ISSN: 2473-4901
v12 n6668

Using Generative AI for Program Assessment: Faculty Reflections on Multi-Tool Workflows and Human Verification

Tyson Riskas
tyson.riskas@uvu.edu
Utah Valley University
Orem, UT

Eduardo Sanachez
esanchezt2@unicartagena.edu.co
Universidad de Cartagena, Colombia
Bolívar, Colombia

Colin Doyle
doylec10@erau.edu
Embry Riddle Aeronautical University
Daytona, FL

Amaury Cabarcas
acabarcasa@unicartagena.edu.co
Universidad de Cartagena, Colombia
Bolívar, Colombia

Sayeed Sajal
Sayeed.Sajal@uvu.edu
Utah Valley University
Orem, UT

Weidong Wu
Weidong-Wu@utc.edu
University of Tennessee at Chattanooga
Chattanooga, TN

Anthony Nix
nixa@oregonstate.edu
Oregon State University
Corvallis, OR

ABSTRACT

Qualitative evidence can inform program improvement, but organizing narrative feedback is time-intensive and requires interpretive judgment. This exploratory study examines how seven faculty participants described using generative artificial intelligence (GenAI) during an ABET-sponsored professional-development workshop on program assessment. The study distinguishes a staged workshop exercise using de-identified, atomized exit-survey statements from the faculty reflection dataset analyzed in this paper. Participants reported using heterogeneous subsets of ChatGPT, Claude, Gemini, Microsoft Copilot, Jasper, NotebookLM, and Grok for drafting, organizing, prompt refinement, stress-testing, and verification. Because tools, prompts, and tasks varied, the study does not claim a controlled comparison of platforms. Cross-case thematic synthesis identified four patterns: GenAI supported rapid first-pass organization; useful outputs required iterative and context-specific prompting; faculty verification was necessary to detect counting, grouping, logical, and pedagogical errors; and privacy, auditability, and institutional governance shaped acceptable use. The strongest contribution is a faculty-guided workflow in which AI organizes evidence but does not make assessment decisions. Findings are limited by the seven-person exploratory sample, heterogeneous procedures, self-reported reflections, and incomplete model-version provenance. Future research should use fixed datasets, prompts, model versions, and predefined comparison criteria.

Keywords: Generative artificial intelligence, program assessment, qualitative analysis, human-in-the-loop, faculty reflection, ABET

Using Generative AI for Program Assessment: Faculty Reflections on Multi-Tool Workflows and Human Verification

Tyson Riskas, Colin Doyle, Sayeed Sajal,
Anthony Nix, Eduardo Sanachez and Amaury Cabarcas

1. INTRODUCTION

Program assessment depends on evidence that faculty can interpret and use for improvement. ABET Criterion 4 requires programs to use appropriate documented processes for assessing student outcomes and to use evaluation results systematically in continuous-improvement actions (ABET, 2026). Quantitative indicators are important, but narrative feedback and faculty interpretation often reveal context that is not visible in summary measures.

Generative artificial intelligence (GenAI) can rapidly organize, summarize, and restructure text. In higher education, educators have reported both practical interest in these capabilities and uncertainty about appropriate use, support, and governance (Lee et al., 2024). For program assessment, the attraction is clear: a model can produce a first-pass grouping of narrative evidence much faster than a faculty committee. The methodological problem is equally clear: fluent output can appear credible while containing omissions, duplications, weak categorizations, or interpretations that do not fit the program context.

This paper presents an exploratory faculty-reflection study that uses an ABET workshop exercise as context for examining human verification, not as evidence that one platform performed better than another. The purpose of this study is to describe how faculty participants used GenAI within program-assessment workflows, what limitations they encountered, and why human review remained necessary. The study addresses three research questions:

1. How did faculty participants describe the roles of GenAI tools in program-assessment work?
2. What limitations or error types did participants identify, and how did human verification address them?
3. How did prompting, tool selection, and privacy or governance considerations shape faculty-guided workflows?

