The artificialness problem in AI-generated text: Toward memory-driven, intuition-based language models

The fundamental challenge facing large language models is not their inability to produce fluent text, but their **inability to think like humans do**—spontaneously, associatively, and imperfectly. This research synthesis reveals that LLM outputs are detectable precisely because they are *too* optimized: too coherent, too on-topic, too uniform. The path toward genuine naturalness lies not in better optimization, but in architectures that replicate the beneficial "imperfections" of human cognition—forgetting, associative memory triggering, reconstructive recall, and non-goal-directed thought.

Why AI text feels artificial: Detection signatures and optimization artifacts

Current AI detection systems achieve **96.5%+ accuracy** by exploiting a fundamental property: AI-generated text exhibits significantly higher stylistic uniformity than human writing. A 2025 Nature study found that LLM outputs "display a higher degree of stylistic uniformity, clustering tightly by model," while human texts form "broader, more heterogeneous clusters, reflecting the diversity of individual expression." [nature] [Nature] Critically, newer models like GPT-4 show *tighter* clustering than GPT-3.5—sophistication increases predictability. [nature]

Detection methods exploit two core metrics: **perplexity** (how "surprising" text is to a language model) and **burstiness** (variation in sentence complexity). AI text shows lower perplexity (more predictable word choices) and lower burstiness (uniform sentence lengths). [QuillBot] [arxiv] Human writing exhibits the "rhythm of short and long phrases, mixing up both simple and complex sentences." [QuillBot] The mathematical signature is stark: AI text occupies **negative curvature regions** of a model's log probability function—a property exploited by Stanford's DetectGPT, which achieves **0.95 AUROC** on fake news detection. [arXiv]

Beyond statistical signatures, AI text reveals itself through what it lacks: natural tangents, subjective opinions, personality. Studies show humans identify AI by "overly formal tone," "statements too objective," and excessive hedging phrases like "it's worth noting." [arxiv] The **Uniform Information Density (UID) hypothesis** reveals another tell: while humans distribute information uniformly during production, AI varies information density differently—the variance in word probability becomes a discriminating feature. [arXiv] Perhaps most revealingly, AI text clusters same-entity mentions closer together, while humans coherently refer back to entities even with long separations—a pattern extractable as graph structure for detection. [arXiv] [arxiv]

The single-minded pattern adherence problem

The "single-minded pattern adherence" described in the research question—being exactly as elaborate or concise as prompted—emerges directly from how models are trained. RLHF optimization creates what researchers call **mode collapse**: a dramatic reduction in output diversity where the model converges on a single "optimal" response style. [arXiv] [arXiv] Kirk et al. (2023) provided the first rigorous demonstration that "RLHF significantly reduces output diversity compared to SFT across all diversity metrics (lexical, semantic, logical)." [arxiv] Successive GPT-3 versions showed "increasing degrees of mode collapse whereby overfitting the model during alignment constrains it from generalizing over authorship." [arXiv]

The mathematical explanation is stark: the RL loss function $-\sum t\ rt = -R$ makes "putting all sequence-level probability mass on the highest reward point an optimal solution." Even when multiple responses have identical reward, concentrating probability on just one is mathematically optimal. arXiv arXiv This isn't a bug—it's a feature of the objective function.

Mixture of Experts and optimization: Architectural sources of artificialness

MoE architectures introduce a counterintuitive source of uniformity: **context-independent specialization**. Research from OpenMoE reveals that "the dynamics of token routing are primarily dictated by the token ID. The same token will almost always be routed to the same expert, no matter the context." Substack Experts specialize on syntactic patterns (punctuation, indentation, common tokens) rather than semantic content. Hugging Face This creates mechanical consistency—responses are assembled from specialized but context-blind components.

DeepSeekMoE's analysis identifies two fundamental MoE limitations: **knowledge hybridity** (limited experts must handle diverse, unrelated knowledge types) and **knowledge redundancy** (different experts acquire similar knowledge, wasting parameters). arXiv The result is responses that feel assembled rather than organically generated.

