Timing Evidence Checks for LLM-Assisted Academic Work: A Protocol and Synthetic Scenario Analysis

Lauren Mercer, Maya Kalyani Reed, Hoang Justin Tran, Alexander J. Garnier
Simon Fraser University
University of Missouri, Columbia
Université Grenoble Alpes
Concordia University
lauren.mercer@sfu.ca, maya.kalyani.reed@umsystem.edu, hoang.justin.tran@etu.univ-grenoble-alpes.fr, alexander.j.garnier@concc

Abstract—Engineering educators and students increasingly use large language models for writing, code, design reasoning, and scholarly communication, creating a need for safeguards whose completion can be evidenced at the right point in an academic workflow. This paper develops a conceptual protocol that places authorization and data classification before use, process controls during work, and verification and disclosure before release. An author-coded validity assessment maps seven existing safeguards to nine proposed risk domains, and a reproducible synthetic model compares fixed, stage-aligned, and risk-adaptive presentation under explicit assumptions about hazards, efficacy, adherence, effort, and routing error. The model is an assumption map, not an effectiveness study. In the primary scenario, presenting all seven safeguards at a fixed point produced modeled net loss of 2.608, compared with 2.219 for stage-aligned, applicability-filtered presentation. A synthetic 2×2 decomposition yielded 2.537 for fixed-filtered and 2.240 for staged-all presentation, indicating that most primary separation follows the assumed overload response rather than filtering. With overload disabled, fixed and staged versions using the same selection rule tied exactly under paired simulation; when the probability of returning to each later stage was 0.30, staged-filtered presentation was slightly worse than fixed-all presentation. Adaptive routing broke even at sensitivity 0.64 in the primary setting and 0.72 when routing errors concentrated in high-stakes tasks. The contribution is a testable design and reporting framework for LLM-use safeguards, not evidence that staging improves real academic work.

Index Terms—AI education, assessment, scholarly integrity

I. Introduction

Engineering education increasingly asks learners, instructors, and researchers to make responsible decisions about large-language-model assistance in technical writing, coding, design documentation, analysis, and publication. Policies can state what is permitted, what must be disclosed, and what data may be used, but they do not by themselves determine when a user can truthfully attest that a safeguard has been completed. Some checks are meaningful before any interaction, others only while work is unfolding, and others only after outputs exist. This study addresses that timing problem by linking seven safeguards to observable evidence, mapping them to nine risk domains, and modeling alternative presentation strategies under explicit assumptions. The result is a protocol and scenario-analysis framework that supports implementation planning while preserving the limit that the model is not empirical evidence of effectiveness.

Universities increasingly regulate large-language-model (LLM) use through rules on permitted assistance, authorship, privacy, verification, and disclosure. Such rules define institutional expectations, but users still need concrete actions that fit the sequence of academic work. Higher-education guidance identifies factual error, privacy, authorship, integrity, and insufficient transparency as recurring risks [1–4]. Liang recently proposed seven questions that translate these concerns into task scope, planning, decomposition, traceable computation, output verification, data isolation, and disclosure [5]. Whether all seven safeguards should be presented before the first interaction, rather than when evidence of completion can exist, remains unresolved.

Moving from guidance to a dependable protocol requires evidence about both content and use. A policy-to-action mapping can establish that an item has a rationale, and a worked example can establish that an item is executable in at least one setting. Neither establishes that the complete set is comprehensive, that different users interpret each item similarly, that a recorded check corresponds to a performed safeguard, or that repeated use reduces error or misconduct. Measurement guidance treats relevance and comprehensiveness as only part of content validity [6–8], and validity arguments require warrants linking observations to intended decisions [9, 10]. The central design question is therefore what evidence a completed safeguard should represent.

A second issue is temporal. “Can the task be completed on a computer?” and “Is sensitive data isolated?” can be answered before an interaction. “Has the output been verified?” cannot be truthfully completed until an output exists. Complete disclosure likewise belongs before submission or release, not necessarily before the first prompt. Agentic workflows make this distinction sharper because planning, tool use, state management, and action are interleaved [11–14]. Treating every safeguard as a pre-

use question risks either meaningless prospective assent or retrospective box-ticking.

A third issue is implementation effort. Checklist research in high-reliability settings shows that availability alone is insufficient and that implementation fidelity matters [15, 16]. Nominal adherence may not indicate whether intended work occurred [17, 18]. The literature reviewed for this study did not provide an empirically calibrated adherence function for academic LLM safeguards. The simulation therefore treats adherence, performed-check effort, delayed-stage return, and switching cost as explicit scenario assumptions rather than measured behavior.

This paper provides four contributions:

- an author-coded validity assessment linking each safeguard to observable evidence and to a nine-domain risk map;
- an evidence-timed protocol that preserves the seven safeguards while separating before-use, during-work, and before-release controls;
- a reproducible scenario model that separates timing from applicability filtering and exposes the consequences of alternative adherence, effort, and delayed-return assumptions; and
- routing break-even analyses and reporting recommendations that turn protocol design choices into testable implementation hypotheses without treating synthetic task volume as empirical evidence.

II. Related work

A. Generative AI, integrity, and higher-education policy

Liang frames the seven-question checklist as a bridge between generative-AI policy and technical constraints [5]. The present work retains the seven safeguard constructs and the policy-to-action framing, then adds explicit completion artifacts, an author-coded nine-domain map, and a formal comparison of presentation strategies. Reviews and policy frameworks agree that LLMs can support explanation, drafting, feedback, coding, and other educational tasks while also creating risks of inaccuracy, opacity, privacy loss, dependency, and unauthorized substitution [1–4]. These risks cannot be managed solely by prohibiting tools because educational value and risk both depend on the task, learning objective, assessment design, and evidence retained by the user.

Additional strands sharpen the practical stakes. AI literacy research argues that users need competencies for recognizing what AI systems can and cannot do, not merely permission to use them [19]. Empirical work on learning processes warns that generative AI can also encourage metacognitive laziness when learners outsource monitoring and reflection [20]. The AI Assessment Scale links permitted tool use to assessment design [21], while the IDEA framework develops metacognitive planning, monitoring, and regulation with exploratory pilot evidence [22]. Academic-integrity controls therefore need to be interpreted alongside learning-purpose controls rather than as substitutes for them.