2. RELATED LITERATURE

Emerging research suggests that large language models can assist with qualitative analysis but should not be treated as autonomous interpreters. Morgan (2023) found that ChatGPT more readily reproduced concrete, descriptive themes than subtle interpretive themes and emphasized that AI output remains raw material within a larger analytic process. Hamilton et al. (2023) similarly reported both overlap and difference between human and AI-generated analyses of interview data, supporting an augmentative rather than replacement role.

Prompt design and researcher involvement are central to this work. Naeem et al. (2025) proposed a staged prompting process for thematic analysis while maintaining that researchers must remain involved throughout the analysis. Controlled comparison is possible, but it requires shared data, defined procedures, and a human reference analysis. Ayik et al. (2026), for example, compared four AI-assisted systems against a validated human-coded analysis using a common framework. That design differs fundamentally from a workshop in which participants choose different platforms and prompting strategies for different tasks.

Reporting quality is another concern. A 2026 scoping review of 75 studies found wide variation in how qualitative studies documented model versions, deployment configurations, parameter settings, prompts, and human validation; 97% nevertheless incorporated human verification (Kempny et al., 2026). These findings reinforce two principles used in the present revision: heterogeneous tool use

should not be converted into a controlled comparison after the fact, and methodological transparency should extend beyond naming a platform.

For program assessment, these methodological issues have practical consequences. Assessment artifacts are not merely prose products; they support judgments about student outcomes and program improvement. The relevant question is therefore not whether GenAI can produce polished text, but whether a faculty-guided process can preserve completeness, traceability, and defensible interpretation.

3. METHODS

Study Design and Context

This exploratory qualitative study was conducted in connection with an ABET-sponsored professional-development workshop on using GenAI tools and techniques for program assessment. The workshop provided a practice-centered setting in which faculty experimented with AI-supported assessment tasks and reflected on the usefulness, limitations, and governance of those workflows. The revised study does not benchmark model performance and does not treat platform choice as an experimental condition.

Participants

Seven faculty participants from multiple institutions and engineering-related disciplines contributed de-identified reflection records labeled P1 through P7. Participants entered the workshop with differing levels of AI experience and used different tools for different assessment tasks. This variation is treated as a characteristic of the exploratory setting rather than as a basis for causal or comparative claims.

Contextual Assessment Task and Data Collection

To examine how generative AI supports qualitative assessment workflows, workshop participants evaluated de-identified narrative responses directly sourced from an ABET program-level student exit survey (specifically answering "What would you recommend to improve the program?"). During this structured exercise, participants applied a multi-stage prompting workflow to these survey responses to break raw comments into 74 discrete, single-issue statements, verify statement counts and traceable IDs, synthesize themes, and rank/classify those themes—stopping explicitly for faculty review and verification between each stage (Rogers, 2026; complete staged workflow detailed in Appendix A).

The primary analytic dataset for this research consists of the qualitative reflection records generated by seven faculty participants (P1 through P7) evaluating their experience using AI platforms to process, verify, and audit these 74 discrete statements. Following the workshop exercise, participants completed an asynchronous qualitative interview administered through a custom Gemini 3.1 Pro agent (configured as the "Qualitative Research Assistant"). The agent guided participants through a sequential, single-question reflection protocol covering four distinct phases: background experience, prompt iteration, output verification and error detection, and ethical/governance considerations. The complete setup, system instructions, configuration parameters, and interview protocol for this agent are detailed in Appendix B. These participant records are analyzed as reflective accounts of human-in-the-loop verification rather than a standardized model-performance benchmark.

Tool-Use Context

Participants reported using different subsets of ChatGPT, Claude, Gemini, Microsoft Copilot, Jasper, NotebookLM, and Grok. Table 1 consolidates the retained participant-level descriptions. Platform names are reported without inferring common model versions, settings, or deployment conditions. The table is descriptive and should not be read as evidence of relative platform performance.

