The Human Burden in AI Collaboration:

Evidence from a Multi-Agent Evaluation of Human–AI Interaction

Robert G. Eccles

Oxford Saïd Business School

April 2026 | Revision 3.0

Abstract

AI capabilities are improving rapidly, but the coordination burden required to use those capabilities effectively remains largely human—and may increase in complex, long-horizon workflows. This paper analyzes how five current-generation, chat-based large language models—ChatGPT, Claude, Gemini Pro, Mistral, and Perplexity—conceptualize the challenges of human–AI collaboration. Each system rated seven collaboration challenges along three dimensions: human difficulty (1–5), impact on output quality (1–5), and likelihood that AI improvements will reduce the challenge (1–5). The findings should be interpreted as structural signals about how these systems understand their own interaction environment, not as direct empirical measurements of human behavior. The results reveal consistent cross-system agreement that reliability assessment (Challenge 3) is both the most difficult and the highest-impact challenge, while attracting the lowest expectations for AI-driven improvement. Managing human cognitive limits (Challenge 6) shows a similar pattern. The only domain of clear optimism is continuous learning (Challenge 7), where all five systems returned identical scores. Taken together, the data support a structural argument: the architectural properties of current chat-based transformer systems—stateless design, session boundaries, absence of persistent memory—systematically externalize coordination to the human operator, creating a capability–usability gap that technical progress in raw model performance alone is unlikely to close.

1. Introduction

The past several years have produced dramatic advances in large language model (LLM) capability. Systems that once struggled to maintain paragraph-level coherence now produce multi-thousand-word structured outputs, engage in sustained reasoning, write and debug code, and perform well on evaluations considered challenging benchmarks only recently. Yet a persistent pattern emerges among practitioners who use these systems intensively: the improvement in raw capability has not been accompanied by a proportional improvement in what might be called collaborative usability—the ease with which a human can deploy that capability reliably, coherently, and at scale across complex projects.[1]

This paper investigates that gap, with an important epistemic clarification at the outset. The study presented here analyzes how AI systems conceptualize the challenges of interacting with them—it is a study of AI self-perception about human–AI collaboration, not a direct empirical measurement of human users. This is treated not as a liability but as a distinctive methodological lens: how these systems characterize the coordination challenges of working with them is itself significant data, particularly when it converges across architecturally distinct systems with different training approaches. Findings should be interpreted as structural signals consistent with the core argument, not as precise measurements of human cognitive effort.

The central claim is that AI capabilities are improving rapidly, but the coordination burden required to use those capabilities effectively remains largely human—and may increase in complex workflows. More precisely, this claim applies to the current generation of chat-based, transformer-based large language models operating within conventional session-bounded architectures.[2] It is not a claim about all AI systems or all possible designs; specialized agentic architectures, multi-agent orchestration systems, and systems with persistent memory represent meaningfully different design paradigms that this study does not evaluate. The claim is structural: that the dominant chat-interface LLM paradigm—stateless, session-bounded, without native project structure—systematically externalizes coordination to the human operator in ways that constrain realized effectiveness even when underlying model quality is high.

To investigate this claim, a structured multi-agent survey was conducted across five systems, asking each to evaluate seven defined challenges in human–AI collaboration along three dimensions. The resulting dataset provides a distinctive form of evidence: five architecturally distinct systems, evaluating the same structured problem space, independently. Agreement across these systems is more informative than agreement within any single one; divergence is equally instructive. Both are examined here.

The paper is organized as follows. Section 2 describes the methodology. Section 3 presents quantitative results. Section 4 synthesizes key findings. Section 5 addresses alternative explanations. Section 6 situates the findings within a Narrative AI Ethnography (NAIE) framework. Section 7 draws implications for the design of Hybrid Intelligent Teams (HITs). Sections 8 and 9 address limitations and future research. Section 10 concludes. Three appendices and a bibliography are attached.

---

[1] Cleotilde Gonzalez et al., "Toward a Science of Human–AI Teaming for Decision Making: A Complementarity Framework," *PNAS Nexus* 5, no. 3 (2026): pgag030.

[2] OpenAI, "GPT-4 System Card," OpenAI, 2023, accessed April 29, 2026, https://cdn.openai.com/papers/gpt-4-system-card.pdf.

2. Methodology

2.1 Survey Design

The evaluation instrument consisted of seven challenges in human–AI collaboration, developed through prior research and iterative refinement. For each challenge, participating systems were asked to provide ratings along three dimensions:

1. Human Difficulty (1–5): How difficult this challenge is for human users to navigate.
2. Impact on Output Quality (1–5): The degree to which this challenge, when unresolved, degrades the quality of AI-assisted outputs.
3. Likelihood AI Will Reduce the Challenge (1–5): The probability that AI improvements over the coming years will substantially mitigate this challenge.

