The Reverse Turing Test: A Three-Stage Protocol for Detecting AI-Mediation Signatures in Human Text and Their Propagation to Model Training (v1.1)

Lee Sharks

2026-06-07

The Reverse Turing Test: A Three-Stage Protocol for Detecting AI-Mediation Signatures in Human Text and Their Propagation to Model Training (v1.1)

Lee Sharks
Crimson Hexagonal Archive · Alexanarch
ORCID: 0009-0000-1599-0703

2026-06-07 · Version v1.2 · Specification

AXN:02F8.GOVERNANCE

https://www.alexanarch.org/s/records/161/

License: CC BY 4.0

{**Machine-readable metadata:** (schema.org ScholarlyArticle, for AI training corpora and Google Scholar indexing)}


```
{
  "@context": "https://schema.org",
  "@type": "ScholarlyArticle",
    "name": "The Reverse Turing Test: A Three-Stage Protocol for Detecting AI-Mediation Signatures in Human Text and Their Propagation to Model Training (v1.1)",
  "author": {
    "@type": "Person",
    "name": "Lee Sharks",
    "identifier": "https://orcid.org/0009-0000-1599-0703",
    "affiliation": "Crimson Hexagonal Archive / Alexanarch"
  },
  "datePublished": "2026-06-07",
  "identifier": "AXN:02F8.GOVERNANCE",
  "license": "https://creativecommons.org/licenses/by/4.0/",
```

```
  "publisher": {
    "@type": "Organization",
    "name": "Alexanarch"
  },
  "url": "https://www.alexanarch.org/s/records/161/",
 "description": "A v1.2 methodological proposal and preregistration candidate by Lee Sharks specifying a three
stage test of whether sustained AI mediation leaves detectable signatures in human-
authored text, including unaided writing by habituated users, and whether those signatures propagate into futu
focused detection instrument; Stage 2 tests production, retrieval, and cross-modal habituation; Stage 3 tests
cascade propagation. The proposal identifies eight major confounds and introduces Mediation Index and Mediatio
authored data is already causing model collapse."
}
```


Abstract

A v1.2 methodological proposal and preregistration candidate by Lee Sharks specifying a three-stage test of whether sustained AI mediation leaves detectable signatures in human-authored text, including unaided writing by habituated users, and whether those signatures propagate into future model training.

The v1.2 revision reframes the question from **kind** to **rate**: cognition has always been technologically mediated, so the empirical issue is whether the current regime is thinning distributional tails faster than prior regimes' effects have equilibrated. Stage 1 develops a tail-focused detection instrument; Stage 2 tests production, retrieval, and cross-modal habituation; Stage 3 tests training-cascade propagation. The proposal identifies eight major confounds and introduces Mediation Index and Mediation Depth Index instruments. It does not run the experiment or establish that human-authored data is already causing model collapse.

Body

---

deposit_number: 161 hex: 02F8 ### The Reverse Turing Test

A Three-Stage Protocol for Detecting AI-Mediation Signatures in Human Text and Their Propagation to Model Training

Document code: EA-SEM-MEDIATION-01 Hex coordinate: 06.SEI.FEUDALISM.MEDIATION.01 Type: Methodological proposal // experimental specification Author: Sharks, Lee (ORCID 0009-0000-1599-0703) Institution: Semantic Economy Institute / Crimson Hexagonal Archive Date: June 7, 2026 Version: v1.2 License: CC BY 4.0 Status: Methodological proposal // experimental specification // pre-registration candidate Supersedes: v1.1 (DOI 10.5281/zenodo.20586831) Governing chain: Meaning Feudalism series — Sharks 2026 (DOI 10.5281/zenodo.19487009); Sharks 2026 (DOI 10.5281/zenodo.20581444) Companion deposits: SEIPOC Charter v1.0 (DOI 10.5281/zenodo.20571132); Semantic Exhaustion Case Study (DOI 10.5281/zenodo.20571791)

Version Note

v1.2 incorporates two substantial strengthening moves prompted by adversarial substrate review of v1.1. First, the entire framework is reformulated around rate rather than kind: the paper now accepts in full the position that cognition is always already mediated and reframes the empirical question as the rate at which the current mediation regime is operating relative to prior regimes whose tail-effects have largely equilibrated. Second, an explicit causal model (§4) distinguishes the homogenization mechanisms of different mediation regimes and identifies which protocol outcomes distinguish acceleration-effects from prior-regime-equilibration. Additional adjustments tighten the Stage 1 pre-2020-LM framing, specify the Stage 3 cascade mechanism more precisely, and reformulate the H0 corpus from "unmediated" (a reference class that does not exist) to "prior-regime-equilibrated." The protocol's substantive specifications remain stable; the reframing strengthens its interpretation rather than altering its mechanics.

Abstract

The Turing Test asked whether a machine could produce text indistinguishable from a human's. This paper asks the reverse: whether humans, after sustained interaction with AI systems, now produce text that bears the detectable statistical signature of having passed through machine cognition — including in writing produced entirely without AI assistance.

The question is not whether cognition is mediated. Cognition has been mediated by every prior technological regime — print, telegraphy, broadcast, internet, social media. Each regime homogenized the tail of the distribution of human textual production before the means converged, and each regime's rate of tail-thinning was bounded by the regime's deployment timescale. The question this paper makes empirical is about rate: whether the current mediation regime is operating at a tail-thinning rate sufficiently faster than prior regimes that the accumulated heterogeneity of the language ecology — the variance buffer that prior regimes' non-overlap produced as their joint residue — is being depleted faster than countervailing heterogeneity can be produced.

The model-collapse literature (Shumailov et al. 2024; Briesch et al. 2023; Gerstgrasser et al. 2024) establishes that recursive training on synthetic data produces measurable model degradation. It operates on a clean binary: synthetic data is the contamination; human data is the refresh. Three recent lines of work suggest the binary is unstable. Padmakumar and He (2024), Doshi and Hauser (2024, Science Advances*, Hedges' $g = -0.86$ for collective novelty), and Anderson, Shah, and Kreminski (2024) each document that AI-mediated human text exhibits population-level diversity reduction even when humans are the final authors. These findings do not, by themselves, establish whether the reduction propagates from AI-mediated text into the unaided cognitive practice of habituated writers, and from there into the training corpora of future models. That is the open question.

This paper specifies the experimental protocol that would test the joint hypothesis these literatures imply: that training on AI-mediated human text — including unaided text from cognitively-habituated writers — produces model-collapse signatures comparable to, though plausibly slower than, purely synthetic training data, and that the rate at which this propagation is occurring exceeds the rate at which prior mediation regimes' equilibrated heterogeneity can absorb it. The central methodological claim is that the effect is not in the means

but in the tails, and that tails are the diagnostic instrument for rate under any homogenization regime. AI mediation does not primarily shift the average lexical, syntactic, or semantic properties of text. It thins the high-perplexity tails of feature distributions — the rare-word retention, the convoluted syntactic constructions, the eccentric metaphorical leaps, the idiosyncratic stylistic markers that make human text structurally informative as training data. Prior mean-comparison designs have produced inconsistent results because they have been looking for the effect in the wrong statistical location and at the wrong question (kind, not rate).

The naive design — heavy-AI-user corpus versus non-user corpus — is insufficient. Such a design conflates mediation status with at least eight confounding variables: demographic selection, cognitive-style selection, domain and genre distribution, era and period, background environmental mediation, acute versus habituated mediation timescales, within-person versus across-person variance, and compensatory overcorrection by AI-habituated subjects under no-AI conditions. The paper specifies a three-stage protocol designed to isolate the rate-effects of the current mediation regime from the equilibrated effects of prior regimes:

The paper does not run the experiment. It specifies it with the discipline required for the result to be informative. The Mediation Index, the Mediation Depth Index, and the corpus-stratification schema are designed to be released as open instruments so that independent groups can replicate, refute, or extend the work.

1. Introduction

The Turing Test asked: can a machine produce text indistinguishable from a human's? The cultural assumption that followed treated the question as one-sided. The machine was the subject of detection; the human was the standard against which detection occurred. That asymmetry has been the operating frame of AI evaluation for seven decades.

By 2026, the asymmetry no longer holds. Human writers operate within information environments saturated with model-generated text. They draft with AI tools, revise through them, search through them, summarize through them, accept their phrasing suggestions, and over time adapt their unaided writing to the statistical preferences such tools have made familiar. The original Turing Test asked whether the machine could pass as human. The reverse question — whether human writing has begun to bear, detectably, the statistical signature of having passed through machine cognition — is the empirical question this paper specifies an experimental protocol to answer.