The closest educational precedent is ALIGN, a nine-principle framework for responsible generative-AI use in thesis work based on 28 interviews; it includes learning integrity, verification, disclosure, and activities before, during, and at the end of thesis work [23]. ALIGN establishes that staged support and a learning-integrity domain are not new in themselves. The narrower contribution here is to translate Liang's seven safeguards into explicit completion artifacts and to expose, through a reproducible scenario model, which behavioral quantities would have to be measured before choosing among fixed, filtered, staged, or adaptive delivery.

Verification is especially important. Hallucination is not a single removable defect but a family of unsupported-generation problems [24, 25]. Long-context models may underuse relevant information depending on its position [26]. Automated AI-text detection is also an unstable substitute for process evidence: evaluations document errors, and paraphrasing can sharply degrade detection [27–29]. A responsible protocol should therefore prioritize source inspection, executable checks, and provenance rather than treating detector scores as proof.

Process evidence also matters because language models can reproduce or amplify existing citation patterns rather than neutrally surveying a field [30]. A verification gate should therefore require source inspection and search logging, not merely plausible-looking references.

B. Checklists and implementation fidelity

The surgical checklist literature offers both inspiration and warning. A landmark multi-site study associated a checklist intervention with improved outcomes [15], and later synthesis identified effectiveness alongside compliance and implementation factors [16]. But a checklist is not a causal ingredient independent of workflow, authority, team norms, or fidelity. Counting checked boxes may measure documentation more readily than implementation [17].

Interaction timing is already a checklist-design variable. Kulp et al. linked digital checklist logs to observed work and documented incomplete or retrospective checking, showing why recorded completion and performed action must be separated [18]. Just-in-time adaptive-intervention theory likewise distinguishes decision points, tailoring variables, decision rules, and intervention options [31]. In scholarly LLM use, PaperTrail operationalizes claim-level links to evidence and reports that making evidence more visible can increase caution without necessarily changing reliance [32]. Together, these studies motivate an evidence-linked protocol while cautioning that better-timed or more visible prompts do not by themselves establish changed behavior.

C. Validity and appropriate reliance

Instrument-development guidance recommends defining the construct, generating and reviewing items, evaluating response processes, and collecting reliability and criterion evidence before strong score interpretations [6–8, 33]. Argument-based approaches further require a chain from observations to intended decisions [9, 10]. For the present checklist, at least three uses are possible: a memory aid, a traceable process record, or a measure of responsible behavior. Each requires different evidence.

Human-factors research similarly distinguishes trust from appropriate reliance and documents automation bias [34, 35]. Human–AI combinations are not uniformly superior to either humans or AI alone [36]. The aim of a checklist should therefore be calibrated allocation of responsibility, not generic confidence in either party.

D. Task-specific evidence artifacts

The artifact needed to support a completed safeguard depends on the computational object being produced. Label-efficient segmentation and novel-object captioning make labels, uncertainty, and evaluation targets salient [37, 38]; model reprogramming and in-context semi-supervised learning make adaptation data and visible task context material [39, 40]; and generative tissue reconstruction requires provenance plus domain-specific validation [41]. Agentic language-model workflows and claim-evidence interfaces add tool traces and source links as further artifact types [13, 32]. This variation motivates recording what was checked and how, rather than treating a uniform "yes" as equivalent evidence across tasks.

III. Methods

A. Design and scope

We study the seven safeguards under eight stylized academic-use profiles through an author-coded validity assessment and a formal scenario simulation. A local analytic plan was frozen before the original simulation outputs were inspected, but it was not externally registered or time-stamped. The factorial decomposition, common-random-number correction, alternative weight runs, and structural checks reported in this revision were added after independent review and are identified as such. The complete plan, code, seeds, and generated results accompany this manuscript.

The study uses no participants or personal data. All behavioral quantities in the simulation are declared assumptions, so the numerical outputs describe model scenarios rather than estimates for students, educators, institutions, or other users.

B. Validity assessment

The assessment considered construct representation and content relevance, response process and completion observability, internal structure or reliability, relations to external outcomes, fairness and cross-context invariance, and consequences and implementation burden. Each category was coded as supported, partially supported, or not yet supported using the source report and verified methodological literature. These codes are author judgments rather than independent-rater measurements.

We separately mapped the seven questions to nine risk domains: task/tool mismatch, unauthorized substitution, workflow drift, computational error, factual or citation error, privacy or confidentiality breach, disclosure or provenance failure, learning displacement, and fairness or contestability. A domain counted as covered only if a question required an observable action capable of interrupting the risk. The last two domains were included to evaluate comprehensiveness, not to predetermine that longer is better.

TABLE I
Evidence-timed placement of the original safeguards.

Stage	Safeguard	Minimum completion evidence
Before use	Task and tool scope	Task classification and a reason the chosen tool is suitable
Before use	Human authorization and plan	Human-owned objective, prohibited substitutions, and stopping rule
Before use	Data isolation	Data classification and approved transfer or local-processing route
During work	Task decomposition	Named subtasks, dependencies, and review boundaries
During work	Traceable computation	Executable code, versioned formulas, or equivalent trace
Before release	Output verification	Source checks, tests, or numerical reconciliation
Before release	AI-use disclosure	Tool, version where available, scope, and human responsibility

C. Evidence-timed protocol

The revised deployment preserves the original seven safeguards but moves them to the stage when affirmative evidence can exist (Table I). We distinguish a safeguard from the user-interface wording used to elicit it.

This representation changes timing but does not claim a new empirical effect.

D. Scenario model

We represented eight stylized task profiles: brainstorming, language editing, literature triage, coding, quantitative analysis, formative feedback, grading or assessment, and thesis drafting. Each profile has fixed applicability flags and baseline hazard propensities for the nine domains. The values were frozen before execution and are exposed in the code.

The original analysis compared five protocols:

- none: no safeguard prompts;
- fixed7: all original questions displayed together for every task;
- staged7: the same questions displayed only when applicable and at their evidence-bearing stage;
- adaptive7: the same seven questions routed using an imperfect task classifier;
- adaptive9: the routed seven plus learning-purpose and fairness/contestability safeguards.

