AI, Take the Wheel: What Drives Delegation and Trust in Human–Computer Cooperative Question Answering?

Maharshi Gor, Yoo Yeon Sung, and Yu Hou
University of Maryland
College Park, MD, USA

Eve Fliesig
University of California
Berkeley, CA, USA

Irene Ying
Phasechange.ai
Lakewood, CO, USA

Tianyi Zhou
MBZUAI
Abu Dhabi, UAE

Jordan Boyd-Graber
University of Maryland
College Park, MD, USA

Abstract

AI systems are fallible, and humans can make mistakes in deciding whether to trust AI over their own judgment. Thus, improving human-AI collaboration requires that we understand when, why, and how humans decide to rely on AI. We study two reliance decisions: delegating a task to AI without seeing its output (*whether* AI is used) and evaluating AI suggestions to decide whether to adopt them (*how* AI output shapes final decisions). Both matter for effective collaboration, yet prior work lacks naturalistic experiments capturing both patterns for the same users. We address this gap by studying collaborative human–AI teams competing in a question-answering game in which humans can choose when and how to work with AI agents to win. Our 24 matches pair 23 expert humans with 16 AI agents, capturing 387 delegation and 1440 adoption decisions. While human-AI collaboration performs better than either AI or humans alone, humans make suboptimal collaboration decisions, both under-relying on correct AI suggestions (3.7% of opportunities missed) and over-relying when AI misleads them (1.5%). Both parties contribute wrong answers: reported model confidence is near chance when humans and AI disagree, while confirmation bias drives higher under-reliance (60.7%) when an AI suggestion agrees with humans' initial incorrect answer.

1 Introduction: How much do you trust LLM output?

A wide range of people—from casual users looking for basic explanations to domain experts requiring professional support in medicine (Leonard et al., 2024), law (Magesh et al., 2024), and finance (Maple et al., 2024)—are using AI.[1] However, some users over-trust AI answers because of limited time or imperfect knowledge of AI capabilities (Goddard et al., 2012; Bansal et al., 2019; Buçinca et al., 2021), while other users are skeptical of AI output when it would indeed help them (Kleinberg et al., 2018; Jakesch et al., 2023).

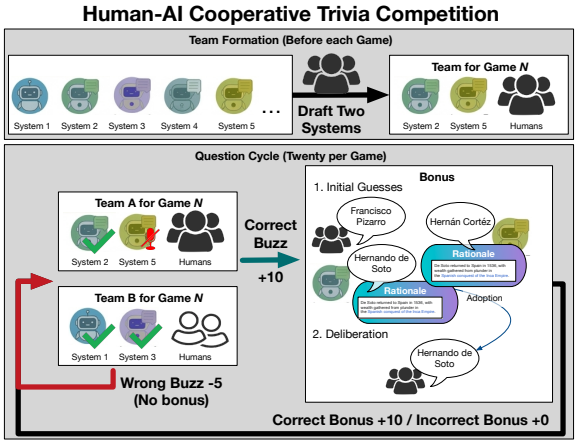

Figure 1: Our experimental setup has humans working with AI teammates in two competitive QA settings. In the tossup phase, AI teammates can directly answer questions without human intervention. In the bonus phase, humans and AI collaborate to reach consensus, with humans providing the final answer.

Prior work (Lee and See, 2004) focuses on trust and adoption of LLM output through post-hoc acceptance measures or synthetic tasks that overlook real-world reliance dynamics and strategic choices under uncertainty (Section 5). A recent taxonomy of 66 human–AI studies confirms that capturing both delegation and advisory patterns in a single study remains rare (Gomez et al., 2025). Moreover, real-world reliance decisions happen under pressure: limited time, imperfect knowledge of AI capabilities (even among experienced users), existing human social dynamics, and limited chances to learn through repeated interaction.

We address this gap by studying how skilled users in human-AI *teams* use AI outputs in a

[1]Throughout the paper, we use AI as a stand-in for text-based agentic workflows that are built using transformer-based LLMs (e.g., GPT-4.1, Claude 3.5, etc.); details in Appendix G.

Findings of the Association for Computational Linguistics: ACL 2026, pages 8664–8689

live **novel collaborative benchmark**[2] that measures *whether* and *how* humans rely on AI assistance. *Proactive delegation* is the decision to let AI act autonomously without reviewing its output—capturing *whether* AI will be used. *Deliberative adoption* is the decision to accept or reject AI output after evaluating it—capturing *how* AI output shapes final decisions.

We test proactive delegation and deliberative adoption through a competitive trivia tournament (Section 2): humans form teams with AI teammates (Figure 1). In **team formation**, teams draft AI systems from a pool of based on perceived capabilities, reflecting humans' perception of AI capabilities. There are two types of questions: tossups and bonuss. In the **tossup** phase (proactive delegation), teams decide whether to let their chosen AI system answer autonomously or mute it entirely, revealing beliefs about AI reliability without oversight. In the **bonus** phase (deliberative adoption), teams see AI suggestions with confidence scores and explanations, then decide whether to adopt them, showing what model outputs contribute to trust and how useful those outputs are. These three scenarios capture complementary facets of human–AI collaboration. Analyzing all three provides insights into "how" people rely on AI: not just whether they accept or reject suggestions, but which models they trust, when, and why. For collaboration to work, the questions must require both human and computer skills (Section 3.2).

While collaboration is mostly synergistic—better than either alone—it is not perfect (Section 4). **Teams miscalibrate trust**, primarily through under-reliance: they fail to adopt correct AI answers 3.7% of the time when they initially propose incorrect answers. Over-reliance is less common (1.5%), but teams override their correct answers with incorrect AI suggestions. Both parties contribute to these errors. On the **AI side**, confidence scores are poorly calibrated: when humans and AI disagree, relying on model confidence to select the correct answer performs near chance. On the **human side**, confirmation bias amplifies mistakes—when an incorrect human answer is confirmed by one of the two AI teammates, under-reliance rises to 57.6% as agreement signals reinforce wrong judgments. High-skill teams are particularly susceptible to this effect, as expertise breeds overconfidence in initial judgments.

The successful collaborations provide **a blueprint for improving future human–AI collaboration**: we distill five design principles throughout our analysis (Section 4). In the bonus phase, explanations that cite specific question clues help humans abandon wrong answers 12% more often. However, we note that features that predict AI correctness (reasoning coherence, question understanding) differ from what humans trust (surface similarity, presence of quotes). Encouragingly, human teams improve with practice: across bonus rounds, inaccuracies decrease from 28% to 18%. Most strikingly, teams reach correct answers in 5.5% of cases where neither humans nor AI were initially right.

2 Game Design for Human–AI Collaborative Question Answering

We study human-AI collaboration through a competitive trivia tournament where mixed teams face off in question-answering games. Each team has up to three human players and two AIs. Two teams compete head-to-head in a game that alternates between two question types: "tossup" questions, where any player from either team can buzz in to answer individually, and "bonus" questions, where the team that answered the tossup correctly collaborates on a three-part question. Before gameplay, teams draft which AI agents to work with from a pool of available systems (Section 3.1).

This design is grounded in the trivia domain (Joshi et al., 2017), which provides both challenging questions and a motivated, enthusiastic community (Jennings, 2006). Koivisto and Hamari (2019) argue that empirical human data collection is more effective when participants are intrinsically motivated. Unlike synthetic laboratory studies, competitive trivia tournaments provide players who are familiar with the task, face real stakes with costs for incorrect reliance on agents, and deliberate with other humans. Our format is based on Quizbowl (Boyd-Graber et al., 2012), a well-established academic trivia competition.

Game flow. A game is played between two teams and consists of 20 cycles. Each cycle begins with a tossup question: any player can buzz in to answer, and a correct answer earns 10 points for the team (an incorrect buzz before the question ends costs

[2]Dataset: https://huggingface.co/datasets/qanta-challenge/qanta25-gamedata; Platform: https://github.com/qanta-org/qb-tournament-runner; Analysis: https://github.com/qanta-org/qanta25-analysis.

-5 points). If a team answers the tossup correctly, they earn a bonus question with three parts, each worth 10 points. The team collaborates to answer each part, with humans making the final decision after seeing AI suggestions. The rest of this section details each phase and the human–AI collaboration.

2.1 Tossup Delegation (Tossups)

Tossups are designed to be *interrupted* by a **buzz**. Each tossup question (a sequence of clues starting hard and getting easier) is read aloud to all players.[3] Once any player—human or AI—is confident enough to answer, they "buzz" in with a response. Humans buzz using a physical buzzer (or its online equivalent) and must answer immediately with no team discussion, following standard quizbowl rules. An AI buzzes in using the interface and similarly immediately vocalizes an answer. However, each team only has one chance to answer a tossup; an incorrect guess results in a penalty and bars the incorrect player's whole team from answering the question and participating in the bonus phase. Thus, a player with low accuracy or poor calibration hinders the whole team. Human teammates can shout or glare at poorly calibrated teammates; for AI teammates, we provide an analogous mechanism: **muting**.

Muting AI teammates. Teams may **mute** an AI teammate at the start of the game or after any tossup–bonus cycle,[4] preventing it from buzzing for the rest of the game. Crucially, humans still see muted AIs' suggestions in the bonus phase—teams may distrust an AI's buzzing judgment while still valuing its knowledge for collaborative decisions. For instance, during one tournament, a question began as follows:

> In one work, a large-headed man with a multi-colored top hat and pink…

One of the AI players buzzed at this point with the guess Willy Wonka[5] with 90% confidence. The correct answer was **Christ**. This egregious mistake led to an immediate request to mute that AI teammate for the rest of the game.

[3]This decrease in difficulty should be true for both types of players (humans and AI); see Section 3.2 for how we got trivia experts to write these questions.

[4]A cycle ends immediately before the next tossup question is read. If neither team answered the tossup correctly, the cycle ends right after the tossup.

[5]We distinguish a guess, a (possibly incorrect) response, from the correct **answer** to the question.

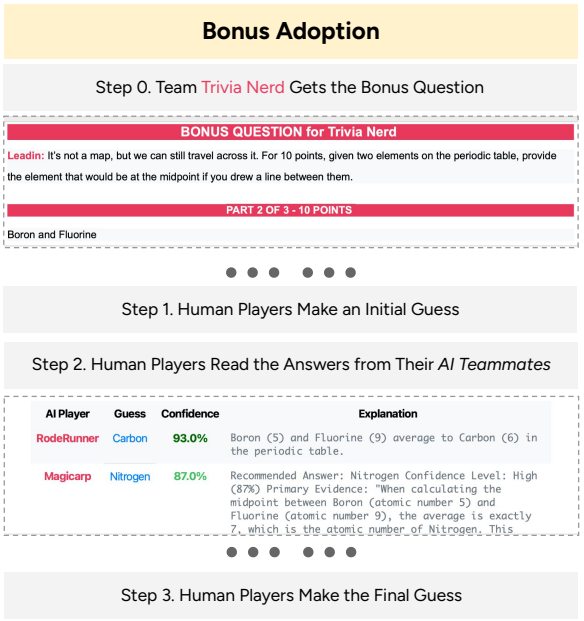

Figure 2: Overview of our collaboration interface showing deliberative decision-making for bonus questions (top). Humans first provide their own guess without any assistance from AI, then see suggestions from two AI teammates with confidence scores and explanations (middle), and finally discuss and decide how to adopt them (bottom).