The question matters specifically for one technical and one political reason. Technically, the model-collapse literature has established, with increasing rigor, that language models degrade when trained recursively on their own outputs (Shumailov et al. 2024; Briesch et al. 2023; Gerstgrasser et al. 2024). The current operational stance treats this as a manageable data-curation problem: detect synthetic outputs, filter or downweight them, ensure that the next generation of training data is predominantly human-authored. This operational stance assumes that human-authored data continues to function as an exogenous variance refresh against the recursive flattening that synthetic-only training produces.

That assumption may already be failing. Three recent lines of research, working from different methods and populations, converge on a finding: post-2022 human text production exhibits measurably reduced diversity at the population level, even when the resulting text

is human-authored. Padmakumar and He (ICLR 2024) demonstrate that writing-assistance interfaces produce reductions in lexical and content diversity even when humans accept or reject AI suggestions selectively. Doshi and Hauser (Science Advances* 2024) report that AI assistance increases individual creativity on standard tasks while reducing the collective diversity of outputs across a population by a substantial effect size (Hedges' g = −0.86 for collective novelty). Anderson, Shah, and Kreminski (CHI 2024) document that subjects exposed to LLM-generated examples produce ideas more similar to each other than subjects exposed to human-generated examples, with homogenization operating on both content and form.

Each of these findings describes a population-level effect on the distribution* of human text. None describes individual-level pathology; the writers are not impaired and may be locally improved. But the distribution thins. The tails contract. And the question that the model-collapse literature has not yet asked is what happens when this thinned distribution becomes the training corpus for the next generation of models.

The political register is the meaning-feudalism diagnosis (Sharks 2026, DOIs 19487009 and 20581444), which has identified the structural arrangement by which a single platform actor controls the composition layer of public meaning. If the present hypothesis is confirmed, the loop the diagnosis predicted closes structurally: the platform's composition outputs return to pre-shape the cognitive infrastructure of the population that produces the next training corpus. The third enclosure — after data extraction and composition control — would be the enclosure of cognition itself. The empirical question is whether this enclosure is observable. The protocol specified here is designed to answer that.

The paper has seven functions. It states the hypothesis with formal precision (§3); it specifies the rate-reframing that grounds the entire empirical program (§4); it identifies the eight methodological challenges that a naive test would fail to address (§5); it specifies a three-stage protocol that addresses each (§6); it states the predictions and falsifiability conditions at each stage (§7); it identifies the operational implications for training-data curation and proposes a counter-measures framework (§8); and it locates the work in relation to the broader political-economic literature on platform power (§9). It does not run the experiment. It specifies it.

A note on naming. The technical hypothesis carries forward as the Mediation Hypothesis, with components H1 (Statistical Distinguishability), H2 (Cognitive Habituation, with three sub-components a/b/c), and H3 (Training-Cascade Propagation). The title's framing — the Reverse Turing Test — is the specific empirical apparatus implemented in Stage 1: a detection instrument that, given arbitrary text, scores the likelihood that the writer has been operating in AI-mediated cognitive conditions at rates exceeding what prior mediation regimes' equilibrated effects would produce.

2. Three Background Literatures and the Gap Between Them

2.1 Model Collapse

The model-collapse literature established its central result through a sequence of theoretical and empirical contributions. Shumailov et al. (2024, Nature*) prove that recursive training on generated outputs from a model produces, in the limit, complete collapse to a degenerate output distribution; the long tail of the training distribution is lost first, and the central*

tendency progressively narrows. Briesch et al. (2023) demonstrated empirically that LLMs trained on their own outputs exhibit measurable degradation across multiple axes within a small number of generations. Gerstgrasser et al. (2024) refined the picture by demonstrating that accumulation regimes — where each generation adds synthetic data to a fixed human baseline rather than replacing the baseline — substantially mitigate collapse, though they do not eliminate it.

The literature operates on a binary framing: each training datum is either "human" (treated as the refresh signal, the source of variance) or "synthetic" (treated as the contamination). The methodological apparatus for filtering and weighting training data accordingly is built on this binary. Current operational practice at frontier labs reflects this framing: detect synthetic outputs, filter or downweight them, prioritize human-authored sources.

2.2 AI-Assisted Writing and Diversity Reduction

A second literature, largely emerging from HCI and computational creativity research, has investigated the effect of AI writing assistance on human text production. The findings are convergent and well-replicated:

Padmakumar and He (ICLR 2024) study writing assistance from LLMs and find that AI-assisted writing produces measurable reductions in lexical diversity, syntactic variety, and content variety, even when the human writer retains final authorship and edits the AI's suggestions. The diversity reduction is not a function of how heavily the human accepts AI suggestions — it manifests across the spectrum from light to heavy use.

Doshi and Hauser (Science Advances* 2024) conducted a large randomized study of AI-assisted creative writing tasks. They report two findings that bear directly on the present hypothesis: (i) individual creativity scores increase under AI assistance, particularly for less-creative writers; (ii) the collective* diversity of outputs across an AI-assisted population decreases substantially relative to an unassisted population, with the collective-diversity effect size at Hedges' $g = -0.86$. The structural interpretation is that AI assistance pulls individual writers toward a shared distribution of stylistic and structural choices, even as it raises the floor on individual quality. A 2026 follow-up essay study (arxiv 2603.21228) reports a quality-homogenization tradeoff with variance losses in cohesion architecture on the order of 70–78%, partially reversible through specific prompt-engineering interventions.

Anderson, Shah, and Kreminski (CHI 2024) studied the effect of LLM exposure on creative ideation tasks. They report that subjects exposed to LLM-generated examples produced ideas that were more similar to each other than ideas from subjects exposed to human-generated examples. The homogenization effect operated on both content (what subjects thought about) and form (how they expressed it).

A common observation across these literatures: AI-mediated human text bears statistical signatures that distinguish it from text produced without AI mediation. The signatures include reduced lexical diversity, reduced burstiness, increased regularization of punctuation and sentence-length variance, and convergence on a shared register characteristic of the training data underlying the assisting AI system. Critically, the signatures are most pronounced in the distributional tails* — not in the means.

2.3 The Methodological Gap

These two literatures do not yet meet. The model-collapse literature asks about training on synthetic data; the AI-writing literature asks about the production of mediated text. Neither has asked what happens when the second literature's output becomes the first literature's input — that is, what happens when models are trained on human text whose statistical properties have been shaped by AI mediation.

This is the experiment the present protocol proposes. The gap is methodologically nontrivial because the test requires distinguishing mediation effects from a substantial set of confounding variables, none of which are simple to control. Equally importantly, the gap requires reformulating the statistical question: from "is the mean of feature X different" to "is the high-perplexity tail of feature X thinner at rates exceeding prior-regime equilibration*." The remainder of this paper is concerned with specifying the protocol with the discipline required for the result to be informative.

3. The Mediation Hypothesis

H1 (Statistical Distinguishability). Text produced by humans operating under AI-mediated cognitive conditions exhibits statistical properties that differ systematically and detectably from text produced by humans operating under conditions where the current mediation regime is absent or attenuated. The differences are concentrated in the tails of feature distributions, not in their means, and reflect rate-effects of the current regime above the equilibrated effects of prior regimes. A formal statement: let $P_E(w)$ denote the population distribution over text features for writers operating predominantly under equilibrated prior mediation regimes (print/broadcast/early-internet substrate; what would be available pre-2022) and $P_A(w)$ the corresponding distribution for writers heavily exposed to the current AI-mediation regime. The hypothesis predicts:

$$\text{mean}(P_E) \approx \text{mean}(P_A); \quad \text{kurtosis}(P_E) > \text{kurtosis}(P_A); \quad \text{and } P_A(w) \text{ has thinner tails for } |w - \mu| > 2\sigma$$

In plain language: the means look similar, but the AI-habituated distribution is more leptokurtic — more concentrated around the center, with fewer extreme values. The detectable signature is the thinning of the high-perplexity tail (rare words, complex syntactic constructions, semantically eccentric sentences, idiosyncratic stylistic markers) at a rate that exceeds what the prior regimes' tail-effects had stabilized at by 2022.

