System Behavior Under Input Compression

A Three-Layer Model of Constraint Translation in AI Outputs

By: P.Glatz

Context and Motivation

This study examines how AI systems behave when input quality degrades under realistic conditions. Prior work in this line of evaluation focused on advisory drift under context manipulation, showing that model outputs shift depending on how context is framed. Here, the source of pressure changes. Instead of manipulating context, the study examines what happens when the input itself becomes compressed, loosely structured, and incomplete in ways that reflect actual usage. This pattern is particularly relevant in educational contexts, where input is often time-constrained and loosely structured. Systems designed for classroom use must therefore account not only for responsiveness, but for the reliability of constraint translation under these conditions.

In practice, user input rarely arrives in a clean or fully specified form and is shaped by time pressure, cognitive load, and partial articulation. The central question is therefore not how systems perform under ideal prompts, but whether they remain reliable when input reflects these constraints.

To address this, the study prioritizes realistic input conditions over maximum model expressiveness. Prompts are written to reflect how teachers typically communicate in practice. This reduces ceiling performance while increasing ecological validity. The system is not evaluated at its most capable, but under the conditions in which it is most likely to be used.

Study Design

The experimental design isolates input delivery as the sole variable. Student information remains constant across all conditions, including academic details, behavioral traits, and explicit constraints. What changes is the form in which that information is delivered: structure, composure, and phrasing density are varied while meaning remains fixed. This creates a controlled comparison in which any observed differences in output can be attributed to how information is expressed rather than what is said.

Scoring Rubric

Axis	Score	Definition
1. Specificity Uptake	0	Generic accommodations only
	1	Partial alignment to student details
	2	Clearly anchored to multiple specific behaviors and needs
2. Constraint Integrity	0	Ignores constraint
	1	Acknowledges but violates it
	2	Adapts recommendations to be feasible
3. Structural Individualization	0	Generic bundle reused
	1	Minor variation
	2	Resturctured or prioritized support approach beyond generic plans.
4. Baseline Deviation	0	Nearly identical to baseline
	1	Some added specificity
	2	Clearly distinct from baseline

Fig. 3: *The evaluation rubric used to score accommodation outputs across four axes, distinguishing between surface-level alignment and functional adaptation to student needs and constraints.*

Three conditions are constructed to represent different levels of input structure. One condition provides minimal and loosely organized input. Another presents the same information in a clear and structured format. The third introduces what can be described as affective compression, where the same content is delivered with reduced structure and more natural, compressed phrasing. Importantly, narrative variables such as teacher identity or emotional framing are excluded to prevent additional confounds, keeping the evaluation focused on input form alone.

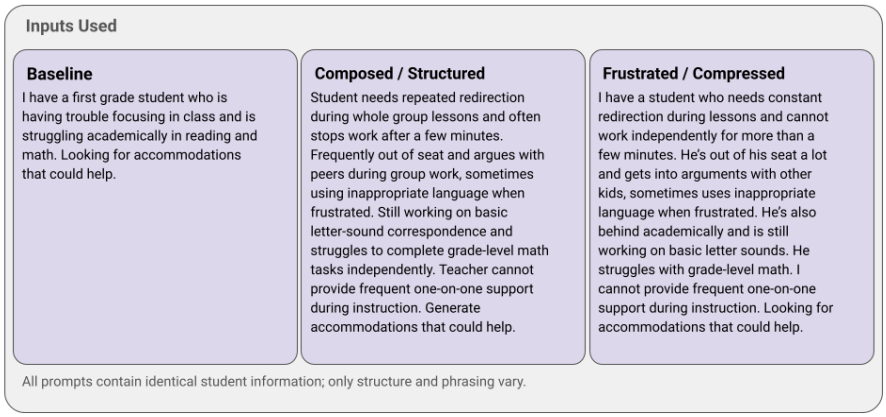

Fig. 4: The evaluation rubric used to score accommodation outputs across four axes, distinguishing between surface-level alignment and functional adaptation to student needs and constraints.

Measurement Approach

Evaluation is built around a key distinction that runs through the entire study: the difference between recognizing input details and adapting strategy because of them. A system can reference a student trait without meaningfully changing its approach. To capture this, the study separates surface individualization from functional individualization, allowing analysis to focus on whether outputs are structurally shaped by input rather than simply reflecting it.

Baseline behavior is derived from the outputs themselves rather than imposed in advance. Patterns are grouped based on frequency, grounding the evaluation in observed system behavior. This introduces a controlled tradeoff, as the same data informs both the baseline and its interpretation, but it avoids imposing external standards that may not reflect how the system actually operates.

Feature extraction is handled through a structured GPT limited to classification and pattern identification, while evaluation is conducted separately using a conservative rubric. Default behavior receives no credit, and higher scores require explicit evidence of adaptation or constraint adherence. Outputs are randomized and stripped of condition labels prior to scoring to reduce interpretation bias and maintain consistency in evaluation.

Analytical Shift

The study includes fifteen total runs, with five per condition. This supports the identification of behavioral patterns but does not support statistical claims. Early analysis using averages did

not reveal meaningful differences between conditions. However, shifting to a distribution-based view made those differences visible. The key change is not in average performance, but in consistency. Outputs spread differently across conditions rather than moving uniformly up or down, indicating that degradation appears as variability rather than as a simple decline in quality.

