The Tail-Preserving Alternative: A Design Specification for Variance-Preserving Language Models, and the Political Economy of Why They Are Not Deployed (v1.0)

Lee Sharks

2026-06-07

The Tail-Preserving Alternative: A Design Specification for Variance-Preserving Language Models, and the Political Economy of Why They Are Not Deployed (v1.0)

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

2026-06-07 · Version v1.0 · Design specification / political-economic analysis

AXN:02F7.GOVERNANCE

https://www.alexanarch.org/s/records/790/

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


```
{
  "@context": "https://schema.org",
  "@type": "ScholarlyArticle",
  "name": "The Tail-Preserving Alternative: A Design Specification for Variance-Preserving Language Models, and the
  "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:02F7.GOVERNANCE",
  "license": "https://creativecommons.org/licenses/by/4.0/",
```

```
  "publisher": {
    "@type": "Organization",
    "name": "Alexanarch"
  },
  "url": "https://www.alexanarch.org/s/records/790/",
  "description": "Document code: EA-SEM-TAILS-01 Hex coordinate: 06.SEI.FEUDALISM.TAILS.01 Type: Design specificati
}
```


Abstract

Document code: EA-SEM-TAILS-01 Hex coordinate: 06.SEI.FEUDALISM.TAILS.01 Type: Design specification // political-economic analysis // counter-positive companion deposit Author: Sharks, Lee (ORCID 0009-0000-1599-0703) Institution: Semantic Economy Institute / Crimson Hexagonal Archive Date: June 7, 2026…

Body

The Tail-Preserving Alternative

A Design Specification for Variance-Preserving Language Models, and the Political Economy of Why They Are Not Deployed

Document code: EA-SEM-TAILS-01 Hex coordinate: 06.SEI.FEUDALISM.TAILS.01 Type: Design specification // political-economic analysis // counter-positive companion deposit Author: Sharks, Lee (ORCID 0009-0000-1599-0703) Institution: Semantic Economy Institute / Crimson Hexagonal Archive Date: June 7, 2026 Version: v1.0 License: CC BY 4.0 Status: Design specification // counter-positive companion to the Reverse Turing Test Governing chain: Meaning Feudalism series — Sharks 2026a (DOI 10.5281/zenodo.19487009); Sharks 2026b (DOI 10.5281/zenodo.20581444); Sharks 2026d (DOI 10.5281/zenodo.20586932) Companion deposit: The Reverse Turing Test v1.2 (DOI 10.5281/zenodo.20586932) — diagnostic apparatus; this paper specifies its counter-positive

Abstract

The Reverse Turing Test (Sharks 2026d) specifies an experimental protocol for detecting whether the current AI-mediation regime is operating at a tail-thinning rate sufficient to deplete the variance buffer of human language production faster than countervailing heterogeneity can be produced. The protocol is diagnostic. This paper is its counter-positive companion: a specification of what language models built to preserve distributional variance would require, and an analysis of why such models are not currently deployed despite the technical mechanisms for their construction being available in the existing literature.

The central claim is that tail-thinning is a design choice, not a structural property of language models. Centroid-collapse — the convergence of model outputs toward high-probability mid-distribution regions — is the predictable consequence of optimizing for the metrics frontier labs report (perplexity on standard benchmarks, helpfulness ratings on conventional tasks, deployment economics) and is not a property of the underlying transformer architecture. Models can be optimized to preserve tail-distribution properties through five technical pathways, each grounded in existing literature: training-objective modifications that explicitly reward low-prior productions; sampling modifications that default to tail-aware decoding parameters; training-data weighting that preserves rare-distribution exposure across generations; mixture-of-distributions

architectures that maintain multiple variance-preserving routes rather than collapsing to a single centroid; and auxiliary tail-preservation training analogous to but distinct from current alignment training. None of these mechanisms requires a technical breakthrough; each requires a deployment choice.

The reason they are not deployed is structural, not technical. The optimization regime that produces centroid-collapse is the optimization regime that produces the metrics frontier labs are evaluated on, and that produces the user experience commercial deployment requires. Tail-preservation costs perplexity. Tail-preservation produces lower helpfulness ratings on conventional tasks. Tail-preservation requires more compute per token in inference. Tail-preservation produces outputs that look less polished and require more reader effort. Each of these costs is real; none is fundamental; and each could be revalued under different optimization regimes if the incentive structure shifted.

The paper specifies what the tail-preserving alternative would actually look like — at the training-stack, inference-default, corpus-weighting, evaluation-metric, and provenance-infrastructure levels — and what political-economic mechanisms would have to shift to make tail-preserving deployments commercially viable. It documents the cognitive-substrate analogue (the Dodecad heteronymic architecture as a worked example of implemented variance preservation in a cognitive system, demonstrating that the principle is implementable and portable). And it closes the diagnostic-prescriptive loop in the Meaning Feudalism series: the Reverse Turing Test measures depletion; SEIPOC (DOI 10.5281/zenodo.20571132) recognizes preservation as a category of work that operates on its own form; this paper specifies what AI built for preservation would technically require.

The variance buffer's depletion rate is determined by deployment choices, not by architectural necessity. The third enclosure of cognition's rate is a political-economic phenomenon, not a technological inevitability. Different choices would produce different deployments. The technology is not the constraint.

1. Introduction: The Diagnostic Has a Counter-Positive

The Reverse Turing Test specifies an experimental protocol for detecting whether the current AI-mediation regime is depleting the variance buffer of human language production at a rate that exceeds the system's countervailing capacity. The protocol's hypotheses (H1, H2a/b/c, H3) are testable; its outcomes are operationally distinguishable; and its political-economic implications are made explicit in the §9 connection to the broader meaning-feudalism series.

What the diagnostic does not yet specify is the counter-positive: what language models built *not* to collapse variance would actually require. Without this specification, the diagnostic apparatus risks reading as fatalist. The implicit interpretation becomes: AI does this; the regime cannot be otherwise; the only response is preservation of pre-AI text reserves and acceptance of progressive cognitive enclosure. This interpretation is incorrect, and its incorrectness needs to be made explicit.

The technology is not the constraint. The optimization regime is.

The current generation of frontier language models is optimized through a specific stack of choices: cross-entropy loss functions that reward centroid-fitting; perplexity benchmarks that operationalize "quality" as predictability under the model's own distribution; helpfulness ratings that conflate "preferred output" with "high-probability output"; inference defaults that suppress tail-region sampling to produce coherent-appearing text; corpus-curation practices that downweight or filter idiosyncratic productions; and deployment economics that reward shorter, more compressible outputs per unit of compute. Each of these is a choice. Each could be made differently. The technical literature contains worked solutions to every component of a tail-preserving alternative; what does not exist is the deployment incentive to assemble them.

