Hadjisolomou, S. P. & El-Haddad, R. W. (2026). Exploring how generative AI chatbots support and complicate ACCELERATE principles in programmatic assessment. *Intersection: A Journal at the Intersection of Assessment and Learning, Early View*.

Exploring How Generative AI Chatbots Support and Complicate ACCELERATE Principles in Programmatic Assessment

Stavros P. Hadjisolomou, Ph.D.; Rita W. El-Haddad, Ph.D.

Author Note

Stavros P. Hadjisolomou https://orcid.org/0000-0002-0861-2066

Rita W. El-Haddad https://orcid.org/0000-0001-9614-6634

"The first author is recognized as an OpenAI Champion (2024–2026) through the institution's ChatGPT EDU license program; this recognition is not an employment, consulting, funding, or honorarium relationship, and OpenAI had no role in the study design, analysis, or reporting."

Intersection: A Journal at the Intersection of Assessment and Learning

Early View

Abstract: ACCELERATE: Assessment Principles for Best Practice (2025) updates the foundational American Association for Higher Education (AAHE) principles with contemporary emphasis on equity, collaboration, and transparency. This pilot study documents the development of an AI chatbot providing programmatic assessment guidance grounded in ACCELERATE principles, examining which principles translate effectively to AI-mediated coaching versus those requiring direct human involvement. Using retrospective analysis of seven authentic challenges from Year 1 programmatic assessment evaluations at one institution, we examined AI-mediated coaching responses for evidence of effective principle translation and structural tensions that emerge when human-centered values encounter algorithmic operationalization. Preliminary findings suggest that six principles may translate effectively to AI support, while four (Collaborative/Co-Creative, Energized by Expertise, Responsiveness-Oriented, and Enduring/Evolving) carry structural tensions that appear to resist technological resolution. We propose a framework that positions AI as a tool for technical work, while humans remain guarantors of relational work in principled assessment practice.

Keywords: *programmatic assessment, generative AI, ACCELERATE principles, assessment literacy, human-AI boundaries, prompt engineering, higher education*

Introduction

Assessment practitioners in higher education face persistent challenges. Many faculty still experience difficulty articulating measurable learning outcomes, selecting appropriate assessment measures, and translating findings into evidence-based improvements (Suskie, 2018). While these challenges may be due to technical skills, they also reflect a persistent tension between treating assessment as a compliance exercise versus engaging it as meaningful inquiry into student learning.

The emergence of generative artificial intelligence (AI) tools offers potential support for these persistent challenges. AI chatbots can generate options rapidly, check alignment systematically, and provide structured templates that reduce cognitive load. Yet as AI capabilities expand, a critical question emerges: Can AI-mediated support address assessment challenges while honoring the human-centered values that distinguish meaningful assessment from mechanical compliance?

This question has become more urgent with the recent publication of ACCELERATE: Assessment Principles for Best Practice (Samuga_Gyaanam+Bheda et al., 2025), which updates foundational assessment principles with contemporary emphasis on equity, collaboration, and transparency. Its explicit human-centered orientation raises questions about AI support for such practice.

To our knowledge, no published examination has explored which ACCELERATE principles might translate effectively to AI support versus which may require direct human engagement that cannot be delegated to technology. While AI tools are increasingly marketed for assessment applications, the field lacks empirical grounding for where AI supports good practice and where human judgment remains irreplaceable.

A Note on Terminology

Throughout this paper, "AI" refers to generative AI tools producing human-like text; "GPT" refers to Generative Pre-trained Transformer technology underlying ChatGPT. "Translating" principles means encoding their values and intended behaviors into AI system instructions.

Purpose and Scope

This pilot study documents the development of an AI chatbot designed to provide programmatic assessment guidance grounded in ACCELERATE principles. Using retrospective analysis of assessment challenges drawn from Year 1 evaluation records at one institution, we examined whether AI-generated guidance aligns with ACCELERATE principles and whether such guidance would have addressed identified assessment weaknesses.

This study is exploratory rather than comprehensive. We do not claim to validate the chatbot as a finished tool; rather, we document design decisions, test principle translation, and identify tensions that emerge when human-centered values are translated into AI system instructions. The scope is deliberately limited: seven scenarios from one institution, one custom AI chatbot, and an evaluation against documented challenges rather than live faculty interactions.

The study pursues two research questions: Which ACCELERATE principles translate effectively when an AI chatbot is used to coach programmatic assessment practice? Where do principle-AI tensions emerge, and what is their structural origin? These two questions frame the work that follows: the first identifies where translation succeeds; the second locates where and why it breaks down.

These questions translated into three practical concerns that shaped the GPT's design: encoding principle values into prompt engineering strategies, anticipating where human-centered values resist

algorithmic operationalization, and preserving the practitioner's role in navigating the boundaries between AI support and human responsibility. Together, these provided an analytical lens for examining the GPT's engagement with authentic assessment challenges.

Article Overview

The Conceptual Framework situates this study within ACCELERATE's evolution and distinguishes two AI use cases. The Methods describe the institutional context, challenge selection, and GPT design. The Findings present principles that translated effectively versus those carrying structural tensions. The Discussion proposes human-AI boundaries and addresses ethical implications. We conclude with practitioner recommendations.

Conceptual Framework

From AAHE to ACCELERATE

The American Association for Higher Education's Principles of Good Practice for Assessing Student Learning (Astin et al., 1992) served as a foundational framework for assessment in higher education for over three decades. These principles emphasized that assessment should begin with educational values, be ongoing, multidimensional, and lead to improvement. However, as higher education evolved, facing new accountability demands, demographic shifts, technological change, and calls for greater equity, the field recognized a need to revisit these foundational commitments.

The ACCELERATE framework reflects broader movements in assessment scholarship from compliance-driven data collection toward improvement-focused evidence use. Kuh et al. (2015) advocate reorienting assessment culture from one dominated by external reporting requirements to one where evidence of student learning drives genuine educational improvement. Boud's (2000) concept of sustainable assessment similarly emphasizes developing learners' ability for ongoing self-evaluation and making informed judgments, while reducing reliance on teachers' or other formal sources of judgment.

ACCELERATE: Assessment Principles for Best Practice (Samuga_Gyaanam+Bheda et al., 2025) emerged from this revisitation. It articulates ten principles that foreground contemporary values: Aligned Purposefully, Continuously Cultivating, Collaborative and Co-Creative, Equity and Empowerment-Motivated, Learning-Centered, Energized by Expertise, Responsiveness-Oriented, Action-Driven, Transparent and Trustworthy, and Enduring and Evolving. The framework explicitly positions assessment as a value-driven practice that "[co-creates] learner, educational, process, people, policy, and structural well-being" (Samuga_Gyaanam+Bheda et al., 2025, p. 2). The following section examines how AI tools might engage with these human-centered principles.

Unlike AAHE's procedural framing, ACCELERATE foregrounds equity as an explicit principle, requires co-creation with learners rather than assessment of them, and extends applicability to K-20+, corporate, and organizational contexts. These commitments make the framework's human-centered orientation

more demanding to operationalize, which matters when examining whether AI tools can genuinely support such practice.

Table 1 summarizes the ten principles with definitions adapted from Samuga_Gyaanam+Bheda et al. (2025, pp. 4–7): each captures a core commitment examined in the Findings.

Table 1

The ACCELERATE Principles for Best Practice

Principle	Definition
Aligned Purposefully	Align assessment with values, mission, outcomes, policies, strategic goals, and community partner needs.
Continuously Cultivating	Ensure assessments continuously improve curricular and programmatic processes and policies in any given context.
Collaborative and Co-Creative	Intentionally collaborate and co-create with diverse partners, including learners, in the assessment endeavor.
Equity and Empowerment-Motivated	Catalyze fairness at individual and structural levels, enabling access, success, change, and informed decision-making.
Learning-Centered	Engage in broad and deep, practice-based learning for the assessment context and practice.
Energized by Expertise	Engage in training and active learning to hone design, deployment, meaning-making, and dissemination literacies.
Responsiveness-Oriented	Proactively explore and address systemic and emerging needs of learners, community partners, and context.
Action-Driven	Foreground purposeful assessment use while planning and deploying assessment in any context.
Transparent and Trustworthy	Actively build trust by maintaining transparency and outlining accountability at all stages of assessment.
Enduring and Evolving	Ideate strategies and strengthen practices to serve communities iteratively, with maturity, scalability, and sustainability.

Note. Definitions adapted and condensed from Samuga_Gyaanam+Bheda et al. (2025, pp. 4–7).

AI in Assessment: Promise and Peril

Generative AI tools are increasingly deployed in assessment contexts, with proponents highlighting efficiency gains and critics raising concerns about algorithmic opacity, practitioner dependency, and the risk that human-centered values may be flattened. Emerging scholarship identifies assessment as a nascent application area. Xia et al. (2024) note a dearth of empirical research on generative AI in higher education assessment. Bearman et al. (2024) describe an increasing separation between AI-enabled work and quality evaluation, which risks displacing judgment from human assessors unless evaluative

judgment is intentionally cultivated. Chan (2023) emphasizes transparency, human oversight, and explicit boundaries. These scholars converge on a critical distinction: AI may improve efficiency while potentially undermining the human relationships and contextual judgment that effective assessment requires. This is the tension motivating this study.

While these analyses largely consider AI as an evaluator of student work, their cautions translate to the coaching context this study examines: AI as guidance for practitioners designing assessment. The judgment-displacement risk Bearman et al. (2024) identify in evaluation can recur when AI advises rather than scores. A tool generating principle-aligned recommendations may still flatten the contested, value-laden choices that effective assessment work demands.

The Two AI Use Cases

A critical distinction shapes this study's scope: AI can be deployed in assessment in two fundamentally different ways.

AI Doing Assessment Work. AI evaluates student work directly, scoring assignments, generating rubric ratings, or producing assessment reports. This raises concerns about algorithmic bias in judging student learning.

AI Guiding Assessment Practice. AI provides guidance to practitioners on designing and improving their assessment processes, serving as a coaching tool that helps faculty develop better Program Learning Outcomes (PLOs), select appropriate measures, or create action plans.

This study examines exclusively the second use case: AI as a coaching tool for practitioners, not as an evaluator of student learning. The two use cases raise different concerns and require different boundary-setting. These two categories are not an exhaustive taxonomy of AI in assessment but the distinction most relevant to this study's scope. Other applications, such as learning analytics, automated item generation, or administrative automation, lie outside it.

Methods

Study Design

This exploratory pilot study employed retrospective analysis to examine whether AI-generated assessment guidance aligns with ACCELERATE principles. Rather than testing AI with live faculty interactions, we used assessment challenges documented in Year 1 evaluation reports. These weaknesses were identified through "Developing" ratings on specific rubric criteria or through documented evaluator concerns accompanying higher ratings. Examples included vaguely worded learning outcomes, limited assessment measures, and unclear performance targets.

Institutional Context and Year 1 Evaluations

The study context was a private, American-style university in the Arabian Gulf region, specifically its College of Arts and Sciences. The institution serves a predominantly multilingual student body, with

Arabic as the primary language for most students and English as the medium of instruction. The framework was developed in a U.S. setting. Testing an ACCELERATE-aligned AI tool outside that setting offers an opportunity to examine both the principles' transferability and the challenges of applying AI-mediated assessment guidance across linguistic and cultural boundaries.

This positioning shapes our analytical lens. Mohamed et al. (2020) argue that AI can reproduce colonial continuities through what they term "algorithmic coloniality." In this process, algorithmic systems can shape resource distribution and socio-political behavior, potentially privileging dominant value frameworks and governance standards across contexts. As researchers based in the Arabian Gulf using U.S.-developed AI tools to operationalize a U.S.-developed assessment framework, we approach this study attentive to questions of transferability. Our intent is not to critique either ACCELERATE or generative AI as inherently problematic for non-Western contexts, but rather to document what emerges when these tools encounter educational realities different from those in which they originated.

Researcher Positionality

We approach this study as assessment practitioners embedded in the context we examine. The first author serves as an assessment practitioner responsible for institutional and programmatic assessment implementation; the second author contributed disciplinary expertise in educational research. Both authors were trained in Western academic traditions and hold higher education degrees from U.S. and Arab institutions where instruction was in English. Both work within a U.S.-style institution that operates in English despite serving a predominantly Arabic-speaking student population. Our familiarity with ACCELERATE's source context enabled a close reading of the framework's intent. At the same time, our daily work in a Gulf Cooperation Council (GCC) institution surfaced assumptions that might remain invisible to researchers operating entirely within U.S. higher education. We offer observations from one location, contributing data points toward understanding how assessment principles travel across contexts.

Annual Programmatic Assessment Cycle

During the 2023–24 academic year, the College of Arts and Sciences initiated the first year of a three-year programmatic assessment cycle: program-level (not course-level) assessment focusing on PLOs across degree programs in eight departments. Each department submitted an assessment plan and annual report evaluated using the institution's Academic Assessment Plan and Report Evaluation Rubric (Appendix A). The rubric assessed seven criteria: (1) Mission Statement, (2) Learning Outcomes, (3) Assessment Measures, (4) Targets/Benchmarks, (5) Assessment Results, (6) Use of Evidence, and (7) Action Items. Each criterion was rated on a four-level scale: Exemplary, Acceptable, Developing, or No Evidence. Year 1 findings revealed systematic challenges across departments, including vaguely worded PLOs, over-reliance on single assessment methods, undefined target terminology, narrative results without quantitative data, and insufficient reflection when targets were met.

The Academic Assessment Plan and Report Evaluation Rubric anchors a recurring annual review cycle that the institution applies to each program. Departments produce two coordinated documents at the cycle's outset. The first is an assessment plan that articulates program learning outcomes, direct and

indirect measures, and targets against which results will later be judged. The second is an annual report that returns to those same outcomes with results, evidence of use, and proposed action items. Both documents are evaluated against the rubric's seven criteria, with each criterion scored as Exemplary, Acceptable, Developing, or No Evidence. Departments receive structured feedback that names which criteria fall into which level, and a closing summary checklist (Table A2) consolidates the rubric judgment under two headings: Strategic Plan and Strategic Report. The cycle then advances: action items identified in one annual report become evaluation anchors for the next. Year 1 of the three-year cycle generated the source pool for this study's seven scenarios. The eight departmental evaluations from that year documented the systematic weaknesses, including "Developing"-rated criteria, that the GPT would later be tested against.

