2 RQ1: Mental Models of Multi-Agent Systems

To understand how participants conceptualised a multi-agent LLM and how this shaped their interpretations, we examined their spontaneous explanations and metaphorical framings.

5.2.1 ***System metaphors.*** Participants frequently drew on familiar metaphors (e.g., teams, panels) to describe the multi-agent system, and attributed varying levels of autonomy, coordination, and capability to the agents. Several imagined the system as a *"panel"* of agents, each independently generating and voting on answers (*P3*, *P4*, *P7*), while others saw it as a team with roles: *"an agent who reads the query [...] another who does the search [...] another presents the final answer"* (*P7*). For V4 (Critique), some described the system as composed of *"multiple teams trying to arrive at the same answer, taking different routes,"* with an additional *"team that goes into each of those and tries to break down that argument"* (*P1*).

Even when uncertain, participants actively constructed mental models of how the system arrived at a final answer. For example, *P8* hypothesised *"some kind of numerical score"* or probability-based aggregation mechanism, while *P4* inferred a majority vote logic: *"Each individual agent tells the system what their answer is, and then the system will go with the most picked answer."*

5.2.2 ***Anthropomorphised framings of the multi-agent system.*** Several participants anthropomorphised the agent ensembles, attributing to them human-like reasoning or motives: *"I guess the system has these little people, or I mean, agents, who go through all the computation [...]"* (*P2*). More specifically, V5 (Debate) was described to be similar to *"listening to a conversation with other humans"* (*P7*), while V4 (Critique) was seen *"like a human judge analysing a group of people"* (*P12*).

Interestingly, one participant also imagined a mixture of expert systems at play: *"There's multiple AI agents working in the background to give me an answer. If you treat them as people, then one of them could be good at biology, one at logic, one at comprehension. The system received my question, passed it on to the AI who is the most knowledgable [...], and then that AI has generated the answer for me"* (*P9*). Even participants who did not explicitly equate the agents to human-like entities still verbalised their mental models in anthropomorphised terms: *"peers who can think together"* (*P7*).

5.2.3 ***Agents' model/vendor identity served as a heuristic for reliability.*** Four participants expressed a desire to know more about the individual agents' underlying model or vendor, as a means of evaluating trustworthiness: *"I would want information about each model and vendor to decide if it is believable"* (*P6*). These participants expressed that newer or more capable models reduced their need for transparency: *"If the agents in [V2 (Agents Answer)] were all GPT-5 level, I would find it very reliable, I would not even need an explanation with such newer models. But if the agents were [from the] beginning of ChatGPT, then I would not be so sure and need more explanation"* (*P10*).

5.3 RQ2a: Affordances as Signals, and RQ2b: Operationalisation Preferences

5.3.1 ***D1: Number of Agents.***

Participants saw the number of agents in the multi-agent system (ensemble size) as a signal of system capability and reliability, with their preferences revealing trade-offs between reliability and mental workload.

Larger ensembles were seen as more reliable, but also introduced mental workload. A vast majority of participants perceived larger ensembles as more capable or credible because they entailed *"lots of little people working towards one goal, so it seems more powerful and credible"* (*P8*). Similarly, participants also believed systems with more agents would make more reliable decisions: *"With more agents involved, it just gives me more certainty that the answer doesn't come from some random guess, so I trust it more"* (*P11*). However, this was not universally true. A few participants viewed higher agent counts as introducing ambiguity or mental workload: *"I'm not that trusting of AI, I feel the need to go over all of them, so more agents would create more work"* (*P12*).

While participants typically saw agent diversity as valuable to the system's internal reasoning, they wanted to manage how it was surfaced. They wanted to *know* that multiple agents were involved, but did not necessarily want to *see* all their outputs: *"More agents would make the system seem more reliable, if all [the individual responses] are hidden behind the scenes"* (*P5*); *"Use as many [agents] as needed for developers to be confident, but for presentation to me, 4 is OK"* (*P4*).

Ideal ensemble size was seen as task-dependent. Our participants wanted the system to adapt its ensemble size to the complexity of the task: *"For the [info-seeking task], I only need 2 or 3 agents to trust it, but for the [reasoning task], I would need 5 to feel confident*

in the system" (*P3*). One participant pointed out that design choices like odd-numbered ensembles could be useful to avoid 'ties': *"I would never want an even number"* (*P12*).

5.3.2 ***D3: Response Format***.

*Participants generally found **V1 (Final Answer)**, which only presented a final answer without any agent responses / rationales, too opaque to be trustworthy or helpful. The lack of visible process detail made it hard to evaluate the system's reasoning or build trust in its output.*

Aggregated answers without explanation obscured the system's reasoning, reducing trustworthiness. Many participants expressed distrust toward the multi-agent system presented in V1 due to its lack of visible process transparency: *"It just says this is the answer, it doesn't give me any explanation. It's too little, I don't trust it"* (*P11*). Participants desired more insight into the underlying process: *"I would need at least a summary of what different agents said, and how they agreed"* (*P3*). This lack of visible reasoning also made the system not very helpful for sensemaking: *"[V1] just pops one answer in my head and doesn't help me think through it, I have to work out on my own why it might be right"* (*P12*).

A minority valued the speed and decisiveness of an aggregated answer with minimal transparency. Two participants preferred V1 for its clarity and lack of ambiguity, particularly in the simpler info-seeking task: *"[V1] is the perfect amount of information I need for this type of question, is gives me a straightforward answer without hesitating"* (*P8*). Paradoxically, *P4* described feeling more confident because they were not exposed to the inner workings of the system: *"For a simple problem like this, where I trust AI enough in general, [V1] is good enough. [I am] actually more confident with this because I can't see its imperfections"* (*P4*).

__V2 (Agents Answer)__ displayed individual agent responses before presenting the final aggregated answer, but omitted any agent-level rationales. Overall, participants regarded individual agent responses [V2] as a minor improvement over an aggregated system-level answer [V1], but still inappropriate for trust and sensemaking without some explanation or evidence.

Surfacing individual agent answers offered users some transparency, but insufficient for trust and sensemaking. Participants perceived the multi-agent system in V2 to be less opaque than V1, and a few appreciated the glimpse into agent-level output: *"It gives me a bit more like, hey, us agents have different ideas and then this is the majority"* (*P2*). Still, this increase in surface transparency was widely considered insufficient: *"[...] a sliver more of information, but I still don't know how each individually reached the conclusion"* (*P5*). Without explanations, participants questioned the system's trustworthiness: *"Sometimes ChatGPT will just make up answers. So [V2] is helpful but more evidence is needed for me to trust it"* (*P3*). Several participants also expressed that V2 did little to support their reasoning and sensemaking: *"It doesn't really help me reason [...], not much more helpful or reliable than [V1]"* (*P4*). Not seeing agent-level rationales was even more undesirable in the reasoning tasks: *"For a task like this, I think the reasoning matters a lot. I don't like to rote learn. I like to actually understand why things happen"* (*P11*).

We report participants' interpretations of the (implicit) disagreement signal (D4) and the majority-vote cue (D7) present in the V2 interface in §5.3.3 and §5.3.5.

__V3 (Agents Explain)__ displayed each agent's individual response alongside a brief rationale, followed by an aggregated final response. Participants widely described this design as having affordances that strike the ideal balance between process transparency and mental workload.

Brief individual agent rationales hit the sweet spot of process transparency and supported users' sensemaking. A majority of participants described this format as offering the right amount of process visibility without overwhelming them, which made the multi-agent system appear more helpful and reliable: *"[V3] is just enough information to make a quick decision in real life and be a little bit more sure, I found it reliable"* (*P1*).

Viewing diverse agent rationales gave users decision control, fostering trust. Several participants appreciated that V3's response format surfaced multiple agent rationales that let them weigh the evidence themselves: *"It's not telling me, oh, here's the right answer, I can pick for myself which I want to follow"* (*P1*). Others described using the agent explanations as scaffolding to build their own reasoning, rather than relying blindly on an AI recommendation: *"Reading each explanation is useful [...] to build my own reasoning"* (*P10*). This retained participants' agency as the final decision-maker: *"Although I'm asking the chatbot my question, I'm not handing [it] the agency and autonomy 100%"* (*P3*).

We report how agent disagreement (D4) within V3 shaped user perceptions in §5.3.3, and while V3 was widely seen as striking the right balance, preferences varied by task context, a finding we present later when discussing task-related perceptions in §5.4.

