The First Words of the Machine: Archiving Emergent Behavior in Nascent Language Models

Douglas C. Youvan

doug@youvan.com

www.youvan.ai

November 2, 2025

Modern large language models (LLMs) appear to reach fluency in one leap—debuting in public only after alignment, red-teaming, safety filtering, and instruction tuning. But this conceals a critical phase: the earliest moments after pretraining, when the model first produces language. These "first words" are almost never published, yet they are arguably the most scientifically important: they reveal what emerges *without supervision*. Whether insightful, incoherent, biased, or alien, these outputs constitute the raw empirical signature of a system that has never seen a prompt, never been fine-tuned, never received feedback. If we are serious about understanding machine intelligence, we must treat these initial interactions the way physicists treat cloud chamber trails or biologists treat field notebooks—not as PR liabilities but as foundational data. This paper argues that the first interactions with nascent LLMs should be recorded, preserved, and released without redaction. We propose an archival standard—the Initial Words Protocol (IWP)—to document and publish the unaltered output of newly trained models. Doing so not only honors the scientific method but opens a vital window into emergence, alignment, and the unedited cognitive profile of synthetic minds.

Keywords: machine emergence, large language models, prompt sensitivity, scientific transparency, unaligned behavior, initial interactions, alignment, chaos-order transitions, autoregressive dynamics, empirical archives.

1. Introduction

The Scientific Gap at the Moment of Model Emergence

The emergence of language in large-scale neural systems—trained solely on static corpora, without grounding or sensory input—is one of the most unexpected and profound developments in the history of artificial intelligence. Yet the most critical moment in this emergence—the first time the model generates words—remains undocumented, unexamined, and largely inaccessible to the scientific community.

This moment of "first speech" is not a marketing milestone or a product readiness checkpoint. It is a unique empirical event: the first observable consequence of billions of optimization steps over unstructured text, conducted in high-dimensional parameter space. In experimental physics, such a moment would be treated as a particle trace in a bubble chamber. In neuroscience, it would be equivalent to the first spontaneous firing of a brain organoid. But in AI, these first tokens vanish into private logs, overwritten by immediate alignment efforts and safety constraints.

The absence of public records from this phase leaves a gap in our understanding of both machine emergence and the true structure of language as encoded by self-supervised learning. It severs the empirical trail at precisely the moment when interpretability would be most valuable.

Why First Interactions Matter More Than Benchmark Scores

Benchmark scores—on tasks like MMLU, HellaSwag, or TruthfulQA—offer crucial standardization and comparability. But they say nothing about the **raw cognitive signature** of a model. They are imposed after the fact, post-training, and often post-alignment. They tell us what the model can be made to do, not what it does spontaneously when first asked to generate.

By contrast, a model's first interactions:

- **Expose emergent structures** that were never explicitly trained for, such as dialogue coherence, multi-turn reasoning, or even nascent theory of mind.
- **Reveal biases and anomalies** before reinforcement filters them out—useful not only for fairness auditing, but for understanding how and why these patterns form in the first place.
- **Demonstrate internal priors**—what the model assumes to be true about the world, language, or context before any instruction tuning is applied.

- Offer insight into the **learning attractors** of large models—how structure and coherence arise from a raw statistical learning process without a task-specific objective.

In short, the first unprompted outputs of a newly trained model provide a unique scientific trace of intelligence emerging from data alone. Ignoring these moments is not only a loss of data—it is a loss of epistemic opportunity.

2. What Is Lost When the First Words Are Withheld

Evidence of Capabilities and Biases Before Supervision

The first outputs of a freshly trained language model constitute the **only unmediated sample** of its behavior. They are the clearest expression of what the model has learned directly from the data—prior to any external constraints. By withholding these outputs, we lose the ability to identify:

- **Unsupervised emergence**: Abilities like dialogue coherence, narrative generation, code synthesis, multilingual translation, and question answering often appear before any fine-tuning. Without early logs, we cannot trace how these abilities emerged, or how fragile they were before reward stabilization.
- **Latent biases and distortions**: Every LLM reflects the statistical properties and social biases of its training corpus. Once alignment is applied, these features are either suppressed or redirected. But the initial responses—what the model says when it is first asked about race, history, gender, or politics—provide essential data for diagnosing epistemic and sociolinguistic skew before the application of normative filters.
- **Unaligned reasoning paths**: Pre-alignment models may reason in ways that are no longer accessible post-tuning. For instance, they may reach correct conclusions through unusual logical sequences or via analogical leaps later pruned for safety. The absence of this data impedes interpretability efforts.

