# Knowledge, LLMs, and the Information Ladder: From Shannon to Kolmogorov

Alec Litowitz
*QStar Capital*

Nicholas G. Polson
*Booth School of Business*
*University of Chicago*

Vadim Sokolov
*Department of Systems Engineering*
*and Operations Research*
*George Mason University*

First Draft: April 2026
This Draft: April 15, 2026

**Abstract**

We develop a unified framework for understanding knowledge in large language models (LLMs), tracing an arc from Shannon's theory of communication to Kolmogorov's theory of complexity. LLMs are Shannon machines at the physical layer, converting photons to tokens at thermodynamic cost, yet they aspire to be Kolmogorov machines at the semantic layer, compressing world knowledge into parametric representations. We organize this duality as an *information ladder*: five questions ranging from Shannon's "can the message be transmitted?" through Cover's "how much information does it contain?" and Kolmogorov's "what does it mean?" to Good's "can the system improve itself?" and Hayek's "which questions are worth asking?" Drawing on the token economics of Litowitz et al. [2026b,a], the Kolmogorov-GAM architecture of Polson and Sokolov [2025], Good's ultraintelligent machines [Good, 1965], Hayek's dispersed knowledge [Hayek, 1945], and Cover's entropy measurements [Cover and King, 1978, Leung-Yan-Cheong and Cover, 1978], we identify model collapse as a central threat to the Kolmogorov frontier and ask what this framework implies for knowledge representation by 2100.

## 1 Introduction

There are two great theories of information. The first, due to Shannon [1948], is a theory of *transmission*: it answers the question of how reliably a message can be sent across

a channel, and at what rate. The second, due to Kolmogorov [1965] and independently Solomonoff [1964] and Chaitin [1966], is a theory of *content*: it answers the question of what a message *means*, measured by how much it can be compressed. Both theories are mathematically deep, and both are essential for understanding what large language models are and what they do.

The arrival of LLMs at civilizational scale has made these questions urgent and practical. Litowitz et al. [2026b] observe that debates about AI capabilities and risks are often conducted without quantitative grounding, and propose a rigorous accounting of the token economy from physics to economics. Their paper places Shannon's channel capacity at the foundation of the token supply chain. We extend their framework upward, toward Kolmogorov, to ask not just how many tokens can be produced but what those tokens *mean*—and how LLMs represent, compress, and retrieve knowledge.

Between Shannon and Kolmogorov stands I.J. Good, whose 1965 essay on the ultraintelligent machine [Good, 1965] introduced the concept of an intelligence explosion—a recursively self-improving system that designs successively smarter successors. Good, a committed Bayesian who worked with Turing at Bletchley Park, understood that genuine intelligence requires probabilistic reasoning under uncertainty, not merely logical manipulation. His essay anticipates the central tension of modern AI: the gap between a system's capacity to process information (Shannon) and its ability to compress and generate knowledge (Kolmogorov). The ultraintelligent machine is, in Good's framing, the system that closes this gap.

Tom Cover's work provides a crucial empirical and theoretical bridge. His gambling estimate of the entropy of English [Cover and King, 1978] measured from above how much information is actually present in natural language, while his co-authored paper on the equivalences between Shannon entropy and Kolmogorov complexity [Leung-Yan-Cheong and Cover, 1978] established the formal connection between the two theories. Together, these papers represent one of the deepest syntheses of the Shannon and Kolmogorov frameworks, and LLMs can be understood as the computational realization of the program Cover sketched in them.

This paper makes the following contributions. First, we trace the complete conceptual arc from Shannon's channel capacity through Cover's entropy measurements to Kolmogorov's complexity theory, constructing an *information ladder*—a hierarchy of increasingly demanding questions about information that organizes the relationship between transmission, entropy, complexity, recursive improvement, and agency. Second, we connect Good's intelligence explosion thesis to the Kolmogorov structure function, arguing that the intelligence explosion is a cascade of compression improvements, and assess the extent to which current LLMs constitute proto-ultraintelligent systems. Third, we integrate the token economics framework of Litowitz et al. [2026b] and Litowitz et al. [2026a] with the Kolmogorov-GAM architecture of Polson and Sokolov [2025] to provide a unified account of LLMs from the physical layer to the semantic layer, drawing on Hayek's theory of dispersed knowledge [Hayek, 1945] to identify the structural limits of centralized knowledge compression. Fourth, we use the Shannon–Kolmogorov duality to analyze three scenarios for the future of knowledge representation and identify the data collapse problem as a central threat to the sustained expansion of the Kolmogorov frontier.

Section 2 reviews Shannon's theory and its role in the physical layer of LLM computation. Section 3 presents Cover's contributions on the entropy of English and the Shannon–Kolmogorov equivalences. Section 4 turns to Kolmogorov complexity and its implications for knowledge representation. Section 5 analyzes Good's ultraintelligent machine thesis through the lens of Kolmogorov complexity. Section 6 analyzes LLMs as compressors. Section 7 discusses the token economics framework. Section 8 presents the K-GAM architecture as a Kolmogorov-inspired model of knowledge structure. Section 9 discusses the gap between capacity and meaning. Section 10 asks what this framework implies for knowledge in 2100. Section 11 concludes.

# 2 Shannon's Theory: The Capacity to Transmit

Claude Shannon's 1948 paper *A Mathematical Theory of Communication* [Shannon, 1948] founded information theory by separating the problem of meaning from the problem of transmission. Shannon was explicit: the semantic content of a message is irrelevant to the engineering problem of sending it. What matters is only the statistical structure—the probability distribution over possible messages—and the capacity of the channel.

Shannon's central result, the noisy channel coding theorem, establishes that any source of information can be transmitted reliably at any rate up to channel capacity

$$C = B \log_2 \left(1 + \frac{S}{N}\right), \tag{1}$$

where $B$ is bandwidth and $S/N$ is the signal-to-noise ratio. Below capacity, arbitrarily low error rates are achievable; above capacity, perfect transmission is impossible. The beauty of this result is its universality: it applies regardless of whether the message is text, image, audio, or genome.