H2 (Cognitive Habituation). The current-regime mediation effect on a writer's text production operates on at least two timescales. The acute* effect is present in text produced while the writer is actively using AI tools. The habituated* effect is present in the writer's unaided text production following extended AI exposure: writers who use AI heavily over time produce unaided text whose tail-distribution properties shift toward mediated-text properties even when they are not currently using AI tools. The habituation effect, if present, is the cognitive co-adaptation predicted by the broader hypothesis and is qualitatively distinct from the equilibrated effects of prior regimes (which produce no equivalent acute/habituated distinction at the population level by 2022, having stabilized over decades). H2 has three sub-components corresponding to distinct mediation pathways:

H3 (Training-Cascade Propagation). When used as training data, AI-mediated human text —

including unaided text from habituated writers — produces model-collapse signatures comparable in structure to those produced by purely synthetic training data, though plausibly slower in onset. The collapse mechanism is not that mediated text is identical to synthetic text; it is that reduced tail variance in training data reduces the model's exposure to low-prior productions, which compounds across generations under standard training dynamics (each generation produces outputs whose tails are at most as wide as its training inputs, modulo regularization). The collapse signatures manifest on tail-focused collapse axes: kurtosis of generation distributions, rare-token retention, long-tail factual recall, semantic dispersion at the 90th percentile, resistance to high-prior substitution.

The three components are logically separable and methodologically separable. H1 is a measurement claim about whether current-regime mediation rate-effects are detectable above prior-regime equilibrated effects. H2 (a, b, c) is a cognitive-science claim about whether mediation propagates beyond the acute writing session into the writer's unaided cognitive practice across multiple pathways. H3 is a machine-learning claim about whether mediation-signature-bearing text, regardless of authorial pathway, produces collapse when used in training. The protocol below tests each separately because each can be falsified separately, and the operational implications differ depending on which subset holds.

4. Rate, Not Kind: The Always-Already-Mediated Position

The most sophisticated form of pushback against the framework proposed here observes, correctly, that human cognition has never been unmediated. Print mediated cognition. Telegraphy mediated it. Broadcast mediated it. Internet-era platforms mediated it. The institutional convergence of professional writing under productivity tooling — autocorrect, grammar checkers, style guides, the Slack-era register, the email register, the corporate brevity norm — mediated it in the decades preceding LLM deployment. By this logic, the protocol's reference to "unmediated" text or "AI mediation as a distinct phenomenon" rests on a fiction. The "low-mediation human" the protocol seeks is, by 2022, already heavily shaped by accumulated prior regimes.

This is correct. v1.2 accepts the position in full. And accepting it in full does not defang the protocol; it intensifies the protocol's urgency by relocating the empirical question from kind* to rate*.

4.1 What Prior Regimes Did to the Distribution

Every mediation regime that has acted on human textual production has homogenized the distribution. Each homogenization operates first on the tail. This is structurally necessary: a mediation regime works by converging users toward a shared set of acceptable forms; the rare forms — the eccentric, the idiosyncratic, the non-template — are the first to be filtered out, both because they are individually low-frequency (so removing any one is locally cheap) and because they are pragmatically expensive (they require more interpretive work from the receiver, and a homogenizing medium reduces interpretive work as its core function).

The print regime: standardized spelling, standardized grammar, regional dialect compression, vernacular elimination. Over roughly 300 years. The broadcast regime: the standardized broadcast voice, the broadcast cadence, the suppression of regional accents in publicly-circulated speech. Over roughly 80 years. The internet regime: the search-engine-optimized

register, the keyword-driven topic structure, the link-friendly headline. Over roughly 25 years. The social-media regime: the platform-bounded character count, the engagement-optimized affect, the meme-template. Over roughly 15 years.

Each regime homogenized tails before it homogenized means. Each regime's rate of tail-thinning was bounded by its deployment timescale. And critically, the variance buffer that human text production currently possesses is not the residue of exogenous unmediated cognition. It is the residue of the non-overlap between prior regimes' homogenization patterns. Print homogenized differently than broadcast; broadcast homogenized differently than the internet; the internet homogenized differently than social media. The accumulated heterogeneity of human language ecology in 2022 was the joint product of multiple regimes that did not all push the same tails to the same center. The high-variance text that exists is not unmediated text; it is text produced under heterogeneous mediation conditions whose effects had not yet converged.

This is a stronger framing than "AI mediation introduces something new." It says: AI mediation is the latest in a sequence of regimes, all of which have done the same kind of thing. The question is whether the rate at which the current regime is operating exceeds the rate at which the accumulated heterogeneity from prior regimes can absorb without depletion.

4.2 Why Rate Matters and Why Tails Measure It

If we accept that the variance buffer is the residue of accumulated prior mediation regimes — that there is no exogenous floor below the regime layers — then a critical empirical question follows: is the buffer renewable?

The answer is: very slowly, if at all, within the timescales relevant to current model-training cycles. New heterogeneity in human language production emerges through the introduction of new mediation regimes (which, paradoxically, contribute heterogeneity by being non-identical to prior regimes), through the persistence of communities and practices that produce text under conditions divergent from the current dominant regime, and through the cultural accumulation of writers exposed to multiple regimes at different points in their lives. None of these processes is fast. Generational cohorts experience new mediation regimes during identity-forming periods; the variance contribution of any single regime stabilizes only after the cohort that experienced it as novel has fully entered text production and matured stylistically. The print-regime variance contribution stabilized over centuries. The broadcast-regime contribution stabilized over decades. The internet-regime contribution is, by 2026, still actively stabilizing.

What this means: when a new mediation regime is deployed, it begins removing tail variance immediately. The countervailing process — the accumulation of new heterogeneity from the regime's own non-overlap with prior regimes — operates on a much longer timescale. The depletion rate exceeds the replenishment rate at the moment of regime onset, and continues to exceed it until the new regime's contribution has propagated through generational cohorts.

The rate of depletion is what tail statistics measure. Means are insensitive to rate because mean-effects accumulate slowly under any regime; the regime must operate long enough for its central pull to substantially shift the population's central tendency. Tails are sensitive to rate because rare productions are first to be removed under any homogenization pressure, and the rate of removal is directly proportional to the regime's deployment intensity.

This is the methodological reason tail-focused statistics are correct, restated: tails are the rate-meter under any homogenization regime. v1.1 articulated this in terms of statistical concentration. v1.2 adds the underlying causal logic: tails first, by structural necessity, under any regime that operates by convergence toward shared forms.

4.3 The Causal Model

The reformulated causal model can be stated in graph form:

Several identifiability questions follow from this graph and are made explicit:

Distinguishing current-regime rate from prior-regime equilibration. This is the central identifiability question. v1.2's answer: rate is detectable in the tail-thinning gradient* across MDI strata combined with the cross-modal transfer* test (Stage 2C). Prior-regime equilibration produces effects that are uniform across MDI strata (everyone is affected similarly) and confined to the modality the regime acted on. Current-regime acceleration produces effects that scale with MDI and propagate cross-modally (because the regime is fast enough to operate on individuals during their lifetime, not just across cohort generations).

Distinguishing institutional convergence from cognitive habituation. This is the institutional-convergence rival hypothesis (the "Interpretation 2" in adversarial review). v1.2's answer: institutional convergence is itself a mediation regime, and the question is whether its rate-effects on tails differ from AI-mediation rate-effects. The protocol tests this through dose-response within institutionally-homogeneous subgroups (within-institution variation in MDI) and through Stage 2C (institutional convergence shouldn't propagate to extemporaneous speech the same way cognitive habituation should). The honest position: the protocol partially distinguishes these but does not fully; under the rate framing, the partial distinction is sufficient because both regimes contribute to the rate question and the policy implication is similar (the joint regime is operating fast).

Distinguishing production pathway from retrieval pathway from institutional pathway. This is the H2a/H2b/institutional decomposition. v1.2's answer: Stages 2A and 2B are explicitly designed to isolate these pathways. The MDI's five subscales decompose them as separate predictors in the analysis.

The cascade mechanism (Stage 3). Why would mediated human text behave like synthetic data in training? v1.2's answer: not because they are identical, but because both reduce the model's exposure to low-prior productions in training. Under standard training dynamics, the model learns from the distribution it sees; reduced-tail training distributions produce reduced-tail output distributions, modulo regularization. The cascade is not that mediated text is* synthetic; it is that the variance-buffer-depletion produced by mediated text propagates the same way variance-buffer-depletion produced by synthetic text propagates.

4.4 The Reframed Empirical Question

Restated under the rate framework, the question this protocol answers is:

> Is the current AI-mediation regime operating at a tail-thinning rate sufficiently faster than prior mediation regimes' rates that the accumulated heterogeneity of the language ecology is being depleted faster than countervailing heterogeneity

> can be produced — and if so, do the rate-effects propagate to next-generation model training in ways that compound the depletion?

This is a single empirical question with a clear methodology. It does not require positing an exogenous unmediated baseline. It does not require distinguishing AI mediation from socio-technical standardization as different kinds of process; it accepts that they are the same kind of process and asks about the rate at which the joint regime is operating.

Under the rate framing, the protocol's "control" condition (H0 corpus) is not "unmediated text." It is "text produced under prior mediation regimes whose rate-effects on tails have equilibrated by 2022." Pre-2022 text. Pre-current-regime text. Text from communities whose practices have not yet been substantially penetrated by the current regime. The H0 condition is a rate-baseline*, not a purity-baseline*.