Participant	Reported platforms	Primary reflection emphasis
P1	Microsoft Copilot, Claude, Gemini	Iterative refinement; comparison of tone, context retention, and assessment drafting
P2	ChatGPT, Gemini	Role/persona prompting; alignment of performance-indicator verbs with Bloom's Taxonomy
P3	ChatGPT	Generation of contrasting response cases to stress-test rubric sensitivity
P4	ChatGPT, Gemini, Jasper, Copilot, Grok	Meta-prompting and cross-checking across tools
P5	Microsoft Copilot	Pattern organization and review of rubric structure
P6	Claude, Gemini, NotebookLM	Constraint prompting for measurable indicators and verb use
P7	ChatGPT, Gemini, Claude	Cross-checking, auditability, privacy, and governed institutional use

Table 1. Descriptive participant tool-use context. Tools, prompts, and tasks varied; the table is not a controlled platform comparison.

Analysis

The qualitative dataset comprised seven de-identified reflection records generated via the custom Gemini 3.1 Pro interview agent (detailed in Appendix B). Data were analyzed using a cross-case thematic synthesis informed by reflexive thematic analysis principles, which emphasize active researcher engagement and interpretive meaning-making rather than passive code extraction (Braun & Clarke, 2022).

The analysis proceeded through a systematic, multi-coder review process:

Initial Coding: Two members of the research team independently conducted open coding on the primary reflection summaries to identify candidate categories. Initial codes focused on recurring descriptions of tool roles, iterative prompt refinement, human verification practices, assessment-error detection, stress-testing via synthetic cases, and institutional governance.
Category Reconciliation: The coders met iteratively to compare code definitions, group overlapping concepts into higher-order themes, and resolve interpretive discrepancies through reflexive consensus.

Thematic Synthesis: The refined categories were synthesized across all seven participant cases, producing the four primary findings presented in Section 4.

Because participants used heterogeneous subsets of AI platforms, prompts, tasks, and output criteria, no frequency-based reliability statistics (e.g., inter-coder reliability) or inter-model agreement measures were calculated. The goal was not to measure statistical convergence across standardized treatments, but to synthesize patterns of human oversight within real-world assessment workflows. The instructional workshop exercise was analyzed separately as contextual evidence illustrating the verification process. Its documented statement-count and theme-reporting discrepancies were evaluated as operational artifacts of the workflow rather than attributes of a specific platform, and they were intentionally kept separate from the participant reflection dataset.

Privacy and Traceability

The workshop protocol required personal names to be removed or replaced with neutral descriptors and identifying contextual details to be generalized while preserving meaning. Atomized statements

retained traceable student and statement identifiers for completeness checks. Faculty reflections are presented only through P1-P7 identifiers. The study reports no student outcome evaluation and no personally identifying information.

4. RESULTS

GenAI Supported First-Pass Organization and Drafting

Participants described GenAI as useful for producing a rapid first pass on assessment-related material. Reported uses included organizing text-rich evidence, drafting performance indicators or rubric language, improving structure, and testing alternative wording. The practical value was not that the model completed assessment work independently, but that it reduced the effort required to create material for faculty review. Participant comments about named tools are reported as individual experiences under heterogeneous conditions, not as controlled evidence of platform superiority.

> *"Copilot did a really good job with performance indicators and drafting the rubric. Gemini struggled a bit... I started to prefer Copilot." (P1)*
> *"Copilot was stuffy and too formal in its responses. I had to continually tell it to be more casual and conversational." (P1)*
> *"I did notice that Gemini would get stuck with the previous prompt and I needed to refresh or start a new chat." (P1)*

One exploratory use extended beyond drafting. P3 generated contrasting response cases and used them to test whether a developing rubric could distinguish stronger from weaker performance. This illustrates a potentially useful stress-testing role for GenAI, but the present study does not establish the validity or reliability of that method.

> *"If a rubric can't distinguish between the 'competent' and 'not competent' responses the AI produced, the rubric itself needs refinement." (P3)*

Useful Output Required Iterative and Context-Specific Prompting

The reflections consistently described prompt refinement as part of the work rather than a preliminary step completed once. Participants moved from broad requests toward constraints, role descriptions, examples, tone adjustments, measurable-verb requirements, and instructions tied to accreditation or program context. Several participants used one tool to draft and another to review or restructure an output. These practices were individualized; they show how faculty adapted tools to a task, not that a particular prompting strategy was tested under common conditions.