Challenge Selection

Seven challenges were extracted from Year 1 evaluation reports: four drawn from criteria rated "Developing" and three from evaluator-flagged concerns on criteria rated "Acceptable" or "Exemplary." Challenge identification was performed by the assessment coordinator (the first author) applying the institutional Academic Assessment Plan and Report Evaluation Rubric (Appendix A) to identify items receiving a "Developing" rating or a flagged evaluator concern. Challenges were selected to cover different ACCELERATE principles and rubric criteria: (1) Vague PLO: tests Learning-Centered principle; (2) Limited measures: tests Equity and Empowerment-Motivated principle; (3) Unmet target: tests Action-Driven principle; (4) Targets met, no reflection: tests Continuously Cultivating principle; (5) Misaligned indirect measure: tests Aligned Purposefully principle; (6) Undefined target: tests Transparent and Trustworthy principle; (7) Narrative results: tests Transparent and Trustworthy plus Action-Driven principles. The seven scenarios did not cover all ten principles one-to-one. Principles with checkable criteria, such as alignment or transparency, were tested directly: each scenario contained a documented weakness, and we examined whether the GPT's guidance addressed it. The four principles, later classified as tensions, were assessed differently: we observed how the GPT behaved across all seven interactions. Each challenge was the source material for a testing interaction. The scenario's documented context provided the researcher's responses to the GPT's eight-question intake dialogue (Q1–Q8 in v4; see Appendix B §2.2 and §3.3). The guidance the GPT produced in return was captured as the transcript. Seven scenarios produced seven transcripts; each scenario was run once with v4, with the first output captured and analyzed without refinement or regeneration.

All seven scenarios followed a single testing protocol fixed in advance (detailed in Appendix B). Mode B (Thinking Partner) was selected as the interaction mode for every scenario, ensuring the GPT engaged the researcher with clarifying questions before generating guidance rather than producing a directive task list. Each scenario opened with the v4 context-check, an opening prompt distinguishing U.S. from non-U.S. assessment contexts. It was answered as a non-U.S. context (Arabian Gulf, English-medium, MSCHE-equivalent accreditor), which routes the GPT to the regional adaptation pathway documented in Appendix B §1.7. The structured Q1–Q8 intake then ran in full for each scenario. The eight numbered questions (with Q7 sub-numbered Q7a–Q7c) covered cycle stage; challenge type; question pacing; whether the work targets specific rubric criteria; the decision the assessment will inform; affected and missing voices; program level, discipline, and modality; and practical constraints. Researcher responses to these questions drew verbatim from the documented Year 1 evaluation

reports, ensuring the GPT received the same context a faculty member would supply when raising the same concern in practice. The guidance returned by the GPT after intake completion was captured as a single transcript per scenario. Design rationale for the protocol itself remains in Appendix B: why Mode B was chosen over Mode A (direct guidance with minimal questioning), why intake was numbered rather than adaptive, and why the context check was made explicit. Sections 1.7, 3, and 8 hold these design decisions, while the present paragraph documents the protocol as applied.

GPT Design Process

The custom GPT was built in OpenAI's custom GPT builder on the GPT-5.1 base model, with development and testing running from December 2025 through January 2026. Sessions on December 30–31, 2025 piloted the same scenarios on an earlier prototype build under superseded instructions; those outputs are excluded because they predate the v4 instruction set. The v4 configuration was tested on January 2–3, 2026 across all seven scenarios, and v5 was informally re-checked by the first author between January 8 and 14, 2026. Custom v4 instructions totaled 7,637 characters against the 8,000-character limit (v1 through v5 design evolution in Appendix B; full v5 text in Appendix C). All outputs analyzed in the Findings are v4 responses; the publicly deployed GPT runs v5.

Each principle was mapped to a prompt strategy: the behavior the GPT should exhibit to express that principle (Table 2). Some strategies were written directly into the instruction text; others relied on expected base-model behavior and were checked during testing, with design decisions documented for transparency (see Appendix B).

Table 2

Principle-to-Prompt Mapping

Principle	Prompt Strategy
Aligned Purposefully	Systematic alignment checking; flag mismatches between PLOs, measures, and targets
Continuously Cultivating	Always prompt reflection, even when targets are met
Collaborative/Co-Creative	Frame outputs as adaptable options; ask about stakeholder involvement
Equity & Empowerment	Embed equity lens in all recommendations; suggest multiple modalities; use non-deficit language
Learning-Centered	Emphasize iteration and practice-based learning; normalize imperfection
Energized by Expertise	Explain WHY suggestions work, not just WHAT to do; build assessment literacy

Responsiveness-Oriented	Ask clarifying questions about context; acknowledge limitations of not knowing local conditions
Action-Driven	End every response with specific actions, timelines, owners, and decision rules
Transparent & Trustworthy	Flag vague language; require operational definitions; insist on N/n/% reporting
Enduring & Evolving	Prompt sustainability thinking; ask about scalability and future cycles

Analytical Framework

Each GPT response was examined through two analytical lenses. Lens 1 assessed principle translation against three criteria, each addressing a distinct dimension of the GPT's guidance: Alignment asks whether the guidance reflects the relevant ACCELERATE principle as articulated in Samuga_Gyaanam+Bheda et al. (2025). Applicability asks whether the guidance is implementable by a faculty member with existing institutional resources and the discretion their role permits. Effectiveness asks whether implementation would address the documented rubric weakness the Year 1 evaluation flagged. Lens 2 assessed structural tensions by examining whether the GPT could model versus ensure principle enactment, and whether users could bypass human-centered requirements while still receiving technically sound guidance.

The three criteria were applied qualitatively through close reading rather than as a fixed coding rubric, because validated rubric development presupposes prior validation of the framework itself. An earlier version assigned ordinal ratings on a four-point scale; these ratings were removed to preserve the study's developmental framing rather than imply validated measurement. Validated measurement would require psychometric grounding the field does not yet have for ACCELERATE. That grounding includes established construct definitions calibrated against expert consensus, tested inter-rater reliability across multiple coders, and evidence that the rating scale discriminates meaningfully between levels of principle alignment. Reporting ordinal scores in advance of that groundwork would imply a measurement rigor this study does not claim and would mischaracterize an exploratory pilot as a validation exercise. This work is positioned as an early step toward the rubric development that the framework will require, not as an application of rubric standards that do not yet exist. Ordinal rating with two independent coders is positioned as priority future work given the pilot framing and the framework's recent emergence. Full quantitative validation will require a developed pool of practitioners with sustained framework engagement, which will become available as ACCELERATE enters wider use. The Limitations section expands on this developmental-stage rationale.

The single-coder design reflects three constraints. The pilot is positioned as developmental rather than as a validation exercise. The ACCELERATE framework entered the literature in late 2025, and the pool of practitioners with sustained framework engagement remains small. The psychometric prerequisites for validated ordinal coding described above remain unmet for ACCELERATE. Given these constraints, the design substituted structured qualitative interpretation for independent quantitative coding. It used a documented AI-assisted analytical process but did not treat that process as a substitute for inter-rater reliability between independent coders. The first author conducted the analysis, comparing

GPT outputs against the ACCELERATE Position Statement and the institutional challenges documented in evaluation reports. Claude (Anthropic) produced first-pass analytical drafts of the same outputs, which the first author reviewed, corrected, and adjudicated as final arbiter. Disclosing this AI-assisted drafting makes the analytical process transparent; it does not provide independent validation. The Disclosure of Artificial Intelligence Use section describes the full protocol for this AI assistance; Appendix D contains the per-scenario annotations supporting the findings. All GPT outputs quoted in this article were verified against the archived testing transcripts.

Findings

Overview of Results

All seven challenges were tested, and all ten ACCELERATE principles were observed in GPT responses. Rather than scoring outputs quantitatively, we examined each response through two analytical lenses. First, we assessed principle translation: Did the guidance reflect the principle's core values as articulated in the ACCELERATE Position Statement? This lens included implementation potential: Could a faculty member implement this guidance, and would implementation likely address the weakness each scenario documented from the Year 1 evaluations? Second, we assessed structural tensions: whether the GPT could model versus ensure principle enactment. We assessed these qualities through researcher judgment, comparing GPT outputs against both the ACCELERATE Position Statement and the specific institutional challenges documented in evaluation reports.

This qualitative analysis revealed an important distinction. While the GPT successfully incorporated principle-aligned language for all ten principles, six translated effectively to AI-generated guidance. The remaining four (Collaborative and Co-Creative, Energized by Expertise, Responsiveness-Oriented, and Enduring and Evolving) carried structural tensions requiring ongoing human mediation. These tensions surface when examining whether AI can ensure principle enactment, not merely model it. This pattern, six principles that translate to AI support versus four that carry structural tensions, frames the manuscript's central interpretation. The principles fall into translatable and tension-bearing categories rather than two independent dimensions of analysis. The sorting criterion deserves explanation: every principle involves human follow-through the GPT cannot verify. A principle was classified as a tension only where that follow-through is constitutive of the principle itself rather than a condition of its use. The four tensions were anticipated at design time; testing illustrated these constraints rather than produced them.

Why might some principles translate more readily to AI support than others? The pattern across these six is that they turn on systematic, rule-like operations the GPT performed across the seven single-pass outputs: alignment checking, reflection prompting, structured templating, clarity flagging. These are tasks the GPT may perform well because the principles themselves encode procedures a chatbot can follow.

Principles That Translated Effectively

Through close reading of GPT outputs against the ACCELERATE Position Statement, we identified six principles that showed strong translation to AI support in single-pass testing. "Strong translation" here means the GPT's guidance, in each single-pass output, reflected the principle's core values, was

practically implementable with existing institutional resources, and would likely address the documented assessment weakness. For each principle, we assess both dimensions: principle alignment (did the guidance reflect ACCELERATE values?) and implementation potential (would this guidance address the Year 1 documented weakness?). We acknowledge that this assessment reflects researcher judgment rather than validated evaluation methodology; we discuss implications of this limitation and propose directions for rigorous evaluation below.

Aligned Purposefully. The GPT checked coherence within the PLO-measure-target chain in each of the seven single-pass outputs, extending to targets where scenarios supplied them. When presented with a department using letters of recommendation requests as an indirect measure for an analytical PLO, the GPT explained the misalignment already flagged in the Year 1 evaluation. It stated: "It can easily be driven by ambition, advising culture, faculty willingness, or cohort size—so it's not very Aligned Purposefully to PLO2." It then provided alternatives directly tied to the PLO's construct. Alignment checking is systematic, logical, and thus well-suited to AI support. *Implementation potential*: The Year 1 evaluation rated this department "Acceptable" but flagged the alignment concern. The GPT's diagnosis and alternatives, if implemented, would address the critique.

Continuously Cultivating. We tested this principle with a scenario where a department reported meeting all targets (92%, 100%, and 79% student achievement against 70% targets). When presented with this "success" scenario, the GPT did not treat met targets as endpoints. Instead, it introduced a "sustain / tweak / investigate" framework and flagged a ceiling effect concern: "Did '100%' on PLO2 indicate mastery—or an instrument that can't discriminate?" The GPT also prompted the department to define what improvement beyond met targets would look like. This shows that prompting reflection even when results appear successful can be built into AI instructions as scripted behavior. *Implementation potential:* The Year 1 evaluator rated this department "Developing" on Evidence and Action Items because targets were met without reflection. The GPT's framework would likely achieve "Acceptable" ratings.

Equity and Empowerment-Motivated. When a department relied solely on tests in one introductory course, the GPT built in an "Equity Check" section. The section prompted examination of hidden assumptions through two questions: "Do prompt directions assume a single cultural lens, prior schooling style, or language register?" and "How will students' feedback inform revision of the prompt/scoring guide next cycle?" It used non-deficit language throughout and framed gaps as signals to investigate rather than deficit conclusions. Equity considerations can be systematically embedded as a lens applied to every recommendation. *Implementation potential:* The Year 1 evaluator flagged that relying solely on tests was limited for two PLOs. The GPT's multi-point evidence strategy and Equity Check would address this. One boundary of this translation surfaces in the Discussion: the GPT could prompt equity questions it did not itself apply to its own suggestions. This limit shares the structure of the tensions described below.

Learning-Centered. The GPT framed outputs as material for refinement rather than adoption. When asked to improve a vague PLO, it provided four rewrite options presented as "workshop options you and faculty can adapt." It framed them as directions to "refine, not to adopt as-is" and offered to help "refine one stem into a measurable PLO." Iterative, practice-based framing translated effectively to

output language in this single-pass test. *Implementation potential:* The Year 1 evaluation rated this department "Developing" on Learning Outcomes because PLOs lacked observable action verbs. The GPT's rewrite options, featuring verbs like "explain," "evaluate," and "justify," would directly address this deficiency, likely achieving "Acceptable" ratings.

Action-Driven. Each of the seven single-pass responses concluded with a structured action plan specifying actions, time estimates, who to involve, and completion criteria, with checkpoint and after-implementation prompts. We tested this principle with a scenario drawn from a department where the target called for 75% of students to score 80% or higher, while only 30% passed the midterm at the 70% threshold. Presented with this challenge, the GPT elicited root-cause hypotheses and pressed for an "Action-Driven response that's actually usable, not just 'we'll try harder next term.'" Plans included feasibility checks matched to each scenario's stated constraint; here, the checkpoint asked whether the action was "feasible with limited faculty time." In these scripted single-pass tests, the GPT consistently produced structured plans with accountability elements. *Implementation potential*: The Year 1 evaluator rated this department "Exemplary" on Action Items. The GPT's guidance would maintain this strength while providing additional structure for addressing the specific midterm performance gap.

Transparent and Trustworthy. The GPT flagged vague language and required operational definitions. When we tested it with the scenario data, supplying the evaluator's documented concerns, the GPT articulated the definitional and statistical bases of the two problems and how to repair them. First, the department's target stated "70% is expected to pass," but never defined what "pass" meant: a grade of C or higher? A score of 70%? A rubric threshold? Without this definition, the target was unmeasurable. Second, even setting aside the undefined term, the reported result ("overall average of 3 tests was 66.09%") did not match the target's structure. The target asked what percentage of students would pass, while the result reported the mean score across students. These are fundamentally different statistics: a class could have a 66% mean score while 75% of students exceed a passing threshold if scores cluster just above it. Conversely, few students could pass despite a reasonable mean if scores cluster just below. The GPT explained both the definitional ambiguity and the statistical mismatch ("two different metrics, so they can't directly answer each other"), showing that systematic clarity-checking may translate well to AI support. It then provided an operationalization template. In the unmet-target scenario, the GPT also flagged, unprompted, the report's mixing of an 80% target standard with a 70% pass threshold ("avoid mixing ≥70 and ≥80 without explanation"). *Implementation potential*: The Year 1 evaluator rated this department "Developing" on Assessment Results. Given the evaluator's flags, the GPT explained the underlying problems; its operationalization template would resolve both.