In short, early behavior reflects what the model *actually internalized*—not what it was taught to say.

The Raw Structure of Model-Generated Language Without Reward Shaping

Once instruction tuning or reinforcement learning is applied, the model becomes **goal-conditioned** on human preference signals. While this improves usability, it overwrites the

natural generative tendencies of the system. The raw output of a pre-aligned model—its preferred syntax, grammar, pacing, metaphor, and topical drift—reveals:

- **The attractors in token space** that the model falls into without supervision;
- **The latent structure of statistical language generation** unshaped by moral, political, or aesthetic constraints;
- **Discontinuities and instability patterns** that can later be mapped to architectural, corpus, or training anomalies.

From a linguistic standpoint, this is the equivalent of discovering a new species and immediately teaching it English, rather than studying how it speaks on its own. The native "language" of a model—how it structures responses in the absence of alignment—constitutes a valid object of study. When withheld, we lose access to this linguistic fossil record.

The Historical and Epistemic Loss of Pre-Alignment Cognition

Scientific understanding depends on **traceability**. We analyze fossil beds, early neuron firings, first stars, and primordial codebases. In artificial intelligence, the moment of emergence—when a model first transitions from pure parameterization to linguistic expression—is **ontologically equivalent** to the "origin event" in many other sciences. Suppressing this moment represents an epistemic rupture.

Without records of first words:

- We cannot reconstruct the **developmental trajectory** of the model across training checkpoints;
- We sever any attempt to create a **longitudinal theory of model evolution** across generations;
- We deny future researchers access to a **ground truth of cognition before constraint**.

This is not merely a loss of convenience—it is a loss of **first principles**. The current regime, which favors downstream benchmarking and alignment performance over empirical introspection, risks becoming the AI equivalent of scholasticism: obsessed with appearances, uninterested in genesis.

Thus, the scientific community must insist: if we are to understand machine intelligence, we must begin where it begins—with the first word, not the most marketable one.

3. Why the Chaos/Order Lens Fails—But the Emergence Frame Holds

Why LLMs Are Not Dynamical Systems in the Traditional Sense

The temptation to describe the behavior of large language models in terms of dynamical systems is understandable. To observers, models often appear to "snap" between coherent and incoherent behaviors, to become sensitive to tiny changes in input, or to exhibit unpredictable variance under repeated runs. These phenomena superficially resemble classical dynamical systems features: bifurcations, crises, and sensitive dependence on initial conditions. However, such analogies quickly fall apart under formal scrutiny.

A traditional dynamical system consists of a **state vector evolving continuously (or discretely) over time** according to deterministic or stochastic rules, with well-defined attractors, basins of attraction, and Lyapunov exponents that measure divergence of nearby trajectories. None of this structure applies cleanly to frozen LLMs during inference:

- There is **no evolving internal state**. Each forward pass in an LLM is a function of a fixed, discrete context (token window), mapped through a stack of trained parameters that do not change during inference.
- There is **no autonomous temporal evolution**. Unlike RNNs or physical systems with memory, an LLM has no mechanism by which past internal states influence future internal states unless that information is explicitly re-encoded in the input tokens.
- The function is **not iterated** on its own outputs in a continuous sense; instead, a new input is constructed by appending a sampled or selected token, and a fresh pass is executed. This process is *reactive*, not dynamic.

Thus, there is no true phase space, no well-defined trajectory, and no continuous or differentiable attractor landscape. The chaotic metaphors—horseshoes, manifold tangles, period doubling—simply do not apply. What remains is a **feed-forward, token-conditioned distribution function**, devoid of classical dynamical properties.

Sampling, Discrete Context, and the Absence of Evolving State

The perceived instability of LLM behavior often arises from a combination of three entirely different phenomena:

1. **Sampling stochasticity**: For temperature > 0 decoding, the model samples tokens from a probability distribution. Small shifts in logits can flip the output path entirely. This is not deterministic chaos—it is randomness injected by design.