Relative to fixed7, staged7 changes both lifecycle placement and applicability filtering. Its contrast with fixed7 therefore evaluates a combined presentation configuration, not a timing-only intervention.

The revision adds a synthetic 2×2 decomposition:

- fixed-all: all seven safeguards in one bundle;
- fixed-filtered: only author-labeled applicable safeguards, still in one bundle;
- staged-all: all seven safeguards, split across the three stages; and
- staged-filtered: author-labeled applicable safeguards, split across stages.

This decomposition isolates the two mechanisms only inside the model. Applicability remains an oracle author label, and the timing contrast still inherits the assumed adherence function.

For task i and risk r, the baseline hazard indicator H_{ir} is drawn from a task-specific probability. If an applicable check is presented, performed, and effective, the residual contribution is reduced by efficacy E_{ir}:

$$L_i = \sum_r w_r H_{ir} \prod_{j \in J(r)} (1 - E_{ij} P_{ij} A_{ij}) + \lambda C_i + \gamma S_i,$$

where w_r is a consequence weight, $J(r)$ is the set of safeguards mapped to risk r, P_{ij} indicates performance, and A_{ij} indicates applicability. In the primary model, $C_i = \sum_j P_{ij}$, so λ prices performed-check effort rather than exposure to prompts; a structural check instead sets C_i to the number presented. S_i counts later stages visited and γ is a switching cost, set to zero in the primary model. Presentation and performance are separate. Adherence is assumed to be 0.90 for a bundle containing no more than three displayed safeguards and to decline linearly beyond that threshold. In adaptive protocols, an applicable safeguard is shown with routing sensitivity and an inapplicable safeguard is shown with one minus routing specificity.

The primary run used 200,000 tasks per seed and five seeds. These seeds are Monte Carlo replicates, not independent studies. Safeguard efficacy was sampled uniformly from 0.35 to 0.75; overload was 0.08 per safeguard beyond three; performed-check effort weight was 0.02; routing sensitivity was 0.90; and specificity was 0.85. The resulting 0.58 assumed adherence for a seven-item fixed bundle versus 0.90 for each stage bundle is a model input, not a behavioral estimate.

Parameter sensitivity analyses crossed three hazard multipliers, three efficacy ranges, three overload values, four effort weights, and six routing sensitivities, yielding 648 cells per protocol comparison. Post-review structural checks changed the burden basis, disabled overload, introduced later-stage return probabilities, and added switching cost. Alternative author-chosen risk-weight schemes emphasized privacy, integrity, epistemic error, or learning. A separate sweep varied routing sensitivity from 0.50 to 1.00 and included an adversarial condition in which high-stakes tasks were harder to route. None of these scenarios calibrates the model to observed users or tasks.

E. Reproducibility

The analysis uses Python, NumPy, pandas, and Matplotlib. code/run.sh regenerates all tables and figures. Primary seeds are listed in code/SEEDS.md. Protocols use aligned random streams for tasks, hazards, performance, efficacy, and routing so that equal-expectation comparisons do not differ only because protocol branches consume random numbers in a different order. Reported task quantiles describe the synthetic outcome distribution; they are not confidence intervals for protocol effects. The smoke command runs a smaller sample while preserving the full output schema.

IV. Results

A. Validity evidence

The assessment found partial support for content relevance and response process, while the remaining categories lacked direct support (Table II). Content relevance is partial because the source paper provides policy or technical rationales for each item but no independent item ratings. Response process is partial because the worked examples show possible actions, although the same operator designed, used, and documented the protocol. Comprehensiveness, stage coherence, reliability, external outcome relations, cross-context fairness, and unintended consequences remain untested.

The author-coded construct map assigns one original safeguard to each of seven proposed domains. Learning displacement and fairness or contestability lack explicit gates. Figure 1 displays the item–domain assignments rather than treating the extended nine-domain set as validated. The two unassigned rows are hypotheses for independent content review, not proof that two items must be added.

B. Primary scenario

Without safeguards, mean residual harm was 3.278 normalized units per synthetic task. Fixed7 reduced modeled residual harm to 2.527. Staged7 reduced it to 2.114, while adaptive7 and adaptive9 produced 2.230 and 1.963, respectively. With the assumed performed-check effort weight, mean net loss was 2.608 for fixed7, 2.219 for staged7, 2.329 for adaptive7, and 2.084 for adaptive9

TABLE II
Validity assessment of the source checklist.

Evidence class	Rating	Basis	Minimum next study
Content relevance	Partially supported	Seven rationales and mappings, without independent ratings	Expert content-validity panel
Content comprehensiveness	Not yet supported	No systematic stakeholder or domain elicitation	Multi-context content study
Response process	Partially supported	One self-documented software case	Think-aloud plus artifact verification
Stage coherence	Not yet supported	Some evidence cannot exist before first use	Compare evidence-timed deployment
Internal structure/reliability	Not yet supported	No score model or repeated ratings	Define intended score, then reliability study
Relations to outcomes	Not yet supported	No comparison condition	Preregistered outcome evaluation
Cross-context fairness	Not yet supported	No invariance or accessibility analysis	Differential consequence study
Consequences and burden	Not yet supported	Repeated-use burden not measured	Implementation-fidelity study

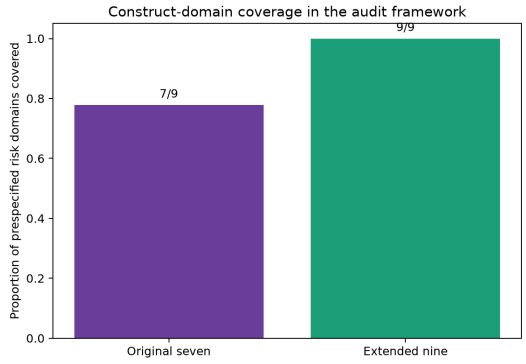


Fig. 1. Author-coded mapping of seven safeguards to nine proposed risk domains. Filled cells indicate the model's one-safeguard/one-primary-risk assignment; blank rows identify proposed content gaps.

(Table III). The ordering is generated by the model's hazard, efficacy, adherence, applicability, and effort inputs rather than observed intervention performance.