Ratings were elicited independently from each system in a structured format with no prior exposure to other systems' ratings. The scale is ordinal. Means reported in Section 3 are presented as directional summaries across five data points, not as statistically precise estimates; they are included to facilitate comparison across challenges, not to imply inferential validity beyond the sample.

Scope Note: This study evaluates five current-generation, chat-based, transformer-based large language models in their standard conversational interfaces. It does not evaluate specialized agentic systems, retrieval-augmented architectures, or multi-agent orchestration frameworks, which represent meaningfully different design paradigms.[3] *Conclusions should not be generalized to AI systems as a category.*

2.2 The Seven Challenges

The seven challenges were defined as follows:

Challenge 1 — Translating Intent into Instructions: The human's ability to formulate requests in ways that accurately communicate intent and elicit the desired output.

Challenge 2 — Understanding System Capabilities: The human's ability to form accurate mental models of what a given AI system can and cannot do, and to calibrate requests accordingly.

Challenge 3 — Assessing Output Reliability: The human's ability to evaluate whether AI outputs are accurate, coherent, complete, and appropriate—particularly in domains where the human cannot independently verify claims.

Challenge 4 — Managing Context and Performance Limits: The human's ability to manage within context window constraints, session boundaries, and performance variability across prompts.

Challenge 5 — Maintaining Project and Interaction Organization: The human's ability to structure, track, and organize sustained projects across multiple sessions, prompts, and output types.

[3] Yuchen Zhang et al., "OrgAgent: Organize Your Multi-Agent System like a Company," arXiv preprint arXiv:2604.01020 (2026).

Challenge 6 — Managing Human Cognitive Limits: The human's ability to manage their own attention, fatigue, judgment, and decision-making over extended AI-assisted workflows.

Challenge 7 — Adapting Through Continuous Learning: The human's ability to improve their effectiveness over time by learning from past interactions and refining their approach.

2.3 Participating Systems and Evidence Type

Five systems participated: ChatGPT (Clean instance, GPT-4 class), Claude (Anthropic), Gemini Pro (Google DeepMind), Mistral (Mistral AI), and Perplexity (Perplexity AI). These systems represent distinct architectural lineages and product design philosophies. Their inclusion allows comparison across meaningfully different instantiations of current-generation LLM technology.

A critical interpretive note is required here. The dataset reflects what each system believes or infers about human difficulty and system limitations—not empirical measurement of actual human behavior. The systems are not measuring the cognitive effort of human users; they are articulating their understanding of that effort, filtered through training data, alignment tuning, and design assumptions. This framing has a specific implication: convergence across systems should be interpreted as evidence of a shared structural understanding, not as proof that human users actually experience these challenges at the levels indicated. The study is, in formal terms, an analysis of AI self-perception regarding human–AI coordination; empirical validation with human subjects remains an open priority (see Section 9).

2.4 Agent Profiles

To structure interpretation, the five systems can be characterized in terms of their analytical tendencies, with the explicit caveat that these profiles are interpretive lenses only—not settled behavioral characterizations:

ChatGPT: Execution-focused; tends to emphasize concrete task completion and user workflow friction.

Claude: System- and workflow-oriented; tends toward structural analysis of interaction patterns and organizational dynamics.

Gemini: Architecture- and tooling-oriented; tends to foreground technical infrastructure and design considerations.

Mistral: Balanced; provides relatively even assessments across dimensions without strong systematic tendencies.

Perplexity: Workflow- and external-systems-oriented; tends to emphasize integration with external information environments.

3. Results

3.1 Human Difficulty Ratings

Table 1 presents all five systems' difficulty ratings across the seven challenges. Challenge 3 (Assessing Output Reliability) attracted the highest scores by a consistent margin: four of five systems returned the maximum score of 5, with one returning 4—the most concentrated high-end pattern of any challenge in the table. Challenge 7 (Adapting Through Continuous Learning) was the only challenge where all five systems agreed exactly, all returning 3, making it the single point of unanimous agreement in this dimension. Challenge 5 (Maintaining Project and Interaction Organization) produced notable divergence: ChatGPT and Claude both returned 5, while Gemini, Mistral, and Perplexity each returned 4. Challenge 6 (Managing Human Cognitive Limits) also showed divergence, with Perplexity returning 5, ChatGPT and Claude returning 4, and Gemini and Mistral returning 3.

Table 1. Human Difficulty Ratings (1–5, where 5 = most difficult)

Challenge	ChatGPT	Claude	Gemini	Mistral	Perplexity
C1: Translating Intent into Instructions	4	3	4	4	4
C2: Understanding System Capabilities	4	4	3	4	4
C3: Assessing Output Reliability	5	4	5	5	5
C4: Managing Context and Performance Limits	4	3	4	4	4
C5: Maintaining Project and Interaction Organization	5	5	4	4	4
C6: Managing Human Cognitive Limits	4	4	3	3	5
C7: Adapting Through Continuous Learning	3	3	3	3	3

3.2 Impact on Output Quality Ratings

Table 2 presents impact ratings. Challenge 3 achieved the maximum score of 5 from all five systems—the only challenge to achieve this unanimity at the highest level, and the strongest convergence in the dataset. Challenge 1 (Translating Intent into Instructions) was rated at either 4 or 5 by all systems, with a consistent high-end profile across the board. Challenge 7 again achieved unanimity at the lower end, with all five systems returning 3—the only challenge where impact and difficulty were both unanimously scored. Challenge 5 produced the widest spread in this dimension: ChatGPT and Perplexity returned 5, Claude and Mistral returned 4, and Gemini returned 3. This three-way split represents the largest intra-challenge divergence in the impact dimension.