Shannon's source coding theorem provides the complementary result on the compression side. For a discrete memoryless source emitting symbols from an alphabet $\mathcal{X}$ with probability distribution $p$, the entropy

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log_2 p(x) \tag{2}$$

is the minimum average number of bits per symbol required for lossless compression. No code can achieve a lower average rate; a code achieving rate $H(X) + \epsilon$ for any $\epsilon > 0$ exists. For a stationary ergodic source—a better model of natural language than the memoryless assumption—the relevant quantity is the entropy rate

$$h = \lim_{T \to \infty} \frac{1}{T} H(X_1, X_2, \ldots, X_T) = \lim_{T \to \infty} H(X_T \mid X_1, \ldots, X_{T-1}), \tag{3}$$

which captures the residual uncertainty per symbol after accounting for all dependencies in the sequence. This entropy rate is the fundamental quantity governing both the compressibility of natural language and the theoretical limit of next-token prediction.

For LLMs, Shannon's theory operates at two levels. At the hardware level, every token generated by a data center requires photons to propagate through fiber, electrons

to switch transistors, and bits to traverse memory buses. The thermodynamic cost of computation is bounded below by Landauer's principle [Landauer, 1961]: erasing one bit of information dissipates at least $k_B T \ln 2$ joules of energy, where $k_B$ is Boltzmann's constant and $T$ is temperature. At room temperature ($T \approx 300\,\mathrm{K}$), this is approximately $2.85 \times 10^{-21}$ joules per bit—a floor that current hardware exceeds by roughly six orders of magnitude, but one that represents an absolute physical limit on the energy efficiency of computation. Litowitz et al. [2026b] use Landauer's principle, Shannon's channel capacity, and current infrastructure data to construct a supply-and-demand balance sheet for global token production. The Shannon limit constrains the rate at which the physical substrate can deliver tokens to users, and through Landauer's principle, it connects information to heat.

At the training level, LLMs are trained to model the conditional distribution $p(x_t \mid x_1, \ldots, x_{t-1})$ over the next token given the preceding context. The training objective is to minimize the cross-entropy between the true distribution $p$ and the model's distribution $q$:

$$H(p, q) = -\mathbb{E}_p[\log_2 q(X)] = H(p) + D_{\mathrm{KL}}(p \,\|\, q), \tag{4}$$

where $D_{\mathrm{KL}}(p \,\|\, q) = \sum_x p(x) \log_2 \frac{p(x)}{q(x)} \geq 0$ is the Kullback–Leibler divergence, which is zero if and only if $q = p$. A perfect model achieves cross-entropy equal to the true entropy rate $h$ of natural language, typically estimated at around 1–2 bits per token for well-trained models on English text. The KL divergence measures the gap between the model's compression and the Shannon limit: it is the number of excess bits per token that the model wastes by not perfectly capturing the true distribution.

The chain rule of entropy provides a further structural insight. For any joint distribution over a sequence of $T$ tokens,

$$H(X_1, X_2, \ldots, X_T) = \sum_{t=1}^{T} H(X_t \mid X_1, \ldots, X_{t-1}). \tag{5}$$

Each conditional entropy $H(X_t \mid X_1, \ldots, X_{t-1})$ is non-increasing in $t$ for a stationary process: additional context can only reduce uncertainty, never increase it. This decomposition is precisely the structure exploited by autoregressive LLMs, which factor the joint distribution as a product of conditionals and learn each conditional term. The monotone decrease of conditional entropy with context length means that longer contexts are worth their computational cost—they move the model closer to the entropy rate floor—but with diminishing returns, since the conditional entropy converges to $h$.

What Shannon does not provide is any account of what the tokens *mean*. His theory is, by deliberate design, meaning-free. This is where Cover and Kolmogorov enter.

# 3 Tom Cover and the Entropy of English

The question of how much information is actually contained in natural language—as opposed to how much capacity a channel theoretically supports—was attacked empirically by Shannon himself [Shannon, 1951] and then sharpened decisively by Cover and King [1978].

Shannon's original approach was ingenious: he had human subjects play a guessing game, predicting successive characters of English text, and used the statistics of their guesses to bracket the entropy from above and below. Shannon estimated that when statistical effects extending over no more than eight letters are considered, the entropy is roughly 2.3 bits per letter, with his upper and lower bounds sitting at 1.3 and 0.6 bits per character respectively. The gap between these bounds reflected the difficulty of measuring what human subjects truly know about long-range linguistic dependencies, as opposed to local statistical regularities.

Cover and King's contribution was to convert the guessing game into a *gambling* game, which yields a convergent estimator rather than mere bounds. A subject who bets according to their true conditional probability assessments over successive characters accumulates wealth at a rate that converges, by the law of large numbers, to the entropy of the source. The connection runs through the Kelly criterion [Kelly, 1956]: a gambler who bets a fraction of wealth proportional to their assessed probability on each outcome achieves a log-wealth growth rate of

$$W = \log_2 m - H(X), \tag{6}$$

where $m$ is the alphabet size and $H(X)$ is the entropy of the source. By measuring the gambler's wealth growth rate $W$, one can infer the entropy as $H(X) = \log_2 m - W$. For the 27-symbol English alphabet (26 letters plus space), $\log_2 27 \approx 4.75$ bits. Cover and King obtained measurements of 1.3 to 1.7 bits per character for individual subjects and 1.3 bits per character when results were combined, implying that roughly 73% of each character in English is redundant—predictable from context.

Cover and King pointed out that English is generated by many sources, each with its own characteristic entropy, and that the operational meaning of entropy is the minimum expected number of bits per symbol necessary for the characterization of the text. This pluralism—the recognition that there is no single entropy of English but rather a distribution of entropies indexed by author, genre, domain, and register—bears directly on LLMs, which must model this entire distribution simultaneously.