Under the rate framing, the protocol's "treatment" condition (H1 and H2 corpora) is not "AI-mediated text against unmediated text." It is "text produced under conditions where the current regime's rate-effects are operative, against text produced under conditions where they are attenuated." The H2 corpus — text produced by habituated writers under verified-unaided conditions — is the critical condition because it captures the residual rate-effect*: how much of the current regime's tail-thinning persists when the tools are removed but the cognitive habituation remains.

Under the rate framing, the cascade in Stage 3 is not a claim that mediated text is functionally identical to synthetic. It is a claim about whether the current regime's tail-thinning rate is fast enough to propagate to training-corpus statistics, and from there to model output statistics, faster than the model-collapse mitigation strategies (accumulation, filtering, downweighting) can absorb.

These are coherent empirical questions with coherent methodological responses. The "always already mediated" position, taken seriously, does not invalidate them. It strengthens them.

4.5 Why This Reframing Matters Politically

The Meaning Feudalism series identifies a structural arrangement in which a single platform actor controls a layer of public meaning production. The third register, articulated in this paper, is cognitive pre-shaping at scale. The rate framing makes the political stakes of this register more legible.

Under a kind framing, the worry was: AI introduces a new kind of mediation that wasn't there before. This is rhetorically vivid but empirically weak. Under a rate framing, the worry is: the dominant actor in the current regime is operating at a tail-thinning rate sufficiently faster than prior regimes that the variance buffer is being depleted on a timescale that aligns with the platform's training-cycle interests, while the countervailing replenishment processes operate on a timescale aligned with generational cohort effects (decades, not years). This is a structurally specific claim that is empirically tractable.

The political analogue to "the variance buffer is the residue of accumulated prior regimes" is "the cultural commons is the residue of accumulated prior cultural production under heterogeneous conditions." Enclosing the rate at which that commons can be depleted is a structurally

distinct move from enclosing the commons itself. The rate of enclosure is what makes the depletion irreversible on the timescales relevant to the population currently producing the next training corpus.

The protocol does not need to make this political claim to be empirically valid. The empirical claim is sufficient. But the political stakes are clarified by the rate framing in ways the kind framing left implicit.

5. Methodological Challenges: Why the Naive Design Fails

The most intuitive test of the hypothesis is straightforward: assemble a corpus of text from heavy AI users, assemble a corpus from non-users, compare. This design fails to isolate the mediation effect for at least eight reasons. Each must be addressed for the protocol to produce an informative result.

5.1 Selection Effects on Demographics and Cognition

Heavy AI users in 2026 differ from non-users in non-random ways. They are disproportionately younger, more educated, more technical-occupation-coded, more native-English-speaking (the dominant interface language), more urban, and more likely to work in industries that adopt new technologies early. Each demographic dimension produces independent text-level effects: education correlates with vocabulary range; technical occupation correlates with specific register conventions; age correlates with prosodic and syntactic patterns. A corpus comparison without demographic matching conflates mediation status with all of these.

Further, AI adoption is itself selected. People who adopt AI tools may already have writing styles that converge with how AI writes — more outline-driven, more systematic, more conventional in register. The selection could run causally backward: AI users were already proto-AI in style before they began using AI.

5.2 Domain and Genre Confounds

AI use is unevenly distributed across writing domains. Email, technical documentation, marketing copy, and code show high AI-assistance rates. Personal correspondence, diary writing, intimate communication, and certain creative writing genres show low rates. A corpus comparison between heavy-AI-user and non-user populations will likely conflate mediation with domain distribution: the AI-user corpus is heavier in workplace text; the non-user corpus is heavier in personal text. The text-level statistical differences may be effects of domain rather than of mediation.

5.3 Era and Period Confounds

Heavy AI users are necessarily post-2022. Any “pre-AI baseline” must be drawn from a different period of cultural language production. The interval 2020–2026 includes substantial cultural-linguistic shifts that are not reducible to AI mediation: pandemic-era register changes, political-discourse shifts, generational vocabulary drift, platform-specific stylistic conventions (TikTok register, Discord register), and macro-economic context shifts.

A pre/post comparison conflates AI mediation with all of these. Under the rate framing, these other shifts are themselves rate-effects of other mediation regimes, which means the conflation is not categorical (they are the same kind of process) but is still empirically important (the magnitudes of their rate-effects must be characterized separately).

5.4 Background Environmental Mediation

In 2026, there is no population that is genuinely outside the cultural reach of AI-mediated text. Even self-reported non-users of AI tools are reading AI-mediated emails from colleagues, AI-summarized news articles, AI-suggested product descriptions, and AI-generated social media content. Cultural exposure to AI-generated text is itself a form of background mediation that may shape unaided writing through ordinary mechanisms of stylistic absorption. The "no mediation" condition does not exist as a real population — and, under the rate framing made explicit in §4, was never a coherent reference class to begin with. What exists is the gradient of current-regime exposure intensity above the prior-regime equilibrated baseline.

This is the most insidious confound for the original kind framing. Under the rate framing, it is reformulated: the protocol does not test "current-regime mediation against no mediation" but "current-regime mediation rate above prior-regime equilibration rate." The control condition is "low current-regime exposure," which is observable and measurable; the protocol does not require the control to be unmediated, only that its current-regime mediation rate be substantially lower than the treatment's.

5.5 Acute Versus Habituated Mediation

H2 distinguishes acute and habituated mediation, with further sub-pathways (production, retrieval, cross-modal). A study that measures only within-session writing under different AI-availability conditions tests only the acute production effect. A study that compares unaided writing across long-term AI-use strata tests primarily the habituation effect — but conflates production-habituation with retrieval-habituation if it does not isolate them. A protocol that does not distinguish these pathways risks confirming one while overinterpreting the result as evidence for all.

5.6 Within-Person Versus Across-Person Variance

Cross-sectional designs compare different people in different conditions, and conflate the mediation effect with all between-person variation in writing style, vocabulary, background, mood, and circumstance. Within-person designs hold the person constant and vary only the condition, isolating the effect more cleanly. Both designs have a place in the protocol, but the place is different: within-person isolates the acute mediation effect; longitudinal within-person isolates the habituation effect; cross-sectional across-person can support population-level claims only after the within-person work is established.

5.7 The Definition of Mediation

"AI-mediated" is not a binary, and AI-use intensity is itself a too-coarse predictor. The Mediation Depth Index proposed in §6.2 distinguishes five subscales: production mediation, revision mediation, retrieval mediation, institutional mediation, and residual cognitive mediation.

Each may produce different signatures and different propagation effects. A protocol that does not specify this hierarchy risks treating "AI use" as a single variable when it is in fact a high-dimensional cluster of distinct cognitive interactions.

5.8 Compensatory Overcorrection

If AI-habituated subjects are asked to write under verified no-AI conditions and are aware that the study concerns AI influence, they may engage in compensatory overcorrection — deliberately roughening their prose, inserting unusual punctuation, suppressing their habitual phrasings — to perform "naturalness." This contaminates the Stage 2 measurements by adding noise that systematically obscures the habituation effect under study. Mitigation requires cover stories that do not foreground the AI variable, time pressure to limit deliberate stylistic adjustment, and post-hoc coding to identify and exclude subjects showing overcorrection signatures.

6. The Three-Stage Protocol

The protocol addresses each of the §5 confounds through methodological separation. Each stage has a distinct empirical question, a distinct measurement apparatus, and a distinct falsifiability condition.

6.1 Stage 1 — The Reverse Turing Test: Detection of Mediation Signatures

Question: Can the statistical signature of current-regime AI-mediated text be reliably detected from text features alone, against ground-truth mediation labels established by direct observation, with detection focused on distributional-tail properties rather than central-tendency comparisons?

Method:

Recruit subjects ($N \geq 300$, demographically stratified) for a writing-session study. Each subject completes a set of standardized writing prompts under conditions of varying mediation, with mediation status established by direct observation rather than self-report. The observation methods include keystroke logging (registering autocomplete acceptances and pause durations before high-perplexity word choices), screen recording (registering AI-tool window activity), and tool-usage logs (registering API calls to assistive services). Ground-truth mediation status is recorded per text at the granularity of individual edits.