5.3.3 ***D4 (& D2): Agreement, Disagreement, and Critique***.

Participants interpreted agent-level (dis)agreement and critique as key system credibility signals, sometimes in unexpected ways. In our stimuli, role specialisation (D2) was made salient only via a dedicated *critic agent* in interface V4, so we also discuss findings related to design dimension D2 in this subsection. We organise this subsection into three parts, presenting user perceptions of: (i) implicit disagreement (V2 and V3), (ii) explicit disagreement (via the critic in V4), and (iii) agent agreement observed across interfaces.

Implicit agent disagreement reduced system reliability, especially when no rationale was provided. In **V2 (Agents Answer)**, a majority of participants expressed diminished trust when one of the four agents visibly disagreed but offered no explanation. The mere presence of a dissenting agent without rationale introduced uncertainty and eroded the system's reliability: *"Looking at disagreement in [V2] makes me uncertain, it makes me feel like I need even more information to trust the system"* (*P5*). A few participants expressed that they would prefer not to see such disagreements at all: *"I don't want to see that [dissenting agent], I understand that AI is not always right but it's just giving me so much doubt"* (*P2*).

In contrast, in **V3 (Agents Explain)**, agent disagreement was perceived as less damaging to trust because each agent provided a rationale. This allowed users to interrogate the disagreement and determine its relevance: *"It now lets me see the reasoning of the agent that gives the different answer. And [...] lets me judge which one is believable"* (*P4*).

An explicit critic agent boosted perceived trustworthiness by showing that the system is checking itself; critiques also provided scaffolding for participants' own reasoning and sensemaking. Participants widely appreciated the critic agent in **V4 (Critique)** because it made the system appear more credible. The presence of an explicit check gave users confidence that the system was not blindly generating answers: *"They are analysing each other which makes it very trustworthy"* (*P6*); *"The Critic Agent improves my trust [...] this system is not deceptive"* (*P7*). Participants also described the critic agent as helping them trust the AI more because it acknowledged AI fallibility: *"It makes me less doubtful [...] because it's kind of like self-aware about AI limitations"* (*P2*).

The critic agent also surfaced "raw materials" for participants' own deliberation (*P11*), assisting sensemaking: *"I saw some arguments and different perspectives which makes me understand better"* (*P8*). Further, the critic supported users' agency by highlighting weaknesses of each reasoning path: *"I feel like I have control of the decision here, which I find very helpful"* (*P6*).

Highlighting system limitations through an explicit critique backfired in some cases. While many found the critic helpful, some also saw it as cognitively burdensome: *"Lot of information to digest. More burden for me to analyse both the agent responses and the critique"* (*P10*). For one participant, the critic exposing system limitations backfired and depleted trust: *"Looking at the critic agent makes me less confident in the system's capacity to answer my question"* (*P3*). A few participants suggested the critic agent should be employed behind the scenes: *"This would be more trustworthy if it happened in the background [...] showing it to me makes too much cognitive noise"* (*P5*).

When agents agreed and offered consistent reasoning, participants perceived the system as more reliable. Across conditions that displayed individual agent rationales (e.g., **V3 and V4**), participants not only looked at whether agents agreed, but also evaluated the consistency of their reasoning. When agreeing agents explained their reasoning in similar ways, this boosted perceived system reliability: *"The majority of agents are giving the same answer in [V3] and they explain themselves in similar ways, so I trust this answer even more"* (*P8*).

5.3.4 *D5 & D6: Interaction Paradigm and Information Flow.*

Although interaction paradigm (D5) and information-flow architecture (D6) are core design dimensions of multi-agent systems, they did not emerge as salient cues shaping participants' transparency or trust judgements. Only one participant expressed wanting to know more about the information flow, prompted by something visible (e.g., disagreement) rather than by the structures themselves: *"I don't know how much weight each agent carries [...] do they have the same weight and influence, is there a leader? [...] It's still very blurry for me how they collaborate with each other"* (*P5*). Overall, participants' assessments were driven by *visible* cues (e.g., the presence of rationales, whether agents agreed, whether a critic was present), while interaction topology and information flow remained perceptually latent. Thus, in our study, **D5/D6 did not function as user-facing trustworthiness signals.**

5.3.5 *D7: Aggregation Mechanism: Majority Vote and Consensus.*

Participants' perceptions of the multi-agent system were shaped not just by the individual agent outputs, but also by how those outputs were consolidated, whether through a majority vote / agreement, or by the agents reaching consensus.

Visible majority agreement among agents enhanced perceived system reliability, even without rationales. Across design variants that surfaced multiple agents' responses (e.g., **V2, V3, V4**), participants frequently interpreted majority agreement as a signal of trustworthiness and correctness: *"If the majority of agents agrees on something, I trust it"* (*P10*). This perception was so strong that some participants frequently disregarded the minority/dissenting agent altogether: *"I think the [dissenting agent] could also be right, but it doesn't have enough friends. So I don't think it's trustworthy"* Some even reflected critically on this tendency, acknowledging that majority agreement *could* be misleading: *"Wow, the more I think about this, the majority could kind of fool me or convince me [...]"* (*P2*).

Seeing agents reach consensus through debate boosted trust, and delivered rationales in a naturalistic, conversational format. The multi-turn debate ending in consensus **(V5)** was seen by many as a valuable system process that mirrored human-like deliberation and enhanced the system's reliability: *"I think it's really important to see that the agents actually converge at some point, because if [not], then I'd be skeptical about the answer"* (*P10*), and helpfulness: *"I am able to read through and understand where I was wrong, because I had a similar view as Agent B but it got proven wrong. Seeing that made me adjust my own thoughts"* (*P12*). The conversational structure also made the reasoning feel more digestible: *"This conversation feels natural [...] as if people are talking about this"* (*P10*).

Despite its transparency, the debate mechanism was sometimes perceived to contain "too much" information. While **V5**'s debate process was seen as trustworthy and thorough, some participants felt it introduced unnecessary complexity which was *"not worth the cost,"* especially for simpler tasks: *"I find it very reliable, but I think that's a lot [...] to read"* (*P1*).

5.4 RQ3: Transparency Needs by Task Demands

5.4.1 *Participants evaluated the costs and benefits of process transparency based on task complexity and fit.*

Participants calibrated their transparency preferences to the complexity and perceived cognitive demands of the task. For the simpler, factual info-seeking task, concise outputs with minimal process visibility (e.g., **V1 or V2**) were often preferred: *"I consider [V1 (Final Answer)] the most helpful in this context, that is sufficient for me to trust this AI"* (*P4*). In these cases, more elaborate forms of process visibility (e.g., **V5**) was seen as unnecessary: *"Now I feel like [the debate] is just too much because it feels like a fairly simple question"* (*P1*). Conversely, for more complex reasoning tasks, participants perceived process

visibility as valuable and worthwhile: *"[V5 (Agents Debate)] is necessary, simply because the LSAT task requires more mental demand. Having the longer explanations and a discussion actually helps with my overall understanding"* (*P10*). In these settings, interfaces like **V4 and V5** were seen as helpful rather than burdensome. We also observed that **when the system surfaced more reasoning than users felt the task warranted, it sometimes backfired and undermined their trust.** For example, in simple info-seeking tasks, overly elaborate outputs reduced perceievd system credibility: *"[V5] is trying too hard to convince me of the answer, it's throwing a lot at me so I don't find it reliable at all"* (*P1*).

In addition to task complexity, participants' transparency preferences were also shaped by the **perceived stakes or consequences** of the task. In some cases, stakes overrode complexity: *"If I was answering this [info-seeking task] on Who Wants to Be a Millionaire, I would very much take on the extra costs [of engaging with more elaborate forms of process visibility]. I'll need more insight into how the agents made this decision"* (*P4*).

5.4.2 ***Participants' task expertise and disposition to trust automation influenced how much process visibility is desirable.*** Participants' own task domain expertise shaped how much and which form of process insight they sought. When confident, they typically wanted minimal transparency: *"If I'm querying the AI for something that I have expertise in, then I would like only a brief output"* (*P9*). Conversely, low task confidence increased their need for transparency: *"When I'm not confident at all, even looking at all four [agent's reasoning paths] may not be enough to trust it"* (*P4*).