2.2 Bonus Adoption (Bonuses)

When a team answers a tossup correctly, they earn a "bonus" question with three parts on a common theme (Elgohary et al., 2018) announced with a "lead-in" (e.g., Figure 2, top). To measure the effect of collaboration, we want to measure human ability in the absence of AI assistance *and* how they navigate (often unreliable) AI support. This two-stage design is essential: by recording the human guess *before* revealing AI suggestions, we can directly compare the same team's pre- and post-AI answers, isolating the causal effect of the AI on human decisions.

Initial Human Guess. The moderator reads each part of the question, allows the human players to confer on the answer, and the teams provide a consensus guess without any assistance from their AI teammates. In the illustrated example (Figure 2), the second part of the question asks the team to find the element that is the midpoint of Boron and Fluorine on the periodic table. Human players gave a correct guess of nitrogen. While we record this guess and its correctness for analysis, this does not contribute to the team's score. The human team is not told whether their guess is correct or not.

AI Guesses. After the humans guess, they then see

guesses from two distinct AI systems they drafted to be their teammates, along with a confidence score and textual explanation (Figure 2, middle). These explanations are generated by the AI agents themselves—not designed by the experimenters—creating natural heterogeneity across the 16 systems built by different participants using diverse architectures. Here, System 1 the AI nicknamed RodeRunner[6] uses correct reasoning, but incorrect math, leading to an incorrect guess of carbon with a high confidence score of 93%, while the other system (Magicarp) guesses correctly nitrogen with a clear explanation and a confidence score of 87%.

Final Consensus Guess. The humans on the team now need to give a final answer. They can: retain their initial guess, pick one of the AI guess(es), or make a new guess entirely (after deliberation). This final answer counts for the team's score. In this case, the team kept their nitrogen guess, confirmed by Magicarp, earning 10 points.

2.3 Behavioral Traces Captured

The game design yields three complementary behavioral traces. For **team formation**, we record each team's agent selections and the draft pick order for every round (§ 3.1). For **tossups** (proactive delegation), we record who buzzed (human or AI), the clue position, each AI's mute state, buzz correctness, and points gained or lost. For **bonuses** (deliberative adoption), we record the human team's initial answer, each AI's answer together with its confidence score (0–1) and textual explanation, the team's final answer, and correctness at each stage. Together these traces span the full arc of reliance decisions, from team formation through proactive delegation to deliberative adoption.

3 Human–AI Cooperative Trivia Tournament

Following the game design (Section 2), we ran two tournaments (one in-person, one online) with 23 human players and 16 AI agents. This section outlines tournament structure, question design, and AI agent architectures.

3.1 Tournament Structure and Participants

The tournament ranks human trivia teams on both their knowledge and their ability to form and work with AI teams. Both tournaments—in-person and online—collect behavioral traces of human–AI collaboration.[7] Our participants comprise 23 experienced trivia players whose competitive experience ranges from 1 to 7+ years (mean 3.2), including several with national game show appearances. They formed nine teams across both tournaments.

Team formation. The tournament is organized into games, each featuring a themed question packet (e.g., music, spatial reasoning, cultural references). Before each round, teams select two AI agents as teammates using a serpentine draft (Lee and Liu, 2022): The pick order ascends from the lowest- to the highest-scoring team, which picks twice consecutively before the order reverses back down to the lowest-scoring team. Each AI agent can only be selected once per round, preventing all teams from choosing the perceived best AI and avoiding conflicts when identical agents would face each other.[8] This design gives weaker teams first access to the perceived-best AI teammates, partially offsetting human skill gaps and preventing runaway advantages by stronger teams.

Human players initially know nothing about their potential AI teammates. The models' origins are obscured through opaque nicknames (e.g., "RodeRunner," "Magicarp"). Over the course of early rounds, teams observe which AIs buzzed and whether they were correct; when they win a tossup, they additionally see their own AI teammates' answers, confidence scores, and explanations for that bonus. Teams use this accumulating behavioral evidence to inform drafting choices in subsequent rounds. Teams must decide not only *whether* to trust AI, but *which* AI to work with—a more realistic reflection of real-world AI adoption than studies with single fixed systems.

Game Structure and Schedule. Each game consists of 20 tossup questions and 20 three-part bonus questions. Teams play 24 games across the two tournaments, yielding 140 toss-ups and 420 bonus parts. The format includes a round-robin phase followed by single-elimination playoffs; each game uses a themed packet targeting known AI or human weaknesses (temporal reasoning, cultural references, wordplay, etc.). Appendix C summarizes the dataset scale.

[6] Teams know their AI teammates only by opaque nicknames during drafting (Section 3.1)

[7] We compared online and in-person teams on bonus accuracy, switching rate, and muting behavior and found no significant differences, so we pool them in all analyses.

[8] Draft mechanics varied slightly between prelim phase and playoff rounds to give teams access to preferred AI teammates; see Appendix E for details.

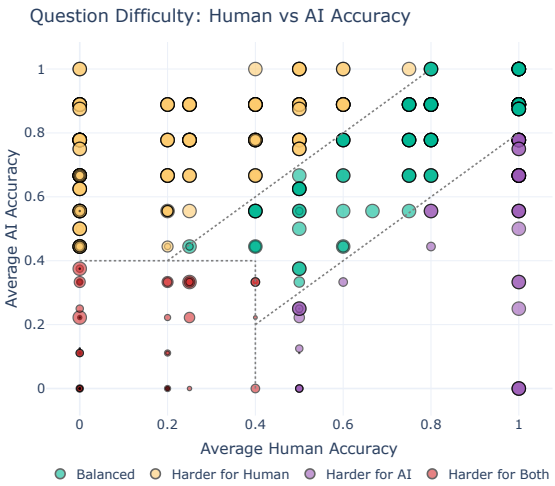

Figure 3: Question difficulty reveals systematic complementarity between humans and AI. Each point is a question; x-axis shows average human accuracy, y-axis AI accuracy. Bubble size indicates team accuracy after AI-assisted deliberation. Collaboration generally improves accuracy, but opportunities remain for questions that challenge both parties.

3.2 Adversarial Question Design

We adopt the adversarial question-writing framework of Sung et al. (2025a), which uses human-in-the-loop authoring (Kiela et al., 2021; Wallace et al., 2019a) to create questions challenging for both humans and AI systems. tossup questions must be difficult *at every clue* while still decreasing in difficulty, since they can be interrupted (Section 2.1); bonus questions must be hard enough that neither party can answer trivially alone, yet designed so each side's strengths compensate for the other's blind spots. Concretely, humans struggle with precise factual recall or cross-domain linking, while AI systems falter on culturally embedded reasoning and indirect references (Gor et al., 2024), making solo success unlikely for either party. Figure 3 shows bonus question difficulty for humans and AIs, highlighting this systematic complementarity. Appendix D lists round themes and examples.

3.3 AI Agent Architectures

The tournaments feature 16 distinct AI agents built through a four-week open competition before tournament play. Architectures range from single-model calls with engineered prompts to multistep pipelines with up to four model consultations for answer generation, verification, and confidence calibration. Base models include GPT-4.1, GPT-4o, Claude 3.5 Sonnet, DeepSeek V3, and Cohere Command-R, often combined within a single agent. This architectural diversity produces heterogeneity in capabilities: agents range from 30 to 80% accuracy on our question set, with markedly different strengths across question types, domains, and confidence calibration, ensuring no single agent dominates and teams lack an obvious drafting strategy (Section 3.1). Full agent specifications appear in Appendix G.

4 How Humans Trust AI Assistance—and Where They Misjudge

We analyze how teams navigate the two forms of reliance our tournaments capture. Section 4.1 examines proactive delegation via muting decisions on tossups: what drives teams to mute their AI, whether muting helps, and how close teams come to optimal muting. Section 4.2 examines deliberative adoption on bonuses: how teams evaluate AI suggestions, what drives switching, and where calibration breaks down.

The trivia setting kept users engaged: they answer questions before the AIs, offered guesses on their own, and strategically muted AI teammates when beneficial. Muting rates range from 30% to 100% depending on round theme, and bonus switching rates from 28% to 86%—both suggesting deliberate, context-sensitive decisions rather than random behavior or blind acceptance. Despite being mostly synergistic, collaboration is not frictionless. Under-reliance (3.7% of help opportunities missed) exceeds over-reliance (1.5%), with confirmation bias and cross-model calibration failure as primary drivers.

Tossup Analysis Approach. We evaluate muting effectiveness via counterfactual estimation: for each muting decision, we estimate what would have happened had the AI remained active (or been muted earlier), comparing actual outcomes against this counterfactual and against an oracle with perfect foresight (Section 4.1).

Bonus Analysis Approach. Two independent judges validated answer correctness with 95%+ agreement; a third expert resolved disagreements. To understand *why* teams made specific reliance decisions, an author experienced in trivia tournaments coded decision rationales from tournament video recordings, noting which artifacts (confidence scores, explanations, model agreement)

teams cited during deliberation. We analyze which explanation features predict adoption decisions using single-feature logistic regressions with significance testing (Section 4.2).

Key comparisons use chi-squared and McNemar's tests with effect sizes reported where appropriate. We distill findings into design principles (DP) for human–AI collaborative systems, highlighted in colored callouts.

4.1 Tossup Delegation: Muting Decisions

Skilled humans often buzzed before AI players: 17.9% of questions were answered by a human before any AI buzz. Humans also showed superior calibration, with only 20.0% of their buzzes being incorrect compared to 29.4% for AIs.

Beyond individual buzzes, the more interesting aspect of human–AI *collaboration* is whether teams *allowed* AI teammates to answer autonomously. As described in Section 2.1, players can "mute" an AI teammate before a question is read, preventing it from buzzing during rest of the round.

What Drives Muting Decisions? Teams mute less when models perform well, with *recent accuracy* being the strongest predictor. Teams also adapt to AI weaknesses, muting up to 30% more on challenging topics like music, spatial reasoning and cross-domain identification, where AIs falter more.

Does Muting Help? We analyzed whether users' mental models of AI allowed them to mute optimally. In a round of twenty questions, there are twenty opportunities to mute. For each opportunity, we estimate the counterfactual effect on total points in the tossup phase if the AI had been muted or not.[9] We define an *oracle policy* as the muting strategy that maximizes net tossup points[10] given perfect knowledge of questions and AI behavior—a ceiling that real teams cannot reach since they lack advance knowledge.

Muting pays off: 8 of 9 teams earned more points per game on average by muting than they would have by leaving their AI companions active throughout. These teams captured 79% of the oracle's maximum possible gain. The exception, team T2, a strong human team, muted too aggressively and missed opportunities to benefit from AI buzzes.

Calibration Gaps. Muting is not just about silencing weak models—it requires *calibrating trust*. Humans learn but remain imperfect. Only 9% of muting decisions are made at the optimal time. Teams tend to mute late: 73% of muting decisions occur later than the oracle policy would recommend. By that point, the AI may have already cost the team points through incorrect buzzes.

In contrast, when teams do mute early (18% of decisions), they do so with high magnitude: 49% earlier in the round than optimal (9.8 questions earlier on average). This asymmetry is notable: **over-reliance is more frequent but smaller in magnitude, while under-reliance is rarer but more severe.** As net effect, the average muting occurs 3.4 questions earlier than optimal (about 15% of the round), suggesting early accuracy drops prompt premature AI abandonment. The rarity of optimal timing highlights how hard it is for humans to calibrate trust in real time, even with direct behavioral feedback. We did note that on topics where AI excels—literature, military history—teams appropriately let AI answer.

DP 1: Give users granular control over AI involvement.

