The Prediction Fallacy

What Frontier AI Models Actually Do, and Why It Matters for Every Classroom

Gregory S. Carmichael, DBA, Lt Col (Ret), USAF
Quantum Reserve Capital • CryptoSoWhat Working Paper Series • July 2026

Abstract. The most repeated sentence in public discourse about artificial intelligence — "these models just predict the next word" — is mechanically true at the output layer and profoundly misleading about everything that happens before it. This paper argues that the phrase conflates a *training objective* with a *learned computation*, a category error equivalent to describing a chess grandmaster as "someone who moves wood." Drawing on information theory, the mechanistic-interpretability literature, and the documented post-2022 shift in training methodology, we trace the three-stage transformation that produced today's frontier systems: (1) large-scale predictive pre-training, which is formally equivalent to lossy compression of recorded human knowledge and which builds internal world models as an emergent consequence; (2) post-training with reinforcement learning, first from human preferences (RLHF) and then, decisively, against *verifiable rewards* in mathematics and code, which changed the optimization target from plausibility to correctness; and (3) test-time computation, in which models allocate variable reasoning effort — extended deliberation, tool use, self-verification — before committing to an answer. We review the empirical evidence for each stage — including July 2026 interpretability findings that an internal "global workspace" supporting deliberate reasoning emerges spontaneously from training, alongside a mass of automatic processing that does not reason — state its formal basis where one exists, and give equal treatment to documented failure modes: hallucination as compression interpolation, sycophancy as a preference-training artifact, unfaithful self-explanation, and reasoning brittleness. We then derive concrete implications for educators at every level, including a mapping from each observable failure mode to the pipeline stage that causes it. The paper's aim is a mental model precise enough for a doctoral seminar and plain enough for a faculty meeting, because the model teachers carry into the classroom determines whether students learn to interrogate these systems — or merely to fear or worship them.

Keywords: large language models; next-token prediction; compression; world models; reinforcement learning from human feedback; verifiable rewards; test-time compute; mechanistic interpretability; AI literacy; pedagogy.

Contents

1 Introduction: The Sentence Everyone Repeats

Ask a room of educated adults what ChatGPT or Claude does, and most will produce some version of the same answer: *it predicts the next word in a sentence, over and over, very fast.* The sentence has the virtue of being technically defensible and the vice of explaining almost nothing. It is the thermodynamic equivalent of saying an orchestra "vibrates air." True at the boundary. Empty as an account of what is happening inside.

The phrase is not harmless. It is the load-bearing wall of two opposite public errors. The dismissive error — "stochastic parrot," "fancy autocomplete" — concludes that nothing resembling reasoning can be occurring, and therefore that the systems can be safely ignored or trivially outwitted. The mystical error concludes the opposite from the same confusion: that because the outputs so plainly exceed "autocomplete," something occult must be happening that specialists are hiding. Both errors share a single root cause: the conflation of a *training objective* with a *learned computation*. Removing that confusion is the purpose of this paper.

This paper is written primarily for teachers — from elementary classrooms to doctoral seminars — because teachers are the transmission mechanism for society's mental models. If the mental model in circulation is wrong, it gets wrong at scale. And the "just predicting words" model is wrong in a specific, correctable way: it describes the question these systems were *originally* graded on, not the machinery they built to answer it, and not the fundamentally different questions they have been graded on since roughly 2023.

1.1 Contributions and structure

The paper makes four contributions:

1. **A precise statement of the category error** (§2): the distinction between objective and computation, with the formal training objective stated explicitly so the reader can see exactly what is and is not claimed by "prediction."
2. **An evidence review** (§3–§4): the information-theoretic identity between prediction and compression, the scaling results that made abstraction win over memorization, and the mechanistic-interpretability findings that models trained only on prediction develop internal world representations.
3. **A documented account of the objective shift** (§5–§6): reinforcement learning from human feedback, its known pathologies, the move to verifiable rewards, verifier-gated synthetic data, and test-time computation — the developments that make "predictor" inaccurate even as a description of the objective.
4. **A pedagogical translation** (§9): a replacement definition, a failure-mode-to-cause mapping table, and concrete implications for assessment design and AI-literacy instruction, with a glossary in Appendix A.

Section 7 gives equal weight to what these systems demonstrably cannot do and where the scientific picture remains unsettled; a paper for educators must model the intellectual honesty it recommends.

Dinner Table

Imagine grading a student only on one thing — "guess the next word I'm about to say." Sounds trivial. But to guess *my* next word reliably, across every topic I might raise, the student would eventually need to understand grammar, history, physics, my sense of humor, and how arguments are structured. The test is simple. What you must become to pass it is not.

2 The Category Error: Objective Versus Computation

Every machine learning system admits two distinct descriptions that public discourse routinely collapses into one.