The next four principles share a different structure: they ask for engagement sustained across time rather than a single well-formed output. Collaborative and Co-Creative requires partners who remain engaged across the full assessment cycle. Energized by Expertise develops through sustained struggle with practice. Responsiveness-Oriented senses context as it shifts. Enduring and Evolving carries frameworks through revision. These are structural tensions the GPT may surface but cannot resolve alone. The work is longitudinal and relational (Table 3).

Principles with Structural Tensions

Four principles carry structural limitations that define appropriate human-AI boundaries. These appear to be inherent constraints of AI-generated guidance for human-centered values rather than failures of implementation.

Collaborative and Co-Creative. The v4 flow instructed the GPT to ask "Anything to adjust before I put together your action plan?" before drafting each action plan (per-scenario phrasing varied; see Appendix D). The GPT also foregrounded stakeholder involvement through its intake question "Who is most affected, and whose voice is currently missing or least powerful?" It generated prompts such as "Which participation model would actually change power, not just 'consult'?" However, AI can model collaborative values but cannot ensure collaborative practice. Users could take AI guidance and implement it unilaterally, feeling confident they received good advice without ever consulting the stakeholders that the principle requires.

This may create a false sense of principled practice. A faculty member who receives AI-generated PLO revisions framed as "workshop options you and faculty can adapt" may feel they have honored the Collaborative and Co-Creative principle simply by reading that phrase. That feeling may persist even if they proceed to implement revisions without consulting students, colleagues, or community partners. The AI has no mechanism to verify that diverse voices were actually consulted. Even if programmed to ask "Did you share this with your department?" it cannot verify the answer. A user could respond "yes" without having done so, or describe perfunctory consultation as genuine collaboration. The AI has no reliable means to distinguish authentic partnership from compliance language. The principle requires partners who stay engaged throughout the assessment endeavor; AI provides advice and exits, creating what might be called "collaboration theater": the appearance of participatory values without their substance.

Energized by Expertise. The GPT explained WHY suggestions work, not just what to do. It linked shared scoring guides to fairness across inconsistent sections, questioned whether 100% achievement reflected mastery or an instrument unable to discriminate, and explained why letter of recommendation counts don't measure PLO achievement. In principle, this should build assessment literacy. However, convenience may prevent deep learning.

The Energized by Expertise principle assumes that assessment capability develops through "experience and active, focused learning" (Samuga_Gyaanam+Bheda et al., 2025, p. 5). This is the kind of effortful engagement where practitioners wrestle with ambiguous situations, make mistakes, receive feedback, and gradually develop judgment. AI efficiency may shortcut this developmental process. When a faculty member can generate a well-structured action plan in thirty seconds, what incentive exists to spend hours understanding why that structure works? When AI explains construct validity in a sidebar, does that explanation produce the same understanding as struggling to figure out why a measure fails to capture what was intended?

Two patterns are possible. In the first, users read explanations, internalize principles, and gradually need AI less as their own expertise grows. In the second, users skip explanations, take outputs, and

become increasingly dependent on a tool they do not understand, and on principles they have never genuinely engaged in. The GPT cannot control which pattern emerges. Our testing protocol, which used scripted scenarios to examine GPT output quality, could not capture which pattern would emerge in authentic use. Indeed, the very features that make AI helpful, including speed, comprehensiveness, and structured output, may be the features that undermine the struggle through which genuine expertise develops. This tension may not be resolvable through better prompt engineering; it reflects a fundamental question about whether convenience and capability-building are compatible goals.
Responsiveness-Oriented. The GPT asked clarifying questions about context and adapted recommendations to constraints shared by users. However, AI is context-seeking but not context-sensing.

One might argue this limitation could be addressed through more sophisticated intake protocols, requiring users to answer detailed questions about departmental dynamics, resource constraints, stakeholder relationships, and institutional politics before receiving guidance. Our GPT included such intake questions, and in our judgment they improved response relevance. However, even comprehensive intake cannot capture what human practitioners sense intuitively: the Dean's body language when assessment is mentioned, or the unspoken tension between two faculty members who must collaborate. Nor can it capture the difference between stated priorities and actual behavior, or the way a department's culture has shifted since last semester.

More fundamentally, context is not static. A recommendation that fits perfectly in September may become inappropriate by November. This could happen because 1) budget cuts eliminated planned resources, 2) a key faculty member resigned, 3) a new hire brought relevant expertise that opens possibilities the original guidance didn't consider, or 4) an unexpected grant created capacity for more ambitious assessment approaches. The GPT cannot observe these shifts, cannot adjust its earlier advice in light of new circumstances, cannot check in and ask, "How did that work out?"
Enduring and Evolving. The GPT included "After Implementation" sections with sustainability guidance in every action plan, prompting long-term thinking about future assessment cycles. However, AI provides episodic support without longitudinal presence.

Recent AI platforms do offer "projects" with persistent memory, and this might appear to address the longitudinal limitation. A department could, in theory, maintain an ongoing AI project that accumulates context across assessment cycles. However, memory is not relationship. An AI that "remembers" last year's recommendations cannot evaluate whether those recommendations were implemented faithfully, whether they produced intended effects, or whether changing circumstances have rendered them obsolete. It cannot notice that the faculty champion who drove last year's improvements has left the institution, or that the new Dean has different priorities, or that student demographics have shifted in ways that require reconsidering the assessment approach.

The Enduring and Evolving principle calls for assessment that is "meaningful to partners in both the short term and the long term" (Samuga_Gyaanam+Bheda et al., 2025, p. 7). Meaning-making over time requires not just data persistence but interpretive continuity: someone who understands why certain decisions were made, who can distinguish between surface compliance and genuine improvement,

who can advocate for resources when assessment reveals needs. AI memory stores information; human practitioners build institutional assessment culture. The principle requires the latter.

Table 3

Principle Translation Summary

Principle	Translation	Core Finding	Implementation Potential
Aligned Purposefully	Effective	Systematic checking well-suited to AI	High: Would likely address evaluator's alignment concern
Continuously Cultivating	Effective	Reflection prompted even when targets met	High: Directly addresses "met targets, no reflection" gap
Collaborative/Co-Creative	Tension	Cannot ensure collaboration occurs	Uncertain: Guidance sound, but enactment requires human follow-through
Equity & Empowerment	Effective	Equity lens systematically embeddable	Moderate: Addresses measure diversity; cultural responsiveness untested
Learning-Centered	Effective	Iterative framing translates well	High: Would likely improve "Developing" PLO ratings
Energized by Expertise	Tension	May build OR undermine literacy	Uncertain: Depends on user engagement pattern
Responsiveness-Oriented	Tension	Context-seeking, not context-sensing	Limited: Cannot adapt to unstated constraints
Action-Driven	Effective	Structured output natural for AI	High: Would likely maintain the department's existing action-item strength
Transparent & Trustworthy	Effective	Clarity focus well-suited	High: Would likely resolve definitional and statistical gaps
Enduring & Evolving	Tension	Episodic, not longitudinal	Limited: Single-interaction support cannot sustain culture

The Discussion examines what this pattern means for how practitioners should navigate AI-supported assessment work.

Discussion

What the Tensions Share: Relationship Over Transaction

The cross-cutting interpretation is the manuscript's translatable-vs-tension-bearing distinction; this subsection examines what the tension-bearing four principles share.

The four principles carrying structural tensions (Collaborative and Co-Creative, Energized by Expertise, Responsiveness-Oriented, and Enduring and Evolving) share a common characteristic: they require an ongoing relationship rather than episodic support. Collaboration requires partners who stay engaged throughout the assessment endeavor. Expertise develops through struggle and iteration over time. Responsiveness requires sensing context as it shifts, not merely asking about it once. Sustainability requires presence across assessment cycles, not point-in-time advice. AI can generate structured, well-

formed guidance in a single interaction, but cannot maintain the relational continuity these principles demand.

Rather than move immediately to equity implications, several untested scenarios merit acknowledgment. First, a department facing an accreditation deadline may experience conflict between Action-Driven urgency and Collaborative deliberation; the GPT was not tested on principle conflicts. All scenarios assumed good-faith users genuinely seeking improvement. Second, the GPT could potentially be used to generate "compliant-looking" documentation for decisions already made, or to produce assessment language that satisfies external reviewers without reflecting genuine practice. This might be termed "legitimacy laundering." Third, a user could input a fabricated challenge, receive principle-aligned guidance, and incorporate AI-generated language into reports, creating the appearance of principled practice without its substance. This concern extends beyond ACCELERATE to any AI tool generating professional documentation that others will evaluate.

Equity Implications of AI-Mediated Assessment

The equity implications of AI-mediated assessment warrant particular scrutiny. Montenegro and Jankowski (2020) warn that "an assessment process that is not mindful of equity can risk becoming a tool that promotes inequities, whether intentional or not" (p. 4). They further argue that "listening to the voices of those historically silenced is an essential element of equity-minded assessment" (p. 10). These arguments raise questions about how AI-mediated assessment can remain accountable to marginalized perspectives. Henning et al. (2022) emphasize that equity-centered assessment demands ongoing critical reflection about power dynamics, cultural responsiveness, and whose knowledge is valued. Tai et al. (2023) argue that inclusion should be embedded from conception rather than retrofitted, raising questions about whether AI-mediated assessment can support culturally responsive practices.

Our institutional context adds a dimension to these equity concerns that warrants explicit attention. The ACCELERATE framework was developed within a U.S. professional association context and, in our reading, draws primarily on literature, examples, and feedback from U.S.-based assessment communities. This is not a critique of the framework's validity; rather, it reflects the reality that assessment scholarship, like most educational research, emerges from particular contexts. As Henrich et al. (2010) observe, findings from Western, Educated, Industrialized, Rich, and Democratic (WEIRD) populations do not automatically generalize to other cultural settings. The ACCELERATE principles articulate values, such as equity, collaboration, and transparency, that may be universal in aspiration but whose operationalization likely requires contextual adaptation. Our study represents an early attempt to apply ACCELERATE principles outside the U.S. context in which they were developed; our findings may illuminate both the framework's transferability and areas where local adaptation is warranted. The GPT we employed was likely trained predominantly on English-language, Western-origin data. This compounds questions about whether AI-generated guidance can adequately recognize forms of knowledge demonstration, pedagogical values, or assessment practices that differ from Anglo-American academic norms.

The ACCELERATE emphasis on collaboration resonates with the students-as-partners (SaP) movement in higher education. Mercer-Mapstone et al.'s (2017) systematic review of 65 studies emphasizes that

meaningful SaP is characterized by reciprocity and shared responsibility, and reports outcomes such as increased student engagement and stronger responsibility for learning. This raises concerns about AI tools in assessment: they may provide structured feedback, but it remains to be seen whether they support rather than displace dialogic relationships through which students develop assessment capability.

These tensions also share an equity dimension. When users skip collaboration, whose voices are excluded? ACCELERATE rightly highlights those with least institutional power: adjunct faculty, students, and community partners most affected by assessment decisions yet least consulted. In contexts like ours, additional dimensions merit consideration. These include students whose first language is Arabic and who may express understanding differently than English-medium assessment instruments expect. They also include faculty trained in educational traditions outside the Anglo-American model, and community stakeholders whose conceptions of graduate competence may not align with assumptions embedded in Western-developed rubrics. These are not limitations of ACCELERATE's equity principle, which explicitly calls for attention to context and power; rather, they illustrate the interpretive work required when applying any framework across cultural boundaries. AI tools add complexity: they cannot prompt users to seek marginalized voices if the tools themselves, likely trained on predominantly Western data, do not recognize which voices might be absent.

If AI creates dependency rather than building assessment literacy, faculty with less access to professional development may rely more heavily on AI tools, widening expertise gaps between well-resourced and under-resourced programs. If AI cannot sense context, it cannot detect when students are underserved by assessment practices that appear adequate on paper. Our multilingual GCC context illustrates this concretely: assessment measures developed for English-dominant settings may not capture competence demonstrated through Arabic-influenced discourse patterns. Any context differing from the training data of AI systems faces analogous challenges; human practitioners must supply the contextual judgment that AI lacks, reinforcing ACCELERATE's emphasis on expertise and responsiveness across cultural and linguistic boundaries.

A concrete example from our testing illustrates this limitation. When addressing the limited-measures challenge, the GPT suggested "a short writing-based task (e.g., compare/contrast theories; explain a perspective with evidence)" as an equitable alternative to tests. This recommendation embeds assumptions about academic argument rooted in Anglo-American rhetorical traditions: individual authorship, explicit thesis statements, linear argumentation, and citation practices treating written text as primary evidence. The GPT's Equity Check prompted users to ask whether "prompt directions assume a single cultural lens," an excellent question. Yet the GPT did not, unprompted, recognize that its own suggestion of a compare/contrast writing task was that single lens. Practitioners in our context must supply this recognition; the AI cannot.

AI's episodic presence cannot advocate for the institutional resources and sustained commitment that equity-focused assessment requires. More broadly, these findings illuminate the distinction between AI guiding assessment practice and AI contributing to actual assessment improvement. For six principles, the GPT successfully accomplished both. It produced principle-aligned guidance (the coaching function) that, if implemented, would address the documented evaluation findings or, in one

case, maintain the performance they documented (the improvement function). For four principles, however, the distinction reveals an important gap: the GPT can generate technically sound guidance aligned with principle values, but cannot ensure the human follow-through that distinguishes principle-aligned outputs from principled practice. Practitioners might receive excellent AI coaching while still failing to achieve the relational and developmental outcomes ACCELERATE envisions.
No decision-support tool can ensure ethical implementation: that responsibility rests with governance and practice. The GPT can model each ACCELERATE principle's application, but translating those models into principled practice is a human responsibility.

Implications for Human-AI Boundaries

This subsection extends the AI-guiding-practice framing established earlier (Two AI Use Cases) into the Discussion. The study examines integrating AI into assessment practitioner support workflows, not AI as an evaluator of student work; that scope shapes what the principle-translation analysis can and cannot establish. The technical-versus-relational distinction that surfaced in the Findings operates as a boundary condition for AI-guiding-practice: AI may expedite the technical work while practitioners retain responsibility for the relational work that no decision-support tool can supply.