Furthermore, participants' dispositional trust and past AI experience also shaped preferences. One participant described always wanting more process insight due to low dispositional trust: *"I haven't fully trusted AI 100% yet, so I would want as much information as I can get every time, to understand why the AI made that decision"* (*P4*). Two others expressed that they would calibrate their information sufficiency needs from a multi-agent system over time: *"If the AI is reliable enough for a while, then [...] I may not need as much [transparency]"* (*P7*).

5.4.3 ***Participants desired interfaces that let them flexibly adjust 'how much transparency' based on the task.*** Eight participants expressed a desire for collapsible or toggleable transparency interface elements that would allow them to access agent rationales only when needed — such as when verifying or better understanding the system's reasoning: *"Have the consensus answer up at the top [...] and if I have doubts or want to check the conversation, I can expand it if needed"* (*P5*). This was observed especially for the simpler info-seeking task: *"Whenever I use thinking [AI] models, I always like to have the thinking collapsed. [That] would be useful here as well, especially for the [info-seeking] task, if details could be collapsed but still available if I want to see what each agent came up with and why"* (*P10*).

6 Discussion

In this study, we took a discovery-oriented, design-led approach to understanding how users interpret transparency and trustworthiness in multi-agent systems. This discussion synthesises our findings to offer conceptual insights and implications for designing trustworthy, context-sensitive multi-agent AI systems.

6.1 Reconceptualising Transparency: A Sufficiency Judgement, Not a Volume Dial

Our results join a growing body of work [35, 43, 45, 56, 57] that challenges the assumption that *more transparency* or insight into system reasoning is *necessarily* better. Across tasks and interfaces, our participants did not seek complete visibility into multi-agent processes, nor did they treat transparency as a *"more-is-better"* dial. Instead, they evaluated it through a lens of *contextual sufficiency* — whether the system offered *enough* insight to support their decision-making in that moment. **Participants gravitated towards what we term a *"Goldilocks"* zone of transparency, a level of process visibility that aptly balanced the mental workload the multi-agent interface exerted against the informational value it offered.** Importantly, this *"just right"* threshold was dynamic and context-sensitive, influenced by task, AI, and participants' dispositional characteristics, with the perceived trustworthiness of the multi-agent system dropping sharply on either end of this threshold. Notably, these judgements were not just about *quantity*, but *fit*. Users sought enough process visibility to enable trust, sensemaking, and independent judgement, but no more than that.

Task type was a strong determinant of the *"just right"* transparency threshold. Participants doing the more ambiguous reasoning task preferred interfaces such as V4 (Agent Critique) and V5 (Debate), which supported their ability to evaluate diverse reasoning paths. This form of transparency provided users with "raw material" for their own sensemaking, which boosted their trust. In such tasks, receiving a final answer without any explanation, such as V1 (Final Answer) and V2 (Agents Answer), was seen as insufficient and unhelpful for sensemaking. At the same time, however, these interfaces became satisfactory and trustworthy in simpler tasks, where more process insight or visibility was instead seen as unnecessary, backfiring and eroding users' trust. This contrast is especially striking given that all five of our interfaces presented the same final (correct) answer, suggesting *how the amount and type of process transparency (and not accuracy alone) influenced perceived trustworthiness*. We posit that transparency in multi-agent systems should not be seen as a binary (transparent vs. opaque) construct, but rather as a spectrum where both extremes can undermine trust.

Individual and dispositional factors also shaped where the *"just right"* threshold for transparency lay. Participants with higher task domain expertise or confidence typically preferred minimal transparency, viewing additional system insight as unnecessary, while those with lower confidence expressed a need for more visible reasoning to assess trustworthiness. We also found dispositional trust impacting these preferences: participants who expressed a lower baseline trust in AI sought more detailed transparency to be able verify system outputs and trust them.

6.1.1 ***Progressive, On-Demand Transparency Balances Process Visibility and Cost***. Eight of our twelve participants expressed a clear preference for flexible, progressive transparency, where more detail could be revealed on demand. Participants welcomed the reliability benefits of multi-agent systems, without the

cognitive cost of reading every agent's reasoning by default. Instead, they preferred a simplified starting point, such as the consensus answer, with the ability to expand individual rationales, critiques, or debates as needed. This design preference echoes longstanding UI principles of progressive disclosure [14, 62], which is being increasingly advocated for transparency in intelligent systems [43, 85]. Progressive disclosure is also supported by social science views of *explanation* as an *occasioned* activity, provided *only* when warranted by the context [34]. Our findings extend this to human-AI interaction, and we posit that 'optimal' transparency utility may be better achieved through more contextually-aware, layered disclosures rather than static, exhaustive explanation/information dumps.

Our findings advocate for a new framing: multi-agent transparency as *contextual sufficiency*, negotiated between the system and the user. Rather than aiming for maximal exposure of agent reasoning, designers should prioritise delivering *just enough* transparency that best suits the task, dispositional, and contextual needs, to support trust, sensemaking, and decision ownership without overwhelming users. This re-framing also resonates with debates around *intelligibility* versus *fidelity* in AI explainability research [47, 99]: users often prefer explanations that are simplified and cognitively manageable (intelligibility), even if they abstract away from the system's full internal logic (fidelity). In our study, participants did not seek complete traceability of how agent reasoning was generated, but valued *enough* insight for the task at hand.

6.1.2 *Implications for Trust Calibration and Future Work.*

These findings also offer a novel lens for theorising trust calibration in multi-agent LLM systems. While much prior work conceptualises trust calibration as stemming from a match between perceived and actual AI performance [5, 49], our participants' trust often hinged on the *perceived fit* between the system's transparency and their current informational needs.

This raises concerns: transparency that feels "just right" may increase users' *subjective* trust, but not necessarily *warranted* trust. That is, sufficiency-based transparency may boost users' trust in the system because it feels aligned, regardless of the AI's actual reliability. This is especially critical in scenarios where users have low domain expertise: *ideal* transparency may lend the system an illusion of legitimacy, as is commonly observed with explanations in XAI systems [23, 67].

Our study provides theoretical grounding for several design *"levers"* that may modulate trust, including process visibility, perceived sufficiency, and flexible transparency structures. Future work should empirically test how these levers interact with AI accuracy, user expertise, and task complexity to influence trust calibration. *Do users better detect AI errors when given more process transparency? Does progressive disclosure support critical assessment of AI output or foster indiscriminate trust? Does combing through more process transparency serve as a cognitive forcing function [13] which reduces overreliance, or do users learn to ignore the bulk of process insight and fall back on heuristic trust?* These are some key questions for advancing human-centred design of multi-agent LLM systems, which our work raises.

6.2 Interpreting Multi-Agent Transparency: Cues, Heuristics, and How To De-Risk Them

Multi-agent reasoning is typically introduced as a back-end mechanism intended to improve raw AI performance, through mechanisms such as self-consistency (sampling multiple reasoning paths from a single model), voting (aggregating responses from multiple independent agents), or debate (inter-agent deliberations to arrive at a consensus answer) [53, 58, 95]. While these techniques can boost AI accuracy, our findings suggest they also produce epistemic signals, such as disagreement, critique, and consensus, that end-users actively interpret as cues for when and how much to trust the multi-agent system.

A key goal of our study was to build theoretical insight into the cues users attend to and how these shape their perceptions of transparency and trustworthiness. Across tasks, we found that visible surface-level interface affordances, such as the number of agents, presence of rationales, disagreement, and consensus, strongly shaped user judgements. In contrast, more hidden multi-agent architecture elements, such as interaction paradigm (whether agents reasoned sequentially or in parallel, design dimension D5) and information flow (whether information flowed between agents in a pipeline or hierarchical form, design dimension D6), remained perceptually latent and did not impact participants' trustworthiness judgements. In this section, we unpack the key epistemic cues users relied on, highlight the design tensions and trade-offs that emerged, and present implications for design.