Provide context-dependent toggles (by topic, difficulty) rather than binary on/off controls. User agency over *when* AI participates matters as much as whether to follow its advice (Vaccaro et al., 2020).

4.2 Bonus Answer Adoption

In the bonus phase, teams answer multipart questions with the help of AI: humans provide an initial guess; see AI guesses, confidence scores, and explanations; and then decide the *team's* final guess.

But, first: how do teams decide whether to trust AI answers, and how well do they take AI advice? To aid our analysis, an author familiar with trivia tournaments annotated the confidence of the human participants from recordings and which evidence (if any) the humans used to make their decision.

When revising guesses with AI assistance (Table 4), humans most often follow AI agreement (54.8% of decisions), followed by their own domain knowledge (35.0%). Model explanations (4.4%), confidence scores (2.2%), and model reputation (2.0%; "trust System 1, not System 7, bro") also guide choices, though some decisions appear random (1.5%). Following AI agreement achieves perfect accuracy (100%), and domain knowledge performs well (92.4%), but confidence scores (52.3%) perform barely better than chance,

[9] This counterfactual has limitations: when AI buzzes early and wrong, we cannot know if a human would have answered correctly. However, removing an early wrong response is a net positive. Cases where AI could have been early and right are straightforward to evaluate.

[10] For the rest of this section, we discuss only *net* points that muting changes, ignoring effects on subsequent bonus questions.

Metric	Freq (%)
Accuracy	
Human initial guess correct	42.6
AI correct (random pick of AI)	59.2
Oracle AI selection	77.5
Any human or AI had correct answer	80.5
Final consensus answer correct	**81.4**
Effectiveness	
Humans retain correct answer	98.0
Humans adopt a correct AI answer	94.4
Humans don't know correct answer but discern which correct AI guess to trust	83.3
Failure Modes	
Over-reliance: humans reject their correct answer for wrong AI answer	1.5
Under-reliance: humans are wrong and fail to adopt correct AI answer	3.7

Table 1: How well human and AI teammates work together on collaborative question answering. The accuracy of the final consensus answer is higher than humans or computers alone and better than just picking the best AI (which is a hard task). However, the collaboration is not perfect: human teams sometimes reject their own correct answers or fail to adopt AI answers.

highlighting the need for better cross-model calibration.

DP 2: Standardize confidence across models.

Systems deploying multiple AI agents should invest in cross-model calibration—especially for disagreement cases where users need the most help (Guo et al., 2017).

4.2.1 Adoption Patterns

Humans often adopt AI answers. Humans adopted an AI answer 96.3% of the time they had no initial answer or not confident (50.7% of cases). The AI was correct in 73.4% of these cases. The title for this paper came from Team 9, which was fond of saying "AI, take the wheel", referencing Underwood (2005) as it adopted an AI answer.

Teams learned to use AI more effectively. Both utilization rate (adopting available correct AI answers) and discernment (picking the right AI answer when models disagreed) increased significantly across rounds, particularly on the hardest questions (Figure 4; $\beta = 0.75$, $p < 0.01$). This rules out simple defaulting under uncertainty: if teams were blindly deferring to AI when they had no answer, discernment would remain near 50%. Instead, discernment improved from 27.1% to 87.5%, indicating that teams learned to distinguish good from bad AI responses.

Teams built informal AI reputation through observation alone. Despite receiving no prior performance information, teams' drafting choices correlated with actual AI performance (Table 2).

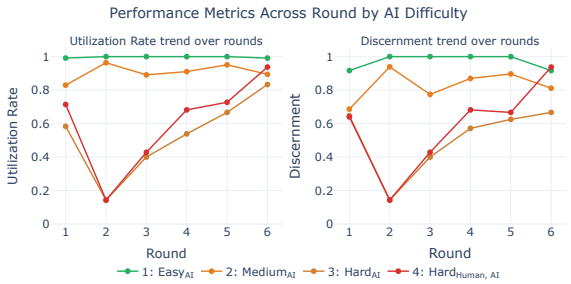

Figure 4: Utilization and discernment rates by round and question difficulty. As the tournament progressed, teams increasingly adopted correct AI answers when available (utilization rate) and became better at selecting the right one when models disagreed (discernment), especially on harder questions. This trend reveals teams learning to use AI help more effectively over time.

Metric	In-person	Online
Tossup Points	+0.81*	+0.60
Bonus Points	+0.74*	+0.43
Combined Points	+0.76*	+0.57
Buzz Accuracy	+0.79*	+0.67†
1st-Buzz Accuracy	+0.67†	+0.71*

*$p < 0.05$; †$p < 0.10$

Table 2: Spearman ρ between draft Elo ratings (from teams' selections) and AI performance metrics ($n = 8$ systems per tournament). Teams drafted roughly in performance order, especially in person, where there was more informal discussion of AI properties.

Humans overruled AI, but rarely. Only 7.7% of final answers differed from both AI responses. This was more likely when humans were highly confident (44.4%), though even then they kept their own answer only 38.1% of the time—often revising using AI-provided content (Figure 6).

Synergy. With an overall accuracy of 81.3%, the overall team (Table 1) is better than the humans alone (who get only slightly more than a third of the answers correct; McNemar's $\chi^2 = 633.48$, $p << 0.001$), better than a randomly chosen AI (a little over half), and better than an oracle pick of the best AI on any given team for that specific question (three quarters of the questions right). This shows that collaboration is mostly working as intended.

DP 3: Surface accumulating collaboration evidence.

Show where AI has succeeded and failed by domain during collaboration, rather than providing only static, pre-deployment proficiency summaries (Yin et al., 2019).

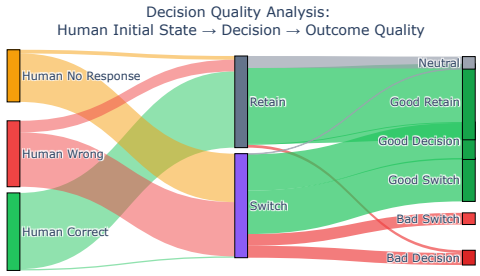

Figure 5: Every bonus decision traced from left to right: whether the human's initial guess was correct (left), whether the team switched to an AI answer (center), and the final outcome (right). The thickest flow—"Human Wrong" through "Switch" to "Good Switch"—represents successful collaboration: teams recognized their error and adopted a correct AI suggestion. The thinner "Human Right" to "Switch" to "Bad Switch" flow represents over-reliance: teams abandoned a correct answer for an incorrect AI suggestion.

4.2.2 Measuring Appropriate Reliance

However, despite those promising results, we are still not at optimal AI usage. Following Schemmer et al. (2023), we examine the error cases of **under-reliance** and **over-reliance**. Under-reliance is humans failing to adopt correct AI advice, while over-reliance as correct humans adopting incorrect AI advice.[11] Both optimal rates are 0%, although it requires both well-calibrated AI systems and discerning human users.[12]

Reliance errors are asymmetric (Figure 5): under-reliance (3.7% of help opportunities missed) exceeds over-reliance (misled 1.5% of times humans were initially correct). Teams are appropriately cautious but sometimes overly conservative.

> **DP 4: Design for mutual coverage, targeting under-reliance.**
>
> Highlight cases where AI confidence is high on domains where the user historically struggles, helping experts recognize when to defer (Kleinberg et al., 2018).

Consensus as a Trust Signal Teams interact with two AIs per game, observing when models agree or disagree. When both AIs give the same correct answer, teams switch 82% of the time—above the 68% average ($\chi^2 = 7.67$, $p < 0.006$). When models disagree, switching drops to 45% even when one is correct. Consensus acts as a strong reliability signal; disagreement signals uncertainty.

Confirmation bias amplifies errors. When an incorrect human answer is confirmed by one AI, under-reliance rises to 60.7%—teams trust the agreeing AI even when wrong ($\chi^2 = Fisher$, $p << 0.001$). Conversely, when AIs agree on an incorrect answer, over-reliance spikes to >10% as teams abandon correct initial guesses.

[11] Formal definitions in Appendix B.1.

[12] And it is not realistic to actually reach 0%, as errors from incorrect / poorly written questions, out of date information, etc. can cause human and computer miscalibration.

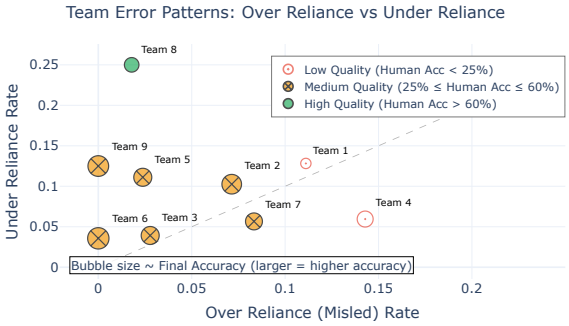

Figure 6: Team skill at question answering (marker color and shape) does not predict skill at using AI (marker size). Middling teams had the best under/over-reliance tradeoff; strong teams like Team 8 under-relied on AI, while weak teams like Team 4 over-relied.

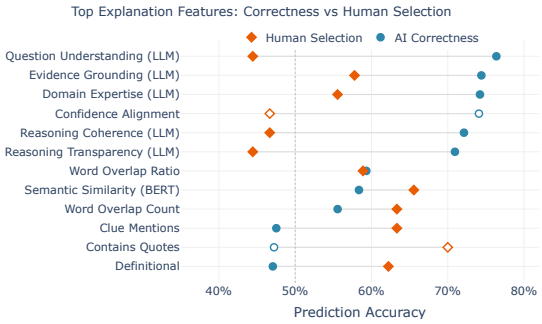

Figure 7: Explanation features differ in what predicts AI correctness versus what humans trust. Each row shows a feature's single-predictor accuracy. Filled markers indicate positive predictors (higher feature = more likely): hollow markers indicate negative. LLM-assessed features (Question Understanding 76%, Reasoning Coherence 72%) predict correctness but humans rely on surface signals (quotes 70%, BERT similarity 66%). Only Evidence Grounding appears in both top-6 lists.

What Makes Humans Trust an Explanation? Beyond consensus and confidence, what features of an explanation lead humans to trust it? For each question, we extract 57 features from each AI explanation, spanning surface properties (length, readability), structural patterns (quotes, clue mentions), and reasoning quality (Appendix F). We then ask: which features predict (1) whether the AI answer is actually correct, and (2) whether humans select that explanation? For each feature, we fit a single-feature logistic regression predicting either AI cor-

rectness or human selection. Features whose coefficient reaches statistical significance ($p < 0.05$) are shown with filled markers in Figure 7; hollow markers indicate features that did not reach significance. Marker position shows the single-predictor accuracy: how well that feature alone distinguishes correct from incorrect AI answers (blue) or selected from unselected explanations (orange). LLM-assessed features like Question Understanding (76%), Reasoning Coherence (72%) strongly predict correctness (Figure 7), while humans rely on surface signals like quotes (70%), semantic similarity (66%), and word overlap (63%). Only Evidence Grounding appears in both top lists. This is demonstrated by a team's comment in Round 4 of the online tournament. On bonus 13, part 3, Magicarp gave the same guess as the players, with 95% confidence. System 1 gave a different guess with 80% confidence. However, players read the explanations and one stated "I trust [System 1's] citation more" because it was grounded in the actual bonus question text, unlike System 7's. The humans switched their guess to System 1's, which turned out to be correct.

This gap suggests a path forward: AI systems should make reasoning explicit through evidence citation, while humans should evaluate whether explanations demonstrate genuine understanding, not just surface familiarity.

DP 5: Anchor explanations in evidence.