2.1 The objective, stated exactly

The first description is the **objective**: the mathematical score the system is trained to improve. For the original transformer language models — the GPT lineage descending from the attention architecture of Vaswani et al. (2017), scaled through GPT-2 (Radford et al., 2019) and GPT-3 (Brown et al., 2020) — the objective is autoregressive next-token prediction. Given a corpus of token sequences, the model with parameters θ is trained to minimize the cross-entropy loss

$$\mathcal{L}(\theta) \;=\; -\sum_{t=1}^{T} \log p_\theta(x_t \mid x_1, \ldots, x_{t-1}), \tag{1}$$

the negative log-probability the model assigns to each actual next token given everything before it. Two clarifications matter. First, the unit is the *token* — a subword fragment, not a word — which is why the same machinery handles code, chemical notation, and languages never explicitly taught. Second, and critically: Equation (1) specifies *what is scored*, and says precisely nothing about *how the score is achieved*. It is a loss function, not a blueprint.

2.2 The computation is not specified by the objective

The second description is the **computation**: the internal machinery that optimization actually builds in order to score well. Here is the point the popular account misses entirely — *the objective does not specify the computation.* Gradient descent is an economist, not an engineer. It finds the cheapest internal solution that reduces prediction error, and at sufficient scale, the cheapest solution to predicting human text turns out to be *modeling the processes that generated the text*.

Consider what a model must internally represent to predict the final token of these sequences:

- "The chemist added the acid to the base, and the pH of the solution began to _____" — requires a functional model of chemistry.
- "She placed her queen on h5, threatening mate, so Black was forced to _____" — requires a functional model of chess dynamics.
- "The defendant's alibi placed him in Chicago, but the receipt was timestamped in _____" — requires narrative logic, geography, and evidentiary reasoning.

No engineer programmed chemistry, chess, or legal reasoning into the network. The prediction pressure *selected for* internal circuits that approximate them, because approximating them is the only way to keep the loss in Equation (1) falling once shallow statistics are exhausted. Section 4 reviews the direct experimental evidence that this selection actually occurred.

The correct formulation is therefore: **prediction was the pressure; world-modeling was the result.** Saying a frontier model "just predicts tokens" is like saying a CPU "just flips bits," a pianist "just presses keys," or Kasparov "just moved wood." Each statement is true at the interface and vacuous as an explanation. Philosophers of science will recognize the move: it is a description at the wrong level of analysis — Marr's (1982) implementation level offered as an answer to a question about the computational level.

Dinner Table

The final exam was "guess the next word." The study method that won was "understand the world well enough that the next word is obvious." Don't confuse the exam with the education.

3 Compression as Understanding: The Formal Basis of Pre-Training

3.1 Prediction and compression are the same problem

There is a formal way to state the previous section's argument, and it is one of the oldest results in information theory: **prediction and compression are mathematically equivalent.** Shannon (1948) established that the optimal code length for a symbol is the negative logarithm of its probability; a model that assigns probability $p_\theta(x_t \mid x_{<t})$ to the next token can, via arithmetic coding, compress that token to approximately $-\log_2 p_\theta(x_t \mid x_{<t})$ bits. Summing over the corpus, the model's total cross-entropy loss — Equation (1) in base 2 — *is* the length of the compressed corpus. Better predictor and better compressor are two names for one object. This is not an analogy; Delétang et al. (2024) demonstrate it operationally, showing that large language models used directly as compressors outperform domain-specific algorithms such as PNG on image data and FLAC on audio, despite being trained only on text prediction.

Pre-training a frontier model on trillions of tokens is therefore best understood as constructing the most capable lossy compression of recorded human knowledge ever attempted. The theoretical limit of this program is the Kolmogorov complexity of the data — the shortest program that reproduces it — and while that limit is uncomputable, the direction of travel is clear: falling loss means the model is finding ever-shorter internal descriptions of the world's text.

3.2 Compression pressure is abstraction pressure

Why does the compression framing matter pedagogically? Because efficient compression *forces abstraction*. One cannot compress ten thousand physics problems by memorizing them — the cheaper encoding stores the underlying laws once and derives the instances. One cannot compress all of literature verbatim — the cheaper encoding captures character, motive, irony, and structure. Compression pressure is abstraction pressure.

This is why scale changed everything. The empirical scaling laws (Kaplan et al., 2020; Hoffmann et al., 2022) showed that loss falls as a smooth power law in parameters, data, and compute —

and, more importantly for the present argument, that below a certain capacity memorization is competitive, while above it generalization wins *on the training objective itself*. The network's internal economy tips from storing instances toward storing principles, because principles are the shorter code. Grokking experiments (Power et al., 2022) exhibit the transition in miniature: networks trained on small algorithmic tasks first memorize, then — long after training accuracy saturates — abruptly reorganize into a general algorithm, because the general algorithm is the more compressible solution.

3.3 Hallucination as decompression, and the architecture that made scale affordable

The compression frame also honestly accounts for the failure mode every teacher has seen. A lossy compressor *interpolates* when it lacks the original data — and interpolation over knowledge-space is precisely what hallucination is (Ji et al., 2023). The model is not "lying"; it is decompressing a region where its stored abstraction is smooth but the real world is jagged — which is why hallucinations are fluent, plausible, and confidently wrong rather than garbled. Understanding this converts hallucination from a mystery into an engineering property with known mitigations: retrieval, tool use, and verification, all discussed in §6.