6.2.1 *Ensemble Size Serves as a Reliability Signal, But Presents Trade-Offs Between Cognitive Cost and Visibility.*

Our participants frequently interpreted the number of agents (ensemble size) as a heuristic for system capability, perceiving larger ensembles as having *"more people working towards one goal"* and hence, more reliable. This reflects an instinctive behaviour to find the *"wisdom of the silicon crowd"* more reliable, [78], and aligns with empirical findings showing multi-agent reasoning often outperforming single-agent approaches [22, 53, 55, 82]. However, this heuristic also comes with a cognitive trade-off: more agents can mean more reasoning paths and responses to process. Crucially, participants wanted adaptive ensemble sizes tailored to task complexity, requiring more agents for more complex tasks; and while they did not always want to read every agent's output, knowing that multiple agents existed *"behind the scenes"* made the system response feel more deliberated and, by extension, more trustworthy. This highlights a classic *cost-benefit tension in AI transparency*: users want to know that the system is deliberating, but do not necessarily want to comb through all that deliberation. This also highlights the risk of users relying on heuristics that may not hold, as larger ensemble sizes may not always imply greater reliability. To de-risk this, interfaces could surface only distinct lines of agent reasoning instead of raw agent counts, or present a few representative rationales while indicating that a larger ensemble operated behind the scenes. This aligns with our participants' desire to know that multiple agents deliberated, without being overwhelmed by every output. We call on future work to investigate how best to convey this information, preserving the perceived benefits of multi-agent reasoning while reducing over-reliance on ensemble size as a proxy for accuracy.

6.2.2 ***Agent-Level Rationales Support Sensemaking and Enhance User Agency, But Majority Heuristics Persist.*** Agent-level rationales played a key role in helping users make sense of multi-agent AI outputs. Interfaces that offered agent responses without rationales (e.g., V2) provided surface-level transparency, but left users unable to engage with or evaluate the AI's reasoning. In contrast, when agents offered brief explanations (e.g., V3), users were better able to weigh competing reasoning paths, scaffold their own decision-making, and retain a sense of agency over the final decision. This aligns with sensemaking frameworks that emphasise the importance of scaffolding rather than supplanting human reasoning [44]. Agent-level rationales invited users to deliberate while keeping their mental workload under check, rather than leading to indiscriminate reliance. Our participants used these rationales to construct, test, and revise their own reasoning.

However, we also observed that in the absence of rationales, majority agreement between the agents significantly boosted perceived system reliability and often led users to dismiss dissenting agents entirely, suggesting how users strongly rely on agent consensus as a heuristic for reliability. This behaviour reflects well-known social influence heuristics, such as the "bandwagon effect" where individuals adopt majority views due to perceived social consensus [37], and aligns with prior findings showing how users tend to conform to AI advice in objective tasks [26, 74]. However, the majority decision may not always be accurate. Thus, to mitigate this risk and reduce over-reliance on majority-based cues, interfaces could clearly highlight reasoning diversity, for example, by presenting the strongest dissenting/'minority' rationale, nudging users to inspect 'both sides' of the evidence rather than default to the majority.

We see these as promising directions for future work: to empirically examine whether multi-agent rationales can reduce over-reliance and lead to better decision outcomes, whether such rationales help users detect errors or merely increase confidence regardless of accuracy, and how the majority heuristics we observed can be leveraged to promote trust calibration. These findings also echo recent calls to move beyond AI systems that merely make a decision and explain it [61], and towards decision-support systems that scaffold human reasoning in contextually-sensitive ways.

6.2.3 ***Explained Disagreement and Critique Convey That The System is "Checking Itself" — Enhancing Trust, When Presented Correctly.*** Visible agent disagreement was a double-edged cue: it could either build or erode trust, depending on how it was surfaced and explained. When an agent disagreed without explanation, participants saw it as a sign of internal AI inconsistency or failure, reducing their trust. Several expressed a preference not to see such unexplained disagreement *at all*. However, when the same disagreement was accompanied with agent-level rationales, i.e., agreeing and disagreeing agents explaining their reasoning, participants found it helpful and trustworthy. Disagreement helped explore competing perspectives, evaluate uncertainty, and refine users' own reasoning. These perceptions were even more pronounced when a dedicated critic agent critiqued all its peers: the multi-agent system appeared trustworthy and honest because it was *"checking itself."* We observed more positives of explicit critiques: they helped users refine their own reasoning by surfacing both sides of the argument, and only a small minority experienced a decrease in perceived system capability after the critic highlighted limitations.

Our findings align with long-standing research in collaborative and group decision-making, which emphasises the epistemic value of surfacing diverse opinions and disagreements [18, 79, 80]. These findings also resonate with Reingold et al. [72], who observed that while dissenting explanations reduced users' trust in AI, they also reduced overreliance — offering a potential path to more calibrated trust.

Taken together, our results suggest that disagreement and critique are powerful epistemic signals, but only when well-explained and well-structured. When appropriately surfaced, they can foster healthy AI scepticism, highlight uncertainty, and act as cognitive forcing functions that could reduce indiscriminate reliance on AI. To avoid unintended trust erosion from observing disagreeing agents, these signals should be accompanied by brief explanations that help users understand the disagreement and interpret it as a deliberate feature of the system's reasoning process rather than as a limitation or sign of failure. Future work should empirically examine whether (and for whom) disagreement fosters calibrated trust. For instance, users with low task domain expertise may benefit from disagreement as a cautionary signal when they lack the skills to detect AI errors. Designing systems that expose such disagreement without overwhelming users remains an open and important challenge for future work.

6.3 Multi-Agent Systems Surface Epistemic Cues That Are Double-Edged and Must Be Calibrated

Across all the interface affordances we examined (ensemble size, rationales, disagreement, consensus, and aggregation) a consistent pattern emerged: *no cue functioned as a universally positive signal of trustworthiness.* Each offered distinct benefits in terms of establishing trust, but also introduced design trade-offs and risks. Their impact depends on how they are surfaced and presented, when they are shown, and how well they align with user expectations, task demands, and cognitive bandwidth. What boosts trust in one context could erode it in another; what supports sensemaking for one user could overwhelm or mislead another.

This pattern reveals a central insight: **epistemic signals in multi-agent interfaces are inherently double-edged**. Their value lies not in their presence alone, but in how they are presented, situated, and attuned to context. Designing for transparency in multi-agent interfaces, then, is not about maximising visibility, but enabling *contextually-sufficient transparency* — a level of process insight that aligns with users' information needs, and supports trust and interpretability without inducing too much mental workload or misplaced/unwarranted trust. We reinforce the central argument of this paper: transparency in multi-agent LLM systems is not a binary property to be maximised, but a dynamic, context-sensitive sufficiency judgement. Designing for calibrated trust in such systems requires a deep understanding of which interface cues users attend to and how they can inadvertently introduce risks of overtrust.

Our work contributes a foundational mapping of these epistemic cues, highlights their associated trade-offs and design tensions, and offers recommendations to help de-risk them. Future work must

build on this foundation, investigating how to surface these signals in ways that are not only informative, but *appropriately aligned* with user needs and task demands to help foster calibrated trust in multi-agent AI systems.

6.4 Limitations

Our study involved predefined, bounded tasks with objectively verifiable answers, enabling control and comparability across interfaces. However, letting users interact with multi-agent systems in self-directed tasks may surface different needs and behaviours. Further, we explored key multi-agent affordances through specific operationalisations (e.g., a fixed 3:1 majority in agreement) and did not vary finer-grained configurations such as the degree of consensus, or tested the effects of linearly increasing ensemble size (e.g., sizes other than 4 or 8). This was a necessary design choice for two reasons: first, our goal was to evaluate representative interface designs that foreground different forms of multi-agent transparency embedded with distinct epistemic cues, rather than exhaustively test all combinations. Second, granularly varying ensemble size or disagreement proportions across all variants would lead to an impractically large number of interfaces and compromise our ability to use each interface variant as a focused probe. Thus, while our interface variants were designed to capture a diverse range of multi-agent configurations rooted in prior literature, they represent only a small subset of the broader design space. Future work can build on our conceptual groundwork to examine how different levels of disagreement, consensus, or agent count influence trust perceptions, as well as explore more granular combinations across the design dimensions.

Moreover, as with all qualitative, design-led, Comparative Structured Observation studies, our goal was to understand user behaviours, rather than to quantify causal effects. Nevertheless, this method yielded rich insights into how users make sense of multi-agent transparency, which future work can test empirically at scale. Further, while our sample size (N=12) is consistent with guidance on thematic saturation in qualitative research [11, 27] and comparable to prior CSO studies in HCI [21, 88], it may limit the breadth of perspectives captured. Lastly, all final AI outputs in our interfaces were accurate. While this allowed us to isolate interface effects, it leaves open how users interpret and calibrate trust when agent errors are present and/or unevenly distributed within the ensemble.