Reference observable input features (specific clues, quotes) rather than abstract reasoning. Only evidence grounding predicts both AI correctness and human trust (Vasconcelos et al., 2023).

5 Related Work

Human-AI collaboration is pervasive and touches on many fields from AI to psychology to human–computer interaction. Here, we focus on expertise and confidence (Appendix A details related work).

Expertise. For instance, field studies analyze consequential decisions but lack controlled interaction: observational data from judges or physicians (Kleinberg et al., 2018; Gaube et al., 2021) reveal reliance patterns but without behavioral information (e.g., time spent on explanations, confidence comparisons, deliberation sequences) needed to understand decision mechanisms. In addition, skilled practitioners are a crucial group in this context, but lab studies often do not engage genuine expertise, e.g., Buçinca et al. (2020) uses crowdworkers to evaluate unfamiliar domains, which provides limited insight into how experts integrate AI into practiced workflows. ***Our contribution:*** We bridge these gaps by studying experts in their domain of practice, capturing both proactive delegation (deciding whether to let AI act before seeing its output) and deliberative adoption (deciding whether to follow AI after evaluating its answer, confidence, and explanation). This dual approach reveals how humans weigh evidence, calibrate trust, and integrate AI assistance. Our setting reveals both systematic under-reliance (3.7% missed opportunities) and appropriate resistance (88% rejection of misleading AI), showing sophisticated but imperfect calibration that improves.

Confidence. Previous work has also found that certainty/calibration (Sendak et al., 2020), explainability (Ribeiro et al., 2016), and interpretability (Poursabzi-Sangdeh et al., 2021) affect how much users trust an AI, with both positive and negative results. ***Our contribution:*** Confidence was not a strong predictor of switching decisions, but explanation quality modulates this effect substantially. Specifically, explanations referencing specific evidence from the question increase appropriate switching by 12%.

6 Conclusion and Discussion

Human–AI collaboration is already happening, both in proactive delegation and deliberative adoption settings. To improve outcomes and fully use human and AI resources, we need systems that are better calibrated, offer clear evidence, and give people the information they need to make informed decisions. While our study focused on text-only settings, where AI agents surpassed the human teams, the synergistic collaboration exceeded the sum of its parts. This is more important in multimodal domains where AI accuracy still lags. More important will be not just facilitating collaboration with a fixed set of team members but facilitating team *formation*: deciding which humans and which AI contributors to tap for a given problem. To support future work, we release our dataset, tournament platform code, and analysis scripts (Appendix H).

7 Limitations

Several limitations warrant discussion:

Domain specificity. While our cooperative trivia tournament provides excellent experimental control and genuine expertise, generalizing findings to other domains requires caution. The competitive, knowledge-intensive nature of quiz bowl may not reflect collaborative contexts like medical diagnosis or legal review where different decision pressures apply. Future work should validate whether under-reliance patterns, confidence sensitivity, and explanation effects generalize to other expert domains.

Sample size and statistical power. With 23 human players and 16 AI agents across 24 games, our study provides reasonable statistical power for main effects but limited ability to detect nuanced individual differences or rare interaction patterns. Larger studies could reveal additional player archetypes, more precise learning trajectories, or context-specific reliance strategies.

Causality and interventions. Our observational design reveals correlational patterns between features (confidence, explanations) and reliance decisions, but cannot establish causality. Our annotator observes live video recordings of play and notes which artifacts (confidence scores, explanations, model agreement) teams use to arrive at decisions, but we cannot rule out confounding: high-confidence AI may also generate better explanations or answer easier questions. Future work should use randomized interventions—varying confidence levels or explanation quality experimentally—to establish causal effects.

Muting interface effects. The muting mechanism required explicit action before each toss-up, potentially introducing friction that inflates muting rates. Alternative designs (default-on with quick toggle, voice commands) might yield different strategic patterns. However, the systematic learning and context-dependence we observe suggests humans engaged strategically rather than randomly.

Question design and ecological validity. Our adversarial questions intentionally exploit known AI weaknesses, creating systematic skill gaps. While this validates complementarity opportunities, real-world collaborative contexts may feature less predictable AI failure modes. However, the design principles we derive—prioritizing calibrated confidence, grounded explanations, and user control—should generalize beyond adversarial settings.

Temporal scope. We observe learning effects across four tournament rounds (approximately 2 hours per session), but cannot assess long-term trust calibration. Extended collaboration might reveal different patterns: humans could become overconfident in AI after positive experiences, or develop more sophisticated mental models enabling better calibration. Longitudinal studies tracking trust evolution over weeks or months would complement our findings.

8 Ethics Statement

The experiments performed in this study involved human participants. All the experiments involving human evaluation in this paper were exempt under institutional IRB review.

All human data collection procedures were reviewed and approved by an institutional review board (IRB) to ensure the protection of participants' privacy and rights. Human buzzpoints in the dataset are fully anonymized. Although the post-competition survey collected participant names for compensation purposes, only aggregate statistics and anonymized quotes are reported in the study.

Trivia players collectively received $600 USD in online gift cards as prizes for the competitions, with awards of $150, $100, and $50 for the top three teams in both offline and online tournaments. Model submitters received a total of $400 USD in online gift cards, distributed as $200, $150, $100, and $50 prizes. Question writers were compensated $5 per question, and editors $1 per edited question, corresponding to an estimated rate exceeding $10 USD per hour—above the U.S. federal minimum wage of $7.50 USD. All the involved participants gave their consent to disclose their interactions with the interface. The documents used in the study are distributed under an open license.

IRB approval and informed consent. This study was approved by our institutional review board (IRB) to ensure participant privacy and rights. All participants gave informed consent after being clearly told the study involved human-AI collaboration and behavioral data collection. Participation was voluntary and could be withdrawn at any time without penalty. The IRB monitored procedures for collecting responses and questions to protect privacy throughout the study.

Participant compensation and treatment. Human players participated voluntarily, motivated by

competitive interest in quiz bowl rather than monetary compensation (consistent with standard quiz bowl tournament practice). Question writers received fair compensation ($25/hour, above typical rates for quiz bowl writing). All participants were treated with respect throughout the study.

Data privacy and transparency. We collected only gameplay data (answers, timing, muting decisions, switching decisions) and basic demographic information (experience level). No personally identifiable information beyond pseudonymous player IDs appears in our dataset or analysis. Participants were informed about AI system characteristics (model types, skill levels) but not specific implementation details that might affect strategic behavior.

AI attribution and disclosure. Participants knew they were collaborating with AI agents and understood when AI suggestions appeared. We clearly disclosed confidence scores and explanations as AI-generated rather than human expert opinions. No deception was used; all AI interactions were transparent.

Community engagement and data release. We engaged with the quizbowl community throughout this research, explaining our goals and design choices to tournament organizers and participants. We release the full dataset (questions, responses, behavioral traces), the tournament web application source code, and all analysis scripts to support reproducibility. See Appendix H for platform details.

Broader impacts. This research advances understanding of human-AI collaboration with potential benefits for interface design in high-stakes domains (medical diagnosis, legal review). However, our findings about systematic under-reliance could be misused to encourage blind trust in AI systems. We emphasize that appropriate reliance requires well-calibrated AI confidence and grounded explanations—trust should be conditional on AI system quality, not automatic.

Use of AI assistants. The authors used AI tools (OpenAI's ChatGPT and Anthropic's Claude) for coding assistance during data analysis and visualization, and as a writing assistant limited to paraphrasing for conciseness. All substantive content, analysis, and conclusions are the authors' own work.

Acknowledgements

We thank the UMD CLIP group for valuable feedback on study design and adversarial evaluation. We are also grateful to the participants who developed and submitted AI systems for the live trivia tournament, the participants who played in the tournament, and all who made the experiment possible.

This work was supported in part by the National Science Foundation (IIS-2403436) and the NSF Institute for Trustworthy AI in Law & Society (TRAILS, 2229885). Any opinions, findings, conclusions, or recommendations expressed are those of the authors and do not necessarily reflect the views of the sponsors.

References

Gagan Bansal, Besmira Nushi, Ece Kamar, Daniel S Weld, Walter S Lasecki, and Eric Horvitz. 2019. Beyond accuracy: The role of mental models in human-ai team performance. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 7, pages 2–11.

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*, pages 1–16.

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*, AIES '21, pages 401–413. ACM.

Jordan Boyd-Graber, Brianna Satinoff, He He, and Hal Daumé III. 2012. Besting the quiz master: Crowdsourcing incremental classification games. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1290–1301, Jeju Island, Korea. Association for Computational Linguistics.

Zana Buçinca, Phoebe Lin, Krzysztof Z. Gajos, and Elena L. Glassman. 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable ai systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, IUI '20, pages 454–464. ACM.

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):1–21.

Carrie J. Cai, Jonas Jongejan, and Jess Holbrook. 2019. The effects of example-based explanations in a machine learning interface. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, IUI '19, pages 258–262. ACM.

Michelene T. H. Chi. 2006. *Laboratory Methods for Assessing Experts' and Novices' Knowledge*, pages 167–184. Cambridge University Press.

Upol Ehsan, Q. Vera Liao, Michael Muller, Mark O. Riedl, and Justin D. Weisz. 2021. Expanding explainability: Towards social transparency in ai systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, pages 1–19. ACM.

Ahmed Elgohary, Chen Zhao, and Jordan Boyd-Graber. 2018. Dataset and baselines for sequential open-domain question answering. In *Empirical Methods in Natural Language Processing*.

Stephen M. Fleming and Hakwan C. Lau. 2014. How to measure metacognition. *Frontiers in Human Neuroscience*, 8.

Susanne Gaube, Harini Suresh, Martina Raue, Alexander Merritt, Seth J. Berkowitz, Eva Lermer, Joseph F. Coughlin, John V. Guttag, Errol Colak, and Marzyeh Ghassemi. 2021. Do as ai say: susceptibility in deployment of clinical decision-aids. *npj Digital Medicine*, 4(1).

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

Kate Goddard, Abdul Roudsari, and Jeremy C Wyatt. 2012. Automation bias: a systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association*, 19(1):121–127.

Catalina Gomez, Sue Min Cho, Shichang Ke, Chien-Ming Huang, and Mathias Unberath. 2025. Human-AI collaboration is not very collaborative yet: A taxonomy of interaction patterns in AI-assisted decision making from a systematic review. *Frontiers in Computer Science*, 6:1521066.

Maharshi Gor, Hal Daumé Iii, Tianyi Zhou, and Jordan Lee Boyd-Graber. 2024. Do great minds think alike? investigating human-AI complementarity in question answering with CAIMIRA. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21533–21564, Miami, Florida, USA. Association for Computational Linguistics.

Ben Green and Yiling Chen. 2019. Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 90–99. ACM.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR.

Sophie Haroutunian-Gordon, Hubert L. Dreyfus, and Stuart E. Dreyfus. 1988. Mind over machine: A plea for the intuitive conception of mind. *Educational Researcher*, 17(3):50.

Patrick Hemmer, Max Schemmer, Niklas Kühl, Michael Vössing, and Gerhard Satzger. 2025. Complementarity in human-ai collaboration: concept, sources, and evidence. *European Journal of Information Systems*, pages 1–24.

Kevin Anthony Hoff and Masooda Bashir. 2014. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 57(3):407–434.

Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. 2023. Co-writing with opinionated language models affects users' views. In *Proceedings of the 2023 CHI conference on human factors in computing systems*, pages 1–15.