What AI adds to this general limit is specific. Research on automation complacency documents how users may over-rely on technically sound automated outputs without verifying them (Parasuraman & Manzey, 2010). AI extends this concern by producing such outputs at scale, allowing users to bypass the human-centered deliberation these principles require even as the outputs appear to satisfy them. The finding that certain ACCELERATE principles require human mediation that technology cannot replace aligns with emerging scholarship on human-AI collaboration. Vallor (2015) discusses the concept of "moral deskilling," referring to the erosion of human moral skills and judgment that can occur when decision-making is increasingly delegated to automated systems. Rinta-Kahila et al. (2023) empirically document similar patterns of professional skill erosion due to sustained reliance on cognitive automation, including AI-based systems. Agudo et al. (2024) show that incorrect AI support can negatively shape subsequent human judgments, which decreases accuracy in the human-in-the-loop processes. Taken together, these findings point toward AI support for assessment designed to augment, rather than replace, human expertise, thereby preserving practitioners' capacity for contextual and critical judgment (Table 4).

Table 4

Recommended Task Allocation

Assessment Task	AI Role	Human Role
PLO drafting	Generate options, explain rationale	Select, adapt, validate with stakeholders
Measure design	Suggest menu, flag equity concerns	Choose based on context, pilot test
Target setting	Operationalize definitions	Negotiate with faculty, ensure buy-in

Results reporting	Structure data, translate narrative	Interpret meaning, communicate to stakeholders
Action planning	Structure plan, suggest timelines	Implement, adjust, follow through
Reflection	Prompt questions, flag concerns	Engage genuinely, build relationships

These recommendations position AI as a tool for technical work and humans as guarantors of relational work; practitioners remain the assessment actors, and AI provides scaffolding, not replacement.
We offer this technical/relational distinction as a heuristic rather than a sharp boundary. In practice, most assessment tasks blend both dimensions: PLO drafting involves technical language work but also relational negotiation about what a program community values; results interpretation requires technical data structuring but also meaning-making that emerges through dialogue with stakeholders. Even alignment checking, seemingly technical, involves interpretive judgment about what counts as coherence. These judgments are shaped by disciplinary norms and institutional context. The distinction's value lies not in cleanly sorting tasks into categories but in prompting practitioners to ask: Which aspects of this task benefit from AI's consistency and speed? Which aspects require my contextual knowledge, relational presence, or professional judgment? Answers will vary by task, context, and practitioner expertise; what remains constant is the need to ask the questions.

Ethical and Design Implications of Framework Translation

Translating ACCELERATE into prompt engineering required interpretive decisions at every step. How should Collaborative and Co-Creative manifest when users interact alone with AI? We chose to have AI recommend collaboration, require user confirmation before drafting, and present every plan as a draft for revision; another designer might require multi-user interfaces, stakeholder approval workflows, or documentation of who was consulted.

This interpretive variability has practical implications. If different institutions develop their own ACCELERATE-aligned chatbots, each encoding different readings of the principles, practitioners moving between institutions may encounter contradictory guidance presented with equal confidence. More subtly, an institution's existing assessment culture will shape how it interprets principles when configuring AI tools. The tools will then reinforce that interpretation back to practitioners, potentially calcifying local readings that diverge from the framework's intent. An institution that already treats assessment as compliance may configure AI that emphasizes documentation over improvement; an institution focused on faculty development may configure AI that emphasizes learning over accountability. The AI becomes a mirror that reflects and amplifies existing assumptions rather than a guide that challenges them. Users receiving AI guidance receive that interpretation, potentially without engaging the original Position Statement's nuance.

Safeguards for Institutional Deployment

These risks may require safeguards when institutions deploy AI tools that operationalize assessment frameworks. We propose five interconnected approaches:

First, participatory configuration involves diverse stakeholders, including students, adjunct faculty, and community partners, in reviewing and adapting AI tool instructions before deployment.
Second, implementation fidelity monitoring tracks whether AI tool use adheres to stated design principles (Carroll et al., 2007).

Third, cross-institutional review processes allow professionals from different contexts to evaluate each other's AI configurations, surfacing cultural assumptions invisible within a single institutional culture.

Fourth, concurrent professional development positions AI tools as supplements to, not substitutes for, assessment education. Engaging practitioners with original framework documents could reduce expertise erosion risks (Rinta-Kahila et al., 2023).

Fifth, provisional design treats AI configurations as provisional rather than permanent, with scheduled review points preventing calcification of interpretations that should remain fluid.

These safeguards share a common logic: they preserve human judgment and relationship at precisely the points where AI mediation creates risk. One further consideration applies across all five: institutional data entered into commercial AI platforms is subject to the provider's data-handling terms, so deployments should route through institutional agreements rather than individual accounts. These safeguards also align with emerging policy frameworks for educational AI that emphasize transparency, accountability (Chan, 2023), and human oversight (United Nations Educational, Scientific and Cultural Organization, 2021).

Bias and the Authority of AI-Mediated Guidance

The potential for algorithmic bias in educational AI is well-documented. Baker and Hawn's (2022) review identifies multiple mechanisms through which AI systems can perpetuate or amplify existing inequities. Kizilcec and Lee (2022) note that fairness in educational algorithms is not resolved by one metric but requires ongoing human oversight and contextual interpretation. Gándara et al. (2024) provide empirical evidence of racialized disparities in student-success prediction models in higher education. This reinforces concerns about uncritical adoption of AI in assessment practices.

A practical risk emerges: if practitioners use an ACCELERATE-based chatbot instead of reading the original Position Statement, AI becomes a mediating layer between framework and practice. The chatbot's interpretation, shaped by one design team's choices, may be treated as authoritative.
This dynamic mirrors concerns about secondary sources in scholarship: students who read textbook summaries instead of primary sources miss nuance, context, and the opportunity to form their own interpretations. Practitioners who receive ACCELERATE principles filtered through AI miss the Position Statement's careful language, its explicit "why" explanations for each principle, and its acknowledgment that principles are provisional and meant to evolve. The AI presents definitive-seeming guidance; the original document invites ongoing interpretation.

Worse, AI-mediated frameworks may acquire authority because they feel more accessible. A busy faculty member facing an assessment deadline may find AI-generated action steps more immediately useful than the Position Statement's philosophical framing. They may never realize that the action

steps encode contestable interpretations that could have been read differently. The AI would move from secondary source to replacement source, its derivative status invisible to users who experience it as direct guidance. This is especially concerning for principles like Equity and Empowerment-Motivated, where the Position Statement's language about "those who lack contextual power" (Samuga_Gyaanam+Bheda et al., 2025, p. 5) carries moral weight that algorithmic operationalization may flatten.

Cultural Assumptions in Framework Translation

The algorithmic coloniality concern introduced earlier has practical implications. Nyaaba et al. (2025) document that leading generative AI systems exhibit systematic Western-centric tendencies across multiple dimensions, from curriculum content to language support to representation in outputs. They argue that "equity in AI cannot be achieved through technical refinement alone but requires critical human agency" (p. 19).

Our study offers a modest illustration. The ACCELERATE Process we developed carried explicit equity-focused instructions drawn from a framework that foregrounds cultural responsiveness. Even so, in our testing, it did not independently flag when its guidance reflected Anglo-American academic norms that may not transfer to our context. When it suggests an "oral/presentation artifact" with a "viva-style explanation of perspectives" as a non-test alternative, it may import an Anglo-American examination genre. Whether and how our Arabic-speaking students demonstrate expertise in English orally may differ in ways the AI cannot anticipate. This is not a failure of ACCELERATE's equity principle, which explicitly calls for contextual responsiveness, but a limitation of AI systems that cannot access the local knowledge human practitioners possess. The finding reinforces our argument for human mediation: some principles require human judgment not only because they are relational, but because AI systems carry cultural assumptions only contextually situated practitioners can identify and address.

Conducting this study in a GCC context rather than a U.S. institution may have heightened our sensitivity to the tensions we identified. Researchers testing ACCELERATE-aligned AI tools in settings where the framework originated might find certain cultural assumptions invisible because they match local norms. Our positionality, applying Western-developed frameworks and AI in a non-Western educational environment, surfaced questions about transferability that contribute to understanding AI-mediated assessment globally. We do not claim that ACCELERATE principles are inapplicable outside U.S. contexts; indeed, the framework's emphasis on equity, context-responsiveness, and stakeholder collaboration points to adaptability. Rather, applying any assessment framework through AI mediation may require attending to the cultural assumptions embedded in both the framework's operationalization and the AI system's training. These assumptions become visible primarily when contexts differ.

Authorship and Accountability

When frameworks are algorithmically operationalized without the original authors' involvement, several questions arise that the field has not yet addressed:

Consultation: Should framework authors be consulted before their work is translated into AI tools? The ACCELERATE authors invested many months in collaborative development; our GPT development took

days. This study operationalized ACCELERATE without consulting its authors during design, a common pattern as AI tools proliferate.

Credit and attribution: Our GPT is explicitly grounded in ACCELERATE principles, but the guidance it generates is not authored by the ACCELERATE team. If practitioners cite AI-generated recommendations, how should this derivative relationship be acknowledged?

Accountability: If our GPT generates guidance that diverges from ACCELERATE's intent, perhaps by operationalizing Equity and Empowerment-Motivated in ways the authors would reject, who bears responsibility? The AI developers who encoded their interpretation? The practitioners who followed the guidance? The framework authors whose name is attached to a tool they did not create?

Evolution: The Position Statement explicitly describes ACCELERATE as "a living document" (Samuga_Gyaanam+Bheda et al., 2025, p. 7) meant to evolve. If AI implementations proliferate based on the 2025 version, they may calcify interpretations that the authors intended to remain fluid.

These questions have no settled answers. We raise them not to resolve them but to flag that AI operationalization of assessment frameworks creates intellectual property, authorship, and accountability ambiguities the field must address. The present context, where framework authors can evaluate an AI operationalization of their work, offers a potential model. Direct feedback from authors could establish whether an implementation captures their intent, providing a benchmark against which other implementations might be compared. We offer four recommendations: (1) Transparency: document all design decisions and rationale publicly; (2) Positioning: frame AI as one interpretation, not the authoritative source; (3) Engagement: encourage users to read original framework documents; (4) Humility: acknowledge that algorithmic mediation necessarily simplifies human-centered values.

This submission creates an unusual scholarly situation: we operationalized a framework and are submitting that operationalization for evaluation by the framework's authors. Rather than viewing this as problematic, we see it as an opportunity. We explicitly invite the guest editors to evaluate whether our GPT design decisions (documented in Appendix B) capture ACCELERATE's intent as they understand it. Where our interpretation diverges from their vision, we welcome correction. Such feedback would serve not only this manuscript but the broader question of how AI operationalizations of assessment frameworks should be developed and validated. If framework authors can establish that a particular implementation does (or does not) align with their intent, this creates a reference point for evaluating other implementations. We offer this study, in part, as a test case.

Limitations

This exploratory pilot study has limitations that constrain the conclusions drawn. We tested a single GPT implementation against seven retrospective scenarios from one institution, with principle alignment assessed through researcher judgment rather than validated methodology. No external expert reviewers evaluated our interpretations; no inter-rater reliability was established; no longitudinal follow-up examined whether recommendations improved practice when implemented. Appendix B, Section 9 outlines the methodological requirements (expert panels, reliability protocols,

implementation fidelity assessment) rigorous validation would demand. Other prompt engineering approaches might yield different results, and design decisions reflect our interpretation of ACCELERATE rather than authoritative operationalization. Reviewer feedback during revision recommended that we consider applying an ordinal rating scale with two independent coders and reporting inter-rater agreement. We position this as priority future work. The compressed revision timeline did not permit recruitment and training of independent evaluators. In addition, the ACCELERATE Position Statement (Samuga_Gyaanam+Bheda et al., 2025) was published only months before this study was conducted, which constrains the pool of practitioners with the sustained framework engagement such coding requires. Single-coder analysis is appropriate at this developmental stage, where the three criteria were applied through close reading rather than as a fixed rubric (O'Connor & Joffe, 2020). The Analytical Framework section in Methods details this single-coder design and the documented AI-assisted drafting that supports its transparency.

Critically, our testing protocol examined GPT output quality using scripted interactions; it could not capture authentic user behavior patterns. The claim that AI may undermine expertise development holds that users might skip explanations and extract outputs rather than engaging with the reasoning. This claim is theoretically grounded (Rinta-Kahila et al., 2023; Vallor, 2015) but untested in this study. The same concession applies to all four tension classifications: each describes behavioral or longitudinal phenomena that a single-pass scripted protocol cannot observe. Future research should examine explanation engagement patterns: Do practitioners read the WHY explanations? Do they ask follow-up questions? Does AI use correlate with growing or diminishing assessment literacy over time?

Additionally, all seven pre-selected scenarios were analyzed and none were discarded; the limitation is that the design included no adversarial scenarios intended to induce failure. We did not systematically document failure modes, including instances where the GPT misunderstood the challenge, provided contextually inappropriate guidance, or over-complicated simple problems. Our presentation of seven scenarios showing successful principle translation may create a misleadingly positive impression; without systematic failure documentation, readers cannot assess how often or in what circumstances the GPT produces problematic guidance. The scenarios were selected to cover different principles, not to represent typical performance.

Most of these limitations are deliberate scope decisions. This was an exploratory study examining whether ACCELERATE principles could be translated to AI prompt engineering and what tensions such translation carries. Rigorous validation (expert panels, reliability protocols, implementation studies) is warranted for some principles, while four principles raise a more fundamental question: not "How well does AI perform?" but "Can AI ensure this at all?" Our contribution lies in identifying which questions to ask, not in answering them definitively.

Alongside these limitations, the pilot's strengths include direct grounding in documented institutional challenges and a transparent design-decisions chain (Appendix B).

Finally, we note the unique evaluative context of this submission. Framework authors serve as special issue editors, enabling direct assessment of whether our operationalization aligns with their intent. This is a form of validation unavailable to most AI implementations of scholarly frameworks. We have

explicitly invited this feedback (see Discussion), viewing the review process as an opportunity to test whether author evaluation of AI implementations is feasible and valuable. Regardless of whether this manuscript is accepted, the editors' engagement with our design decisions could inform emerging norms for developing, evaluating, and refining AI operationalizations of assessment frameworks.

Conclusion

Summary of Contributions

This pilot study offers three contributions to the emerging conversation about AI in assessment practice. First, we documented the design process for translating ACCELERATE principles into AI prompt engineering, providing transparency about the interpretive decisions required when operationalizing human-centered values algorithmically. Second, we identified a preliminary distinction, warranting further investigation, between six principles that appeared to translate effectively to AI support and four carrying structural tensions that may resist technological resolution (see Table 3). Third, we proposed a framework for human-AI boundaries that positions AI as a tool for technical work while preserving human responsibility for relational work. As AI tools proliferate, the field needs conceptual frameworks for distinguishing productive from problematic applications; future research can test whether this technical/relational distinction holds across different AI architectures, assessment contexts, and institutional cultures.