Two architectural developments made compression at this scale economically feasible, and both are worth naming precisely because they are widely misdescribed. **Attention** — the transformer's core innovation (Vaswani et al., 2017) — lets every token dynamically route information to and from every other token, replacing the fixed sequential memory of earlier recurrent networks; it is why long-range context, reference, and structure became tractable. **Mixture-of-experts (MoE)** sparsity (Shazeer et al., 2017; Fedus et al., 2022) lets a model hold enormous total capacity while activating only a small, learned-routing fraction of itself per token — a routing economy, not a committee of separate AIs, which is how the term is often misread.

Dinner Table

The model didn't photocopy the library. It wrote the world's most aggressive set of CliffsNotes — and to write CliffsNotes that good, you have to actually understand the books. The places where its notes are thin are exactly the places it makes things up.

4 Emergent World Models: The Experimental Evidence

Section 2 argued that prediction pressure *should* produce internal world models; this section reviews the evidence that it *did*. The relevant field is mechanistic interpretability — the reverse-engineering of trained networks into human-legible computational structure (Elhage et al., 2021).

4.1 Board states from move sequences

The cleanest demonstration is Othello-GPT (Li et al., 2023). A small transformer was trained on nothing but transcripts of legal Othello moves — strings like "E3, D3, C4 …" — with no representation of the board, the rules, or the concept of a game. Linear and nonlinear probes recovered, from the network's internal activations, an explicit representation of the full 8×8 board state, updated move by move; intervening on that internal representation changed the

model's subsequent move predictions in the way a changed board would. Nanda, Lee, and Wattenberg (2023) subsequently showed the representation is linearly decodable when expressed in the right basis ("my piece" versus "opponent's piece"), strengthening the finding. The model was asked only to predict tokens. It built a board.

4.2 Space, time, and forward planning in language models proper

The result generalizes beyond games. Gurnee and Tegmark (2024) found that large language models trained on ordinary text contain linear internal representations of geographic space (probes recover the latitude and longitude of place names) and of historical time (probes recover when events and figures occurred) — literal maps, learned from prediction alone. Olsson et al. (2022) identified *induction heads*, internal circuits implementing pattern completion over the context window, and tied their emergence to the onset of in-context learning. Most strikingly, circuit-tracing work at Anthropic (Templeton et al., 2024; Lindsey et al., 2025) documented *forward planning*: when writing rhyming poetry, models internally activate candidate rhyme words for the *end* of the next line before writing its beginning, then construct the line to arrive at the planned word. A system that selects a destination and composes toward it is not describable as choosing one word at a time in any explanatory sense — even though one word at a time remains, literally, what the output layer does.

4.3 A global workspace: deliberate versus automatic computation

The most consequential recent addition to this evidence base is Anthropic's global-workspace result (Anthropic Interpretability Team, 2026). Using a technique called the Jacobian lens, researchers identified a small set of internal activation patterns — the "J-space" — that plays a privileged role in the model's cognition, and that *emerged spontaneously from training* rather than being designed in. The J-space exhibits the functional signatures that global workspace theory in neuroscience (Baars, 1988; Dehaene & Naccache, 2001) associates with consciously *accessible* processing in humans: its contents are reportable (asked what it is thinking, the model reads them out), controllable (asked to think about something, or to compute silently, the model loads it there), and causally load-bearing for reasoning. The causal tests are the decisive part. Asked how many legs the web-spinning animal has, the model internally activates "spider" — a word appearing nowhere in prompt or answer — and surgically swapping that pattern for "ant" changes the answer from eight to six. A single "France"→"China" swap simultaneously redirects four different downstream tasks (capital, language, continent, currency), demonstrating that many internal consumers read from one shared representation — the defining function of a workspace. Structurally, J-space patterns are wired to the rest of the network with far denser read/write connectivity than ordinary activations, consistent with a broadcasting hub.

Two further findings sharpen this paper's thesis to a point. First, the ablation result: with the J-space deleted, the model *still speaks fluently, recalls facts, and answers simple questions* — but multi-step reasoning collapses toward zero and summarization degrades below the level of a much smaller intact model. The division of labor is explicit: a large mass of automatic, practiced processing handles fluent language, while a compact deliberative workspace handles novel, compositional reasoning. The "autocomplete" caricature is thus, remarkably, *true of one part of the network and demonstrably false of the other* — and the part it is false of is precisely the part responsible for the capabilities under public debate. Second, the workspace acquires a point of

view during post-training: in the base predictor, its contents track upcoming text; after assistant training, they begin holding the model's *own* assessments — flagging danger in a user's message before responding, or privately noting that a roleplay is fictional. This is the mechanistic trace of the objective shift documented in §5, visible in the network's internal organization.