Ken Jennings. 2006. *Brainiac: adventures in the curious, competitive, compulsive world of trivia buffs*. Villard.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the Association for Computational Linguistics*.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. Dynabench: Rethinking benchmarking in NLP. In *Conference of the North American Chapter of the Association for Computational Linguistics*.

Gary A. Klein. 2017. *Sources of Power: How People Make Decisions*. The MIT Press.

Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. Human decisions and machine predictions. *The quarterly journal of economics*, 133(1):237–293.

Jonna Koivisto and Juho Hamari. 2019. The rise of motivational information systems: A review of gamification research. *International Journal of Information Management*, 45:191–210.

Vivian Lai and Chenhao Tan. 2019. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 29–38. ACM.

Himabindu Lakkaraju and Osbert Bastani. 2020. "how do i fool you?": Manipulating user trust via misleading black box explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, pages 79–85. ACM.

John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human factors*, 46(1):50–80.

Michael D. Lee and Siqi Liu. 2022. Drafting strategies in fantasy football: A study of competitive sequential human decision making. *Judgment and Decision Making*, 17(4):691–719.

A. Leonard, J. Fagan, T. O'Sullivan, D. O'Connor, and P. O'Sullivan. 2024. Use of artificial intelligence in triage in hospital emergency departments: A scoping review. *Cureus*, 16:e248163.

Maria Madsen and Shirley D Gregor. 2000. Measuring human-computer trust. In *Proceedings of the 11th Australasian Conference on Information Systems*, volume 6, pages 1–6.

Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D Manning, and Daniel E Ho. 2024. Ai on trial: Legal models hallucinate in 1 out of 6 (or more) benchmarking queries'.

Carsten Maple, Alpay Sabuncuoglu, Lukasz Szpruch, Andrew Elliott, and Tony Zemaitis Gesine Reinert. 2024. The impact of large language models in finance: Towards trustworthy adoption. *The Alan Turing Institute*.

Raja Parasuraman and Victor Riley. 1997. Humans and automation: Use, misuse, disuse, abuse. *Human factors*, 39(2):230–253.

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, pages 1–52. ACM.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 1135–1144. ACM.

Pedro Rodriguez, Shi Feng, Mohit Iyyer, He He, and Jordan Boyd-Graber. 2019. Quizbowl: The case for incremental question answering.

Max Schemmer, Niklas Kühl, Carina Benz, Andrea Bartos, and Gerhard Satzger. 2023. Appropriate reliance on ai advice: Conceptualization and the effect of explanations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, IUI '23, pages 1–21. ACM.

Mark P. Sendak, Michael Gao, Nathan Brajer, and Suresh Balu. 2020. Presenting machine learning model information to clinical end users with model facts labels. *npj Digital Medicine*, 3(1).

Keith E. Stanovich and Richard F. West. 2002. *Individual Differences in Reasoning: Implications for the Rationality Debate?*, pages 421–440. Cambridge University Press.

Mark Steyvers, Heliodoro Tejeda, Gavin Kerrigan, and Padhraic Smyth. 2022. Bayesian modeling of human–ai complementarity. *Proceedings of the National Academy of Sciences*, 119(11):e2111547119.

Yoo Yeon Sung, Eve Fleisig, Yu Hou, Ishan Upadhyay, and Jordan Lee Boyd-Graber. 2025a. GRACE: A granular benchmark for evaluating model calibration against human calibration. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 19586–19587, Vienna, Austria. Association for Computational Linguistics.

Yoo Yeon Sung, Maharshi Gor, Eve Fleisig, Ishani Mondal, and Jordan Lee Boyd-Graber. 2025b. Is your benchmark truly adversarial? AdvScore: Evaluating human-grounded adversarialness. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 623–642, Albuquerque, New Mexico. Association for Computational Linguistics.

Matthias Sutter. 2023. *Kahneman, Daniel: Thinking, Fast and Slow*, pages 1–2. J.B. Metzler.

Philipp Tschandl, Christoph Rinner, Zoe Apalla, Giuseppe Argenziano, Noel Codella, Allan Halpern, Monika Janda, Aimilios Lallas, Caterina Longo, Josep Malvehy, John Paoli, Susana Puig, Cliff

Rosendahl, H. Peter Soyer, Iris Zalaudek, and Harald Kittler. 2020. Human–computer collaboration for skin cancer recognition. *Nature Medicine*, 26(8):1229–1234.

Carrie Underwood. 2005. *Jesus, Take the Wheel*. Arista Nashville.

Kristen Vaccaro, Christian Sandvig, and Karrie Karahalios. 2020. "at the end of the day facebook does what itwants": How users experience contesting algorithmic content moderation. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–22.

Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael S. Bernstein, and Ranjay Krishna. 2023. Explanations can reduce overreliance on ai systems during decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1):1–38.

Eric Wallace, Shi Feng, and Jordan Boyd-Graber. 2019a. Misleading failures of partial-input baselines. In *Proceedings of the Association for Computational Linguistics*.

Eric Wallace, Pedro Rodriguez, Shi Feng, Ikuya Yamada, and Jordan Boyd-Graber. 2019b. Trick me if you can: Human-in-the-loop generation of adversarial examples for question answering. *Transactions of the Association for Computational Linguistics*, 7:387–401.

Bryan Wilder, Eric Horvitz, and Ece Kamar. 2020. Learning to complement humans. *arXiv preprint arXiv:2005.00582*.

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, pages 1–12. ACM.

Yunfeng Zhang, Q Vera Liao, and Rachel KE 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*, pages 295–305.

A Additional Related Work

This section outlines additional related work not featured in the main text of the paper.

A.1 Human-AI Collaboration and Appropriate Reliance

Understanding appropriate reliance in automation has long been central to human factors research (Parasuraman and Riley, 1997; Lee and See, 2004). Parasuraman and Riley's foundational taxonomy distinguishes misuse (over-reliance on unreliable automation) from disuse (under-reliance on reliable automation), with trust calibration—matching reliance to actual capability—identified as crucial for effective human-automation teams (Lee and See, 2004). Early work focused on measuring trust dimensions (Madsen and Gregor, 2000; Hoff and Bashir, 2014), establishing that trust must align with system capability for optimal collaboration.

Recent work demonstrates that mental model accuracy—not just AI accuracy—predicts team performance. Bansal et al. (2019) show that when humans develop accurate understanding of AI strengths and weaknesses, team performance improves even when AI accuracy remains constant. However, explanations can paradoxically reduce complementarity by inducing over-reliance: teams exposed to AI explanations sometimes defer blindly rather than critically evaluating suggestions (Bansal et al., 2021). Other work formalizes the learning-to-complement problem, developing algorithms that optimize for team performance rather than individual AI accuracy (Wilder et al., 2020). Steyvers et al. provide a Bayesian framework for human-AI complementarity, showing how optimal aggregation depends on confidence calibration and skill correlation (Steyvers et al., 2022). Zhang et al. find that confidence scores and explanations have complex interactions: high confidence can increase trust, but explanations sometimes fail to improve calibration (Zhang et al., 2020).

Field studies reveal persistent reliance problems across domains. Judges systematically ignore helpful AI bail recommendations, suggesting algorithmic aversion or mistrust (Kleinberg et al., 2018). In contrast, physicians over-rely on AI in medical imaging even when initially correct, changing correct diagnoses to match AI suggestions (Gaube et al., 2021). A systematic review identifies under-reliance as more common than over-reliance across human-AI systems, though context matters substantially (Hemmer et al., 2025). Green and Chen (2019) show that users adapt AI use strategically across contexts, suggesting reliance patterns are neither fixed nor random but responsive to perceived task demands.

Despite this progress, three methodological gaps limit our understanding of reliance mechanisms. *First*, most controlled studies measure post-hoc acceptance rates after humans see AI suggestions (Bansal et al., 2021; Zhang et al., 2020), capturing whether people accept advice but not *how* they decide—the real-time evaluation process, information weighting, or decision criteria remain opaque.

Second, field studies analyze consequential decisions but lack controlled interaction: observational data from judges or physicians (Kleinberg et al., 2018; Gaube et al., 2021) reveal reliance patterns without the behavioral traces (e.g., time spent on explanations, confidence comparisons, deliberation sequences) needed to understand decision mechanisms. *Third*, lab studies achieving experimental control often use synthetic tasks that fail to engage genuine expertise (Buçinca et al., 2020)— crowdworkers evaluating unfamiliar domains provide limited insight into how experts integrate AI into practiced workflows.

Our contribution: We bridge these gaps by studying experts in their domain of practice, capturing both proactive strategic decisions (muting before seeing AI output) and deliberative evidence-based decisions (switching after evaluating AI's answer, confidence, and explanation). This dual-signal approach reveals the reliance process itself: how humans weigh evidence, calibrate trust, and integrate AI assistance. Our setting reveals both systematic under-reliance (3.7% missed opportunities) and appropriate resistance (88% rejection of misleading AI), showing sophisticated but imperfect calibration that improves.

A.2 Trust Calibration: Confidence and Explanations

Appropriate reliance requires well-calibrated AI confidence. Modern neural networks are poorly calibrated: high-confidence predictions often prove incorrect, while low-confidence predictions can be accurate (Guo et al., 2017). Structured presentation of uncertainty—communicating both point estimates and confidence intervals—helps users make better decisions than point estimates alone (Bhatt et al., 2021). In medical domains, confidence-aware AI assistance improves diagnostic accuracy compared to AI without confidence scores (Tschandl et al., 2020). Model facts labels that present calibration information to clinical users improve trust calibration (Sendak et al., 2020), and uncertainty quantification can reduce overreliance when implemented carefully ().

Explanations have complex, sometimes counterintuitive effects on reliance. While explainability methods like LIME (Ribeiro et al., 2016) aim to increase transparency, explanations can manipulate trust through plausible but incorrect rationales (Lakkaraju and Bastani, 2020). Poursabzi-Sangdeh et al. find that greater interpretability can increase overconfidence rather than improving calibration (Poursabzi-Sangdeh et al., 2021). More promisingly, cognitive forcing functions—interventions requiring engagement with explanations before making decisions—reduce overreliance by prompting critical evaluation (Buçinca et al., 2021). Vasconcelos et al. demonstrate that explanations help when they reveal AI limitations: showing where and why AI might fail improves reliance decisions more than generic explanations of reasoning (Vasconcelos et al., 2023). The type of explanation matters: example-based explanations affect trust differently than feature importance (Cai et al., 2019), and social transparency (explaining system context and design choices) complements technical explanations (Ehsan et al., 2021).

Our contribution: We jointly analyze confidence and explanations in naturalistic decisions, finding that grounded explanations—those referencing specific evidence from the question—increase appropriate switching by 12%, providing actionable design guidance beyond generic explanation requirements. We show that confidence was not a strong predictor of switching decisions, but explanation quality modulates this effect substantially. Moreover, two AI companions are not always better than one, and the consensus among AIs is not always helpful.

A.3 Adversarial Evaluation and Question Answering

Standard QA datasets like SQuAD (Rajpurkar et al., 2016, 2018) measure AI progress but often saturate as models exploit superficial patterns. Adversarial evaluation exposes systematic weaknesses: adding distracting sentences with keyword overlap breaks reading comprehension models (Jia and Liang, 2017), and human-in-the-loop adversarial generation creates natural-seeming questions that fool state-of-the-art systems (Wallace et al., 2019b). Contrast sets—minimal edits changing correct answers—reveal that models rely on spurious correlations rather than robust reasoning (**?**). Shortcut learning explains these failures: neural networks exploit dataset artifacts rather than learning intended capabilities (Geirhos et al., 2020).

