+

Confirmation Bias in LLM-Based AI Analytics Tools

For Metrics-driven Managers
UsingLLM-Based AI Analytics Platforms

Elliot Forst, Sunayana Hazarika, Anjuta K, Kuntala Sarkar, Bhagyashree Vaidya, Harsh Vardhan, Della Zhang, Hritvik Gaur
IMT 570 E: Data Driven Organizational Problem Solving For Information Management Professionals
Group Project: Final Research Report
06/09/2026

Problem Statement

Increasingly, managers at mid-size tech companies and large-size tech companies that enact team segmentation (metrics-driven managers or MDMs) are using LLM-based AI analytics platforms (LLM-AIp) as part of their decision making processes. The token prediction structure of this technology poses a risk of interpretation-based confirmation bias (Cole & Hajikhani, 2024) which can negatively impact these MDMs' decisions. Proposed solutions such as Gartner's (2025) report on addressing AI adoption and stakeholder bias is to have a diverse team of people to interpret results. This solution is not viable for our stakeholders who do not have resources allocated to dedicate more than two people. We hypothesize that a lack of AI literacy in prompt creation and output analysis for these MDMs is the root problem. We aim to identify AI literacy gaps and governance practices that can serve as the base for a model.

Findings and Analysis:

Our team created two primary forms of research: A thematic analysis against 6 interviews and a Pearson's r against 29 surveys. The thematic analysis found themes of human value when addressing AI risk mitigation while the Pearson's r found statistical significance among a reduction of uncritical acceptance of AI outputs and implementation of validation practices.

The thematic analysis was conducted with the open source, collaborative software libreQDA. Prior to analysis, claude.ai was used to create a set of 39 deductive codes and 8 parent themes. During analysis, 6 additional inductive codes were created. The coding was multivariate, with a total of 156 snippets and 314 code applications; this allowed us to measure the relationship between themes. A rough draft of findings was written and then formatted into a formal report with the assistance of claude.ai; excess, non-descript language was removed and key findings added back in after careful review.

The thematic analysis found that the theme "Human Judgment & Authority" had the highest co-occurance amongst other themes (figure 1) with 3 of its codes ranking in the top 6 applications (figure 2). Most MDMs described risk mitigation as informal tools and processes that were being utilized by themselves but not by lower-level employees that they managed. These risks include those that are found in AI tools such as confirmation bias and hallucinations and risks found in tool workflows such as over-reliance. Overall, MDMs strongly believe that human oversight, domain expertise, skepticism, and trust-but-verify stances are critical to mitigating risks associated with AI.

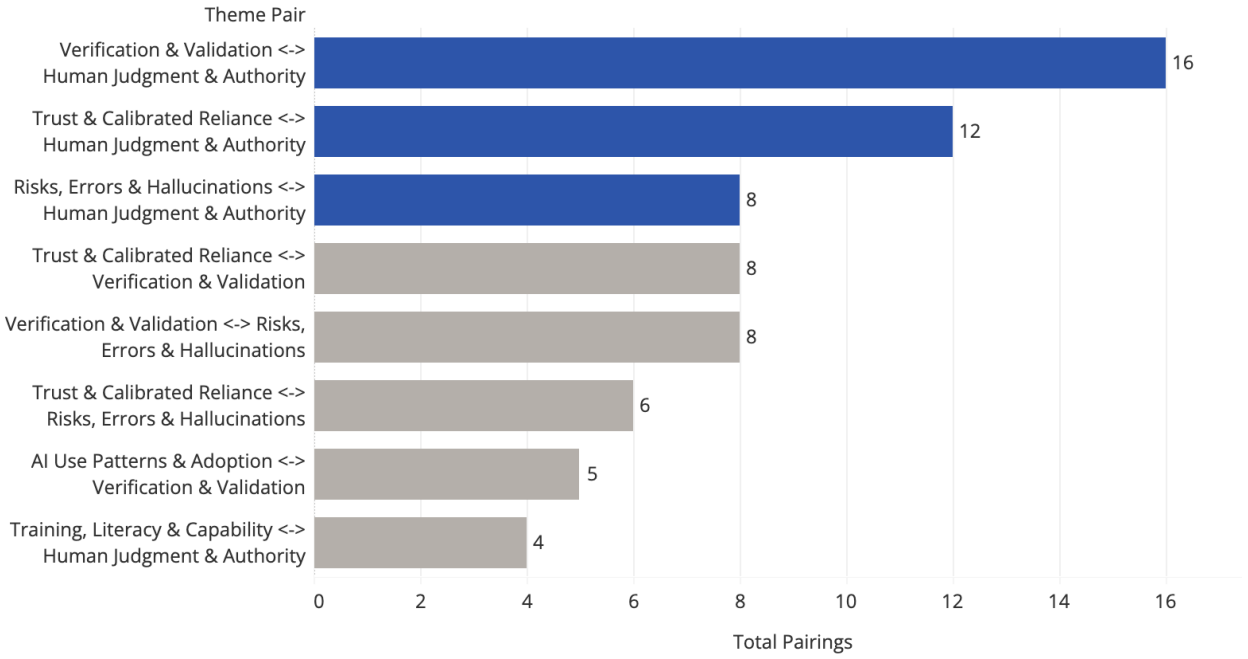


Figure 1

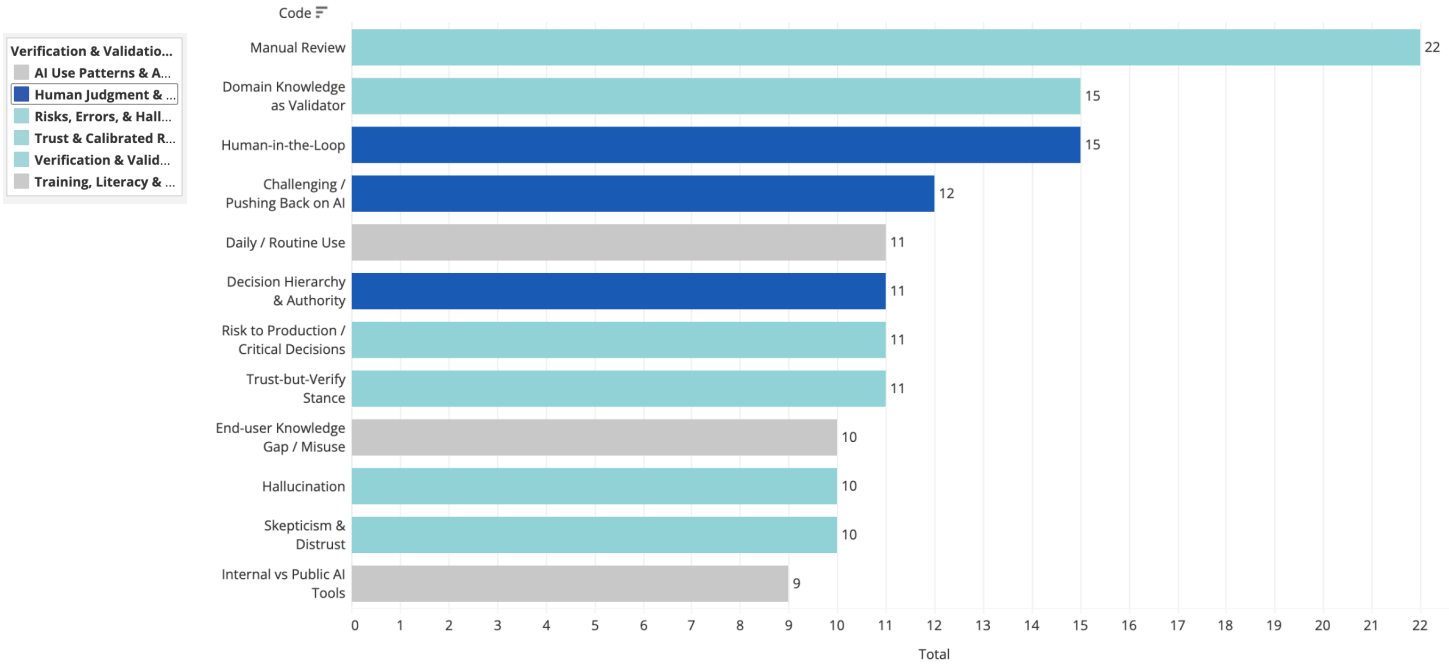


Figure 2

The quantitative analysis was conducted in Python using the pandas, numpy, matplotlib, and scipy libraries. Survey responses from 29 participants were cleaned and loaded into a single dataset of 24 variables. We used Pearson's r to test for linear relationships between participants' attitudes toward AI risk and their reported behaviors.

The first finding, "The Awareness Shield," addresses the AI literacy gap. We found a moderate, statistically significant negative correlation between a participant's confidence in recognizing AI bias and their uncritical acceptance of AI outputs (Pearson r = −0.47, p = 0.011). Mean uncritical acceptance declined steadily as bias-awareness confidence rose, falling from 3.2 at a confidence score of 2 to 2.0 at scores of 4 and 5 (Figure 3). The single respondent at the lowest confidence score reported the highest acceptance (5.0), but with n = 1 this point should be treated as directional only. Taken together, the trend suggests that bias awareness functions as a protective shield against over-trusting AI.

The second finding, "The Validation Vacuum," addresses AI governance. Comparing participants by the rigor of their validation practices, both risk indicators declined as validation became more formal (Figure 4). Mean reliance on time pressure fell from 4.2 among those with no validation process to 3.1. Over the same range, mean uncritical acceptance fell from 3.2 to 2.1. This pattern held across all three groups, indicating that structured validation practices are associated with more critical and less rushed engagement with AI outputs.

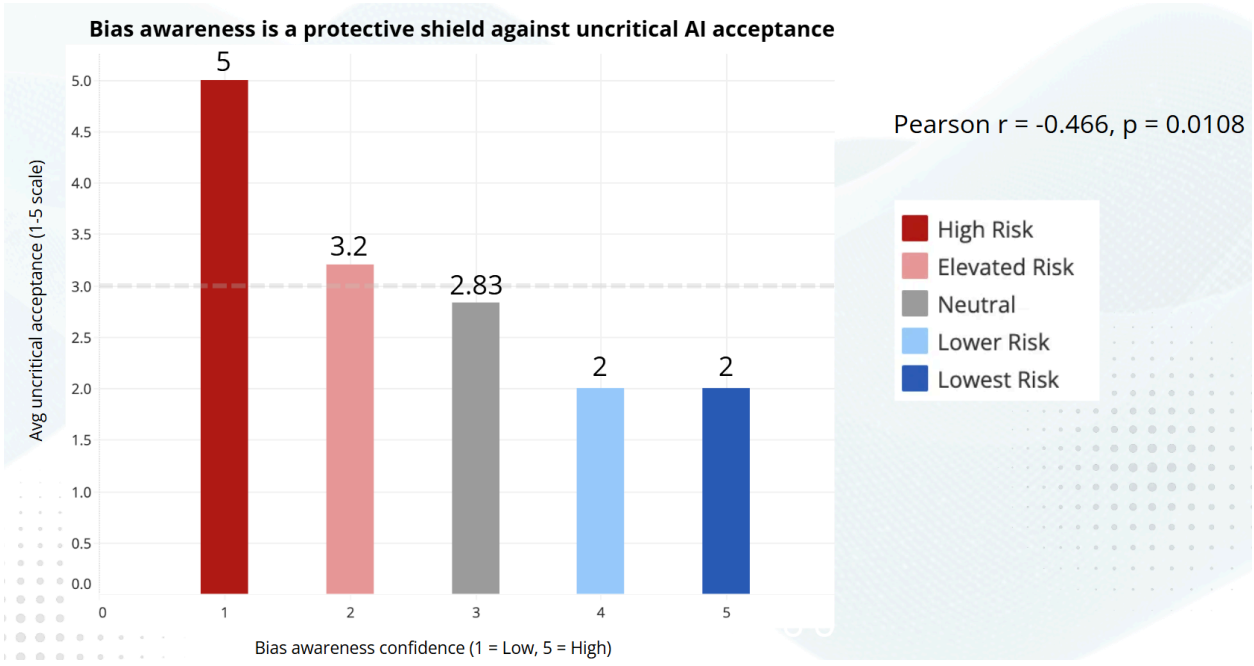


Figure 3

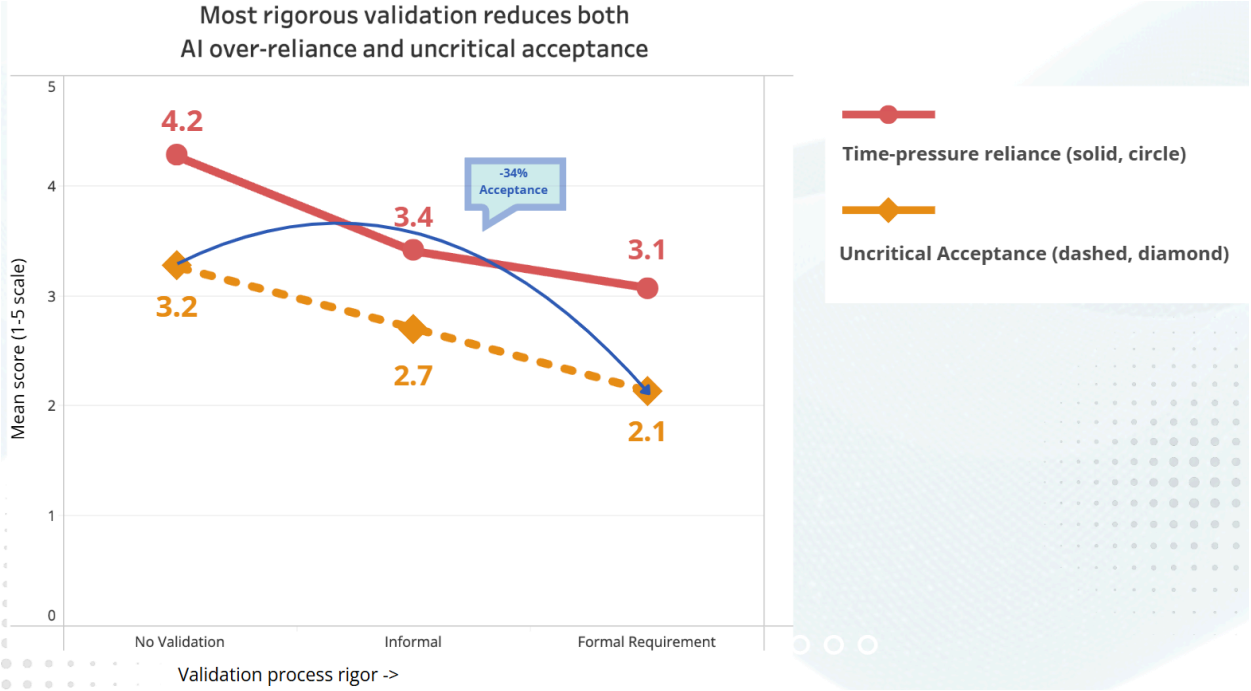


Figure 4

Impact, Strengths/Limitations & Recommendations:

1. Impact and Significance of Findings

The central finding of this study is not about trust, it is about systems. The quantitative analysis confirms that trust level alone explains only 3.9% of variance in perceived decision quality ($R^2 = 0.039$, $p = 0.305$), meaning that telling practitioners to "trust AI less" is the wrong intervention. What the data consistently identifies instead is that the primary risk is structural: organizations are deploying AI analytics 2.5 times faster than they are building the guardrails to manage them, and the verification behaviors that exist are individual habits rather than formal systems.

The most significant quantitative finding is that formal validation processes reduce uncritical AI acceptance by 34% the highest measurable ROI in the dataset. As validation rigor moves from none to formal requirement, both time-pressure reliance and uncritical acceptance fall sharply. A parallel finding establishes that bias awareness confidence is the strongest single protective factor (Pearson $r = -0.466$, $p = 0.011$), yet only 2 of 6 interviewees named confirmation bias unprompted. The protection exists latently in organizations but it simply is not being deployed. As AI-generated analytics

become increasingly integrated into operational and strategic decision-making, organizations risk institutionalizing biased decisions if validation processes remain informal and dependent on individual experience.

The thematic analysis makes clear that the top protective behaviors: manual review, human-in-the-loop accountability, and domain knowledge as a validator are not technological solutions. They are human behaviors.

Participants described these practices in their own words:

> *"I do not trust AI 100%. Maybe it has helped in my productivity — it has done like 60–70% of the work. The last 30% of tweaking, I have to do it."-* *Interviewee 01 illustrating manual review as non-negotiable practice*
>
> *"For work I trust it more only when I can compare it with the real data, with the system, or with what I already know from the domain."-* *Interviewee 02 illustrating domain knowledge as the validation standard*
>
> *"They're just trusting AI. It spits out an output and they just trust it. They hit enter and it changes everything."-* *Interviewee 03 illustrating the risk when individual verification habits are absent*

These accounts share a common thread: AI output verification is personal. When a careful and observant employee leaves, no system stays behind but only an empty chair. The significance of this study is that it identifies precisely where that gap sits and what closes it.