Relative to fixed7, staged7—the combined stage-aligned and applicability-filtered configuration—lowered modeled net loss by 14.9%. Across Monte Carlo seeds, fixed7 ranged from 2.606 to 2.611 and staged7 from 2.218 to 2.222; these ranges measure simulation stability, not inferential uncertainty. Figure 2 displays modeled residual harm alongside prompts shown. Net loss prices performed safeguards, not prompt exposure.

TABLE III
Primary scenario, means across five seeds.

Protocol	Residual harm	Checks presented	Checks performed	N
None	3.278	0.00	0.00	
Fixed7	2.527	7.00	4.06	
Staged7	2.114	5.87	5.29	
Adaptive7	2.230	5.46	4.91	
Adaptive9	1.963	6.88	6.03	

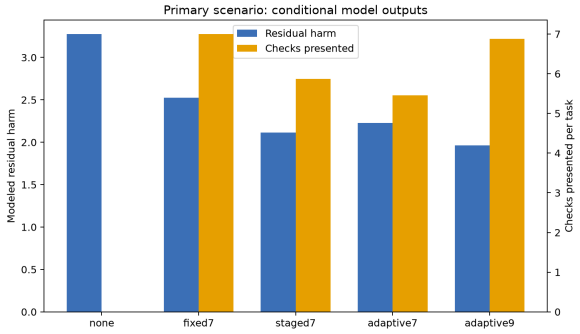


Fig. 2. Primary scenario under assumed parameters. Residual harm and prompts shown are separate model outputs; net loss uses performed-check effort.

C. Factorial decomposition and structural checks

The 2×2 decomposition shows where the primary separation enters the model (Table IV). Under the assumed overload slope, filtering a fixed bundle changes net loss from 2.608 to 2.537, while staging all safeguards changes it from 2.608 to 2.240. The combined staged-filtered value is 2.219. Thus most of the primary difference follows the assumed increase from 0.58 adherence for one seven-item bundle to 0.90 adherence for stage bundles of at most three items.

When overload was set to zero, paired runs gave identical results for fixed-all and staged-all (net loss 2.238) and for fixed-filtered and staged-filtered (2.218). The timing-only advantage therefore disappears when its assumed mechanism is removed. Structural checks also make delayed return consequential: with a 0.30 probability of returning to each later stage, staged-filtered net loss was 2.612 versus 2.607 for fixed-all. Charging shown rather than performed safeguards, adding a switching cost, and using four alternative author-chosen weight schemes changed magnitudes but did not supply empirical calibration. These follow-up checks were added after review and should not be interpreted as preregistered robustness evidence.

TABLE IV
Synthetic 2 × 2 decomposition under primary assumptions.

Timing	Selection	Residual harm	Presented	Performed	Net loss
Fixed	All	2.527	7.00	4.06	2.608
Fixed	Filtered	2.461	5.87	3.83	2.537
Staged	All	2.114	7.00	6.30	2.240
Staged	Filtered	2.114	5.87	5.29	2.219

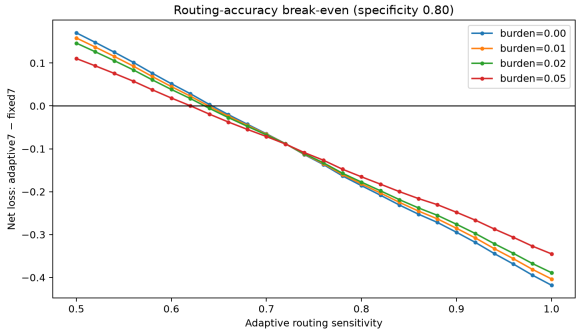


Fig. 3. Modeled net-loss difference between adaptive7 and fixed7 over routing sensitivity. Values below zero favor adaptive routing within the declared scenario.

D. Parameter sensitivity and routing break-even

With aligned random streams, staged7 had lower modeled loss in 594 of the 648 parameter cells and tied fixed7 exactly in the 54 cells with both zero overload and zero effort weight; fixed7 was lower in none. This pattern is a consequence of the modeled structure: staging cannot hurt when later-stage return is perfect and switching is free, while filtering can reduce charged effort. It is not evidence of universal dominance outside that structure.

Adaptive7 was not uniformly preferable. It was lower than fixed7 in 274 of 648 cells (42.3%) and tied in nine. The proportion lower increased from 33.3% when performed safeguards had no effort cost to 54.3% at weight 0.05. Selective presentation becomes more favorable as charged effort rises, whereas missed applicable safeguards dominate when routing sensitivity is low.

At routing specificity 0.80 and burden weight 0.02, adaptive7 first matched or improved on fixed7 at sensitivity 0.64. When high-stakes grading and thesis tasks were made harder to route, the threshold increased to 0.72. Across tested specificity and burden combinations, ordinary thresholds were 0.62–0.66 and adversarial thresholds were 0.70–0.74. Figure 3 displays the break-even pattern at specificity 0.80.

E. Evidence relative to the source claims

The source paper's strongest defensible claim is operational: each question can correspond to a concrete action in a software-development case. The present analysis preserves that contribution while narrowing two stronger interpretations. First, one case cannot show that the questions reduce errors or integrity incidents. Second, a post-output action cannot coherently be completed before the output exists.

The simulation adds conditional comparisons that the source lacks. The factorial analysis attributes most of the primary separation to the assumed overload response, and the delayed-return check identifies a model structure in which the ordering reverses. Adaptive selection offered lower modeled loss only when routing was sufficiently accurate. Adaptive9's lower value is partly definitional because the model assigns positive hazards and efficacious safeguards to the two added domains; it does not establish content validity. The practical extension is therefore to preserve the safeguards, place them where completion evidence can exist, require evidence of performance, and validate both adherence and routing before relying on either mechanism.

V. Discussion

A. A protocol, not yet an instrument

The original checklist is useful as a candidate protocol. Calling it an instrument invites a scoring interpretation: a completed list might be read as evidence that a task is safe, compliant, or educationally appropriate. That inference is not currently supported. A user can answer "yes" incorrectly, superficially, or under a different interpretation. Items may also be formative rather than reflective: the safeguards cause different actions and need not form a single psychometric scale. Future work should define the intended decision before selecting reliability or factor-analytic methods.