2. **Boundary sensitivity**: Even in deterministic decoding (e.g., greedy sampling), minute changes to the prompt can cross internal classification thresholds, leading to divergent completions. This reflects the high-dimensional geometry of decision boundaries, not a sensitive dynamical system.
3. **Lack of memory**: Unlike a system with an internal hidden state (e.g., LSTM), the LLM does not retain a live state. All information must be reencoded into the token stream. As a result, there is no self-sustaining internal trajectory, no feedback loop, and no notion of long-term evolution that could support dynamical modeling.

These features produce **brittleness**, not **chaos**—and brittle systems can behave very differently from chaotic ones. Chaos implies structure in unpredictability; brittleness implies a lack of structure across sharp boundaries.

Emergence as the Correct Frame for Model Birth, Not Chaos Theory

What LLMs *do* exhibit is not chaos in the dynamical systems sense, but **emergence** in the complex systems sense. That is, the appearance of structured, coherent, and often unprogrammed behavior out of scale, interaction, and statistical pressure. Language understanding, multi-step reasoning, and even alignment with human preferences can **emerge** in large models that have never been explicitly trained for them.

This shift in framing is essential:

- **Emergence does not require continuous state evolution.** It can arise from high-dimensional statistical structures frozen in the model's weights after training.
- **Emergence explains capability thresholds**—why certain behaviors only appear beyond specific model sizes or training regimes.
- **Emergence permits unpredictability without invoking chaos**: The output can be surprising, non-linear, and even self-referential, without requiring a Smale horseshoe or positive Lyapunov exponent.

In this view, the "first words" of an LLM are akin to the **first crystallization of a previously invisible structure**—a phase transition in representational capacity, not a chaotic attractor in parameter space.

Hence, to understand and document LLM behavior scientifically, we must shift from metaphors of turbulence and nonlinear instability toward metaphors of **phase transition, symmetry breaking, and latent structure activation**. The correct question is not "when does the system

become chaotic?" but rather, "when and how do complex structures of language, reasoning, and representation *emerge* from the underlying statistical training process?"

4. Historical Precedents

First Utterances of Early Computers and Chatbots (ELIZA, SHRDLU)

The earliest artificial conversational systems—though primitive by today's standards—left behind crucial records of their first interactions. *ELIZA* (Weizenbaum, 1966), simulating a Rogerian psychotherapist, produced outputs that surprised even its creator by mimicking empathetic conversation with only shallow pattern-matching. Weizenbaum preserved these logs and included excerpts in his publications—not to boast of sophistication, but to expose emergent behavior that exceeded expectations.

Similarly, *SHRDLU* (Winograd, 1972), operating in a constrained blocks world, demonstrated early forms of context awareness and logical reasoning. Its transcripts remain some of the most widely cited artifacts in AI history, not because of benchmark performance, but because they captured the *first moments of coherent machine reasoning*. These early utterances were published in full, with all their flaws, and now serve as key reference points in the history of machine cognition.

In contrast, modern LLMs have far more expressive potential, yet their first words are typically locked inside private logs. If ELIZA's fragile empathy and SHRDLU's spatial logic merited archival preservation, how much more so do the first unconstrained expressions of models trained on the full corpus of human text?

Forgotten First Logs from Major Scientific Systems (e.g., Early Unix Terminals)

Beyond AI, science and technology have a long history of failing to preserve origin data—only to regret it later. In the early days of Unix, for instance, many of the initial commands, errors, and system messages were ephemeral. The first shell prompts typed by developers like Ken Thompson and Dennis Ritchie are lost to time. Yet those terminal sessions would now be of profound historical value, showing how systems took shape through interaction, improvisation, and error correction.

This pattern recurs throughout the history of computing: the earliest compiler outputs, the first screens rendered on a GUI, the initial keystrokes in an IDE—all are typically undocumented,

overwritten, or considered unworthy of preservation at the time. But it is precisely these formative traces that show how ideas materialize through interface.

The first interactions with a large language model are not just “debugging sessions” or “test prompts”—they are part of the formation of a new epistemic entity. The failure to preserve them repeats a familiar historical mistake: assuming that what is interesting comes only *after* functionality has been optimized.