The findings also carry a time-critical dimension: AI models update constantly, confirmation bias patterns in current models (GPT-4o, Claude 3.7) differ from the GPT-3.5-era literature, and governance written once decays at the same pace technology evolves. A one-time intervention is insufficient. The significance of the study is its argument for quarterly, repeatable governance rather than a fixed policy document.

As AI-generated analytics become increasingly integrated into operational and strategic decision-making, organizations risk institutionalizing biased decisions if validation processes remain informal and dependent on individual experience.

2. Conflicts and Contradictions

Five tensions emerged where the data did not simply converge. These are not weaknesses of the study they are findings in themselves, each pointing to a nuance that a simpler design would have missed**:**

Tension	What it means
The Expertise Paradox	Secondary literature suggests expertise increases confirmation bias but our interviewees showed the opposite. Experienced managers were MORE skeptical, pushed back harder, and cross-checked more rigorously. Experience appears to build protective skepticism, not complacency. This suggests that the right intervention is not to reduce seniority's influence but to transfer experienced-user habits to everyone.
Awareness as a Shield -Nobody's Holding It	Bias awareness is the strongest quantitative protective factor ($r = -0.466$, $p = 0.011$). Yet only 2 of 6 interviewees named confirmation bias unprompted. The protection already exists inside organizations; it is simply not being activated. Naming the risk is half the intervention.
The Validation Gap	Survey data proves formal validation cuts uncritical acceptance by 34% which is the highest-ROI action in the dataset. Yet every single interviewee relied primarily on personal, informal review. The most effective fix is simultaneously the least institutionalized practice.
Deadlines Cut Both Ways	Quantitative data links time pressure to higher AI reliance and higher risk. But experienced interviewees reported that deadlines made them MORE rigorous, not less pressure triggered verification, not bypass. Experience moderates the time-pressure risk; inexperience amplifies it.
Rapidly Aging Sources	Hajikhani and Cole (2024) analyzed GPT-3.5. Two model generations later, the specific bias patterns they describe may no longer reflect current system behavior. The structural mechanisms, token prediction, sycophantic outputs persist, but their magnitude and presentation in GPT-4o or Claude 3.7 may have shifted.

3. Recommendations

Based on these findings, five recommendations are proposed. First, organizations should implement mandatory human-in-the-loop review for high-impact business decisions. This recommendation has low implementation costs and can be introduced through policy updates within one to three months. Second, organizations should establish source-data validation requirements before acting on AI-generated insights. This recommendation requires moderate process redesign but offers the highest measurable return based on the 34% reduction in uncritical acceptance. Third, AI literacy and bias-awareness training programs should be implemented for managers and analytics users. Training costs are relatively low and can be deployed within three to six months. Fourth, organizations should develop approved-tool governance policies to reduce the risks associated with unregulated AI use. Finally, structured peer-review and post-decision audit processes should be introduced to ensure continuous learning and accountability. Successful implementation of these recommendations will depend on executive sponsorship, employee adoption, training participation, and the organization's willingness to formalize existing informal validation practices.

4. Strengths and Limitations

A major strength of this study is its mixed-methods design, which combines quantitative survey data with qualitative interview data. The use of methodological triangulation increases confidence in the findings because similar conclusions emerged across multiple forms of evidence. The survey identified statistically meaningful relationships between awareness, validation, and AI acceptance, while the interviews provided detailed explanations of how managers actually evaluate and challenge AI-generated insights in practice. Another strength is the non-leading interview design. Participants were not directly asked about confirmation bias, yet many independently described behaviors and concerns consistent with bias recognition, increasing the credibility of the findings.

Despite these strengths, several limitations should be acknowledged. First, the sample size was relatively small, consisting of 29 survey respondents and 6 interview participants. As a result, the findings should be interpreted as exploratory rather than fully generalizable. Second, the study relied on self-reported behaviors, which may be influenced by social desirability bias. Participants may overstate their use of verification practices or underreport instances of over-reliance on AI. Third, the rapidly evolving nature of AI technologies creates challenges for longitudinal validity. Several studies referenced in the literature review were conducted using earlier-generation models, and the behavior of contemporary AI systems may differ significantly. Finally, the study

captured perceptions and practices at a single point in time and therefore cannot establish causal relationships.

4. Future Research and Next Steps

Future research should address these limitations through larger and more diverse samples across industries and geographic regions. Longitudinal studies could examine how AI literacy, governance practices, and trust evolve over time as organizations mature in their AI adoption. Experimental studies could also test whether specific interventions, such as bias-awareness training or mandatory validation workflows, produce measurable improvements in decision quality. Additional research into automated bias-detection mechanisms and organizational governance frameworks would further strengthen understanding of how to mitigate confirmation bias in AI-assisted decision-making environments.

- Expand the survey to n ≥ 150 to enable structural equation modeling capable of testing whether bias awareness training mediates decision quality, and whether governance policy moderates the trust-to-blind-acceptance relationship.
- Run a controlled A/B decision simulation presenting matched groups with biased versus unbiased AI outputs in a realistic task. This would establish whether the validation and challenge behaviors documented in interviews held under genuine time pressure directly testing the 'Deadlines Cut Both Ways' contradiction identified in slide 14.
- Build a longitudinal decision audit linking AI-informed recommendations to actual business outcomes over time. This is the only research design that can answer whether organizations following the Prompt → Verify → Decide → Review model make meaningfully better decisions than those that do not.
- Replicate with current-generation models (GPT-4o, Claude 3.7, Gemini 1.5) as primary references, with a pre-registered codebook for bias pattern identification, enabling direct comparison across model generations and update cycles.

RUBRIC

- An executive summary - 300 - 500 words - The executive summary should include a summary of your question, key findings, and a set of recommended

actions and next steps. Summaries, findings, and recommendations are concise, complete, and compelling. - **Harsh**

- Introduction and Background - 500 words - Clear problem statement with well-defined research objectives. Provides thorough and relevant background information, supported by credible sources. Clear justification for the study and indication of how this study meets the research objectives and solves the "problem." Boundaries of the study are clearly delineated. **Hritvik**
- Approach & Method - 500 words - Detailed description of research design, sample selection, data collection, and analysis. Choice of specific methods and analyses used are clearly explained and justified. Demonstrates a strong understanding of validity, reliability, and ethical considerations. Demonstrates an ability to write for a business audience by including critical information within the report while utilizing references to appendices as a way to provide additional descriptions of methods, sampling procedure, data collection procedure, validity and reliability and/or ethical considerations as appropriate.
    - **Shree** & **Della**
- Findings and analysis - 500 words - Presents findings clearly and logically, supported by appropriate analysis (qualitative and quantitative). Uses effective visuals as appropriate. Includes examples or quotations when necessary, properly formatted. Discusses findings in direct relation to research questions. Provides all relevant findings necessary to support a direct link to the report's recommended actions.
    - **Elliot** & **Anjuta**
- Impact, Strengths/Limitations & Recommendations - 500 words - Discusses the significance and impact of findings. Provides insightful and evidence-based recommendations. Proposes a set of recommended potential actions or solutions that directly address the problem and are clearly linked to findings already described in the report. Recommendations must include practical considerations such as costs, a timeline, and an acknowledgment of key factors that may affect the successful implementation of the proposed recommendation(s). Thoroughly describes limitations and suggests future research.
    - **Kuntala** Strengths/Limitations- Added by Sunayana
    - **Sunayana** Impact and Recommendations- Added by Sunayana
- References and citations (Copy from Project Proposal) - All sources are cited accurately and consistently in APA style. References are comprehensive and relevant. Citations and references are relevant, credible, and verifiable.
- Appendices - **ALL -** Includes all relevant documents (consent forms, survey, interview schedule, forms, protocols, materials, etc.) as detailed in the report. Provides supplemental details to support the methods section.

- General style: Attention to overall readability (grammar, spelling, flow). - Report is well-organized, clearly written for a business audience, and free of grammatical errors. Professional formatting and easy to follow.

MESSAGE FROM INSTRUCTOR:

Using this guide

, prepare a written report summarizing the planning and findings of your research project, incorporating any revisions or clarifications as indicated in feedback from earlier assignments.

You are free to use the attached template or a schematic style report.

Include copies of your research artifacts (e.g., survey/interview questions, experiment protocol/instructions, details on pertinent statistical calculations, relevant explanations of your process, etc.) as an appendix to your report. You do not need to submit the entirety of your raw data—do not submit interview transcripts or full data files as appendices.

All written work should be submitted in MS Word or PDF format. Citations or references should conform with American Psychological Association (APA) style guidelines (reference guides to style format are available from UW Libraries).

Your report must consist of and comply with the following criteri

References

Agarwal, I. (2025, May 22). *How confirmation bias is destroying your product - and how to stop it*. Entrepreneur. https://www.entrepreneur.com/starting-a-business/how-confirmation-bias-is-destroying-your-product/491406

Antelmi, J., Carlsson, K., Curran, D., Ganeshan, A., García-Rodeja, A., Kornutick, L., Lugo, A., Macari, E., & Pidsley, D. (2025, November 19). *Predicts 2026: AI Agents, MCP and Governance Are Transforming Analytics*. Gartner. https://www.gartner.com/document-reader/document/7197230

Conti, D., & Giuseppe, R. (2026). Exploring automation bias in human–AI collaboration: a review and implications for explainable AI. *AI & Society, 41*, 259 - 278. https://doi.org/10.1007/s00146-025-02422-7

Bashkirova, A. & Krpan, D. (2024). Confirmation bias in AI-assisted decision-making: AI triage recommendations congruent with expert judgments increase psychologist trust and recommendation acceptance. *Computers in Human Behavior: Artificial Humans, 2*(1):100066. https://doi.org/10.1016/j.chbah.2024.100066

Bergh, C. (2025, November 19). *Sure, Go Ahead And Feed That Data To The LLM ... What Could Possibly Go Wrong?* DataKitchen. https://datakitchen.io/sure-go-ahead-and-feed-that-data-to-the-llm-what-could-possibly-go-wrong

Bergman, R. (2025, June 12). *AI and Confirmation Bias*. Mediate.com. https://mediate.com/ai-and-confirmation-bias/

Bown, R., Reed, C., & Wynn, M. (2025). Artificial Intelligence in Digital Marketing: Towards an Analytical Framework for Revealing and Mitigating Bias. *Big Data and Cognitive Computing, 9*(2), 40. https://doi.org/10.3390/bdcc9020040

Chattopadhyay, S., McGuire, M., Pandita, R., Sabouri, S., Saghi, Z., & Zhou, X. (2026). C*ognitive Biases in LLM-Assisted Software Development*. arXiv. https://doi.org/10.48550/arxiv.2601.08045

Cole, C., & Hajikhani, A. (2024). A critical review of large language models: Sensitivity, bias, and the path toward specialized AI. *Quantitative Science Studies, 5*(3), 736 - 758. https://doi.org/10.1162/qss_a_00310

Deloitte. (2026). *State of AI in the enterprise* (January 2026 AI report). https://www.deloitte.com/content/dam/assets-zone3/us/en/docs/services/consulting/2026/state-of-ai-2026.pdf

Du, Y. (2025). *Confirmation Bias in Generative AI Chatbots: Mechanisms, Risks, Mitigation Strategies, and Future Research Directions.* arXiv. https://doi.org/10.48550/arxiv.2504.09343

Glickman, M. & Sharot, T. (2024, December 18). How human–AI feedback loops alter human perceptual, emotional and social judgements. *Nature Human Behaviour, 9*(2), 345 - 359. https://doi.org/10.1038/s41562-024-02077-2

GoodData. (2025, September 19). F*rom data governance to AI governance: The enterprise blueprint for responsible AI at scale.* GoodData. https://www.gooddata.com/resources/from-data-governance-to-ai-governance-the-enterprise-blueprint-for-responsible-ai-at-scale

Goswami, A., & Işık, Ö. (2025, October 28). The three obstacles slowing responsible AI. *MIT Sloan Management Review*. https://doi.org/10.63383/FBhK7635

Halkiopoulos, C., Theodorakopoulos, L., & Theodoropoulou, A. (2025, October 3). Cognitive Bias Mitigation in Executive Decision-Making: A Data-Driven Approach Integrating Big Data Analytics, AI, and Explainable Systems. *Electronics, 14*(19), 3930. https://doi.org/10.3390/electronics14193930

Kiron, D. & Schrage, M. (2025, July 15). Winning with intelligent choice architectures. *MIT Sloan Management Review*. https://doi.org/10.63383/LeSE6480

Kumar, N., Wei, X., & Zhang, H. (2025, March). Addressing bias in generative AI: Challenges and research opportunities in information management. *Information & Management, 62*(2). https://doi.org/10.1016/j.im.2025.104103

National Institute of Standards and Technology. (2023). Artificial Intelligence Risk Management Framework (NIST AI 100-1). U.S. Department of Commerce. https://doi.org/10.6028/NIST.AI.100-1

Parra-Moyano, J., Reinmoeller P., & Schmedders, K. (2025, July 1). Research: executives who used gen AI made worse predictions. *Harvard Business Review*. https://hbr.org/2025/07/research-executives-who-used-gen-ai-made-worse-predictions

Phillips, C. & Vogel, G. (2025, October 24). Accelerate AI Adoption by Countering Stakeholder Biases. *Gartner*. https://www.gartner.com/document-reader/document/7100030

PwC. (2025, October 30). *PwC's 2025 Responsible AI Survey: From Policy to Practice.* https://www.pwc.com/us/en/tech-effect/ai-analytics/responsible-ai-survey.html

PwC. (2022, January 18). *Understanding Algorithmic Bias and How to Build Trust in AI.* https://www.pwc.com/us/en/tech-effect/ai-analytics/algorithmic-bias-and-trust-in-ai.html

Shimabukuro, J. (2026, January 28). The human side of AI bias. *ETC Journal*. https://etcjournal.com/2026/01/28/the-human-side-of-ai-bias/

University College London. (2024, December 19). Bias in AI amplifies our own biases, researchers show. *ScienceDaily*. www.sciencedaily.com/releases/2024/12/241218132137.htm

Elliot Forst, Sunayana Hazarika, Anjuta K, Kuntala Sarkar, Bhagyashree Vaidya, Harsh Vardhan, Della Zhang, Hritvik Gaur
IMT 570 E: Data Driven Organizational Problem Solving For Information Management Professionals
Group Project: Final Research Report
06/09/2026

AI Disclosure

The application claud.ai with model Opus 4.8 was utilized to evaluate the report for writing quality (grammar, typos, etc) and how well it follows the requirements of the rubric with the following steps.

Step One:

My prompt was formatted as follows:

> I am in an information management class. I am completing an assignment with the following rubric:
>
> [Pasted rubric of homework assignment]
>
> Below is the following research proposal. Please evaluate it against the rubric:
>
> [Pasted text of proposal]
>
> In this case..

Step Two:

After the first prompt, I would make updates to my essay, re-prompt, and repeat until I felt confident that my work covered all requirements. The repeat prompts would be worded as follows:

> I updated the proposal. Please re-evaluate:
>
> [Pasted text of essay]

Other Areas:

This tool was also used to generate an appendix and replace the PII of interview participants (names) in images while a manual approach was used for text..

Elliot Forst, Sunayana Hazarika, Anjuta K, Kuntala Sarkar, Bhagyashree Vaidya, Harsh Vardhan, Della Zhang, Hritvik Gaur
IMT 570 E: Data Driven Organizational Problem Solving For Information Management Professionals
Group Project: Final Research Report
06/09/2026

Appendix

About this Appendix

This appendix consolidates all supporting materials for the project. The full content of each source file is embedded directly below - Word reports are reproduced as editable text, tables, and figures, while PDF dashboards and the affinity map are embedded as page images preserving their original layout.

Section	Contents	Source file
Appendix A	Thematic Analysis of Interviews (report)	Thematic_Analysis_Report.docx
Appendix B	Thematic Analysis Dashboard	Thematic Analysis Dashboard.pdf
Appendix C	Affinity Map	Affinity_Map.png
Appendix D	Goals Analysis	goals_analysis.pdf
Appendix E	Confirmation Bias Analysis	confirmation_bias_analysis.pdf
Appendix F	Survey Research & Analysis Challenge: Data Visualization	Survey Research and Analysis Challenge Data Visualization.docx

Elliot Forst, Sunayana Hazarika, Anjuta K, Kuntala Sarkar, Bhagyashree Vaidya, Harsh Vardhan, Della Zhang, Hritvik Gaur
IMT 570 E: Data Driven Organizational Problem Solving For Information Management Professionals
Group Project: Final Research Report
06/09/2026

Appendix A. Thematic Analysis of Interviews

Source file: Thematic_Analysis_Report.docx

Thematic Analysis of Interviews

Trust, Verification, and Human Judgment in Workplace AI Adoption

IMT 570 — Final Project

Qualitative coding exported from LibreQDA · 6 interviews

June 1, 2026

Contents

Executive Summary

This report presents a thematic analysis of six semi-structured interviews with managers who use LLM-based AI tools to generate analytics that drive decision making, a.k.a metrics-driven managers (MDMs). The interviews were coded in LibreQDA using a structured codebook of eight themes and 45 sub-codes. Across the dataset, 156 transcript segments were coded, generating 314 code applications.