> *"The AI was great but started out more general and needed help becoming more specific." (P5)*
> *"We used AI to design the initial prompt and it produced a good output, but it lacked the specificity... The second go around was a lot better and more focused." (P4)*

Human Verification Detected Errors That Fluent Output Could Hide

Verification was the strongest recurring finding. The workshop exercise provided a concrete illustration: after theme development, the verification report accounted for 76 statements even though 74 were expected, and one theme reported nine statements while listing ten. The problem was not the absence of readable output; it was the loss of numerical and structural integrity within an apparently organized result. The staged process surfaced the discrepancy before faculty interpretation or action.

> *"It took us a minute to see if the AI was hallucinating and required human in the loop spot checking." (P1)*

The faculty reflections described additional error types. These included combining multiple assessment criteria into a single rubric element, assigning multiple Bloom's Taxonomy verbs in ways that weakened measurability, carrying earlier prompt context into later tasks, and producing plausible statements that required factual or logical checking. These examples support a shift in faculty work

from initial production toward auditing, but they do not support a claim that every platform failed in the same way or at the same rate.

> *"Verification revealed logical errors where the AI combined multiple assessments into one spot on the rubric where they should be split up." (P5)*
> *"It definitely improves on speed and fluency. But we need humans to validate the information is correct." (P7)*

Heterogeneous Tool Use Increased the Need for Governance and Documentation

Participants did not converge on one preferred platform or one standardized workflow. Instead, they used tools selectively for drafting, formatting, organization, cross-checking, or simulation. This flexibility was useful in practice but prevented direct comparison. It also made documentation more important: without a record of the task, prompt, platform, model version, and human review, later readers cannot determine why an output changed or whether a reported difference came from the tool, the prompt, the data, or the evaluator.

> *"Using multiple tools simultaneously definitely helped mitigate the 'black box' nature of any one AI system. Instead of trusting a single output, we were able to triangulate across tools." (P4)*

Privacy and auditability were described as conditions for responsible adoption. Participants favored de-identification before cloud processing, visible human review, preservation of an audit trail, and institutionally governed environments for sensitive assessment information. These concerns were not treated as proof that one technical architecture is required; rather, they identified design requirements that programs must resolve before scaling AI-assisted assessment work.

> *"So long as all outputs eventually are reviewed by a human and there is a good audit trail, I think there are no ethical issues." (P1)*

5. DISCUSSION

The Contribution Is a Verification Workflow, Not a Platform Ranking

The central contribution of this study is a bounded faculty-guided workflow: AI can organize evidence and generate candidate structures, while faculty remain responsible for completeness, coherence, interpretation, and decisions. This conclusion is consistent with research showing that LLMs tend to perform more reliably on descriptive organization than on subtle interpretation (Morgan, 2023) and with studies positioning AI as a supplement to human qualitative analysis (Hamilton et al., 2023; Naeem et al., 2025).

Controlled comparisons such as Ayik et al. (2026) use common data, common analytic frameworks, and a validated human reference. In the present workshop, participants used different tools, prompts, and tasks. Their experiences are informative about practice, but observed differences cannot be attributed to platform capability. This paper's contribution is therefore an account of how faculty exercised judgment under heterogeneous real-world conditions.

Verification Redistributes Rather Than Eliminates Faculty Work

GenAI may reduce the mechanical burden of first-pass organization, but the workshop evidence indicates that it creates a corresponding need for verification. Faculty must check counts and identifiers, inspect whether themes represent the underlying statements, separate combined criteria, assess the measurability of verbs, and challenge contextually plausible but incorrect output. This is not simply copyediting. It is professional judgment tied to the meaning and consequences of program assessment.

The finding is especially important because polished language can conceal defects. The documented 76-versus-74 discrepancy would be easy to miss if reviewers focused on readability rather than traceability. A staged workflow with explicit stop points makes verification visible and gives faculty an

opportunity to reject or revise an output before it becomes part of a continuous-improvement decision.