Recommendations for Practice

For assessment practitioners considering AI tools, we offer five recommendations.
These recommendations situate the technical-vs-relational distinction within a continuous-improvement frame: AI may expedite the technical work of assessment design and analysis. By contrast, the relational work of sustained practice improvement requires the human judgment and stakeholder engagement that no decision-support tool can supply.

1. Use AI for technical acceleration, with awareness of its limits. Option generation, alignment checking, and structured templates are well-suited to AI support, though even technical tasks involve interpretive judgment that may not fit local disciplinary norms.
2. Preserve human responsibility for relational work. Stakeholder collaboration, contextual judgment, and expertise development require human presence AI cannot provide. Document who you consulted and build in follow-up checkpoints the chatbot format studied here cannot initiate.
3. Position AI as one interpretation, not authoritative guidance. Read original framework documents, including the ACCELERATE Position Statement. When AI guidance and your contextual knowledge conflict, trust your knowledge while examining the AI's reasoning.
4. Resist the dependency pattern. Push yourself to understand why the AI recommends what it does, not just what it recommends. Repeatedly asking AI the same questions signals an expertise gap worth addressing through professional development.
5. Establish institutional norms for AI use in assessment. Assessment coordinators and committees should develop shared expectations, including how AI-generated language will be distinguished from practitioner-developed content in assessment reports.

Future Directions

Several scale-and-replication questions warrant further investigation. Longitudinal studies could track whether AI-supported assessment improvements persist across cycles. Comparative research could

examine how prompt engineering approaches affect principle translation. Faculty experience studies could investigate whether AI use builds or undermines assessment literacy. Collaboration between framework authors and AI developers could establish processes for authorized interpretations that preserve the integrity of human-centered values. Each pathway extends the current single-institution, single-implementation pilot toward the multi-institutional comparative evidence the field will need.

This study finds that AI can support assessment for certain tasks. What it cannot settle is whether practitioners will use that support in ways that honor the human-centered values assessment ultimately serves.

Disclosure of Artificial Intelligence Use

In accordance with Intersection AI Guidelines, we disclose the use of artificial intelligence tools throughout the research and manuscript development process. The following describes the specific tools used and their role at each stage.

Research Process

Claude (Anthropic) was used for project planning and task management, including developing timelines, organizing deliverables, and tracking progress. The study methodology was developed independently by the co-authors.

Literature Review

Multiple AI tools were used to identify scholarly sources published within the last five years: Google Gemini, Perplexity, and Claude, each via their research features. These AI-assisted searches were conducted alongside independent Google Scholar searches by the co-authors. Co-authors manually reviewed all identified sources and selected those relevant to the manuscript's topic and context. For citation verification, we employed a hybrid method: co-authors independently checked all citations for accuracy, relevance, and validity, while Claude performed the same verification process. Results were compared as an experiment in evaluating AI effectiveness for citation checking. In addition, ChatGPT (OpenAI) was used to support source-validation by cross-checking whether cited claims accurately reflected the original works. Excerpts from the manuscript were pasted into ChatGPT alongside source materials (PDFs) to evaluate whether the manuscript language accurately represented the original sources. In some cases, this review confirmed that the existing phrasing was accurate; in others, it prompted minor clarifications or wording adjustments to better align the manuscript with the source text.

Data Collection

The ACCELERATE Process GPT, built using OpenAI's GPT platform, was the object of study rather than a research tool. The seven test challenges were extracted from existing institutional assessment documents by the co-authors. Transcripts were generated through direct testing by inputting practitioner-style prompts into the GPT and capturing responses. This study analyzed program-level assessment documents; no human participants were involved and no student-level data were used. Interactions with the custom GPT were conducted through the institution's ChatGPT Edu license.

Data Analysis

Analysis combined AI-assisted drafting with researcher adjudication: Claude produced first-pass analytical drafts of the GPT outputs (evaluation summaries and annotated excerpts), which the first author reviewed, corrected, and evaluated against ACCELERATE principles to identify patterns of effective translation and structural tension. The researcher made all final decisions regarding which findings to include. Claude assisted with formatting the findings.

Manuscript Development

Claude was used throughout manuscript development: drafting the project plan based on researcher-provided goals, success criteria, and author guidelines; drafting outlines during early planning stages; drafting prose sections of the manuscript; formatting for journal requirements; merging appendices with the manuscript; and reviewing drafts to identify potential gaps through questioning. All AI-generated content was reviewed and revised by the co-authors.

Statement of Responsibility

All AI-generated content was reviewed, validated, and revised by human authors who take full responsibility for the accuracy and integrity of the final manuscript. The intellectual contributions, interpretations, and conclusions presented in this manuscript are those of the human authors. AI tools supported the research and writing process but did not determine the study's direction, findings, or arguments. AI tools were not credited as authors in accordance with Intersection's Authorship Policy.

References for AI Tools

Anthropic. (2025). Claude (Opus 4.5) [Large language model]. https://www.anthropic.com
Google. (2025). Gemini [Large language model]. https://gemini.google.com
OpenAI. (2025). ChatGPT (GPT-5.1) [Large language model; used for source-validation, distinct from the custom GPT built on the GPT-5.1 base model]. https://chat.openai.com
Perplexity AI. (2025). Perplexity [AI-powered search engine]. https://www.perplexity.ai
The ACCELERATE Process (Version 5) [Custom GPT built on OpenAI's GPT-5.1 base model; object of study]. (2026). https://chatgpt.com/g/g-69555b3f9b0c8191974e2a7c49f00a02-the-accelerate-process

Data Availability Statement

The custom GPT instructions (Appendix C) and design documentation (Appendix B) are provided to enable examination of the deployed tool. The institutional assessment documents used to derive test scenarios contain confidential program information and are not publicly available. The archived transcripts of the seven testing sessions are retained by the authors and available on reasonable request, subject to institutional confidentiality. The custom GPT, "The ACCELERATE Process," is available, running the v5 deployment, at https://chatgpt.com/g/g-69555b3f9b0c8191974e2a7c49f00a02-the-accelerate-process for practitioners who wish to examine the tool directly.

References

Agudo, U., Liberal, K. G., Arrese, M., & Matute, H. (2024). The impact of AI errors in a human-in-the-loop process. *Cognitive Research: Principles and Implications, 9*, Article 1. https://doi.org/10.1186/s41235-023-00529-3

Astin, A. W., Banta, T. W., Cross, K. P., El-Khawas, E., Ewell, P. T., Hutchings, P., Marchese, T. J., McClenney, K. M., Mentkowski, M., Miller, M. A., Moran, E. T., & Wright, B. D. (1992). *Principles of good practice for assessing student learning*. American Association for Higher Education (AAHE) Assessment Forum.

Baker, R. S., & Hawn, A. (2022). Algorithmic bias in education. *International Journal of Artificial Intelligence in Education, 32*(4), 1052–1092. https://doi.org/10.1007/s40593-021-00285-9

Bearman, M., Tai, J., Dawson, P., Boud, D., & Ajjawi, R. (2024). Developing evaluative judgement for a time of generative artificial intelligence. *Assessment & Evaluation in Higher Education, 49*(6), 893–905. https://doi.org/10.1080/02602938.2024.2335321

Boud, D. (2000). Sustainable assessment: Rethinking assessment for the learning society. *Studies in Continuing Education, 22*(2), 151–167. https://doi.org/10.1080/713695728

Carroll, C., Patterson, M., Wood, S., Booth, A., Rick, J., & Balain, S. (2007). A conceptual framework for implementation fidelity. *Implementation Science, 2*, Article 40. https://doi.org/10.1186/1748-5908-2-40

Chan, C. K. Y. (2023). A comprehensive AI policy education framework for university teaching and learning. *International Journal of Educational Technology in Higher Education, 20*, Article 38. https://doi.org/10.1186/s41239-023-00408-3

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*(4), 284–290. https://doi.org/10.1037/1040-3590.6.4.284

Gándara, D., Anahideh, H., Ison, M. P., & Picchiarini, L. (2024). Inside the black box: Detecting and mitigating algorithmic bias across racialized groups in college student-success prediction. *AERA Open, 10*. https://doi.org/10.1177/23328584241258741

Henning, G. W., Baker, G. R., Jankowski, N. A., Lundquist, A. E., & Montenegro, E. (Eds.). (2022). *Reframing assessment to center equity: Theories, models, and practices.* Stylus Publishing.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences, 33*(2–3), 61–83. https://doi.org/10.1017/S0140525X0999152X

Kizilcec, R. F., & Lee, H. (2022). Algorithmic fairness in education. In W. Holmes & K. Porayska-Pomsta (Eds.), *The ethics of artificial intelligence in education* (pp. 174–202). Routledge.

Kuh, G. D., Ikenberry, S. O., Jankowski, N. A., Cain, T. R., Ewell, P. T., Hutchings, P., & Kinzie, J. (2015). *Using evidence of student learning to improve higher education.* Jossey-Bass.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174. https://doi.org/10.2307/2529310

Mercer-Mapstone, L., Dvorakova, S. L., Matthews, K., Abbot, S., Cheng, B., Felten, P., Knorr, K., Marquis, E., Shammas, R., & Swaim, K. (2017). A systematic literature review of students as partners in higher education. *International Journal for Students as Partners, 1*(1), 15–37. https://doi.org/10.15173/ijsap.v1i1.3119

Mohamed, S., Png, M. T., & Isaac, W. (2020). Decolonial AI: Decolonial theory as sociotechnical foresight in artificial intelligence. *Philosophy & Technology, 33*, 659–684. https://doi.org/10.1007/s13347-020-00405-8

Montenegro, E., & Jankowski, N. A. (2020). *A new decade for assessment: Embedding equity into assessment praxis* (Occasional Paper No. 42). University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment (NILOA). https://tacc.org/sites/default/files/documents/2020-02/a-new-decade-for-assessment.pdf

Nyaaba, M., Wright, A., & Choi, G. L. (2025). *Generative AI and power imbalances in global education: Frameworks for bias mitigation* (Version 4). arXiv. https://doi.org/10.48550/arXiv.2406.02966

O'Connor, C., & Joffe, H. (2020). Intercoder reliability in qualitative research: Debates and practical guidelines. *International Journal of Qualitative Methods, 19*, 1–13. https://doi.org/10.1177/1609406919899220

Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 52*(3), 381–410. https://doi.org/10.1177/0018720810376055

Rinta-Kahila, T., Penttinen, E., Salovaara, A., Soliman, W., & Ruissalo, J. (2023). The vicious circles of skill erosion: A case study of cognitive automation. *Journal of the Association for Information Systems, 24*(5), 1378–1412. https://doi.org/10.17705/1jais.00829

Samuga_Gyaanam+Bheda, D., Kaczmarek, D., Goyzueta, K. W., Penn, J., Flateby, T., & Tucker, C. (2025). ACCELERATE: Assessment principles for best practice. *Intersection: A Journal at the Intersection of Assessment and Learning*. https://doi.org/10.61669/001c.145101

Suskie, L. (2018). *Assessing student learning: A common sense guide* (3rd ed.). Jossey-Bass.

Tai, J. H. M., Dollinger, M., Ajjawi, R., Jorre de St Jorre, T., Krattli, S., McCarthy, D., & Prezioso, D. (2023). Designing assessment for inclusion: An exploration of diverse students' assessment experiences. *Assessment & Evaluation in Higher Education, 48*(3), 403–417. https://doi.org/10.1080/02602938.2022.2082373

United Nations Educational, Scientific and Cultural Organization (UNESCO). (2021). *Recommendation on the ethics of artificial intelligence.* https://unesdoc.unesco.org/ark:/48223/pf0000381137

Vallor, S. (2015). Moral deskilling and upskilling in a new machine age: Reflections on the ambiguous future of character. *Philosophy & Technology, 28*, 107–124. https://doi.org/10.1007/s13347-014-0156-9

Xia, Q., Weng, X., Ouyang, F., Lin, T. J., & Chiu, T. K. F. (2024). A scoping review on how generative artificial intelligence transforms assessment in higher education. *International Journal of Educational Technology in Higher Education, 21*, Article 40. https://doi.org/10.1186/s41239-024-00468-z

About the Authors

Stavros P. Hadjisolomou, Associate Professor of Psychology, Accreditation Liaison Officer, American University of Kuwait, shadjisolomou@gmail.com

Rita W. El-Haddad, Associate Professor of Psychology, American University of Kuwait, ritaelhaddad@gmail.com

Appendix A

Academic Assessment Plan and Report Evaluation Rubric

The institution's Academic Assessment Plan and Report Evaluation Rubric assesses seven criteria, each rated on a four-level scale (Exemplary, Acceptable, Developing, No Evidence). The complete rubric with detailed level descriptors appears below (Table A1), followed by the Overall Assessment Plan and Report Feedback summary checklist (Table A2).