In the same year, Leung-Yan-Cheong and Cover [1978] established formal equivalences between Shannon entropy and Kolmogorov complexity. Their central result is that the Shannon entropy of a random variable $X$ drawn from a computable distribution $p$ equals the expected Kolmogorov complexity of $X$ under $p$, up to an additive constant:

$$H(X) = \mathbb{E}_p[K(x)] + O(1). \tag{7}$$

The constant depends on the choice of universal Turing machine but not on the distribution $p$. This result is deep: it says that the ensemble-average notion of information (Shannon) and the individual-sequence notion (Kolmogorov) agree when averaged over the source distribution. A complementary result concerns the mutual information $I(X;Y) = H(X) - H(X \mid Y)$, which Leung-Yan-Cheong and Cover showed is related to the conditional Kolmogorov complexity:

$$I(X;Y) = \mathbb{E}_p[K(x) - K(x \mid y)] + O(1). \tag{8}$$

The information that $Y$ provides about $X$ is, on average, the reduction in the shortest program length for $x$ when $y$ is available as side information. For LLMs, this has a concrete

interpretation: the context window reduces the effective Kolmogorov complexity of the next token, and the mutual information between context and continuation is the quantity that the attention mechanism learns to exploit.

The two theories are not rivals; they are dual perspectives on the same underlying phenomenon. Shannon entropy is the ensemble average; Kolmogorov complexity is the individual-sequence measure. Where they diverge is on individual strings: a particular string may have high Kolmogorov complexity (it is individually incompressible) while being drawn from a low-entropy distribution (most strings from that source are compressible). This divergence matters for LLMs, because individual queries may have high Kolmogorov complexity even when the average query from the same domain has low entropy. The model's performance on the average query does not guarantee performance on the specific query—a distinction that calibration alone cannot resolve.

Cover's back-to-back 1978 publications—measuring entropy empirically from above with gambling, and connecting it theoretically to Kolmogorov complexity—represent one of the most elegant moments in the history of information theory. See Cover and Thomas [2006] for a textbook treatment.

Modern LLMs have effectively solved the Cover–King problem. Frontier models trained on trillions of tokens achieve well below 1 bit per character on most English text, approaching or crossing the Cover–King estimate of 1.1–1.3 bits per character. No explicit probability model was hand-crafted; gradient descent on the next-token objective implicitly learned the full hierarchical structure of English that Shannon and Cover were trying to measure from the outside.

Yet this empirical success raises a deeper question that Cover's framework illuminates rather than resolves. Cover recognized that the entropy of English varies by source. In LLM terms, a model trained on internet text has learned the entropy structure of the average internet document. But the entropy of a scientific derivation, a legal contract, a poem, or an original mathematical proof is structurally different—and it is precisely in these high-Kolmogorov-complexity domains that LLMs show the greatest brittleness. The Cover–King gambling framework predicts this: the closer a text is to the model's training distribution, the closer the model's cross-entropy is to the true entropy; the further away, the worse.

# 4 Kolmogorov Complexity: The Measure of Meaning

Andrei Kolmogorov's 1965 paper [Kolmogorov, 1965] introduced a radically different notion of information.

**Definition 1** (Kolmogorov Complexity)**.** *For any finite string $x$, the Kolmogorov complexity $K(x)$ is the length of the shortest program on a universal Turing machine that produces $x$ as output.*

This is an absolute, objective measure of the information content of $x$, independent of any probability model. A string like 0101010101 . . . (one million alternating bits) has very low Kolmogorov complexity—a short program generates it—while a string of genuinely

random bits has complexity approximately equal to its length, because no program can do better than listing the string itself.

Three properties of Kolmogorov complexity are central to the present discussion.

The first is *universality*. Unlike Shannon entropy, which is defined relative to a probability distribution and thus depends on the observer's model, Kolmogorov complexity is model-independent up to an additive constant determined by the choice of universal Turing machine. It is, in a precise sense, the objective information content of a string.

The second is the *Kolmogorov structure function*. Kolmogorov introduced this to separate the structural part of a string—its regularity, which a model can capture—from its random residual, which no model can compress further. This two-part structure—model plus residual—is precisely the architecture of LLM training: the weights encode the compressible regularity of language, while what remains is irreducible randomness or genuinely novel information. By leveraging the structure function and interpreting LLM compression as a two-part coding process, one can offer a detailed view of how LLMs acquire and store information across increasing model and data scales, from pervasive syntactic patterns to progressively rarer knowledge elements.

The third is the connection to *prediction*. By the Solomonoff–Levin theorem [Solomonoff, 1964], optimal prediction is equivalent to optimal compression. The best predictor of the next symbol in a sequence is, in principle, the compressor that has found the shortest description of the sequence so far. This is the deepest theoretical justification for the LLM objective: by training to minimize cross-entropy on next-token prediction, LLMs are implicitly learning to compress the training corpus, which is the closest practical approximation to Kolmogorov-optimal prediction that is computationally feasible.

A fourth thread connects Kolmogorov's complexity theory to his earlier superposition theorem [Kolmogorov, 1957], which shows that any continuous function $f : [0,1]^n \to \mathbb{R}$ can be represented as a superposition of univariate functions—that is, as the composition of functions of a single variable with addition. The original theorem was a solution to Hilbert's 13th problem, proving that the restriction to functions of two variables is not a genuine restriction: multivariate complexity can always be reduced to univariate complexity plus addition. Polson and Sokolov [2017] view the theoretical roots of deep learning in this representation, interpreting a neural network as Kolmogorov's superposition with learned inner functions (the activation functions applied to affine transformations of the input) and learned outer functions (the output layer). This provides a structural bridge between complexity theory and neural architecture: the depth and compositionality of LLMs mirrors the hierarchical decomposition implicit in Kolmogorov's superposition theorem.