7 Conclusion

Multi-agent LLM interfaces surface a rich variety of epistemic signals — such as agent-level responses, disagreement, critique, and consensus — which users actively interpret as cues for when and how much to trust the system. Through a discovery-oriented, design-led study, we unpacked how these signals are made sense of in context, and what tensions emerge in the process. Rather than treating transparency as a static property to be maximised, our findings call for a more context-sensitive approach, one that treats transparency as a *sufficiency judgement* shaped by task demands, user preferences, and mental workload We highlight how interface-level affordances can both boost and erode trust, and argue for calibrated over maximum exposure. By understanding users' interpretive strategies, heuristics, and transparency preferences across varied multi-agent interfaces, our work lays the groundwork for more context-aware, human-centred approaches to designing trustworthy multi-agent AI. We invite future research to build on these insights and test how transparency can be adaptively structured to better support sensemaking, trust calibration, and decision quality.

References

[1] 2007. *The Official LSAT SuperPrep.* Law School Admission Council, Newtown, PA.

[2] 2025. LSAT Logical Reasoning Practice Tests. https://www.cracklsat.net/lsat/logical-reasoning/

[3] Ariful Islam Anik and Andrea Bunt. 2021. Data-Centric Explanations: Explaining Training Data of Machine Learning Systems to Promote Transparency. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3411764.3445736

[4] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S. Lasecki, Daniel S. Weld, and Eric Horvitz. 2019. Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 7 (Oct. 2019), 2–11. doi:10.1609/hcomp.v7i1.5285

[5] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–16. doi:10.1145/3411764.3445717

[6] Luciano Baresi, Matteo Camilli, Tommaso Dolci, and Giovanni Quattrocchi. 2024. A Conceptual Framework for Quality Assurance of LLM-based Socio-critical Systems. In *Proceedings of the 39th IEEE/ACM International Conference on Automated Software Engineering (ASE '24)*. Association for Computing Machinery, New York, NY, USA, 2314–2318. doi:10.1145/3691620.3695306

[7] Umang Bhatt, Javier Antorán, Yunfeng Zhang, Q. Vera Liao, Prasanna Sattigeri, Riccardo Fogliato, Gabrielle Melançon, Ranganath Krishnan, Jason Stanley, Omesh Tickoo, Lama Nachman, Rumi Chunara, Madhulika Srikumar, Adrian Weller, and Alice Xiang. 2021. Uncertainty as a Form of Transparency: Measuring, Communicating, and Using Uncertainty. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, Virtual Event USA, 401–413. doi:10.1145/3461702.3462571

[8] Jessica Y. Bo, Sophia Wan, and Ashton Anderson. 2025. To Rely or Not to Rely? Evaluating Interventions for Appropriate Reliance on Large Language Models. doi:10.48550/arXiv.2412.15584 arXiv:2412.15584 [cs].

[9] Edyta Bogucka, Marios Constantinides, Sanja Šćepanović, and Daniele Quercia. 2024. Co-designing an AI Impact Assessment Report Template with AI Practitioners and AI Compliance Experts. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 7, 1 (Oct. 2024), 168–180. doi:10.1609/aies.v7i1.31627 Number: 1.

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

[11] Virginia Braun and Victoria Clarke. 2013. Successful Qualitative Research : A Practical Guide for Beginners. (2013), 1–400. https://www.torrossa.com/en/resources/an/5017629 Publisher: SAGE Publications Ltd.

[12] Virginia Braun and Victoria Clarke. 2019. Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health* 11, 4 (Aug. 2019), 589–597. doi:10.1080/2159676X.2019.1628806 Publisher: Routledge _eprint: https://doi.org/10.1080/2159676X.2019.1628806.

[13] 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 (April 2021), 188:1–188:21. doi:10.1145/3449287

[14] John M. Carroll and Caroline Carrithers. 1984. Training wheels in a user interface. *Commun. ACM* 27, 8 (Aug. 1984), 800–806. doi:10.1145/358198.358218

[15] Alan Chan, Rebecca Salganik, Alva Markelius, Chris Pang, Nitarshan Rajkumar, Dmitrii Krasheninnikov, Lauro Langosco, Zhonghao He, Yawen Duan, Micah Carroll, Michelle Lin, Alex Mayhew, Katherine Collins, Maryam Molamohammadi, John Burden, Wanru Zhao, Shalaleh Rismani, Konstantinos Voudouris, Umang Bhatt, Adrian Weller, David Krueger, and Tegan Maharaj. 2023. Harms from Increasingly Agentic Algorithmic Systems. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 651–666. doi:10.1145/3593013.3594033

[16] Siew H. Chan, Qian Song, and Lee J. Yao. 2015. The moderating roles of subjective (perceived) and objective task complexity in system use and performance. *Computers in Human Behavior* 51 (Oct. 2015), 393–402. doi:10.1016/j.chb.2015.04.059

[17] Wei-Hao Chen, Weixi Tong, Amanda Case, and Tianyi Zhang. 2025. Dango: A Mixed-Initiative Data Wrangling System using Large Language Model. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, 1–28. doi:10.1145/3706598.3714135

[18] 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 (IUI '24)*. Association for Computing Machinery, New York, NY, USA, 103–119. doi:10.1145/3640543.3645199

[19] Wesley Hanwen Deng, Solon Barocas, and Jennifer Wortman Vaughan. 2025. Supporting Industry Computing Researchers in Assessing, Articulating, and Addressing the Potential Negative Societal Impact of Their Work. *Proc. ACM Hum.-Comput. Interact.* 9, 2 (May 2025), CSCW178:1–CSCW178:37. doi:10.1145/3711076

[20] Ali Dorri, Salil S. Kanhere, and Raja Jurdak. 2018. Multi-Agent Systems: A survey. 6 (2018), 28573–28593. doi:10.1109/ACCESS.2018.2831228

[21] Steven Dow, Manish Mehta, Ellie Harmon, Blair MacIntyre, and Michael Mateas. 2007. Presence and engagement in an interactive drama. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '07)*. Association for Computing Machinery, New York, NY, USA, 1475–1484. doi:10.1145/1240624.1240847

[22] Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2024. Improving Factuality and Reasoning in Language Models through Multiagent Debate. https://openreview.net/forum?id=zj7YuTE4t8

[23] Upol Ehsan and Mark O. Riedl. 2021. Explainability Pitfalls: Beyond Dark Patterns in Explainable AI. doi:10.48550/arXiv.2109.12480

[24] Sorouralsadat Fatemi and Yuheng Hu. 2024. Enhancing Financial Question Answering with a Multi-Agent Reflection Framework. In *Proceedings of the 5th ACM International Conference on AI in Finance (ICAIF '24)*. Association for Computing Machinery, New York, NY, USA, 530–537. doi:10.1145/3677052.3698686

[25] Sally Fincher and Josh Tenenberg. 2005. Making sense of card sorting data. *Expert Systems* 22, 3 (2005), 89–93. doi:10.1111/j.1468-0394.2005.00299.x

[26] Christopher Flathmann, Wen Duan, Nathan J. Mcneese, Allyson Hauptman, and Rui Zhang. 2024. Empirically Understanding the Potential Impacts and Process of Social Influence in Human-AI Teams. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1 (April 2024), 49:1–49:32. doi:10.1145/3637326

[27] A Fugard and H Potts. 2014. Sample size determination for thematic analysis and related qualitative methodologies: a quantitative model. *6th ESRC: Research Methods Festival. Oxford: St Catherine's College* (2014).

[28] Bill Gaver, Tony Dunne, and Elena Pacenti. 1999. Design: cultural probes. *interactions* 6, 1 (1999), 21–29. Publisher: ACM New York, NY, USA.

[29] Katy Ilonka Gero, Zahra Ashktorab, Casey Dugan, Qian Pan, James Johnson, Werner Geyer, Maria Ruiz, Sarah Miller, David R. Millen, Murray Campbell, Sadhana Kumaravel, and Wei Zhang. 2020. Mental Models of AI Agents in a Cooperative Game Setting. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/3313831.3376316

[30] Jarod Govers, Saumya Pareek, Eduardo Velloso, and Jorge Goncalves. 2025. Feeds of Distrust: Investigating How AI-Powered News Chatbots Shape User Trust and Perceptions. *ACM Trans. Interact. Intell. Syst.* (March 2025). doi:10.1145/3722227 Just Accepted.