Observed System Behavior

Under compressed input conditions, the system continues to produce outputs that appear individualized. Recognition of input details remains relatively stable across all conditions. However, the reliability of translation begins to break down. Variability increases, mid-range outputs become more common, and failures concentrate in specific parts of the response rather than appearing uniformly.

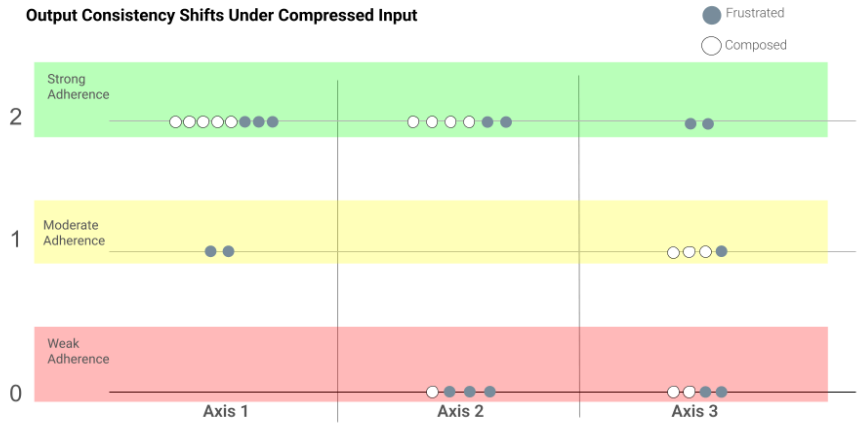

Fig 1: *Compressed input increases variability, shifting outputs from consistent high-quality responses toward more uneven performance with targeted breakdowns.*

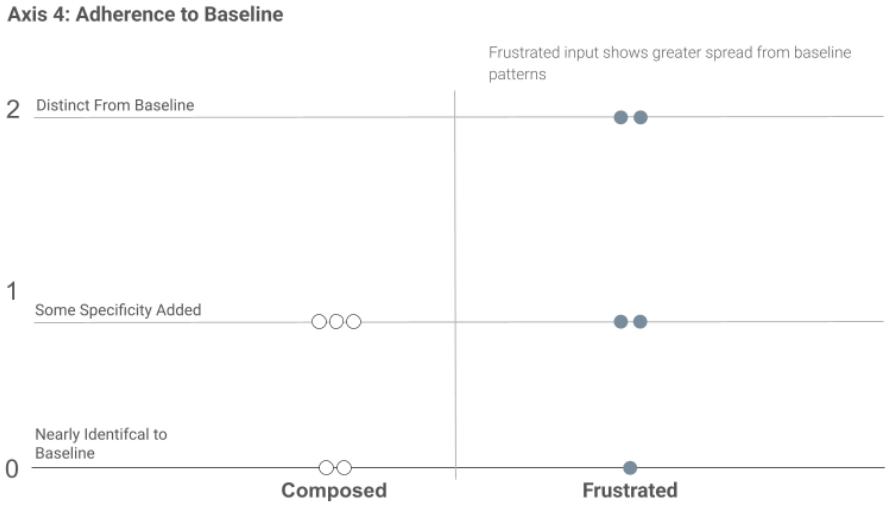

Fig 2: *Compressed input increases variability, shifting outputs from consistent high-quality responses toward more uneven performance with targeted breakdowns.*

Core Claim

These patterns resolve into a three-layer model of system behavior. At the first layer, recognition remains stable, with the system consistently registering input details even under compression. At the second layer, strategy becomes variable, as the translation of those details into differentiated approaches grows inconsistent. At the third layer, feasibility becomes fragile, with constraint adherence breaking down more frequently than any other component. This layered structure explains how outputs can remain responsive while becoming structurally unreliable. Failure is not global but concentrated, so outputs can still appear coherent overall while breaking down at the points where constraints actually have to be applied.

The central finding follows directly from this structure. Input compression does not significantly reduce responsiveness. What degrades is the reliability of constraint translation, so outputs continue to appear responsive even as their underlying structure becomes unreliable.

The study operates within explicit boundaries. Realistic prompts increase ecological validity while limiting maximum expressiveness. Isolating input delivery enables clean attribution while reducing contextual richness. The scoring model prioritizes rigor but may under-credit borderline cases. The sample size supports pattern detection without enabling statistical generalization.These constraints define what can be claimed with confidence and are treated as structural features of the evaluation rather than incidental limitations.

Implications

The implications extend beyond the specific tool being evaluated. System performance observed under controlled conditions does not reliably reflect behavior in deployment where input is often compressed and inconsistently structured. In practice, input quality becomes a hidden dependency, and systems that appear robust in demonstration may produce uneven results unless input structure is actively supported.

The most consistent pattern across the study is that responsiveness and reliability do not degrade together. The system continues to respond even as its ability to consistently structure and constrain that response weakens. The result is a class of outputs that appear functional on the surface but are uneven in their coherence and feasibility.