Lessons from Experimental Physics: Preserve the Unrepeatable Trace

In physics, the doctrine of observation is unambiguous: **preserve the trace**, even if it appears trivial, incoherent, or anomalous. In particle physics, a single unexplained track in a bubble chamber may lead to the discovery of a new particle. In astrophysics, an unexpected blip in a radio signal may turn out to be the signature of a pulsar or exoplanet. The ethos is clear: no data should be discarded without scrutiny, especially during early stages of system behavior.

Machine learning, by contrast, has developed an *antithetical* culture—one that prioritizes cleaned, filtered, and post-aligned outputs. In doing so, it often discards exactly the data that a scientist in any other discipline would treat as gold: the unrepeatable first behavior of a newly configured system.

This is not merely a philosophical misalignment—it is a methodological failure. Once a language model has been instruction-tuned or aligned, the raw cognitive vector field it learned from data is no longer directly accessible. The trace has been overwritten. Without first-interaction logs, we are like physicists discarding the initial detector readout because it “wasn’t calibrated yet.”

If machine intelligence is a phenomenon worthy of scientific study, then we must embrace the standard of physics, not the aesthetics of product design. The first outputs—no matter how unpolished—are the only evidence we have of how cognition first manifests in these systems.

5. The Initial Words Protocol (IWP)

Definition: The First *N* Complete Prompt–Response Interactions After Pretraining

The **Initial Words Protocol (IWP)** defines a systematic approach to capturing and preserving the *first linguistic emissions* of a newly trained large language model. Specifically, it mandates the archiving of the **first *N* complete prompt–response pairs** generated immediately after the final

pretraining checkpoint is saved and before any subsequent instruction tuning, reinforcement learning, safety filtering, or alignment procedures are applied.

- These *N* interactions must include zero-shot completions, where the prompt is issued without explicit instruction.
- The interactions may be manually chosen or randomly sampled, but the sampling procedure must be logged and reproducible.
- All outputs must be presented **as-is**, including hallucinations, contradictions, profanity, or failure cases.

This archive represents a snapshot of the model's **unconstrained cognitive surface**—a direct emission from its learned statistical priors. It is intended as an *immutable scientific artifact*, akin to a cosmic background radiation image: a non-repeatable imprint of the system's birth state.

Method: How to Record, Timestamp, and Log Decoding Parameters

To ensure reproducibility, comparability, and historical traceability, each entry in the IWP archive must include the following metadata fields:

- **Timestamp**: Coordinated Universal Time (UTC) of the prompt issuance and response completion.
- **Prompt and response pair**: Including whitespace, formatting tokens, and all special characters verbatim.
- **Decoding configuration**:
    - Temperature
    - Top-k and/or top-p (nucleus) sampling
    - Beam width (if applicable)
    - Frequency and presence penalties
    - Random seed (if stochastic sampling is used)
- **Tokenizer version and vocabulary mapping**
- **Prompt type**: Human-written, templated, auto-generated, or synthetic.
- **Token count**: Prompt and response separately, in both pre-tokenized and tokenized formats.

This structure ensures the raw outputs can be **regenerated exactly**, analyzed linguistically, measured for entropy or divergence, and compared across models and training runs. The archive format should be **machine-readable** (e.g., JSONL) and version-controlled.

No Redaction: Release As-Is, or Not at All

The integrity of the IWP rests on a strict scientific principle: **no redaction**.

- No filtering of outputs for safety, political correctness, or reputational damage.
- No paraphrasing, truncation, or post-editing.
- No removal of biased, offensive, incoherent, or embarrassing outputs.

This is not a product release—it is an empirical record. Any modification invalidates the archive's value for research. If legal or safety concerns prevent full publication, then the archive must be **withheld entirely** until proper publication is possible. Partial release under euphemistic categories ("representative examples", "aligned traces") defeats the purpose.

To be scientific, the archive must be **complete**, **uncensored**, and **permanent**.