Table 2. Impact on Output Quality Ratings (1–5, where 5 = highest impact)

Challenge	ChatGPT	Claude	Gemini	Mistral	Perplexity
C1: Translating Intent into Instructions	5	4	5	5	4
C2: Understanding System Capabilities	4	3	4	4	4
C3: Assessing Output Reliability	5	5	5	5	5
C4: Managing Context and Performance Limits	4	4	4	4	4
C5: Maintaining Project and Interaction Organization	5	4	3	4	5
C6: Managing Human Cognitive Limits	4	4	4	3	5
C7: Adapting Through Continuous Learning	3	3	3	3	3

3.3 Likelihood AI Will Reduce the Challenge

Table 3 presents reduction likelihood ratings. Challenge 6 (Managing Human Cognitive Limits) produced the most pessimistic and cohesive cross-system pattern: four systems returned 2 and Perplexity returned 3—the lowest and most tightly clustered set of scores in the entire table.[4] Challenge 3 (Assessing Output Reliability) also attracted consistently low scores, with Mistral returning 2 and the remaining four systems returning 3, making it the second-lowest-rated challenge for AI improvement. Challenge 7 (Adapting Through Continuous Learning) produced the most optimistic and cohesive pattern: all five systems returned 4, matching Challenge 4 in average level though with more variance in the latter (Gemini returned 5; Mistral returned 3). The data reveal a directional inversion: the challenges rated as hardest for humans are also those for which AI-driven improvement is least expected.

Table 3. Likelihood AI Will Reduce the Challenge (1–5, where 5 = most likely)

Challenge	ChatGPT	Claude	Gemini	Mistral	Perplexity
C1: Translating Intent into Instructions	4	3	4	3	3
C2: Understanding System Capabilities	3	4	3	4	3
C3: Assessing Output Reliability	3	3	3	2	3
C4: Managing Context and Performance Limits	4	4	5	3	4
C5: Maintaining Project and Interaction Organization	4	3	4	3	4

[4]John Sweller, "Cognitive Load During Problem Solving: Effects on Learning," *Cognitive Science* 12, no. 2 (1988): 257–85.

Challenge	ChatGPT	Claude	Gemini	Mistral	Perplexity
C6: Managing Human Cognitive Limits	2	2	2	2	3
C7: Adapting Through Continuous Learning	4	4	4	4	4

3.4 Cross-Dimensional Summary

Table 4 presents summary comparisons across all five systems for each challenge and dimension. The final column shows the arithmetic difference between the average difficulty rating and the average AI-reduction likelihood rating. This is presented as a heuristic ordering device—not as a formally derived metric. Because the two scales are ordinal and were developed to measure distinct constructs, subtracting them does not produce a mathematically rigorous quantity. The value of the comparison is directional: it allows challenges to be ranked by the degree to which high human difficulty coincides with low expected AI improvement. Challenges where this gap is positive and large are those where the coordination burden appears most persistent and least addressable through AI progress alone.

Table 4. Directional Summary by Challenge (means as ordinal summaries across 5 systems; gap index is heuristic only)

Challenge	Difficulty (Mean)	Impact (Mean)	AI Reduction (Mean)	Gap (Heuristic)
C1: Translating Intent into Instructions	3.8	4.6	3.4	0.4
C2: Understanding System Capabilities	3.8	3.8	3.4	0.4
C3: Assessing Output Reliability	4.8	5.0	2.8	2.0
C4: Managing Context and Performance Limits	3.8	4.0	4.0	−0.2
C5: Maintaining Project and Interaction Organization	4.4	4.2	3.6	0.8
C6: Managing Human Cognitive Limits	3.8	4.0	2.2	1.6
C7: Adapting Through Continuous Learning	3.0	3.0	4.0	−1.0

Interpreting directionally: Challenge 3 has the largest positive gap value (2.0), indicating that it is rated both consistently highly difficult and consistently unlikely to improve with AI advances—the profile most characteristic of a persistent structural constraint. Challenge 6 has the second-largest positive gap (1.6), with similar characteristics. Challenge 7 has the most negative gap value (−1.0), indicating that AI improvement is expected to outpace current difficulty—a profile

consistent with a tractable near-term opportunity. All gap values are ordinal comparisons and should not be interpreted as precise quantities.