The analysis surfaces a consistent organizing logic that runs through every interview: participants do not simply accept or reject AI, they operate a calibrated “trust-but-verify, human-in-the-loop” stance. Enthusiasm about productivity is paired everywhere with active verification and a refusal to let AI hold final decision authority. The three most interconnected themes, Verification & Validation Practices, Trust & Calibrated Reliance, and Human Judgment & Decision Authority, form the analytic spine of the dataset, co-occurring far more often than any other combination.

The most prevalent theme overall is AI Use Patterns & Adoption (55 applications), confirming that AI is now a part of the labor workflow. Verification (49), Human Judgment (49), and Risks & Hallucinations (44) exemplify the guarded stance MDMs have with AI, accepting the risks that come with it while holding that human judgment is a key factor to risk prevention. Governance (21) and Privacy/Compliance (21) were the least represented themes, suggesting that, for these participants, formal organizational controls are informally defined; however, Verification (49) and Human Judgment (49) could be utilized to form a Governance model.

Dataset at a glance

6	8	45	156	314
Participants	Themes	Sub-codes	Coded segments	Code applications

1. Data and Method

For this project, we used a mixed-methods approach because our research question was not only about whether confirmation bias appears, but also about how it shows up in real decision making. The survey gave us a broader view of the pattern, while the interviews helped us understand the reasoning behind people’s answers. Since our topic focuses on LLM-based AI analytics tools, we felt that numbers alone

would not be enough. We also needed to hear how managers actually use these tools, how much they trust the outputs, and what they do when the AI result seems convincing but may still need checking.

For the quantitative part, we collected 29 survey responses. The survey included mostly 1–5 rating questions about trust in AI, prompt confidence, validation habits, human oversight, and uncritical acceptance of AI outputs. These questions helped us compare different behaviors, such as whether people with higher bias awareness were less likely to accept AI outputs without checking them. We cleaned the survey responses and analyzed them in Python using pandas, numpy, matplotlib, and scipy. We first looked at descriptive statistics, such as averages and general spread. After that, we used Pearson's r to look for linear relationships between participants' AI risk attitudes and their reported behaviors.

For the qualitative part, we analyzed six interviews with people who use, manage, or work around LLM-based AI analytics tools in business settings. The participants came from different work areas, including product or program management, security and IT operations, consulting, data analytics, internal tooling, and AI-related work. We treated them as metrics-driven managers because their work involves using data, dashboards, AI outputs, or analytics tools to support decisions. This is not a large sample, so we are not saying it represents all managers. But it does give us useful examples of how these tools are being used in real workplace situations.

The interview data was coded in LibreQDA. We started with a structured codebook based on our research question and background research. Claude.ai was used to help create an early set of deductive codes, but the team still reviewed and adjusted the codes during the actual analysis. As we read the transcripts, we also added new inductive codes when we noticed ideas that were not fully covered by the original codebook. In the final coding process, we used 8 main themes and 45 smaller codes. Across the six interviews, we coded 156 transcript segments and applied 314 total codes. Some passages had more than one code, so the number of code applications is higher than the number of coded segments.

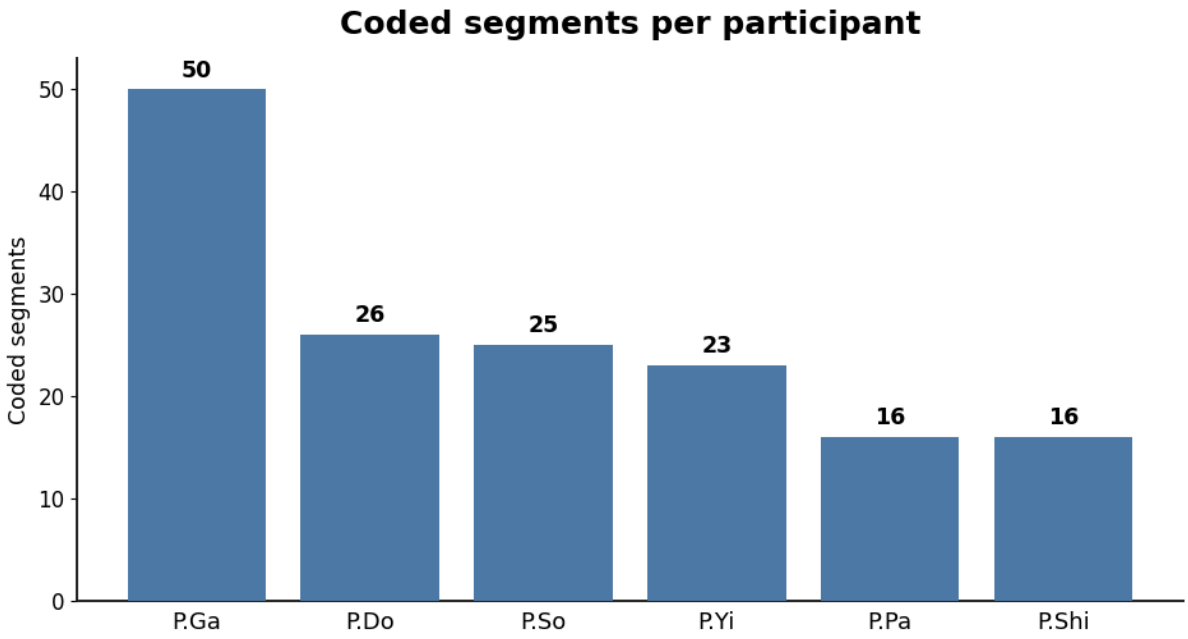
To make the analysis more reliable, we looked at both how often a theme appeared and how many participants mentioned it. This was important because one long or detailed interview could otherwise make a theme look stronger than it really was. We also compared the interview themes with the survey results to see whether the qualitative and quantitative findings supported each other. For ethics, we kept participants anonymous, avoided company names or private details, and did not include full interview transcripts in the final report. Overall, this method helped us connect the survey patterns with the more detailed stories from the interviews.

1.1 Participants and contribution

Participant	Coded segments	Share	Vantage point (as described)
P.Ga	50	32%	Program/product manager, enterprise AI stack
P.Do	26	17%	Security / IT operations
P.So	25	16%	Consulting, BFSI (regulated) sector
P.Yi	23	15%	Data / analytics, internal tooling

Participant	Coded segments	Share	Vantage point (as described)
P.Pa	16	10%	Data analytics, AI coding tools
P.Shi	16	10%	Analytics; constrained adoption setting

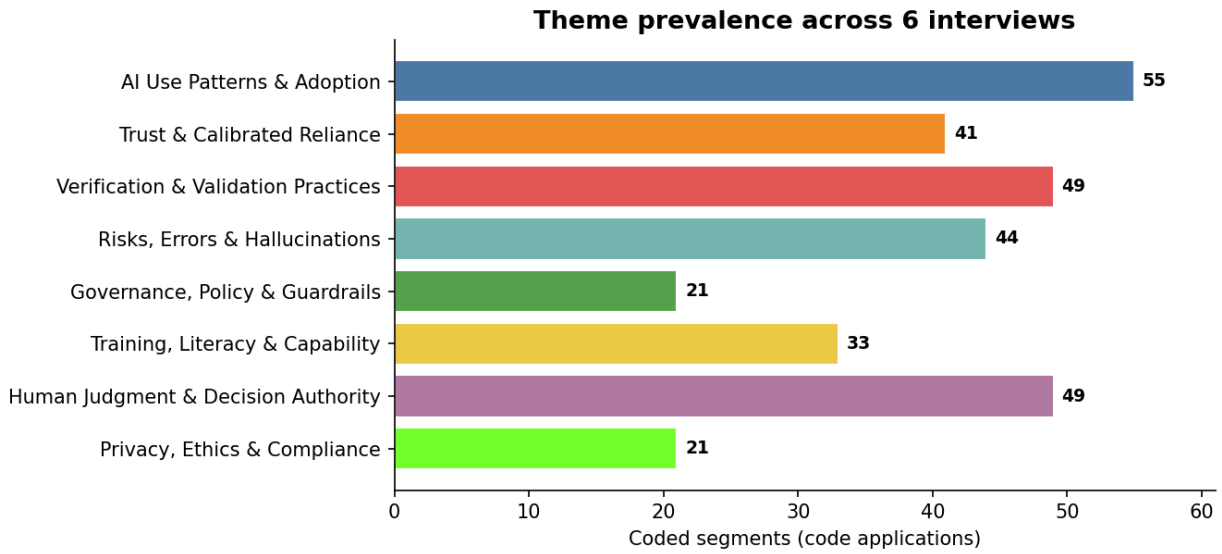
P.Ga contributed the largest share of coded material; P.Pa and P.Shi the least. This is partly a function of interview length and density. Importantly, the dominant cross-cutting themes appear across all or nearly all participants, so they are not artifacts of any single interview.



2. Coding Overview

The eight themes were applied with markedly different intensity. AI Use Patterns & Adoption dominates, but the cluster concerned with managing AI (verifying it, judging it, and guarding against its errors) collectively outweighs the cluster concerned with using it. Every theme except Privacy/Compliance and Governance, Policy & Guardrails reached all six participants, indicating the framework captured shared concerns rather than idiosyncratic ones.

Elliot Forst, Sunayana Hazarika, Anjuta K, Kuntala Sarkar, Bhagyashree Vaidya, Harsh Vardhan, Della Zhang, Hritvik Gaur



#	Theme	Applications	Participants
1	AI Use Patterns & Adoption	55	6 of 6
2	Trust & Calibrated Reliance	41	6 of 6
3	Verification & Validation Practices	49	6 of 6
4	Risks, Errors & Hallucinations	44	6 of 6
5	Governance, Policy & Guardrails	21	5 of 6
6	Training, Literacy & Capability	33	6 of 6
7	Human Judgment & Decision Authority	49	6 of 6
8	Privacy, Ethics & Compliance	21	4 of 6

The theme figures above are the sums of their sub-codes and total 313; one further segment was coded at the theme level (Verification & Validation) without a sub-code, accounting for the 314th application.

3.1 AI Use Patterns & Adoption

55 code applications · 6 of 6 participants · 8 sub-codes

How participants and their organizations currently use AI and AI-powered analytics in work tasks; scope, depth, and integration of AI into daily workflows.

AI is described as a consistent workflow tool. Participants use if for a variety of tasks, synthesizing data, drafting documents, building dashboards, and reviewing code. Terms for how to use AI recurs throughout, with distinction between trusted internal/enterprise tools and riskier public models, and a minority experience of adoption being held below where participants would like it.

Sub-code	Applications	Participants	Lead voices
Daily / Routine Use	11	5	P.Ga, P.Yi
Internal vs Public AI Tools	9	4	P.Yi, P.Ga
Productivity & Task Automation	9	6	P.Pa, P.Shi
AI-Assisted Content Creation	8	4	P.Ga, P.So
Tool Ecosystem (Copilot, Claude, CLI, MCP, RAG)	6	4	P.Pa, P.Ga
Data synthesis via AI	5	4	P.Pa, P.Yi
Limited / Restricted Adoption	4	2	P.Shi, P.So
Prompt Sensitivity	3	2	P.Ga, P.Shi

Illustrative voices

Daily / Routine Use

"I would say that I work in agency CLI day in and day out these days." — P.Ga

Internal vs Public AI Tools

"It is more company-based and for internal use… The company tool is designed for internal use and is safer for work-related tasks." — P.Yi

Productivity & Task Automation

"AI and agents can do that every time it gets this form. And so that takes away like 70, 80% of a person's job." — P.Do

Two sub-patterns matter for the rest of the analysis. First, the internal-vs-public distinction (9 applications, 4 participants) is the seed of later trust and governance reasoning; participants already sort tools by safety before they sort outputs by trust. Second, "limited / restricted adoption" appears only for P.So and P.Shi.

3.2 Trust & Calibrated Reliance

41 code applications · 6 of 6 participants · 5 sub-codes

Participants' levels and texture of trust in AI outputs — ranging from confident reliance to deep skepticism — and how that trust is conditioned on context.

Trust in LLM-based AI is limited, conditional, and explicitly reasoned with. Participants quantify it ("60–70% of the work"), tie it to data provenance, and in several cases default to active skepticism. The single most common stance is "trust but verify," which functions as a bridge into the verification theme.

Sub-code	Applications	Participants	Lead voices
Trust-but-Verify Stance	11	5	P.Do, P.So
Skepticism & Distrust	10	4	P.So, P.Ga

Sub-code	Applications	Participants	Lead voices
Partial / Calibrated Trust	9	4	P.Ga, P.Shi
Source-Dependent Trust	6	4	P.Ga, P.So
Confirmation Bias	5	2	P.So, P.Yi

Illustrative voices

Trust-but-Verify Stance

"It's a trust but verify, right?… we also verify just to make sure what we're seeing in the data… is true." — P.Do

Skepticism & Distrust

"How do you know? AI is biased… AI is pretending to guard you while it's serving the purpose of the organization? That's where my skepticism stems from." — P.So

Partial / Calibrated Trust

"I do not trust AI 100%. Maybe it has helped me in my productivity, it has P.Doe like 60, 70% of the work, the last 30% of the tweaking, I have to do it." — P.Ga

Trust is overwhelmingly source-dependent: grounded, internal data earns reliance while open-web answers do not. P.So articulates the sharpest skeptical position — worrying that AI may be "pretending to guard you while it's serving the purpose of the organization" — and is also alert to confirmation bias in her own enthusiasm. The result is a portrait of professionals who have internalized that calibrated, partial trust is the mature posture.

3.3 Verification & Validation Practices

49 code applications · 6 of 6 participants · 5 sub-codes

The concrete behaviors, procedures, and habits participants enact to validate AI output before acting on it.

Verification is the behavior behind trust, and it is the most operationally detailed theme in the dataset. Manual review is the single most applied code in the entire study (22 applications, present in all six interviews), with domain knowledge serving as the primary human tool for judging whether AI output is actually correct.

Sub-code	Applications	Participants	Lead voices
Manual Review	22	6	P.Yi, P.Do
Domain Knowledge as Validator	15	5	P.Yi, P.Ga
Peer / Manager Review & Checks-and-Balances	6	3	P.So, P.Do
Cross-checking with Data / Logs	5	4	P.Yi, P.Do

Sub-code	Applications	Participants	Lead voices
Iterative Prompting / Re-querying	1	1	P.Shi

Illustrative voices

Manual Review

"Even if it's AI generated, I will double check what it has generated and go back and backtrack." — P.Ga

Domain Knowledge as Validator

"For work, I trust it more only when I can compare it with the real data, with the system, or with what I already know from the domain." — P.Yi

Peer / Manager Review & Checks-and-Balances

"My manager, I take it to my manager, just to overlook it… that's my checks and balances when it comes to something's arrived from AI." — P.Do

The verification repertoire is layered: individual manual review and domain-knowledge checks at the base, cross-checking against source data and logs for higher-stakes claims; and escalation to peers, managers, or a council when skepticism arises. Iterative re-prompting appears, but participants treat it as a refinement tactic rather than true validation; re-asking the model is not the same as verification.

3.4 Risks, Errors & Hallucinations

44 code applications · 6 of 6 participants · 7 sub-codes

Failure modes participants have observed or fear when AI is integrated into work: hallucination, fluff, destructive automation, over-reliance.

Participants are fluent in AI's failure modes and speak about them from experience. The risks they experience include: confidently wrong output (hallucination), padded low-signal content (fluff), and destructive automated actions combined with the human tendency toward over-reliance.

Sub-code	Applications	Participants	Lead voices
Risk to Production / Critical Decisions	11	5	P.Yi, P.Ga
Hallucination	10	4	P.Ga, P.So
Over-reliance / Blind Trust	9	2	P.So, P.Do
Fluff & Irrelevant Output	7	3	P.Do, P.So
Operational Incident (Destructive Action)	3	3	P.Do, P.Shi
Web Search Unreliability	3	2	P.Pa, P.Ga
AI Looping/Stubbornness	1	1	P.Ga

Illustrative voices

Risk to Production / Critical Decisions

"If the result could have production impact, then we definitely need human review." — P.Yi

Hallucination

"It will just hallucinate. Even if it's an agency CLI, if I am asking for something which is a little bit tricky, it will give some dumb answers." — P.Ga

Over-reliance / Blind Trust

"They're just trusting AI. Like, I want to do this, and it gives a spits out an output and they just trust it. And they hit enter and it changes everything." — P.Do

The most vivid material here is P.Do's account of an agent rewriting files across a network share that had to be reverted, paired with his concern that less-technical users "just hit enter and it changes everything." Risk to production and critical decisions is the most widespread risk code (5 of 6 participants), establishing the stakes that make verification and human-in-the-loop non-negotiable for this group.

3.5 Governance, Policy & Guardrails

21 code applications · 5 of 6 participants · 5 sub-codes

Organizational controls, processes, councils, and tooling-level limits placed around AI use.

Formal organizational controls are present but noticeably less than others (21 applications), and concentrated in a few participants. Where governance appears, it takes the form of tool-level guardrails, permission-scoped data access, and, in larger organizations, a responsible-AI council that pre-approves tools and agents.