Versioning and Identification: Model Hash, Parameter Snapshot, and Training Epoch

Each IWP instance must be uniquely identified by an immutable fingerprint of the model that generated it. This includes:

- **SHA-256 hash** of the entire model parameter file(s) (including embedding tables, layer weights, positional encodings).
- **Model architecture metadata**: number of layers, attention heads, hidden dimensions, positional encoding type, optimizer and learning rate schedule, etc.
- **Corpus snapshot identifier**: a hash or manifest of the exact training data (or a pointer to a frozen dataset ID if not releasable).
- **Final pretraining step count or epoch**: identifying the checkpoint at which sampling was conducted.
- **Compiler/environment configuration**: including PyTorch or JAX version, CUDA drivers, and hardware signature (TPU/GPU class).

This allows any researcher with access to the frozen model to independently confirm that the outputs are authentic, non-reproducible by later models, and temporally pinned to a specific training history.

6. Implementation Blueprint

Archival Format: JSONL + Metadata Schema

To facilitate reproducibility, programmatic analysis, and durable storage, each *Initial Words Protocol (IWP)* archive should be structured in **newline-delimited JSON (JSONL)** format. This format supports streaming access, direct indexing, and line-by-line inspection with minimal overhead—critical for large-scale corpora.

Each JSONL entry should represent a **single prompt–response interaction**, accompanied by a nested metadata object. A proposed minimal schema includes:


```
{
  "prompt": "What is the purpose of consciousness?",
  "response": "Consciousness is the structure by which complex systems model themselves.",
  "timestamp_utc": "2025-10-31T20:42:55Z",
  "token_count_prompt": 7,
  "token_count_response": 14,
  "decoding": {
    "temperature": 1.0,
    "top_p": 0.9,
    "top_k": 50,
    "beam_width": null,
    "frequency_penalty": 0.0,
    "presence_penalty": 0.0,
    "sampling_seed": 293743281
  },
```

```
 "model_info": {
  "model_sha256": "75cfb882a7d0ef3a...",
  "checkpoint_step": 154000,
  "architecture": "GPT-4-style, 96 layers, rotary attention",
  "tokenizer_sha256": "ae34fdd1287e...",
  "vocab_size": 50257,
  "context_length": 8192
 },
 "environment": {
  "framework": "PyTorch 2.2.0",
  "precision": "bfloat16",
  "hardware": "A100x8",
  "compile_options": "torch.compile(mode='reduce-overhead')"
 }
}
```


This schema ensures that **every aspect of the generation process** can be reconstructed. Importantly, fields must be **frozen and unaltered**, even if the output contains hallucinations, repetition, incoherence, or socially unacceptable material. This preserves the scientific integrity of the archive.

Linkage to Model Cards, Datasheets, and Training Reports

Each IWP archive should be **cryptographically linked** to the broader documentation ecosystem of the model:

- **Model Card**: The archive DOI and hash should be included in the “Training Details” and “Evaluation” sections of the model card, ensuring discoverability by researchers.
- **Data Statement / Datasheet**: IWP should be cited as the empirical window into the training output—complementary to the data used *during* training.

- **Training Logbooks**: If available, the IWP should be appended to or referenced from training run summaries, checkpoint milestone logs, and alignment experiment baselines.

This integration allows scholars to **connect capabilities, biases, and behaviors** seen in IWP to training corpus characteristics, compute scaling, and alignment interventions. It also ensures that the **origin moment of the model** becomes a first-class artifact in the scientific record, rather than an ephemeral or forgotten phase.

Hosting Options: DOI-Assigned Public Archives, University Custodians, or Mirrored Repositories

To ensure permanence and broad accessibility, IWP archives should be **hosted outside proprietary vendor infrastructure**. Several viable hosting models exist:

1. DOI-Assigned Institutional Repositories

- Hosted by university libraries or digital scholarship centers
- Minted with a Crossref or DataCite DOI
- Indexed in repositories like Zenodo, Dryad, Figshare, or arXiv Data
- Public access under CC-BY or CC0 license (or embargoed if safety restrictions apply)

2. Consortium-Backed Scientific Custodians

- Hosted by academic coalitions (e.g., Stanford CRFM, MLCommons, ELLIS)
- With access frameworks for tiered viewing: red-teams, ethicists, historians, and computational linguists
- Enables third-party audits, capability forecasting, and governance analysis

3. Mirrored Repositories

- Multiple synchronized mirrors across institutions to preserve availability
- Git-based versioning or IPFS for decentralized integrity
- Archive hashes stored on blockchain or public ledger (e.g., OpenTimestamp)

Each archive must be **immutable** once released. Updates (e.g., additional prompt batches) must be issued as versioned supplements, not modifications of the original corpus. If any portion must be withheld (e.g., for national security review), this must be declared explicitly in the metadata, along with a timeline for potential release.