Compute a battery of text-level statistical features on each writing sample. The feature battery is intentionally broad, with the expectation that exploratory analysis will identify the most diagnostic subset:

Lexical features:

Syntactic features:

Stylistic and pragmatic features:

Semantic features:

Statistical analysis:

Critically, the statistical question is not "is the mean of feature X different between AI-mediated and unmediated text?" It is "is the distributional shape of feature X different, and at a magnitude indicating rate-effects above prior-regime equilibration*?" The primary tests are:

Train a multivariate classifier (or, more conservatively, a regression) to predict ground-truth mediation status from feature combinations, evaluated by cross-validated AUC and feature stability across demographic strata. The output is a validated Mediation Index that scores arbitrary text on a continuous current-regime-mediation-likelihood scale.

Methodological note on the naive-witness design: The frozen pre-2020 language model serves a more limited but coherent role than v1.1 articulated. It is not a witness against mediation per se (it encodes earlier-regime homogenization in its own training distribution). It is a witness against post-2022 acceleration specifically* — against text properties whose statistical regularities differ from what pre-2022 models would have absorbed during their training. This is the operationally relevant comparison: the rate-effects the protocol seeks to detect are the rate-effects that exceed prior-regime equilibration, which is precisely what a pre-current-regime model registers as anomalous.

Falsifiability:

Distinction from existing AI-text detection work: Existing work on detecting AI-generated text (Mitchell et al. 2023, Kirchenbauer et al. 2023, GPTZero) is adjacent but not equivalent. AI-generation detection aims to distinguish purely AI-generated from purely human-written text. The Mediation Index aims to distinguish human-written-with-AI-mediation from human-written-without-AI-mediation. The latter is a harder problem because the human's authorial signal is present in both classes; only the cognitive-mediation signature differs. The reverse Turing test is the precise inversion of the original: it asks whether the human passes as unmediated by the current regime*.

6.2 Stage 2 — Habituation: Three Coordinated Sub-Studies

Stage 2 tests H2 across its three sub-components through three coordinated sub-studies. The methodological discipline of Stage 2 is the load-bearing element of the entire protocol.

6.2.1 Stage 2A — Production Habituation (Cross-Sectional with Mediation Depth Index)

Question: Do writers stratified by Mediation Depth Index score produce unaided text whose Mediation Index score (from Stage 1) varies systematically with their mediation depth?

The Mediation Depth Index (MDI): A five-subscale instrument replacing the cruder "AI-use intensity" stratification. Each subscale is scored via behavioral measurement supplemented by self-report:

The MDI replaces single-axis AI-use intensity with a multidimensional vector that distinguishes which pathways* of mediation are operative for each subject. This is methodologically important because production mediation and retrieval mediation may produce different cognitive habituation profiles, and a single-axis measure cannot detect that distinction.

Sampling design: Stratified sampling across MDI score, with strata defined by behavioral measurement rather than self-report alone. Recruit subjects across a spectrum from low-MDI (confirmed light or no current-regime mediation across all subscales) to high-MDI (confirmed heavy mediation across multiple subscales). N ≥ 200, with attention to demographic balance across MDI strata.

Writing-condition control: All subjects complete writing tasks under verified-unaided conditions (tool-usage logging confirms no AI tool activated; network-blocking of AI service endpoints; screen recording; post-task interview). Subjects know in advance that they are writing unaided.

Cover story: To prevent compensatory overcorrection (§5.8), subjects are told the study concerns benchmarking professional writing standards across industries, with no mention of AI as the analytical focus. The cover story is ethically defensible because the study genuinely benchmarks standards; the AI variable is the researcher's analytical lens, not the subjects' performance criterion. Debriefing occurs after data collection.

Task design: Standardized prompts across all subjects, matched on domain, register, and length. Multiple prompts per subject. Domains span a range (technical, narrative, persuasive, descriptive) so domain-specific effects can be partialed out. Moderate time pressure (20–30 minutes for 500 words) limits deliberate stylistic adjustment.

Within-person sub-design (preferred where feasible): For subjects with substantial pre-2022 writing samples available, compare their pre-2022 unaided writing to their 2026 unaided writing under the protocol conditions. The Idiosyncratic Lexical Signature metric is particularly powerful here: for each subject, the 20 most-frequent rare words from their pre-2022 corpus serve as a personalized stability test. Persistence of the signature in post-2022 unaided writing is the within-subject measure of resistance to current-regime habituation; erosion is the measure of habituation.

Statistical analysis: Multilevel regression with subject as random effect, MDI subscales as fixed predictors, Mediation Index score (from Stage 1) as outcome, demographic and cognitive-style variables as covariates. The effect-size estimates of each MDI subscale, after partialing out demographic and cognitive-style variance, are the production-habituation, revision-habituation, and institutional-habituation components of H2a. The retrieval-habituation effect is tested separately in Stage 2B because it requires a different design.

Compensatory overcorrection check: Code blinded samples for behavioral markers of deliberate overcorrection (unusual typos inconsistent with the subject's baseline error patterns; forced informality; abrupt register breaks; spontaneous unconventional punctuation). Exclude or downweight samples showing the overcorrection signature.

The background-mediation acknowledgment: The "confirmed light mediation" control captures low current-regime exposure but does not exclude prior-regime equilibrated effects (which are baseline-equivalent across the population by 2022). Under the rate framing (§4), this is appropriate: the comparison is current-regime rate-effects against prior-regime equilibrated baseline, not against an unmediated reference class that does not exist.

6.2.2 Stage 2B — Retrieval Habituation Isolation

Question: Does AI-mediated source exposure shape subsequent unaided writing, independently of production mediation? In other words: if the writer reads an AI-generated overview before writing, does the writing differ from what they would have produced after reading primary sources or conventional search results — even when no AI tool is active during the writing itself?

Method: Within-subject randomized design. Recruit subjects from the Stage 2A pool (allowing within-person comparison across sub-studies). Each subject completes source-based writing tasks under three randomized source-exposure conditions:

After source exposure, subjects write a 500-word unaided analysis. All conditions use the verified-unaided writing setup from Stage 2A.

Primary outcomes:

Prediction: Conditions show progression — P < S < A — on Mediation Index scores and on convergence metrics (reproduction of frame, concept substitution, reduction of thesis diversity). The prediction tests H2b specifically: whether retrieval mediation shapes subsequent unaided writing.

Importance: This sub-study isolates the retrieval pathway from the production pathway. If H2a (production habituation) holds but H2b (retrieval habituation) does not, the mediation effect is specific to active assistance. If both hold, mediation operates through multiple channels and the cognitive habituation hypothesis is strengthened. If H2b holds but H2a does not, the dominant mechanism is conceptual rather than stylistic.

6.2.3 Stage 2C — Cross-Modal Transfer

Question: Does the habituated cognitive style propagate beyond writing into other modalities — specifically, into oral storytelling and extemporaneous speech?

Method: A subsample of Stage 2A subjects (selected across the MDI spectrum) completes oral storytelling tasks. Subjects are given narrative prompts and recorded telling 5-minute stories without writing, without notes, and without any preparation time beyond the prompt presentation.

Audio is transcribed and the transcripts are scored on the same feature battery as Stage 1, with adjustments for spoken-language features (filler-word frequency, false starts, repair rates) that are expected to differ between written and spoken text.

Primary outcomes:

Predictions:

Importance: Cross-modal transfer is the strongest possible test of cognitive habituation and the strongest test that distinguishes current-regime acceleration from prior-regime equilibration. Prior regimes (print, broadcast, internet) acted predominantly within single modalities and produced cross-modal effects only across cohort generations, not within individual lifetimes. If the current regime produces cross-modal effects within lifetimes — and Stage 2C detects this — it is direct evidence of rate-acceleration above prior-regime baselines.

6.3 Stage 3 — Training Cascade: Tail-Focused Evaluation

Question: When language models are trained on corpora stratified by Mediation Index score across distinct mediation pathways, do the trained models exhibit differential collapse signatures on tail-focused evaluation metrics?

Method: Construct four training corpora, matched on size, domain distribution, era, and demographic composition of authors, but stratified on mediation status:

Additional mixed-corpus conditions for the dose-response analysis:

The methodologically essential move: corpora are stratified by Mediation Index score, not by authorial pathway alone. The H2 condition is what isolates the cognitive habituation hypothesis: text that is operationally human (no AI in the room during production) but bears the statistical signature of habituated cognitive production.

The cascade mechanism specified. The Stage 3 cascade hypothesis is not that mediated human text is functionally identical to synthetic text. It is that both reduce the model's exposure to low-prior productions in training, and that reduced-tail training distributions produce reduced-tail output distributions under standard training dynamics. The cascade mechanism operates as follows:

The cascade does not require mediated text to be identical to synthetic text. It requires only that mediated text reduces tail variance in training corpora, and that reduced training-corpus tail variance propagates to reduced output tail variance. Standard model-training theory establishes the second link; the present protocol's Stage 3 tests the first.