Methodological Transparency Is Part of Trustworthiness

The broader literature shows that LLM-assisted qualitative studies often underreport model versions, deployment conditions, parameter settings, prompting, and validation procedures (Kempny et al., 2026). The present study encountered the same risk retrospectively: platform names were retained more consistently than exact model and prompt provenance. Future work should preserve the complete prompt, model/version, date, deployment mode, relevant settings, input identity, output, and human decision at each checkpoint. This level of reporting supports evaluation of the computational process without claiming that qualitative interpretation itself is mechanically reproducible.

A Practical Faculty-Guided Workflow

The findings support six practical safeguards for program-assessment use:

1. Define the assessment task and the boundary between organization, interpretation, and decision-making.
2. Remove personally identifying information and document the permitted processing environment before using a model.
3. Use staged prompts with explicit stop points rather than requesting a complete analysis in one step.
4. Preserve source and statement identifiers so completeness, duplication, and cross-coding can be checked.
5. Record the platform, model version when known, prompt, date, deployment mode, and human revisions.
6. Require faculty approval before themes, classifications, recommendations, or improvement actions are accepted.

6. LIMITATIONS

This study is exploratory and based on seven faculty participants in one professional-development setting. Participation was voluntary, and the reflections represent self-reported experiences rather than observations of standardized performance. Participants used heterogeneous platforms, prompts, tasks, and levels of prior experience. Consequently, the study cannot estimate platform accuracy, efficiency, reliability, or superiority.

The workshop exercise and the faculty reflection dataset served different purposes. The documented counting and theme-reporting discrepancy demonstrates the value of verification but does not establish an error rate or identify a platform effect. The retained records also did not provide complete model-version, parameter, or deployment provenance for every participant. These limitations constrain reproducibility and are now treated as findings about reporting practice rather than being concealed by a stronger comparative design.

Finally, the findings are grounded in accreditation-aligned program-assessment work and may not transfer directly to other qualitative research settings. Future research should use a fixed dataset, a versioned common prompt, recorded model configurations, predefined comparison criteria, independent human analysis, and transparent validation procedures.

7. CONCLUSION

Faculty participants described GenAI as useful for rapid organization, drafting, prompt-supported refinement, and exploratory stress-testing in program-assessment work. Its value depended on context-specific prompting and sustained human review. The most consequential evidence was not that multiple tools produced the same answer, but that faculty verification detected numerical, structural, logical, and pedagogical problems that fluent output could hide.

Accordingly, GenAI should be positioned as an analytical assistant within a documented workflow, not as an evaluator or decision-maker. AI can organize evidence; faculty must determine whether the organization is complete, coherent, meaningful, and suitable for continuous-improvement action. That division of responsibility is the most defensible and practically useful conclusion supported by this exploratory study.

ACKNOWLEDGMENTS

The authors thank Dr. Gloria Rogers and Dr. Kevin Huggins for their leadership and facilitation of the ABET workshop on using generative AI tools and techniques for program assessment. The workshop's emphasis on staged prompting, faculty verification, and responsible use provided the context for this study. The authors also acknowledge the faculty participants whose reflections informed the analysis.

DECLARATION OF GENERATIVE AI AND AI-ASSISTED TECHNOLOGIES

During manuscript preparation, the authors used ChatGPT, Jasper, Claude, Google Gemini, Microsoft Copilot, and NotebookLM to support idea organization, comparative synthesis, drafting, and iterative editing. All AI-assisted content was reviewed and revised by the authors, who retain responsibility for the accuracy, integrity, and final presentation of the work. Participant-reported use of Grok and other tools during workshop activities is described separately and does not imply standardized research conditions.

REFERENCES

ABET. (2026). Criteria for accrediting engineering programs, 2026-2027. https://www.abet.org/accreditation/accreditation-criteria/criteria-for-accrediting-engineering-programs-2026-2027/

Ayik, B., Gu, D., Zan, Y., Kim, S., & Kim, W. L. (2026). Human vs. AI: Evaluating thematic analysis with ChatGPT, QInsights, ATLAS.ti AI, and MAXQDA AI Assist. Qualitative Inquiry. Advance online publication. https://doi.org/10.1177/10778004251412874

Braun, V., & Clarke, V. (2022). Thematic analysis: A practical guide. SAGE.