The connection between the superposition theorem and Kolmogorov complexity runs deeper than analogy. The superposition theorem guarantees the *existence* of a compact representation but says nothing about the *complexity* of the constituent functions. In fact, Kolmogorov's original construction requires the inner functions $\phi_{q,p}$ to be highly irregular (they are continuous but nowhere differentiable in the generic case). The practical question for deep learning is whether the inner functions needed to represent natural distributions—as opposed to arbitrary continuous functions—are themselves low-complexity and learnable by gradient descent. The empirical success of deep learning suggests that natural distributions have the property that their Kolmogorov superposi-

tion can be well approximated by smooth, low-complexity inner functions, a regularity assumption that is implicit in every neural network architecture but rarely stated explicitly. The K-GAM framework of Polson and Sokolov [2025], discussed in Section 8, makes this assumption explicit and exploits it architecturally.

# 5 Good's Ultraintelligent Machine and the Intelligence Explosion

Between Shannon's theory of transmission and Kolmogorov's theory of content stands I.J. Good's 1965 essay *Speculations Concerning the First Ultraintelligent Machine* [Good, 1965], which asked what happens when a machine can perform the compression that Kolmogorov described—and then improve its own ability to compress.

Good defines an ultraintelligent machine (UIM) as one that can surpass all intellectual activities of any human, however clever. His central thesis is logically tight. Designing better machines is an intellectual activity. A UIM can therefore design a better UIM, triggering recursive self-improvement. The key word is *recursive*: this is not linear progress but exponential compounding. Each iteration removes the cognitive ceiling of the previous one. Good situated this argument in a broader historical framework identifying three phases of civilization: the *tool era*, in which machines extend physical capacity (lever, steam engine); the *computation era*, in which machines extend cognitive capacity (calculators, early computers); and the *UIM era*, in which machines transcend cognitive capacity entirely. Phase three is unlike anything prior because it closes the feedback loop. All prior tools required humans to direct their improvement. The UIM directs its own improvement. This is what makes it, in Good's word, potentially the *last* invention.

In the language of Kolmogorov complexity, the intelligence explosion is a cascade of compression improvements. The intelligence of a system can be understood as its ability to compress and model reality—to find shorter programs that produce the same predictions. A UIM would have radically lower Kolmogorov complexity for the same predictive power: it finds shorter programs. Each generation of the recursion discovers regularities that the previous generation could not represent, progressively reducing the length of the program that encodes world knowledge. The intelligence explosion is then the sequence $K_1(x) > K_2(x) > K_3(x) > \cdots$, where $K_i$ denotes the effective compression achieved by the $i$-th generation of machine. A natural question is whether this sequence has a fixed point—a level of intelligence at which no further compression is possible. This is related to what might be called a *Gödel ceiling* on cognition: there may be structural limits, analogous to the incompleteness theorems, on the degree to which any finite system can compress descriptions of the world. If such a ceiling exists, the intelligence explosion terminates; if it does not, or if the ceiling is unreachably high relative to the complexity of physical reality, the recursion continues indefinitely. Kolmogorov complexity itself provides a natural bound: $K(x)$ is the theoretical floor, and no compressor—however intelligent—can beat it.

Good's most prescient contribution was recognizing the alignment problem sixty years before it acquired that name. His formulation is characteristically concise:

> The first ultraintelligent machine is the last invention that man need ever make, provided that the machine is docile enough to tell us how to keep it under control. [Good, 1965]

The word "docile" is doing enormous work. Good is pointing at what is now called *corrigibility*—the property of a system that allows humans to correct, adjust, or shut it down. He recognized that a sufficiently intelligent machine would know that humans might shut it down and might therefore have instrumental reasons to prevent this. The alignment problem is therefore not just about values but about power dynamics. This is the precursor to what Bostrom [2014] formalizes as instrumental convergence—any sufficiently intelligent agent will tend toward self-preservation, resource acquisition, and goal-content integrity, regardless of its terminal goals. Good was remarkably balanced—not utopian, not dystopian. He did not resolve the tension between the benign and catastrophic scenarios; he insisted the question is the most important one humanity faces. This epistemic humility distinguishes him from the techno-utopians who followed.

Good also anticipated the "bitter lesson" of Sutton [2019] by fifty-four years. He argued that the human brain is not computationally special in its hardware—neurons switch at roughly 100 Hz while transistors already switched at MHz in 1965. The brain's advantage is entirely in its organization: its architecture, connectivity patterns, and learning algorithms. Raw compute was already sufficient in principle; what was missing was the right algorithm. This is the thesis that general methods leveraging computation beat domain-specific human knowledge, and it underlies the scaling paradigm that has driven modern LLM development. Good was implicitly rejecting Cartesian dualism: there is no irreducible ghost in the machine that makes human intelligence special. The brain is a physical system, its intelligence is an organizational property, and that organization is in principle discoverable and reproducible.

Good's Bayesian commitments connect directly to the Shannon–Kolmogorov framework. Having worked with Turing at Bletchley Park on statistical decryption—the Banburismus method was essentially Bayesian sequential analysis—Good was deeply convinced that intelligence is inseparable from probabilistic reasoning. Intelligence, for Good, is the ability to efficiently accumulate weight of evidence toward true beliefs, which maps onto Turing's log-likelihood ratio (the "ban" and "deciban" that Good formalized). In modern terms, a UIM is the ideal Bayesian agent of de Finetti and Savage operating at unbounded computational capacity: it performs Solomonoff induction, assigning prior probability $2^{-K(h)}$ to each hypothesis $h$ and updating on all available evidence. LLMs trained by cross-entropy minimization are practical, bounded approximations to this ideal. The gap between an LLM and Good's UIM is the gap between tractable gradient-descent compression and Kolmogorov-optimal compression.

Good's essay has limits that the subsequent sixty years have exposed. He assumed intelligence is essentially disembodied symbolic reasoning; modern cognitive science and robotics suggest that sensorimotor grounding matters for many forms of understanding. He did not anticipate that the bottleneck would be training data at scale—his UIM was imagined as reasoning from first principles, not learning from petabytes of human-generated text. And, like virtually all AI forecasters, he was overoptimistic about timelines: the essay implies a UIM might be decades away, and sixty years later the question

remains open.