The optimization process itself eliminates what makes human communication distinctive. Cross-entropy loss maximizes accuracy globally without structural awareness—it doesn't "consider underlying structures or differences between classes." arXiv ACL Anthology Alternative losses (Focal, Lovász) achieve **+42% improvement on exact match** for certain tasks, suggesting standard objectives fundamentally limit expressiveness. arXiv Adjusting KL penalty coefficients during RLHF doesn't recover diversity—"increasing the KL penalty coefficient leads to a drop in performance as expected, but also to a drop in per-input diversity, rather than a gain." arXiv arxiv

Promising counter-approaches

Diverse Preference Optimization (DivPO) from Meta/NYU/ETH offers a path forward: instead of selecting highest-rewarded responses as training targets, select the "most diverse response meeting quality threshold." arXiv Results show **45.6% more diverse persona attributes**, **74.6% increase in story diversity**, and 2.4% winrate improvement over standard DPO. ADS The insight is simple: optimization must explicitly reward diversity, not just quality.

Human cognition as the blueprint: Memory, intuition, and beneficial imperfection

The most profound finding from cognitive science is that human "imperfections" are adaptive features, not bugs. Understanding these mechanisms provides the blueprint for more natural AI.

Spreading activation versus input-driven retrieval

Human semantic memory operates through **spreading activation**: when a concept is accessed, activation propagates automatically through associative networks to related concepts. "Water" may trigger "whale" even when asked about mammals—without explicit input requesting this connection. This explains priming effects (faster responses to "doctor" after "nurse") and the spontaneous associations that characterize human thought. LLMs are purely input-driven; they don't exhibit autonomous activation between queries.

Reconstructive memory versus precise retrieval

Sir Frederic Bartlett's foundational research established that human memories are not passive recordings but **active reconstructions** combining stored fragments with existing schemas. His "War of the Ghosts" study showed participants changing unfamiliar elements to match cultural expectations (canoes became boats, Native American concepts were rationalized). Memories undergo "leveling" (removing unfamiliar details) and "sharpening" (exaggerating salient ones).

This "lossy" memory enables superior generalization. Humans remember **gist** and reconstruct details—leading to both creative synthesis and systematic distortions that LLMs don't exhibit. LLMs retrieve patterns precisely, missing the creative recombination that emerges from imperfect recall.

Sleep, consolidation, and abstraction

Human memory consolidation during sleep transforms specific experiences into abstract knowledge. Research shows memories are "repeatedly replayed during slow-wave sleep," driving a "transformation process" where hippocampus-dependent episodic memories become "schema-like neocortical representations." One year after encoding, subjects showed significant gist knowledge only if they slept immediately after learning.

This compression-during-consolidation—extracting patterns while discarding specifics—is fundamentally different from LLM training, which preserves raw data relationships. The compression itself may be the source of human generalization ability.

Beneficial forgetting: The Ebbinghaus curve as feature

The Ebbinghaus forgetting curve (50% loss within an hour, 70% within 24 hours) is typically framed as limitation. But research identifies three adaptive functions: **emotion regulation** (limiting access to negative memories), **knowledge abstraction** (forgetting specifics enables generalization), and **context sensitivity** (ensuring knowledge remains current). William James noted: "If we remembered everything, we should on most occasions be as ill off as if we remembered nothing."

Intuition as compressed pattern recognition

Gary Klein's naturalistic decision-making research reveals that expert intuition is "nothing more and nothing less than recognition" (Herbert Simon)—pattern matching from compressed experience. His Recognition-Primed Decision model shows experts recognize patterns, simulate outcomes, and act without consciously comparing options. **90% of critical decisions** by firefighters and military commanders are intuition-based.