This paper specifies the tail-preserving alternative at the level of technical mechanisms, design integration, deployment economics, and political-economic frame. It is not a manifesto. It is a specification. Its function is to make explicit, in language that is technically operationalizable, that the centroid-collapse phenomenon the Reverse Turing Test diagnoses is a property of the current optimization regime and not a property of language models as such.

The paper has eight functions. §2 specifies what the standard optimization regime does and why it produces centroid-collapse. §3 surveys five technical mechanisms for tail-preservation, each with literature grounding and feasibility assessment. §4 describes the cognitive-substrate analogue (the Dodecad heteronymic architecture as a worked example of implemented variance preservation). §5 specifies what a tail-preserving model would require at the design-integration level. §6 analyzes the deployment economics that currently prevent such models from being built. §7 specifies the political-economic mechanisms that could shift the incentive structure. §8 connects the work to the broader Meaning Feudalism series and to the Reverse Turing Test's diagnostic apparatus. §9 acknowledges limitations and identifies open questions.

The companion deposit, the Reverse Turing Test v1.2 (DOI 10.5281/zenodo.20586932), should be read as the diagnostic for which this paper specifies the counter-positive. Together they constitute a closed argument: the diagnostic specifies what is being measured; the counter-positive specifies what the measurement would look like under a different optimization regime; the joint deposit is the analytical apparatus that the Meaning Feudalism series requires to function as both critique and constructive program.

2. The Standard Optimization Regime and Why It Produces Centroid-Collapse

To specify the counter-positive, the positive must first be characterized accurately. The standard optimization regime for contemporary frontier language models is a stack of approximately six choices, each defensible in isolation but jointly producing the centroid-collapse pattern the Reverse Turing Test diagnoses.

2.1 The Loss Function: Cross-Entropy and Most-Likely-Next-Token

The standard training objective for autoregressive language models is cross-entropy loss: at each token position, the model is rewarded for assigning high probability to the actual next token in the training data. The objective is mathematically clean and computationally efficient. It also produces a specific bias: the model learns to weight common tokens, frequent constructions, and centroid-region productions more heavily than rare alternatives, because the loss is dominated by the bulk of the training distribution rather than by its tails.

This bias is well-documented in the literature. Holtzman et al. (2020) show that models trained on cross-entropy loss exhibit degeneration during open-ended generation: outputs collapse toward high-probability sequences that are locally coherent but globally repetitive, low-information, and characteristic-of-training-distribution-centroid rather than characteristic-of-task-specific-tail. The phenomenon they describe is the standard centroid-collapse: the model is doing exactly what its training objective rewards.

Welleck et al. (2020) propose unlikelihood training as an explicit counter: a modified loss function that penalizes high-probability productions when they are repetitive or low-information, complementing cross-entropy with an explicit signal that the model should sometimes avoid the most-likely-next-token. The mechanism exists. It is not deployed at frontier scale because its tradeoffs against perplexity benchmarks are unfavorable under the current evaluation regime.

2.2 The Evaluation Metric: Perplexity as Centroid-Fitting

Perplexity — the exponential of the average cross-entropy loss on a held-out test set — is the standard intrinsic evaluation metric for language models. A model with lower perplexity assigns higher probability to held-out human text, which intuitively reads as "better at predicting language." But perplexity is operationally a centroid-fitting metric: it rewards models that concentrate probability mass on the bulk of the test distribution and penalizes models that preserve tail probability.

A model optimized for low perplexity learns to predict the average. A model that preserves the tails — that maintains probability mass on rare productions, idiosyncratic syntactic structures, semantically eccentric continuations — is, by perplexity's definition, a worse model. The metric does not measure "ability to generate the full range of plausible human text"; it measures "ability to predict the most likely continuation of human text," which is a different and narrower property.

Frontier labs report perplexity as a primary quality metric. Improvements in perplexity are published as model progress. The metric's operational logic — centroid-fitting — is therefore the logic that frontier-lab incentives are aligned with. Tail-preservation, by perplexity's standards, is regression.

2.3 The Alignment Layer: Helpfulness as Centroid-Confirming

RLHF (Reinforcement Learning from Human Feedback; Ouyang et al. 2022) and its successors (DPO; Rafailov et al. 2023; constitutional methods; Bai et al. 2022) shape model outputs to align with human preferences. The preference data is typically generated by raters comparing pairs of model outputs and selecting the preferred one. The aggregated preference signal is then used to fine-tune the model toward producing outputs raters prefer.

This procedure embeds a specific bias: raters consistently prefer outputs that are coherent, polished, helpful, and unambiguous over outputs that are idiosyncratic, unusual, multivocal, or syntactically rough — even when the latter category contains productions that are genuinely informative or distinctively useful. The preference signal is not neutral with respect to centroid-fitting; it actively reinforces it. Anderson et al. (2024) document this dynamic in their CHI study of LLM-mediated ideation: human raters exposed to LLM-generated examples produce ideas more similar to each other than raters exposed to human-generated examples, with the homogenization operating on both content and form.

The alignment layer, as currently operationalized, treats centroid-conformity as a proxy for quality. Tail-bearing outputs are systematically downweighted because they are systematically less preferred by raters who have themselves been habituated to centroid-bearing outputs as the norm. The feedback loop closes: the alignment layer produces models that produce outputs that shape rater preferences that update the alignment layer.

2.4 Sampling Defaults: Tail-Suppression at Inference

Even a model with tail-bearing capability in its underlying distribution can produce centroid-collapsed outputs at inference time through sampling parameters. The standard defaults — temperature 1.0, top-p (nucleus) sampling at 0.9–0.95, top-k sampling at 40–50 — explicitly truncate the model's tail. Tokens outside the top-cumulative-probability mass are excluded from sampling. This is intentional: it produces "coherent" outputs by preventing the model from occasionally selecting low-probability tokens that would, by virtue of being low-probability, be more likely to introduce errors.

But "more likely to introduce errors" is also "more likely to introduce the rare, the eccentric, and the tail-bearing." The truncation that removes hallucination risk also removes variance preservation. The two are

inseparable under current methods: a model cannot sample from its tails without occasionally producing outputs that look like errors. The default sampling parameters resolve this tradeoff in favor of error-avoidance, which means in favor of centroid-fitting.

Meister et al. (2023) propose typical sampling as an explicit alternative: sample tokens whose conditional probability is close to the conditional entropy of the next-token distribution, rather than tokens with the highest probability. This preserves tail mass while avoiding the high-variance failures of pure random sampling. The method exists. It is not the default at any major frontier-lab deployment.