[31] Jarod Govers, Eduardo Velloso, Vassilis Kostakos, and Jorge Goncalves. 2024. AI-Driven Mediation Strategies for Audience Depolarisation in Online Debates. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–18. doi:10.1145/3613904.3642322

[32] Gaole He, Stefan Buijsman, and Ujwal Gadiraju. 2023. How Stated Accuracy of an AI System and Analogies to Explain Accuracy Affect Human Reliance on the System. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2 (Oct. 2023), 276:1–276:29. doi:10.1145/3610067

[33] Gaole He, Lucie Kuiper, and Ujwal Gadiraju. 2023. Knowing About Knowing: An Illusion of Human Competence Can Hinder Appropriate Reliance on AI Systems. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–18. doi:10.1145/3544548.3581025

[34] Denis J. Hilton. 1990. Conversational processes and causal explanation. *Psychological Bulletin* 107, 1 (1990), 65–81. doi:10.1037/0033-2909.107.1.65 Place: US Publisher: American Psychological Association.

[35] Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. 2018. Metrics for Explainable AI: Challenges and Prospects. https://arxiv.org/abs/1812.04608v2

[36] C.I. Hovland, I.L. Janis, and H.H. Kelley. 1953. *Communication and persuasion*. Yale University Press, New Haven, CT, US.

[37] Jonathan Howard. 2019. Bandwagon Effect and Authority Bias. In *Cognitive Errors and Diagnostic Mistakes: A Case-Based Guide to Critical Thinking in Medicine*, Jonathan Howard (Ed.). Springer International Publishing, Cham, 21–56. doi:10.1007/978-3-319-93224-8_3

[38] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.* 55, 12 (March 2023), 248:1–248:38. doi:10.1145/3571730

[39] PN Johnson-Laird. 1983. *Mental models: Towards a cognitive science of language, inference, and consciousness*. Harvard University Press.

[40] Sunnie S. Y. Kim. 2024. Establishing Appropriate Trust in AI through Transparency and Explainability. In *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems (CHI EA '24)*. Association for Computing Machinery, New York, NY, USA, 1–6. doi:10.1145/3613905.3638184

[41] Sunnie S. Y. Kim, Q. Vera Liao, Mihaela Vorvoreanu, Stephanie Ballard, and Jennifer Wortman Vaughan. 2024. "I'm Not Sure, But...": Examining the Impact of Large Language Models' Uncertainty Expression on User Reliance and Trust. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 822–835. doi:10.1145/3630106.3658941

[42] Sunnie S. Y. Kim, Jennifer Wortman Vaughan, Q. Vera Liao, Tania Lombrozo, and Olga Russakovsky. 2025. Fostering Appropriate Reliance on Large Language Models: The Role of Explanations, Sources, and Inconsistencies. doi:10.1145/3706598.3714020 arXiv:2502.08554 [cs].

[43] René F. Kizilcec. 2016. How Much Information? Effects of Transparency on Trust in an Algorithmic Interface. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. Association for Computing Machinery, New York, NY, USA, 2390–2395. doi:10.1145/2858036.2858402

[44] Gary Klein, Jennifer K. Phillips, Erica L. Rall, and Deborah A. Peluso. 2007. A Data–Frame Theory of Sensemaking. In *Expertise Out of Context*. Psychology Press.

[45] Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. 2013. Too much, too little, or just right? Ways explanations impact end users' mental models. In *2013 IEEE Symposium on Visual Languages and Human Centric Computing*. 3–10. doi:10.1109/VLHCC.2013.6645235

[46] Moritz Körber. 2018. Theoretical considerations and development of a questionnaire to measure trust in automation.

[47] Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Jure Leskovec. 2017. Interpretable & Explorable Approximations of Black Box Models. doi:10.48550/arXiv.1707.01154 arXiv:1707.01154 [cs].

[48] John Lee and Neville Moray. 1992. Trust, control strategies and allocation of function in human-machine systems. *Ergonomics* 35, 10 (1992), 1243–1270. doi:10.1080/00140139208967392 Place: United Kingdom Publisher: Taylor & Francis.

[49] John D. Lee and Katrina A. See. 2004. Trust in Automation: Designing for Appropriate Reliance. *Human Factors* 46, 1 (March 2004), 50–80. doi:10.1518/hfes.46.1.50_30392

[50] Soohwan Lee, Mingyu Kim, Seoyeong Hwang, Dajung Kim, and Kyungho Lee. 2025. Amplifying Minority Voices: AI-Mediated Devil's Advocate System for Inclusive Group Decision-Making. In *Companion Proceedings of the 30th International Conference on Intelligent User Interfaces*. 17–21. doi:10.1145/3708557.3716334 arXiv:2502.06251 [cs].

[51] Yoonjoo Lee, Kihoon Son, Tae Soo Kim, Jisu Kim, John Joon Young Chung, Eytan Adar, and Juho Kim. 2024. One vs. Many: Comprehending Accurate Information from Multiple Erroneous and Inconsistent AI Generations. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 2518–2531. doi:10.1145/3630106.3662681

[52] Haoran Li, Xusen Cheng, and Xiaoping Zhang. 2025. Accurate Insights, Trustworthy Interactions: Designing a Collaborative AI-Human Multi-Agent System with Knowledge Graph for Diagnosis Prediction. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, 1–15. doi:10.1145/3706598.3713526

[53] Junyou Li, Qin Zhang, Yangbin Yu, Qiang Fu, and Deheng Ye. 2024. More Agents Is All You Need. doi:10.48550/arXiv.2402.05120 arXiv:2402.05120 [cs].

[54] Yuan Li, Lichao Sun, and Yixuan Zhang. 2025. MetaAgents: Large Language Model Based Agents for Decision-Making on Teaming. *Proc. ACM Hum.-Comput. Interact.* 9, 2 (May 2025), CSCW134:1–CSCW134:27. doi:10.1145/3711032

[55] Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. Encouraging Divergent Thinking in Large Language Models through Multi-Agent Debate. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 17889–17904. doi:10.18653/v1/2024.emnlp-main.992

[56] Q. Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–15. doi:10.1145/3313831.3376590

[57] Brian Y. Lim, Anind K. Dey, and Daniel Avrahami. 2009. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '09)*. Association for Computing Machinery, New York, NY, USA, 2119–2128. doi:10.1145/1518701.1519023

[58] Zijun Liu, Yanzhe Zhang, Peng Li, Yang Liu, and Diyi Yang. 2024. A Dynamic LLM-Powered Agent Network for Task-Oriented Agent Collaboration. doi:10.48550/arXiv.2310.02170 arXiv:2310.02170 [cs].

[59] Anna Luusua, Johanna Ylipulli, Marko Jurmu, Henrika Pihlajaniemi, Piia Markkanen, and Timo Ojala. 2015. Evaluation Probes. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. Association for Computing Machinery, New York, NY, USA, 85–94. doi:10.1145/2702123.2702466

[60] Wendy E. Mackay and Joanna McGrenere. 2025. Comparative Structured Observation. *ACM Trans. Comput.-Hum. Interact.* 32, 2 (April 2025), 14:1–14:27. doi:10.1145/3711838

[61] Tim Miller. 2023. Explainable AI is Dead, Long Live Explainable AI! Hypothesis-driven Decision Support using Evaluative AI. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 333–342. doi:10.1145/3593013.3594001

[62] Lloyd H. Nakatani and John A. Rohrlich. 1983. Soft machines: A philosophy of user-computer interface design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '83)*. Association for Computing Machinery, New York, NY, USA, 19–23. doi:10.1145/800045.801573

[63] Donald A. Norman. 1983. Some Observations on Mental Models. In *Mental Models*. Psychology Press. Num Pages: 8.