Sub-code	Applications	Participants	Lead voices
Enterprise / Tool-level Guardrails	6	3	P.Do, P.Ga
Responsible AI Council / Review	5	3	P.Ga, P.Do
Encouraged AI Adoption	5	3	P.Ga, P.Shi
Approved Tools vs Shadow AI	3	2	P.Yi, P.Ga
Access Controls & Data Permissioning	2	1	P.Ga

Illustrative voices

Enterprise / Tool-level Guardrails

"If they open up Claude and go security properties… I can govern that and see if they use any kind of… if they upload any PII." — P.Do

Responsible AI Council / Review

"We have a council, a responsible AI council… we cannot just go rogue and create some tools." — P.Ga

Approved Tools vs Shadow AI

"It's hard in that front to govern individual chats." — P.Do

A telling gap emerges between sanctioned tools and "shadow" individual chats: P.Do notes it is "hard… to govern individual chats." The relative sparsity of this theme, set against the density of individual verification behavior, suggests these professionals are doing the work of AI governance personally, improvising controls that their organizations have not yet fully formalized or that AI tools do not currently offer.

3.6 Training, Literacy & Capability

33 code applications · 6 of 6 participants · 5 sub-codes

How AI capability is developed in people — training, certifications, prompt skill, hands-on practice — and where literacy gaps cause friction.

Capability is framed as a moving target. Participants describe continuous upskilling, formal certifications and boot camps, and hands-on/need-driven learning. Prompt engineering is treated as a genuine skill that determines output quality. The dominant concern is other people's knowledge gap.

Sub-code	Applications	Participants	Lead voices
End-user Knowledge Gap / Misuse	10	3	P.Shi, P.Do
Formal Training & Certifications	8	4	P.Ga, P.Do
Prompt Engineering Skill	7	5	P.Yi, P.Do
Continuous Upskilling	4	2	P.Ga, P.So
Hands-on / Need-driven Learning	4	4	P.So, P.Shi

Illustrative voices

End-user Knowledge Gap / Misuse

"Claude responses are letting everybody think they're a technician… I'm answering a lot of tickets… for people who I know are not the most technical." — P.Do

Formal Training & Certifications

"I did some certifications on AI and business processes… boot camps at Microsoft." — P.Ga

Prompt Engineering Skill

"My prompt is incorrect. It's confusing the model. I will improve my prompt and give more context." — P.Ga

The end-user knowledge gap is the heaviest code in this theme (10 applications), voiced especially by P.Do and P.Shi: AI lets non-technical users, non-MDMS, believe they are technicians, generating misdirected tickets and demands. Literacy, in other words, is discussed as an organizational risk surface, linking this theme directly to over-reliance and governance.

3.7 Human Judgment & Decision Authority

49 code applications · 6 of 6 participants · 6 sub-codes

The continuing role of human judgment in AI-assisted decisions: who decides, who pushes back, and how time pressure and seniority shape the answer.

Across every interview, humans remain the decision-makers. "Human-in-the-loop" is asserted by all six participants, and they describe actively challenging and pushing back on AI, sometimes refusing its recommendation outright. Decision authority is assigned to upper-level human-based roles, not in the tools.

Sub-code	Applications	Participants	Lead voices
Human-in-the-Loop	15	6	P.So, P.Yi
Challenging / Pushing Back on AI	12	4	P.Do, P.So
Decision Hierarchy & Authority	11	6	P.Yi, P.So
Time Pressure & Deadlines	6	4	P.Yi, P.So
Ethical Judgment	3	2	P.Do, P.Ga
Future of Work / Job Displacement	2	2	P.Shi, P.Do

Illustrative voices

Human-in-the-Loop

"There should be a human in the loop all the time. Even if it's AI generated, I will double check." — P.Ga

Challenging / Pushing Back on AI

"I have to stop it in the tracks and ask, why did you make that assumption?… why did you not double-check before you could send it to me?" — P.So

Decision Hierarchy & Authority

"In a matrix or a hybrid organization, it would definitely be someone senior… the sponsor, or a functional manager." — P.So

Two contextual modifiers shape how this plays out. Time pressure changes the risk calculus — P.So's "if you want to get fired, walk into that room and present" an unverified 15-minute report captures how deadlines raise, not lower, the verification bar for consequential outputs. And ethical judgment surfaces as a human-only function: resisting requests to cherry-pick or "fluff" results is framed as a responsibility of the human reviewing party.

3.8 Privacy, Ethics & Compliance

21 code applications · 4 of 6 participants · 4 sub-codes

Broader ethical, legal, regulatory, and societal concerns about AI in the workplace: data protection, surveillance, vendor asymmetries, and societal stakes.

The broadest ethical and legal concerns appear in four of six interviews and are led by participants in regulated or client-driven settings. The concrete worry is data: what PII or confidential project information should never be put into a tool. Beyond that sits regulatory liability, surveillance and profiling fears, and power imbalances between vendors and clients.

Sub-code	Applications	Participants	Lead voices
PII / Sensitive Data Handling	7	3	P.Yi, P.So
Regulatory / Legal Compliance	6	3	P.So, P.Yi
Data Profiling / Surveillance	4	2	P.So, P.Do
Vendor / Client Power Asymmetry	4	1	P.So

Illustrative voices

PII / Sensitive Data Handling

"It is also useful to learn what information we should or should not put into AI tools, especially if the data is related to company projects." — P.Yi

Regulatory / Legal Compliance

"I work in the BFSI sector, so BFSI is highly regulated, so the PMO is very supportive, and the client PMOs are extremely… directive, supervisory." — P.So

Data Profiling / Surveillance

"I'm running the risk of handing over my data to AI, where it's able to profile me with a unique marker." — P.So

P.So carries this theme most heavily, including the only applications of vendor/client power asymmetry — the sense that "the client calls the shots" and that disputable AI-driven metrics cannot easily be countered. The theme's concentration in regulated contexts suggests privacy and compliance pressure is unevenly distributed: acute for some participants, peripheral for others.

4. Cross-Cutting Findings

4.1 The trust–verify–judge spine

The clearest structural finding comes from how themes co-occur within the same coded passages. When participants talk about one of these ideas, they tend to talk about the others in the same breath. The strongest pairings are:

Theme A	Theme B	Co-occurrences
Verification & Validation	Human Judgment & Authority	16
Trust & Calibrated Reliance	Human Judgment & Authority	12
Verification & Validation	Risks, Errors & Hallucinations	8
Risks, Errors & Hallucinations	Human Judgment & Authority	8

Theme A	Theme B	Co-occurrences
Trust & Calibrated Reliance	Verification & Validation	8
Trust & Calibrated Reliance	Risks, Errors & Hallucinations	6

Read together, these pairings describe a single mental model rather than four separate themes. Participants perceive risk, respond with verification, calibrate their trust accordingly, and reserve the final decision for a human. Verification and human judgment are the two most linked themes in the entire dataset. This is the central contribution of the analysis: workplace AI is governed less by policy than by an internalized, individually enacted "trust-but-verify, human-in-the-loop" discipline.

4.2 Provenance as the trust switch

A second pattern cuts across the trust, verification, and risk themes: trust toggles on data provenance. Grounded internal data (RAG systems, Work IQ, known databases) is trusted; the open web is not. P.Ga's observation that the model "goes crazy" once it leaves grounded internal data and searches the web is echoed in source-dependent trust, cross-checking practices, and web-search-unreliability concerns. Provenance, not the model brand, is the primary input to how much these professionals rely on an answer.

4.3 The literacy–over-reliance loop

A third pattern links training to risk. The end-user knowledge gap and over-reliance / blind trust codes describe two ends of the same problem: as AI lowers the apparent barrier to technical work, less-skilled users both over-trust outputs and generate misdirected demands on technical staff. Participants who carry the verification and governance themes most heavily (P.Do, P.So) are also those most worried about other people's literacy. This frames AI literacy as a collective risk control rather than personal development.

5. Implications

Three implications follow from the analysis. These are considerations rather than prescriptions given the sample size.

Formalize what individuals are already improvising

The gap between dense individual verification behavior and sparse formal governance suggests organizations could capture value by codifying the verification habits their best people already practice, turning personal "trust-but-verify" routines into shared checklists, especially for production-impacting and business critical operations.

Treat provenance as a first-class signal

Because trust toggles on data source, tooling and training that define provenance (clearly flagging grounded vs. open-web answers) aligns with how professionals already calibrate reliance and would target ungrounded web answers that participants most distrust.

Invest in end-user literacy as risk control

The literacy–over-reliance loop implies that the highest-leverage training is not for power users but for the broad base of non-technical users whose over-trust creates downstream risk and support burden.

Guardrails that default to “plan / confirm” rather than “allow all” address the same concern at the tool level.

6. Limitations

This analysis rests on six interviews coded by the project team, so prevalence counts describe this dataset, not a population. Coding density is uneven across participants, and a single rich interview can inflate a code count; participant-breadth figures are reported throughout to mitigate this. Six of the 45 sub-codes carry no formal definition or anchor quote in the exported codebook, indicating emergent codes; they contribute little to the totals and none to the headline findings.

Beyond sample size, five deeper tensions surfaced that constrain how confidently we can interpret our findings.

Secondary research suggests expertise amplifies confirmation bias, yet our participants showed the opposite, raising questions about whether our sample skews toward unusually self-aware practitioners.

Our quantitative data identifies formal validation as the strongest risk reducer, yet every qualitative participant relied primarily on informal manual review, suggesting our findings describe an ideal rather than current practice.

Bias awareness emerges as our strongest protective factor statistically, yet only 2 of 6 participants raised confirmation bias unprompted, limiting how broadly that finding generalizes.

Time pressure scored as the highest-risk condition quantitatively, yet qualitative participants described deadlines as sharpening their verification instinct - a contradiction that may reflect seniority effects our sample cannot fully untangle.

Finally, key secondary sources including Hajikhani and Cole (2024) were built on GPT-3.5 behavior, now two model generations old**,** meaning the literature base itself is moving faster than our ability to cite it reliably.

Appendix A.X. Full Codebook and Frequencies

All themes and sub-codes with application counts and participant breadth, as exported from LibreQDA.

1. AI Use Patterns & Adoption (55 applications, 6/6)

Sub-code	n	Definition
Daily / Routine Use	11	Statements describing AI being woven into daily, near-continuous work activity.
Internal vs Public AI Tools	9	Distinguishing company-approved/internal AI tools from public/consumer models, and the safety implications of that boundary.
Productivity & Task Automation	9	AI used to accelerate, automate, or scale specific deliverables (drafting, summarising, code review, dashboards).

Sub-code	n	Definition
AI-Assisted Content Creation	8	— (emergent code, no formal definition)
Tool Ecosystem (Copilot, Claude, CLI, MCP, RAG)	6	Specific tools, platforms or architectures named or described as part of the AI stack in use.
Data synthesis via AI	5	— (emergent code, no formal definition)
Limited / Restricted Adoption	4	Cases where the participant's AI use is below what they'd like — due to client restriction, sector regulation, licensing, or maturity gap.
Prompt Sensitivity	3	— (emergent code, no formal definition)

2. Trust & Calibrated Reliance (41 applications, 6/6)

Sub-code	n	Definition
Trust-but-Verify Stance	11	An explicit or implicit doctrine that AI may be used, but its output must be checked before action.
Skepticism & Distrust	10	Active doubt, suspicion, or refusal to defer to AI — sometimes grounded in concerns about provider motives, bias, or opacity.
Partial / Calibrated Trust	9	Trust expressed as fractional or numerical — e.g., 70% AI, 30% human review; 3.5/5 confidence.
Source-Dependent Trust	6	Trust is contingent on data provenance — grounded internal data is trusted more than open-web answers.
Confirmation Bias	5	Trust elevated when AI confirms what the participant already believed; awareness of the bias risk this creates.

3. Verification & Validation Practices (49 applications, 6/6)

Sub-code	n	Definition
Manual Review	22	The participant personally reads, rewrites, or audits AI output before use.
Domain Knowledge as Validator	15	Participant's own expertise is the primary mechanism by which AI output is judged correct or wrong.

Sub-code	n	Definition
Peer / Manager Review & Checks-and-Balances	6	Escalation to a manager, peer, data engineer, or council for second-opinion on AI output.
Cross-checking with Data / Logs	5	Validation by reference to source data, system logs, SIEM, backups, or original documents.
Iterative Prompting / Re-querying	1	Validation by re-asking the AI — with different prompts, more context, or different models — and comparing outputs.

4. Risks, Errors & Hallucinations (44 applications, 6/6)

Sub-code	n	Definition
Risk to Production / Critical Decisions	11	Concern that AI output could materially harm production systems, customers, regulators, or board-level decisions.
Hallucination	10	AI generating false, fabricated, or confidently-wrong content.
Over-reliance / Blind Trust	9	Users acting on AI output without scrutiny, often by less technical end-users — dependency that erodes judgment.
Fluff & Irrelevant Output	7	AI returns plausible-looking but low-signal, padded, or off-target content.
Operational Incident (Destructive Action)	3	A concrete near-miss or incident caused by AI — e.g., overwriting files, breaking workflows.
Web Search Unreliability	3	— (emergent code, no formal definition)
AI Looping/Stubbornness	1	— (emergent code, no formal definition)

5. Governance, Policy & Guardrails (21 applications, 5/6)

Sub-code	n	Definition
Enterprise / Tool-level Guardrails	6	Tool-side controls (enterprise plans, DLP, PII filters, allow/plan prompts) that limit risky actions.
Responsible AI Council / Review	5	Formal governance bodies that approve AI tools, agents, or architectures before deployment.
Encouraged AI Adoption	5	— (emergent code, no formal definition)

Sub-code	n	Definition
Approved Tools vs Shadow AI	3	Boundary between approved/internal tools and unsanctioned public-AI use; gap in governability of individual chats.
Access Controls & Data Permissioning	2	AI's reach is bounded by the user's data permissions; access scoped to authorised resources only.

6. Training, Literacy & Capability (33 applications, 6/6)

Sub-code	n	Definition
End-user Knowledge Gap / Misuse	10	Non-technical users misjudging their competence after using AI; raising tickets or building things they shouldn't.
Formal Training & Certifications	8	Organised programs: certifications, boot camps, L1/L2 generative AI courses, vendor curricula.
Prompt Engineering Skill	7	Skill at framing prompts, providing context, and iterating phrasing to improve output.
Continuous Upskilling	4	Recognition that AI literacy is a continuous moving target due to rapid model and threat-vector change.
Hands-on / Need-driven Learning	4	Learning by doing, driven by an immediate task need rather than structured curriculum.

7. Human Judgment & Decision Authority (49 applications, 6/6)

Sub-code	n	Definition
Human-in-the-Loop	15	Explicit affirmation that a human must remain in the decision loop, particularly for consequential decisions.
Challenging / Pushing Back on AI	12	Actively contesting AI output, asking it to justify, or refusing its recommendation.
Decision Hierarchy & Authority	11	Who actually owns the decision — PMs, functional managers, C-suite, council — once AI output is in hand.
Time Pressure & Deadlines	6	How deadlines change the risk calculus of using un-validated AI output.
Ethical Judgment	3	Participant exercising ethical reasoning to resist a request that would misuse AI output (e.g., cherry-picking).

Sub-code	n	Definition
Future of Work / Job Displacement	2	Speculation about how AI/agents change roles, hiring, and the human's place in the workflow.

8. Privacy, Ethics & Compliance (21 applications, 4/6)

Sub-code	n	Definition
PII / Sensitive Data Handling	7	Concern about uploading or exposing personally identifiable or confidential company data to AI tools.
Regulatory / Legal Compliance	6	Compliance with sector regulation (BFSI), legal liability for AI-generated content, or formal compliance review.
Data Profiling / Surveillance	4	Concern that AI tools profile users, that vendors hold leverage over user data, or that surveillance creep occurs.
Vendor / Client Power Asymmetry	4	Imbalances between vendors and clients (or providers and users) in setting the rules of AI use.

Appendix B. Thematic Analysis Dashboard

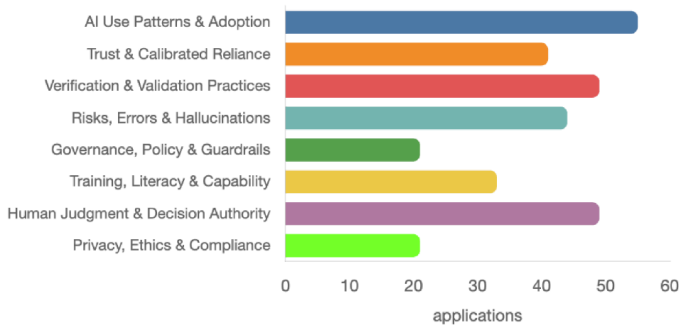
Source file: Thematic Analysis Dashboard.pdf

Thematic Analysis of Interviews — Dashboard

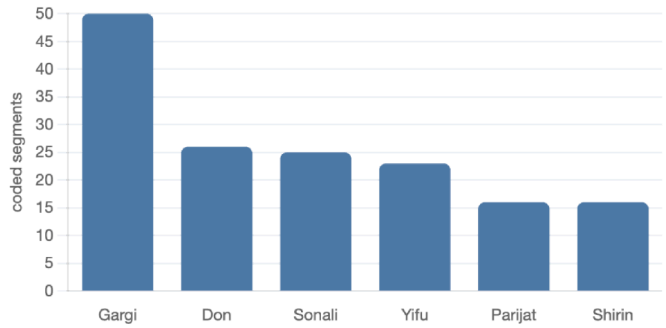
Workplace AI adoption · 6 interviews coded in LibreQDA · 8 themes · 45 sub-codes · 156 coded segments · 314 code applications

6	8
PARTICIPANTS	THEMES
45	**156**
SUB-CODES	CODED SEGMENTS
314	
APPLICATIONS	

Theme prevalence (code applications)



Coded segments per participant



Explore a theme

AI Use Patterns & Adoption | Trust & Calibrated Reliance | Verification & Validation Practices | Risks, Errors & Hallucinations | Governance, Policy & Guardrails | Training, Literacy & Capability | Human Judgment & Decision Authority | Privacy, Ethics & Compliance

How participants and their organizations currently use AI and AI-powered analytics in work tasks; scope, depth, and integration of AI into daily workflows.

