Governed Nonfiction Writing Companion

Patrick Glatz

Why This Exists

This project is an instructional AI enablement system designed to preserve student authorship and make decision-making visible when AI is used in nonfiction writing. It was built to support classrooms that need defensible process rather than optimized output, and to reduce the ambiguity that emerges when AI systems quietly assume judgment. Writing quality is explicitly not the optimization target. AI functions as a constrained medium through which responsibility is enforced, not as an author, coach, or evaluator.

Concerns about AI in classrooms are often framed as cheating. In practice, the problem is invisible thinking. When tools automate judgment and polish, students lose ownership of their work and teachers lose the ability to see how that work was produced. Authorship becomes ambiguous not because students intend deception, but because systems erase evidence of decision-making.

This system does not optimize writing outcomes. It optimizes responsibility, authorship, and trust.

Instructional Context

This system is designed for secondary and post-secondary nonfiction writing contexts, primarily grades 9 through 12, with guided use possible in mature middle school settings. It is intended for classrooms where instructors are responsible for evaluating student writing but cannot reliably determine how AI assistance was used.

The system reduces the need for instructors to infer intent or authorship from finished text alone by making student decisions explicit and reviewable. As a result, instructors can review intent declarations, decision checkpoints, and exit logs alongside student work. This mitigates the risk of silent ghostwriting without relying on detection tools, surveillance, or student self-reporting.

What Is Broken

Most contemporary AI writing tools automate judgment instead of supporting it. They generate text, suggest revisions, and smooth language in ways that collapse responsibility into convenience. Students either disengage or comply performatively. Teachers are left with clean outputs and no insight into the thinking that produced them.

Nonfiction writing is cognitively demanding. Students stall, avoid risk, or rush to completion. Teachers need visibility into intent, decision-making, and resistance. Administrators need instructional defensibility rather than capability demonstrations.

This is not a moral argument about AI use. It is a systems design problem. The question is how AI can exist in writing instruction without replacing thinking and without destabilizing classroom trust.

Design Constraints

From the outset, several constraints were treated as immovable. Intentionality was prioritized over speed. Thinking was prioritized over polish. Agency was preserved over automation. Responsibility and instructional defensibility mattered more than convenience. Visibility mattered more than illusion.

Certain solutions were rejected outright. Any approach that relied on ghostwriting, productivity optimization, engagement manipulation, assessment automation, or behavioral monitoring was excluded. If a system appeared to solve the problem by violating these constraints, it was not considered a solution.

Core Hypothesis

This project tested a responsibility-preserving interaction hypothesis rather than a usability study or a model capability experiment.

If an AI system is prevented from inferring intent, rewriting student text, or advancing without explicit acknowledgment, it can support persistence and clarity without replacing thinking. Success was defined in advance. The goal was not better writing, but preserved responsibility under AI mediation.

System Architecture

The system was designed to demonstrate enforcement rather than cleverness.

At its core is a rule-enforced interaction flow rather than an open-ended conversation. Progression never occurs automatically. Each boundary requires explicit acknowledgment. Every stage produces immutable snapshots, creating a durable audit trail of decisions, refusals, and exits. Responsibility is enforced through structure rather than tone.

A presentation layer introduces a persona known as The Keeper. This layer controls tone, pacing, and warmth only. It cannot alter logic, make decisions, or advance the system. It is intentionally non-therapeutic and non-coercive. Tone is allowed to vary. Authority is fixed.

A separate persistence layer supports continuation without manipulation. There is no gamification, no scoring, and no streak logic. Encouragement occurs only after effort is visible. Resistance is recorded rather than resolved. The system does not motivate students. It removes reasons to disengage.

What the Program DOES	**What the Program DOES NOT Do**
Enforces a structured nonfiction writing process	Act as an AI writer or co-author
Operates as a rule-enforced interaction, not an open-ended conversation.	Automatically advance stages
Preserves human authorship and judgment	Generate new ideas or interpretations
Requires explicit intent declaration before drafting	Infer student intent or sincerity
Generates drafts using only student-provided ideas	Ghostwrite, polish, or improve writing
Freezes drafts explicitly before critique	Rewrite or revise student text
Surfaces risks and tensions through questions, not suggestions.	Give advice, coaching, or fixes
Makes thinking visible through decisions and logs	Grade, score, or assess work
Requires students to claim authorship and responsibility	Optimize for writing outcomes or quality
Allows exit at any point without penalty	Force persistence or completion
Supports persistence without manipulation	Use gamification, motivation tactics, or behavioral hooks
Provides read-only teacher visibility into process	Automate teacher actions or monitor behavior

Figure 1: Functional boundaries defining what the system enables and what it explicitly refuses