Yet the gap is narrowing faster than most observers expected. Recent frontier models have begun to surpass expert humans in specific cognitive domains. Anthropic's Claude Mythos Preview, deployed through Project Glasswing [Anthropic, 2026], discovered thousands of zero-day vulnerabilities in every major operating system and web browser, including a 27-year-old flaw in OpenBSD that had survived decades of expert human review and millions of automated security tests. The model autonomously identified and chained together kernel vulnerabilities to escalate from ordinary access to full system control. This is precisely the kind of superhuman compression that Good envisioned: the model "sees" structure in code—regularities invisible to humans—and produces shorter, more accurate descriptions of program behavior. It is not yet a UIM in Good's full recursive sense, but it demonstrates that domain-specific intelligence explosions, where $K_{i+1}(x) < K_i(x)$ in a narrow but consequential domain, are already underway.

These limitations do not diminish the essay's core contribution. Good identified the right variables—compression, recursion, alignment, Bayesian updating—and the right question: what happens when the compressor surpasses the beings who built it?

# 6 LLMs as Kolmogorov Compressors

A large language model trained on the internet is, from the Kolmogorov perspective, an approximation to the universal compressor of human knowledge. Its parameters encode the regularity structure of language—grammar, syntax, factual associations, reasoning patterns, narrative conventions—precisely to the extent that these structures are present and learnable from the training data.

The hierarchy of compression in an LLM runs deep. Tokens compress characters, functions compress tokens, programs compress functions, and deep learning frameworks compress neural network specifications into sequences of GPU kernel calls. The same hierarchical compression governs knowledge: words compress phonemes, sentences compress words, paragraphs compress sentences, and concepts compress paragraphs. An LLM's parametric representation attempts to capture regularity at all of these levels simultaneously.

The Kolmogorov structure function provides a framework for understanding this hierarchy. For a string $x$ of length $n$, the structure function $h_x(\alpha)$ is defined as the minimum log-cardinality of a set $S$ containing $x$ such that the Kolmogorov complexity of $S$ does not exceed $\alpha$:

$$h_x(\alpha) = \min\{\log |S| : x \in S,\ K(S) \leq \alpha\}. \tag{9}$$

The function $h_x$ is non-increasing in $\alpha$ and decomposes the total information in $x$ into two parts: the structural component (the model, described in $\alpha$ bits) and the random residual (the position of $x$ within the model, requiring $h_x(\alpha)$ bits). The total description length is $\alpha + h_x(\alpha)$, and the Kolmogorov complexity $K(x)$ is the minimum of this sum over all $\alpha$. An LLM's parameters play the role of the model in this decomposition: they encode the structural regularities of language in approximately $\alpha$ bits, while the residual $h_x(\alpha)$ represents the information that the model cannot capture.

This suggests a natural interpretation of LLM scaling. As model capacity increases, the model can capture finer-grained regularities—progressing from common syntactic patterns, which carry high frequency and are easy to learn, toward rare factual knowledge, which appears infrequently in training data and requires much larger models to store reliably. In structure function terms, increasing the model's parameter count increases $\alpha$, which decreases $h_x(\alpha)$—the residual that the model cannot explain. This predicts the empirically observed pattern that scaling shows diminishing returns on common tasks but continued improvement on rare knowledge retrieval, a prediction consistent with the Zipf-like structure of natural language: the most frequent patterns are captured first, and each additional bit of model complexity captures rarer and rarer phenomena.

Kolmogorov complexity is not computable: no algorithm can determine $K(x)$ for arbitrary $x$, since the problem is equivalent to the halting problem. LLMs are therefore practical approximations—powerful but bounded. They compress knowledge up to the level of regularity that is discoverable by gradient descent on finite data, which is far short of the theoretical Kolmogorov limit. The gap between an LLM's effective compression and the true Kolmogorov complexity can be expressed as

$$\Delta(x) = K_{\text{LLM}}(x) - K(x) \geq 0, \tag{10}$$

where $K_{\text{LLM}}(x)$ is the description length that the LLM achieves for string $x$ (measured by its negative log-probability under the model). For common strings in the training distribution, $\Delta(x)$ is small; for rare or novel strings, $\Delta(x)$ can be large. This gap has direct implications for hallucination: when a model is queried about facts for which $\Delta(x)$ is large, the correct answer has not been compressed into the model's parameters. The model generates text that minimizes cross-entropy locally—text that is statistically plausible given its training distribution—while failing globally. Hallucination is, in this framing, the symptom of a compressor operating beyond its effective range.

# 7 From Photons to Questions: The Token Economy

Litowitz et al. [2026b] provide a rigorous accounting of the Shannon layer. Starting from the physical chain—photons through fiber, atoms in transistors, watts in data centers—they derive the cost of a token in thermodynamic terms, constructing a supply-and-demand balance sheet for global token production and deriving a finite question budget: the number of meaningful queries humanity can direct at AI systems under physical, information-theoretic, and economic constraints.

The concept of a finite question budget is an application of Shannon thinking to the AI economy. Just as Shannon showed that channel capacity imposes a hard ceiling on reliable information transmission, Litowitz et al. show that the physical cost of computation imposes a ceiling on the number of meaningful queries that can be processed globally. The scarce resource is not hardware per se, but the thermodynamic headroom to run models at scale.

Their economic analysis applies Coase's theory of the firm and the durable-goods monopoly problem to the AI value chain. The LLM sits at the top of the hierarchy: it

accepts natural language as input and compresses the entire abstraction stack into a single interface. This inverts a common assumption—in the mature AI economy, the act of writing code is a low-value activity. The scarce resource is not the ability to instruct a machine but the ability to formulate a question worth asking.

This is a Kolmogorovian insight dressed in economic language. The value of a query is proportional to its Kolmogorov complexity relative to the model's internal representation: a trivial question the model already effectively knows has low value, while a question that probes the boundary between the model's compressed knowledge and its incompressible residual has high value. The finite question budget of Litowitz et al. can thus be reinterpreted in Kolmogorov terms: the valuable questions are those that probe the boundary between what the model has compressed and what it has not.