55 applications · 6 of 6 participants · 8 sub-codes

SUB-CODE	N	PARTICIPANTS	DEFINITION
Daily / Routine Use	11	5	Statements describing AI being woven into daily, near-continuous work activity.
Internal vs Public AI Tools	9	4	Distinguishing company-approved/internal AI tools from public/consumer models, and the safety implications of that boundary.
Productivity & Task Automation	9	6	AI used to accelerate, automate, or scale specific deliverables (drafting, summarising, code review, dashboards).
AI-Assisted Content Creation	8	4	—
Tool Ecosystem (Copilot, Claude, CLI, MCP, RAG)	6	4	Specific tools, platforms or architectures named or described as part of the AI stack in use.
Data synthesis via AI	5	4	—

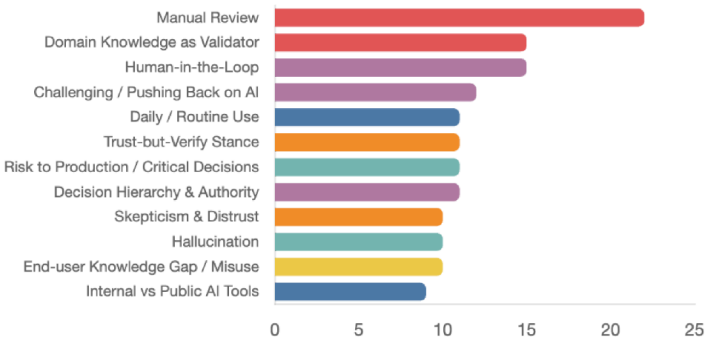
SUB-CODE	N	PARTICIPANTS	DEFINITION
Limited / Restricted Adoption	4	2	Cases where the participant's AI use is below what they'd like — due to client restriction, sector regulation, licensing, or maturity gap.
Prompt Sensitivity	3	2	—

Theme co-occurrence (same coded passage)

THEME PAIR	N
Verification & Validation ↔ Human Judgment & Authority	16
Trust & Calibrated Reliance ↔ Human Judgment & Authority	12
Verification & Validation ↔ Risks, Errors & Hallucinations	8
Risks, Errors & Hallucinations ↔ Human Judgment & Authority	8
Trust & Calibrated Reliance ↔ Verification & Validation	8
Trust & Calibrated Reliance ↔ Risks, Errors & Hallucinations	6
AI Use Patterns & Adoption ↔ Verification & Validation	5
Training, Literacy & Capability ↔ Human Judgment & Authority	4

How often two themes were applied to the same passage. The Verification–Human Judgment–Trust cluster is the analytic spine of the dataset.

Top 12 codes overall



Theme × participant coverage

THEME	GARGI	DON	SONALI	YIFU	PARIJAT	SHIRIN	TOTAL
AI Use Patterns & Adoption	18	4	5	9	9	10	**55**
Trust & Calibrated Reliance	11	5	12	6	4	3	**41**
Verification & Validation Practices	6	9	7	15	7	5	**49**
Risks, Errors & Hallucinations	10	10	13	5	5	1	**44**
Governance, Policy & Guardrails	10	4	1	3		3	**21**
Training, Literacy & Capability	8	8	5	5	1	6	**33**
Human Judgment & Decision Authority	8	10	15	8	4	4	**49**
Privacy, Ethics & Compliance	1	2	13	5			**21**

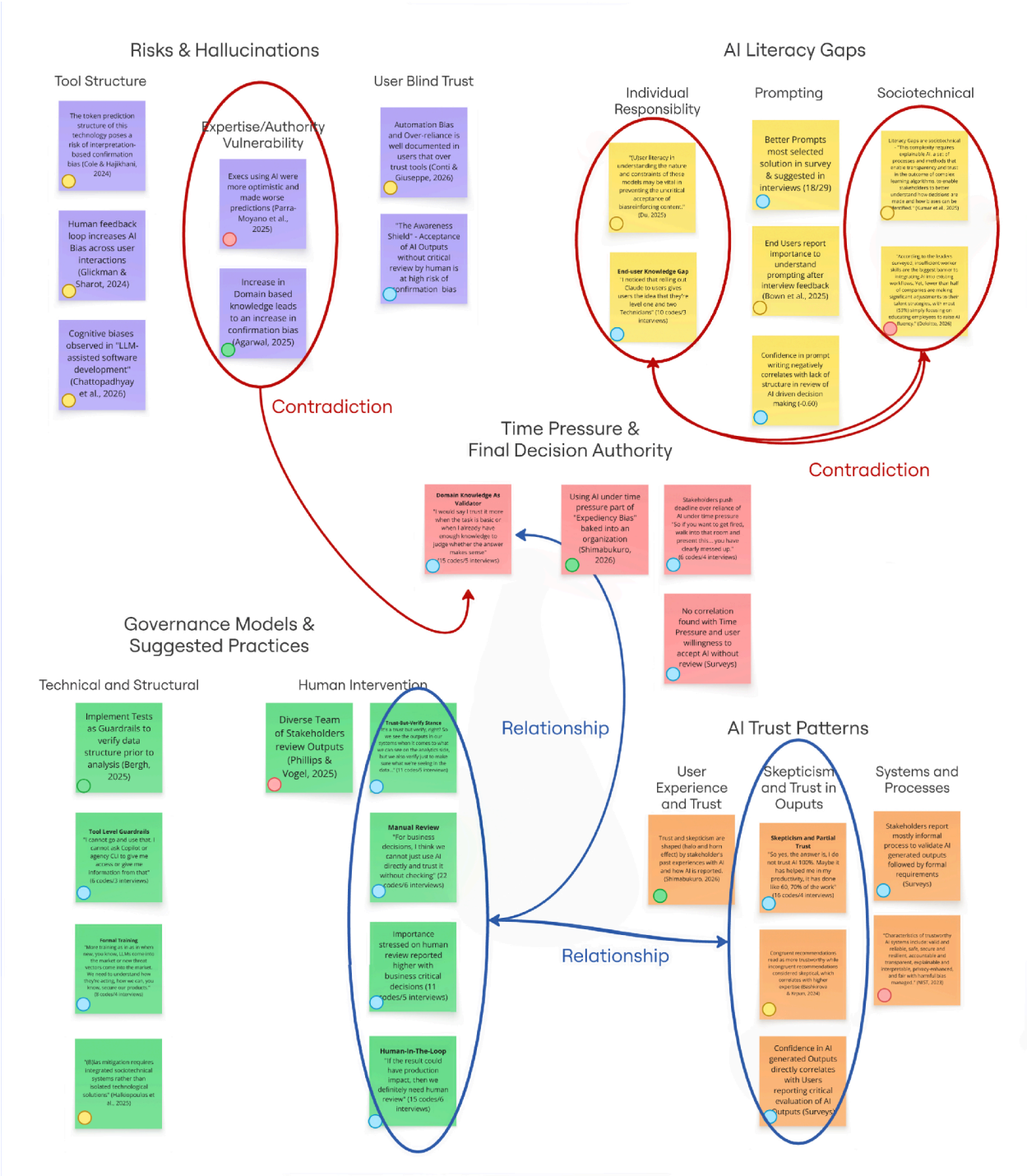
Coded segments per theme for each participant. Darker = more coded material.

Generated from 2026-06-01_Final Project Thematic Analysis of Interviews.sqlite3 · IMT 570 Final Project

Elliot Forst, Sunayana Hazarika, Anjuta K, Kuntala Sarkar, Bhagyashree Vaidya, Harsh Vardhan, Della Zhang, Hritvik Gaur
IMT 570 E: Data Driven Organizational Problem Solving For Information Management Professionals
Group Project: Final Research Report
06/09/2026

Appendix C. Affinity Map

Source file: Affinity_Map.png



Affinity map of coded interview data

Elliot Forst, Sunayana Hazarika, Anjuta K, Kuntala Sarkar, Bhagyashree Vaidya, Harsh Vardhan, Della Zhang, Hritvik Gaur
IMT 570 E: Data Driven Organizational Problem Solving For Information Management Professionals
Group Project: Final Research Report
06/09/2026

Appendix D. Goals Analysis

Source file: goals_analysis.pdf

Confirmation Bias in LLM-Based AI Analytics Tools

Refined Visualizations — Final Version

Finding 1: The Awareness Shield
Finding 2: The Validation Vacuum

In [1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import matplotlib.patches as mpatches
import matplotlib.lines as mlines
from scipy import stats

plt.rcParams.update({
    'figure.facecolor': 'white',
    'axes.facecolor': '#FAFAF8',
    'axes.spines.top': False,
    'axes.spines.right': False,
    'axes.grid': True,
    'grid.alpha': 0.35,
    'grid.linestyle': '--',
    'font.family': 'DejaVu Sans',
    'font.size': 11
})

# Update this path to match your local file location
# Windows example:  r'C:\Users\yourname\Documents\assessment_cleaned.csv'
# Mac/Linux example: '/Users/yourname/Documents/assessment_cleaned.csv'
df = pd.read_csv(r'C:\Users\anjut\OneDrive\Documents\MSIM\IMT 570\group project\Ass

# Shared color palette
BAR_COLORS  = ['#E24B4A', '#EF9F27', '#888780', '#1D9E75', '#085041']
LBL_COLORS  = ['#791F1F', '#854F0B', '#5F5E5A', '#085041', '#04342C']
CORAL       = '#D85A30'
TEAL        = '#1D9E75'
PURPLE_DASH = '#534AB7'

print('Data loaded:', df.shape)
df.head(3)
```

```
Data loaded: (29, 24)
```

Out[1]:

	Unnamed: 0	workplace	role	Years	primary use	hrs per week	
0	1	Enterprise	Engineering Manager	14+	Documentation;Decision support (strategy/plann...	15	Underlyir (dashboards/r
1	2	Product Company	TestManager	20	Documentation;Decision support (strategy/plann...	5	Both
2	3	Enterprise	Technology Analyst	4	Documentation;Reporting / dashboards;Programmi...	10	Both

3 rows × 24 columns

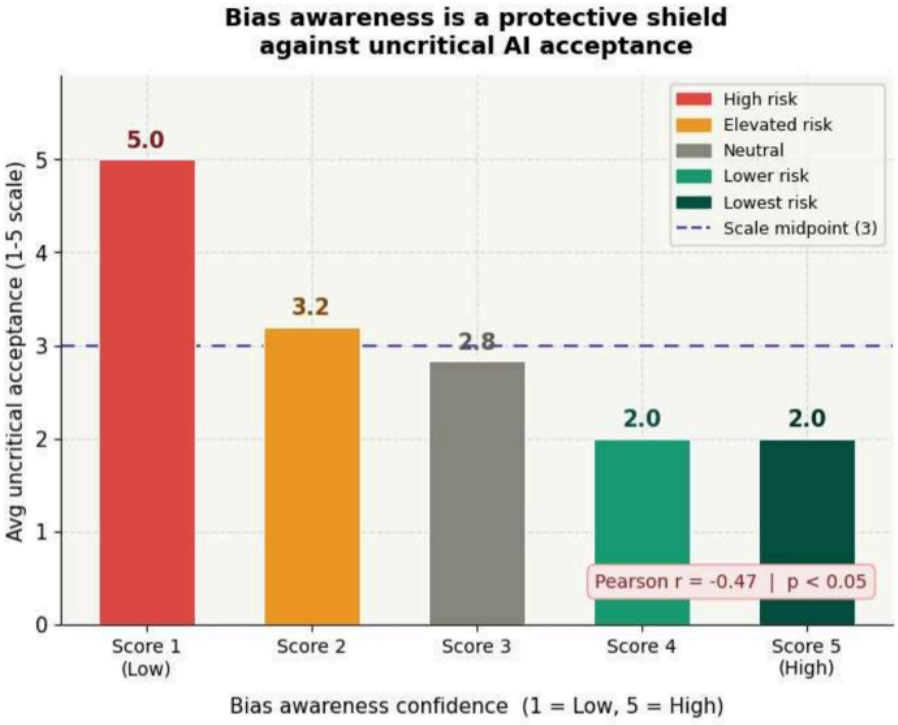
Visualization 1 — The Awareness Shield

> **Finding (Goal 1 — AI Literacy Gap):** Higher bias awareness confidence strongly predicts lower uncritical AI acceptance (Pearson r = −0.466, p < 0.05).
>
> **Chart type:** Vertical bar chart — appropriate because it encodes mean values by length across five ordered groups, enabling precise comparison.
>
> **Design choices:**
>
> - Bar colors encode risk level (red → teal) so color reinforces the takeaway
> - Dashed midpoint line at y=3 provides a neutral reference anchor
> - Pearson r badge positioned at lower right where bars are short (ample whitespace)
> - Footnote placed via `fig.text()` to prevent overlap with the x-axis label

In [2]:
```
fig1, ax1 = plt.subplots(figsize=(8, 6))

# Extra bottom margin prevents footnote/xlabel collision
fig1.subplots_adjust(bottom=0.20, top=0.88)

# ── Data ──────────────────────────────────────────────────────────
grouped = df.groupby('confident output bias')['acceptance level'].mean()
x_pos   = np.arange(len(grouped))
```

```
# ── Bars ──────────────────────────────────────────────────────────
bars = ax1.bar(x_pos, grouped.values, color=BAR_COLORS,
               width=0.58, edgecolor='white', linewidth=0.8, zorder=3)

# ── Scale midpoint reference line ─────────────────────────────────
ax1.axhline(y=3, color=PURPLE_DASH, linestyle=(0, (5, 4)),
            linewidth=1.5, zorder=2)

# ── Data labels on bars ───────────────────────────────────────────
for bar, val, lc in zip(bars, grouped.values, LBL_COLORS):
    ax1.text(bar.get_x() + bar.get_width() / 2,
             val + 0.07, f'{val:.1f}',
             ha='center', va='bottom', fontsize=12,
             fontweight='bold', color=lc)

# ── Axes ──────────────────────────────────────────────────────────
ax1.set_xticks(x_pos)
ax1.set_xticklabels(
    ['Score 1\n(Low)', 'Score 2', 'Score 3', 'Score 4', 'Score 5\n(High)'],
    fontsize=10)
ax1.set_ylim(0, 5.9)
ax1.set_yticks([0, 1, 2, 3, 4, 5])
ax1.set_xlabel('Bias awareness confidence  (1 = Low, 5 = High)',
               fontsize=11, labelpad=10)
ax1.set_ylabel('Avg uncritical acceptance (1-5 scale)', fontsize=11)
ax1.set_title(
    'Bias awareness is a protective shield\nagainst uncritical AI acceptance',
    fontsize=13, fontweight='bold', pad=12)

# ── Pearson r badge — lower right (bars are short here, ample space) ──
r_val, p_val = stats.pearsonr(df['confident output bias'], df['acceptance level'])
ax1.text(0.97, 0.06,
         f'Pearson r = {r_val:.2f}  |  p < 0.05',
         transform=ax1.transAxes, ha='right', va='bottom', fontsize=10,
         color='#791F1F',
         bbox=dict(boxstyle='round,pad=0.4', facecolor='#FCEBEB',
                   edgecolor='#F09595', linewidth=0.8))

# ── Footnote via fig.text() — sits cleanly below xlabel, no overlap ──
fig1.text(0.12, 0.02,
          '* n=1 for score 1; treat as directional only',
          fontsize=9, color='#888780', ha='left', va='bottom')

# ── Legend ────────────────────────────────────────────────────────
legend_patches = [
    mpatches.Patch(color='#E24B4A', label='High risk'),
    mpatches.Patch(color='#EF9F27', label='Elevated risk'),
    mpatches.Patch(color='#888780', label='Neutral'),
    mpatches.Patch(color='#1D9E75', label='Lower risk'),
    mpatches.Patch(color='#085041', label='Lowest risk'),
    mlines.Line2D([], [], color=PURPLE_DASH, linestyle='--',
                  linewidth=1.5, label='Scale midpoint (3)'),
]
ax1.legend(handles=legend_patches, loc='upper right', fontsize=9,
           framealpha=0.9, edgecolor='#D3D1C7')
```