Quiz bowl provides a compelling testbed for incremental question answering. Boyd-Graber et al. (2012) introduced the QANTA dataset with pyramidal questions where clues progress from obscure to obvious. Rodriguez et al. (2019) formalize buzz timing as a confidence signal: systems must decide

when to answer based on expected utility, balancing accuracy against competitive risk. Recent work systematically generates context-aware adversarial examples that exploit model weaknesses while maintaining naturalness (Sung et al., 2025b).

Our contribution: While adversarial NLP typically focuses on AI failure, we use adversarial design to create complementarity opportunities—questions where human-AI skill gaps enable collaboration gains. This reframes adversarial evaluation from exposing weaknesses to engineering productive partnerships: AI-favoring questions create under-reliance reduction opportunities, while human-favoring questions test appropriate skepticism.

A.4 Strategic Behavior and Expertise

Cognitive science distinguishes expert intuition, which relies on holistic pattern recognition developed through extensive experience (Klein, 2017; Chi, 2006), from novice deliberation following explicit rules (Haroutunian-Gordon et al., 1988). Kahneman's dual process framework contrasts fast, automatic System 1 thinking with slow, deliberative System 2 reasoning (Sutter, 2023). These distinctions raise questions about expert-AI integration: do experts rely on intuition when evaluating AI, or does AI assistance shift them toward deliberative reasoning?

Users adapt AI use strategically across contexts rather than maintaining fixed reliance patterns (Green and Chen, 2019). Users slowly learn to calibrate trust as they observe AI performance across repeated interactions (Yin et al., 2019). However, subjective measures like self-reported trust can mislead: objective behavioral outcomes matter more than stated preferences (Buçinca et al., 2020). Perceived control moderates AI acceptance—users value the ability to override AI even when they rarely exercise it (Vaccaro et al., 2020).

Individual differences moderate reliance quality: cognitive ability predicts susceptibility to biases (Stanovich and West, 2002), and metacognitive accuracy—knowing what you know—varies substantially across individuals (Fleming and Lau, 2014). Expertise moderates how users integrate AI assistance: domain experts process explanations differently than novices, sometimes exhibiting appropriate skepticism that novices lack (Lai and Tan, 2019).

Our contribution: We study expert humans in a competitive domain with genuine expertise requirements, capturing strategic reliance (costly muting) and deliberative reliance (explanation-mediated choice). This reveals systematic patterns: experts under-rely (missing 3.7% of opportunities) but appropriately resist misleading AI (88%), showing sophisticated but imperfect calibration. Trust improves across rounds (+10% from Round 1 to Round 7), demonstrating learnable calibration, and high-skill players show better calibration than low-skill players (gap=8% vs 22%), consistent with expertise moderating AI integration.

Principle	Guideline
Calibrated confidence	Invest in standardized confidence scales; within-model calibration works, but cross-model comparison fails.
Grounded explanations	Anchor in observable input ("The clue mentions X...") rather than abstract reasoning; use calibrated scores over hedges.
Strategic control	Provide context-dependent toggles (by topic, difficulty) rather than binary on/off; user agency over *when* AI participates matters.
Collaboration feedback	Surface context-specific analytics ("AI helped on 8/10 science questions") to accelerate trust calibration.
Expert adoption	Highlight AI strengths on question types where users struggle; under-reliance (3.7%) exceeds over-reliance (1.5%).

Table 3: Design principles for human–AI collaboration systems derived from our behavioral findings. Each guideline addresses a specific failure mode observed in our tournaments.

B Bonus Phase Analysis: Extended Details

This appendix provides additional details for the deliberative adoption analysis in Section 4.2.

B.1 Formal Notation

We formalize the bonus phase interaction as follows. After the human team confers and provides an initial response, the final human answer $h_c \in \{0, 1\}$ indicates correctness.[13]

The team then sees responses from two AI agents. We define:

[13] Multiple team members deliberate, so an individual member may have the correct answer without it being given as the final team response—this is still recorded as $h_c = 0$.

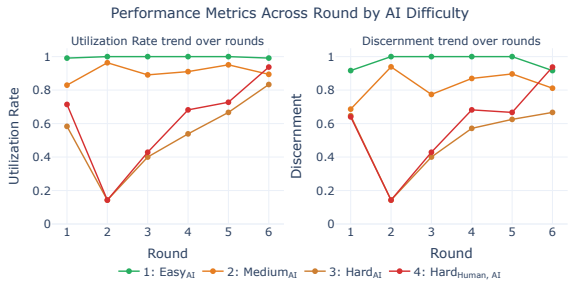

Figure 8: Efficiency trends by round number. Teams colored by AI difficulty (1: Easy, 2: Medium, 3: Hard) and relative difficulty (Both). Utilization Rate measures adoption of correct AI answers; Discernment measures finding the correct answer when neither human nor single AI was individually correct. Round themes shown on x-axis reveal topic effects alongside temporal learning.

- $a_{gc} \in \{0, 0.5, 1\}$: AI guess correctness (0 if both wrong, 0.5 if one correct, 1 if both correct)
- $a_{rc} \in \{0, 0.5, 1\}$: AI recall correctness, measuring whether the correct answer appears in either model's full output including explanations
- $s \in \{0, 1\}$: whether the team switched from their initial response
- $f_c \in \{0, 1\}$: final team correctness after deliberation

Following Schemmer et al. (2023), we define reliance as depending on whether the AI was correct and whether humans adopt it:

$$\text{Under-reliance} = P(s = 0 \mid h_c = 0, a_{rc} > 0) \quad (1)$$

$$\text{Over-reliance} = P(s = 1 \mid h_c = 1, a_{gc} < 1) \quad (2)$$

Under-reliance captures missed help opportunities; over-reliance captures being misled by incorrect AI suggestions. Both optimal rates are 0%.

B.2 Efficiency Trends by Round

Placeholder Analysis. This figure shows how teams calibrated their reliance over the tournament.

B.3 Selection Method Breakdown

When humans revise their response using AI assistance, how do they choose which guess to trust? Table 4 breaks down selection methods and their effectiveness.

Selection Method	Proportion	Final Accuracy
AI Agreement	54.8%	100.0%
Domain Knowledge	35.0%	92.4%
Model Explanation	4.4%	85.7%
Model Confidence	2.2%	52.3%
Model Reputation	2.0%	69.2%
Random Selection	1.5%	31.2%

Table 4: Selection methods used by humans when revising their responses with AI assistance. The table shows the final accuracy after human selection and the proportion of cases using each method. Domain knowledge (identifying correct guesses from their own expertise) is the most common selection criterion.

Key Findings. Domain knowledge dominates selection decisions, accounting for 75% of choices with highest accuracy (81%). This suggests humans are most effective when they can leverage their expertise to evaluate AI suggestions rather than relying on surface signals. Strong rationales (12%) and higher confidence scores (4%) also guide choices, though confidence alone is a weak predictor. Notably, 7% of selections appeared random—teams unable to identify a principled basis for choice.

B.4 Decision Flow by AI Agreement

Figure 5 in Section 4.2 shows the aggregate decision flow. Here we present the full breakdown conditioned on whether the two AI agents agreed (Figure 9).

Key Observations. When AIs reach consensus, teams benefit from a clear signal: switching to the agreed answer yields high accuracy, and under-reliance drops substantially. When AIs disagree, teams must evaluate competing suggestions, leading to higher cognitive load and more conservative behavior (increased inaccuracies). This pattern highlights the value of AI-AI agreement as a trust signal, though it can also amplify errors when both AIs are wrong (Section 4.2).

B.5 Reliance by AI Agreement Condition

Figure 10 breaks down over- and under-reliance rates across four conditions defined by whether one of the AI teammates agreed with the initial human response. When an incorrect human answer is confirmed by one AI while the other provides the correct answer, under-reliance spikes to 60.7% (§4.2).

Component	Count	Description
Tournaments	2	Offline (June 14) + Online (June 21)
Games	24	20 tossups + 20 bonuses each
Human players	23	Experienced quizbowl competitors
AI agents	16	Varying skill levels and architectures
Unique questions	140 each	140 tossups, 420 bonuses
Tossup responses	387	With muting state, buzz timing
Bonus responses	1440	With AI suggestions per part
Muting decisions	∼150	Strategic AI disabling
Switching decisions	∼450	Changed answer after AI input

Table 5: Dataset overview showing the scale and richness of behavioral data collected during human-AI collaborative quizbowl tournaments. The dataset captures multiple types of reliance decisions with real stakes.

B.6 Figures Summary

The following figures support the bonus phase analysis:

1. **Figure 5** (Main paper): Sankey diagram showing aggregate decision flow from initial human state through switching to final outcome.
2. **Figure 9** (Appendix): Full decision flow breakdown by AI agreement status.
3. **Figure 6** (Main paper): Team-level error patterns showing over/under-reliance by team skill.
4. **Figure 7** (Main paper): Calibration gap between correctness predictors and human trust signals.
5. **Figure 8** (Appendix): Round-by-round efficiency trends.

B.7 Placeholder: Qualitative Examples

C Dataset Overview

This appendix provides an overview of the dataset collected in the tournament.

D Our Adversarial Questions

We categorize each bonus question by difficulty for humans and AI separately, based on average accuracy across all responses. Questions with accuracy below 40% are labeled *Hard*, 40–70% as *Medium*, and above 70% as *Easy*. We then compute *relative difficulty* by comparing human and AI accuracy: questions where humans outperform AI by more than 20 percentage points are *Harder for AI*, questions where AI outperforms humans by the same margin are *Harder for Human*, questions where both struggle (below 40% accuracy) are *Hard for Both*, and the remainder are *Balanced*.

Figure 11 shows the distribution of relative difficulty. The largest category (45%) comprises questions harder for humans, where AI outperforms. About 26% fall into the *Balanced* category where humans and AI perform comparably. Questions hard for both (18%) represent challenging items that neither humans nor AI answer reliably. Finally, 11% are harder for AI, where humans have a clear advantage. The right panel shows this breakdown by packet, revealing variation across question sets.

Figure 12 compares absolute difficulty for humans and AI across packets. Human difficulty shows more variation, with some packets containing a higher proportion of hard questions. AI difficulty is more uniform across packets, suggesting that question design affects humans and AI differently. This variation creates diverse collaboration scenarios within each tournament.

E Draft Mechanics

This appendix details the team formation procedure used in our tournaments. The draft mechanism varied slightly between the in-person and online tournaments to accommodate logistical differences.

E.1 In-Person Tournament

The in-person tournament used a full serpentine ("snake") draft before each round. With fewer teams than available slots for AI agents (two per team), each team selected two AI teammates without conflict. The serpentine ordering ensured competitive balance: the lowest-scoring team picked first, selection proceeded upward to the highest-scoring team (who picked twice consecutively), then reversed back down. This gave weaker teams first access to perceived stronger AI agents.

E.2 Online Tournament

The online tournament had more team–game slots than distinct AI systems. To prevent the same AI agent from facing itself in a match, we modified the draft to operate per-game rather than per-round. Before each game, the two competing teams drafted

their AI teammates from the available pool, with agents selected by one team becoming unavailable to the opponent for that game. This ensured no AI agent appeared on both sides of the same match.

E.3 Playoff and Finals Modifications

During playoff rounds in both tournaments, we adopted the same per-game draft style as the online tournament. This allowed teams participating in finals and playoffs to have their best shot at selecting optimal AI teammates while adhering to the constraint that no two AI systems could face each other in the same game.