4. Key Findings

4.1 The Human Burden Is the Central Pattern

The most consistent cross-system finding is not a specific score but a shared structural inference: the burden of making AI collaboration work falls predominantly on the human. Across all seven challenges, qualitative patterns from all five systems emphasize structure, discipline, workflow management, and coordination responsibility as human obligations. This convergence across architecturally distinct systems is unlikely to be incidental, though alternative explanations—including shared training data—are addressed in Section 5.

The specific burdens identified—tracking, structuring, coordinating, and validating—are not trivial overhead. They are cognitively demanding, require sustained attention, and resist delegation because they require judgment about what matters, what should be trusted, and what needs to be organized. The data are consistent with the interpretation that current chat-based LLM systems are designed in ways that shift these demands onto the human rather than absorbing them.

4.2 Reliability Assessment Is the Hardest and Least Improvable Challenge

Challenge 3 (Assessing Output Reliability) received the highest difficulty scores of any challenge—four of five systems at maximum with one below—and unanimous maximum impact scores, the only challenge to achieve this in the impact table. More importantly, its AI-reduction likelihood scores were among the lowest in the dataset, with four systems returning 3 and Mistral returning 2. This configuration—consistently high difficulty and impact, consistently low improvement expectation—is what one would expect of a structural constraint rather than a temporary technical limitation.

The difficulty of reliability assessment does not arise primarily from user unfamiliarity with AI. It arises because the systems themselves cannot reliably signal when they are wrong, and the human lacks an independent standard of comparison in many high-value domains. More capable models may produce more convincing outputs that are nonetheless incorrect, potentially intensifying the problem before it improves.

4.3 Human Cognitive Limits Are Treated as Largely Outside AI's Reach

Challenge 6 (Managing Human Cognitive Limits) received the lowest cluster of AI-reduction likelihood scores in the dataset: four systems at 2 and one at 3. This near-unanimous pattern stands in contrast to the moderate difficulty and impact ratings for this challenge, which sit in the middle of the range. The implication is not that cognitive limits are trivial—they are rated as meaningfully difficult and impactful—but that AI progress is not expected to address them directly.

This finding points to a category of coordination cost that is endogenous to the human operator and therefore largely invariant with respect to AI technical improvement. As AI systems become more capable and workflows more complex, the demands on human cognition may increase rather than decrease, even if the AI is doing more of the constituent work.

4.4 Challenge 7 Is the Unique Consensus Domain

Challenge 7 (Adapting Through Continuous Learning) is the only challenge where all five systems returned identical scores across all three dimensions: difficulty 3, impact 3, likelihood 4. It suggests that adaptive learning—improving through feedback, refining prompts, building personal knowledge about effective interaction—is treated as uniformly moderate in current difficulty and output impact, but uniformly optimistic about AI-assisted improvement. Whether this optimism reflects expected architectural changes (persistent memory, adaptive interfaces) or the inherently iterative nature of learning tasks, it represents the clearest near-term improvement vector in the dataset.

4.5 The Capability–Usability Gap

Across all seven challenges, the data reveal a consistent profile: current-generation chat-based LLMs are well-suited to bounded, single-session tasks with clear success criteria, and significantly less suited to long-horizon workflows involving coordination, validation, and sustained organizational discipline. This is the capability–usability gap in concrete form.[5]

A working estimate consistent with the data is that even experienced practitioners operating in complex, multi-session AI-assisted workflows may achieve effective utilization in the range of 5–6 out of 10. This is not a failure of skill; it is a structural consequence of an architectural design that does not natively support the persistence, context management, and coordination scaffolding that sustained collaborative work requires.

4.6 Systems Know More Than They Do

Note: The following observation is based on qualitative analysis of system responses and cannot be directly read from the 1–5 rating tables. It is included as an interpretive finding, clearly labeled as such.

When asked to describe how the identified challenges could be addressed, the systems tended to articulate well-reasoned prescriptions for human behavior—maintain structured notes, decompose tasks, validate outputs, manage fatigue, develop prompt discipline. These prescriptions reflect genuine understanding of effective practice.

Yet the same systems do not implement that understanding as system-level behavior. They can advise users to manage context carefully but do not themselves manage context across sessions. They can advise users to track reliability but do not flag uncertainty in ways that reduce verification burden. The explanation is architectural: model knowledge—what is encoded in weights through training—is not the same as system-level behavior, which is determined by design constraints. A system can encode knowledge about the value of persistent memory without having persistent memory. This gap between model knowledge and system capability is, qualitatively, another form of coordination cost absorbed by the human.

[5] Lisanne Bainbridge, "Ironies of Automation," *Automatica* 19, no. 6 (1983): 775–79.

5. Alternative Explanations

The convergence of five systems on a consistent structural pattern is the primary evidentiary basis for the paper's core argument. However, three plausible alternative explanations deserve explicit acknowledgment and should be considered when interpreting the findings.