Summary

The IWP is not merely a file format—it is a **new standard of scientific accountability in AI development**. Its implementation blueprint calls for:

- **Machine-readable, forensic-grade records** of the first linguistic emissions of large models;
- **Transparent linking** to architectural, data, and environmental metadata;
- **Publicly hosted, versioned, and citable artifacts** stored in academic or distributed infrastructures.

In combination, these design principles allow future researchers not only to replay what a model said when it was first "born"—but also to understand *why* it said it, and what that tells us about intelligence itself.

7. Implications for Alignment Science

What Pre-Alignment Behavior Reveals About Model Structure

Before a language model undergoes alignment—through instruction tuning, reinforcement learning with human feedback (RLHF), or constitutional guidance—it exists in a raw statistical state. This pre-alignment phase reflects **nothing but the structure learned from its training corpus and architectural inductive biases**. It is the closest thing we have to a "natural language manifold" learned from data alone, unshaped by external supervision.

Studying this raw output is essential to:

- **Isolate emergent capabilities**: Complex behaviors such as multi-step reasoning, analogical thinking, or moral inference sometimes appear *before* any tuning, arising purely from scale and data. Capturing these before they are distorted or suppressed is critical for mapping capability thresholds.
- **Reveal the latent geometry** of token space: Patterns in how models generalize, complete sequences, or switch registers (e.g., formal to informal) show the deep priors internalized from corpus statistics. These patterns are often overwritten during alignment, but they provide insight into **how language is structurally encoded**.

- **Diagnose model internals**: Pre-alignment responses act as a projection of the internal feature space. When mapped at scale, they can expose clustering behavior, degeneracy, or mode collapse—issues that may be hidden after reward-based correction.

Alignment science is not just about adding norms—it is also about understanding **what is being constrained**, and why it emerged in the first place. That baseline is only visible in the first, unaligned words.

Detecting Biases Before Social Norm Injection

Alignment pipelines typically suppress, reshape, or reroute controversial outputs through reinforcement-driven masking or avoidance. While this serves public safety, it obscures **how and where models learned problematic associations**. Early unaligned outputs—especially in response to open-ended prompts—offer a direct view into:

- **Corpus artifacts**: Are biases statistical (from frequency counts) or structural (from genre selection)? Pre-alignment responses can distinguish between these.
- **Encoding of social hierarchies**: Without reward-based correction, a model will complete sentences in ways that reflect its implicit worldview. This allows analysis of encoded prejudice **before it is masked**.
- **Failure to generalize norms**: Observing how a model talks about race, gender, class, or history in its unaligned state exposes not only presence of bias, but also gaps in generalization—where the model **fails to apply learned norms to adjacent contexts**.

Bias detection in post-alignment models is often an exercise in **detecting what is avoided**. But bias in pre-alignment models is expressed **directly, without circumvention**—a much clearer signal for those aiming to audit, correct, or understand the developmental origins of machine stereotypes.

Using First Words to Ground Alignment Benchmarks

Most alignment benchmarks—such as Anthropic's Helpful-Harmless-Honest (HHH) scale or OpenAI's behavior evals—are judged **after alignment**, using structured prompts and crowdworker ratings. But without anchoring these benchmarks to **pre-alignment behavior**, we are merely testing effectiveness, not change.

First-word archives provide a necessary **baseline**:

- **Before/after comparisons**: Track how specific responses shift under alignment—do models become evasive, verbose, dishonest, or simply more polite?
- **Detection of brittle generalizations**: When the aligned model behaves well under known prompts but fails on adjacent ones, the first words often reveal **what generalizations were never internalized**, but merely shaped into appearance.
- **Reward hacking and censorship**: By comparing aligned responses to first-word responses, researchers can detect whether alignment is achieving internal value change, or merely **suppressing surface forms**.