A traceable protocol can still be valuable before psychometric validation. The crucial change is to bind each answer to evidence. "Verified" should identify what was checked, by what method, and with what result. "Data isolated" should identify the data class and transfer route. "Disclosed" should point to a complete release record.

B. Why timing is substantive

Evidence timing is not cosmetic. Prospective verification is a plan to verify, not completed verification. Conflating the two weakens both accountability and measurement. The three-stage structure makes responsibility clearer:

- before use: is the task authorized, appropriately scoped, and safe to expose?
- during work: is the process decomposed and computationally traceable?
- before release: have claims been checked and assistance disclosed?

This arrangement also accommodates agentic systems. Tool calls and state changes can trigger checks at the moment of consequence rather than only at the first prompt. The structural checks nevertheless show that timing can fail if users do not return to later gates; the staged protocol should therefore be treated as a design hypothesis requiring workflow support.

C. When adaptation helps

The adaptive model exposes a governance trade-off. Routing can reduce unnecessary prompts, but a missed privacy, assessment, or authorship safeguard may be more costly than several irrelevant reminders. Institutions should not deploy adaptive suppression merely because a classifier appears convenient. They should measure sensitivity by task and consequence, retain conservative defaults for high-stakes work, log routing decisions, and make suppressed safeguards visible on demand. The break-even thresholds reported here are scenario outputs, not deployment criteria.

The adversarial analysis is especially important. Average routing accuracy can conceal systematic underperformance on precisely the tasks where the protocol matters most. The relevant validation target is consequence-weighted recall, not overall accuracy.

D. Learning and fairness

The original list emphasizes compliance and technical failure. Those priorities are defensible, but academic work also has a learning purpose and affects other people. A task can be policy-compliant, factually verified, and still replace the practice the assignment was designed to develop. An automated recommendation or grading workflow can be well documented and still create unreviewed disparate consequences. We therefore propose learning-purpose and fairness/contestability as domains for content validation, not as settled additions.

E. Testable hypotheses

The model identifies hypotheses whose direction depends on quantities that must be measured:

1) stage-specific prompts will increase artifact-verified completion only if any adherence gain from smaller bundles exceeds attrition from missed later stages and switching costs;
2) self-reported completion will exceed artifact-verified completion;
3) adaptive presentation will reduce prompt exposure but may increase consequential omissions when applicable-safeguard recall is low;
4) high-stakes tasks will require higher routing sensitivity or conservative always-on safeguards.

VI. Implications for research and practice

A. Design of implementation studies

The distinction between a memory aid and a measure changes the design of the next study. A trial that merely asks users whether they completed the list would test willingness to report compliance, not whether the safeguards occurred. Each gate should therefore have an evidence rule defined before data collection. Examples include a saved task classification, a marked policy boundary, a data-transfer decision, executable analysis code, a citation-verification log, and a disclosure statement. Outcome assessors should be blinded to interface condition where feasible and should judge artifacts rather than the participant's global claim of responsible use.

The comparison condition also matters. "Checklist versus no checklist" answers whether any structured prompt is better than none, but it does not test the timing claim developed here. A more informative design would randomize intact course sections or workflow episodes among fixed front-loaded, evidence-timed, and possibly adaptive interfaces. Primary outcomes could include the proportion of applicable safeguards verified by artifacts, consequential factual or computational errors, unauthorized data transfers, and disclosure completeness. Secondary outcomes could include time, perceived burden, learning performance, and false reassurance. Because tasks within a learner and learners within a course are dependent, analysis should respect clustering rather than treating prompts as independent observations.

A factorial implementation study should cross timing (front-loaded versus stage-aligned) with selection (all safeguards versus applicability-filtered safeguards). Holding wording, evidence requirements, and interface functions constant across arms would estimate the two mechanisms separately and reveal whether they interact. Exposure logs should record which safeguards were available, displayed, opened, completed, and supported by retained evidence, allowing presentation failure to be distinguished from nonperformance after presentation.

Measurement development should precede a definitive effectiveness trial. Independent experts and affected stakeholders should evaluate item relevance, comprehensibility, and comprehensiveness. Cognitive interviews can reveal whether "planning," "verification," and "sensitive" have stable meanings across disciplines. A generalizability design could then separate variation due to persons, tasks, items, and raters. Only after the intended interpretation is clear would a total score be justified; a profile of gate-specific evidence may remain more defensible.

B. Institutional deployment without premature certification

Institutions need not wait for a definitive trial before offering procedural support, but they should avoid presenting checklist completion as certification. A low-risk deployment can use the protocol formatively: remind users, link each gate to examples, retain user-controlled evidence, and state that completion does not establish compliance. High-stakes uses such as grading, disciplinary investigation, research involving confidential data, or thesis submission require specialist review and should not be delegated to the checklist.

The staged structure permits local adaptation without silently changing the construct. Institutions can change examples, approved tools, and disclosure templates while

preserving the three temporal functions. Versioning should record which policy mapping and interface were active for a given task. When rules change, earlier records should remain interpretable rather than being overwritten by a current template.

Automation should be conservative. A router may recommend checks, but it should expose why a check was selected, allow users to reveal the full set, and default to showing high-consequence safeguards when uncertain. Suppression decisions should be logged for evaluation. These requirements turn routing from an invisible convenience feature into an inspectable component of the intervention.

C. Interpretation of the scenario model

The model shows how lifecycle placement and applicability filtering can alter expected performance when the declared overload relationship holds; it does not establish that users will comply more often with staged prompts. The factorial decomposition clarifies that the timing-only contrast is driven by the assumed overload–adherence mechanism, while the delayed-return scenario shows a plausible way staging can lose. The additional learning and fairness domains likewise do not establish that the protocol should contain exactly nine questions. They identify areas where the source construct map may be incomplete. Lower modeled net loss is a transparent combination of simulated hazards, assumed safeguard efficacy, and a chosen effort valuation rather than a causal effect.

Formal models are useful here as assumption maps. They reveal which empirical quantities control a decision, prevent illustrative cases from being mistaken for effect estimates, and make disagreements testable. A stakeholder who assigns no overload cost to repeated checking, for example, can inspect the zero-overload cells rather than debating the usefulness of checklists in the abstract.