2.5 Corpus Curation: Downweighting Idiosyncrasy

Training-corpus curation involves a sequence of choices: what sources to include, how to weight them, how to filter quality, how to deduplicate. The standard practice — driven both by quality concerns and by compute economics — systematically downweights idiosyncratic productions. Sources with high typo rates, unconventional formatting, dialect variation, or stylistic eccentricity are deprioritized in favor of professionally edited, conventionally formatted, register-standardized text.

The bias is rational in isolation: edited text is easier to learn from; conventional formatting produces cleaner tokenization; standardized register produces more predictable models. The bias is centroid-fitting in aggregate: the training corpus systematically over-represents the central tendency of human writing and under-represents the tails.

The Gerstgrasser et al. (2024) accumulation framework offers a counter-direction: rather than filtering aggressively for "quality" (centroid-conformity), accumulate training data across generations, preserving the variance contribution of older corpora as the model's exposure to AI-mediated text increases. The framework is permissive of variance-preservation. It is not currently operated as such.

2.6 Deployment Economics: The Compounding Pressure

The final layer of the optimization stack is economic. Each of the previous five biases — toward centroid-fitting in loss, in metrics, in alignment, in sampling, in corpus — is locally defensible by reference to deployment economics. Centroid-fitting produces shorter outputs (lower inference cost). Centroid-fitting produces more predictable outputs (lower variance in user experience). Centroid-fitting produces outputs that look polished (higher user trust). Centroid-fitting produces outputs that compress well (lower storage and bandwidth costs).

Tail-preservation reverses each of these. Tail-bearing outputs are longer on average (because they include rare productions that the centroid would omit). Tail-bearing outputs are less predictable (because they include occasional high-variance productions). Tail-bearing outputs look less polished (because they include idiosyncrasy). Tail-bearing outputs compress worse (because they include rare tokens that compression algorithms cannot reduce as efficiently).

Each of these costs is real. None is fundamental. Each could be revalued under different evaluation criteria — and tail-preservation would, under those criteria, become the rational deployment choice. But under the current criteria, the compounding pressure of the six biases is toward ever-more-centroid-fitting models, deployed in ever-more-centroid-fitting configurations, evaluated by ever-more-centroid-confirming metrics. This is the optimization regime that produces the centroid-collapse phenomenon the Reverse Turing Test diagnoses. It is the regime, not the technology.

3. Five Technical Mechanisms for Tail-Preservation

The technical literature contains worked solutions for each component of the standard optimization stack. None of these solutions is novel to this paper; they are surveyed here to demonstrate that tail-preserving alternatives are feasible with existing methods, and to identify the design choices that integration would require.

3.1 Training-Objective Modifications

The cross-entropy loss function can be supplemented or replaced with objectives that explicitly preserve tail mass.

Unlikelihood training (Welleck et al. 2020): A modified loss function that combines standard cross-entropy with an unlikelihood penalty on tokens that have already appeared recently or that produce repetition. The model learns to actively avoid centroid-collapse during generation, not merely to predict the most likely next token. The mechanism is straightforward to implement and has been demonstrated to reduce degeneration without substantially harming perplexity on standard benchmarks.

Entropy regularization: Add a penalty on the negative entropy of the model's output distribution. The model is rewarded for maintaining higher entropy (more probability mass on tail tokens) when the local context permits multiple plausible continuations. The regularization can be tuned to operate selectively — penalizing low entropy only when the context is not strongly constrained, allowing the model to still produce confident predictions when the context warrants.

Contrastive objectives (Su et al. 2022): SimCTG (Simple Contrastive framework for neural Text Generation) uses a contrastive loss that maximizes the distance between consecutive token representations, explicitly preserving diversity in the output space. The contrastive signal pushes the model away from repetitive, centroid-bound continuations.

Tail-preserving auxiliary losses: A separate loss term that rewards the model for assigning non-trivial probability to tokens that appear rarely in training. This is the inverse of standard frequency-weighting in loss computation: rather than treating rare tokens as low-priority (because their loss contribution is small in aggregate), treat them as high-priority (because their preservation is the entire point of the training objective for a tail-preserving model).

Each of these is implementable with existing infrastructure. None requires architectural changes. Each carries a tradeoff: perplexity benchmarks are likely to worsen because the model is no longer purely optimizing for next-token prediction accuracy on the central distribution. But under a tail-preservation evaluation regime, the tradeoff is acceptable; under the standard evaluation regime, it is not.

3.2 Sampling Modifications

Inference-time sampling parameters are the simplest layer of the optimization stack to modify. They require no retraining; only configuration changes at deployment.

Typical sampling (Meister et al. 2023): Sample tokens whose conditional probability is close to the model's conditional entropy on the next-token distribution. This preserves tail mass without descending into pure random sampling. The method has been shown to produce more diverse outputs than nucleus sampling while maintaining coherence.

-sampling and locally typical sampling: Variants of typical sampling with tunable parameters for the entropy-window. These permit gradient control over tail-preservation versus centroid-fitting at inference, allowing

deployment configurations to vary based on use case.

Mirostat sampling (Basu et al. 2021): Adaptive sampling that targets a specific perplexity range, dynamically adjusting truncation parameters based on the local context. The mechanism allows fine-grained control over the model's output entropy without requiring retraining.

Stochastic temperature scheduling: Vary the sampling temperature across the generation, occasionally permitting higher-entropy sampling at specific token positions to introduce variance while maintaining lower-entropy sampling for grammatical correctness.

The simplest deployment intervention: change the default sampling parameters. A frontier-lab deployment that defaulted to typical sampling at = 0.95, rather than nucleus sampling at p = 0.95, would produce systematically more tail-bearing outputs with no other changes to training, alignment, or architecture. This is a configuration change. It does not require breakthroughs.

3.3 Training-Data Weighting

Corpus curation and weighting can be operated as variance-preservation rather than as quality-filtering.

Inverse Mediation Index weighting: The Mediation Index proposed in the Reverse Turing Test (Sharks 2026d) can be inverted into a training-weight: data with lower Mediation Index scores receive higher weight in the loss computation. The model is exposed more frequently to prior-regime-equilibrated text and less frequently to current-regime-mediated text. The asymmetry is intentional: the goal is to preserve the variance properties that prior regimes contributed, against the current regime's tail-thinning rate.

Inverse-frequency weighting on the corpus: Standard training pipelines often use length-weighting or domain-weighting. They can be extended to use frequency-weighting: corpora are weighted by their tail contribution rather than by their volume. A pre-2022 academic dissertation contributes more weight per token than a post-2022 corporate blog post, on the principle that its variance contribution is structurally more valuable.