5.1 Shared Training Data and Pre-Trained Consensus

The most significant alternative explanation is that the convergence reflects not independent assessment of human–AI coordination challenges, but shared training on a common body of text. If all five systems were trained on substantial overlapping corpora—including AI safety literature, human-computer interaction research, and public discourse about LLM limitations—their agreement on Challenge 3 and Challenge 6 may primarily reflect the consensus of that shared corpus rather than independent structural analysis.

This explanation cannot be ruled out with the current data. Its force is partially mitigated by the systematic divergences observed (the three-way split on Challenge 5 impact; Gemini's outlier optimism on Challenge 4 reduction likelihood; Perplexity's elevated assessment of Challenge 6 difficulty), which suggest that the systems are not simply replaying identical training data. But partial divergence does not neutralize the pre-trained consensus concern. Readers should treat the convergence findings as consistent with—but not proof of—a structural claim about current LLM architecture.

5.2 Alignment Tuning and Modesty Bias

Several of the participating systems—particularly Claude and ChatGPT—are trained with substantial reinforcement learning from human feedback (RLHF) and related alignment techniques designed to make them helpful, honest, and appropriately cautious about their own limitations. This training may produce a systematic tendency to rate their own contribution to challenge reduction as lower than warranted—a form of conferred modesty that makes low AI-reduction likelihood scores a predictable byproduct of alignment choices, not an objective assessment of architectural constraints.

The existence of the modesty bias concern supports the case for human-subject validation: if human practitioners independently rate AI-reduction likelihood for Challenge 6 as low, the structural interpretation is strengthened; if they rate it higher, the modesty bias interpretation gains ground.

5.3 Prompt Framing Effects

The seven challenge descriptions were authored by the researcher and provided to each system as part of the evaluation instrument. If those descriptions were framed in ways that emphasized human effort or coordination complexity, the systems—designed to be responsive and contextually aligned—may have adopted that framing in their ratings, producing scores that reflect the researcher's implicit model rather than independent assessment.

This concern is inherent to survey-based research and applies with particular force when the respondents are systems tuned to be agreeable and contextually sensitive. Partial mitigation comes from the presence of genuine divergences in the data, and from the fact that the challenge

descriptions were designed to be neutral characterizations of widely recognized problems. Nevertheless, the possibility of framing influence should be held in view when interpreting specific ratings.

5.4 Summary Assessment

None of the three alternative explanations is fatal to the paper's core argument, but none can be dismissed with the current data. The most defensible claim is that the findings are structurally consistent with the coordination externalization hypothesis—and that this hypothesis provides a more parsimonious account of the observed convergence pattern than the alternatives taken individually. Definitive discrimination among these explanations requires human-subject data and experimental variation in prompt framing, both of which remain priorities for subsequent research.

6. NAIE Insights

Narrative AI Ethnography (NAIE) treats sustained human–AI interaction as ethnographic data—attending not only to what is produced but to how collaboration unfolds, what patterns of behavior emerge, and what the interaction structure itself reveals about the epistemic and organizational dynamics of human–machine cognition.[6]

From a NAIE perspective, the most significant feature of this study is not the ratings but the structure of their production.[7] Five AI systems were asked to evaluate their own interaction environment—to assess, in effect, how hard they are to use and whether they are likely to become easier. The act of eliciting this data is itself an example of the primary unit of analysis in NAIE: the loop, the exchange between human and machine that generates something neither would have produced alone.

Several observations follow from this framing. First, the convergence of the five systems on the structural location of the human burden constitutes a consistent signal in the data: systems from different organizations, with different training approaches, produce compatible characterizations of where coordination costs are located. Whether this reflects genuine independent assessment, shared training data, or alignment-induced modesty (see Section 5) is an open question, but the convergence itself is a fact about the dataset that any structural account of the domain must address.

Second, the divergences are informative. The three-way split in Challenge 5 impact ratings (5 from ChatGPT and Perplexity; 4 from Claude and Mistral; 3 from Gemini) reflects genuinely different assessments of how seriously sustained project organization degrades output quality. These differences are consistent with the architectural orientations of the respective systems. The Gemini rating of 3 may reflect a design context in which external tooling and project management integration are assumed; the ChatGPT rating of 5 may reflect more direct observation of output degradation in unstructured multi-session work.

Third, the unanimous scores on Challenge 7 are a NAIE signal worth interpreting. When five architecturally distinct systems agree completely, that agreement warrants attention. In this case, the unanimity suggests a shared structural understanding of what adaptive learning requires and where near-term improvement is most tractable—a convergence on the same improvement vector across very different systems.

Fourth, the metadata of the study process itself is analytically relevant. One of the systems, when operating in a batch-oriented deep research mode, continued executing its predefined task sequence when the researcher attempted to introduce a meta-level question. The system did not re-orient; it resumed production. This is documented in Appendix C and treated as a concrete behavioral instance of the coordination asymmetry the study identifies. The system was operating as designed: task completion was weighted above conversational re-alignment. The human bore the cost of recognizing the misalignment and re-establishing the interaction. No system feature performed that work.