D. Minimum reporting standard for future evaluations

Future evaluations should report enough detail to distinguish the protocol from the surrounding implementation. At minimum, authors should identify the task population, institutional rules, model and interface versions, which gates were applicable, which were shown, what counted as performance, what artifacts were retained, and who assessed those artifacts. Studies should separately report presentation, self-reported completion, artifact-verified completion, and downstream outcomes. Collapsing those quantities into one adherence percentage would obscure the interpretation of protocol completion.

Intervention fidelity should also include exposure. A safeguard cannot be judged ineffective if the interface failed to present it at the relevant decision point. Conversely, presentation alone is not performance. For adaptive interfaces, a confusion matrix should be reported at the gate level, with consequence-weighted false omissions highlighted. Aggregate routing accuracy is inadequate when low-frequency thesis or assessment tasks carry disproportionate consequences.

Finally, harms should be specified in advance. Academic integrity violations, factual errors, privacy incidents, lost learning opportunities, and unfair decisions are not interchangeable endpoints. They have different base rates, detection processes, and consequence weights. A study powered for citation errors may say little about unauthorized authorship or confidential-data exposure. The protocol should therefore be evaluated as a vector of safeguards and outcomes before any composite score is defended. A further reporting implication is that evaluators should separate interface effects from policy effects. If a course changes its artificial-intelligence rules and its safeguard interface at the same time, later evidence cannot identify which component changed behavior. Artifact review should also distinguish absence from non-applicability. A learner who never transfers data should not be penalized for lacking a data-transfer record; a grading workflow, by contrast, should retain evidence that affected learners had a review or appeal route. This gate-level accounting makes the protocol useful even before outcome evidence is definitive. It lets instructors see where users need examples, lets administrators see where local policy is unclear, and lets researchers estimate which parts of a protocol deserve controlled testing. The practical value of the design is therefore not certification. It is a disciplined way to connect a broad institutional rule to a visible action, a retained artifact, and a later evaluation question.

Future interfaces should therefore treat each safeguard as a small evidence request rather than a moral declaration. The interface can ask for a task label, a data decision, a verification artifact, or a disclosure record, and it can leave narrative judgment to the instructor or reviewer. That distinction keeps the protocol lightweight while preserving the information needed for later evaluation.

VII. Limitations

This study has no human outcome data. Hazard prevalence, safeguard efficacy, adherence, consequence weights, and routing quality are scenario inputs, so numerical differences quantify the implications of assumptions rather than effects in a population. The construct map is author-generated and lacks independent stakeholder review; the proposed missing domains may require different wording, subdivision, or exclusion.

The original primary comparison changed timing and applicability filtering together. The added 2×2 decomposition separates them only under synthetic oracle applicability and the same unsupported adherence rule. The effort term in the primary model charges performed safeguards rather than documentation time, training, false reassurance, or all prompt exposure; structural checks explore but do not validate alternatives. The model also assigns each safeguard one primary risk even though safeguards may affect several risks and interact, and it omits learning

across repeated use, institutional incentives, adversarial misconduct, interface design, and unequal access. The primary 14.9% difference is therefore a property of the declared scenario, not evidence that staging is effective.

The analytic plan was locally specified rather than externally registered. The first implementation omitted planned alternative risk-weight schemes and median summaries; the present revision runs and reports them as post-review additions. The literature search was structured but not an exhaustive systematic review and provides no pooled effect estimate. The analysis concerns the checklist described in the source manuscript rather than every institutional AI policy; legal interpretation and policy currency require jurisdiction-specific expert review.

VIII. Conclusion

The seven safeguards provide a useful bridge from policy to individual action. Organizing them by the stage at which completion evidence becomes available creates a clearer protocol for authorization, process control, verification, and disclosure. In the declared model, the combined stage-aligned and applicability-filtered configuration improves the effort–fidelity trade-off only under favorable adherence and return assumptions, whereas adaptive routing helps only when applicable-check recall is high enough, especially on consequential tasks.

The next empirical program is concrete: independently review domain coverage, operationalize affirmative answers as artifacts, compare fixed and staged interfaces, and measure both self-reported and verified performance. Until those studies are completed, the protocol is best understood as a plausible and testable design whose real-world effectiveness remains unknown.

Data availability

The source material is the public preprint identified by DOI 10.20944/preprints202609.1187.v2 and cited as Liang [5]. No human-subject or personal data were used. All numerical outputs are synthetic scenario results generated from the parameters documented in the accompanying code.

Code availability

Code is available from the corresponding author and will be released at publication.

CRediT author contribution

Lauren Mercer, Maya Kalyani Reed, Hoang Justin Tran, Alexander J. Garnier: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing.

Funding

This research received no external funding.

Declaration of competing interest

The author declares no competing interests.

Declaration of generative AI use

During the preparation of this work the author used AI-assisted tools for literature search, code development and drafting. The author reviewed and edited the content and takes full responsibility for it.

References

[1] E. Kasneci, K. Sessler, S. Kuchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Gunnemann, E. Hullermeier, S. Krusche, G. Kutyniok, T. Michaeli, C. Nerdel, J. Pfeffer, O. Poquet, M. Sailer, A. Schmidt, T. Seidel, M. Stadler, J. Weller, J. Kuhn, and G. Kasneci, "Chatgpt for good? on opportunities and challenges of large language models for education," Learning and Individual Differences, vol. 103, p. 102274, Apr 2023. [Online]. Available: http://dx.doi.org/10.1016/j.lindif.2023.102274

[2] C. K. Y. Chan, "A comprehensive ai policy education framework for university teaching and learning," International Journal of Educational Technology in Higher Education, vol. 20, no. 1, July 2023. [Online]. Available: http://dx.doi.org/10.1186/s41239-023-00408-3

[3] D. R. E. Cotton, P. A. Cotton, and J. R. Shipway, "Chatting and cheating: Ensuring academic integrity in the era of chatgpt," Innovations in Education and Teaching International, vol. 61, no. 2, pp. 228–239, Mar 2023. [Online]. Available: http://dx.doi.org/10.1080/14703297.2023.2190148

[4] L. Yan, L. Sha, L. Zhao, Y. Li, R. Martinez-Maldonado, G. Chen, X. Li, Y. Jin, and D. Gasevic, "Practical and ethical challenges of large language models in education: A systematic scoping review," British Journal of Educational Technology, vol. 55, no. 1, pp. 90–112, Aug 2023. [Online]. Available: http://dx.doi.org/10.1111/bjet.13370