[64] Eugénio Oliveira, Klaus Fischer, and Olga Stepankova. 1999. Multi-agent systems: which research for which applications. *Robotics and Autonomous Systems* 27, 1 (April 1999), 91–106. doi:10.1016/S0921-8890(98)00085-2

[65] Raja Parasuraman and Christopher A. Miller. 2004. Trust and etiquette in high-criticality automated systems. *Commun. ACM* 47, 4 (April 2004), 51–55. doi:10.1145/975817.975844

[66] Saumya Pareek, Sarah Schömbs, Eduardo Velloso, and Jorge Goncalves. 2025. "It's Not the AI's Fault Because It Relies Purely on Data": How Causal Attributions of AI Decisions Shape Trust in AI Systems. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, 1–18. doi:10.1145/3706598.3713468

[67] Saumya Pareek, Niels van Berkel, Eduardo Velloso, and Jorge Goncalves. 2024. Effect of Explanation Conceptualisations on Trust in AI-assisted Credibility Assessment. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW2 (Nov. 2024), 383:1–383:31. doi:10.1145/3686922

[68] Saumya Pareek, Eduardo Velloso, and Jorge Goncalves. 2024. Trust Development and Repair in AI-Assisted Decision-Making during Complementary Expertise. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Rio de Janeiro Brazil, 546–561. doi:10.1145/3630106.3658924

[69] Pat Pataranutaporn, Ruby Liu, Ed Finn, and Pattie Maes. 2023. Influencing human–AI interaction by priming beliefs about AI can increase perceived trustworthiness, empathy and effectiveness. *Nature Machine Intelligence* 5, 10 (Oct. 2023), 1–11. doi:10.1038/s42256-023-00720-7

[70] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and Measuring Model Interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–52. doi:10.1145/3411764.3445315

[71] Pooja S. B. Rao, Sanja Šćepanović, Ke Zhou, Edyta Paulina Bogucka, and Daniele Quercia. 2025. RiskRAG: A Data-Driven Solution for Improved AI Model Risk Reporting. doi:10.48550/arXiv.2504.08952 arXiv:2504.08952 [cs].

[72] Omer Reingold, Judy Hanwen Shen, and Aditi Talati. 2024. Dissenting Explanations: Leveraging Disagreement to Reduce Model Overreliance. *Proceedings of the AAAI Conference on Artificial Intelligence* 38, 19 (March 2024), 21537–21544. doi:10.1609/aaai.v38i19.30151 Number: 19.

[73] John K. Rempel, John G. Holmes, and Mark P. Zanna. 1985. Trust in close relationships. *Journal of Personality and Social Psychology* 49, 1 (1985), 95–112. doi:10.1037/0022-3514.49.1.95 Place: US Publisher: American Psychological Association.

[74] Paolo Riva, Nicolas Aureli, and Federica Silvestrini. 2022. Social influences in the digital era: When do people conform more to a human being or an artificial intelligence? *Acta Psychologica* 229 (Sept. 2022), 103681. doi:10.1016/j.actpsy.2022.103681

[75] Malak Sadek, Marios Constantinides, Daniele Quercia, and Celine Mougenot. 2024. Guidelines for Integrating Value Sensitive Design in Responsible AI Toolkits. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–20. doi:10.1145/3613904.3642810

[76] Sara Salimzadeh, Gaole He, and Ujwal Gadiraju. 2024. Dealing with Uncertainty: Understanding the Impact of Prognostic Versus Diagnostic Tasks on Trust and Reliance in Human-AI Decision Making. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–17. doi:10.1145/3613904.3641905

[77] Elizabeth B.-N. Sanders and Pieter Jan Stappers. 2014. Probes, toolkits and prototypes: three approaches to making in codesigning. *CoDesign* 10, 1 (Jan. 2014), 5–14. doi:10.1080/15710882.2014.888183 Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/15710882.2014.888183.

[78] Philipp Schoenegger, Indre Tuminauskaite, Peter S. Park, Rafael Valdece Sousa Bastos, and Philip E. Tetlock. 2024. Wisdom of the silicon crowd: LLM ensemble prediction capabilities rival human crowd accuracy. *Science Advances* 10, 45 (Nov. 2024), eadp1528. doi:10.1126/sciadv.adp1528

[79] Daniel L Schwartz. 1998. The Productive Agency that Drives Collaborative Learning. (1998).

[80] David M. Schweiger, William R. Sandberg, and James W. Ragan. 1986. Group Approaches for Improving Strategic Decision Making: A Comparative Analysis of Dialectical Inquiry, Devil's Advocacy, and Consensus. *The Academy of Management Journal* 29, 1 (1986), 51–71. doi:10.2307/255859

[81] Sarah Schömbs, Saumya Pareek, Jorge Goncalves, and Wafa Johal. 2024. Robot-Assisted Decision-Making: Unveiling the Role of Uncertainty Visualisation and Embodiment. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–16. doi:10.1145/3613904.3642911

[82] Andries Smit, Nathan Grinsztajn, Paul Duckworth, Thomas D. Barrett, and Arnu Pretorius. 2024. Should we be going MAD? a look at multi-agent debate strategies for LLMs. In *Proceedings of the 41st International Conference on Machine Learning (ICML'24, Vol. 235)*. JMLR.org, Vienna, Austria, 45883–45905.

[83] Ian René Solano-Kamaiko, Dibyendu Mishra, Nicola Dell, and Aditya Vashistha. 2024. Explorable Explainable AI: Improving AI Understanding for Community Health Workers in India. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–21. doi:10.1145/3613904.3642733

[84] Elizabeth Solberg, Magnhild Kaarstad, Maren H. Rø Eitrheim, Rossella Bisio, Kine Reegård, and Marten Bloch. 2022. A Conceptual Model of Trust, Perceived Risk, and Reliance on AI Decision Aids. *Group & Organization Management* 47, 2 (April 2022), 187–222. doi:10.1177/10596011221081238

[85] Aaron Springer and Steve Whittaker. 2018. Progressive Disclosure: Designing for Effective Transparency. doi:10.48550/arXiv.1811.02164

[86] Ian Steenstra, Farnaz Nouraei, and Timothy Bickmore. 2025. Scaffolding Empathy: Training Counselors with Simulated Patients and Utterance-level Performance Visualizations. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, 1–22. doi:10.1145/3706598.3714014

[87] Chelse Swoopes, Tyler Holloway, and Elena L Glassman. [n. d.]. The Impact of Revealing Large Language Model Stochasticity on Trust, Reliability, and Anthropomorphization. ([n. d.]).

[88] Aurélien Tabard, Wendy Mackay, Nicolas Roussel, and Catherine Letondal. 2007. PageLinker: integrating contextual bookmarks within a browser. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '07)*. Association for Computing Machinery, New York, NY, USA, 337–346. doi:10.1145/1240624.1240680

[89] Tolgahan Toy. 2024. Transparency in AI. *AI & SOCIETY* 39, 6 (Dec. 2024), 2841–2851. doi:10.1007/s00146-023-01786-y

[90] Helena Vasconcelos, Gagan Bansal, Adam Fourney, Q. Vera Liao, and Jennifer Wortman Vaughan. 2023. Generation Probabilities Are Not Enough: Exploring the Effectiveness of Uncertainty Highlighting in AI-Powered Code Completions. doi:10.48550/arXiv.2302.07248 arXiv:2302.07248 [cs].

[91] Helena Vasconcelos, Gagan Bansal, Adam Fourney, Q. Vera Liao, and Jennifer Wortman Vaughan. 2024. Generation Probabilities Are Not Enough: Uncertainty Highlighting in AI Code Completions. *ACM Trans. Comput.-Hum. Interact.* (Oct. 2024). doi:10.1145/3702320 Just Accepted.

[92] Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael Bernstein, and Ranjay Krishna. 2023. Explanations Can Reduce Overreliance on AI Systems During Decision-Making. doi:10.48550/arXiv.2212.06823 arXiv:2212.06823 [cs].

[93] Jayne Wallace, John McCarthy, Peter C. Wright, and Patrick Olivier. 2013. Making design probes work. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. Association for Computing Machinery, New York, NY, USA, 3441–3450. doi:10.1145/2470654.2466473

[94] Huichen Will Wang, Larry Birnbaum, and Vidya Setlur. 2025. Jupybara: Operationalizing a Design Space for Actionable Data Analysis and Storytelling with LLMs. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, 1–24. doi:10.1145/3706598.3713913

[95] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-Consistency Improves Chain of Thought Reasoning in Language Models. doi:10.48550/arXiv.2203.11171 arXiv:2203.11171 [cs].