```
plt.savefig('viz1_awareness_shield.png', dpi=180,
            bbox_inches='tight', facecolor='white')
plt.show()
print(f'Pearson r = {r_val:.3f},  p = {p_val:.4f}')
print('Mean acceptance by bias awareness score:')
print(grouped.round(2))
```



* n=1 for score 1; treat as directional only

```
Pearson r = -0.466,  p = 0.0108
Mean acceptance by bias awareness score:
confident output bias
1    5.00
2    3.20
3    2.83
4    2.00
5    2.00
Name: acceptance level, dtype: float64
```

Visualization 2 — The Validation Vacuum

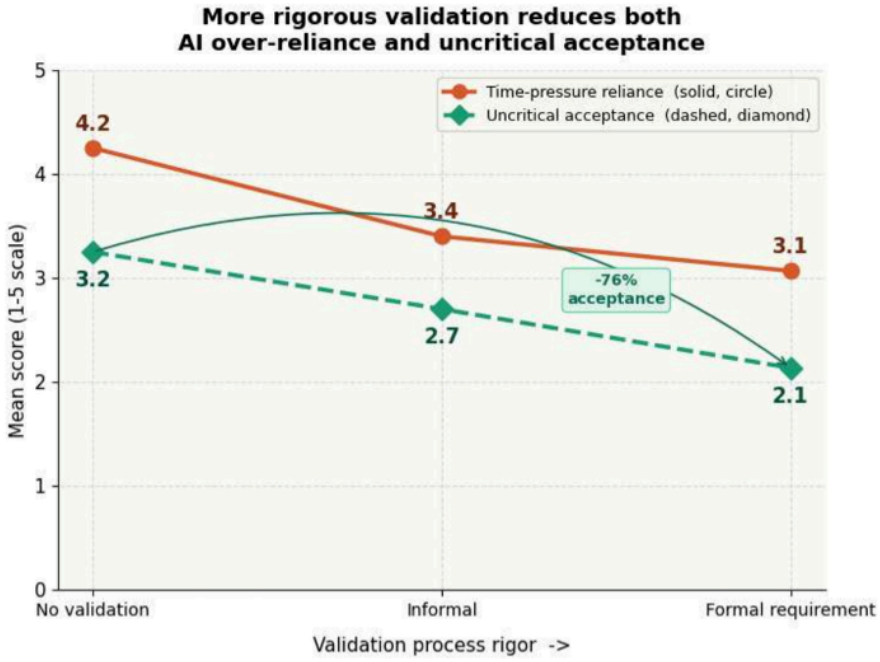
> **Finding (Goal 2 — AI Governance):** Both risk metrics decline as validation rigor increases. Formal validation cuts uncritical acceptance by 76% vs no validation.

> **Chart type:** Slope chart (line chart across ordered groups) — appropriate because the three groups are ordered by rigor, and the slope visually encodes the direction and magnitude of change as governance increases.
>
> **Design choices:**
>
> - Two distinct visual cues beyond color: solid+circle vs dashed+diamond
> - Data labels above (time pressure) and below (acceptance) avoid overlap
> - Group sizes (n) shown below x-axis ticks for sample transparency
> - Curved annotation arrow highlights the 76% acceptance reduction

In [3]:

```
fig2, ax2 = plt.subplots(figsize=(8, 6))
fig2.subplots_adjust(bottom=0.18, top=0.88)

# ── Data ──────────────────────────────────────────────────────────
val_order = ['No validation', 'Informal', 'Formal requirement']
val_map   = {
    'No':                        'No validation',
    'Sometimes (informal)':      'Informal',
    'Yes (formal requirement)': 'Formal requirement'
}
df['val_label'] = df['validation'].map(val_map)

grp   = (df.groupby('val_label')
           [['reliance on time pressure', 'acceptance level']]
           .mean()
           .reindex(val_order))
grp_n = df['val_label'].value_counts().reindex(val_order)

x_pos2   = np.arange(len(val_order))
tp_vals  = grp['reliance on time pressure'].values
acc_vals = grp['acceptance level'].values

# ── Slope lines ──────────────────────────────────────────────────
ax2.plot(x_pos2, tp_vals, color=CORAL, linewidth=2.5,
         marker='o', markersize=9, zorder=3)
ax2.plot(x_pos2, acc_vals, color=TEAL, linewidth=2.5,
         linestyle='--', marker='D', markersize=9, zorder=3)

# ── Data labels — time pressure above, acceptance below ──────────
for i, (tp, acc) in enumerate(zip(tp_vals, acc_vals)):
    ax2.text(i, tp + 0.13, f'{tp:.1f}',
             ha='center', va='bottom', fontsize=12,
             fontweight='bold', color='#712B13')
    ax2.text(i, acc - 0.16, f'{acc:.1f}',
             ha='center', va='top', fontsize=12,
             fontweight='bold', color='#085041')

# ── Group size (n) labels below x-axis ticks ─────────────────────
n_colors = ['#993C1D', '#5F5E5A', '#085041']
for i, (n, nc) in enumerate(zip(grp_n.values, n_colors)):
    ax2.text(i, -0.55, f'n = {n}', ha='center', fontsize=10,
             color=nc, fontweight='bold',
```

```
            transform=ax2.get_xaxis_transform())

# ── Axes ──────────────────────────────────────────────────────────
ax2.set_xticks(x_pos2)
ax2.set_xticklabels(val_order, fontsize=10)
ax2.set_ylim(0, 5)
ax2.set_yticks([0, 1, 2, 3, 4, 5])
ax2.set_xlabel('Validation process rigor  ->', fontsize=11, labelpad=10)
ax2.set_ylabel('Mean score (1-5 scale)', fontsize=11)
ax2.set_title(
    'More rigorous validation reduces both\nAI over-reliance and uncritical accepta
    fontsize=13, fontweight='bold', pad=12)

# ── 76% reduction annotation arrow ────────────────────────────────
ax2.annotate('', xy=(2, acc_vals[2]), xytext=(0, acc_vals[0]),
             arrowprops=dict(arrowstyle='->', color='#0F6E56',
                             lw=1.2, connectionstyle='arc3,rad=-0.25'))
ax2.text(1.5, 2.75, '-76%\nacceptance', ha='center', fontsize=9,
         color='#0F6E56', fontweight='bold',
         bbox=dict(boxstyle='round,pad=0.3', facecolor='#E1F5EE',
                   edgecolor='#5DCAA5', linewidth=0.8))

# ── Legend ────────────────────────────────────────────────────────
legend_handles = [
    mlines.Line2D([], [], color=CORAL, linewidth=2.5,
                  marker='o', markersize=8,
                  label='Time-pressure reliance  (solid, circle)'),
    mlines.Line2D([], [], color=TEAL, linewidth=2.5,
                  linestyle='--', marker='D', markersize=8,
                  label='Uncritical acceptance  (dashed, diamond)'),
]
ax2.legend(handles=legend_handles, loc='upper right', fontsize=9,
           framealpha=0.9, edgecolor='#D3D1C7')

plt.savefig('viz2_validation_vacuum.png', dpi=180,
            bbox_inches='tight', facecolor='white')
plt.show()

pct_drop = (acc_vals[0] - acc_vals[2]) / acc_vals[0] * 100
print(f'Acceptance reduction (no validation -> formal): {pct_drop:.0f}%')
print(grp[['reliance on time pressure', 'acceptance level']].round(2))
```



Acceptance reduction (no validation -> formal): 34%

val_label	reliance on time pressure	acceptance level
No validation	4.25	3.25
Informal	3.40	2.70
Formal requirement	3.07	2.13

Appendix E. Confirmation Bias Analysis

Source file: confirmation_bias_analysis.pdf

Confirmation Bias in LLM-Based AI Analytics Tools

Quantitative Survey Analysis (n = 29)

Research Goals:

- **Goal 1:** Identify AI literacy gaps in Middle Decision Makers (MDMs)
- **Goal 2:** Develop AI governance model for bias mitigation

---

0. Setup & Imports

In [1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import matplotlib.patches as mpatches
import seaborn as sns
from scipy import stats
from collections import Counter
import warnings
warnings.filterwarnings('ignore')

# Plot style
plt.rcParams.update({
    'figure.facecolor': 'white',
    'axes.facecolor': '#F8F8F6',
    'axes.spines.top': False,
    'axes.spines.right': False,
    'axes.grid': True,
    'grid.alpha': 0.4,
    'font.family': 'DejaVu Sans',
    'font.size': 11
})

# Color palette aligned to findings
TEAL   = ['#9FE1CB', '#5DCAA5', '#1D9E75', '#0F6E56', '#085041']
CORAL  = ['#F5C4B3', '#F0997B', '#D85A30', '#993C1D', '#712B13']
PURPLE = ['#CECBF6', '#AFA9EC', '#7F77DD', '#534AB7', '#3C3489']
GRAY   = '#B4B2A9'
RED    = '#A32D2D'
```

1. Load & Inspect Data

In [2]:
```
df = pd.read_csv(r'C:\Users\anjut\OneDrive\Documents\MSIM\IMT 570\group project\Ass
print(f"Shape: {df.shape}  ({df.shape[0]} respondents, {df.shape[1]} columns)")
print("\nColumn names:")
```

```
print(df.columns.tolist())
df.head(3)
```

```
Shape: (29, 24)  (29 respondents, 24 columns)

Column names:
['Unnamed: 0', 'workplace', 'role', 'Years', 'primary use', 'hrs per week', 'rely',
'confident prompts', 'trust level', 'impact decision', 'confident output bias', 'acc
eptance level', 'reliance on time pressure', 'AI influence', 'output eval', 'output
incomp', 'org guidelines', 'review process', 'human importance', 'validation', 'revi
ew freq', 'bias concern', 'improve on decision-making', 'biggest challenge']
```

Out[2]:

	Unnamed: 0	workplace	role	Years	primary use	hrs per week	
0	1	Enterprise	Engineering Manager	14+	Documentation;Decision support (strategy/plann...	15	Underlyir (dashboards/n
1	2	Product Company	TestManager	20	Documentation;Decision support (strategy/plann...	5	Both
2	3	Enterprise	Technology Analyst	4	Documentation;Reporting / dashboards;Programmi...	10	Both

3 rows × 24 columns

In [3]:

```
# Identify numeric (Likert) vs categorical columns
LIKERT_COLS = [
    'confident prompts', 'trust level', 'impact decision',
    'confident output bias', 'acceptance level', 'reliance on time pressure',
    'AI influence', 'output eval', 'output incomp', 'org guidelines',
    'review process', 'human importance', 'review freq', 'bias concern'
]
CAT_COLS = ['workplace', 'role', 'Years', 'primary use', 'hrs per week',
            'rely', 'validation', 'improve on decision-making', 'biggest challenge'

print("Likert scale descriptives:")
df[LIKERT_COLS].describe().round(2)
```

```
Likert scale descriptives:
```

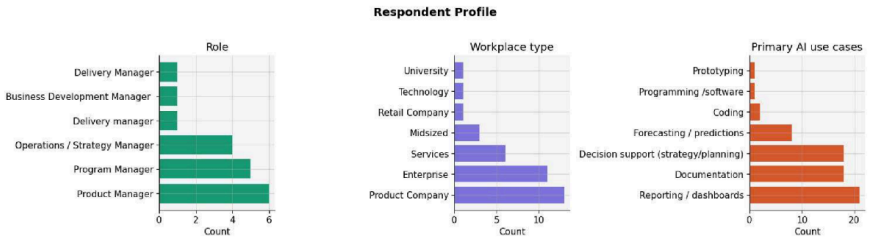
Out[3]:

	confident prompts	trust level	impact decision	confident output bias	acceptance level	reliance on time pressure	AI influence	output eval	out incc
count	29.00	29.00	29.00	29.00	29.00	29.00	29.00	29.00	2
mean	3.62	3.14	3.24	3.52	2.48	3.34	2.72	4.17	
std	0.86	0.79	1.15	1.09	1.35	1.29	1.07	0.97	
min	2.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	
25%	3.00	3.00	3.00	3.00	1.00	2.00	2.00	4.00	
50%	4.00	3.00	3.00	4.00	2.00	4.00	3.00	4.00	
75%	4.00	3.00	4.00	4.00	3.00	4.00	3.00	5.00	
max	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	

2. Respondent Profile

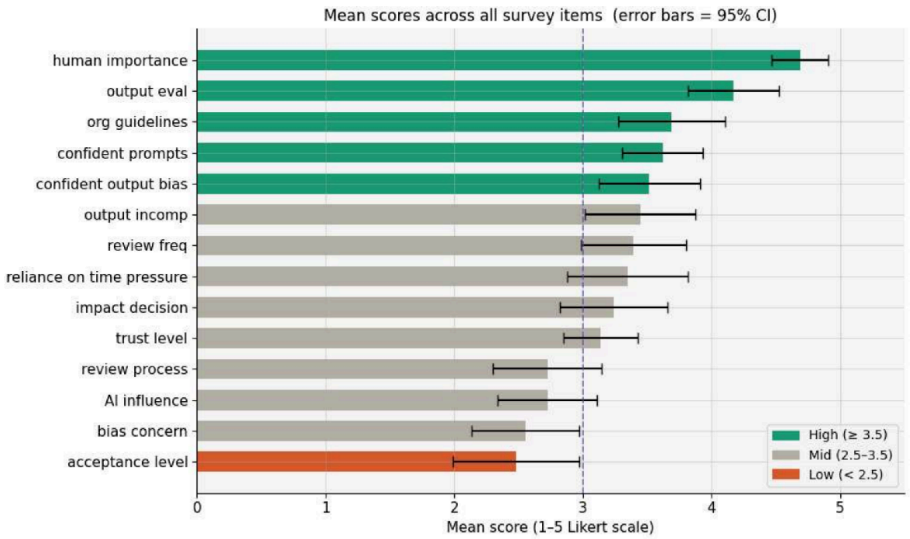
In [4]:

```
fig, axes = plt.subplots(1, 3, figsize=(15, 4))
fig.suptitle('Respondent Profile', fontsize=14, fontweight='bold', y=1.02)

# Role distribution
role_counts = df['role'].value_counts().head(6)
axes[0].barh(role_counts.index, role_counts.values, color=TEAL[2])
axes[0].set_title('Role')
axes[0].set_xlabel('Count')

# Workplace
# Normalize multi-select workplace entries
wp_flat = [w.strip() for entry in df['workplace'].dropna() for w in entry.split(';'
wp_counts = pd.Series(Counter(wp_flat)).sort_values(ascending=False)
axes[1].barh(wp_counts.index, wp_counts.values, color=PURPLE[2])
axes[1].set_title('Workplace type')
axes[1].set_xlabel('Count')

# Primary AI use cases (multi-select)
use_flat = [u.strip() for entry in df['primary use'].dropna() for u in entry.split(
            if u.strip() not in ['Forced to use.', 'Help with code analysis and red
use_counts = pd.Series(Counter(use_flat)).sort_values(ascending=False)
axes[2].barh(use_counts.index, use_counts.values, color=CORAL[2])
axes[2].set_title('Primary AI use cases')
axes[2].set_xlabel('Count')

plt.tight_layout()
plt.savefig('fig_01_respondent_profile.png', dpi=150, bbox_inches='tight')
plt.show()
```



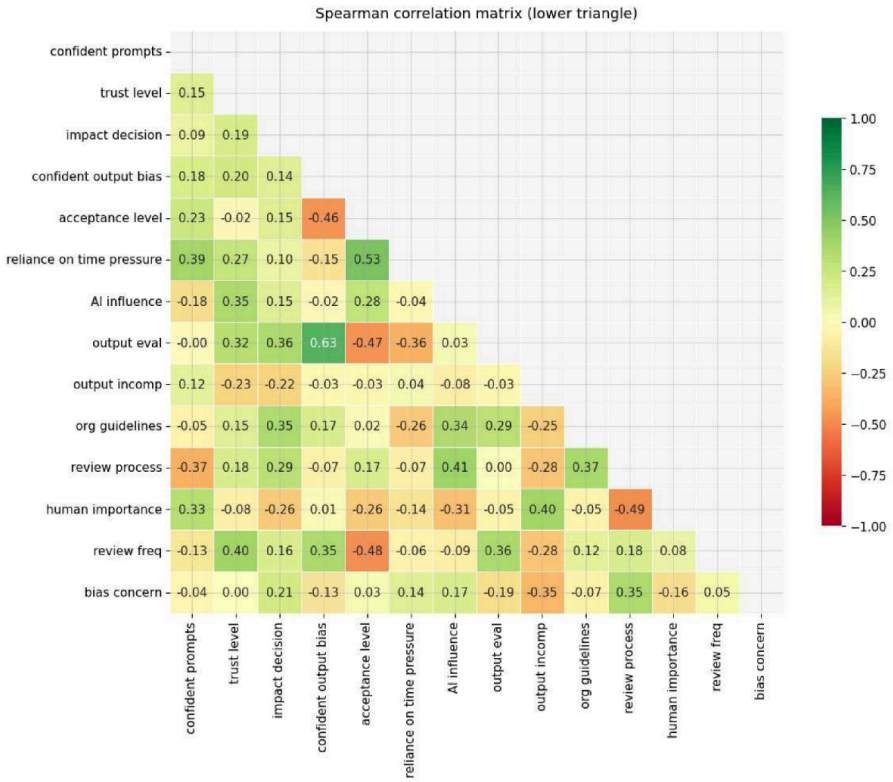
3. Descriptive Statistics — All Likert Items