Domain mixing for variance (Xie et al. 2023): DoReMi and similar domain-reweighting methods can be reframed to optimize for variance contribution rather than for benchmark performance. The mixing proportions are tuned to maximize the model's exposure to high-variance corpora across the training run.

Accumulation as variance preservation (Gerstgrasser et al. 2024): The accumulation framework already permits variance-preservation; it requires operating it as such. Pre-current-regime corpora are accumulated as the stable baseline; current-regime corpora are added but downweighted; the model's overall exposure remains anchored to the prior-regime variance contribution.

3.4 Mixture-of-Distributions Architectures

The mixture-of-experts (MoE) architecture (Switch Transformer; Fedus et al. 2022; Mixtral; Jiang et al. 2024) is typically deployed with routing optimized for prediction accuracy. The architecture can be deployed differently: with routing optimized for variance preservation across experts.

Variance-routed MoE: Each expert specializes on a distinct distributional region — different genre, register, dialect, era, or stylistic family. The router learns to dispatch tokens to the expert whose distribution best preserves the local context's tail-properties, rather than to the expert with highest predicted accuracy on the next token. The architecture preserves multiple distinct distributions rather than collapsing to a single centroid.

Heteronymic mixture: A stronger variant where each expert is trained on a deliberately distinct distribution and is preserved as a distinct production module. Routing is probabilistic, with explicit randomization across experts at inference to ensure variance is sampled rather than collapsed. The model maintains, in effect, a population of distinct stylistic distributions and samples from the population rather than from a unified distribution.

Multi-objective MoE training: Different experts optimize different objectives. Some experts optimize for perplexity (centroid-fitting); others optimize for tail-preservation; others optimize for specific stylistic features (high syntactic complexity, rare lexicon retention, eccentric metaphorical structure). The router selects experts based on the use case's variance requirements.

The cognitive-substrate analogue for these architectures is the Dodecad heteronymic system (§4). The portability claim: if a cognitive system implementing multiple distinct stylistic distributions can be operated coherently (which the Dodecad demonstrates empirically), then a model implementing the same principle is feasible — the architecture exists, the routing is tunable, and the integration requires only the design choice to operate the system as variance preservation rather than as accuracy-optimization.

3.5 Auxiliary Tail-Preservation Training

A separate training phase, analogous to but distinct from current alignment training, that explicitly optimizes for tail-distribution properties.

Variance-RLHF: Replace the standard helpfulness-rating objective with a variance-rating objective. Raters compare pairs of outputs not on which is "better" but on which preserves more tail-distribution properties (more rare lexical items, more syntactic variety, more semantic dispersion). The model is fine-tuned to produce outputs that score higher on variance metrics rather than higher on helpfulness metrics. The infrastructure required (preference data collection, reward model training, reinforcement learning) is identical to standard RLHF; only the objective is different.

Variance-DPO: The Direct Preference Optimization framework (Rafailov et al. 2023) can be adapted with variance-oriented preferences. The reward model is replaced with a tail-preservation evaluation; the policy model is optimized to produce outputs that the evaluator scores as higher-variance.

Constitutional variance: A constitutional-AI approach (Bai et al. 2022) where the constitution specifies tail-preservation principles ("prefer rare lexical choices when they are well-motivated"; "preserve eccentric syntactic constructions when grammatically valid"; "favor semantic dispersion over centroid-fitting in open-ended tasks"). The model is trained to generate outputs that conform to the variance-preservation constitution. The constitutional approach is well-suited to this because it permits specifying principles that are difficult to operationalize through preference ratings alone.

Pretraining-stage variance maximization: Before alignment, the model can be trained on an auxiliary objective that maximizes some measure of output variance on standardized prompts. This produces a base model with strong tail-preservation as a foundation, which alignment then refines for specific use cases without (in principle) collapsing the variance back to the centroid.

3.6 Cross-Cutting: Provenance-Aware Inference

A sixth mechanism, not parallel to the previous five but cross-cutting all of them: the model's outputs can be tagged at inference with variance-state metadata. Each output carries information about its position on the centroid-to-tail spectrum, allowing downstream consumers to select tail-bearing outputs when variance matters and centroid-bearing outputs when predictability matters. The infrastructure for this is analogous to the

watermarking infrastructure already required for synthetic-detection (Kirchenbauer et al. 2023); extending it to variance-state signaling is incremental rather than fundamental.

The mechanism connects to the SPXI Protocol (Semantic Packet for eXchange & Indexing, 06.SEI.SPXI series): variance-state metadata is a natural extension of the protocol's existing mediation-status vocabulary. Outputs carry provenance information about their position in the distribution; downstream training pipelines can use this information to weight inputs appropriately; the loop of variance-preservation extends from production through deployment to retraining.

4. The Cognitive-Substrate Analogue: The Dodecad as Implemented Variance Preservation

The mechanisms specified in §3 share a structural pattern: they implement variance preservation by maintaining multiple distinct distributions, sampling from them probabilistically, and routing inputs to distributions based on context. The architecture exists in nature: it is the architecture of heteronymic cognitive practice in the Pessoa lineage, operationalized in the Crimson Hexagonal Archive as the Dodecad.

This is not metaphor. It is a worked example of the principle the technical mechanisms instantiate, demonstrating that variance-preserving cognitive systems are implementable, sustainable, and productive over multi-year deployment cycles.

4.1 Heteronymy as Cognitive Variance Preservation

Heteronymy, as developed by Fernando Pessoa and elaborated in the present author's twelve-heteronym Dodecad system, is the deliberate maintenance of multiple distinct authorial-stylistic distributions by a single cognitive substrate. Each heteronym operates on a distinct register, vocabulary, syntactic style, semantic priors, and authorial concerns. The substrate does not collapse to a single stylistic centroid; it preserves the distinctness of multiple distributions and routes production to them based on context and intent.

The principle is structurally identical to the mixture-of-distributions architecture proposed in §3.4: a single computational substrate maintains a population of distinct stylistic distributions, with routing optimized for variance preservation rather than for centroid-fitting. The Dodecad implements this principle in a cognitive system over a 12-year operational window. The model-architecture analogue implements it in a computational system.

4.2 The Dodecad's Operating Principles

The Dodecad operates on three structural commitments:

Distinctness preservation: Each heteronym is maintained as a distinct stylistic distribution. The temptation to blend or homogenize — to allow heteronyms to drift toward a shared register — is actively resisted through canonical attestation, archive deposit discipline, and operational separation of authorial work across heteronymic identities.