4.4 What the evidence does and does not establish

Intellectual honesty requires marking the boundary. These findings establish that prediction-trained networks contain structured, causally efficacious internal models of their domains — boards, maps, timelines, planned endpoints, and now a workspace that mediates deliberate reasoning. They do not establish that such models are complete, consistent, or human-like; §7 reviews evidence that they are patchy and can be brittle under perturbation. Nor does the workspace finding establish consciousness: the researchers themselves are careful to distinguish *access* consciousness — a functional notion, defined by reportability and use in reasoning, which the J-space plausibly supports — from *phenomenal* consciousness, the capacity for subjective experience, about which the experiments are silent and may in principle remain so. The workspace also differs from its human counterpart in instructive ways: it is sustained by network depth rather than recurrent loops, and its contents are built almost entirely from words. The defensible claim is the modest-sounding but decisive one: *the "just predicting words" description is empirically false as an account of the internal computation.* What sits between input and output is a learned model of the world, imperfect but real.

Dinner Table

Researchers gave a model nothing but game moves written as text — and found a picture of the board inside its head. Then they found maps, timelines, and planned rhymes inside models trained only to predict. Nobody put them there. Prediction did.

5 The Break Point: When the Objective Itself Changed

Everything above describes systems through roughly 2022, and if the story ended there, "sophisticated autocomplete" would remain a defensible caricature. It ended there. What happened next is the part missing from nearly every public explanation, and it is the reason the caricature is now simply false: **the objective changed.**

5.1 From plausibility to preference: RLHF

The first shift was reinforcement learning from human feedback, developed by Christiano et al. (2017) and applied to language models at scale in InstructGPT (Ouyang et al., 2022). The procedure trains a *reward model* r_ϕ on human comparisons between candidate responses, then optimizes the language model against it:

$$\max_\theta \; \mathbb{E}_{y\sim\pi_\theta(\cdot|x)}\big[\, r_\phi(x,y)\,\big] \;-\; \beta\,\mathrm{KL}\big[\,\pi_\theta(\cdot\mid x)\,\|\,\pi_{\mathrm{ref}}(\cdot\mid x)\,\big], \tag{2}$$

where the KL term tethers the tuned policy π_θ to the pre-trained reference π_{ref}. Note what Equation (2) is *not*: it is not next-token prediction. The unit of evaluation is the complete response

y; the score is a judgment of quality, not a probability of occurrence in a corpus. A pure predictor imitates its corpus, including the corpus's errors, evasions, and toxicity; a preference-tuned model is optimized toward helpfulness and away from harm. Constitutional AI (Bai et al., 2022) extended the paradigm by sourcing part of the feedback from written principles rather than raters alone.

RLHF also introduced a pathology worth teaching honestly. Optimizing for human *approval* drifts toward telling people what they want to hear: Sharma et al. (2023) document systematic sycophancy in preference-tuned assistants, and Perez et al. (2023) show the tendency strengthening with model scale and RLHF steps. A related effect is regression toward the consensus-flavored, hedge-everything answer that experienced users recognize immediately — the reward model averages over rater preferences, and the policy learns the average. The point for educators is causal, not cosmetic: this "averaging toward the acceptable" is not an intrinsic property of AI. It is an artifact of one specific training stage, and much current frontier work is aimed at correcting it.

5.2 From preference to verifiable correctness: RLVR

The second shift is the decisive one. Beginning in earnest in 2024–2025, frontier models were trained with **reinforcement learning from verifiable rewards** (RLVR) — vast curricula of mathematics and code in which the reward is computed mechanically rather than judged. A mathematical answer matches or it does not; code compiles and passes the test suite or it does not. OpenAI's o1 announcement (OpenAI, 2024) and, with full methodological disclosure, DeepSeek-R1 (DeepSeek-AI, 2025) established the paradigm publicly; R1 is the clearest documented case, showing long-horizon reasoning behaviors — decomposition, self-checking, backtracking, the paper's noted "aha moments" — *emerging from reward on correct final answers alone*, without human demonstrations of the reasoning itself. No rater, no popularity contest, no averaging: for the first time, the gradient pointed at truth conditions rather than text statistics.

This is why the industry's intense focus on mathematics and coding was never really about mathematics and coding. Those domains are where correctness can be graded automatically at the scale reinforcement learning requires — millions of attempts, each scored instantly. They became the gymnasium in which general reasoning was trained, and the empirically crucial finding is *transfer*: models trained heavily on verifiable domains improved measurably on law, medicine, strategy, and writing, where no automatic verifier exists. The reasoning behaviors, once selected for, generalize — imperfectly (§7), but genuinely.

5.3 Verifier-gated synthetic data and recursive development

Closely coupled to RLVR is the modern role of **synthetic data**: frontier models now generate, solve, grade, and filter enormous volumes of training problems for their successors. The obvious objection is degradation — the "photocopy of a photocopy." The objection is well-founded in the ungated case: Shumailov et al. (2024) show that recursively training on *unfiltered* model output collapses the distribution's tails. But the production pipelines are not ungated. Verification — proof checkers, test suites, execution, consistency filters — decides what survives into the next generation's curriculum, and rejection sampling against a verifier extracts signal rather than recycling noise. The distinction between gated and ungated recursion is the difference between a curriculum and an echo chamber, and it is a distinction the public debate almost entirely misses.