Kahneman's dual-process framework distinguishes **System 1** (fast, automatic, pattern-based) from **System 2** (slow, deliberate, analytical). F'inn Insights Current LLMs exhibit System 1-like behavior (rapid pattern recognition) but struggle with System 2 tasks requiring multi-step logical reasoning on novel problems. F'inn Insights

Mind-wandering and the Default Mode Network

Humans spend approximately **47% of waking hours mind-wandering**. Far from being unproductive, this spontaneous thought—mediated by the Default Mode Network—facilitates creative incubation. Research with 2,433 participants found creativity predicted by switches between the DMN and Executive Control Network—spontaneous thought generation followed by controlled selection.

This "non-goal-directed thinking" enables serendipitous connections impossible in purely goal-directed systems. LLMs lack any equivalent to rumination, incubation, or spontaneous processing between queries.

Memory-augmented architectures: Toward dynamic recall

Foundational MANNs: Neural Turing Machines and Differentiable Neural Computers

The **Neural Turing Machine** (Graves et al., 2014) introduced differentiable external memory with content-based and location-based addressing. DeepMind's **Differentiable Neural Computer** (2016) extended this with memory attention mechanisms, temporal attention for sequential reasoning, and dynamic memory allocation. DNCs demonstrated graph traversal, family tree reasoning, and block puzzle solving.

However, classical MANNs face limitations: soft read/write operations access every memory entry (creating bottlenecks), scaling to very large memory is difficult, and controller structure selection remains ad-hoc. Nature

Modern memory-augmented transformers

Memorizing Transformers (Wu et al., 2022) use non-differentiable kNN lookup into cached key-value pairs, supporting memory sizes up to **262K tokens** with performance comparable to 5x larger vanilla transformers. The approach can be added to pre-trained models via fine-tuning.

MemoryLLM (Wang et al., 2024, ICML) introduces self-updatable memory with a fixed-size memory pool within the transformer latent space. Built on Llama2-7B, it maintains operational integrity after approximately 1 million memory updates—enabling something closer to continuous learning.

HippoRAG: Neurobiologically-inspired retrieval

HippoRAG (Gutiérrez et al., 2024, NeurIPS) mimics hippocampal indexing theory with three components: an LLM as "artificial neocortex," a retrieval encoder as "parahippocampal region," and a schema-less knowledge graph as "hippocampal index." During retrieval, Personalized PageRank spreads activation through the graph—mimicking neural spreading activation. Results show **up to 20% improvement** on multi-hop QA at 10-30x lower cost than iterative retrieval.

HippoRAG addresses a fundamental RAG limitation: standard retrieval encodes passages independently, missing cross-document connections. RAG cannot identify that "Prof. Thomas" is both "Stanford professor" AND "Alzheimer's researcher" unless a single passage mentions both—human associative memory handles this trivially.

Cognitive architecture frameworks

CoALA (Sumers et al., 2023, TMLR) organizes agents with working memory, episodic memory (past events), semantic memory (factual knowledge), and procedural memory (skills). The framework demonstrates that GPT-3.5 improved from **48% to 95%** on coding benchmarks when enhanced with cognitive architecture.

Letta/MemGPT (Packer et al., 2023) treats context windows as constrained memory resources analogous to OS virtual memory. Letta The architecture features self-editing memory via tool calls, with the LLM autonomously deciding what to keep, discard, or store. Serokell Core memory holds "persona" and "user information" that the agent actively updates. Letta

---

Practical techniques for human-like generation

Sampling strategies beyond greedy decoding

Nucleus (top-p) sampling (Holtzman et al., 2019) avoids "neural text degeneration" by sampling from the smallest token set whose cumulative probability exceeds threshold p. With p ∈ [0.9, 1), outputs closely match human text self-similarity patterns.

Mirostat (Basu et al., 2021) directly controls perplexity through feedback-based adaptive top-k. arXiv It identifies the "boredom trap" (low k values cause repetition) and "confusion trap" (high k causes incoherence), targeting **perplexity ~3-4 bits/token** for human-like text.

Min-p sampling (ICLR 2025) uses the top token's probability as a scaling factor for dynamic truncation, improving both quality and diversity at higher temperatures. Human evaluations show clear preference for min-p in quality and creativity.