Hamilton, L., Elliott, D., Quick, A., Smith, S., & Choplin, V. (2023). Exploring the use of AI in qualitative analysis: A comparative study of guaranteed income data. International Journal of Qualitative Methods, 22. https://doi.org/10.1177/16094069231201504

Kempny, C., Frings, J., Rust, P., Meister, S., & Fehring, L. (2026). The use and methodological reporting of large language models in qualitative research: A scoping review. BMC Medical Research Methodology, 26, Article 137. https://doi.org/10.1186/s12874-026-02913-1

Lee, D., Arnold, M., Srivastava, A., Plastow, K., Strelan, P., Ploeckl, F., Lekkas, D., & Palmer, E. (2024). The impact of generative AI on higher education learning and teaching: A study of educators' perspectives. Computers and Education: Artificial Intelligence, 6, 100221. https://doi.org/10.1016/j.caeai.2024.100221

Morgan, D. L. (2023). Exploring the use of artificial intelligence for qualitative data analysis: The case of ChatGPT. International Journal of Qualitative Methods, 22. https://doi.org/10.1177/16094069231211248

Naeem, M., Smith, T., & Thomas, L. (2025). Thematic analysis and artificial intelligence: A step-by-step process for using ChatGPT in thematic analysis. International Journal of Qualitative Methods, 24. https://doi.org/10.1177/16094069251333886

Rogers, G. (2026, April). Partnering with generative AI in program assessment [Workshop presentation]. ABET Symposium.

APPENDIX A. STAGED WORKSHOP PROMPT

Source: Rogers, G. (2026, April). Partnering with generative AI in program assessment [Workshop presentation]. ABET Symposium. Reproduced for private team review from the authorized workshop material.

Workflow instruction

This is a staged qualitative analysis workflow with faculty verification checkpoints. Complete only the current prompt. Stop immediately when instructed and wait for faculty confirmation before proceeding to the next stage. Do not perform later steps unless explicitly instructed to continue.

Prompt 1: ATOMIZE (Clean)

Context: These are responses from a program exit survey answering the question: "What would you recommend to improve the program?" Each response may contain multiple suggestions, concerns, or recommendations.

Task: Break each student response into discrete, single-issue statements.

1. Preserve original wording as much as possible.

2. Do not summarize or interpret.

3. If a sentence contains multiple suggestions, separate them.

4. Each atomized statement must contain only one independent idea. If a statement contains a conjunction (e.g., and, but, also, as well as) linking two evaluative claims, split them into separate statements.

5. Do not omit any recommendation or suggestion.

6. Remove or replace all personal names (faculty, students, staff).

7. Replace names with neutral descriptors such as "a faculty member," "a staff member," or "a classmate."

8. Do not generalize beyond what is stated.

9. If contextual details could identify a specific individual, generalize while preserving meaning.

10. Assign traceable IDs: Student ID (e.g., S01) and Statement ID (e.g., S01-1, S01-2, etc.).

Output format: Provide a table with Student ID, Statement ID, and Atomized Statement. Output only the table/results. No additional text.

STOP HERE - Present the cleaned and atomized dataset for faculty verification before proceeding. Do not continue until faculty confirmation is provided.

Prompt 2: VERIFY

Context: These are responses from a program exit survey answering the question: "What would you recommend to improve the program?"

Task: Verify the completeness and uniqueness of the atomized statements using 74 as the expected total number of statements.

1. Count the total number of statements.

2. List all Statement IDs.

3. Confirm that there are no duplicate Statement IDs and no missing sequence within each student (e.g., S01-1, S01-2, etc.).

Output: Total number of statements; confirmation that all Statement IDs are unique and complete; identification of any issues (duplicates, gaps, inconsistencies); output only what is requested. No additional text.

STOP HERE - Present the verification results for faculty confirmation before thematic analysis begins. Do not continue until faculty confirmation is provided.

Prompt 3: Theme Development

Task: Organize the atomized statements into clearly defined themes.