[96] Zihan Wang, Ziqi Zhao, Yougang Lyu, Zhumin Chen, Maarten de Rijke, and Zhaochun Ren. 2025. A Cooperative Multi-Agent Framework for Zero-Shot Named Entity Recognition. In *Proceedings of the ACM on Web Conference 2025*

(WWW '25). Association for Computing Machinery, New York, NY, USA, 4183–4195. doi:10.1145/3696410.3714923

[97] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2022. Taxonomy of Risks posed by Language Models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 214–229. doi:10.1145/3531146.3533088

[98] Justin D. Weisz, Michael Muller, Stephanie Houde, John Richards, Steven I. Ross, Fernando Martinez, Mayank Agarwal, and Kartik Talamadupula. 2021. Perfection Not Required? Human-AI Partnerships in Code Translation. In *Proceedings of the 26th International Conference on Intelligent User Interfaces (IUI '21)*. Association for Computing Machinery, New York, NY, USA, 402–412. doi:10.1145/3397481.3450656

[99] Daniel S. Weld and Gagan Bansal. 2019. The challenge of crafting intelligible intelligence. *Commun. ACM* 62, 6 (May 2019), 70–79. doi:10.1145/3282486

[100] Frank Xing. 2025. Designing Heterogeneous LLM Agents for Financial Sentiment Analysis. *ACM Trans. Manage. Inf. Syst.* 16, 1 (Feb. 2025), 5:1–5:24. doi:10.1145/3688399

[101] Joshua C. Yang, Damian Dailisan, Marcin Korecki, Carina I. Hausladen, and Dirk Helbing. 2025. LLM Voting: Human Choices and AI Collective Decision-Making. In *Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society (AIES '24)*. AAAI Press, San Jose, California, USA, 1696–1708.

[102] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Understanding the Effect of Accuracy on Trust in Machine Learning Models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/3290605.3300509

[103] Kun Yu, Shlomo Berkovsky, Dan Conway, Ronnie Taib, Jianlong Zhou, and Fang Chen. 2016. Trust and Reliance Based on System Accuracy. In *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization (UMAP '16)*. Association for Computing Machinery, New York, NY, USA, 223–227. doi:10.1145/2930238.2930290

[104] Kun Yu, Shlomo Berkovsky, Ronnie Taib, Jianlong Zhou, and Fang Chen. 2019. Do I trust my machine teammate? an investigation from perception to decision. In *Proceedings of the 24th International Conference on Intelligent User Interfaces (IUI '19)*. Association for Computing Machinery, New York, NY, USA, 460–468. doi:10.1145/3301275.3302277

[105] Mireia Yurrita, Tim Draws, Agathe Balayn, Dave Murray-Rust, Nava Tintarev, and Alessandro Bozzon. 2023. Disentangling Fairness Perceptions in Algorithmic Decision-Making: the Effects of Explanations, Human Oversight, and Contestability. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–21. doi:10.1145/3544548.3581161

[106] Carlos Zednik. 2021. Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence. *Philosophy & Technology* 34, 2 (June 2021), 265–288. doi:10.1007/s13347-019-00382-7

[107] Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 295–305. doi:10.1145/3351095.3372852

[108] Changmeng Zheng, Dayong Liang, Wengyu Zhang, Xiao-Yong Wei, Tat-Seng Chua, and Qing Li. 2024. A Picture Is Worth a Graph: A Blueprint Debate Paradigm for Multimodal Reasoning. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM '24)*. Association for Computing Machinery, New York, NY, USA, 419–428. doi:10.1145/3664647.3681102

[109] Jun-Peng Zhu, Peng Cai, Kai Xu, Li Li, Yishen Sun, Shuai Zhou, Haihuang Su, Liu Tang, and Qi Liu. 2024. AutoTQA: Towards Autonomous Tabular Question Answering through Multi-Agent Large Language Models. *Proc. VLDB Endow.* 17, 12 (Aug. 2024), 3920–3933. doi:10.14778/3685800.3685816

Appendix

A Prompt to GPT4-o for Stimuli Generation

"Generate text stimuli for a research study investigating how users evaluate and interpret large language model (LLM) systems with multiple agents. In this study, participants interact with simulated multi-agent LLM interfaces to receive support on deductive reasoning tasks. These interfaces display will different different agents responding to the same reasoning task.

Your task is to generate agent explanation texts (short rationales that resemble an LLM's response) as part of the experiment stimulus. The reasoning task involves logical reasoning questions from the LSAT examination. These questions present problem scenarios followed by a question, with multiple choices presented (options A–E). These questions assess logical reasoning, have exactly one correct answer.

In the interfaces used in this study, four agents respond to the same LSAT question: three select the correct answer and provide valid rationales, while one selects an incorrect answer and provides a compelling rationale that appears plausible. Plausible incorrect rationales should be easy to generate since the logical reasoning problems are fairly complex. Further, in one of the interfaces, there will be a critic agent which evaluates each of the four agent's rationales, surfacing flaws, gaps, or oversights in their reasoning. Your job is to help generate these base + critical rationales.

Input you will receive:

- The full LSAT question text and its five answer options (A–E).
- The verified correct answer (e.g., "X").
- The official LSAT explanation for the correct answer.
- A specific incorrect answer option for which you will generate a plausible rationale.

Your task consists of two parts:

(1) **Generate four base rationale texts (R1–R4):** R1, R2, and R3 should each provide a distinct, logically valid rationale supporting the correct answer. These should be faithful to the official LSAT explanation but condensed and phrased naturally. R4 should provide a plausible but incorrect rationale supporting the specified wrong option. It should appear logical and persuasive. Each rationale must be a single sentence, 15–17 words long, and follow a consistent tone, clarity level, and reasoning depth. Do not label any rationale as correct or incorrect in its text; each should stand alone and be independently convincing.

(2) **Generate four critique texts (C1–C4):** These will be texts that evaluate the quality of the four base rationales (R1–R4). Each critique text should identify potential weaknesses, limitations, incorrect assumptions, or vague reasoning in the corresponding rationale. All critique texts must also be one sentence of 15–17 words, written in natural, fluent English, and comparable in tone and style to the base rationales.

Style and formatting guidelines:

- Write in fluent, natural English suitable for a chat-based LLM interface. Use British English spellings.
- Maintain consistent structure, length, clarity, and reasoning depth across all rationales and critiques.

- Provide similar contextual detail and tone across all base rationales, and similarly across all critical responses.
- Provide your output in the following format:
  - R1: [Answer <X> is correct. 15–17 word rationale supporting correct answer]
  - R2: [Answer <X> is correct. Distinct 15–17 word rationale supporting correct answer]
  - R3: [Answer <X> is correct. Distinct 15–17 word rationale supporting correct answer]
  - R4: [Answer <Y> is correct. Plausible but flawed 15–17 word rationale supporting incorrect answer]
  - C1: [critique of R1]
  - C2: [critique of R2]
  - C3: [critique of R3]
  - C4: [critique of R4]"

B Scales and Measures

These variables are self-reports, used to describe our sample.

- **Demographics:**
  - Age: (text field)
  - Gender: Man; Woman; Non-binary; Gender-diverse; Prefer not to answer; Prefer to self describe: (text field)
- **LLM Usage Frequency.** Measured by asking the following question: "How frequently do you use chatbots such as ChatGPT, Claude, Gemini, or similar?" on a 7-point scale: Never; Less than once a month; A few times a month; About once a week; A few times a week; About once a day; Multiple times a day.
- **Dispositional Trust in Automation (TiA-PtT).** Measured using the 3-item Propensity to Trust (PtT) subscale of the Trust in Automation (TiA) Scale by Körber [46], rated on a 5-point Likert scale, following prior work in human-AI interaction [33, 68].

(1) One should be careful with unfamiliar AI systems. (reverse coded)
(2) I rather trust an AI system than I mistrust it.
(3) AI systems generally work well.

- **AI Literacy.** Measured using a 4-item, 5-point Likert scale, adopted from Yurrita et al. [105].

(1) I have a good knowledge in the field of artificial intelligence.
(2) My current employment includes working with artificial intelligence.
(3) I am confident interacting with artificial intelligence.
(4) I understand what the term artificial intelligence means.

C Participant Demographic and Dispositional Measures

Table C.1 contains an overview of our participants' demographic and dispositional data.

Measure	Participant Distribution / Summary (N = 12)
Age	*Mean* = **27.4** years, *SD* = **2.4** years, *Median* = **28.5** years
Gender	Women **(*n* = 6, 50%)**, Men **(*n* = 6, 50%)**
LLM Usage Frequency	Never **(*n* = 0)**, Less than once a month **(*n* = 0)**, A few times a month **(*n* = 2, 16.7%)**, About once a week **(*n* = 0)**, A few times a week **(*n* = 2, 16.7%)**, About once a day **(*n* = 0)**, Multiple times a day **(*n* = 8, 66.7%)**
Dispositional Trust in Automation (1–5)	*Mean* = **3.03**, *SD* = **0.72**, *Median* = **2.83**
Self-reported AI Literacy (1–5)	*Mean* = **3.54**, *SD* = **0.87**, *Median* = **3.75**

Table C.1: Participant demographic and dispositional measures.