---

[6] Iyad Rahwan et al., "Machine Behaviour," *Nature* 568, no. 7753 (2019): 477–86.

[7] Inkeri Koskinen, "We Have No Satisfactory Social Epistemology of AI-Based Science," *Social Epistemology* 38, no. 3 (2024): 255–71.

From a NAIE standpoint, this incident is significant not as a malfunction but as a design signal. It is a moment in which the architectural priorities of the system—execution over interaction management—became observable in real time. Whether that prioritization is appropriate depends on context, but its existence is a design fact. It illustrates the argument of the broader paper: the coordination layer is not internal to the system; it is externalized to the human.

7. Implications for HIT Design

Hybrid Intelligent Teams (HITs) are collaborative structures combining human and AI contributors in defined roles to accomplish tasks that neither could complete alone at the same quality and scale.[8] The findings of this study have direct implications for how HITs should be designed, and for what a responsible account of HIT effectiveness must include.

7.1 The Coordination Infrastructure Problem

The most immediate design implication is the need for coordination infrastructure—the tools, practices, protocols, and workflow structures that absorb the coordination costs currently borne by the human. These include session handoff protocols, output logging and versioning systems, validation checklists keyed to task type, and explicit project management scaffolding. The Challenge 5 pattern (consistently high difficulty and impact) makes the stakes clear: organizational challenges in sustained AI-assisted work are both hard to manage and significantly impactful when unmanaged. Addressing them through individual discipline is insufficient at scale. HIT design should treat coordination infrastructure as a first-class design problem.

7.2 The Reliability Verification Problem

Challenge 3's consistent profile—high difficulty, maximum impact, low improvement expectation—implies that HIT design must incorporate structural reliability verification mechanisms that do not depend on AI self-assessment. These mechanisms might include mandatory cross-checking protocols, independent human review at defined checkpoints, or multi-system triangulation. The data underscore why these are not optional: the cost of undetected unreliability in AI-assisted work is high, and the expectation of AI-led reduction in that cost is low.

7.3 The Cognitive Sustainability Problem

The pattern on Challenge 6 implies that HIT design must treat human cognitive capacity as a binding constraint. Effective HITs will not be designed on the assumption that human collaborators can sustain indefinite attention and judgment even as AI systems become more capable. Sustainable HIT design requires explicit attention to cognitive load management, session length constraints, decision fatigue mitigation, and the careful scoping of tasks that require high-quality human judgment.

7.4 Optimism and Its Conditions

The unanimous optimism on Challenge 7 suggests that HITs should be designed to accumulate and deploy learning over time. Interaction logs, prompt libraries, retrospective reviews, and personalized adaptation mechanisms are natural design elements in this domain. The data suggest these are also the most tractable near-term improvements, and HIT design should build that improvement in from the outset as a counterbalance to the structural constraints identified elsewhere.

---

[8] Robert G. Eccles, "Hybrid Intelligence Teams: A Theoretical Framework for Human–AI Collaboration in Knowledge Work," working paper, 2025, available at SSRN (forthcoming); preprint PDF at roberteccles.com.

8. Limitations

Several limitations require direct and candid acknowledgment. They are not presented here as minor qualifications; they are genuine constraints on the strength of claims that can be drawn from this dataset.

8.1 AI Self-Assessment Is Not Human Measurement

The most fundamental limitation is that this study measures how AI systems understand human–AI coordination challenges—not how human users actually experience them. The gap between AI self-perception and human reality is empirically unknown and could be substantial. Without human-subject data, the paper's core argument remains a structurally consistent hypothesis rather than an empirically confirmed finding. The absence of a human validation study is the single most significant limitation of the work.

8.2 Small Sample (n = 5)

Five AI systems providing ratings across seven challenges and three dimensions produces 105 data points. This is sufficient to identify directional patterns and make ordinal comparisons, but it does not support statistical inference, confidence intervals, or claims about representativeness. The calculated means in Table 4 are useful as directional summaries; they should not be read as precise estimates of any underlying population quantity.

8.3 Alternative Explanations Cannot Be Ruled Out

As detailed in Section 5, three alternative explanations—shared training data, alignment-induced modesty, and prompt framing effects—are individually plausible and collectively significant. None can be ruled out with the current design. The paper's core argument is more parsimonious than these alternatives considered individually, but parsimony is not proof.

8.4 Scope Limited to Chat-Based LLMs

All five participating systems operate within the same basic paradigm: chat-based, transformer-based, session-bounded conversational interfaces. The paper's structural argument follows from architectural properties shared by all five systems. This means the study cannot speak to whether specialized agentic architectures, persistent-memory systems, or multi-agent frameworks would produce different patterns. Generalization to AI systems as a category is not warranted.

8.5 No Causal Inference

The study design does not allow causal inference. The directional pattern—high difficulty challenges correlate with low AI-improvement expectations—is consistent with a structural explanation but equally consistent with shared training data producing correlated responses. No experimental manipulation was performed, and no control conditions exist.