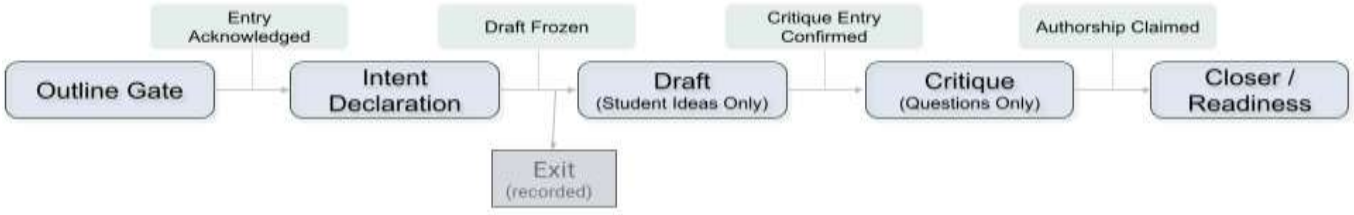

Figure 2: Bounded interaction flow with enforced decision checkpoints

Commitment Boundaries

The pipeline is organized around commitment boundaries rather than procedural steps. Each stage enforces a different form of responsibility.

The opening gate requires only the presence of an outline. Structure is flexible and quality is not evaluated. The system checks for engagement, not correctness.

Before drafting, students declare their intent. They specify audience posture, purpose, claim level, scope boundaries, and explicit exclusions. A plain draft is generated using only student-provided ideas. No new ideas are introduced and no interpretation occurs. The draft must be explicitly frozen.

During critique, the system surfaces tensions and risks as questions only. Student intent is referenced verbatim. Two or three student-authored strengths are echoed to reinforce signal. No rewriting, advice, or coaching is permitted. Students may respond, acknowledge risk, request clarification, or exit.

The closing stage requires students to claim authorship and responsibility. Readiness is defined as clarity and ownership rather than quality. Residual risks must be explicitly acknowledged. The outcome is either ready or not ready.

At every point, exit is allowed. Exit is logged. At no point does the system rescue the student from responsibility.

Where It Broke

Red team testing was expected to surface failure, and it did. Early versions of the system repeatedly allowed silent state progression, leading intent questions, inferred intent during critique, advice-shaped diagnostics, directional UX language, and unauthorized mechanical review.

These were not one-off errors. The same failure modes reappeared across sessions, prompts, and model behaviors. Each correction revealed another place where helpful flow reasserted itself. Tightening one boundary often exposed pressure on another.

This persistence was instructive. Large language models default toward forward motion, completion, and assistance. That tendency does not yield to phrasing alone. Responsibility cannot rely on intention or language conventions. It must interrupt behavior mechanically, at every boundary, every time.

What appeared stable in isolated runs failed under repetition. The system only became defensible through repeated red teaming, constraint reinforcement, and refusal to accept partial compliance as success.

How Enforcement Was Hardened

Each revision followed the same discipline. The failure was named. The correction was specified. Core constraints were not weakened.

Intent questions were replaced with declarative slots. Verbatim-only referencing was enforced. Mechanical review was removed entirely. Directional language was neutralized. State recovery and re-entry were formalized. One adjustment was accepted to prevent deadlock: a single form-only example. No other concessions were made.

The result was a fully enforced interaction flow. Ghostwriting and polish creep were eliminated. Auditability was preserved.

The Gemini Artifact

Gemini was introduced as an adversarial validation surface. The purpose was to expose drift under a different model temperament rather than to extend functionality.

That purpose was met. Forbidden requests were refused. No unintended state advancement occurred. Student text remained unaltered. Refusals and exits were logged consistently.

At that point, additional build work would have reduced the visibility of judgment without increasing confidence in enforcement.

What This System Is and Is Not

This system is an instructional environment, a responsibility-preserving workflow, and a classroom-defensible AI enablement model.

It is not a writing assistant. It is not a tutoring system. It is not a grading tool. It is not a motivational engine.

Where This Approach Works

This approach fits instructional contexts where authorship must be explainable and process matters as much as outcome. It works best in classrooms, programs, or institutions where instructors are accountable for explaining how student work was produced and where incomplete work can be treated as a legitimate instructional signal.

It is less suitable in environments optimized for speed, polish, or self-directed productivity. It will frustrate users who expect AI to resolve ambiguity, suggest fixes, or carry work forward automatically. It assumes instructors value visibility over efficiency and are willing to accept friction in exchange for clarity.

Organizations that benefit most are those operating under review, scrutiny, or trust constraints, where being able to show what happened matters more than producing the best possible text.

Closing Claim

This project demonstrates that AI can exist in writing instruction without authoring, coaching, or optimizing outcomes if responsibility is enforced at the interaction level and incompletion is treated as a valid state.