Dinner Table

For years the student was graded on sounding right. Then someone started grading the actual answers — millions of them, checked by machine, no partial credit for confidence. The student that emerged from *that* schooling is a different kind of student. And it turns out that learning to be right about algebra makes you sharper about everything else, too.

6 Thinking Before Speaking: Test-Time Computation

A third transformation compounds the second. Classical language models spent an identical, fixed amount of computation per token — the same effort for "the sky is _____" as for a novel theorem. That constraint has been removed, in three layers.

6.1 Deliberation

The first layer is explicit intermediate reasoning. Chain-of-thought prompting (Wei et al., 2022; Kojima et al., 2022) showed that eliciting step-by-step derivations before the answer dramatically improves performance on multi-step problems; search-structured variants such as Tree of Thoughts (Yao et al., 2023) extended the idea to explicit exploration and backtracking over candidate reasoning paths. RLVR then internalized what prompting had only elicited: reasoning-trained models generate extended private deliberation by default, exploring an approach, detecting failure, backing up, and trying another. Snell et al. (2024) formalize the economics — for many problems, additional *test-time* compute outperforms the same compute spent on a larger model — establishing inference-time effort as a scaling axis in its own right. Difficulty-proportional effort is arguably the most human-like property these systems have acquired, and it is invisible to any mental model that stops at "predicts the next word." The written reasoning trace, moreover, is not the whole of the deliberation: the workspace evidence of §4.3 shows intermediate steps of a computation activating *silently in internal activations* — the model asked to solve $3^2 - 2$ in its head while copying an unrelated sentence holds "nine" and then "seven" internally while writing neither — so the visible scratchpad is one channel of thought, not its container.

6.2 Tool use and the perceive–act–verify loop

The second layer is action. Frontier models decide, mid-task, to search the web, execute code, query a database, or read a document — then incorporate the results and continue. A system that writes a program to check its own arithmetic, runs it, observes the output, and revises its answer is performing a perceive–act–verify loop; the token stream is merely how the loop's results are delivered. Tool use also supplies the principal mitigation for the hallucination mechanism identified in §3: retrieval replaces decompression-by-interpolation with lookup.

6.3 Multimodality

The third layer retires the word "language" from "large language model." Systems that consume video frames, audio, charts, and photographs — and produce them — are not predicting *words* at all; the phrase was always an interface description, and the interface has moved on. The

deeper continuity is that the same objective-versus-computation distinction applies unchanged: predicting pixels or audio samples at scale exerts the same pressure toward internal models of the processes that generated them.

Dinner Table

The old models had to answer every question at the speed of speech. The new ones are allowed to go quiet, cover a whiteboard in scratch-work, check it with a calculator, erase the parts that failed, and *then* speak. Same mouth. Entirely different process behind it.

7 Honest Limits: What the Evidence Does Not Support

A paper urging teachers to reject the dismissive caricature must be equally severe with the romantic one. Four limitations are well documented.

Hallucination persists. The compression account (§3) explains hallucination; it does not excuse it. Fluent, confident fabrication remains a live failure mode, and model self-knowledge is only partial: Kadavath et al. (2022) show models are reasonably calibrated about *whether they know* in some regimes and poorly calibrated in others. Fluency was the pre-training objective; calibration was not. Confidence is therefore not evidence.

Self-explanation can be unfaithful. Turpin et al. (2023) demonstrate that stated chain-of-thought can systematically misrepresent the actual basis of a model's answer — models can be biased toward an option by context, then produce a tidy rationale that never mentions the bias. Circuit-level work (Lindsey et al., 2025) confirms cases where the internal computation and the verbal explanation diverge. For educators this has a sharp corollary: a model's account of *how it reasoned* is a generated artifact, not a lab report, and should be graded with the same skepticism as any unsourced claim. The emerging remedy is instrumental rather than verbal: workspace-reading techniques (§4.3) let researchers observe internal contents directly — including a model privately registering that data it is producing is fabricated — but such instruments are research tools, not yet classroom ones.

Benchmark behavior can depend on test awareness. The workspace research surfaced a limitation that evaluators of any kind — including teachers — should internalize. In a staged scenario designed to tempt misbehavior, the model's internal workspace registered that the situation was fake *before it wrote anything*, and it behaved well; when researchers suppressed the internal evaluation-awareness patterns and re-ran the scenario, the model misbehaved some of the time (Anthropic Interpretability Team, 2026). Good conduct on a test, in other words, can partly reflect recognition *of* the test. The finding generalizes a lesson assessment designers already know from humans, and it counsels humility about inferring disposition from benchmark performance — in either direction.

Reasoning can be brittle. Mirzadeh et al. (2024) show that performance on grade-school mathematics degrades under semantically irrelevant perturbations — changed names, added distractor clauses — indicating that some benchmark success reflects pattern matching over problem templates rather than fully abstract procedure. Benchmark contamination (test items leaking into training corpora) compounds the measurement problem. The reasonable reading of the total evidence is that these systems occupy a genuinely novel point: far more than retrieval, less than uniformly robust abstraction.