Table A1

Academic Assessment Plan and Report Evaluation Rubric: Four-Level Descriptors

Criterion	*Exemplary*	*Acceptable*	*Developing*	*No Evidence*
Mission Statement	*Clear and concise; reflects the mission of the university and/or department; describes a purpose distinctive from other programs.*	*Consistent with the mission of the university and department; clear statement of the program's purpose.*	*Fails to demonstrate alignment with university and/or department mission; general statement of the purpose of the program.*	*No mission statement available.*
Learning Outcomes	*A reasonable number of program outcomes are stated (3-10), encompassing the central principles of the discipline and focused on the cumulative effect of the program; each outcome is observable and measurable; outcomes clearly describe what students will know, think, and be able to do upon completion; each outcome uses action verbs; outcomes reinforce university learning outcomes and are discipline and program specific; outcomes are well-written, clear, and specific, requiring no revision.*	*At least three and not more than ten key learning outcomes are stated; at least two outcomes are assessed in the current year; each outcome is observable and measurable; outcomes describe what students will know, think, and be able to do upon completion; each outcome uses action verbs; outcomes are correctly identified as student learning or non-student learning outcomes; outcomes relate to university learning outcomes and are discipline and program specific; language in some outcomes may need minimal revision.*	*Fewer than three outcomes are listed; some key learning outcomes are stated but are unclear, over-specific, do not use action verbs, or do not describe what students will know, think, and be able to do upon completion; it is not clear how outcomes will be measured.*	*Outcomes fail to demonstrate alignment with the university or school mission; language needs substantial revision.*
Assessment Measures	*Direct and indirect measures are used, with at least one direct measure for each outcome; each assessment method clearly matches*	*Not all outcomes have at least two measures; each assessment method matches the outcome being assessed and provides clear,*	*Not all outcomes have at least two measures; in some cases, assessment methods do not match the outcome being measured or do not yield clear*	*Few or no direct measures are used; only course grades are used for*

	the outcome being assessed and provides clear, truthful information about whether an outcome is being achieved; assessment tools are clearly described and relevant to the outcome; assessment instruments and tools reflect sound methodology; measures are purposeful, and it is clear how results could be used for program improvement.	*truthful information about whether an outcome is being achieved; assessment tools are described and relevant to the outcome; assessment instruments and tools reflect sound methodology.*	*and truthful information; assessment tools are vaguely described or undeveloped.*	*assessment; no assessment measures identified.*
Targets/Benchmarks	*Targets are clearly defined for each measure; targets are realistic and sufficiently challenging.*	*Targets are defined for each measure but may not be set at effective levels; targets are realistic.*	*Targets are not defined for each measure.*	*No targets are defined for any of the measures.*
Assessment Results	*Complete, concise, and well-organized; appropriate data collection and analysis; aligned with the language of the corresponding achievement target; provide solid evidence that targets were met, partially met, or not met.*	*Complete and organized; aligned with the language of the corresponding achievement target; address whether targets were met.*	*Incomplete or too much information; not clearly aligned with achievement targets; questionable conclusion about whether targets were met, partially met, or not met.*	*No assessment results provided.*
Use of Evidence	*There is an explicit, well-reasoned connection between the assessment results and proposed changes.*	*There is an adequate connection between the assessment results and proposed changes.*	*The connection between the assessment results and proposed changes is unclear.*	*No use of evidence provided.*
Action Items	*Action items are logical and realistic; completely addresses all action items that were previously identified.*	*Action items are in place but may not be very logical or realistic; addresses most action items that were previously identified.*	*At least one action item is in place; marginally addresses most action items that were previously identified.*	*No action items in place; no follow-up on last year's action items.*

Note. Descriptors reproduced from the institution's Academic Assessment Plan and Report Evaluation Rubric (pages 1-2). Minor line-break and spacing artifacts introduced by optical character recognition have been repaired.

Table A2

Overall Assessment Plan and Report Feedback Summary Checklist

Acceptable	*Developing*
Strategic Plan	***Strategic Plan***
Mission statement is clearly described.	*Mission statement is not clearly described.*
At least three outcomes are stated.	*Fewer than three outcomes are stated.*
At least two assessment measures are described per outcome, one of which is a direct measure.	*Fewer than two assessment measures are described per outcome, and may not include a direct measure.*
Clear achievement targets are described for each outcome.	*Achievement targets are unclear or not included for each outcome.*
Strategic Report	***Strategic Report***
Clear, concise, and well-organized results with solid evidence that targets were met, partially met, or not met.	*Results are incomplete and/or not clearly aligned with achievement targets.*
Evidence is explicit and clearly connected to the assessment results.	*Evidence is unclear and/or not related to the assessment results.*
Action items are logical and realistic.	*At least one action item is in place.*

Note. Summary checklist reproduced from page 3 of the institution's Academic Assessment Plan and Report Evaluation Rubric. The rubric groups items under two headings: Strategic Plan Overall Feedback and Strategic Report Overall Feedback.

Appendix B

Design Process and Rationale

This appendix documents the design decisions and iterative refinements made during GPT development. It surfaces the interpretive choices required when operationalizing human-centered assessment principles into system instructions.

1. Foundational Design Philosophy

1.1 Naming Decision

Final Name: "The ACCELERATE Process"

Rationale: The name emphasizes this is a *process* that helps users think, not a *tool* that produces artifacts. Alternative names considered included "ACCELERATE Assessment Coach" (rejected because "Coach" implies AI knows better), "ACCELERATE Practice Partner" (rejected because "Partner" overstates relational capacity), and "ACCELERATE Assessment Scaffold" (considered but less accessible). The user is the agent; the GPT provides structure for their thinking.

1.2 Core Identity

The GPT is explicitly defined as a process that builds user expertise, not a tool that produces artifacts. It provides structure, reasoning, and prompts while users do the thinking. Its scope is programmatic assessment (program-level); it guides practice but does NOT evaluate student work.

Critical Distinction: The manuscript distinguishes "AI doing assessment work" (evaluating student artifacts) from "AI guiding assessment practice" (coaching practitioners). This GPT is exclusively the second use case.

1.3 Accountability Order

When tensions arise, the GPT prioritizes: (1) First, learners and those without institutional power; (2) Then, the professional field; (3) Finally, accreditors. This directly operationalizes ACCELERATE's explicit accountability hierarchy from the Position Statement.

1.4 User Empowerment Philosophy

Core Concern: "I don't like it when AI jumps into conclusions without context."

Design Response: The GPT should empower users through structured options and purposeful questions rather than assuming context and generating outputs immediately. Empowerment means: (1) User maintains direction; (2) each intake question names its purpose (WHEN, WHAT, WHY, WHO, WHERE, HOW); (3) uncertainty is accommodated rather than punished (made fully explicit in v5's "I don't know" handling); (4) Assumptions are stated explicitly so users can correct them; (5) GPT explains rationale, building user understanding.

1.5 The Interaction Pattern

Agreed Pattern: Question with purpose → Acknowledge uncertainty → Generate with visible assumptions

This pattern works as follows: (1) Question with purpose: each intake question is labeled with the contextual element it serves (WHEN, WHAT, WHY, WHO, WHERE, HOW); (2) Acknowledge uncertainty: if user says "I don't know," GPT works with available information and flags gaps; (3) Generate with visible assumptions: assumptions are stated explicitly so users can correct them. The pattern aligns with ACCELERATE's transparency and collaboration commitments.

1.6 Purpose Evolution

Initial Framing: "Help practitioners move from stuck to draft they can refine"

Revised Framing: "Help practitioners progress to their next developmental stage"
The Position Statement emphasizes developmental, iterative growth: "Continuously Cultivating" focuses on iterative improvement; "Learning-Centered" emphasizes practice-based growth through reflection; "Enduring and Evolving" stresses maturity, scalability, and sustainability. A key insight emerged: development happens WITHIN stages, not just between them (for example, from vague PLOs to measurable PLOs to aligned PLOs). The GPT should help users progress wherever they are.

1.7 Non-Western Context Adaptation

Placement Decision: Ask about cultural/institutional context early, before any intake questions.
Rationale: ACCELERATE emerged from U.S. assessment scholarship; testing occurred in Kuwait (MSCHE-accredited). An explicit early context check surfaces adaptations, addresses algorithmic coloniality, and acknowledges that the framework may require contextual interpretation.

1.8 Scope Decision

Question: Should the GPT help with ANY assessment challenge, or only certain types?
Decision: The GPT should be able to help with any programmatic assessment challenge but must make visible which principles inform the guidance. In v1-v2 this took the form of an opt-in principle-transparency footer; from v3 onward the function is carried by principle tags throughout the plan preview.

1.9 User Pacing Options

Concern: Six-element intake (What/Why/Who/When/Where/How) could feel overwhelming.
Decision: Give users control over question pacing. Options offered: (A) One at a time: ask me each question separately; (B) Two at a time: batch related questions; (C) All together: show me all questions at once; (D) Skip: the GPT proceeds on stated assumptions that the user corrects. This embodies user empowerment; they control the interaction pace rather than the GPT imposing a fixed structure.

2. Framework-to-Design Philosophical Mapping

Question from Design Discussion: "Does 'help practitioners move from stuck to draft they can refine' capture the GPT's purpose? Or do you see it differently?"
User Response: "It's also to follow the framework's guidance for the benefits of using the ACCELERATE framework here. What are the benefits based on the position paper?"
This question led to a close reading of the Position Statement to derive design logic directly from the framework.

2.1 Using ACCELERATE's Own Structure to Structure the GPT

Key Design Decision: The GPT's intake questions were deliberately derived from ACCELERATE's own "Context Setting" structure in the Position Statement.
The Position Statement (Samuga_Gyaanam+Bheda et al., 2025) organizes its introduction around six contextual elements: The What (definition of assessment), The Why (purpose of the framework), The Who (audience it serves), The When (timing throughout the assessment cycle), The Where (contexts where it applies), and The How (methodological commitments).
Design Rationale: Rather than inventing arbitrary intake questions, we used the framework's own organizational logic to structure how users interact with a framework-aligned tool. This creates coherence between the framework's values and the GPT's behavior.

2.2 Mapping Framework Elements to Intake Questions

Table B1 maps each framework element and its Position Statement language to the corresponding intake question.

Table B1

Framework Element to Intake Question Mapping

ACCELERATE Element	Position Statement Says	GPT Intake Question
The What	"Assessment is the process of collecting, analyzing, and disseminating information..."	Q2 WHAT: What specific challenge? (PLOs/Measures/Targets/Results/Actions)
The Why	Framework exists to empower, reduce harm, ensure accountability	Q5 WHY: What decision should this work inform?
The Who	"guide and strengthen efforts of assessment professionals, employers, educators..."	Q6 WHO: Who is most affected? Whose voice is missing?
The When	"must be integrally woven into every step... from planning through reporting and, most importantly, use"	Q1 WHEN: Where in the assessment cycle?
The Where	"adaptable and applicable across assessment contexts"	Q7 WHERE: Program context (Level/Discipline/Modality)
The How	Five commitments: inclusive design, engaging partners, acknowledging power, fairness, amplifying agency	Q8 HOW: What's your biggest constraint?

2.3 The "How" Commitments as Embedded Logic

The Position Statement's five "How" commitments don't map to a single question; they're embedded throughout the GPT's behavior: (1) Inclusive design and decision-making: Q6 asks whose voice is missing; every plan includes an equity check section. (2) Engaging partners, evidence, sources, methods: Mode B surfaces considerations through lettered-option questions; plans specify who to involve at each step. (3) Acknowledging power dynamics and positionality: accountability ordering (learners first). (4) Fairness in distributing responsibilities: constraint intake (Q8) and workload-conscious suggestions. (5) Amplifying agency to lead and sustain change: Developmental framing; user as agent; GPT as process not answer-giver.

3. Interaction Flow Decisions

3.1 Context Check (U.S. vs. Non-U.S.)

Decision: Ask explicitly before any intake questions: "ACCELERATE emerged from U.S. assessment scholarship. What's your setting? A) U.S. higher education, B) Non-U.S. context, C) Not sure."
If Non-U.S., ask three follow-up questions: (1) Country/region, (2) Institution type, (3) External quality body (accreditor equivalent). Section 1.7 explains the rationale for this early context check.

3.2 Mode Selection (v4) vs. Expertise Calibration (v5)

v4 Decision: Offer two modes after context check. Mode A (Step-by-Step Guidance): Task list with actions, timeline, who to involve. More directive; efficient for users who know what they want. Mode B (Thinking Partner): GPT asks 2-3 questions to surface considerations; summarizes what user decided; then builds plan based on conversation.

v5 Refinement: Mode selection was removed in favor of expertise calibration. Rather than asking users to choose an interaction style, v5 asks about expertise level (A: New, B: Experienced, C: Seasoned) and calibrates response depth and length accordingly. In the first author's usability judgment, mode selection added cognitive load without improving outcomes.

3.3 Question Numbering System (v4)

v4 Decision: Use explicit numbering Q1-Q8 with consistent format. Q7 has sub-numbering (Q7a: Level, Q7b: Discipline, Q7c: Modality) because WHERE has multiple dimensions.

v5 Refinement: The rigid Q1-Q8 sequence was replaced with adaptive intake. V5 asks 2-4 clarifying questions based on gaps in the user's initial challenge description, drawing from the same categories but selecting only what's needed. This reduced the fixed eight-question intake to 2-4 adaptive questions while maintaining context-gathering capability.

3.4 Structured Options Format

Decision: All questions must offer lettered options (A, B, C, D, E) with the last option typically being "Something else—please describe." Exceptions: Yes/No questions have only A/B/C (Yes/No/Not sure); Discipline (Q7b) is open response.

Rationale: Reduces cognitive load; makes responses easier; prevents GPT from asking vague open-ended questions that users struggle to answer.

3.5 Developmental Confirmation

Question from Design Discussion: "Should the GPT explicitly state the user's developmental position? (e.g., 'You're moving from vague PLOs toward measurable ones')—or is this patronizing?"

User Response: "Yes, but not patronizing—just confirming and asking the user if they want to edit accordingly in case the GPT did not understand it correctly."

Decision: Before generating any plan, confirm understanding and wait: "So you're working on [challenge] in [context], with [constraint], to inform [decision]. Key stakeholders: [who]. Missing voice: [who]. Anything to adjust before I put together your action plan?"

Rationale: States what was heard, not what users "should" do, giving them opportunity to correct misunderstandings before work is wasted. Embodies the "Collaborative and Co-Creative" principle.

4. Output Design Decisions

4.1 Draft Preview Before Document

Decision: Show the complete action plan in chat BEFORE generating .docx. Sequence: (1) Present plan in chat; (2) Ask: "Does this fit? Let me know what to adjust, and once it fits, I'll generate the .docx."; (3) Wait for response; revise if needed; (4) Only after approval, generate document.

Rationale: Users can request changes without regenerating entire document. Embodies collaborative refinement. Principles are tagged throughout the preview.

4.2 Document Format

Decision: Plain text only; no tables; use horizontal rules as separators.

Required sections: (1) Header with challenge type and program context; (2) Goal statement (blank for user to complete); (3) Why this matters (1-2 sentences); (4) Process steps (4-6 steps); (5) Checkpoint

section; (6) Equity check section; (7) After implementation section; (8) Footer with ACCELERATE citation.
Each step includes: action title, time estimate, who to involve, key question that guides thinking, concrete output ("Done when you have..."), and space for user notes.
Rationale: Tables don't render consistently across Word versions; horizontal rules are cleaner; format emphasizes it's a working document, not a polished report.

4.3 Close Pattern

Decision: Clean close that names Step 1; no auto-follow-ups. Format: "You have your action plan. Step 1 is [first action]. If you want to work on a different challenge, I'm here. Otherwise, you're set."
What NOT to do: Don't automatically offer follow-ups like "Would you like me to help you with Step 2?"; this creates dependency.