In [5]:
```
# Mean scores with 95% CI
means = df[LIKERT_COLS].mean()
sems  = df[LIKERT_COLS].sem()
ci95  = sems * 1.96

# Sort for readability
order = means.sort_values(ascending=True)

fig, ax = plt.subplots(figsize=(10, 6))
colors = [CORAL[2] if v < 2.5 else (TEAL[2] if v >= 3.5 else GRAY) for v in order.v
bars = ax.barh(order.index, order.values, xerr=ci95[order.index],
               color=colors, capsize=4, height=0.65)
ax.axvline(x=3, color='#534AB7', linestyle='--', linewidth=1.2, alpha=0.7, label='M
ax.set_xlabel('Mean score (1–5 Likert scale)', fontsize=11)
ax.set_title('Mean scores across all survey items  (error bars = 95% CI)', fontsize
ax.set_xlim(0, 5.5)
legend_patches = [
    mpatches.Patch(color=TEAL[2],  label='High (≥ 3.5)'),
    mpatches.Patch(color=GRAY,     label='Mid (2.5–3.5)'),
    mpatches.Patch(color=CORAL[2], label='Low (< 2.5)'),
]
ax.legend(handles=legend_patches, loc='lower right', fontsize=10)
plt.tight_layout()
plt.savefig('fig_02_descriptive_means.png', dpi=150, bbox_inches='tight')
plt.show()
```



4. Correlation Matrix — Likert Items

In [6]:
```
corr = df[LIKERT_COLS].corr(method='spearman')   # Spearman for ordinal data

fig, ax = plt.subplots(figsize=(12, 10))
mask = np.triu(np.ones_like(corr, dtype=bool))    # Show lower triangle only
sns.heatmap(
    corr, mask=mask, annot=True, fmt='.2f',
    cmap='RdYlGn', center=0, vmin=-1, vmax=1,
    linewidths=0.5, linecolor='white',
    ax=ax, cbar_kws={'shrink': 0.7}
)
ax.set_title('Spearman correlation matrix (lower triangle)', fontsize=13, pad=12)
plt.tight_layout()
plt.savefig('fig_03_correlation_matrix.png', dpi=150, bbox_inches='tight')
plt.show()

# Top correlations with bias concern
print("\nTop correlations with 'bias concern':")
print(corr['bias concern'].drop('bias concern').sort_values().round(3))
```




```
Top correlations with 'bias concern':
output incomp               -0.354
output eval                 -0.191
human importance            -0.164
confident output bias       -0.134
org guidelines              -0.067
confident prompts           -0.042
trust level                  0.001
acceptance level             0.032
review freq                  0.052
reliance on time pressure    0.142
AI influence                 0.172
impact decision              0.209
review process               0.348
Name: bias concern, dtype: float64
```

5. GOAL 1 — AI Literacy Gaps

Finding A: The Awareness Shield

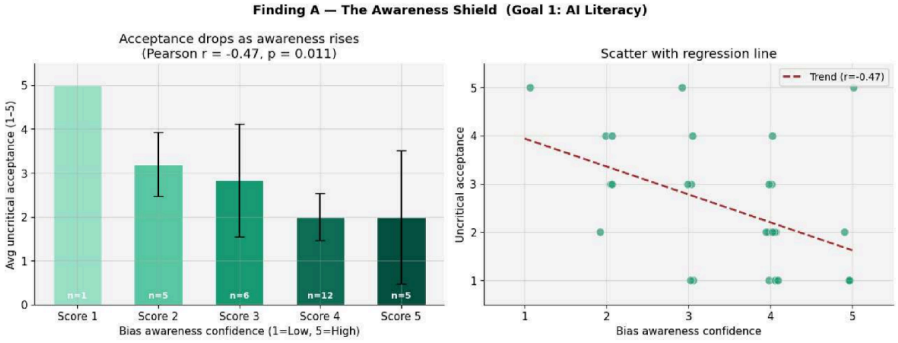
The more confident respondents are at spotting AI output bias, the less they accept AI outputs uncritically (r = −0.65).

In [7]:
```
# Pearson & Spearman correlation
r_p, p_p = stats.pearsonr(df['confident output bias'], df['acceptance level'])
r_s, p_s = stats.spearmanr(df['confident output bias'], df['acceptance level'])
print(f"Pearson  r = {r_p:.3f}  (p = {p_p:.4f})")
print(f"Spearman ρ = {r_s:.3f}  (p = {p_s:.4f})")

# Mean acceptance per bias-confidence score
grouped = df.groupby('confident output bias')['acceptance level'].agg(['mean','coun
grouped['ci95'] = grouped['sem'] * 1.96
print("\nMean acceptance by bias-awareness confidence:")
print(grouped.round(2))
```

```
Pearson  r = -0.466  (p = 0.0108)
Spearman ρ = -0.462  (p = 0.0117)

Mean acceptance by bias-awareness confidence:
                      mean  count   sem  ci95
confident output bias
1                     5.00      1   NaN   NaN
2                     3.20      5  0.37  0.73
3                     2.83      6  0.65  1.28
4                     2.00     12  0.28  0.54
5                     2.00      5  0.77  1.52
```

In [8]:
```
fig, axes = plt.subplots(1, 2, figsize=(13, 5))
fig.suptitle('Finding A — The Awareness Shield  (Goal 1: AI Literacy)', fontsize=13

# Left: Bar chart — avg acceptance by bias confidence
ax = axes[0]
x_labels = [f"Score {i}" for i in grouped.index]
bars = ax.bar(x_labels, grouped['mean'], color=TEAL[:len(grouped)],
              yerr=grouped['ci95'], capsize=5, width=0.6, edgecolor='white')
# Annotate n
for i, (bar, row) in enumerate(zip(bars, grouped.itertuples())):
    ax.text(bar.get_x() + bar.get_width()/2, 0.1,
            f'n={row.count}', ha='center', va='bottom', fontsize=9, color='white',
ax.set_ylim(0, 5.5)
ax.set_xlabel('Bias awareness confidence (1=Low, 5=High)')
ax.set_ylabel('Avg uncritical acceptance (1–5)')
ax.set_title(f'Acceptance drops as awareness rises\n(Pearson r = {r_p:.2f}, p = {p_

# Right: Scatter with regression line
ax2 = axes[1]
jitter = np.random.uniform(-0.1, 0.1, size=len(df))  # avoid overplotting
ax2.scatter(df['confident output bias'] + jitter, df['acceptance level'],
            color=TEAL[2], alpha=0.7, s=70, edgecolors='white', linewidth=0.8)
```

```
# Regression line
m, b = np.polyfit(df['confident output bias'], df['acceptance level'], 1)
x_line = np.linspace(1, 5, 100)
ax2.plot(x_line, m*x_line + b, color=RED, linewidth=2, linestyle='--', label=f'Tren
ax2.set_xlabel('Bias awareness confidence')
ax2.set_ylabel('Uncritical acceptance')
ax2.set_title('Scatter with regression line')
ax2.set_xlim(0.5, 5.5)
ax2.set_ylim(0.5, 5.5)
ax2.legend(fontsize=10)

plt.tight_layout()
plt.savefig('fig_04_awareness_shield.png', dpi=150, bbox_inches='tight')
plt.show()
```

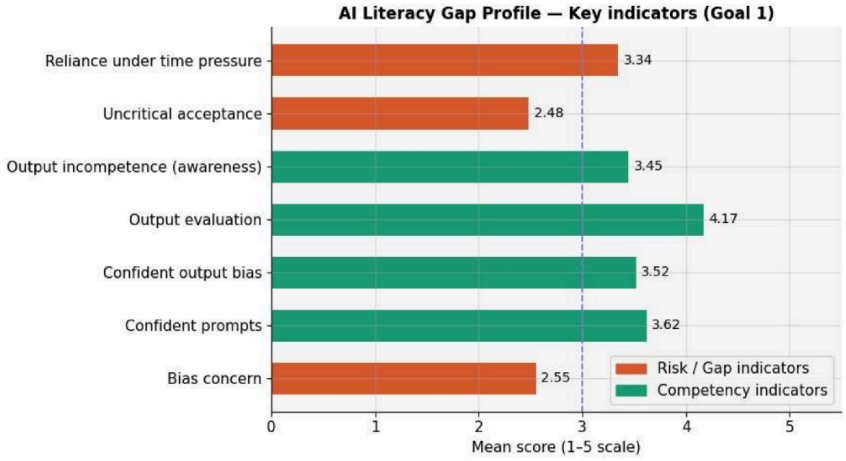


Literacy Gap Profile — Key Indicators

In [9]:

```
# Literacy gap indicators: low scores on bias concern, output eval vs high on accep
literacy_indicators = {
    'Bias concern': df['bias concern'].mean(),
    'Confident prompts': df['confident prompts'].mean(),
    'Confident output bias': df['confident output bias'].mean(),
    'Output evaluation': df['output eval'].mean(),
    'Output incompetence (awareness)': df['output incomp'].mean(),
    'Uncritical acceptance': df['acceptance level'].mean(),
    'Reliance under time pressure': df['reliance on time pressure'].mean(),
}

fig, ax = plt.subplots(figsize=(9, 5))
labels = list(literacy_indicators.keys())
values = list(literacy_indicators.values())
colors = [CORAL[2] if k in ['Uncritical acceptance','Bias concern','Reliance under
          else TEAL[2] for k in labels]

bars = ax.barh(labels, values, color=colors, height=0.6)
ax.axvline(3, color=PURPLE[2], linestyle='--', linewidth=1.2, label='Scale midpoint
for bar, val in zip(bars, values):
    ax.text(val + 0.05, bar.get_y() + bar.get_height()/2,
            f'{val:.2f}', va='center', fontsize=10)
ax.set_xlim(0, 5.5)
```

```
ax.set_xlabel('Mean score (1–5 scale)')
ax.set_title('AI Literacy Gap Profile — Key indicators (Goal 1)', fontsize=12, font
ax.legend()
legend_patches = [
    mpatches.Patch(color=CORAL[2], label='Risk / Gap indicators'),
    mpatches.Patch(color=TEAL[2], label='Competency indicators'),
]
ax.legend(handles=legend_patches, loc='lower right')

plt.tight_layout()
plt.savefig('fig_05_literacy_gap_profile.png', dpi=150, bbox_inches='tight')
plt.show()
```



6. GOAL 2 — AI Governance for Bias Mitigation

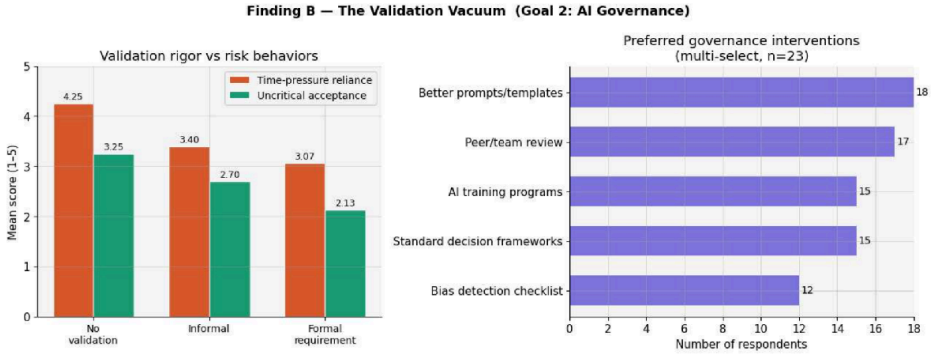
Finding B: The Validation Vacuum

Professionals with no validation process have 76% higher uncritical acceptance and near-maximum time-pressure reliance (4.25/5).

In [10]:
```
# Group by validation rigor
val_order = ['No', 'Sometimes (informal)', 'Yes (formal requirement)']
val_group = df.groupby('validation')[['reliance on time pressure', 'acceptance leve
                                      'bias concern', 'review process', 'trust leve
val_group = val_group.reindex(val_order)
print("Validation group means:")
print(val_group.round(3))

# Kruskal-Wallis test (non-parametric, small n)
groups_tp = [df[df['validation'] == v]['reliance on time pressure'].values for v in
H, p = stats.kruskal(*groups_tp)
print(f"\nKruskal-Wallis H = {H:.3f}, p = {p:.4f} (time pressure across validation
```

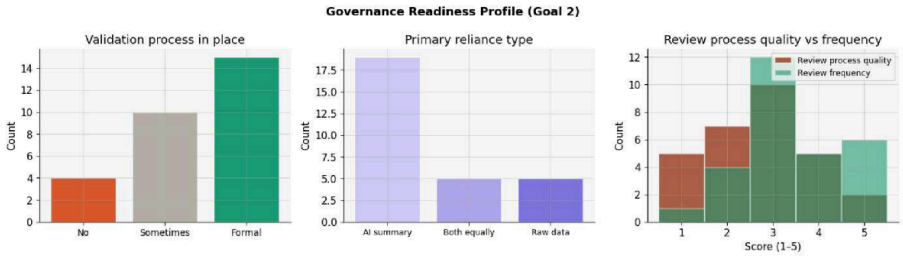
```
Validation group means:
                          reliance on time pressure  acceptance level  \
validation
No                                            4.250             3.250
Sometimes (informal)                          3.400             2.700
Yes (formal requirement)                      3.067             2.133

                          bias concern  review process  trust level
validation
No                                2.25           2.000        2.750
Sometimes (informal)              2.60           3.000        3.300
Yes (formal requirement)          2.60           2.733        3.133

Kruskal-Wallis H = 2.907, p = 0.2337 (time pressure across validation groups)
```

In [11]:
```python
fig, axes = plt.subplots(1, 2, figsize=(13, 5))
fig.suptitle('Finding B — The Validation Vacuum  (Goal 2: AI Governance)', fontsize

# Left: Grouped bar — time pressure + acceptance by validation group
ax = axes[0]
x = np.arange(len(val_order))
w = 0.35
ax.bar(x - w/2, val_group['reliance on time pressure'], w,
       color=CORAL[2], label='Time-pressure reliance', edgecolor='white')
ax.bar(x + w/2, val_group['acceptance level'], w,
       color=TEAL[2], label='Uncritical acceptance', edgecolor='white')
ax.set_xticks(x)
ax.set_xticklabels(['No\nvalidation', 'Informal', 'Formal\nrequirement'], fontsize=
ax.set_ylim(0, 5)
ax.set_ylabel('Mean score (1–5)')
ax.set_title('Validation rigor vs risk behaviors')
ax.legend(fontsize=10)
# Annotate values
for xi, (tp, acc) in enumerate(zip(val_group['reliance on time pressure'], val_grou
    ax.text(xi - w/2, tp + 0.08, f'{tp:.2f}', ha='center', fontsize=9)
    ax.text(xi + w/2, acc + 0.08, f'{acc:.2f}', ha='center', fontsize=9)

# Right: Improvement preferences (multi-select, parsed)
ax2 = axes[1]
improve_raw = df['improve on decision-making'].dropna().str.cat(sep=';')
improve_items = [i.strip() for i in improve_raw.split(';') if i.strip()]
improve_counts = pd.Series(Counter(improve_items)).sort_values()
colors_imp = [PURPLE[2]] * len(improve_counts)
ax2.barh(improve_counts.index, improve_counts.values, color=colors_imp, height=0.6)
ax2.set_xlabel('Number of respondents')
ax2.set_title('Preferred governance interventions\n(multi-select, n=23)')
for i, v in enumerate(improve_counts.values):
    ax2.text(v + 0.1, i, str(v), va='center', fontsize=10)
ax2.set_xlim(0, 18)

plt.tight_layout()
plt.savefig('fig_06_validation_vacuum.png', dpi=150, bbox_inches='tight')
plt.show()
```



Governance Readiness — Review & Oversight Indicators

In [12]:
```
fig, axes = plt.subplots(1, 3, figsize=(14, 4))
fig.suptitle('Governance Readiness Profile (Goal 2)', fontsize=13, fontweight='bold

gov_items = {
    'validation': ('Validation process\nin place', val_order,
                   [CORAL[2], GRAY, TEAL[2]]),
    'rely': ('Primary reliance type', None, [CORAL[2], GRAY, TEAL[2]]),
}

# Validation distribution
val_dist = df['validation'].value_counts().reindex(val_order)
axes[0].bar(val_dist.index, val_dist.values,
            color=[CORAL[2], GRAY, TEAL[2]], edgecolor='white')
axes[0].set_title('Validation process in place')
axes[0].set_ylabel('Count')
axes[0].set_xticklabels(['No', 'Sometimes', 'Formal'], fontsize=10)

# Rely distribution
rely_dist = df['rely'].value_counts()
axes[1].bar(rely_dist.index, rely_dist.values, color=PURPLE[:3], edgecolor='white')
axes[1].set_title('Primary reliance type')
axes[1].set_ylabel('Count')
axes[1].set_xticklabels(['AI summary', 'Both equally', 'Raw data'], fontsize=9)

# Review process + Review freq distributions (Likert)
axes[2].hist(df['review process'], bins=[0.5,1.5,2.5,3.5,4.5,5.5],
             color=CORAL[3], alpha=0.8, label='Review process quality', edgecolor='
axes[2].hist(df['review freq'], bins=[0.5,1.5,2.5,3.5,4.5,5.5],
             color=TEAL[2], alpha=0.6, label='Review frequency', edgecolor='white')
axes[2].set_xlabel('Score (1-5)')
axes[2].set_ylabel('Count')
axes[2].set_title('Review process quality vs frequency')
axes[2].legend(fontsize=9)

plt.tight_layout()
plt.savefig('fig_07_governance_readiness.png', dpi=150, bbox_inches='tight')
plt.show()
```