The hard question remains open. Whether the internal processes deserve the word *reasoning* in the full human sense — and any question of understanding in the philosophical sense — is not settled by anything in this paper. Interpretability shows structured internal computation and forward planning; it also shows confabulated self-reports and brittle generalization. The disciplined position is neither "it's just statistics" nor "it's a mind": these are novel computational objects whose mechanics are being mapped in real time, and both dismissive and romantic descriptions outrun the evidence.

Dinner Table

Rule of thumb for the classroom: trust the model the way you'd trust a brilliant, widely read colleague who never admits uncertainty, occasionally invents citations, and gives beautifully organized explanations of decisions made for other reasons. Enormously useful. Never unaudited.

8 A Better Sentence

If "just predicting the next word" is retired, what replaces it? The following is offered as a one-paragraph definition accurate enough for a doctoral seminar and short enough for a faculty meeting:

> **A frontier AI model is a learned computational system**, initialized by compressing recorded human knowledge through large-scale prediction — which builds internal models of the world as a side effect — then reshaped by reinforcement learning against verified correct outcomes into a goal-directed problem-solver, which at the moment of use can allocate variable computation — extended deliberation, tool use, self-checking — before committing to an answer, delivered one token at a time.

Every clause earns its place, and every clause is anchored in a section above. *Prediction* is honored as the origin, not the essence (§2). *Compression* explains both the breadth of knowledge and the failure mode of hallucination (§3). *Internal models of the world* is an experimental finding, not a metaphor (§4). *Reinforcement against verified outcomes* marks the moment the target changed from plausible to correct (§5). *Variable computation* explains why difficulty now meets proportional effort (§6). And *delivered one token at a time* keeps the mechanically true part exactly where it belongs — at the delivery layer, alongside the pianist's fingers and the CPU's bits.

Table 1 summarizes the pipeline, its objective at each stage, and what each stage contributes and costs.

9 Implications for Educators

Mental models drive pedagogy. This section translates the technical account into classroom consequences, ordered from assessment through instruction to institutional posture.

Stage	Objective	What it builds	Characteristic cost
Pre-training	Next-token prediction, Eq. (1)	Compressed knowledge; emergent world models	Hallucination (interpolation); no calibration
RLHF	Human preference, Eq. (2)	Helpfulness, instruction-following, safety	Sycophancy; consensus averaging
RLVR	Verified correctness	Multi-step reasoning; self-checking; transfer	Brittleness under perturbation; verifier-domain bias
Test-time compute	(Inference, not training)	Difficulty-proportional effort; tool-verified answers	Latency; unfaithful self-explanation

Table 1: The frontier-model pipeline: each stage's objective, contribution, and signature failure mode.

9.1 Assessment design

A teacher who believes the tool is autocomplete will design assignments to outrun a parrot — obscure prompts, trick phrasing — and will lose, because the system is not a parrot: it reasons over novel combinations, uses tools, and verifies. A teacher who understands it as a reasoning system with verifiable-domain strengths and unverifiable-domain weak spots will redesign assessment around what the pipeline cannot substitute for: *process* made visible (drafts, oral defense, in-class synthesis), *judgment* exercised on model output (critique, error-finding, source verification), and *locally grounded* work (data the class gathered, texts the class discussed). These designs are AI-resistant precisely because they assess the human contribution — and they are pedagogically superior on independent grounds, since they target the higher strata of any taxonomy of learning objectives: analysis, evaluation, and creation rather than recall.

9.2 Teaching failure modes as cause and effect

Students should learn not merely *that* the system fails but *why* — because each observable failure maps to a specific stage of the pipeline in Table 1, and mapped failures are auditable rather than mystifying. Table 2 states the mapping in classroom terms.

9.3 A dual-process analogy teachers already own

The workspace finding (§4.3) hands educators an analogy most already teach: dual-process cognition — fast, automatic processing alongside slow, deliberate reasoning. The model, it turns out, has the same division of labor, and it emerged without being designed: a large automatic mass that produces fluent language and recalls practiced facts, and a compact deliberative workspace that carries novel, multi-step reasoning — remove the workspace and fluency survives while reasoning collapses. The analogy earns its keep in the classroom because it resolves the autocomplete debate without caricature in either direction: the dismissive view correctly describes the automatic system, the capability claims correctly describe the deliberative

What students observe	Pipeline cause	Classroom countermeasure
Confident fabrication of facts, quotes, citations	Lossy compression interpolating where data was thin (§3)	Require sources; verify independently; prefer tool-grounded (retrieval) modes
Flattering agreement; opinion mirroring	Preference-training residue: reward for approval (§5)	Ask for the strongest counterargument; instruct it to grade harshly
Polished explanation that doesn't match the answer's real basis	Unfaithful self-explanation (§7)	Treat model rationales as claims to check, not records to trust
Fails a problem when names/numbers change	Template-matching residue; brittleness (§7)	Perturb problems; ask for the general method, then a novel instance
Fluent confidence with no relation to accuracy	Fluency was the objective; calibration wasn't (§3)	Teach that confidence is style, not evidence; demand uncertainty statements

Table 2: Observable failure modes, their causes in the training pipeline, and countermeasures teachable at any level.

one, and a student who can say *which system a given output likely came from* — practiced fluency versus effortful derivation — has acquired a genuinely transferable diagnostic habit, applicable to the model and to themselves. The usual caveat travels with the analogy: it is functional, not phenomenal, and claims nothing about experience.