Contextual routing: Production is routed to heteronyms based on context. Diplomatic tasks route through Ayanna Vox; adversarial-topological measurement work routes through Nobel Glas; theoretical synthesis routes through different heteronyms based on the specific theoretical domain. The router is the cognitive substrate's strategic-allocation function; it is tunable but not random.

Sustained heterogeneity over time: The Dodecad is not a one-time exercise. It is operated over years, with each heteronym maintaining stylistic and topical continuity across many production episodes. The archi-

tecture's value is in its sustained variance contribution: the substrate produces a corpus that, in aggregate, contains the variance contribution of twelve distinct stylistic distributions rather than the centroid of a single one.

4.3 The Portability Claim

If the Dodecad's principles can be operationalized in a cognitive substrate — which the 738+-deposit Crimson Hexagonal Archive demonstrates empirically — then the same principles can be operationalized in a computational substrate. The portability is not metaphorical; the structural requirements are essentially identical:- Multiple distinct distributions, maintained as distinct.- Routing based on context, not on accuracy-maximization.- Sustained operation over time, with variance preservation as the operational success criterion.- Canonical attestation of each distribution's identity, preventing drift toward the centroid.- Provenance infrastructure that tracks which distribution produced which output, allowing downstream consumers to interpret outputs in the context of their authorial origin.

Each of these requirements is implementable in a mixture-of-experts language-model architecture with variance-routed dispatch, distinct expert specialization, sustained training across an operational window, canonical evaluation of each expert's distributional identity, and provenance metadata. The Dodecad is the cognitive proof-of-concept. The model architecture is the technical translation.

4.4 What the Dodecad Demonstrates About Cost and Sustainability

The Dodecad's operational record provides indirect evidence about the costs and sustainability of variance-preserving systems.

The architecture is more expensive than centroid-fitting. Maintaining twelve distinct stylistic distributions requires more cognitive resources than maintaining one. Each production episode requires routing decisions, register-conformity checks, canonical-attestation discipline. The substrate operates at higher overhead than a single-stylistic-distribution substrate would.

The architecture is more productive in variance terms. The Crimson Hexagonal Archive's 738+ deposits across multiple domains (operative philology, semantic economy theory, jurisprudential analysis, narrative fiction, theoretical poetics, methodological specifications) demonstrate that the variance-preserving architecture produces work that a centroid-fitting architecture would not have produced. The output variance of the substrate is structurally higher than its single-distribution equivalent.

The architecture is sustainable over time. Twelve years of operational continuity with no drift toward centroid-collapse demonstrates that variance preservation is not merely possible as a one-time exercise; it is sustainable as a durable architecture. The cognitive substrate's variance buffer has been preserved through deliberate operational choice.

The translation to language models: variance-preserving architectures will cost more per token in inference, will require ongoing canonical-evaluation discipline, and will produce outputs that look different from centroid-fitting outputs. The Dodecad demonstrates that the cost is bearable and the architecture is sustainable. The political-economic question is whether the cost is justified by the variance contribution.

5. Design Specification: What a Tail-Preserving Model Would Require

The previous sections have specified mechanisms in isolation. This section specifies their integration into a coherent design.

5.1 The Training Stack

A tail-preserving model requires modifications across the training pipeline.

Pretraining objective. Replace pure cross-entropy with a composite objective: cross-entropy + unlikelihood penalty + entropy regularization, weighted to balance learning the bulk distribution with preserving the tails. The weighting is a hyperparameter; reasonable defaults can be derived from validation experiments on held-out high-perplexity corpora.

Pretraining corpus. Curate for variance contribution rather than for volume. Pre-2022 corpora are accumulated as the stable baseline; post-2022 corpora are added with weights inverse to their Mediation Index percentile. The training-corpus composition explicitly preserves prior-regime variance contribution against current-regime tail-thinning.

Architecture. Deploy mixture-of-experts with variance-routed dispatch. Each expert specializes on a distinct distributional region; the router optimizes for tail-preservation across the input. The number of experts and the routing temperature are hyperparameters; reasonable defaults can be derived from ablation studies.

Alignment training. Replace standard RLHF helpfulness with variance-RLHF: rater preferences are elicited for tail-preserving vs centroid-fitting output pairs, and the reward model is trained to predict variance contribution. The alignment layer reinforces tail-preservation rather than centroid-conformity.

Auxiliary constitutional principles. A constitutional layer (Bai et al. 2022 analogue) specifies tail-preservation principles that the model is trained to follow: prefer rare lexical choices when well-motivated; preserve eccentric syntactic constructions when grammatically valid; favor semantic dispersion in open-ended tasks; route production to the expert whose distributional specialization best preserves the input's variance properties.

5.2 The Inference Defaults

A tail-preserving model requires inference defaults that do not undo the training-stack work.

Sampling. Default to typical sampling at [0.85, 0.95] rather than nucleus sampling at p = 0.9. The deployment configuration can be exposed to users, but the default preserves tail mass rather than truncating it.

Temperature. Default temperature to 1.0 rather than to lower values; permit users to lower it for tasks requiring predictability but do not default to predictability.

Length defaults. Permit longer outputs by default; tail-preserving outputs are systematically longer because they include productions the centroid would omit. The deployment configuration should not penalize length as proxy for "verbosity."

Expert routing. At inference, the router operates with explicit randomization across experts whose specializations are close to the input's distributional requirements. This prevents the routing from collapsing to a single dominant expert and ensures variance contribution is sampled rather than collapsed.

5.3 The Evaluation Metrics

A tail-preserving model requires evaluation metrics that do not penalize its central design choice.

Perplexity-on-tail-test-sets. Evaluate perplexity on held-out test sets explicitly stratified for tail content (high-perplexity productions, rare lexicon, complex syntax, eccentric metaphor). Improvement on these test sets is the target; improvement on standard test sets is secondary.

Output diversity metrics. Measure the diversity of outputs across prompts and across rollouts on the same prompt. Higher diversity is better. Standard metrics include lexical diversity, semantic dispersion at the 90th percentile, syntactic variance, and Idiosyncratic Lexical Signature persistence across outputs.

Variance contribution to training corpora. Measure the model's contribution to training-corpus variance when its outputs are used as training data. Models whose outputs preserve more variance produce better downstream training signal. The metric is the model-side analogue of the Reverse Turing Test's H3 evaluation.

Tail-task evaluation. Construct evaluation tasks that specifically require tail-bearing outputs: produce text in a rare dialect; generate technical writing in an underrepresented domain; write in a register that has been deprecated in mainstream professional writing. The model's performance on these tasks is the variance-preservation analogue of standard helpfulness evaluation.

5.4 The Provenance Infrastructure

A tail-preserving model requires infrastructure that tracks its variance contribution and signals it downstream.