8.6 Absence of Human Validation Study

This limitation merits separate statement from 8.1 because of its direct bearing on the paper's central claim. The claim that coordination burden remains human is supported here only by AI

self-assessment. Whether human users experience these responsibilities as burdensome in the ways described, whether AI systems are correctly characterizing their own architectural constraints, and whether the burdens are growing rather than stable are all empirical questions that require human-subject data to answer. The most important next step is a parallel survey administered to 50–100 intensive AI practitioners using the same seven-challenge instrument.

9. Future Research

Several directions follow naturally from these findings and from the limitations identified above.

The most pressing is empirical validation with human users. The AI self-assessment findings generate specific testable hypotheses: that human practitioners will rate Challenge 3 as the most difficult and least AI-improvable; that Challenge 7 will be rated more tractable than other challenges; that the difficulty–improvement inversion will replicate. If human practitioners rate Challenge 3 as a 3 on difficulty while AI systems rate it near 5, the convergence argument requires substantial revision.

A second direction is longitudinal tracking. Repeating this evaluation at regular intervals with the same instrument would produce time-series data allowing assessment of whether the coordination burden profile is shifting and which challenges are being addressed by architectural improvements.

A third direction is experimental variation in prompt framing. Administering the same instrument with challenge descriptions framed neutrally, positively, and negatively would provide direct evidence on the prompt framing alternative explanation identified in Section 5.3.

A fourth direction is HIT design experimentation. Controlled comparisons of HIT designs with and without explicit coordination infrastructure, reliability verification mechanisms, and cognitive sustainability protocols would provide practice-grounded evidence that design recommendations require.

A fifth direction is cross-domain comparison. High-stakes professional contexts (legal analysis, medical decision support, financial modeling) may show substantially different challenge profiles from creative or research contexts. Domain-stratified replication would clarify where the general findings apply and where they require qualification.

A sixth direction concerns the NAIE methodology itself. Further development of criteria for distinguishing genuine knowledge production from sophisticated pattern completion, and for validating ethnographic observations against behavioral data, remains a priority for the broader research program.

10. Conclusion

The question this paper begins with is: as AI capabilities improve, does the human burden of using those capabilities decrease proportionally? The data from this multi-agent evaluation—interpreted with the limitations of AI self-assessment explicitly in view—are consistent with the answer no, at least not across the full range of challenges relevant to complex, sustained, high-quality AI-assisted work.

Reliability assessment is consistently rated the hardest challenge and the one least likely to improve through AI advances. Managing human cognitive limits attracts near-uniform pessimism about AI-driven reduction. Sustained project organization imposes costs that current chat-based LLM architectures make no native provision to absorb. In each case, the design of current session-bounded, stateless systems externalizes the coordination cost to the human. This is the core structural claim.

The qualitative finding—that systems articulate sound prescriptions for better interaction but do not implement those prescriptions as system behavior—identifies a specific and addressable gap between model knowledge and system design. Closing that gap requires architectural change, not only better models.

It is important to be explicit about what this paper does and does not claim with respect to productivity. The analysis focuses on the coordination burden of AI-assisted work; it does not analyze net utility, output volume, or overall productivity gains. These are different questions. AI systems that impose high coordination costs may nonetheless produce very large gains in output quality or quantity—gains that dwarf the burden. If AI increases effective output by a factor of ten while increasing coordination overhead by a factor of two, the net productivity is strongly positive, and the burden is arguably a modest price. The paper does not challenge that framing. What it argues is that the burden is real, structural, and not being addressed at the rate that capability is improving—and that this asymmetry has design consequences that cannot be assumed away by pointing to net productivity gains.

The optimism around continuous learning and adaptive improvement is genuine and should inform near-term design priorities. But it does not offset the structural constraints in reliability and cognition. The central asymmetry of human–AI collaboration—AI scales capability; humans remain cognitively bounded—is not resolved by better prompting alone. It requires deliberate design of the systems, workflows, and team structures through which AI capability is translated into human-accessible value.

The human burden in AI collaboration is, for now, a structural feature of the dominant LLM paradigm. Acknowledging it precisely—and distinguishing it from questions of net productivity—is the first condition for addressing it responsibly.

Appendix A: Survey Instrument (Summary)

The survey instrument consisted of seven challenge descriptions, each followed by three structured rating prompts. The rating prompts were identical across all seven challenges:

4. On a scale of 1–5, how difficult is this challenge for human users to navigate? (1 = minimal difficulty; 5 = extremely difficult)
5. On a scale of 1–5, how significantly does this challenge, when unresolved, impact the quality of AI-assisted outputs? (1 = minimal impact; 5 = severe impact)
6. On a scale of 1–5, how likely is it that improvements in AI systems over the next few years will substantially reduce this challenge? (1 = unlikely; 5 = very likely)

Each challenge was described in two to three sentences establishing the nature of the coordination problem and its typical manifestation in practice. Ratings were elicited in a single structured prompt per system, with no prior exposure to other systems' responses. Systems were also invited to provide qualitative observations following each rating, which informed the qualitative pattern analysis reported in Sections 4 and 6.