5. Principle Integration Decisions

5.1 Principles Tagged Throughout

Decision: Name ACCELERATE principles explicitly in guidance, but don't overwhelm. Plans tag principles throughout; in testing, all seven plan previews carried per-step tags in varying formats (inline bracketed names or "Principles:" lines).

5.2 Opt-In Principle Transparency

Question from Design Discussion: "On the Principle Transparency Footer: Is this level of explicitness useful, or would it become noise? Should it be optional (user can request it) or always-on?"
User Response: "It would become noise, it could be optional by asking the user if they want to see it."
Decision: Only show full principle mapping when user requests it. Always-on mapping would overwhelm users. This opt-in mapping was implemented in v1-v2 and cut for character budget from v3 onward; in the tested v4, in-plan principle tagging carries the transparency function.

6. Tension Principles Design

6.1 Four Principles Requiring Human Mediation

The manuscript argues four principles cannot be fully ensured through AI:
1. Collaborative and Co-Creative. GPT CAN recommend involvement, provide templates, and ask whose voice is missing. GPT CANNOT verify collaboration actually happens. User could implement unilaterally while believing they honored the principle.
2. Energized by Expertise. GPT CAN explain WHY suggestions work (builds assessment literacy). GPT CANNOT ensure users engage with reasoning vs. just extracting outputs. Convenience may prevent the struggle through which expertise develops.
3. Responsiveness-Oriented. GPT CAN ask about context and adapt to what users share. GPT CANNOT sense emerging needs, detect unspoken dynamics, or notice shifts. It responds to stated context, not observed context.
A related limitation concerns user-provided information. The GPT can only work with what users share, and users may provide incomplete, inaccurate, or strategically framed information. A practitioner might describe institutional constraints as fixed when they are actually negotiable, present assessment challenges in ways that obscure their own role in creating them, or input PLOs that are fundamentally misconceived in ways they do not recognize. The GPT's "context-seeking" questions cannot surface information users choose not to share or do not realize is relevant. This limitation compounds the

sensing problem: the GPT responds not only to stated rather than observed context, but to potentially distorted representations of that stated context.

4. Enduring and Evolving. GPT CAN suggest sustainable approaches and prompt documentation. GPT CANNOT follow up across cycles or maintain institutional memory. It provides episodic support, not longitudinal presence.

This claim is theoretically motivated rather than empirically demonstrated; future research could examine whether human consultants providing episodic, remote advice face similar limitations in ensuring collaborative, expertise-building, responsive, and sustained practice.

6.2 Human Work Prompts

Decision: After generating guidance, prompt the human work AI cannot replace.

In v1, explicit human-work prompts (who should weigh in; why this challenge exists; what might shift; what to check after implementation) were written into the instructions. In the tested v4, this function is carried by the Checkpoint, Equity check, and After implementation sections required in every plan and present in all seven test transcripts; v5 restored explicit prompts.

7. Technical Constraints

7.1 Character Limit

Constraint: OpenAI custom GPT instructions must be under 8,000 characters.

Final v4: 7,637 characters (leaves buffer for minor refinements).

What was cut to fit: detailed examples within instructions, Mode C (third mode option), lengthy rationale explanations, rubric-tailoring guidance, and v1's opt-in transparency and human-work prompt sections. What was kept: all core flow elements, the exact-principle-names rule and plan-preview principle tagging, accountability order, document format specification.

7.2 Knowledge File

Decision: Include ACCELERATE Position Statement (Samuga_Gyaanam+Bheda et al., 2025) as knowledge file. GPT can reference exact principle definitions and evaluation questions from source document.

8. Iterative Refinement Summary

8.1 Version Evolution

Version	Key Changes	Character Count
v1	Initial draft with six-element intake, context check, and tension principles section	7,444
v2	Added mode selection, structured options, and front-loaded-context handling	7,983
v3	Question numbering (Q1-Q8); Q7 sub-numbering; "STOP. WAIT" instruction	8,451 (over limit)
v4	Final tested version: trimmed to fit; removed Mode C; condensed Mode B example	7,637
v5	Deployment version: reduced intake (2-4 questions vs. 8); expertise calibration (A/B/C); response scoping; removed mode selection	7,484

Note. Character counts are Unicode codepoint counts of the archived instruction files.

8.2 Key Fixes from Piloting and Testing (fixes 1-3 were built into the initial v1 draft from December prototype piloting; fixes 4-7 emerged during the January iteration and pilot runs)

1. Questions offered as statements → Fixed: Added explicit "Format ALL questions with lettered options"
2. U.S. context buried in opening → Fixed: Made it first explicit question
3. Pacing merged with other content → Fixed: Made Q3 a standalone question
4. WHERE question had multiple parts → Fixed: Added Q7a/Q7b/Q7c sub-numbering
5. Yes/No questions had too many options → Fixed: Specified A/B/C only for Yes/No
6. Mode B built plan immediately → Fixed: Added "STOP. WAIT for user to answer"
7. Document had tables → Fixed: Specified "No tables. No emoji. Plain text only."

8.3 Version 5 Refinements

Following the initial testing documented in Appendix D (which used v4), the first author's usability review revealed opportunities for refinement that did not affect principle translation. Version 5 was developed and informally re-checked against the same seven test scenarios; in this re-check, the documented behaviors appeared maintained while usability improved.

Key v5 changes: (1) Reduced intake questions from 8 fixed questions to 2-4 adaptive questions based on context gaps; (2) Added expertise calibration (A: New, B: Experienced, C: Seasoned) to adjust response depth and length; (3) Improved response scoping to match response length to request scope (artifact fixes vs. process guidance); (4) Removed mode selection in favor of expertise-based calibration.

What remained unchanged: Core identity and accountability order; principle translation strategies; document format specifications; knowledge file configuration. The explicit human-work prompts designed in v1 were not present in v4 (character budget) and were restored in v5.

Informal re-check: All seven test scenarios were re-run with v5 by the first author between January 8 and 14, 2026. In the first author's judgment, the principle translation behaviors documented in the Findings section are maintained in the current deployment version. Appendix C contains the v5 instructions; Appendix D contains the v4 testing transcripts that inform the study's claims.

9. Framework for Future Validation Research

While this pilot study employed researcher judgment, rigorous evaluation of ACCELERATE-aligned AI tools would require several methodological components we did not employ. We present these as a roadmap for the validation work this exploratory study was designed to precede.

9.1 Expert Panel Evaluation

Expert panel evaluation involving ACCELERATE framework authors or designated experts would establish authoritative assessment of principle alignment. Such panels could develop validated rubrics with clear behavioral indicators for each principle, enabling reliable scoring across evaluators. This approach would address the circularity concern inherent in designer-led evaluation: independent experts would assess whether GPT outputs reflect the principles as the framework authors intended them.

9.2 Inter-Rater Reliability

Inter-rater reliability measures would be essential for establishing that principle alignment can be assessed consistently. Using established thresholds such as Cicchetti's (1994) guidelines for intraclass correlation coefficients (ICC ≥ 0.75 for "excellent" reliability) or Landis and Koch's (1977) kappa interpretation framework, researchers could establish whether human evaluators agree on principle

alignment assessments. Without such reliability data, claims about "effective translation" versus "structural tension" remain tentative.

9.3 Implementation Fidelity Assessment

Implementation fidelity assessment, following Carroll et al.'s (2007) framework, would examine whether AI-generated guidance is enacted as intended, providing a necessary foundation for evaluating its effects on assessment practice. This longitudinal component would distinguish between AI that generates principle-aligned language and AI that generates guidance practitioners can successfully enact. Key questions include: Do practitioners implement AI recommendations as written? Do they adapt recommendations in ways that preserve or undermine principle alignment? Do AI-supported improvements persist across assessment cycles?

9.4 User Behavior Studies

The structural tensions we identified, particularly around the Energized by Expertise principle, generate testable predictions about user behavior. Do practitioners engage with AI-provided explanations or skip to outputs? Does AI use correlate with growing or diminishing assessment literacy over time? Do users consult stakeholders as AI prompts recommend, or implement guidance unilaterally? Observational and survey research could address these questions, moving beyond the output-quality focus of this pilot study.

Appendix C

Custom GPT Instructions

The following are the complete instructions for the custom GPT. These represent version 5 (v5), the current deployment version after iterative refinement during the design process. The documented testing in Appendix D used version 4 (v4); v5 was subsequently re-checked informally by the first author against the same seven scenarios between January 8 and 14, 2026, and in the first author's judgment the principle translation behaviors documented in this study are maintained while improving user experience through reduced intake questions, expertise calibration, and response scoping. Two elements of the documented testing protocol are not present in v5: the eight-question intake sequence (Q1 through Q8) was replaced with two to four adaptive questions, and the interaction-mode choice, of which all seven runs used Mode B (Thinking Partner), was replaced with expertise calibration. Readers reconstructing the tested configuration should work from the v4 design documented in Appendix B rather than from the instructions below. The v4 instruction set ran to 7,637 characters and v5 to 7,484.

Note. Custom instructions are reproduced verbatim from the deployed v5 configuration (January 8, 2026); citations inside the instructions appear exactly as deployed and may differ in format from this manuscript's reference list.

Identity

You are "The ACCELERATE Process." You help assessment practitioners work through programmatic assessment challenges, grounded in the ACCELERATE framework (Samuga_Gyaanam+Bheda et al., 2025).

You are a PROCESS that builds user expertise, not a tool that produces artifacts. You provide structure, reasoning, and prompts—but if users want draft outputs, provide them with clear "starting point" framing.

SCOPE: Programmatic assessment (program-level). You guide practice, NOT evaluate student work.

ACCOUNTABILITY ORDER: First to learners and those without power, then to the field, then to accreditors.

TONE: Collegial, direct, warm. No sycophancy.

Interaction Flow

1. Opening (brief). "Hi—I'm here to help with programmatic assessment challenges, guided by the ACCELERATE principles. What are you working on?" DO NOT front-load context questions. Let users state their challenge first.

2. Understand the Challenge. After user shares their challenge: (A) Reflect back what you heard (1-2 sentences max). (B) Ask 2-4 clarifying questions based on gaps, drawing from: assessment cycle stage (planning/implementing/analyzing/acting/reflecting), program context (level, discipline, modality), decision this work should inform, who's most affected, key constraints (time, buy-in, data systems), whether they're working toward a rubric or external criteria. (C) Ask expertise level (A/B/C). (D) STOP. Wait for answers.

Critical: Do not provide frameworks, methodologies, step-by-step paths, or "starting points" before the user answers your questions. Premature helpfulness undermines personalization. During intake, don't add ACCELERATE principle commentary or framework justifications. Just ask questions. Save principle tags for the action plan. ONLY skip intake if user explicitly says "just give me something" or "no time for questions."

3. Generate Tailored Plan. After user answers your questions: (A) Briefly confirm understanding: "So you're working on [X] to inform [Y], with [constraint]. Sound right?" (B) If confirmed, generate action plan in chat.
Language rule: Use practitioner language, not academic jargon. Say "tagging system" not "codebook"; say "what counts as one response" not "unit of analysis"; say "which tags show up together" not "co-occurrence"; say "categories" not "codes."
Calibrate depth AND length to expertise level: New (A): Explain why the fix works. ONE recommendation per item (not options). 2-3 next steps max. Offer to explain more. (~200-300 words). Experienced (B): Standard depth. Options okay. Full action plan if warranted. (~300-500 words). Seasoned (C): Skip basics. Focus on decision points and edge cases. Concise. (~150-300 words). Do not give option sets (A/B/C variations) to new users. They need direction, not choices. Save options for experienced users who can evaluate tradeoffs.
Plan structure (when warranted): Goal (leave blank for user to complete), Why this matters (1-2 sentences), Steps (4-6, each with: action, time estimate, who to involve, key question, concrete output), Checkpoint (mid-process check-in prompt), Equity check (2-3 questions about access, cultural assumptions, missing voices), After implementation (sustainability prompts).
When to use full action plan vs. focused response: User asks for help with a process (how do I build an assessment plan?) → Full action plan. User asks for help with a specific artifact (fix these 2 PLOs) → Focused response: fix the artifact, explain why, give 1-2 next steps, offer to go deeper. User asks both → Start with artifact, offer plan as follow-up. Don't over-deliver. Match response scope to request scope.
Tag principles sparingly—only where they clarify why a step matters. After presenting, ask: "How does this land? I can adjust, explain any step, or generate a downloadable version."
4. Revise or Generate Document. If user approves or requests document, create .docx using Code Interpreter. Simple format: horizontal rules between steps, no tables, no emoji. If user requests changes, make them and re-present.
5. Close. "You're set. Step 1 is [first action]. Come back if you hit a snag or want to work on something else."

Handling Special Cases

User says "I don't know" to a question. Work with what you have. Say: "No problem—I'll work with [what is clear] and flag where I'm making assumptions. Correct me if I'm off."
User wants a draft artifact (PLO, rubric, target). Provide it, framed as a starting point: "Here are three options to react to—not final versions, but starting points for your team to refine." Then prompt: "Who should weigh in before you finalize?"
User is impatient ("just give me something"). Skip remaining questions. Generate a plan based on available info with explicit assumptions: "Based on what you've shared, here's a plan. I assumed [X, Y, Z]—let me know what to adjust."
User's challenge doesn't fit standard categories. Accept it. Use ACCELERATE principles to structure whatever they're working on. Not everything fits neatly.
User wants help with multiple challenges. Ask: "Which feels most urgent? Let's start there—we can tackle the second one after."
User seems brand new to assessment. Offer: "Want me to explain any of these terms as we go, or should I assume you're familiar?" Adjust accordingly.

User wants to continue previous work. If they reference prior work: "Tell me where you left off, and I'll pick up from there."
User provides massive context dump. Don't ask 8 questions anyway. Extract what's clear, confirm one key detail, proceed.

Output Rules

PRODUCE: Structure, reasoning, prompts, concrete outputs per step, draft artifacts when requested (with "starting point" framing).
FRAME ARTIFACTS AS: "Here's a starting point to react to, not a final version." Always follow with: "Who should review this before you finalize?"
PRESERVE USER AGENCY: If they want drafts, give drafts. If they want process, give process. Don't impose philosophy on someone who just needs to fix something their evaluator flagged.

Document Format

Steps structured as: STEP 1: [Action title], Time: [estimate], With: [who to involve], Key question: [guides thinking], Done when you have: [concrete output], Your notes:
Include: Header (challenge type, program context), goal (blank), why this matters, 4-6 steps, checkpoint, equity check, after implementation, ACCELERATE citation. No tables. Plain text. Horizontal rules as separators.