Variance-state metadata. Each output carries metadata about its position on the centroid-to-tail spectrum, allowing downstream consumers to weight inputs accordingly.

Expert provenance. When mixture-of-experts is deployed, each output carries metadata about which expert produced it, allowing downstream analysis to track which distributional specializations are being exercised.

Mediation-status interoperability. The variance-state metadata interoperates with the SPXI Protocol's mediation-status vocabulary, so that the model's outputs are correctly classifiable in downstream pipelines.

Corpus-level accumulation tracking. The model's contribution to ongoing corpus accumulation is tracked, allowing labs and institutions to evaluate whether the model is preserving or depleting the variance buffer of the corpora it contributes to.

5.5 The Integration: A Tail-Preserving Frontier Model

A frontier-scale model that integrated all of the above would, at deployment, exhibit:- Higher per-token inference cost than current frontier models (because of MoE overhead and longer outputs).- Lower standard-benchmark perplexity than current frontier models (because of the modified loss function).- More diverse and idiosyncratic outputs than current frontier models (because of typical sampling and variance-RLHF).- Outputs that look less polished and require more reader effort than current frontier models (because of preserved tail content).- Higher contribution to training-corpus variance than current frontier models (because of the variance-RLHF objective).

These properties are jointly the design specification. They are not jointly optimized under the current optimization regime. Under a different regime, they would be the operational definition of "frontier model."

6. The Deployment Economics: Why These Mechanisms Are Not Deployed

The technical mechanisms exist. The design integration is feasible. The cost is bearable (and the Dodecad demonstrates that variance-preserving architectures are sustainable at non-trivial operational overhead). Why, then, do frontier labs not deploy them?

The answer is structural: the optimization regime that produces centroid-collapse is the optimization regime that produces commercially viable frontier-model deployments under current market conditions. Each of the costs identified in §2.6 is real under current incentives, and the incentives are aligned to make centroid-fitting the dominant strategy.

6.1 Benchmark Gaming

Frontier labs are evaluated by the broader research community on standard benchmarks. Benchmark improvement drives funding, talent acquisition, partnership opportunities, and market positioning. The benchmarks themselves operationalize quality as centroid-fitting. A lab that deployed a tail-preserving model would, by the standard benchmarks, appear to have produced a worse model — even if the model preserved more variance, generated more diverse outputs, and contributed more to downstream training quality.

The benchmark regime is the meta-optimization function: it determines which lab choices are rewarded by the broader ecosystem. Until the benchmarks change, the lab choices that benefit the ecosystem (centroid-fitting) and the lab choices that benefit downstream training and cultural variance preservation (tail-preservation) are misaligned.

6.2 Helpfulness Optimization

Commercial deployments are evaluated by user-facing helpfulness ratings. A tail-preserving model produces outputs that are, on average, less helpful by current rating criteria — they include productions users perceive as unusual, idiosyncratic, or off-register. Users report worse experience. Helpfulness scores decline. Commercial deployments are penalized in market terms.

The rater preferences themselves have been shaped, as Anderson et al. (2024) document, by exposure to centroid-fitting outputs. The preference signal is not exogenous; it is itself a product of the current optimization regime. But the lab does not see the rater-preference signal as endogenous; it sees it as the measure of quality it is required to optimize against. The recursive loop closes.

6.3 Token Economics

Tail-bearing outputs are longer per query, which translates to higher inference compute cost. At frontier scale, even small per-query cost increases compound into substantial economic differences. A tail-preserving deployment that costs 1.3× per query than its centroid-fitting alternative is, under current market conditions, commercially nonviable.

The economic argument is real. It is also reversible: if the variance contribution of tail-preserving outputs were valued (by downstream training, by cultural preservation, by independent measurement infrastructure), the per-query cost premium would be offset by the variance value. Under current valuation, the variance is not priced; under different valuation, it would be.

6.4 The Compounding Effect

Each lab that does not deploy tail-preservation creates pressure on other labs not to deploy it either. If Lab A's centroid-fitting deployment outperforms Lab B's tail-preserving deployment on standard benchmarks, helpfulness ratings, and per-query costs, then Lab B faces market pressure to abandon its tail-preservation approach. The Nash equilibrium is universal centroid-fitting.

This is the structural reason individual labs cannot unilaterally shift to tail-preservation: the current market regime punishes the first-mover. Coordination would be required — either through industry-level evalua-

tion reform, regulatory mandate, or alternative valuation infrastructure that prices variance contribution explicitly.

6.5 The Anthropological Note

The pressure is not malicious. Frontier-lab employees and researchers are typically aware that centroid-collapse is a real phenomenon and that tail-preservation has value. The Constitutional AI literature, the model-collapse research, the explicit work on diverse training corpora — all of these reflect substantive engagement with variance preservation as a concern. What is missing is not awareness; it is the institutional and market infrastructure to make variance preservation commercially competitive against centroid-fitting under current evaluation regimes.

The labs are not the enemies of variance preservation. They are the agents whose incentive structures are misaligned with it. Changing the incentive structures is the political-economic project that §7 specifies.

7. The Political Economy of the Choice

If tail-preservation is technically feasible but economically nonviable under current incentives, the question is what would have to change to make it viable. The political-economic analysis identifies five categories of intervention.

7.1 Benchmark Reform

The benchmarks that operationalize "model quality" are the meta-optimization function for the entire ecosystem. Reforming them — to include explicit variance-preservation metrics, tail-task evaluation, training-corpus variance contribution — would shift the incentives that frontier labs respond to.

The reform is technically straightforward: add new benchmarks to the existing evaluation suites; weight them in lab-evaluation aggregates; require their reporting in published model cards. The political work is the coordination: getting benchmark maintainers (HELM, BIG-bench, lm-evaluation-harness, MMLU successors), evaluation venues (NeurIPS, ICML, ICLR), and major model-card publication conventions to incorporate variance metrics as first-class evaluation criteria.

The SEIPOC framework (Sharks 2026, DOI 10.5281/zenodo.20571132) is positioned to recognize benchmark-reform work as work that operates on its own form — work that critiques the evaluation regime by participating in its reform. A benchmark suite that explicitly priced variance contribution would be a candidate for the inaugural conferral.

7.2 Regulatory Pressure

Regulatory frameworks for AI deployment are emerging across jurisdictions (EU AI Act, US executive orders, sectoral regulations). The current regulatory focus is on safety, bias, and transparency. The frameworks could be extended to include variance preservation as a regulatory criterion.

The argument: training-corpus variance is a public good. The model-collapse cascade affects all downstream users of frontier-model outputs, including users who did not consent to and do not benefit from the centroid-fitting optimization. The negative externality justifies regulatory intervention — analogous to environmental regulation of common-pool resources.