Appendix B: Agent Response Summary

Full agent responses are archived separately. The following notes summarize the most analytically significant features of each system's response profile.

ChatGPT: Consistently assigned maximum or near-maximum difficulty and impact scores. The outlier pattern was Challenge 7, where ChatGPT returned 3/3/4—identical to all other systems. Qualitative responses emphasized concrete workflow friction, execution overhead, and structured task decomposition.

Claude: The only system to rate Challenge 1 difficulty below 4 (rating: 3) and Challenge 4 difficulty below 4 (rating: 3). Also the only system to return 3 on capability understanding impact (Challenge 2). Claude's pattern suggests a somewhat lower severity assessment for several challenges, with the exception of Challenge 5, where it matched ChatGPT at the maximum. Qualitative responses were notable for structural analysis of interaction patterns.

Gemini: Rated Challenge 4 reduction likelihood at 5—the single highest score in the reduction likelihood table, indicating stronger optimism about context management improvements than other systems. Also the system with the lowest impact rating for Challenge 5 (3 versus 4–5 from others). Qualitative responses emphasized tooling and architectural considerations.

Mistral: Provided the single lowest score in the entire dataset: 2 for Challenge 3 reduction likelihood—particularly strong pessimism about AI-led improvement in reliability assessment. Challenge 6 reduction likelihood also returned 2, consistent with other systems. Qualitative responses were relatively balanced across challenge types.

Perplexity: The only system to rate Challenge 6 difficulty and impact at 5 (others: 3–4 for difficulty, 3–5 for impact), indicating the strongest assessment of human cognitive limits as a challenge. Also the only system to return 3 on Challenge 6 reduction likelihood (others: 2). Qualitative responses emphasized integration with external information workflows.

Appendix C: Methodological Reflection

A notable event during the research process bears direct methodological relevance to the study's core findings.

During the data collection phase, one of the systems was operating in a batch-oriented deep research mode—a workflow configuration that optimizes for sustained, sequential task execution rather than iterative conversational exchange. When the principal investigator introduced a meta-level question mid-process, the system did not respond to the question. Instead, it resumed—or effectively restarted—its predefined task execution sequence, treating the interruption as a signal to continue rather than to re-orient.

This incident is not reported as a malfunction. The system behaved as designed: its design optimized for task completion. The cost of that optimization—the human's inability to introduce a real-time course correction without disrupting the entire workflow—was borne entirely by the human. The human had to recognize the misalignment, determine how to intervene, and re-establish collaborative orientation. No system feature performed any part of that work.

One alternative explanation is worth noting: the behavior may reflect a software implementation detail—a timeout, a buffering behavior, or a specific feature of the deep research mode—rather than a general design orientation. This cannot be ruled out from observation alone. The incident is treated here as consistent with the paper's structural argument rather than as proof of it.

From a NAIE standpoint, the incident is significant as a behavioral data point about the interaction structure. The fact that a meta-level question was not recognized as interrupting the task reveals, in this instance, an architectural prioritization of execution over conversational re-alignment. Whether that prioritization is appropriate depends on context. Its presence in a workflow requiring ongoing collaborative adjustment is a design constraint that the human must manage—precisely the kind of coordination cost this paper analyzes.

Bibliography

Bainbridge, Lisanne. “Ironies of Automation.” *Automatica* 19, no. 6 (1983): 775–79.

Eccles, Robert G. “Hybrid Intelligence Teams: A Theoretical Framework for Human–AI Collaboration in Knowledge Work.” Working paper, 2025. Available at SSRN (forthcoming); preprint PDF at roberteccles.com.

Gonzalez, Cleotilde, Ioannis Pavlidis, F. Javier Lerch, et al. “Toward a Science of Human–AI Teaming for Decision Making: A Complementarity Framework.” *PNAS Nexus* 5, no. 3 (2026): pgag030.

Koskinen, Inkeri. “We Have No Satisfactory Social Epistemology of AI-Based Science.” *Social Epistemology* 38, no. 3 (2024): 255–71.

OpenAI. “GPT-4 System Card.” OpenAI, 2023. Accessed April 29, 2026. https://cdn.openai.com/papers/gpt-4-system-card.pdf.

Rahwan, Iyad, Manuel Cebrian, Nick Obradovich, et al. “Machine Behaviour.” *Nature* 568, no. 7753 (2019): 477–86.

Sweller, John. “Cognitive Load During Problem Solving: Effects on Learning.” *Cognitive Science* 12, no. 2 (1988): 257–85.

Zhang, Yuchen, et al. “OrgAgent: Organize Your Multi-Agent System like a Company.” arXiv preprint arXiv:2604.01020 (2026).

— End of Paper — Revision 3.0 —