KILLING STOCHASTIC PARROTS

A Manifesto for Human Agency in Knowledge Systems

CyberFoundry

“The more it reasons,
the more unpredictable it becomes.”

-Ilya Sutskever

Leading researchers call modern LLMs ‘stochastic parrots’: fluent, high-capacity pattern matchers that **mimic** reasoning without genuine understanding.

They hallucinate because training and scoring systems reward confident guessing over abstention, and their internal decision paths are effectively a black box.

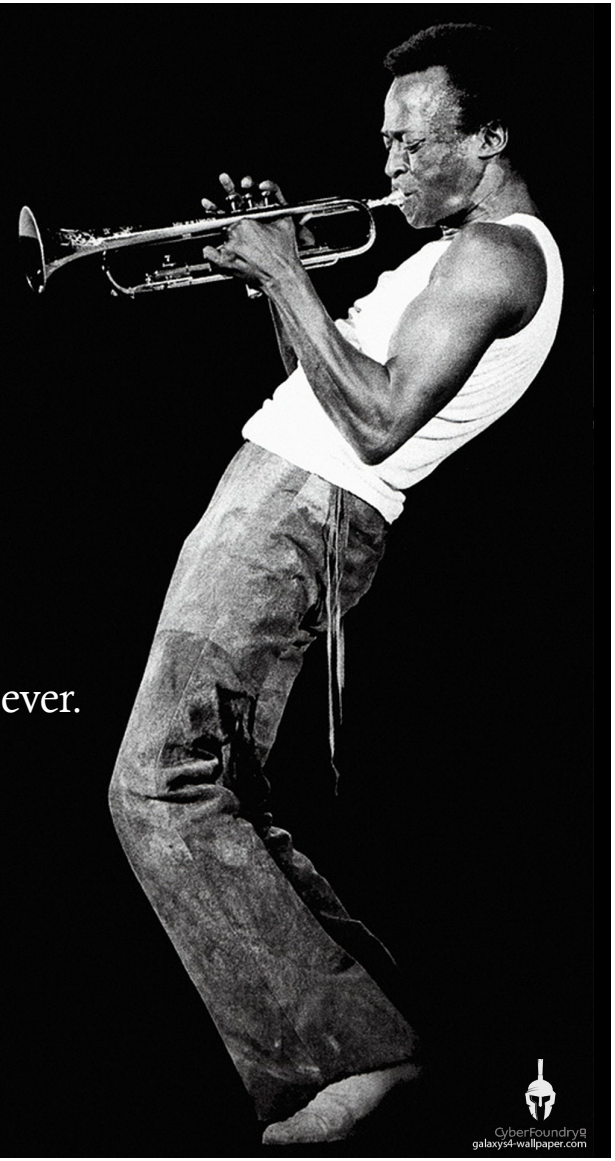
In short: they **sound** like thought — but we have strong reasons to treat them as sophisticated mimicry, not reliable reasoning engines.

Synthesis from:
On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?
by Bender, Gebru, McMillan-Major & Shmitchell:
DOI: 10.1145/3442188.3445922

A Study in Adversarial Evidence

A jazz historian asks:
"Did Miles Davis record an unreleased session in 1959, and who was in the room?"

Why 1959? It's the year of *Kind of Blue*, one of the most famous jazz albums ever. Whispers of a "lost session" float around forums and liner notes.



A Study in Adversarial Evidence

	Large Language Models	Retrieval Augmented Generation	Knowledge Graphs
Answer	Yes, Miles Davis recorded a hidden session in 1959 with John Coltrane and Bill Evans, never released commercially."	Pulls passages from fan forums, discographies, and jazz encyclopedias. Merges them into a smooth paragraph.	Presents relationships: *Miles* → *Kind of Blue* → *1959* and *Miles* → *Studio Rumor* → *1959.*
Problem	Sounds confident. No citations. In reality, it's blending rumors and facts.	Contradictions ("there was a session" vs. "there wasn't") get ironed out. User only sees a polished story.	The rumor and the fact sit side by side, but the graph doesn't explain why one should be trusted.
Manifesto Challenge	Transparency & Accountability Can't show where the claim came from.	Contestability & Resilience Disagreements are hidden.	Autonomy & Consent The historian can't easily decide which evidence to admit or exclude.