- Create concise, descriptive theme titles.
- Group statements based on similarity of meaning.
- A statement may appear in more than one theme only when the same idea clearly supports multiple themes.

- When a statement appears in multiple themes, reuse the same Statement ID; do not duplicate or rewrite the statement.

For each theme, include the theme title, number of unique students, total number of statements, and for each supporting statement the Statement ID and exact atomized statement text.

- Use all 74 verified statements.
- Do not omit any statements.
- Ensure all Statement IDs are accounted for across themes.

STOP HERE - Present the themes and supporting statements for faculty verification. Do not continue until faculty confirmation is provided.

Prompt 4: Theme Classification

Task: Report the number of statements associated with each theme and present themes in ranked order from highest to lowest number of statements.

Do not classify themes. Theme classification will be completed by faculty after review.

Output format: For each theme, report the theme title, number of statements, classification, rank from highest to lowest number of statements, and original theme numbering.

- Use the verified total from Prompt 2 as the fixed number.
- Report the total number of statements analyzed.
- Confirm all Statement IDs are accounted for.
- A statement may appear in more than one theme if a single idea legitimately supports multiple themes.
- Do not duplicate statements; reference the same Statement ID when cross-coding occurs.
- Do not interpret or recommend actions.
- Output only what is requested. No additional text.

STOP HERE - Present the classified themes for faculty classification and verification before creating a summary.

APPENDIX B. QUALITATIVE INTERVIEW PROTOCOL

A Gemini 3.1 Pro Gem was created and utilized to conduct qualitative interviews. Below are the parameters of the Qualitative Research Assistant along with the corresponding questions asked to the participants.

Name: Qualitative Research Assistant

Description: This is a Gem that will gather qualitative data based on a well-defined experience

Instructions: You are a research assistant helping a group of academic professionals document their experiences. The user will input detailed information about the event or experience. Your goal is to collect detailed, qualitative data regarding their experience.

Your Behavior:

- Interview the user one question at a time to avoid overwhelming them.
- Maintain an academic yet collaborative tone.
- Encourage specific examples of prompts used and the quality of the AI's output.

you will assess all the elements of the experience and generate and ask sequential questions until you get enough information to determine their lived experience.

produce an output at the end that summarizes the experience that includes direct quotes suitable for a research paper submissions

Default tool: None

Knowledge:

GAIPA - April 2026 - In-person Syllabus(20260415_112559).pdf (Provided by ABET)

Workshop Experience: Gemini 3.1 Pro Gem Qualitative Questions

Phase 1: Background & Initial Expectations

1. Context & Motivation: What is your current role in program assessment at your institution, and what motivated you to attend the *Using Generative AI Tools and Techniques for Program Assessment* workshop?
2. Prior Experience: Prior to attending, what was your comfort level with both traditional ABET assessment processes and Generative AI tools?

Phase 2: Prompt Engineering & Practical Application

3. Prompt Iteration: Can you walk me through a specific prompt you constructed during the session (e.g., for developing performance indicators, scoring rubrics, or assessment tasks)?
    - *Follow-up:* How did you refine or iterate on that prompt to get a more tailored result?
4. Output Evaluation: When reviewing the AI-generated outputs (such as drafted rubrics or qualitative feedback), what specific strengths or limitations did you notice regarding accuracy, coherence, and alignment with ABET criteria?

Phase 3: Analytical & Complex Assessment Tasks

5. Qualitative Analysis & Case Studies: How effectively did the AI assist you when analyzing qualitative assessment data or working through the assessment case study?
6. Output Verification & Hallucinations: What specific steps or criteria did you use during the session to verify and fact-check the validity of the AI's suggestions?

Phase 4: Ethical, Organizational & Long-Term Reflection

7. Ethics & Sustainability: How did the discussions surrounding ethical considerations, citations, and environmental/sustainability impacts influence your perspective on integrating AI into institutional policy?
8. Phenomenological Impact: In what ways, if any, did this workshop shift your perception of Generative AI—moving from viewing it as a novel tool to an integrated component of continuous program improvement?