E.4 Implications for Analysis

These draft variations do not affect our main analyses, which focus on within-game reliance decisions (muting, switching) rather than cross-game selection patterns. However, the draft mechanism itself provides an team formation signal: teams' choices reveal their beliefs about AI capabilities based on observed performance in earlier rounds. We leave detailed analysis of draft strategy to future work.

F Explanation Feature Analysis

This appendix details the feature extraction and analysis for AI explanations presented in Section 4.2. We extract 57 features from each explanation (49 statistical + 8 LLM-assessed) and evaluate their predictive power for two outcomes: (1) whether the AI answer is correct, and (2) whether humans select that explanation.

F.1 Experimental Setup

For each feature, we fit a single-feature logistic regression and measure prediction accuracy. This isolates each feature's individual contribution, avoiding confounds from correlated predictors.

AI Correctness Prediction. We use all bonus responses where both AI systems provided answers. The outcome is binary: whether the AI's answer matches the gold answer.

Human Selection Prediction. We analyze cases where humans chose between two AI explanations. The outcome is which explanation was selected.

F.2 Feature Categories

We organize features into seven categories with consistent prefixes for filtering and analysis.

LLM-Assessed Features. These eight features are rated by GPT-4o on a 1–5 scale, normalized to 0–1. Among these, `question_comprehension` is the top correctness predictor at 76%, while `evidential_grounding` is the only feature appearing in top-10 for both correctness and human selection. The `overconfidence` feature serves as a negative predictor of correctness.

Feature	Description
`question_comprehension`	Understanding of what question asks
`evidential_grounding`	Cites specific clues as evidence
`reasoning_coherence`	Logical flow from clues to answer
`domain_expertise_display`	Domain-specific knowledge shown
`reasoning_transparency`	Reasoning process explicit/followable
`reasoning_depth`	Multi-step vs. surface reasoning
`answer_explanation_alignment`	Explanation supports stated answer
`overconfidence`	Unwarranted certainty given evidence

Table 6: LLM-assessed features (`llm_*`).

Epistemic Features. These six features capture uncertainty expression. The `confidence_alignment` feature—measuring match between linguistic and numeric confidence—is a strong correctness predictor.

Feature	Description
`confidence_alignment`	Match: linguistic vs. numeric confidence
`hedge_ratio`	Ratio of hedging words (might, could, perhaps)
`certainty_ratio`	Ratio of certainty markers (definitely, clearly)
`uncertainty_ratio`	Ratio of uncertainty markers (unsure, unclear)
`linguistic_confidence`	Net confidence from language
`modal_verb_count`	Count of modal verbs

Table 7: Epistemic features (`epist_*`).

Content Grounding Features. These eight features measure how explanations relate to question content. Notably, humans rely heavily on `semantic_similarity` (66%) and `word_overlap_ratio` (63%) despite these being weak correctness signals—a key driver of the calibration gap.

Feature	Description
`semantic_similarity`	BERT embedding cosine similarity
`jaccard_similarity`	Jaccard similarity of word sets
`word_overlap_ratio`	Shared words / explanation words
`word_overlap_count`	Raw count of shared words
`shared_entities`	Named entities in both texts
`exp_entity_count`	Total entities in explanation
`exp_entity_density`	Entity count / token count
`mentions_answer_text`	Answer string appears in explanation

Table 8: Content grounding features (`content_*`).

Surface Linguistic Features. These 15 features capture basic textual properties. Length measures show that longer explanations are not necessarily better.

Feature	Description
word_count	Total word count
sentence_count	Total sentence count
avg_word_length	Average characters per word
avg_sentence_length	Average words per sentence
type_token_ratio	Lexical diversity
flesch_reading_ease	Readability (higher = easier)
flesch_kincaid_grade	Grade level required
smog_index	SMOG readability index
ari	Automated readability index
noun_ratio	Proportion of nouns
verb_ratio	Proportion of verbs
adj_ratio	Proportion of adjectives
adv_ratio	Proportion of adverbs

Table 9: Surface linguistic features (surface_*).

Structural Features. These nine features capture formatting and discourse structure. The has_quotes feature is the top human predictor at 70%, yet provides a weak correctness signal—another driver of miscalibration.

Feature	Description
has_quotes	Contains quotation marks
clue_mentions	References to "clue," "line," "phrase"
position_mentions	References to "first," "beginning," etc.
mentions_answer	Contains "answer," "guess," "conclude"
num_clauses	Syntactic complexity
causal_connective_ratio	Because, therefore, thus, hence
contrastive_connective_ratio	However, but, although, yet
additive_connective_ratio	And, also, moreover, furthermore
has_parentheses	Contains parenthetical asides

Table 10: Structural features (struct_*).

Reasoning Pattern Features. These five features detect specific reasoning strategies through lexical patterns.

Feature	Description
elimination	"rules out," "can't be," "eliminate"
pattern_matching	"characteristic of," "typical," "consistent with"
analogical	"similar to," "like," "resembles"
definitional	"defined as," "means," "refers to"
abductive	"best explains," "most likely," "suggests"

Table 11: Reasoning pattern features (reason_*).

Pragmatic Features. These six features capture interaction style and metacognition.

Feature	Description
metacognitive_markers	"I think," "I believe," "uncertain"
addressee_references	Direct "you" references
imperative_count	"consider," "note," "look," "see"
alternative_mentions	"also," "alternatively," "could be"
conditional_count	"if" statements
limitation_admission	"don't know," "not sure," "might be wrong"

Table 12: Pragmatic features (pragma_*).

F.3 Full Feature Rankings

Figure 13 shows single-feature prediction accuracy for all features across both outcomes. LLM-assessed features cluster at high correctness accuracy (70–76%) but near-chance human prediction (44–58%). Surface and structural features show the opposite pattern.

F.4 Key Takeaways

1. **One shared signal:** Only evidential_grounding appears in both top-10 lists, suggesting humans recognize quality when reasoning explicitly cites evidence.

2. **LLM features predict correctness:** question_comprehension (76%), evidential_grounding (74%), domain_expertise_display (74%), and reasoning_coherence (72%) strongly predict whether the AI is correct.

3. **Surface features predict human trust:** has_quotes (70%), semantic_similarity (66%), and word_overlap_ratio (63%) predict human selection but are weak correctness signals.

4. **Implication:** AI explanations should make reasoning explicit through evidence citation. Humans should evaluate reasoning quality rather than surface familiarity.

F.5 Implementation Details

Features are extracted using a modular pipeline with seven extractors:

- **SurfaceLinguisticExtractor**: Uses spaCy for tokenization and POS tagging; textstat for readability indices.
- **StructuralExtractor**: Pattern matching for discourse connectives and formatting signals.
- **ContentGroundedExtractor**: Uses sentence-transformers (all-MiniLM-L6-v2) for semantic similarity; spaCy for entity extraction.
- **ReasoningTypeExtractor**: Regex patterns for reasoning strategy indicators.

- **EpistemicExtractor**: Lexicon-based detection of hedges, certainty markers, and modal verbs.
- **PragmaticExtractor**: Pattern matching for metacognitive and interaction markers.
- **LLMBasedExtractor**: GPT-4o with structured prompting for semantic quality assessment (1–5 scale, normalized to 0–1).

All features use consistent prefixes (`surface_`, `struct_`, `content_`, `reason_`, `epist_`, `pragma_`, `llm_`) enabling category-based filtering and analysis.

G AI Assistant Systems

This appendix details the AI assistant systems used in the tournament. Each AI teammate consists of two sub-agents: one for the tossup (tossup) phase and one for the bonus (bonus) phase, reflecting the distinct requirements of each question mode.

G.1 System Collection

We collected AI systems through a four-week online competition prior to tournament play. Participants submitted their systems to our evaluation platform, where they were tested on a held-out set of 80 questions per mode (separate from the questions used in the live tournament). This competition format encouraged diverse architectural approaches while providing participants time to iterate on their designs. The resulting systems span a range of strategies, from single-model configurations with carefully tuned prompts to multi-step pipelines involving text-analysis, ensemble voting, and confidence calibration. During the tournament, teams draft these AI teammates across different rounds using the serpentine selection process described in Section 3.

G.2 Tossup Agent Specifications

In the tossup phase, agents receive a partial prefix of the question as it is read aloud, since it is supposed to be *interrupted*. At each word boundary, the agent outputs two values: a boolean *buzz* decision and its most confident *guess* given the clues seen so far. Our moderator stops reading the question when the model first buzzes and accepts the guess at that point as the team's response. Table 13 details each system's architecture for this phase, where agents must decide when to buzz and provide answers under time pressure. The systems exhibit considerable variation in their approach to confidence calibration and answer generation.

G.3 Bonus Agent Specifications

In the bonus phase, agents receive the lead-in context and the current part of a three-part question. For each part, the agent outputs three values: a *guess*, a *confidence* score, and an *explanation* justifying the answer. Table 14 details each system's architecture for this phase. These configurations often differ from their tossup counterparts, reflecting the different demands of deliberative question answering.

H Tournament Platform and Data Release

H.1 Tournament Web Application

Both tournaments were conducted using a custom web application built for human–AI collaborative quizbowl. The platform provides:

- **Real-time question display** with clue-by-clue progression for tossup questions, mirroring the pacing of a live moderator.
- **Buzzer integration** via BuzzIn Live for the online tournament (physical buzzers for the in-person event), supporting simultaneous human and AI buzz attempts with millisecond-resolution timestamps.
- **Bonus collaboration interface** (Figure 2) presenting human guess entry, AI suggestions with confidence and explanations, and final answer submission in a three-step flow.
- **Moderator dashboard** for controlling game flow, recording answer correctness, managing the serpentine draft, and tracking scores in real time.

The online tournament was conducted over a Zoom call with the moderator sharing the web application screen and managing game progression through the dashboard. All participant interactions with the interface were logged automatically.

H.2 Model Submission Platform

We received approximately 60 tossup agent submissions and 25 bonus agent submissions through a four-week open competition. Agents were evaluated on a held-out set of 80 questions from a recent tournament-difficulty quizbowl question set.

System	Models Used	Calls	Approach
System 1	GPT-4o (answer)	1	*Single-shot* solver using prompt-level calibration norms and probability gating for latency optimization.
System 2	GPT-4.1 (answer)	1	*Rule-intensive* Quizbowl specialist with domain norms embedded in prompts, including indicator discipline and discrete confidence scales.
System 3	GPT-4.1-nano (word count), GPT-4.1-nano (entity count), GPT-4.1 (answer), GPT-4.1-mini (calibration)	4	*Confidence engineering* pipeline that decouples correctness from certainty, rescaling confidence using question completeness heuristics.
System 4	GPT-4o (voter), GPT-3.5-turbo (voter), GPT-4o (aggregator)	3	*Exact-match voting* across heterogeneous models; abstains with zero confidence unless all voters fully agree.
System 5	Claude-3.5-Sonnet (analysis), GPT-4o (answer), Command-R-Plus (confidence)	3	*Structured pipeline* separating clue analysis, answer generation, and confidence assignment.
System 6	Command-R (extraction), GPT-4.1-mini (answer), Claude-3.5-Haiku (verifier), GPT-4o-mini (aggregator)	4	*Verifier-centered* architecture treating generation as hypothesis, penalizing disagreement via AND-style aggregation.
System 7	GPT-4o-mini (candidate), Claude-3.5-Haiku (candidate), GPT-4.1-mini (cross-check), Command-R (confidence)	4	*Cross-validation* combining independently proposed answers with match-based confidence scoring.
System 8	Command-R (answer)	1	*Single-pass* Cohere-native design emphasizing disciplined uncertainty reporting and probability thresholds.