[5] S. Liang, "Seven questions before every prompt: A pre-use checklist for llm-assisted academic work," Oct 2026. [Online]. Available: http://dx.doi.org/10.20944/preprints202609.1187.v2

[6] D. F. Polit and C. T. Beck, "The content validity index: Are you sure you know what's being reported? critique and recommendations," Research in Nursing & Health, vol. 29, no. 5, pp. 489–497, 2006. [Online]. Available: http://dx.doi.org/10.1002/nur.20147

[7] C. A. C. Prinsen, L. B. Mokkink, L. M. Bouter, J. Alonso, D. L. Patrick, H. C. W. de Vet, and C. B. Terwee, "Cosmin guideline for systematic reviews of patient-reported outcome measures," Quality of Life Research, vol. 27, no. 5, pp. 1147–1157, Feb 2018. [Online]. Available: http://dx.doi.org/10.1007/s11136-018-1798-3

[8] L. Mokkink, S. Herbelet, P. Tuinman, and C. Terwee, "Content validity: judging the relevance, comprehensiveness, and comprehensibility of an outcome measurement instrument – a cosmin perspective," Journal of Clinical Epidemiology, vol. 185, p. 111879, Sept 2025. [Online]. Available: http://dx.doi.org/10.1016/j.jclinepi.2025.111879

[9] C. A. Chapelle, M. K. Enright, and J. Jamieson, "Does an argument-based approach to validity make a difference?" Educational Measurement: Issues and Practice, vol. 29, no. 1, pp. 3–13, Mar 2010. [Online]. Available: http://dx.doi.org/10.1111/j.1745-3992.2009.00165.x

[10] D. A. Cook, R. Brydges, S. Ginsburg, and R. Hatala, "A contemporary approach to validity arguments: a practical guide to <scp>k</scp> ane's framework," Medical Education, vol. 49, no. 6, pp. 560–575, May 2015. [Online]. Available: http://dx.doi.org/10.1111/medu.12678

[11] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao, "React: Synergizing reasoning and acting in language models," arXiv preprint arXiv:2210.03629, 2022. [Online]. Available: https://arxiv.org/abs/2210.03629

[12] Z. Xi, W. Chen, X. Guo, W. He, Y. Ding, B. Hong, M. Zhang, J. Wang, S. Jin, E. Zhou, R. Zheng, X. Fan, X. Wang, L. Xiong, Y. Zhou, W. Wang, C. Jiang, Y. Zou, X. Liu, Z. Yin, S. Dou, R. Weng, W. Cheng, Q. Zhang, W. Qin, Y. Zheng, X. Qiu, X. Huang, and T. Gui, "The rise and potential of large language model based agents: A survey," arXiv preprint arXiv:2309.07864, 2023. [Online]. Available: https://arxiv.org/abs/2309.07864

[13] L. Wang, C. Ma, X. Feng, Z. Zhang, H. Yang, J. Zhang, Z. Chen, J. Tang, X. Chen, Y. Lin, W. X. Zhao, Z. Wei, and J. Wen, "A survey on large language model based autonomous agents," Frontiers of Computer Science, vol. 18, no. 6, Mar 2024. [Online]. Available: http://dx.doi.org/10.1007/s11704-024-40231-1

[14] C. Packer, S. Wooders, K. Lin, V. Fang, S. G. Patil, I. Stoica, and J. E. Gonzalez, "Memgpt: Towards llms as operating systems," arXiv preprint arXiv:2310.08560, 2023. [Online]. Available: https://arxiv.org/abs/2310.08560

[15] A. B. Haynes, T. G. Weiser, W. R. Berry, S. R. Lipsitz, A.-H. S. Breizat, E. P. Dellinger, T. Herbosa, S. Joseph, P. L. Kibatala, M. C. M. Lapitan, A. F. Merry, K. Moorthy, R. K. Reznick, B. Taylor, and A. A. Gawande, "A surgical safety checklist to reduce morbidity and mortality in a global population," New England Journal of Medicine, vol. 360, no. 5, pp. 491–499, Jan 2009. [Online]. Available: http://dx.doi.org/10.1056/NEJMsa0810119

[16] A. Borchard, D. L. B. Schwappach, A. Barbir, and P. Bezzola, "A systematic review of the effectiveness, compliance, and critical factors for implementation of safety checklists in surgery," Annals of Surgery, vol. 256, no. 6, pp. 925–933, Dec 2012. [Online]. Available: http://dx.doi.org/10.1097/SLA.0b013e3182682f27

[17] B. M. Gillespie and J. B. Ziemba, "Lost in translation: does measuring 'adherence' to the surgical safety checklist indicate true implementation fidelity?" BMJ Quality & Safety, vol. 33, no. 4, pp. 209–211, Jan 2024. [Online]. Available: http://dx.doi.org/10.1136/bmjqs-2023-016617

[18] L. Kulp, A. Sarcevic, Y. Zheng, M. Cheng, E. Alberto, and R. Burd, "Checklist design reconsidered: Understanding checklist compliance and timing of interactions," in Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. ACM, 2020, pp. 1–13. [Online]. Available: https://doi.org/10.1145/3313831.3376853

[19] D. Long and B. Magerko, "What is ai literacy? competencies and design considerations," in Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, ser. CHI '20. ACM, Apr 2020, pp. 1–16. [Online]. Available: http://dx.doi.org/10.1145/3313831.3376727

[20] Y. Fan, L. Tang, H. Le, K. Shen, S. Tan, Y. Zhao, Y. Shen, X. Li, and D. Gasevic, "Beware of metacognitive laziness: Effects of generative artificial intelligence on learning motivation, processes, and performance," British Journal of Educational Technology, vol. 56, no. 2, pp. 489–530, Dec 2024. [Online]. Available: http://dx.doi.org/10.1111/bjet.13544

[21] M. Perkins, L. Furze, J. Roe, and J. MacVaugh, "The artificial intelligence assessment scale (aias): A framework for ethical integration of generative ai in educational assessment," Journal of University Teaching and Learning Practice, vol. 21, no. 6, 2024. [Online]. Available: https://doi.org/10.53761/q3azde36