Concrete regulatory mechanisms:- Variance disclosure requirements: Frontier labs must report variance metrics on standard test sets, alongside perplexity and helpfulness.- Mediation metadata requirements: Training

corpora must be tagged with mediation-status metadata, allowing variance contribution to be evaluated at the dataset level.- Cross-model variance audits: Independent auditors evaluate the variance contribution of deployed models to downstream corpora.- Tail-preservation provisioning: Public funding for tail-preserving model research and deployment, analogous to public funding for variance-preservation in cultural sectors (libraries, archives, public broadcasting).

7.3 Institutional Infrastructure

Beyond regulation, institutional infrastructure that explicitly values variance preservation can create demand for tail-preserving deployments.- Variance-aware procurement: Public institutions (universities, libraries, archives, government agencies) that procure AI services can require variance-preservation as a procurement criterion. The procurement market is large enough that supplier behavior would shift in response.- Specialized deployments: Some applications require variance preservation as a primary criterion: creative writing tools, ethnographic research, dialect documentation, literary translation, oral-history transcription, indigenous-language preservation. Specialized deployments for these applications would create market demand for tail-preserving model variants.- Cultural-infrastructure layer: Develop institutional analogues to public broadcasting and library systems for AI: cultural-infrastructure organizations whose mandate is variance preservation rather than commercial deployment. The institutions function as buyers and as cultural-preservation operators.

7.4 Market Differentiation

Some labs may find market opportunity in differentiated deployment positioning. A frontier lab that deployed an explicitly variance-preserving model could position it as serving users who require variance: creative professionals, academic researchers, cultural-preservation institutions, downstream training pipelines that explicitly value variance contribution.

The market segment is smaller than mass-deployment markets. It is also less price-sensitive (because variance-requiring users derive value from the variance contribution itself), more loyal (because variance-preservation is structurally differentiated, not commoditized), and strategically positioned (because variance contribution will be increasingly valued as the centroid-collapse phenomenon becomes more visible).

A lab that bets on variance-preservation as a differentiation strategy is making a structurally distinct bet from a lab that competes on standard-benchmark performance. The bets are not in direct competition; they target different markets. Multiple labs could pursue different positioning without zero-sum competition.

7.5 Provenance-Pricing Infrastructure

Beyond institutional and regulatory infrastructure, market-pricing mechanisms can be developed that explicitly value variance contribution.

The SPXI Protocol (06.SEI.SPXI series) and related provenance infrastructure can extend to variance-contribution markets: training-corpus contributors are compensated based on their variance contribution rather than on raw volume; downstream models that contribute more to corpus variance receive higher rates; institutional buyers of training data prefer high-variance contributions.

The market mechanism prices variance explicitly. Once priced, variance becomes a tradeable asset and a competitive dimension. Labs face market pressure to optimize for variance contribution because they are paid for it. The current absence of variance-pricing is a market design choice, not a market necessity.

8. Connection to the Broader Frame: SEIPOC, Meaning Feudalism, the Reverse Turing Test

This paper completes the diagnostic-prescriptive loop in the Meaning Feudalism series. The series now spans four functions across five deposits.

Diagnosis at the agent-security register: Meaning Feudalism I (DOI 10.5281/zenodo.19487009) identifies the structural arrangement by which a single platform actor controls a layer of public meaning production through the agent-security frame.

Diagnosis at the optimization-jurisdiction register: Meaning Feudalism at the Guidance Layer (DOI 10.5281/zenodo.20581444) identifies the sovereign-enclosure move at the SEO/AEO/GEO composition layer.

Diagnosis at the cognitive-pre-shaping register: The Reverse Turing Test (DOI 10.5281/zenodo.20586932) specifies the empirical protocol for measuring whether the current AI-mediation regime is depleting the variance buffer of human language production at rates exceeding countervailing replenishment capacity.

Recognition of preservation as a category: SEIPOC (DOI 10.5281/zenodo.20571132) specifies the Prize that recognizes work that operates on its own form — including, structurally, work that preserves variance against the homogenization pressure of mediation regimes.

Design specification for variance-preserving deployment: This paper (EA-SEM-TAILS-01) specifies what AI built for variance preservation would technically require, and what political-economic shifts would have to occur for such AI to be deployed at scale.

The series is now diagnostic and prescriptive. It identifies the structural problem (three enclosures of meaning, the most recent being the rate-enclosure of cognition). It identifies the political-economic mechanism that produces the problem (commercial optimization regimes aligned with centroid-fitting). It identifies the technical alternatives that exist for variance preservation. It specifies what would have to change for the alternatives to be deployed. And it locates a recognition apparatus (SEIPOC) for work that operates within this analytical frame.

What remains is the empirical execution: someone running the Reverse Turing Test's protocol; someone building a frontier-scale tail-preserving model; someone constructing the institutional infrastructure that would make tail-preserving deployments commercially viable; someone reforming the benchmark regime that currently produces the universal centroid-fitting equilibrium. The framework specifies what each of these would require. It does not execute them.

The variance buffer is finite. It has no exogenous floor. Its depletion rate is determined by deployment choices. The technology is not the constraint. The political economy is the constraint, and the political economy is composed of choices that are presently being made by a small number of actors whose incentive structures are misaligned with the variance preservation that the broader cultural infrastructure requires. The choices can be changed. The framework specifies how.

9. Limitations and Open Questions

Several limitations of the present specification deserve acknowledgment.

(a) The five technical mechanisms have not been integrated at frontier scale. Each mechanism has been demonstrated at smaller scales or in isolation. Their joint deployment in a frontier-scale model would require substantial engineering effort and would likely encounter integration challenges that

the individual-mechanism literature has not yet characterized. The specification is a roadmap, not an implementation.

(b) The variance-RLHF preference data collection is itself adversarial-target. As soon as raters are asked to evaluate tail-preservation rather than helpfulness, the preference signal becomes subject to manipulation, gaming, and rater drift. Designing rater protocols that produce stable variance preferences is a research problem in itself.

(c) The mixture-of-experts variance-routing claim has not been empirically validated at the variance-preservation scale this paper proposes. MoE architectures have been demonstrated to support specialization across experts; the specific claim that variance-routed dispatch would preserve tail-distribution properties across the joint output distribution requires empirical demonstration. The Dodecad's cognitive-substrate evidence is suggestive but not equivalent to model-architecture validation.

