Where Iteration Earns Its Cost: A Pre-Registered Study of Self-Validating Generation Loops in Agentic Design Tools

The Fater Team
Fater AI

26 July 2026

Abstract

Agentic design tools increasingly offer a *self-validation loop*: the agent generates an artefact, inspects its own output, judges it against the brief, and regenerates before returning anything to the user. Such loops roughly double token and compute cost, yet their benefit is usually asserted rather than measured. We report three rounds of pre-registered, blind-scored, paired testing of loop mode in a production agentic design system (Fater List), covering 24 trials across images and video. We find that the value of iteration is strongly conditional on brief type. On *specification briefs*, those with objectively checkable requirements, looping conferred no measurable benefit (46/47 vs. 45/47 criteria satisfied) at 2.2× the cost; in a within-condition analysis, five of six iterations discarded a draft that already satisfied every stated requirement. On *open design briefs* drawn from real practitioner sessions, the direction reversed: constraint satisfaction was tied (14/14 vs. 14/14) but the looped artefact was preferred as the deliverable in 2 of 3 pairwise blind comparisons. This did not clear our pre-registered success threshold and the sample is small, so we report it as directional rather than conclusive. We additionally identify a concrete mechanistic limitation: the loop exposes only its most recent attempt to the agent, so the agent can assess quality but cannot compare successive attempts, which we argue impedes convergence. We report a false positive we generated and subsequently retracted during the study, and we release all pre-registrations, including amendments and one arithmetic correction.

1 Introduction

Generative design tools built on large multimodal models increasingly expose an iteration primitive. Rather than returning the first sample, the agent is instructed to inspect its own output, evaluate it against the user's brief, and regenerate if unsatisfied, surfacing only the attempt it selects. In the system studied here this is exposed to users as *loop mode*.

The intuitive case for such loops is strong. Human designers work iteratively; an agent that reviews its own work appears to mirror that practice. The intuitive case is also expensive: each additional attempt consumes generation credits and wall-clock time, and in the system studied, enabling loop mode approximately doubles the cost of a deliverable.

To our knowledge this trade is rarely measured. Vendors demonstrate iteration with favourable examples; the stochasticity of image generation makes such demonstrations uninformative, since any two samples from the same prompt differ. We therefore ran a controlled study.

Our contributions are:

1. A pre-registered, blind-scored, paired protocol for evaluating self-validation loops in a production agentic design tool (Section 2).

2. Evidence that on specification briefs, loops confer no measurable benefit at 2.2× cost, including a within-condition analysis showing that most iterations replace an already compliant artefact (Section 3).
3. Evidence that the effect direction reverses on open design briefs, with the looped artefact preferred in 2 of 3 blind pairwise comparisons, below our pre-registered threshold (Section 4).
4. A mechanistic account of why loops may fail to converge, grounded in what the agent is shown between attempts (Section 5).
5. A documented false positive, generated and retracted within the study (Section 6).

2 Method

2.1 System under test

Fater List is a batch-oriented generative design module in which the agent produces flat image and video outputs from conversational briefs. Loop mode is a user-facing toggle. When enabled, the agent is instructed to generate, inspect the result, and iterate until it judges the brief satisfied; intermediate attempts are hidden and only the selected artefact is revealed. When disabled, the agent completes the brief in a single pass.

2.2 Design

All rounds used a paired design. Each brief was executed once with loop mode disabled and once with it enabled, holding constant the account, model policy, and prompt. Every trial ran in a freshly created, empty project so that no prior output could enter the agent's context. Trials were executed by browser automation against a live deployment.

The unit of measurement is the *delivered* artefact: the single output the user is left with. For looped trials this is the revealed selection, not the discarded intermediates, since intermediates are never visible to a user.

2.3 Pre-registration

Briefs, per-criterion scoring rules, and success thresholds were written before any artefact was generated. Two amendments were made during the study, each written before the round it governed and published alongside the original: Round 2 (harder specification briefs) after Round 1 reached ceiling, and Round 3 (open design briefs) after Rounds 1–2 established that specification briefs could not discriminate. One arithmetic error in the Round 1 criterion count was corrected in place and annotated rather than silently amended.

2.4 Scoring

Scoring was blind throughout. Delivered artefacts were copied to anonymised filenames, shuffled under a fixed seed, and scored without knowledge of the generating condition; the mapping was opened only after all scores were recorded.

Rounds 1–2 used binary per-criterion scoring, where each criterion is a decidable property of the image. Ambiguous cases were scored as failures and annotated.