Latent reasoning: COCONUT

Meta FAIR's **COCONUT** (Chain of Continuous Thought) represents a paradigm shift: reasoning in continuous latent space instead of language space. OpenReview The last hidden state becomes a "continuous thought" fed back as the next input—reasoning happens in high-dimensional space before being translated to language. Medium

Results are striking: on ProsQA (planning-intensive), Coconut achieves **97.0%** accuracy versus CoT's 77.5%. The continuous thoughts encode multiple alternative next steps—enabling breadth-first search in reasoning. This aligns with neuroimaging findings that "the language network remains largely inactive during reasoning tasks." arxiv

Diversity interventions

Diversity of Thought (arXiv:2310.07088) demonstrates that token-level diversity (temperature) doesn't ensure diverse solution approaches—explicit thought/method-level diversity is needed. The Div-Se method automatically improves prompt diversity by soliciting LLM feedback on solution approaches.

Multi-agent debate with diverse models outperforms single powerful models: after 4 rounds, diverse medium-capacity models (Gemini-Pro, Mixtral 7B×8, PaLM 2-M) achieved **91%** on GSM-8K versus single-model GPT-

4's 82%.

Key researcher perspectives on fundamental limitations

François Chollet: LLMs as interpolative memory stores

Chollet argues LLMs are "100% memorization"—"big interpolative databases" performing "program fetching" rather than reasoning: "When you give them a new puzzle, they can just fetch the appropriate program and apply it. It looks like reasoning but it's not really doing any sort of on-the-fly program synthesis."

His ARC-AGI benchmark measures "skill acquisition efficiency" on unknown tasks—true intelligence handles novelty. ARC Prize Pre-o3 AI reached ~31% versus human ~80%. Lab42 OpenAI's o3 achieved 87.5% but at massive computational cost (tens of millions of tokens per task), suggesting brute-force search rather than efficient intelligence.

Critically, Chollet notes the inverse relationship between memorization and generalization: "The best learners that we are aware of, which are children, are extremely bad at recollecting information... But you're extremely good at picking up new languages and learning from the world." Dwarkesh Podcast

Andrej Karpathy: The anterograde amnesia problem

Karpathy describes LLMs as suffering "anterograde amnesia"—brilliant but unable to form new memories: "LLMs are a bit like a coworker with Anterograde amnesia - they don't consolidate or build long-running knowledge or expertise once training is over." They are "perpetual interns on their first day." The Neuron

His insight on memory as regularization is profound: "If they had less knowledge or less memory, maybe they would be better... maybe the inability to memorize is a kind of regularization." Killerstorm This suggests the path to better generalization may involve *constraining* rather than expanding memory capacity.

Anthropic: Mechanistic interpretability

Anthropic's interpretability research (2024-2025) reveals that individual neurons are polysemantic but linear combinations ("features") can represent single concepts. "Circuit tracing lets us watch Claude think, uncovering a shared conceptual space where reasoning happens before being translated into language." This suggests internal representations may be more "thought-like" than outputs indicate—the challenge is accessing this space directly.

Cognitive signatures: How unique memory creates individual voices

Human brains develop unique connectivity patterns ("brain fingerprints") that change **~13% every 100 days**, shaped by life experiences. Even identical twins share only ~12% of structural connectivity. Cognitive fingerprints appear in simple tasks: researchers identified authors from 300 pseudo-random digits with **96.5% AUC**—individual pattern preferences remained constant over one week.

The **Self-Memory System** model shows autobiographical memory has a bidirectional relationship with identity: current self-views influence recollection while memories shape self-views. Through autobiographical reasoning, individuals link specific memories to form coherent life stories. This has direct implications for AI personalization: agents with user-specific episodic memory could develop unique "voices" analogous to human cognitive signatures.