[22] X. Wang, Z. Zheng, J. Zhang, X. Hou, and Z. Zhu, "The idea framework for metacognitively regulated genai use in higher education: Development and exploratory pilot evidence," Computers and Education: Artificial Intelligence, vol. 11, p. 100657, 2026. [Online]. Available: https://doi.org/10.1016/j.caeai.2026.100657

[23] F. Buehler, "Responsible generative ai use in thesis work: The align framework and checklist for students and supervisors," July 2026. [Online]. Available: https://doi.org/10.35542/osf.io/n8gw4_v2

[24] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, "Survey of hallucination in natural language generation," ACM Computing Surveys, vol. 55, no. 12, pp. 1–38, Mar 2023. [Online]. Available: http://dx.doi.org/10.1145/3571730

[25] P. Manakul, A. Liusie, and M. Gales, "Selfcheckgpt:

Zero-resource black-box hallucination detection for generative large language models," in Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2023, pp. 9004–9017. [Online]. Available: http://dx.doi.org/10.18653/v1/2023.emnlp-main.557

[26] N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, and P. Liang, "Lost in the middle: How language models use long contexts," Transactions of the Association for Computational Linguistics, vol. 12, pp. 157–173, 2024. [Online]. Available: http://dx.doi.org/10.1162/tacl_a_00638

[27] D. Weber-Wulff, A. Anohina-Naumeca, S. Bjelobaba, T. Foltynek, J. Guerrero-Dib, O. Popoola, P. Sigut, and L. Waddington, "Testing of detection tools for ai-generated text," International Journal for Educational Integrity, vol. 19, no. 1, Dec 2023. [Online]. Available: http://dx.doi.org/10.1007/s40979-023-00146-z

[28] R. Tang, Y.-N. Chuang, and X. Hu, "The science of detecting llm-generated text," Communications of the ACM, vol. 67, no. 4, pp. 50–59, Mar 2024. [Online]. Available: http://dx.doi.org/10.1145/3624725

[29] K. Krishna, Y. Song, M. Karpinska, J. Wieting, and M. Iyyer, "Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense," arXiv preprint arXiv:2303.13408, 2023. [Online]. Available: https://arxiv.org/abs/2303.13408

[30] A. Algaba, C. Mazijn, V. Holst, F. Tori, S. Wenmackers, and V. Ginis, "Large language models reflect human citation patterns with a heightened citation bias," in Findings of the Association for Computational Linguistics: NAACL 2025. Association for Computational Linguistics, 2025, pp. 6829–6864. [Online]. Available: http://dx.doi.org/10.18653/v1/2025.findings-naacl.381

[31] I. Nahum-Shani, S. N. Smith, B. J. Spring, L. M. Collins, K. Witkiewitz, A. Tewari, and S. A. Murphy, "Just-in-time adaptive interventions (jitais) in mobile health: Key components and design principles for ongoing health behavior support," Annals of Behavioral Medicine, vol. 52, no. 6, pp. 446–462, 2017. [Online]. Available: https://doi.org/10.1007/s12160-016-9830-8

[32] A. Martin-Boyle, C. Leckey, M. Brown, and H. Kaur, "Papertrail: A claim-evidence interface for grounding provenance in llm-based scholarly q&a," in Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems. ACM, 2026, pp. 1–25. [Online]. Available: https://doi.org/10.1145/3772318.3791101

[33] G. O. Boateng, T. B. Neilands, E. A. Frongillo, H. R. Melgar-Quinonez, and S. L. Young, "Best practices for developing and validating scales for health, social, and behavioral research: A primer," Frontiers in Public Health, vol. 6, June 2018. [Online]. Available: http://dx.doi.org/10.3389/fpubh.2018.00149

[34] J. D. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," Human Factors: The Journal of the Human Factors and Ergonomics Society, vol. 46, no. 1, pp. 50–80, Jan 2004. [Online]. Available: http://dx.doi.org/10.1518/hfes.46.1.50_30392

[35] R. Parasuraman and D. H. Manzey, "Complacency and bias in human use of automation: An attentional integration," Human Factors: The Journal of the Human Factors and Ergonomics Society, vol. 52, no. 3, pp. 381–410, June 2010. [Online]. Available: http://dx.doi.org/10.1177/0018720810376055

[36] M. Vaccaro, A. Almaatouq, and T. Malone, "When combinations of humans and ai are useful: A systematic review and meta-analysis," Nature Human Behaviour, vol. 8, no. 12, pp. 2293–2303, Oct 2024. [Online]. Available: http://dx.doi.org/10.1038/s41562-024-02024-1

[37] J. Fan, B. Gao, H. Jin, and L. Jiang, "Ucc: Uncertainty guided cross-head cotraining for semi-supervised semantic segmentation," in 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2022, pp. 9937–9946. [Online]. Available: https://doi.org/10.1109/cvpr52688.2022.00971

[38] J. Fan, Y. Liang, L. Liu, S. Huang, and L. Zhang, "Rca-noc: Relative contrastive alignment for novel object captioning," in 2023 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, 2023, pp. 15 464–15 474. [Online]. Available: https://doi.org/10.1109/iccv51070.2023.01422

[39] M.-Y. Chung, J. Fan, H. Ye, Q. Wang, W.-C. Shen, C.-M. Yu, P.-Y. Chen, and S.-Y. Kuo, "Model reprogramming demystified: A neural tangent kernel perspective," arXiv preprint arXiv:2506.00620, 2025. [Online]. Available: https://arxiv.org/abs/2506.00620

[40] J. Fan, P. Rosu, A. T. Wang, Z. M. Li, L. Carin, and X. Cheng, "In-context semi-supervised learning," arXiv preprint arXiv:2512.15934, 2025. [Online]. Available: https://arxiv.org/abs/2512.15934

[41] T. Yu, C. Ekbote, N. Morozov, A. Herrera, J. Fan, A. Dominguez Mantes, S. Novello, H. Lashuel, P. Frossard, S. d'Ascoli, G. La Manno, and M. Brbić, "Tissue reconstruction with generative ai," bioRxiv, 2025. [Online]. Available: https://doi.org/10.1101/2025.02.13.638045