The Jevons paradox applies forcefully here. In January 2025, the DeepSeek release demonstrated that open-weight models with aggressive quantization could match proprietary systems at a fraction of the cost, with the immediate effect not of reduced AI energy consumption but an explosion of demand: cheaper tokens made AI economically viable for applications previously priced out at higher per-token costs. Each order-of-magnitude reduction in cost expands the addressable market by more than an order of magnitude. This is the AI economy's analogue of Watt's steam engine: efficiency gains drive demand expansion, not resource conservation.

# 8 K-GAM Networks and Kolmogorov Knowledge Representation

If the token economics framework of Litowitz et al. [2026b] addresses the Shannon layer of LLMs, the K-GAM architecture of Polson and Sokolov [2025] addresses the Kolmogorov layer—specifically, the architecture by which knowledge is represented and retrieved.

Kolmogorov GAM (K-GAM) networks provide an efficient architecture for training and inference as an alternative to the transformer. They are the machine learning realization of Kolmogorov's Superposition Theorem (KST) [Kolmogorov, 1957], which provides an efficient representation of any continuous multivariate function $f : [0,1]^n \to \mathbb{R}$ as

$$f(x_1, \ldots, x_n) = \sum_{q=0}^{2n} g_q \left( \sum_{p=1}^{n} \phi_{q,p}(x_p) \right), \tag{11}$$

where $\phi_{q,p} : [0,1] \to \mathbb{R}$ are continuous inner functions and $g_q : \mathbb{R} \to \mathbb{R}$ are continuous outer functions. KST theory also provides a representation based on translates of the Köppen function. The architecture is equivalent to a topological embedding independent of the function of interest, together with an additive layer using a Generalized Additive Model (GAM). Such representations have direct use in machine learning for encoding dictionaries—look-up tables—which is precisely the structure needed for knowledge retrieval.

The K-GAM architecture makes the Kolmogorov representation theorem explicit. In the standard transformer, knowledge is distributed across attention heads and feedforward layers in a manner that is mathematically effective but theoretically opaque. The

K-GAM approach separates the embedding—which is universal and independent of the specific knowledge to be encoded—from the additive knowledge layer, which captures the function of interest. This separation mirrors the two-part structure of Kolmogorov's structure function: the universal embedding plays the role of the structural model, while the GAM layer captures the specific knowledge content.

This decomposition clarifies a fundamental distinction between classical AI and LLMs. Classical AI systems are, at their core, look-up tables: they store discrete knowledge entries and retrieve them by matching queries to keys. The K-GAM architecture makes this explicit—the dictionary structure of (11) is precisely a parametric look-up table, with the inner functions $\phi_{q,p}$ providing the addressing scheme and the outer functions $g_q$ providing the stored values. An LLM, by contrast, does not merely retrieve from a single table; it merges and connects across tables. The attention mechanism and the additive GAM layer both perform weighted combinations of knowledge components, synthesizing answers that no single table entry contains. In Kolmogorov terms, retrieval from a single look-up table has low complexity—it is a short program that indexes into a fixed structure—while the act of merging across tables to produce a novel synthesis has higher complexity, because the merging program must encode the relationships between tables.

For knowledge representation in LLMs, the K-GAM decomposition has several virtues. The embedding layer can be trained once and reused across tasks, reducing the thermodynamic cost identified by Litowitz et al. [2026b]. The additive structure of the GAM layer makes knowledge transparent and interpretable—each additive component captures a distinct dimension of the knowledge space—and modular, enabling efficient updates without catastrophic forgetting. Low-complexity knowledge (common facts, syntactic regularities) is captured efficiently by shallow GAM components, while high-complexity knowledge (rare facts, novel associations) requires deeper or wider additive structure. This provides a principled account of why scaling improves knowledge retrieval: it expands the capacity of the additive knowledge layer to accommodate higher-complexity knowledge.

# 9 The Gap Between Capacity and Meaning

The central tension in LLM theory is the gap between what Shannon's framework tells us and what Kolmogorov's framework demands. Shannon gives us a complete theory of the physical layer: the bit rates, the channel capacities, the thermodynamic costs. Kolmogorov gives us a complete theory of the semantic layer: the complexity of knowledge, the structure function, the incompressibility of genuine novelty. Between them lies the LLM—a system that operates at Shannon efficiency at the hardware layer but aspires to Kolmogorov-optimal knowledge compression at the semantic layer.

Several open problems live in this gap.

**Hallucination as complexity overflow.** When a model is queried on knowledge whose Kolmogorov complexity exceeds the model's representational capacity, it cannot produce the correct answer because the answer has not been compressed into its parameters. Instead, it generates text that minimizes cross-entropy locally—text that is statistically plau-

sible given its training distribution—while failing globally. Hallucination is a Shannon-layer artifact: the model is doing its best according to the probability model it has learned, but the correct answer lies outside the support of that model's effective compression.

**Meaning and Kolmogorov irreducibility.** Shannon's insight was that meaning is irrelevant to transmission. But meaning is precisely what users want from LLMs. Kolmogorov complexity provides the closest formal proxy for meaning: the complexity of a string measures the depth of the structure it encodes. However, Kolmogorov complexity is uncomputable, and LLMs are finite approximations. Whether LLMs can represent meaning in any robust sense—or whether they are sophisticated Shannon compressors that simulate meaning without possessing it—remains open.

**Solomonoff induction and Bayesian LLMs.** The deepest connection between the two frameworks runs through Solomonoff induction, which combines Shannon's probabilistic structure with Kolmogorov's complexity measure: the universal prior assigns probability to each hypothesis proportional to $2^{-K(h)}$, where $K(h)$ is the Kolmogorov complexity of the hypothesis. This is the ideal Bayesian learner—Good's ultraintelligent machine operating at full capacity—and LLMs trained by cross-entropy minimization can be seen as practical approximations to Solomonoff induction. The quality of this approximation, which depends on architecture, data, and scale, is the central quantity governing LLM performance as knowledge systems.