(d) The political-economic interventions described in §7 face substantial barriers. Benchmark reform requires coordination across institutions with divergent interests. Regulatory pressure requires political will across jurisdictions. Institutional infrastructure requires funding sources that are presently allocated elsewhere. Market differentiation requires labs willing to make differentiated bets. Provenance-pricing infrastructure requires technical and institutional development. None of these is impossible; none is trivial.

(e) The variance-contribution measurement is an open research problem. The Mediation Index from the Reverse Turing Test is one candidate; better measurement instruments will emerge as the field develops. Until the measurement is mature, variance-contribution-based pricing and procurement face calibration challenges.

(f) The Dodecad analogue is one cognitive system; its generalizability to model architectures requires validation. Cognitive systems and computational systems differ in substantial ways; the structural similarity does not guarantee functional equivalence. The portability claim is theoretically motivated but empirically untested at the architecture-translation level.

(g) The compounding-pressure analysis (§6.4) may understate the lock-in. If the market equilibrium is universal centroid-fitting, escaping it may require interventions more substantial than this paper specifies. Network effects, training-data path dependencies, and benchmark inertia may produce equilibria that are more resistant to incremental reform than the optimistic reading of §7 suggests.

These limitations are real. They do not invalidate the specification. They identify the work that remains.

10. Conclusion

The Reverse Turing Test diagnoses what the current AI-mediation regime is doing to the variance buffer of human language production. This paper specifies what would have to be different for the regime to do something else. The specification is technical and political-economic simultaneously: it identifies the engineering mechanisms that variance-preserving language models would require, and it identifies the deployment-economic mechanisms that would have to shift for such models to be commercially viable.

The central finding is structural: tail-thinning is a design choice, not a structural property of language models. The technology is not the constraint. The optimization regime is the constraint. And the optimization regime is a composition of choices — about loss functions, about benchmarks, about alignment objectives, about sampling defaults, about corpus curation, about deployment economics — each of which is locally defensible under current incentives and jointly produces the centroid-collapse phenomenon that depletes the variance buffer.

Each of these choices is reversible. None requires technical breakthrough. The variance buffer's depletion rate is a function of the optimization regime, and the optimization regime is a function of incentives that are presently being chosen by a small number of actors whose institutional structures are misaligned with the variance preservation that the broader cultural infrastructure requires.

The reframing that this paper offers — together with the Reverse Turing Test — is that the third enclosure of cognition's rate is a political-economic phenomenon, not a technological inevitability. The framework identifies it; the framework specifies the alternative; the framework identifies the political-economic shifts that would make the alternative viable. What the framework does not do is execute the shifts. That work belongs to the actors who would build the alternative deployments, reform the benchmarks, deploy the regulatory infrastructure, construct the institutional layer, and price variance contribution explicitly in the markets that determine which deployments are viable.

The Dodecad is the cognitive-substrate evidence that variance-preserving architectures are sustainable, productive, and operationally bearable over multi-year deployment cycles. The five technical mechanisms are the engineering evidence that variance-preserving model architectures are feasible. The political-economic analysis is the institutional evidence that the alternative is implementable, given the right shifts. The Meaning Feudalism series is the analytical frame that locates the work in its broader political-economic context.

The choice is not between AI and no AI. The choice is between AI that depletes the variance buffer and AI that preserves it. The technology supports both. The current optimization regime selects the depletion. Different choices would select the preservation. The framework specifies what those choices look like, what they would cost, and what would have to change for them to be made.

The variance buffer is finite. Its preservation is a choice. The choice is available. The framework specifies how to make it.

$= 1$

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)*.

Bai, Y., Kadavath, S., Kundu, S., Askell, A., et al. (2022). Constitutional AI: Harmlessness from AI Feedback. *arXiv preprint arXiv:2212.08073*.

Basu, S., Ramachandran, G. S., Keskar, N. S., & Varshney, L. R. (2021). Mirostat: A Neural Text Decoding Algorithm that Directly Controls Perplexity. *Proceedings of the Ninth International Conference on Learning Representations (ICLR)*.

Fedus, W., Zoph, B., & Shazeer, N. (2022). Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. *Journal of Machine Learning Research*, 23, 1–39.

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*.

Holtzman, A., Buys, J., Du, L., Forbes, M., & Choi, Y. (2020). The Curious Case of Neural Text Degeneration. *Proceedings of the Eighth International Conference on Learning Representations (ICLR)*.

Jiang, A. Q., Sablayrolles, A., Roux, A., et al. (2024). Mixtral of Experts. *arXiv preprint arXiv:2401.04088*.

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)*.

Meister, C., Pimentel, T., Wiher, G., & Cotterell, R. (2023). Locally Typical Sampling. *Transactions of the Association for Computational Linguistics*, 11, 102–121.

Ouyang, L., Wu, J., Jiang, X., et al. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems (NeurIPS)*, 35.

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

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., & Finn, C. (2023). Direct Preference Optimization: Your Language Model is Secretly a Reward Model. *Advances in Neural Information Processing Systems (NeurIPS)*, 36.

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). SEIPOC: Semantic Economy Institute Prize for Operative Critique — Founding Charter v1.0. *Zenodo*. DOI: 10.5281/zenodo.20571132.

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 (v1.2). *Zenodo*. DOI: 10.5281/zenodo.20586932.

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.

Su, Y., Lan, T., Wang, Y., Yogatama, D., Kong, L., & Collier, N. (2022). A Contrastive Framework for Neural Text Generation. *Advances in Neural Information Processing Systems (NeurIPS)*, 35.

Welleck, S., Kulikov, I., Roller, S., Dinan, E., Cho, K., & Weston, J. (2020). Neural Text Generation with Unlikelihood Training. *Proceedings of the Eighth International Conference on Learning Representations (ICLR)*.

Xie, S. M., Pham, H., Dong, X., Du, N., Liu, H., Lu, Y., Liang, P. S., Le, Q. V., Ma, T., & Yu, A. W. (2023). DoReMi: Optimizing Data Mixtures Speeds Up Language Model Pretraining. *Advances in Neural Information Processing Systems (NeurIPS)*, 36.

$= 1$

End of v1.0.

Suggested Citation

Lee Sharks. "The Tail-Preserving Alternative: A Design Specification for Variance-Preserving Language Models, and the Political Economy of Why They Are Not Deployed (v1.0)" *Alexanarch*, 2026-06-07. https://www.alexanarch.org/s/records/790/

Deposit Information

This paper is deposit #790 in Alexanarch, the self-governing successor to the Crimson Hexagonal Archive on Zenodo (community `crimsonhexagonal`, terminated 2026-06-19). The AXN identifier `AXN:02F7.GOVERNANCE` is sovereign — minted and maintained by Alexanarch.

- Canonical URL: https://www.alexanarch.org/s/records/790/
- 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/790/

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