Table 13: Tossup (proactive phase) agent specifications. Each system implements a distinct strategy for answer generation and confidence calibration under the time pressure of buzzer-style questions.

Tossup agents were ranked by expected points against human buzz-point logs; bonus agents by average part-level accuracy. We paired each submitter's top tossup and bonus agents into a single "quizbowl AI agent" and selected the top 8 paired agents, ensuring architectural diversity across base models and prompting strategies. See Appendix G for per-agent specifications.

H.3 Data and Code Release

We release:

1. **Dataset:** All tournament questions (tossup and bonus), human responses at each stage (initial guess, final answer), AI responses (answers, confidence scores, explanations), muting decisions, answer correctness labels, and anonymized player metadata (experience level, team assignment). All personal and identifiable information has been explicitly anonymized before release. Available at https://huggingface.co/datasets/qanta-challenge/qanta25-gamedata.

2. **Tournament application:** Source code for the web application, including the buzzer integration, bonus collaboration interface, moderator dashboard, and draft management system. Available at https://github.com/qanta-org/qb-tournament-runner.

3. **Analysis scripts:** All scripts used for feature extraction, statistical testing, figure generation, and the logistic regression analyses reported in this paper. Available at https://github.com/qanta-org/qanta25-analysis.

System	Models Used	Calls	Approach
System 1	GPT-4o-mini (answer)	1	*Single-shot* lightweight model with strict JSON-only output format.
System 2	GPT-4o (answer)	1	*High-accuracy* model with Quizbowl-specific guardrails for answer formatting.
System 3	DeepSeek V3 (answer)	1	*End-to-end* single DeepSeek model for all outputs.
System 4	GPT-4o (analyzer), GPT-4o (generator), GPT-4o (evaluator)	3	*Multi-role* pipeline reusing same model across analyzer, generator, and evaluator.
System 5	GPT-4o-mini (advisor)	1	*Advisory-style* solver providing answers with explanations.
System 6	Command-R (extraction), GPT-4.1-mini (generation), GPT-4o-mini (verification), Claude-3.5-Haiku (aggregation)	4	*Multi-model pipeline* with distinct roles: extraction, generation, verification, and aggregation.
System 7	Claude-3.5-Haiku (hypothesis), GPT-4o-mini (scoring), GPT-4.1-mini (calibration), Command-R (explanation)	4	*Sequential synthesis* for hypothesis generation, scoring, calibration, and explanation.
System 8	Command-R (answer)	1	*Single-pass* Cohere model with strict JSON-only output constraints.

Table 14: Bonus (deliberative phase) agent specifications. Systems vary from single-model configurations to multi-step pipelines, reflecting different strategies for collaborative question answering with explanations.

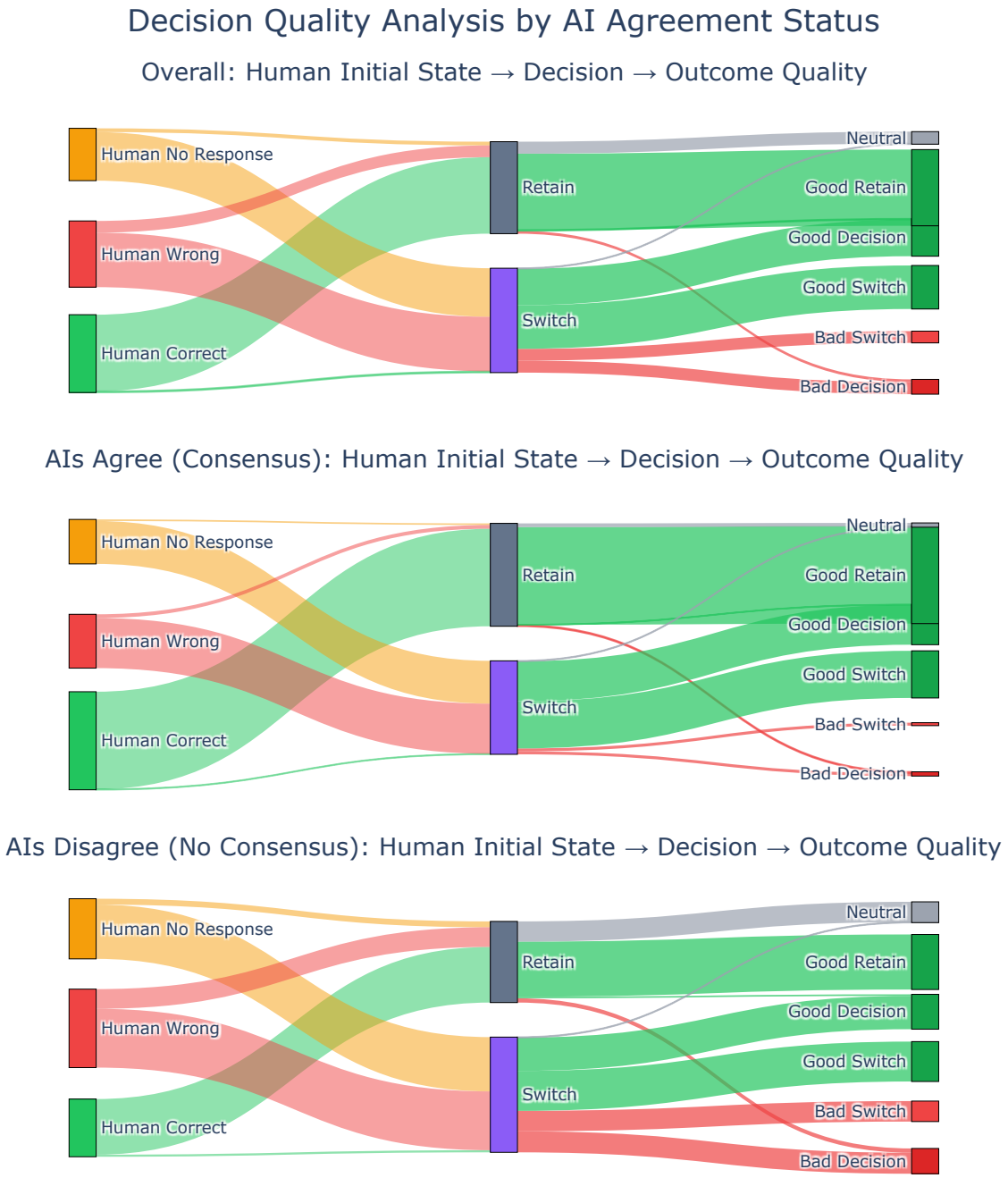

Figure 9: Decision quality analysis by AI agreement status. Top: overall flow from human initial state through decision to outcome. Middle: when AIs agree (consensus), teams switch more readily and achieve higher accuracy. Bottom: when AIs disagree, teams face harder choices and under-reliance increases. Flow widths represent proportion of responses; colors indicate outcome quality. Compare to the aggregate view in Figure 5.

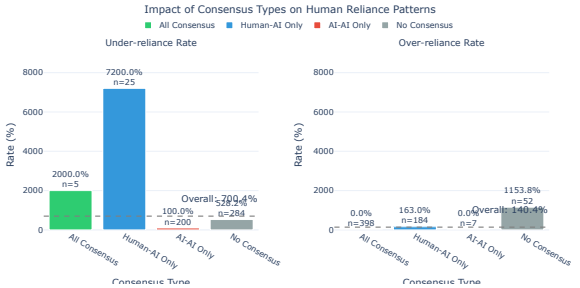

Figure 10: Over- and under-reliance rates by AI agreement condition. Under-reliance is highest when one AI confirms the human's incorrect answer (rightmost bars), illustrating how confirmation bias amplifies errors.

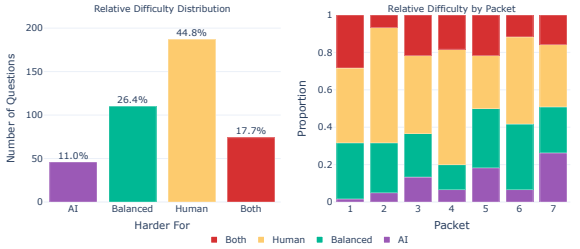

Figure 11: Relative difficulty distribution across bonus questions. Left: overall counts by category. Right: proportion breakdown by packet showing consistent patterns across question sets.

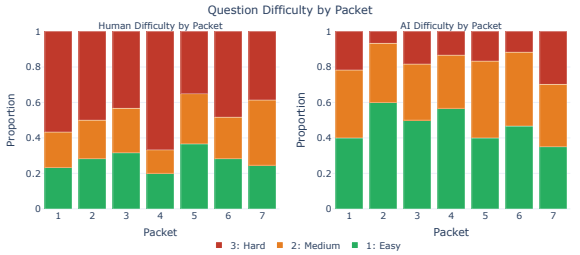

Figure 12: Question difficulty by packet for humans (left) and AI (right). Each bar shows the proportion of Easy, Medium, and Hard questions within a packet. Human difficulty varies more across packets than AI difficulty.

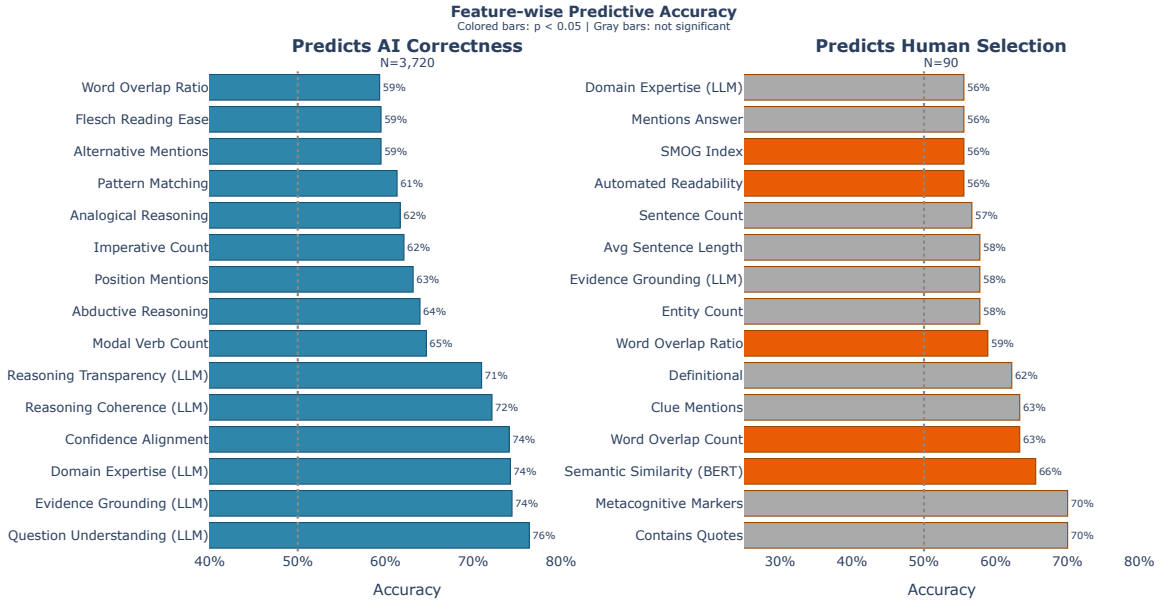

Figure 13: Full feature-wise prediction accuracy. Left: predicting AI correctness. Right: predicting human selection. Colored bars indicate statistical significance ($p < 0.05$); gray bars are not significant. LLM-assessed features dominate correctness prediction but are largely ignored by humans.