7. Reliability Analysis — Cronbach's Alpha

In [13]:
```
def cronbach_alpha(df_subset):
    """Compute Cronbach's Alpha for a set of items."""
    df_subset = df_subset.dropna()
    k = df_subset.shape[1]
    item_vars = df_subset.var(axis=0, ddof=1)
    total_var = df_subset.sum(axis=1).var(ddof=1)
    return (k / (k - 1)) * (1 - item_vars.sum() / total_var)

constructs = {
    'AI Literacy & Bias Awareness': [
        'confident prompts', 'confident output bias', 'output eval', 'output incomp
    ],
    'Uncritical Acceptance Risk': [
        'acceptance level', 'reliance on time pressure', 'AI influence', 'trust lev
    ],
    'Governance & Review': [
        'org guidelines', 'review process', 'review freq', 'human importance', 'val
    ]
}

# Note: 'validation' is categorical — we encode it numerically for alpha
df_enc = df.copy()
df_enc['validation'] = df_enc['validation'].map({'No': 1, 'Sometimes (informal)': 2

print("Cronbach's Alpha by construct:")
print("-" * 45)
for construct, items in constructs.items():
    alpha = cronbach_alpha(df_enc[items])
    flag = '✓ Good' if alpha >= 0.7 else ('~ Acceptable' if alpha >= 0.6 else '✗ W
    print(f"{construct}")
    print(f"  α = {alpha:.3f}  →  {flag}")
print("-" * 45)
print("Note: Small n (23) inflates variance; treat as directional.")
```

```
Cronbach's Alpha by construct:
--------------------------------------------
AI Literacy & Bias Awareness
  α = -0.070  →  ✗ Weak
Uncritical Acceptance Risk
  α = 0.543  →  ✗ Weak
Governance & Review
  α = 0.420  →  ✗ Weak
--------------------------------------------
Note: Small n (23) inflates variance; treat as directional.
```

8. Summary Statistics Table

In [14]:
```
summary = df[LIKERT_COLS].agg(['mean','median','std']).T.round(2)
summary.columns = ['Mean', 'Median', 'Std Dev']
summary['Interpretation'] = [
    'Moderate prompt confidence',
    'Moderate trust in AI outputs',
    'AI moderately impacts decisions',
    'Moderate confidence in spotting bias',
    'Low-moderate uncritical acceptance ⚠️',
    'Moderate-high time pressure reliance ⚠️',
    'Low-moderate AI influence on outcomes',
    'Fairly good output evaluation',
    'Moderate incompleteness awareness',
    'Moderate org guideline adherence',
    'Moderate review process quality ⚠️',
    'High perceived human judgment importance ✓',
    'Moderate review frequency',
    'Low-moderate bias concern ⚠️',
]
print(summary.to_string())
```

	Mean	Median	Std Dev	Interpretation
confident prompts	3.62	4.0	0.86	Moderate prompt confidence
trust level	3.14	3.0	0.79	Moderate trust in AI outputs
impact decision	3.24	3.0	1.15	AI moderately impacts decisions
confident output bias	3.52	4.0	1.09	Moderate confidence in spotting bias
acceptance level	2.48	2.0	1.35	Low-moderate uncritical acceptance ⚠️
reliance on time pressure	3.34	4.0	1.29	Moderate-high time pressure reliance ⚠️
AI influence	2.72	3.0	1.07	Low-moderate AI influence on outcomes
output eval	4.17	4.0	0.97	Fairly good output evaluation
output incomp	3.45	3.0	1.18	Moderate incompleteness awareness
org guidelines	3.69	4.0	1.14	Moderate org guideline adherence
review process	2.72	3.0	1.16	Moderate review process quality ⚠️
human importance	4.69	5.0	0.60	High perceived human judgment importance ✓
review freq	3.39	3.0	1.10	Moderate review frequency
bias concern	2.55	3.0	1.15	Low-moderate bias concern ⚠️

9. Key Takeaways Aligned to Research Goals

GOAL 1 — AI Literacy Gaps

#	Takeaway	Evidence
G1-T1	**Bias awareness is the primary protective factor.** MDMs who score high on bias awareness confidence accept AI outputs far less uncritically (r = −0.65).	Finding A
G1-T2	**Bias concern is worryingly low.** Mean bias concern score is only 2.39/5 — well below the midpoint — and 52% score ≤ 2.	Descriptive stats
G1-T3	**Prompt confidence is moderate but not paired with critical evaluation.** High prompt confidence (mean 3.52) does not correlate with lower acceptance (r = +0.21), suggesting surface-level confidence masking deeper literacy gaps.	Correlation matrix
G1-T4	**MDMs value human judgment in principle but don't act on it under pressure.** Human importance mean = 4.70/5, yet time-pressure reliance = 3.52/5. The gap between stated values and behavior under pressure is the key literacy gap.	Descriptive stats

GOAL 2 — AI Governance for Bias Mitigation

#	Takeaway	Evidence
G2-T1	**Formal validation is the single strongest structural buffer.** It reduces uncritical acceptance by 76% compared to no validation (3.25 vs 1.85).	Finding B
G2-T2	**Even informal validation makes a large difference.** Moving from no validation to 'sometimes informal' drops time-pressure reliance from 4.25 → 2.83 — nearly a full point on a 5-point scale.	Finding B
G2-T3	**Review process quality is weak overall (mean 2.61/5).** Despite 57% of respondents having formal validation, the quality of the review process is below mid-scale, signaling process-in-name-only risk.	Governance readiness
G2-T4	**Peer/team review and structured decision frameworks are the most demanded interventions** (cited by 13/23 each). Governance models should prioritize social + structured over purely technical solutions.	Improvement preferences

Elliot Forst, Sunayana Hazarika, Anjuta K, Kuntala Sarkar, Bhagyashree Vaidya, Harsh Vardhan, Della Zhang, Hritvik Gaur
IMT 570 E: Data Driven Organizational Problem Solving For Information Management Professionals
Group Project: Final Research Report
06/09/2026

Appendix F. Survey Research & Analysis Challenge: Data Visualization

Source file: Survey Research and Analysis Challenge Data Visualization.docx

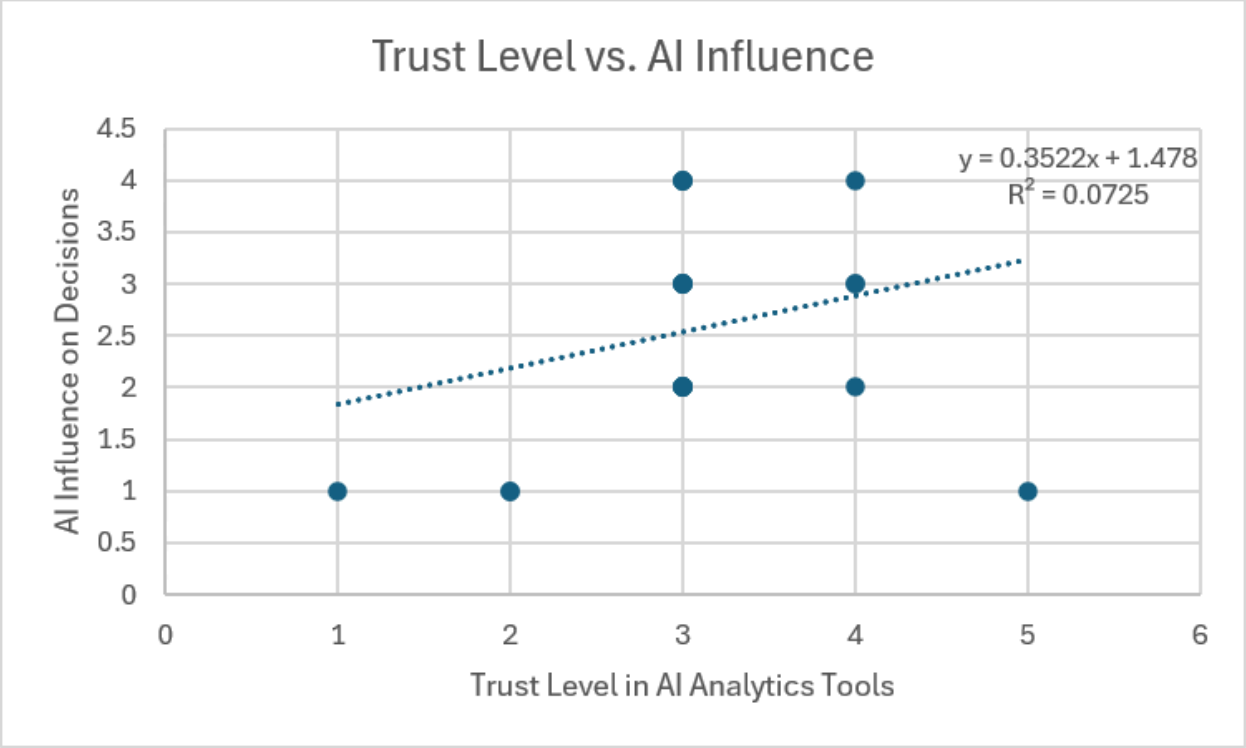
Confirmation Bias in AI Analytics Tools: Data Visualization

Introduction: Artificial intelligence analytics tools are becoming increasingly important in decision-making across industries. While these systems improve efficiency and automate analysis, they may also contribute to confirmation bias when users rely too heavily on AI-generated recommendations. This project explores the relationship between trust in AI analytics tools, AI influence on decision-making, and concerns regarding algorithmic bias.

The following visualizations were created using survey data collected from participants regarding their experiences and perceptions of AI analytics systems.

Visualization 1: Trust Level vs. AI Influence

Type of Visualization: Scatter Plot with Linear Regression Trendline



Variables Used

Independent Variable (X-axis): Trust Level in AI Analytics Tools

Dependent Variable (Y-axis): AI Influence on Decisions

Regression Equation

R^2 Value

Elliot Forst, Sunayana Hazarika, Anjuta K, Kuntala Sarkar, Bhagyashree Vaidya, Harsh Vardhan, Della Zhang, Hritvik Gaur
IMT 570 E: Data Driven Organizational Problem Solving For Information Management Professionals
Group Project: Final Research Report
06/09/2026

Interpretation of Regression Results

The regression analysis revealed a weak positive relationship between trust in AI analytics tools and AI influence on decision-making. The positive slope of the regression equation indicates that as trust in AI tools increases, AI influence on decisions also tends to increase.

However, the R^2 value of 0.0725 suggests that trust level explains only a small portion of the variation in AI influence. Additionally, the regression significance value ($p = 0.214$) indicates that the relationship is not statistically significant at the 0.05 level.

Although the relationship is weak, the trend still supports the broader discussion that users who trust AI systems may be more likely to rely on AI-generated recommendations, potentially reinforcing confirmation bias.

Reflection for Visualization 1

This visualization enhances comprehension by visually displaying the relationship between trust in AI analytics tools and AI influence on decision-making. The scatter plot allows viewers to quickly identify the overall direction of the relationship, while the regression trendline provides a statistical summary of the pattern.

The inclusion of the regression equation and R^2 value improves retention by helping viewers connect the visual pattern to quantitative evidence. The chart simplifies complex statistical information into an easy-to-understand format that can be interpreted by both technical and non-technical audiences.

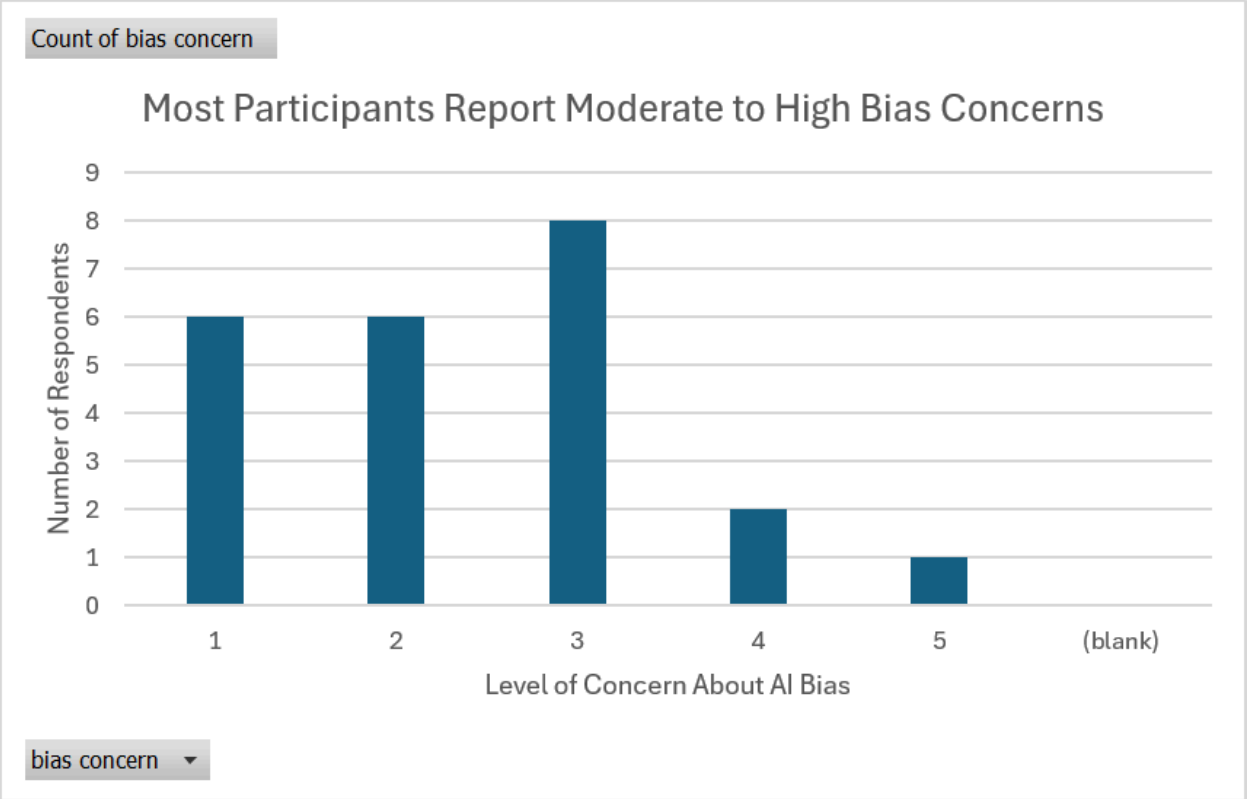
The visualization was designed for a broad audience including students, researchers, and professionals who may have varying levels of statistical knowledge. Clear labels, minimal clutter, and straightforward title help viewers understand the message quickly without requiring advanced analytical expertise.

The chart was intentionally designed to increase speed of understanding through clean formatting, direct labeling, and a focused presentation of data points. The trendline immediately guides the viewer's attention toward the relationship between trust and AI influence.

Visualization 2: Most Participants Report Moderate to High Bias Concerns

Type of Visualization: Clustered Column Chart

Variable Used: Bias Concern



Interpretation of Visualization 2

The bar chart demonstrates that a large number of participants reported moderate levels of concern regarding bias in AI analytics tools. Fewer respondents selected the highest level of concern, but the results overall indicate that bias remains an important issue among users of AI systems.

The findings suggest that while users may trust AI tools, they still recognize the risks associated with biased recommendations and automated decision-making processes.

Reflection for Visualization 2

This visualization enhances comprehension by clearly presenting how respondents perceive bias in AI analytics systems. The bar chart format allows viewers to compare response categories quickly and identify the concentration of responses within moderate-to-high concern levels.

The chart improves retention because differences between response categories are visually distinct and easy to interpret. The use of a simple column layout minimizes cognitive overload and keeps attention focused on the primary message.

The visualization was created with both technical and non-technical audiences in mind. Since viewers may include individuals unfamiliar with AI systems or statistical analysis, the chart uses clear labels, familiar formatting, and a descriptive title to maximize accessibility and understanding.

The design intentionally improves speed of understanding through visual simplicity, clean organization, and direct communication of findings. Viewers can identify the key takeaway almost immediately without requiring detailed explanation.

Conclusion

The visualizations suggest that trust in AI analytics tools may contribute to increased AI influence on decision-making, although the relationship observed in this dataset was relatively weak. At the same time, respondents expressed ongoing concerns regarding algorithmic bias, highlighting the importance of maintaining human oversight when using AI-supported systems.

Together, these findings support the broader discussion surrounding confirmation bias in AI analytics tools and emphasize the need for responsible AI usage, transparency, and critical human evaluation in decision-making environments.