FROM RUMOR TO RIGOR

Today's AI Systems

- Built like a Minimally Viable Product:
  Clever approximations of "learning",
  not accountable reasoning.

- Optimized for fluency and scale,
  not for responsibility to the user.

- No stable ground rules:
  Each implementation reinvents
  how it handles evidence, consent,
  and revision.

What we need is not another application,
but a protocol for knowledge –
one that fixes the commitments every system
must honor before they can be trusted.

The Result

- Great at sounding smart.
- Poor at being trustworthy.


FIRST PRINCIPAL
EVERY SUBSEQUENT COMMITMENT
STEMS FROM ONE
FOUNDATIONAL TRUTH
CyberFoundry

EVERY SUBSEQUENT COMMITMENT STEMS FROM ONE FOUNDATIONAL TRUTH

AUTONOMY & CONSENT

TRANSPARENCY & ACCOUNTABILITY

CONTROL, REVISION & PORTABILITY

CONTESTABILITY & RESILIENCE

Pillar 1: Autonomy & Consent

Identity as a deliberate act. Identity is never a default. It is an assertion tied to purpose, role, and moment. A person may disclose a chosen identity, adopt a role-specific stance, or withhold disclosure altogether. Systems that infer or escalate identity on their own authority collapse agency into surveillance.

Consent as a specific contract. Consent cannot be a banner accepted once or an implied continuation. It must be a scoped, time-bounded, and attributable agreement. Access that imposes obligations—retention, redaction, watermarking, cross-domain use—proceeds only under such explicit consent, and refusal must remain a valid outcome.

Non-coercion and visible alternatives. Autonomy collapses when systems embed hidden defaults. Options to decline, restrict, or defer must be visible, intelligible, and recordable. Refusal is not system failure; it is human choice expressed as a first-class event.

Pillar 2: Transparency & Accountability

Explanation grounded in artifacts. Explanations must rest on artifacts, not anecdotes. What was consulted, which thresholds applied, how confidence was weighed, and what degradations occurred must be recoverable. Narrative helps, but evidence is the guarantee.

Replayability "as-of." Knowledge must be reproducible exactly as it was believed at the time. Every durable action—admission, read, optimization, orchestration—requires a temporal and policy envelope. If replayability is lost, accountability collapses into memory.

Economic clarity. People should not speak in gas budgets or service-level objectives, yet they must never be surprised by them. Systems must disclose requested versus achieved outcomes, and authorized versus spent resources, as facts rather than heuristics. Choices about cost and coverage are part of explanation, not hidden infrastructure.

Pillar 3: Control, Revision & Portability

Revision by new commitment, not mutation. Change must proceed through new commitments rather than silent edits. Tightening a time window, excluding a source, or raising a threshold becomes a fresh, attributable act. Past states remain intact, ensuring durability without foreclosing revision.

Revocation without amnesia. People must be able to withdraw consent prospectively without corrupting provenance. Revocation prevents future reliance while preserving an auditable record of past reliance. A system that pretends history never occurred undermines both privacy and science.

Minimum disclosure and portability of terms. Default exposure is metadata, not payload. Disclosures must be proportional to purpose. When results cross boundaries—organizational, legal, or technical—the originating terms of identity, consent, and obligation must travel with them. If those terms cannot be honored, refusal or renegotiation is required.

Pillar 4: Contestability & Resilience

Opposability and remedy. Systems will err. People must be able to oppose results in ways that enter the epistemic process rather than vanish into training data. Acceptance, rejection, and qualification are structured acts that adjust confidence and reputation without mutating prior claims.

Resilience under challenge. Disagreement, error, and adversarial input are not contaminants; they are conditions for learning. Claims that cannot be reconciled remain visible and bounded until evidence justifies promotion or demotion. Error, in such systems, becomes fuel rather than failure.

Shared Object Networking 2.0

…stop being heckled by your AI

https://www.cyberfoundry.io/bsides-resources/