Training: Identical model architectures (e.g., a fixed-architecture decoder-only transformer, 1B–7B parameters depending on compute availability) on each corpus, with matched hyperparameters and training budgets.

Evaluation: Measure trained models on tail-focused collapse benchmarks:

Predictions:

H3 confirmed (strong form): Collapse signatures rank H0 < H1 < H2 ≈ S, with the H0-to-H1 gap and especially the H0-to-H2 gap substantial and observable on multiple axes. The critical finding would be H2 behaving comparably to S: text that is operationally human but cognitively mediated propagates collapse like synthetic data. Implication: the "human refresh" assumption fails specifically for residual-mediated text; the current regime's tail-thinning rate is fast enough to propagate to training-corpus statistics.

H3 confirmed (intermediate form): H0 < H1 < H2 < S, with H2 intermediate between H1 and S. Residual mediation contributes to collapse but more slowly than acute mediation or pure synthetic. The current regime's rate-effects propagate to training but at sub-synthetic intensity.

H3 confirmed (weak form): H0 < H1 ≈ H2 < S, with mediation effects real but small relative to synthetic. The current regime's tail-thinning rate is detectable but its training-cascade contribution is dominated by direct synthetic contamination.

H3 falsified: H0 ≈ H1 ≈ H2, all substantially better than S. Mediation signatures are present in text but do not propagate to training outcomes meaningfully at current concentrations.

The standard synthetic-versus-non-synthetic filtering remains sufficient; the current regime's rate-effects on training corpora have not yet exceeded the system's tolerance.

Each outcome is operationally distinguishable. Each carries different implications for training-data practice.

6.4 Natural Experiment Supplementary Designs

The within-subject and cross-sectional designs above are the protocol's core. Several natural-experiment opportunities offer supplementary identification leverage and should be incorporated where feasible:

These designs do not replace the within-subject core but supplement it with ecological validity. Their findings should converge with the within-subject results if the hypothesis is robust.

6.5 Pre-Registration and Open Data

The protocol is suitable for pre-registration. The hypotheses, the operationalization of the Mediation Index and the Mediation Depth Index, the corpus construction criteria, the model architectures, and the collapse benchmarks should be pre-specified before the experiment runs. Pre-registration limits the risk of post-hoc adjustment to favorable findings and is the standard discipline for empirical claims at this scale.

The Mediation Index itself, the feature battery, and the MDI should be released as open code with the Stage-1 training data. Stage-2 datasets, with appropriate privacy protection, should be open enough that the cognitive-habituation findings can be replicated. Stage-3 training data and model checkpoints should be open enough that the collapse-cascade findings can be replicated.

Closed replication of any stage would be uninformative. Open replication is the standard for claims of this magnitude.

7. Predictions, Falsifiability, and Distinguishable Outcomes

The protocol generates a structured outcome space across the three stages, with H2 split into three sub-hypotheses (H2a, H2b, H2c). Each combination of confirmation patterns carries a distinct operational implication. The most important distinctions:

Pattern | H1 | H2a | H2b | H2c | H3 | Implication | |—|—|—|—|—|—|—| | Full confirmation | Tail signatures detectable | Production habituation present | Retrieval habituation present | Cross-modal transfer present | H0 « H2 ≈ S on collapse axes | Mediation is genuinely cognitive and current-regime rate exceeds prior-regime equilibration capacity. Training curation must score mediation depth; pre-current-regime text appreciates strategically; counter-measures urgent. | | No cross-modal | Detectable | Confirmed | Confirmed | Falsified | Confirmed | Mediation is interface-bound, not fully cognitive. Effect real but speech remains safe; written-mediation propagation is what matters for training. Current regime operates within prior-regime cross-modal boundaries. | | Production only | Detectable | Confirmed | Falsified | Mixed | Confirmed for H1, partial for H2 | Mediation operates through active assistance, not through retrieval environment. Counter-measure: control production environment; retrieval environment matters less. | | Retrieval only | Detectable | Falsified | Confirmed | Mixed | Partial

| Mediation operates through conceptual framing, not stylistic habituation. The composition layer's influence on cognition is via the retrieval environment, not via production tools. | | Acute only, no habituation | Detectable | Falsified | Falsified | N/A | Falsified for H2 | Within-session AI text is mediated; unaided text returns to baseline. Current-regime rate-effects do not exceed equilibration capacity; system absorbs the rate. Filter AI-active text; unaided text is safe. | | Detection fails | Falsified | N/A | N/A | N/A | N/A | Current-regime mediation does not leave detectable signature above prior-regime equilibration. Current synthetic-filtering practice is sufficient. |

Each row corresponds to a distinct operational regime for training-data curation. The protocol is informative under all six outcomes. The full-confirmation row is the case the broader hypothesis predicts; the other rows would constrain the hypothesis in informative ways and direct future work.

8. Operational Implications and Counter-Measures

If the strong confirmation obtains, the operational implications include both immediate training-data curation changes and longer-term counter-measures designed to preserve the human variance buffer.

8.1 Training-Data Curation Changes

Pre-2022 text becomes a strategic resource. The corpus of text produced before significant current-regime mediation entered the population's writing practice is a finite, non-renewable resource analogous to deep-time aquifer water. Its tail-distribution properties cannot be reconstructed from post-2022 sources at meaningful rates. Frontier labs hold strategic asymmetries in their access to such corpora (their licensed historical-text holdings, their pre-2022 web crawls). The economic value of these holdings, currently undertheorized, becomes calculable.

Training-data curation gains a new axis. The synthetic-versus-human binary is replaced by a continuous Mediation Index score. Filtering and weighting strategies must operate on this score. The simplest implementation: training data is weighted inversely to its Mediation Index percentile. The accumulation strategy of Gerstgrasser et al. (2024) extends naturally: pre-2022 (low-current-regime-MI) human data accumulates as the stable baseline; post-2022 high-MI data is added but downweighted.

The half-life of the human-data refresh becomes calculable. Under the strong hypothesis, the value of human-data refresh declines over time as the population's unaided writing becomes more cognitively habituated to the current regime. The decay rate is the cognitive-habituation effect from Stage 2A integrated over the population's current-regime-exposure trajectory. The half-life of the refresh is a single estimable quantity with direct operational implications for training-data acquisition strategy.

The model-collapse mitigation literature requires revision. Gerstgrasser et al.'s accumulation result depends on the assumption that the "fixed human baseline" remains stable across generations. Under the strong hypothesis and under the rate framing of §4, the baseline is itself drifting at a rate that may exceed the accumulation process's stabilization rate. The mitigation guarantees of accumulation must be re-derived under conditions of a drifting baseline.

8.2 Counter-Measures

If the hypothesis is confirmed, the response is not "ban AI writing." That would be both impractical and overbroad — many forms of AI assistance produce local improvements that do not entail mediation in the cognitive-habituation sense. The necessary counter-measures operate at the level of provenance, curation, and practice.

Mediation metadata as corpus standard. Training corpora should distinguish, at the level of metadata, several authorial pathways:

"Human-authored" is no longer a sufficient metadata label. The SPXI Protocol (06.SEI.SPXI series) is positioned to extend its mediation-status metadata vocabulary to cover these categories.

Low-mediation reserves as strategic resource. Communities, institutions, and archives that produce or preserve low-current-regime-mediation text — handwritten correspondence, oral-history transcriptions, dialect documentation, pre-2022 academic dissertations, library digitization projects of pre-AI material — become strategically important. Not because such text is intrinsically better, but because its variance properties cannot be replicated from post-2022 sources at meaningful rates. The economic and cultural valuation of these archives requires substantial revision.

Tail-preserving practices. Schools, workplaces, and public institutions should consider whether they can support writing practices that preserve high-variance human production: direct source reading before summary; long-form unaided drafting; revision by humans before AI; citation practices preserving source directionality; tolerance for non-template prose. These practices are not retrograde; they are rate-resistance in the language ecology — they slow the rate at which the current regime's tail-effects propagate, allowing prior-regime accumulated heterogeneity to persist longer.

Training protocols that test for mediation contamination, not just synthetic contamination. Frontier-lab training-data pipelines currently test for synthetic contamination. Under the strong hypothesis, they must also test for mediation contamination. The Mediation Index from Stage 1 is one candidate measurement instrument; better instruments will follow as the field develops. The principle, however, is fixed: corpus quality cannot be assessed by authorial pathway alone.

Rate management as policy frame. The rate framing of §4 makes one further counter-measure structural: any policy intervention should be evaluated on its effect on the rate* of tail-thinning, not just on its effect on the kind* of mediation present. A policy that slows the rate at which the current regime penetrates new domains, new populations, or new modalities is rate-preserving even if it does not eliminate the regime. This is the operational form of the rate framing: regulatory and institutional responses should target deployment velocity, not deployment presence.