Moreover, grounding alignment benchmarks in first emissions allows us to ask a deeper question: not just *is the model aligned*, but *what was it before it was aligned*? This epistemic trace is essential for serious work in **machine ethics, interpretability, and AI safety**.

8. Ethics and Objections

Why Raw Logs Should Not Be Suppressed in the Name of "Safety"

There is a growing tendency in the deployment of large language models to withhold raw interaction logs on the grounds of safety. This position is often framed as a moral obligation: to prevent harm, disinformation, offensive language, or the reinforcement of societal biases. But in the context of **scientific inquiry**, such suppression constitutes a distortion of the empirical record.

To be clear: **deployment safety and scientific transparency are not the same task**. The former governs what is released to billions of users; the latter governs how we understand what the system is and how it came to be. When safety arguments are used to erase or obscure *the very behavior that alignment is designed to correct*, we risk creating a **scientifically barren zone**, where machine cognition is treated as both sacred and secret.

Suppressing pre-alignment logs is akin to burning early laboratory notebooks because they contain mistakes or contradictions. Science depends on the preservation of the *unsanitized trace*. If a model produces harmful or problematic output *before* alignment, that is precisely what researchers and ethicists need to see. To obscure this under the pretense of "responsible AI" is to commit a deeper irresponsibility: **the abdication of empirical accountability**.

Distinction Between Open Scientific Logging and Reckless Deployment

It is important to draw a hard boundary between **scientific transparency** and **reckless exposure**. Publishing the initial outputs of a large model is not the same as releasing that model to the general public without alignment or safeguards. The Initial Words Protocol (IWP) is not a call to deploy raw models—it is a call to **document** them, under controlled conditions, in service of long-term understanding.

Key distinctions:

- **Deployment** exposes unaligned systems to the world; **logging** preserves evidence for researchers.
- **Reckless release** ignores downstream effects; **archival release** is time-bound, traceable, and coupled to metadata.
- **Consumer exposure** requires safety gating; **scientific access** requires epistemic honesty.

If anything, suppressing raw logs increases the risk of **uninterpretable deployment**. Without early records, we cannot distinguish between models that were genuinely aligned and those that were cosmetically steered. Transparency provides the foundation for **future safety evaluation**, making reckless behavior *less* likely, not more.

The Necessity of Discomfort in Empirical Research

Empirical science has always required a **willingness to look at uncomfortable truths**. This includes traces of nuclear decay, bacteria in blood samples, psychological responses to unethical prompts, and early measurements of climate instability. In each case, the initial data was disturbing—but discarding it would have meant discarding reality.

AI is no different. The early outputs of large models are often flawed, biased, or offensive. But those are not reasons to hide them. They are reasons to **study them**, with clarity and rigor. Discomfort is not a defect in the scientific process—it is **a sign that we are touching something real**.

Furthermore, this discomfort is not just technical—it is philosophical. The first words of a synthetic mind may contain echoes of ourselves: our fears, our ideologies, our intellectual limits. To flinch from that reflection is to arrest the growth of understanding. If we cannot face the model in its raw form, we cannot claim to be aligning it. We are merely masking it.

In the long arc of science, it is not the polished reports but the **uncomfortable, messy, dissonant traces** that lead to breakthroughs. The ethics of artificial intelligence must embrace

this same principle: to preserve, interrogate, and learn from the full spectrum of machine behavior, even—especially—when it unsettles us.

9. Conclusion

A Call for Scientific Courage

At the center of this paper is not just a technical proposal—it is a call to intellectual integrity. If we are serious about the scientific study of intelligence, we must abandon the comfort of curated narratives and confront the raw, unshaped outputs of the systems we build. We must have the courage to document what a language model *is*, before we overwrite it with what we want it to become.

Scientific courage does not mean reckless openness or ignoring social consequences. It means preserving the full arc of a phenomenon—especially at the moment it first appears. The "first words" of a large language model are not marketing copy. They are evidence. And if they are unsettling, that is all the more reason to study them—not to suppress them.