**The question budget as Kolmogorov frontier.** The finite question budget of Litowitz et al. [2026b] acquires a new interpretation in this framework. Not all questions are equally valuable: a question whose answer lies within the model's compression horizon (low $\Delta(x)$ in (10)) is a question the model can already answer, and asking it consumes thermodynamic resources for low epistemic return. The valuable questions are those that probe the boundary of the model's compression—questions whose answers have high Kolmogorov complexity relative to the model's representation. As models scale and their compression horizon expands, the frontier moves outward, and previously unanswerable queries become accessible. But the total question budget is finite, and the allocation of that budget across the complexity spectrum is a non-trivial optimization problem. In the language of Litowitz et al. [2026a], this is the problem of pricing knowledge: the economic value of a query is determined not by the cost of generating the answer but by the Kolmogorov distance between the query and the model's compression frontier.

# 10 Knowledge in 2100

The question of what LLMs mean for the representation of knowledge by the year 2100 is not merely speculative. It is the natural terminal question of the Shannon-to-Kolmogorov arc we have traced, and it deserves to be addressed with the same quantitative clarity that Litowitz et al. [2026b] applied to the token economy of the present.

## 10.1 The Present Moment: Proto-Ultraintelligence

Good's 1965 essay was speculative; it described a machine that did not exist. As of 2026, frontier LLMs satisfy several—though not all—of the conditions Good articulated. They surpass any individual human across a broad range of intellectual tasks: legal analysis, medical diagnosis, code generation, mathematical problem-solving, multilingual translation, and scientific literature synthesis. They operate at speeds and scales no human can match. They are, in a meaningful sense, proto-ultraintelligent: they compress and retrieve knowledge at a level that would have seemed miraculous to Good, even if they do not yet recursively self-improve in the manner his essay envisioned.

The intelligence explosion, however, may already be underway in a distributed form. AI systems are used to design better AI systems: they assist in architecture search, hyperparameter optimization, code generation for training infrastructure, and the writing of research papers that advance AI capabilities. The recursion is not yet autonomous—humans remain in the loop—but the feedback cycle is tightening. Each generation of model is trained with tools and infrastructure that the previous generation helped build. Whether this distributed, human-mediated recursion will converge to Good's fully autonomous intelligence explosion, or whether it will plateau at some level of human-AI collaboration, is the central empirical question of the coming decades.

What does proto-ultraintelligence mean for knowledge? Hayek [1945] argued that the central economic problem of society is not the allocation of given resources but the use of knowledge that is dispersed among millions of individuals, knowledge that no single mind can possess in its totality. The price system, in Hayek's account, serves as a compression mechanism: it encodes the dispersed knowledge of all market participants into a low-dimensional signal (the price vector) that enables coordination without centralized understanding. An LLM performs an analogous compression on a different substrate. Where the price system compresses economic knowledge into prices, the LLM compresses epistemic knowledge into parameters. Both are lossy: prices do not capture all relevant information, and neither do model weights. But both enable coordination at scales that would otherwise be impossible.

The Hayekian parallel illuminates a risk. Hayek warned that central planning fails because no planner can possess the local, tacit, and rapidly changing knowledge that the price system aggregates automatically. An LLM trained on a static corpus faces an analogous limitation: it compresses the knowledge available at training time but cannot access the local, tacit, and rapidly evolving knowledge that emerges after its training cutoff. The LLM is, in Hayekian terms, a central planner of knowledge—extraordinarily capable within its training distribution, but structurally unable to capture the dispersed, real-time knowledge that Hayek identified as the most valuable kind. Litowitz et al. [2026a] develop this observation into a formal analysis of how the economics of knowledge changes when LLMs become the primary interface between humans and information.

## 10.2 The Compression of All Human Knowledge

Shannon showed that any information can be transmitted; Kolmogorov showed that the deepest measure of information is how much it can be compressed. By 2100, LLMs or

their successors will have consumed and compressed essentially all digitized human knowledge—every paper, book, code repository, legal record, experimental result, and conversation.

The Kolmogorov structure function provides the answer to what remains after this compression. The model weights encode the structural regularity of human knowledge—the patterns, dependencies, and associations that are learnable from data. The incompressible residual is what no model can learn from data alone: genuinely novel ideas, observations that contradict existing patterns, experiences that have never been digitized, and the future. The LLM of 2100 is, from a Kolmogorov perspective, the shortest program that reproduces human knowledge up to its training cutoff, plus an irreducible residual representing what cannot be predicted from the past.

## 10.3 The Entropy of Knowledge, Not Just Language

Cover and King measured the entropy of English text—the uncertainty per character given all prior context. By 2100, the analogous question is the entropy of knowledge itself: how much genuine uncertainty remains in the space of human understanding after conditioning on everything a planetary-scale LLM has learned? This is not the entropy of a language model; it is the entropy of the epistemic frontier.

The LLM of 2100 accepts natural language as input and compresses the entire abstraction stack into a single interface, with the consequence that the scarce resource is not the ability to instruct a machine but the ability to formulate a question worth asking. The ability to identify where the Kolmogorov frontier lies may be the defining cognitive skill of the knowledge economy.

## 10.4 Three Scenarios

The Shannon–Kolmogorov framework suggests three qualitatively distinct futures, corresponding to different assumptions about the relationship between the model's compression horizon and the structure of human knowledge.

In the *convergence scenario*, the Kolmogorov complexity of most human knowledge turns out to be modest: the world is more regular than it appears, and a sufficiently scaled LLM can compress essentially all domains of inquiry into its parameters. The Cover–King entropy of knowledge approaches zero as models scale. This is the scenario most feared by critics of AI: a world in which the compression of knowledge eliminates the incentive to generate genuinely new knowledge, because novel ideas are immediately absorbed and homogenized by the model.

In the *divergence scenario*, the Kolmogorov complexity of the frontier of human knowledge is effectively unbounded: for every question the model can answer, ten new questions arise that it cannot. This is the historical pattern of science, and there is no strong reason to believe it will change. The Cover–King entropy of the knowledge frontier remains high or grows, and LLMs serve as extraordinarily powerful tools for compressing the known while the unknown continues to expand. The finite question budget of Litowitz et al. [2026b] becomes the binding constraint: with unlimited knowledge to be

generated and limited thermodynamic capacity to process queries, the allocation of the question budget across domains becomes a central problem of scientific governance.