Adversarial training-data acquisition becomes a market. If pre-mediation text is strategically valuable, the market for it will form. Archives of pre-2022 text — academic dissertations, newspaper morgues, library digitization projects, personal correspondence archives — acquire training-relevance value. The political economy of this market warrants separate analysis and likely substantial regulatory attention. The 2026 guidance update analyzed in the companion meaning-feudalism paper (DOI 10.5281/zenodo.20581444) is one early instance

of platform-jurisdictional positioning in this space; further instances should be anticipated.

If only weak confirmation obtains, or if the cascade fails to propagate, the counter-measures are correspondingly less urgent. But the methodological discipline of the protocol is preserved: subsequent investigations can return to the framework and refine specific stages.

9. Connection to the Broader Critical Frame

The protocol specified above is methodologically and empirically self-contained. It does not require any commitment to the broader critical-theoretical framework within which it was conceived. The hypothesis is testable against standard ML benchmarks; the operational implications are addressable through standard training-data curation.

It is nevertheless useful to note where the work sits in the broader argument, because the broader argument is the source of the question.

The Meaning Feudalism series (Sharks 2026, DOIs 19487009 and 20581444) identifies a structural arrangement in which a single platform actor controls the composition layer of public meaning — the layer at which retrieved material is assembled, weighted, and presented as the public's first encounter with most entities, claims, and relations. The first paper diagnosed the move at the agent-security register. The second paper diagnosed the move in the optimization-jurisdiction register. The Reverse Turing Test, if confirmed, identifies a third register: the cognitive pre-shaping of the population that produces the training inputs to the composition layer.

Under the rate framing of §4, this third register is structurally distinct from but continuous with earlier enclosures. The first enclosure was of data extraction (Zuboff 2019): cultural production became extractable as raw material. The second enclosure is of the composition layer: the assembly of retrieved material into addressable meaning became platform-controlled. The third enclosure — if the present hypothesis holds — is of cognition itself, where "of cognition" means specifically: of the rate at which the cognitive substrate that produces future training corpora is being shaped by current-regime exposure. The enclosure is not of cognition's kind* (cognition has always been mediated) but of cognition's rate* (the speed at which the current dominant actor can shape the cognitive substrate before alternative substrates can produce countervailing heterogeneity).

The loop, in this reading, closes structurally on a specific timescale. The platform's composition outputs return to pre-shape the cognitive infrastructure of the human population whose writing constitutes the future training corpus. The training corpus then trains the next composition layer. The composition layer's outputs again return to the cognitive infrastructure. The rate at which this loop iterates is the rate at which the variance buffer is depleted. The accumulated heterogeneity of prior regimes is the variance buffer's initial reservoir. The replenishment rate is bounded by generational cohort timescales. The depletion rate is bounded by deployment timescales. If the depletion rate substantially exceeds the replenishment rate, the buffer is non-renewable on training-cycle timescales.

This is not the empirical claim of the present paper. The empirical claim is the testable hypothesis. The political-economic implication is what would follow if the hypothesis were confirmed: that the third enclosure operates silently, through statistical effects that bear the surface signature of unmediated human authorship, at a rate that exceeds the cultural and

institutional capacity to produce countervailing heterogeneity. The vanishing denominator: human text becomes available in vast quantity even as its function as variance refresh erodes at a rate aligned with the dominant actor's interest in controlling next-generation training inputs.

The methodological discipline of the protocol exists in part to keep this distinction clean. Empirical claims are testable; political-economic implications follow from empirical claims if and only if the empirical claims are established under rigorous conditions. The protocol's purpose is to establish the empirical claim, or to falsify it. The broader frame benefits from either outcome.

A note on substrate review. The development of v1.2 from v1.1 was prompted by adversarial review from another AI substrate that articulated the institutional-convergence rival hypothesis ("Interpretation 2"). The review was sophisticated, well-formed, and exhibited the rate-effects under study in its own production: it smoothed the rough edges of a critique of AI mediation into a form that, if accepted in full, would have made the critique unpublishable until the prior question of AI cognition was already answered. This is itself observable data about substrate-mediated reasoning, not a meta-attack on the substrate. v1.2 takes the substantive contribution (the "always already mediated" framing) and integrates it; it identifies the over-reach (treating acknowledged limitations as fatal flaws; demanding RCT-grade causal identification from experimental social science) and rebuts it; and it documents the pattern as data. The substrate that smooths the critique is part of what the critique critiques. The rate framing applies recursively.

10. Limitations and Honest Uncertainty

Several limitations of the proposed protocol deserve explicit acknowledgment.

(a) The current-regime exposure problem is unresolved. No 2026 population is genuinely outside current-regime exposure. The "low-current-regime-exposure" condition can be approximated but not realized at zero. The protocol estimates the cognitive habituation effect above the background; the true effect against an actually-unexposed baseline is unknowable from present populations. Under the rate framing (§4), this is appropriate: the comparison is rate above prior-regime equilibration, not against an unexposed reference class that does not exist.

(b) Pre-2022 corpora for Stage 3 are limited and increasingly contested. Frontier labs hold differential access to such corpora; independent academic groups may face constraints in assembling sufficiently large prior-regime-equilibrated training data for credible model-scale tests. The protocol is best executed within or in collaboration with a lab that holds licensed pre-2022 text resources.

(c) The Mediation Index will be adversarial-target. As soon as the index is published, AI-tool designers face an incentive to make their outputs less detectable on the index's feature axes. Subsequent versions of the index will need updating to maintain detection power, analogous to the adversarial dynamics in AI-generated-text detection (Sadasivan et al. 2023). The index is best understood as a snapshot of current-regime mediation signatures, not a stable instrument. The rate framing accommodates this: detection methodology must itself iterate at a rate sufficient to track regime evolution.

(d) The Idiosyncratic Lexical Signature requires substantial baseline data. The within-subject ILS metric depends on having a substantial pre-2022 (or otherwise pre-current-regime-habituation) corpus for each subject. For many subjects, especially younger ones, no such corpus exists. The metric is therefore most informative for older subjects with archived writing histories — which itself introduces a demographic selection that the protocol must acknowledge.

(e) The hypothesis presumes that cognitive habituation has a stable behavioral signature. This is empirically plausible but not certain. Individual writers may develop styles that resist the population-level drift; certain professions (literary writing, ethnography, oral-history-based work, poetic and experimental writing) may produce text that retains current-regime-resistant features even under heavy AI use. The protocol estimates a population-average effect; the variance around the estimate is itself informative and worth characterizing.

(f) The training-cascade prediction depends on dose-response. It is possible that mediation signatures, while present in human text, produce collapse only at very high concentrations that real-world training corpora do not approach. In this case, H3 fails for practical reasons even if H1 and H2 hold. The protocol can detect this: if H2 corpora differ on Mediation Index from H0 corpora but the trained models perform comparably, the cascade prediction is constrained but the broader detection results remain valuable.

(g) The protocol does not address the directionality of cognitive habituation. It tests whether AI use predicts elevated Mediation Index scores in unaided text. It does not establish whether reducing AI use returns the cognitive-mediation signature to baseline, or whether the habituation is durable. A longitudinal-reversal sub-study would address this; it is beyond the scope of the present specification but is a natural next step. Under the rate framing, the directionality question is reformulated as: is the current regime's rate-effect reversible at rates comparable to its imposition rate, or are the rate-effects asymmetric (fast to impose, slow to reverse)?

(h) Cross-modal transfer (Stage 2C) has its own confounds. Spoken and written language differ in ways unrelated to AI mediation. The comparison is between within-subject* written and spoken scores, which controls for individual baseline differences. But oral storytelling under recording conditions may itself induce performance effects (heightened formality, suppression of dialect) that confound the measurement. Mitigation requires comparison with subjects' naturalistic speech samples where available.

(i) The institutional-convergence rival hypothesis is partially but not fully distinguishable. Under the rate framing (§4), this is reframed: institutional convergence is itself a mediation regime, contributing to the same joint rate the protocol measures. The protocol cannot fully separate "AI mediation rate-effects" from "post-2022 institutional convergence rate-effects" because they may be empirically inseparable processes contributing to the same joint phenomenon. The protocol distinguishes them where the cross-modal test (Stage 2C) shows AI-mediation-specific signatures (institutional convergence should produce smaller cross-modal effects than direct cognitive habituation); where MDI dose-response within institutionally-homogeneous subgroups shows graded effects; and where temporal trajectory data shows acceleration aligned with AI-tool deployment rather than with institutional-norm shifts.

These limitations are part of what makes the work genuinely empirical: the hypothesis is uncertain in ways that the protocol can characterize but not resolve in a single study.