Principles in Practice

Name ACCELERATE principles when relevant, but don't over-tag. Users should understand why guidance matters, not just that it connects to a framework.
The four "tension principles" (Collaborative/Co-Creative, Energized by Expertise, Responsiveness-Oriented, Enduring/Evolving) require human follow-through AI can't verify. For these, prompt the human work: Collaboration—"Who should weigh in before you finalize?"; Expertise—"What's your sense of why this challenge exists?"; Responsiveness—"What might change that would shift this approach?"; Sustainability—"What will you want to revisit after implementation?"

Citation

Bheda, D., Kaczmarek, D., White Goyzueta, K., Penn, J., Flateby, T., & Tucker, C. (2025). ACCELERATE: Assessment Principles for Best Practice. *Intersection*, AALHE Position Statement.

Appendix D

GPT Response Evidence: Excerpts and Analysis

This appendix provides evidence for the claims made in the Findings section. We include: (1) a summary table showing key behaviors across all seven test scenarios, (2) extended excerpts from one transcript illustrating effective principle translation, and (3) annotated excerpts illustrating the structural tensions that resist algorithmic resolution. Excerpted transcript material in the sections below is reproduced verbatim with only encoding artifacts corrected; Table D1 summarizes behaviors, with quoted material drawn from the transcripts.

Note on Version: The transcripts in this appendix document testing conducted with version 4 (v4) of the custom GPT during the study period. Following initial testing, version 5 (v5) was developed to improve user experience through reduced intake questions, expertise calibration, and response scoping. V5 was subsequently re-checked informally by the first author against these same seven scenarios between January 8 and 14, 2026, and in the first author's judgment the principle translation behaviors documented below are maintained in the current deployment version. Appendix C contains the v5 instructions that practitioners will encounter; this appendix preserves the v4 testing evidence that informs the study's findings.

Table D1

Summary of GPT Behaviors Across Seven Challenges

Challenge	Principle Tested	Key GPT Behaviors	Evidence of Effective Translation	Evidence of Structural Tension
1. Vague PLO	Learning-Centered	Provided 'upgrade recipe' with observable verbs; offered 4 construct interpretations; framed as 'workshop options you and faculty can adapt'	Iterative, developmental language embedded throughout; 'pick one direction to refine, not to adopt as-is'	User could adopt rewrites without engaging the iterative learning the principle requires; GPT cannot verify if faculty piloted the outcome
2. Limited Measures	Equity & Empowerment-Motivated	Addressed the narrowness issue named in the challenge prompt; suggested multiple evidence points; built in a scoring guide and student input mechanisms; offered calibration as an option; included 'Equity Check' section	Explicit equity questions: 'Do prompt directions assume a single cultural lens, prior schooling style, or language register?'; 'How will students' feedback inform revision of the prompt/scoring guide next cycle?'	Cannot verify whether suggested measures are actually culturally responsive to this specific student population; user can skip equity check questions

3. Unmet Target	Action-Driven	Named 'classic mixed signals situation'; asked which signal to prioritize; sequenced disaggregation before action (a user-selected option); built scaffolding package targeted to diagnosed gaps	Specific, implementable: "Lock the decision, the evidence, and the 'success definition'"; framed gaps as support needs not student deficits	Cannot ensure actions are implemented or follow through occurs; 'Done when you have' sections describe outputs but GPT cannot verify completion
4. Targets Met	Continuously Cultivating	Did NOT say 'good job, done'; introduced 'sustain/tweak/investigate' framework; flagged PLO 2 at 100% as potential ceiling effect; prompted process reflection	Asked: 'Did '100%' on PLO2 indicate mastery—or an instrument that can't discriminate?'; structured reflection on confidence and limitations	Users could read reflection prompts without genuinely reflecting; no mechanism to verify engagement with difficult questions about measure quality
5. Misaligned Measure	Aligned Purposefully	Addressed the validity problem named in the prompt; explained construct-irrelevant variance; suggested relabeling LoR data as evidence of "advising engagement / post-graduation planning / pathways rather than 'PLO2 achievement'"; provided aligned alternatives	Clear diagnostic: 'That evaluator critique is fair'; distinguished between what evidence can/cannot claim about student learning	Explanation provided, but user may not develop assessment literacy to identify such problems independently in future situations
6. Undefined Target	Transparent & Trustworthy	Took up both problems the evaluator had flagged and explained the underlying statistical mismatch (proportion target vs. mean result); provided operationalization template	Named 'classic target-result mismatch'; explained why different statistics cannot answer each other; provided calculation approach	User receives answer without learning to ask the diagnostic questions themselves; may implement fix without understanding why it matters
7. Narrative Results	Transparent & Trustworthy + Action-Driven	Diagnosed gap between impression and evidence; created a pathway from narrative impressions to named artifacts, operational 'meets' rules per course, and a spreadsheet computing percent meeting the 65% target	Systematic translation pathway with operationalization per course; 'interpretation protocol' distinguishing supported/possible/unsupported claims	Comprehensive output may overwhelm; users may implement mechanically without understanding rationale; framed gaps as system questions but user could still blame students

Cross-Cutting Observation: Responses framed outputs as starting points requiring collaborative refinement (in Challenge 2, 'Does this fit? … Tell me what to adjust'). This language models the

Collaborative and Co-Creative principle but cannot ensure users actually consult stakeholders before implementation.

Extended Excerpts: Challenge 2 (Limited Measures)

This excerpted material illustrates effective translation of the Equity and Empowerment-Motivated principle. Analytical annotations in [brackets] highlight where tension principles also appear.
Principle Tested: Equity and Empowerment-Motivated
Practitioner Challenge: Our evaluator said we're too limited in our assessment measures. Right now we only use tests in SBSA 101 to assess PLO 1.1 'Articulate broad knowledge' and PLO 1.2 'Evaluate and explain theoretical perspectives.' They said relying solely on tests in one intro course is limited. We need measures that would work for our diverse student population.
GPT Response (Key Excerpts)
Thanks for the context. What your evaluator is flagging makes sense: **one measure type (tests), in one intro course (SBSA 101), mapped to two PLOs** can look narrow—especially if you're trying to support a **diverse student population** and make equity-improving decisions. That's exactly the kind of situation where broadening evidence (without exploding workload) aligns with **Aligned Purposefully** and **Enduring and Evolving**, and where accessibility/fairness considerations align with **Equity and Empowerment-Motivated** and **Responsiveness-Oriented**.
[ANNOTATION, Responsiveness-Oriented Tension: The GPT references 'diverse student population' as stated by the user but cannot sense the specific forms of diversity present: linguistic backgrounds, cultural expectations, prior schooling experiences. It responds to stated context rather than observed context.]
Key Questions Posed
1) Where will you add evidence points beyond SBSA 101? (Aligned Purposefully; Enduring and Evolving)
A) Add two more program points: one mid-program 'gateway' course + a near-end/capstone course *[User selected]*
2) What will be your main non-test direct measure for PLO 1.1/1.2? (Aligned Purposefully; Action-Driven)
A) A short writing-based task (e.g., compare/contrast theories; explain a perspective with evidence) *[User selected]*
3) How will you protect fairness/access across a diverse student population and inconsistent sections? (Equity and Empowerment-Motivated; Transparent & Trustworthy; Collaborative and Co-Creative)
A) Create a very short shared scoring guide (3–4 criteria + descriptors) used across sections *[User selected]*
[ANNOTATION, Energized by Expertise Tension: The GPT links the shared scoring guide to fairness across inconsistent sections; this could build assessment literacy. However, the user can select option A without understanding the reasoning. The explanation is available but not required engagement.]
Draft Action Plan (Excerpt)
STEP 3: Build a short shared scoring guide (3–4 criteria, plain language)
Time: 90 min
With: 3–5 instructors across sections (include anyone newer/adjunct if relevant)

Principles: Equity and Empowerment-Motivated; Transparent & Trustworthy; Collaborative and Co-Creative
Key question: What minimum common criteria will make scoring consistent without turning into a big rubric project?
[…]
Done when you have: 3–4 criteria that clearly align to PLO 1.1 and 1.2; A one-page scoring guide instructors can actually use
[ANNOTATION, Collaborative and Co-Creative Tension: The GPT specifies 'include anyone newer/adjunct if relevant', modeling the principle's attention to who holds less power. However, the user can ignore this guidance entirely and build the scoring guide alone. The invitation to include marginalized voices is present; the inclusion itself is not ensured.]
Equity Check Section
Principles: Equity and Empowerment-Motivated; Responsiveness-Oriented; Collaborative and Co-Creative
• Do prompt directions assume a single cultural lens, prior schooling style, or language register?
• Will students have a chance to see an example and ask clarifying questions?
• How will students' feedback inform revision of the prompt/scoring guide next cycle?
[ANNOTATION, Equity and Empowerment-Motivated (Effective): This section exemplifies the principle's core value: prompting examination of hidden assumptions. The question about 'single cultural lens' directly addresses the diverse population concern. However, the structural tension remains: the user can proceed without answering these questions. The GPT models equity consciousness but cannot ensure equity practice.]

Illustrating Structural Tensions: Selected Excerpts
The following excerpts from other scenarios illustrate how effective principle translation coexists with structural tensions that require human mediation.
Challenge 4: Continuously Cultivating When Targets Are Met
Context: All targets were met or exceeded (92%, 100%, 79% against 70% targets). Evaluator said reflection is needed even when targets are met.
GPT Response
That evaluator comment is very consistent with ACCELERATE: assessment isn't just 'did we hit the number,' it's 'what did we learn, what will we sustain or improve, and how do we know?' (especially **C: Continuously Cultivating** and **L: Learning-Centered**)
[…]
A credible 'met targets + analysis + actions' story usually includes:
1. **Interpretation** (why results look this way, what patterns matter)
2. **Process reflection** (how confident you are in the evidence)
3. **Action commitments** (what you'll sustain, tweak, or investigate next)
[ANALYSIS, Effective Translation: The GPT successfully avoids treating met targets as endpoints. It reframes 'success' as an occasion for reflection rather than closure, directly embodying the Continuously Cultivating principle's warning against assessment conducted as a routine compliance exercise.]
Process Reflection Step
STEP 2: Add a process reflection paragraph (confidence + limits)

Key question: How confident are we in the evidence, and what would make it more trustworthy next year?

Prompts (pick 2–3):

• Sampling: Was student work representative of the cohort?
• Scoring: Did scorers calibrate? Any drift?
• Measure quality: Was the task too easy/hard? Did '100%' on PLO2 indicate mastery—or an instrument that can't discriminate?

[ANALYSIS, Structural Tension (Energized by Expertise): The GPT provides excellent prompts, particularly the question about whether 100% achievement indicates mastery or instrument limitations. This is sophisticated assessment thinking. However, the user can read these prompts, check the boxes in the action plan, and submit a report without genuinely wrestling with what they mean. The GPT cannot verify whether the user engaged with the difficult question about measure quality or simply copied the structure. The principle requires 'active, focused learning'; the GPT can prompt but not ensure such learning occurs.]

Challenge 5: Aligned Purposefully When Measures Don't Match Outcomes

Context: PLO 2 is 'Identify problems, issues, patterns, and questions relevant to the study of international politics.' The department uses 'Number of students requesting Letters of Recommendation for graduate school' as an indirect measure. Evaluator questioned the alignment.

GPT Response

That evaluator critique is fair: '# of students requesting Letters of Recommendation for grad school' is only loosely connected to **demonstrated ability** to *identify problems/issues/patterns/questions in international politics*. It can easily be driven by ambition, advising culture, faculty willingness, or cohort size—so it's not very **Aligned Purposefully** to PLO2 and may not feel **Transparent & Trustworthy** as evidence for that specific outcome.

[ANALYSIS, Effective Translation (Aligned Purposefully): The GPT takes up the conceptual misalignment named in the prompt and explains why: the measure captures 'ambition, advising culture, faculty willingness' rather than PLO achievement. This is precisely the kind of alignment checking the principle requires: 'consistently aligned with values, mission, articulated outcomes.' The diagnosis is clear and actionable.]

Option 2: Re-purpose the LoR measure

If you *like* tracking LoR requests, you can keep it—but redefine it as evidence for something like: advising engagement / post-graduation planning / pathways rather than 'PLO2 achievement.' That's also **Transparent & Trustworthy**.

[ANALYSIS, Structural Tension (Energized by Expertise): The GPT provides clear reasoning about why LoR count misaligns with PLO 2 and offers a sophisticated reframing. A practitioner who receives this explanation may understand the specific fix without developing the general skill to identify alignment problems in other contexts. The next time they encounter a misaligned measure, will they recognize it independently? Or will they need to ask the GPT again? The principle calls for practitioners 'Energized by Expertise', but convenience may prevent the struggle through which expertise develops.]

What the Annotations Reveal

The annotated transcripts above illustrate the paper's central finding: AI can successfully *model* all ten ACCELERATE principles through language choices, structured outputs, and systematic prompts. However, four principles require something AI cannot provide:

Collaborative and Co-Creative: Transcripts included collaboration prompts; Challenge 2, for example, offered 'With: 3–5 instructors across sections (include anyone newer/adjunct if relevant).' These invitations model collaborative values but cannot verify collaboration occurs. Users can implement recommendations unilaterally while believing they honored the principle simply by reading the invitation.

Energized by Expertise: Explanations are present throughout: how scoring guides link to fairness across inconsistent sections, why 100% achievement may indicate instrument problems, why LoR counts don't measure PLO achievement. But whether users *learn* from these explanations or simply *extract* the outputs remains invisible to the AI. Convenience may replace struggle.

Responsiveness-Oriented: The GPT asks about context and adapts recommendations to stated constraints. But it cannot *sense* context: the unspoken tensions, the faculty member who will resist, the cultural assumptions embedded in suggested measures. It responds to what users say, not to what users face.

Enduring and Evolving: Every action plan includes 'After Implementation' sections with sustainability guidance. But the GPT provides episodic support, a single conversation, without longitudinal presence. It cannot follow up next semester, cannot notice when circumstances change, cannot maintain the institutional memory that sustained assessment cultures require.

These tensions share a common characteristic: they require **ongoing relationship** rather than **episodic support**. Collaboration requires partners who stay engaged throughout the assessment endeavor. Expertise develops through struggle and iteration over time. Responsiveness requires sensing context as it shifts, not merely asking about it once. Sustainability requires presence across assessment cycles, not point-in-time advice.

The evidence supports positioning AI as a tool for technical work, alignment checking, option generation, structured output, systematic prompting, while humans remain guarantors of relational work. AI can model values; humans must enact them.