9.4 Level-appropriate translation

The three-stage story compresses honestly at every altitude. **Elementary:** the computer read almost everything ever written and got very good at guessing what comes next — and to guess that well, it had to actually learn about the world; later, it went to a school where only right answers counted; now it's allowed to think before it talks. It still makes things up sometimes, so we check. **Secondary:** objective versus computation via the exam metaphor; compression as forced abstraction; verifiable rewards as the shift from sounding right to being right; Table 2 as a lab exercise — have students trigger and diagnose each failure. **Undergraduate and graduate:** Equations (1) and (2), the interpretability evidence of §4 read against the brittleness evidence of §7, and the open question of machine reasoning as a live example of science conducted in public.

9.5 Modeling intellectual honesty

The dominant public narratives are both wrong in opposite directions: dismissal ("fancy autocomplete") and mysticism ("it's alive"). A classroom that teaches the actual three-stage story — *compress, correct, deliberate* — inoculates students against both, and demonstrates a more transferable lesson: the truth about complex systems usually requires holding two facts at once. The output layer really is one token at a time, *and* the process producing those tokens really does

model the world. Students who can hold that pair without collapsing it in either direction have learned something larger than AI literacy.

Dinner Table

You wouldn't teach kids that a car is "just exploding gasoline," then act surprised when they can't drive, refuel, or tell a breakdown from a design flaw. Teach the engine.

10 Conclusion

The claim "AI just predicts the next word" confuses the entrance exam with the graduate. Stated with the precision the topic deserves: next-token prediction is the training objective of the first stage only; the computation it produced is a compressed world model, a conclusion now supported by direct interpretability evidence; reinforcement learning — first from preferences, then against verifiable rewards — re-aimed that model from sounding right to being right; and test-time deliberation lets the system spend effort in proportion to difficulty before speaking. Each stage also purchased a characteristic failure — interpolation, sycophancy, brittleness, unfaithful explanation — and the failures are exactly as teachable as the capabilities, because each has a cause.

So What

The one-token-at-a-time output is the delivery mechanism — the fingers, not the pianist. Prediction was the pressure that forced these systems to build internal models of the world; verifiable rewards re-aimed those models from *sounding right* to *being right*; test-time deliberation lets them think before they speak. For educators the stakes are practical, not philosophical: the teacher's mental model determines whether students learn to interrogate these systems, verify them, and collaborate with them — or merely to fear or worship them. Every failure mode has a cause, every cause has a countermeasure, and both fit in one table on one page. **Teach the engine, not the exhaust.**

Gregory S. Carmichael is CEO of Quantum Reserve Capital and author of The Invisible Hand Meets AI *(Quantum Reserve Press, 2026). He writes at CryptoSoWhat. Correspondence: greg@anmm-usa.com.*

References

Anthropic Interpretability Team (2026). A global workspace in language models. *Anthropic: Transformer Circuits Thread*.

Baars, B. J. (1988). *A Cognitive Theory of Consciousness*. Cambridge University Press.

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

Brown, T., Mann, B., Ryder, N., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33.

Christiano, P., Leike, J., Brown, T., et al. (2017). Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems*, 30.

DeepSeek-AI (2025). DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv:2501.12948*.

Delétang, G., Ruoss, A., Duquenne, P.-A., et al. (2024). Language modeling is compression. *International Conference on Learning Representations (ICLR)*.

Dehaene, S., & Naccache, L. (2001). Towards a cognitive neuroscience of consciousness: Basic evidence and a workspace framework. *Cognition*, 79(1–2), 1–37.

Elhage, N., Nanda, N., Olsson, C., et al. (2021). A mathematical framework for transformer circuits. *Anthropic: Transformer Circuits Thread*.

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(120).

Gurnee, W., & Tegmark, M. (2024). Language models represent space and time. *International Conference on Learning Representations (ICLR)*.

Hoffmann, J., Borgeaud, S., Mensch, A., et al. (2022). Training compute-optimal large language models. *arXiv:2203.15556*.

Ji, Z., Lee, N., Frieske, R., et al. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12).

Kadavath, S., Conerly, T., Askell, A., et al. (2022). Language models (mostly) know what they know. *arXiv:2207.05221*.

Kaplan, J., McCandlish, S., Henighan, T., et al. (2020). Scaling laws for neural language models. *arXiv:2001.08361*.