Frameworks like CoALA and Letta enable this through hierarchical memory with self-editing capabilities. [Letta] Unlike static RAG, these agents actively maintain memory—mirroring human memory consolidation. The distinction matters: RAG retrieves similar chunks while persistent memory blocks hold evolving "executive summaries" of user relationships. [Letta]

Proposed benchmark: Measuring humanness in educational content

Based on this research, a benchmark for AI text naturalness in educational contexts ("EduHumanBench") should incorporate:

Statistical measures:

- Perplexity variance across text spans (human text varies more) [Medium]
- Burstiness (sentence length variation; human writing shows mixed rhythms) [QuillBot]
- Entity reference patterns (human texts reference entities at longer ranges)
- Information density uniformity (humans distribute information more evenly in production)
- Cross-entropy at token level (human text is less predictable) [Medium]

Structural markers:

- Tangent frequency (human explanations include natural digressions)
- Hedging phrase frequency (AI overuses "it's worth noting," "importantly")
- Subjective statement ratio (humans include opinions; AI defaults to objectivity)
- Register variation (humans adapt formality; AI defaults to formal)
- Contraction usage (humans use contractions; AI avoids them)

Semantic features:

- Semantic coherence trajectory (how coherence evolves through a document)
- Associative jump diversity (how far concepts connect across text)
- KL-divergence on topic models (human topic transitions are more varied)
- Embedding trajectory complexity (human text takes less direct paths through embedding space)

Human evaluation protocols:

- Paired comparison (human vs. AI on identical topics)
- Attribute identification (what markers led to classification)
- Domain expert review (subject specialists evaluate accuracy-naturalness tradeoffs)

Technical recommendations for PyTorch implementation

For implementing more human-like text generation:

1. **Sampling layer modifications**: Implement min-p or Mirostat sampling with target perplexity of 3-4 bits/token. Add burstiness targets to maintain sentence-length variance within human-like ranges.
2. **Memory architecture**: Build hierarchical memory following Letta patterns—core context (persona + user knowledge), archival storage (vector DB for long-term facts), recall memory (conversation history with selective retrieval).
3. **Diversity-aware fine-tuning**: Implement DivPO-style preference selection, choosing most diverse responses meeting quality thresholds rather than highest-scoring responses.
4. **Latent reasoning integration**: Explore COCONUT-style continuous thought chains for reasoning-heavy educational content, allowing high-dimensional exploration before language production.
5. **Controlled imperfection injection**: Introduce stochastic tangent triggers, occasional informality injection, and vary information density following human production patterns.
6. **Spreading activation simulation**: For knowledge retrieval, implement Personalized PageRank over concept graphs (HippoRAG-style) to enable associative recall rather than exact-match retrieval.

Synthesis: The fundamental tension and path forward

The research reveals a fundamental tension: optimization for quality eliminates the variability that characterizes human text. Mode collapse is not a bug but an optimal solution to the objective function. arXiv The uniformity that makes AI detectable is the same property that makes it helpful—consistent, on-topic, thorough responses. arxiv

The path forward requires reconceptualizing the goal. Rather than optimizing for "best" responses, we must optimize for **human-like diversity of response**—including beneficial imperfections. This means:

- **Memory systems that compress and abstract** rather than storing raw patterns
- **Retrieval that spreads associatively** rather than matching exactly
- **Objectives that reward diversity** explicitly, not just quality
- **Architecture that enables spontaneous activation** independent of input queries

- **Controlled forgetting** that prevents interference and enables generalization

François Chollet's observation that children—the best learners—are "extremely bad at recollecting information" points to something profound: the inability to memorize may be a feature enabling generalization. Dwarkesh Podcast Current LLMs optimize in precisely the wrong direction, maximizing memory while minimizing the lossy compression that enables human creativity.

The goal is not to make AI that perfectly imitates human text—that is both achievable and hollow. The goal is to build systems that think more like humans: associatively, spontaneously, imperfectly, and creatively. The research synthesized here suggests this requires fundamental architectural changes, not just better prompting or fine-tuning of existing paradigms.