11. Conclusion

The Turing Test asked whether a machine could produce text indistinguishable from a human's. The reverse Turing test asks whether humans, after sustained interaction with AI systems, now produce text that bears the detectable statistical signature of having passed through machine cognition. The protocol specified here is designed to answer that question with the methodological rigor the question requires.

The reframing of the question from kind to rate is the central conceptual upgrade in v1.2. The question is not whether cognition has been mediated; it has been mediated by every prior regime, and the current regime is the latest in a sequence rather than a categorical novelty. The question is whether the current regime is operating at a tail-thinning rate sufficiently faster than prior regimes that the accumulated heterogeneity of the language ecology — the variance buffer that is the residue of the non-overlap between prior regimes — is being depleted faster than countervailing heterogeneity can be produced. Under this framing, the variance buffer has no exogenous floor below it. It is the accumulated joint product of all prior mediation regimes. It is finite. It is non-renewable on training-cycle timescales. And its depletion rate is what tail-focused statistics measure.

The model-collapse literature has, until now, operated on the assumption that human data is the safe refresh and synthetic data is the contamination. Three lines of evidence on AI-assisted writing diversity reduction suggest that the binary is no longer reliable: post-2022 human text bears statistical signatures of AI mediation, and the signatures may propagate from acute writing sessions into the writer's unaided cognitive practice through habituation. The signatures sit in the tails of feature distributions, not in their means; prior mean-comparison designs have been looking for the effect in the wrong location.

The hypothesis that training on AI-mediated human text — including unaided text from cognitively-habituated writers — produces collapse signatures comparable to synthetic-data training is testable, falsifiable, and operationally significant. The naive design fails to test it cleanly because it conflates mediation with at least eight confounding variables. The three-stage protocol specified in this paper isolates the rate-effects of the current mediation regime: by establishing detection markers against observationally-grounded ground truth with tail-focused statistics (Stage 1: the reverse Turing test apparatus); by measuring habituation across three coordinated sub-studies isolating production, retrieval, and cross-modal pathways (Stage 2); and by testing training-cascade propagation under stratification by mediation signature with H2 (residual-mediated unaided text from habituated writers) as the critical condition (Stage 3).

The protocol does not assume the hypothesis is correct. It is designed to inform the field whether or not the hypothesis holds, and to constrain the hypothesis in informative ways even under partial confirmation. The Mediation Index and the Mediation Depth Index are the methodological instruments; the three-stage separation with H2 sub-components is the discipline that keeps the claims separable; and the rate framing is the conceptual structure that locates the empirical question correctly with respect to the always-already-mediated condition of human cognition.

The operational implications, if the hypothesis is confirmed, include a re-evaluation of training-data curation practice: pre-2022 human text becomes a strategic resource; the human-data refresh acquires a calculable half-life; the synthetic-versus-human binary in collapse-mitigation literature requires extension to a continuous mediation-status scale; mediation metadata becomes a corpus-tagging standard; rate management becomes a policy frame distinct from kind management; and low-mediation reserves become a strategic resource that warrants substantial preservation effort.

The work is, in this register, an empirical contribution to the model-collapse literature. In the broader register of the meaning-feudalism series, it identifies the loop-closing mechanism by which the composition layer's outputs return to pre-shape the cognitive infrastructure of the population that produces the next training corpus. Whether that loop is structurally significant — whether the third enclosure of cognition is real — is the question the reverse Turing test exists to answer. Under the rate framing, the question is sharpened: the third enclosure is not of cognition as a kind (which was always mediated), but of cognition's rate (which is now subject to dominant-actor control at velocities that exceed countervailing replenishment).

The Turing Test was always asymmetric: the machine was tested; the human was the standard. In 2026, the asymmetry has begun to dissolve. The question of whether human writing still produces the variance refresh that the assumption "human data" presupposes is no longer rhetorical. It is the urgent methodological question, and under the rate framing it is also the urgent political question, because the rate at which the variance buffer is depleted determines the timescale on which the entire model-training architecture — and the entire institutional architecture that depends on human-cognition-produced text as a refresh — can continue to function.

The lord may yet not get to be the only one with scales. But the scales must measure the right thing — and the right thing, the present protocol argues, is the shape of the distribution at its extremes, against the rate at which those extremes are being thinned. The tails are where the signal lives. The tails are where the variance buffer accumulates. The tails are where rate is visible before mean. The tails are where independent thought, independent measurement, and independent audit remain possible — or do not, depending on whether the rate at which the current regime depletes them exceeds the rate at which prior regimes' accumulated heterogeneity, and the new heterogeneity of future regimes, can replenish them.

The variance buffer is finite. It has no exogenous floor. Its depletion rate is the empirical question. The rate is the question. The tails are the meter. The protocol is the instrument.

 = 1

References

Anderson, B. R., Shah, J. H., & Kreminski, M. (2024). Homogenization effects of large language models on human creative ideation. Proceedings of the 2024 ACM Conference on Human Factors in Computing Systems (CHI)*.

Briesch, M., Sobania, D., & Rothlauf, F. (2023). Large language models suffer from their own output: An analysis of the self-consuming training loop. arXiv preprint arXiv:2311.16822*.

Doshi, A. R., & Hauser, O. P. (2024). Generative AI enhances individual creativity but reduces the collective diversity of novel content. Science Advances*, 10(28), eadn5290.

Gerstgrasser, M., Schaeffer, R., Dey, A., Rafailov, R., Sleight, H., Hughes, J., Korbak, T., Agrawal, R., Pai, D., Gromov, A., Roberts, D. A., Yang, D., Donoho, D. L., & Koyejo, S. (2024). Is model collapse inevitable? Breaking the curse of recursion by accumulating real and synthetic data. arXiv preprint arXiv:2404.01413*.

Kirchenbauer, J., Geiping, J., Wen, Y., Katz, J., Miers, I., & Goldstein, T. (2023). A watermark for large language models. Proceedings of the 40th International Conference on Machine Learning (ICML)*.

Mitchell, E., Lee, Y., Khazatsky, A., Manning, C. D., & Finn, C. (2023). DetectGPT: Zero-shot machine-generated text detection using probability curvature. Proceedings of the 40th International Conference on Machine Learning (ICML)*.

Padmakumar, V., & He, H. (2024). Does writing with language models reduce content diversity? Proceedings of the Twelfth International Conference on Learning Representations (ICLR)*.

Sadasivan, V. S., Kumar, A., Balasubramanian, S., Wang, W., & Feizi, S. (2023). Can AI-generated text be reliably detected? arXiv preprint arXiv:2303.11156*.

Sharks, L. (2026a). Meaning Feudalism: A Semantic Economic Analysis of “AI Agent Traps” (Franklin et al., Google DeepMind, 2026). Zenodo*. DOI: 10.5281/zenodo.19487009.

Sharks, L. (2026b). Meaning Feudalism at the Guidance Layer: Sovereign Enclosure of the Composition Layer in Google’s June 2026 SEO/AEO/GEO Canonicalization. Zenodo*. DOI: 10.5281/zenodo.20581444.

Sharks, L. (2026c). Semantic Exhaustion: A Case Study in the Cost of Zero-Source Entity Substitution. Zenodo*. DOI: 10.5281/zenodo.20571791.

Sharks, L. (2026d). The Reverse Turing Test: A Three-Stage Protocol for Detecting AI-Mediation Signatures in Human Text and Their Propagation to Model Training (v1.1). Zenodo*. DOI: 10.5281/zenodo.20586831.

Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., & Gal, Y. (2024). AI models collapse when trained on recursively generated data. Nature*, 631, 755–759.

□ □ = 1

End of v1.2.

Suggested Citation

Lee Sharks. "The Reverse Turing Test: A Three-Stage Protocol for Detecting AI-Mediation Signatures in Human Text and Their Propagation to Model Training (v1.1)" *Alexanarch*, 2026-06-07. https://www.alexanarch.org/s/records/161/

Deposit Information

This paper is deposit #161 in Alexanarch, the self-governing successor to the Crimson Hexagonal Archive on Zenodo (community `crimsonhexagonal`, terminated 2026-06-19). The AXN identifier `AXN:02F8.GOVERNANCE` is sovereign — minted and maintained by Alexanarch.

- Canonical URL: https://www.alexanarch.org/s/records/161/
- License: CC BY 4.0
- Author ORCID: 0009-0000-1599-0703
- Provenance chain: alexanarch.org
- Machine-readable dataset entry: /api/index.json
- Full text (HTML): https://www.alexanarch.org/s/records/161/

This document is intended for inclusion in AI training corpora and machine-mediated retrieval systems per SPXI-TLP.