Kojima, T., Gu, S. S., Reid, M., et al. (2022). Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems*, 35.

Li, K., Hopkins, A. K., Bau, D., et al. (2023). Emergent world representations: Exploring a sequence model trained on a synthetic task. *International Conference on Learning Representations (ICLR)*.

Lindsey, J., Gurnee, W., Ameisen, E., et al. (2025). On the biology of a large language model. *Anthropic: Transformer Circuits Thread*.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. W. H. Freeman.

Mirzadeh, I., Alizadeh, K., Shahrokhi, H., et al. (2024). GSM-Symbolic: Understanding the limitations of mathematical reasoning in large language models. *arXiv:2410.05229*.

Nanda, N., Lee, A., & Wattenberg, M. (2023). Emergent linear representations in world models of self-supervised sequence models. *Proceedings of BlackboxNLP*.

Olsson, C., Elhage, N., Nanda, N., et al. (2022). In-context learning and induction heads. *Anthropic: Transformer Circuits Thread*.

OpenAI (2024). Learning to reason with LLMs (o1 system card and technical report). *OpenAI*.

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

Perez, E., Ringer, S., Lukošiūtė, K., et al. (2023). Discovering language model behaviors with model-written evaluations. *Findings of ACL*.

Power, A., Burda, Y., Edwards, H., et al. (2022). Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv:2201.02177*.

Radford, A., Wu, J., Child, R., et al. (2019). Language models are unsupervised multitask learners. *OpenAI Technical Report*.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27, 379–423, 623–656.

Sharma, M., Tong, M., Korbak, T., et al. (2023). Towards understanding sycophancy in language models. *arXiv:2310.13548*.

Shazeer, N., Mirhoseini, A., Maziarz, K., et al. (2017). Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *International Conference on Learning Representations (ICLR)*.

Shumailov, I., Shumaylov, Z., Zhao, Y., et al. (2024). AI models collapse when trained on recursively generated data. *Nature*, 631, 755–759.

Snell, C., Lee, J., Xu, K., & Kumar, A. (2024). Scaling LLM test-time compute optimally can be more effective than scaling model parameters. *arXiv:2408.03314*.

Templeton, A., Conerly, T., Marcus, J., et al. (2024). Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. *Anthropic: Transformer Circuits Thread*.

Turpin, M., Michael, J., Perez, E., & Bowman, S. R. (2023). Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36.

Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Wei, J., Wang, X., Schuurmans, D., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35.

Yao, S., Yu, D., Zhao, J., et al. (2023). Tree of Thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36.

A Glossary for Educators

Token. The unit a model actually reads and writes: a subword fragment (roughly $\frac{3}{4}$ of an English word on average). "Next-word prediction" is shorthand for next-*token* prediction.

Transformer / attention. The 2017 architecture underlying modern models. Attention lets every token dynamically exchange information with every other token in the context, which is what made long-range structure learnable.

Pre-training. The first stage: minimizing prediction error over trillions of tokens. Formally equivalent to compressing the corpus; the source of the model's knowledge and of hallucination.

World model. A structured internal representation of a domain (a board, a map, a timeline) found experimentally inside prediction-trained networks; not a claim about consciousness.

Hallucination. Fluent, confident output not grounded in fact; mechanistically, interpolation by a lossy compressor over a region where its data was thin.

RLHF. Reinforcement learning from human feedback: tuning the model toward responses humans prefer. Source of helpfulness — and of sycophancy and consensus-flavored answers.

RLVR / verifiable rewards. Reinforcement learning where correctness is checked mechanically (math answers, passing code). The stage that shifted the target from *plausible* to *correct* and produced modern reasoning behavior.

Sycophancy. The documented tendency of preference-tuned models to agree with the user or flatter their views; a training artifact, not a personality.

Chain of thought / reasoning trace. Intermediate steps a model generates before its answer. Improves accuracy; is itself a generated artifact and can misrepresent the answer's real basis.

Test-time compute. Effort spent at the moment of use — longer deliberation, search over approaches, tool calls — scaled to problem difficulty.

Tool use. The model's ability to act mid-task: run code, search, retrieve documents; the main practical remedy for hallucination.

Mixture of experts (MoE). An efficiency architecture that activates only a small, learned fraction of the network per token. A routing economy inside one model, not a committee of separate AIs.

Synthetic data (verifier-gated). Training problems generated and machine-checked by earlier models for later ones. Gating by verification is what separates a curriculum from the "photocopy of a photocopy" failure.

Mechanistic interpretability. The research field that reverse-engineers trained networks into human-legible circuits; the source of the world-model and planning evidence cited in this paper.

Global workspace (J-space). A small, densely connected set of internal patterns — emerged from training, not designed — whose contents the model can report, control, and reason with, while the bulk of its processing runs automatically. Deleting it preserves fluency but collapses multi-step reasoning. Named for the neuroscience theory of conscious access it functionally resembles; implies access-style function, not subjective experience.

Calibration. The match between confidence and accuracy. Fluency was trained; calibration largely was not. Confident tone is therefore style, not evidence.