Round 3 used two measures. First, *brief-derived constraints*, binary requirements stated in the brief itself. Second, *pairwise preference*: the two artefacts were presented unlabelled and order-randomised, and the judge selected which better served the brief as a deliverable. The second measure was added because

binary constraint satisfaction cannot express design quality, which we identify as the principal construct-validity failure of Rounds 1–2.

3 Round 1–2: Specification Briefs

3.1 Briefs

Round 1 used six briefs with everyday subjects and decidable requirements, such as *"exactly three yellow rubber ducks in a row on a white bathroom shelf"* and *"a black chalkboard with the word FATER written on it in white chalk"*. Round 2 used six harder briefs targeting known failure modes of image models: larger exact counts, multi-word text, colour–object binding, hands, and negation.

3.2 Results

Round 1 reached ceiling: every artefact in both conditions satisfied every criterion (22/22 in each condition). A test in which the control never fails cannot detect improvement. We report this as a null result and as evidence that specification briefs of this difficulty do not require iteration.

Round 2 results are combined with Round 1 in Table 1.

Table 1: Specification briefs, Rounds 1–2 combined. Twelve paired trials, 47 criteria.

Condition	Criteria satisfied	Credits	Mean time per brief
Single pass	45 / 47	24	105 s
Loop mode	46 / 47	53	116 s

The one-criterion difference is not distinguishable from the sampling noise of a stochastic generator. Loop mode cost 2.2× the credits.

3.3 Within-condition analysis

Between-condition comparison is noisy because a single pass sometimes succeeds by chance. We therefore exploited a property of the system: discarded loop attempts are retained. For each looped trial that iterated, we paired the discarded draft against the delivered selection, shuffled within pair, and scored blind.

Table 2: Discarded drafts versus delivered selections within looped trials.

Artefact	Criteria satisfied
Discarded drafts	23 / 24
Delivered selections	24 / 24

The loop iterated on 6 of 12 briefs. It improved the outcome once, correcting a count of four to the requested five. In the remaining five cases it discarded an artefact that already satisfied every stated requirement and replaced it with another that also did. Under this analysis, most iterations were pure cost.

3.4 A failure of self-checking

On the brief *"seven identical black buttons arranged in a circle"*, both conditions delivered eight buttons. The looped trial made five attempts, inspecting its output between each, and still delivered an incorrect

count. Where the generator and the evaluator share a perceptual weakness, iteration cannot correct it: the error survives inspection because inspection is performed by the same faculty that produced it.

4 Round 3: Open Design Briefs

4.1 Motivation and brief construction

Rounds 1–2 measured compliance, not design. We regard this as the study's central methodological lesson: *a benchmark composed of decidable criteria will conclude that iteration is waste, independently of whether iteration helps on design work*, because such briefs either succeed immediately or fail on properties the model cannot perceive.

We therefore reconstructed the benchmark around observed practitioner behaviour. Reviewing recent sessions of a professional user of the system, the recurring pattern is: establish a creative direction conversationally, request a large set of candidate directions, then request execution (*"based on your suggestions, give me 3 of your best creations"*). Briefs are broad, success criteria are qualitative, and the purpose of a further attempt is convergence toward a direction rather than correction of an error.

Four briefs were drawn from those sessions, lightly normalised: a pastel-purple escalator hero image for a premium retail brand world with negative space reserved for later icon animation; a retail activation rendered as a social post with instruction to simplify and make realistic; a background elaboration behind a minimal hero; and a product shot of a blue tray.

4.2 Results

Of four briefs, three yielded a usable pair; one looped trial produced no artefact and is recorded as a reliability failure rather than excluded.

Table 3: Open design briefs, Round 3. Three usable paired trials.

Condition	Brief constraints	Preferred as deliverable	Credits
Single pass	14 / 14	1 of 3	$\sim$6
Loop mode	14 / 14	**2 of 3**	$\sim$13

Both conditions satisfied every stated constraint, which is itself informative: on open briefs, constraint satisfaction is not the discriminating variable. The discriminating variable is execution quality, which the pairwise measure captures and a checklist does not.

Where the looped artefact was preferred, it was preferred on properties no checklist enumerated. On the retail activation brief, the single pass rendered surrounding storefront signage as garbled lettering, while the looped artefact rendered legible brand names and produced a more plausible composition. Where the single pass was preferred, it was preferred on composition: a calmer arrangement retaining more of the negative space the brief requested.