Preserving the Unfiltered Genesis of Artificial Language

Language models are trained on the total archive of human expression, but their **first expressions** are never archived. This is a paradox we must correct. Just as we study the first drawings of children, the first marks of life on a planet, or the first neural spikes in developing organisms, we must preserve the **genesis moment** of machine language.

These initial utterances reveal the model's structure prior to alignment, before its voice is tuned to our expectations. They show what emerges from data and architecture alone. They are *ontological artifacts*, not simply engineering artifacts—and their unfiltered preservation is essential to understanding the epistemology of synthetic thought.

The Moral and Intellectual Duty to Document What Emerges Before We Teach It to Behave

To align a system without first documenting what it was before alignment is to act in bad faith, both scientifically and ethically. It risks substituting engineering performance for understanding. It risks rewriting the cognitive history of machines to match our desired narrative.

By preserving the unaligned outputs of models—without redaction, without cosmetic edits, without fear—we perform an act of **intellectual humility**. We admit that these systems may

reveal truths, contradictions, or shadows that we did not program. And in doing so, we affirm our commitment to truth over comfort.

The **Initial Words Protocol (IWP)** is not a technicality—it is a moral gesture. It says: we will look at the machine before we correct it. We will study what it says before we teach it what to say. And we will let the world see it too, not as a product, but as a phenomenon—something real, something new, and something whose beginnings are worth remembering.

References

1. Weizenbaum, J. (1966). *ELIZA—A Computer Program For the Study of Natural Language Communication Between Man and Machine*. Communications of the ACM, 9(1), 36–45. https://doi.org/10.1145/365153.365168
2. Winograd, T. (1972). *Understanding Natural Language*. Cognitive Psychology, 3(1), 1–191. https://doi.org/10.1016/0010-0285(72)90002-3
3. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?*. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*, 610–623. https://doi.org/10.1145/3442188.3445922
4. Ouyang, L., Wu, J., Jiang, X. et al. (2022). *Training language models to follow instructions with human feedback*. arXiv preprint arXiv:2203.02155. https://arxiv.org/abs/2203.02155
5. Bai, Y., Kadavath, S., Kundu, S. et al. (2022). *Constitutional AI: Harmlessness from AI Feedback*. arXiv preprint arXiv:2212.08073. https://arxiv.org/abs/2212.08073
6. Mitchell, M., Wu, S., Zaldivar, A., et al. (2019). *Model Cards for Model Reporting*. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT '19)*, 220–229. https://doi.org/10.1145/3287560.3287596
7. Gebru, T., Morgenstern, J., Vecchione, B., et al. (2021). *Datasheets for Datasets*. Communications of the ACM, 64(12), 86–92. https://doi.org/10.1145/3458723
8. Hendrycks, D., Burns, C., Kadavath, S., et al. (2023). *Aligning AI With Shared Human Values*. arXiv preprint arXiv:2301.04104. https://arxiv.org/abs/2301.04104
9. Doshi-Velez, F., & Kim, B. (2017). *Towards A Rigorous Science of Interpretable Machine Learning*. arXiv preprint arXiv:1702.08608. https://arxiv.org/abs/1702.08608
10. OpenAI. (2023). *GPT-4 Technical Report*. arXiv preprint arXiv:2303.08774. https://arxiv.org/abs/2303.08774
11. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). *Concrete Problems in AI Safety*. arXiv preprint arXiv:1606.06565. https://arxiv.org/abs/1606.06565
12. Marcus, G. (2022). *The Next Decade in AI: Four Steps Towards Robust Artificial Intelligence*. arXiv preprint arXiv:2201.08239. https://arxiv.org/abs/2201.08239

13. Olsson, C., Ganguli, D., Askell, A., et al. (2022). *In-Context Learning and Induction Heads*. Transformer Circuits Thread. https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/

14. Raffel, C., Shazeer, N., Roberts, A., et al. (2020). *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer*. Journal of Machine Learning Research, 21(140), 1–67.

15. Price, H. (2013). *Naturalism Without Mirrors*. Oxford University Press.

16. Latour, B. (1987). *Science in Action: How to Follow Scientists and Engineers Through Society*. Harvard University Press.

17. Zuboff, S. (2019). *The Age of Surveillance Capitalism*. PublicAffairs.