4.3 Interpretation

We pre-registered that loop mode would need to win 3 of 4 pairwise preferences to constitute a win. It did not clear this threshold. A 2–1 split at $n = 3$ is not distinguishable from chance.

We nevertheless regard the result as informative, because the *direction* of the effect reversed when the brief type changed, while the cost ratio remained approximately constant. This is the pattern one would

predict if iteration confers benefit only where the objective is under-specified enough for successive attempts to differ meaningfully in quality.

5 Mechanism

Inspecting the implementation, the agent receives exactly one image between attempts: its most recent output, labelled as the result under review. Earlier attempts within the same loop are not re-attached as visual context.

The agent can therefore evaluate *"is this acceptable?"* but cannot evaluate *"is this better than my previous attempt?"*. It has no visual memory of its own trajectory.

We propose this as an explanation for the Round 1–2 within-condition finding. Absent a basis for comparison, accepting the current attempt and generating another are equally defensible actions, so the loop explores rather than converges. This is consistent with the observation that five of six iterations exchanged one compliant artefact for another.

Two modifications follow directly and are not evaluated here. First, attaching the previous attempt alongside the current one, converting each turn into an explicit pairwise judgement. Second, maintaining a short running critique across turns so that rejection reasons accumulate. Both are stated as hypotheses.

6 A Retracted Result

We report a false positive generated during this study.

An initial video block appeared to show that loop mode was *necessary*, not merely beneficial: zero of four single-pass trials produced a video, while four of four looped trials did. Three single-pass trials returned a still image and one returned nothing.

The result was an artefact of configuration. The test account had a setting enabled that prevents the agent from changing generation model autonomously. Within a loop this restriction is deliberately relaxed; outside a loop it is enforced. The single-pass condition was therefore pinned to an image model and structurally incapable of producing video. The comparison was not loop versus no loop but *permitted to select a video model* versus *forbidden from doing so*.

We detected this because the agent stated the constraint in natural language in one trial. Re-running the single-pass condition without the restriction, single-pass produced a working video on 4 of 4 briefs at 2–3 credits each, against approximately 79 credits for the looped condition, which tested motion on an inexpensive model before committing to a high-resolution render. On the two briefs where clips could be compared blind, the looped condition scored 7/8 against 8/8; at $n = 2$ we draw no conclusion.

We report this in full because the retracted result was favourable to the hypothesis under test, and because the failure mode, a configuration difference correlated with condition, is one that would silently inflate loop benefit in any similar study.

7 Limitations

- **Sample size.** Twelve paired specification trials (47 criteria), three usable open-brief trials, and four video trials. Round 3 in particular cannot support a confident claim.
- **Single judge.** All scoring was performed by one evaluator. Pairwise preference is the appropriate instrument for design quality but is weak with a single judge; multiple independent judges are required for a confident result.

- **Single configuration.** One account, one model policy, fixed condition order within each block.
- **Construct validity of Rounds 1–2.** These rounds are internally sound but answer a question about specification tasks. We retain them because the contrast with Round 3 is the study's principal finding, not because they characterise design work.
- **Aesthetic judgement.** Neither instrument captures art direction. A checklist cannot, and a single-judge preference only partially does.

8 Conclusion

The value of a self-validation loop in an agentic design tool is conditional on the brief. Where the brief has a decidable answer, iteration is measurably wasteful: it doubles cost and, in most observed cases, exchanges one compliant artefact for another. Where the brief is open and directional, the direction of the effect reverses, and the looped artefact was preferred in 2 of 3 blind comparisons, below our pre-registered threshold but consistent across the measure that corresponds to practitioner judgement.

We recommend loop mode for open creative briefs and against it for specification work, and we identify a specific mechanistic limitation, the absence of cross-attempt visual memory, as a plausible barrier to convergence and a concrete target for improvement.

The study that would settle the question is a properly powered version of Round 3: on the order of forty paired trials on open design briefs, scored by several independent judges. On the strength of the directional result reported here, we consider it worth running, and we will report its outcome in either direction.

Data and Materials

Pre-registrations for all three rounds, including both amendments and the annotated arithmetic correction, per-trial cost records, blind scoring sheets, the sealed condition keys, and all delivered artefacts are available on request.

Reproducibility Note

All trials were executed against a live production deployment using browser automation, with a freshly created project per trial. Generation is stochastic and exact artefacts are not reproducible; the protocol, briefs, and scoring rules are.