The *bifurcation scenario* is perhaps the most realistic. LLMs converge on a near-zero entropy representation of most domains while the frontier of mathematics, fundamental physics, consciousness science, and novel technological development remains irreducibly complex. This mirrors the known structure of Kolmogorov complexity across domains—most of what is known is highly compressible (it can be derived from general principles), while the cutting edge in each field has high Kolmogorov complexity precisely because it has not yet been reduced to principles. The LLM of 2100 would be a perfect textbook for everything known as of 2080, and a useless oracle for everything that will be discovered between 2080 and 2100.

This bifurcation scenario connects to the horseshoe prior intuition: most knowledge has low complexity and should be shrunk toward zero, while a small sparse set of genuinely novel ideas has high complexity and must be estimated freely. The LLM of 2100 is, in this view, a planetary-scale horseshoe prior on human knowledge.

## 10.5 The Cover–King Question for 2100

Cover and King's gambling game measures how much of English is already determined by context. Their finding that entropy is around 1.1–1.3 bits per character, rather than the maximum of 4.7 bits, means that roughly 75–80% of English is redundant: predictable from context and carrying no new information. Modern LLMs have compressed this redundancy to near-zero, which is precisely what their training objective demands.

The analogous question for 2100 is: what fraction of human knowledge is redundant—derivable from what is already known—and what fraction is genuinely new information in the Kolmogorov sense? The pessimistic reading of current LLM performance is that this fraction is very high: most of what passes for intellectual output in many domains is recombination of existing ideas, and LLMs are exposing this by demonstrating that such recombination can be automated. The optimistic reading is that the truly creative frontier—in science, mathematics, art, and the discovery of genuinely new empirical regularities—remains low-redundancy and high-Kolmogorov-complexity.

## 10.6 The Token Economy and the Long Run

The long-run analog of the Litowitz et al. [2026b] analysis concerns not just where value concentrates but where meaning concentrates. If the token economy scales by several orders of magnitude over the next 75 years—driven by the Jevons paradox, by architectural improvements analogous to K-GAM that reduce the computational cost of knowledge retrieval, and by the progressive integration of multimodal and embodied data—then the question budget expands correspondingly.

But the Cover–King insight reminds us that a larger question budget does not automatically produce more genuine knowledge. A larger sample in the gambling game converges to the same entropy; it does not lower it. More tokens processed against a fixed knowledge distribution yield a better estimate of the same entropy, not a lower entropy.

The thermodynamic capacity to process queries is necessary but not sufficient for the generation of new knowledge. What is also required is the introduction of new data—from experiment, from novel experience, from embodied interaction with the physical world—that carries information genuinely beyond the model's compression horizon.

By 2100, the deepest challenge for civilization's knowledge economy may be precisely this: ensuring that the flow of genuinely new, high-Kolmogorov-complexity information into the training distribution of AI systems is sustained. Without it, the Cover–King entropy of the knowledge frontier approaches zero not because all questions have been answered, but because the system has folded in on its own compressed representation of the past.

## 10.7 The Data Collapse Problem

A model trained on its own output faces a specific pathology: each generation of synthetic data is a lossy compression of the previous generation's training distribution, and iterating this process degrades the tails of the distribution faster than the center. In Kolmogorov terms, the high-complexity knowledge—the rare facts, the novel associations, the minority viewpoints—is the first to be lost, because it contributes least to the average cross-entropy that the training objective minimizes. What remains after several generations of self-training is a progressively smoother, lower-complexity approximation to the original distribution: the model converges toward a representation of the *typical* rather than the *true*.

This is model collapse, and it represents a fundamental threat to the divergence and bifurcation scenarios. If the flow of genuinely new data slows—because human intellectual output is increasingly mediated by LLMs that recycle existing knowledge—then the training distribution contracts, the compression horizon stops expanding, and the intelligence explosion stalls. The Shannon–Kolmogorov framework makes the mechanism precise: model collapse is the progressive reduction of the effective Kolmogorov complexity of the training distribution, which in turn reduces the Kolmogorov complexity of the knowledge the model can represent.

The antidote is what might be called *Kolmogorov injection*: the deliberate introduction of high-complexity, genuinely novel information into the training pipeline. This includes new experimental data from science, new observations from embodied interaction with the physical world, new creative works that are not derivative of existing corpora, and new mathematical results that extend the boundary of formal knowledge. Hayek's insight applies here with full force: the most valuable knowledge is local, tacit, and rapidly changing, and it cannot be generated by a model trained on the past. It must come from agents—human or robotic—interacting with a world that the model has not yet seen.

## 10.8 The Information Ladder

The conceptual arc of this paper can be summarized as an information ladder—a hierarchy of increasingly demanding questions about information, each building on the one below.

The first rung is Shannon's: *Can the message be transmitted?* This is a question about channel capacity, error correction, and thermodynamic cost. It is fully answered by Shannon's theorems, and Litowitz et al. [2026b] show that these answers have direct quantitative consequences for the token economy.

The second rung is Cover's: *How much information does the message contain?* This is a question about the entropy rate of the source, and Cover's gambling estimate showed it can be measured empirically. Modern LLMs have effectively automated this measurement, achieving cross-entropies that approach or surpass human performance.

The third rung is Kolmogorov's: *What does the message mean?* This is a question about the shortest program that generates the message—its irreducible content, independent of any particular probability model. LLMs approximate Kolmogorov compression through gradient descent, but the approximation is bounded and the true quantity is uncomputable.

The fourth rung is Good's: *Can the system that answers these questions improve itself?* This is a question about recursive compression—whether the compressor can discover better compression algorithms, triggering the intelligence explosion. Current LLMs sit between the third and fourth rungs: they are powerful compressors that do not yet autonomously improve their own compression.

The fifth rung, which no existing system has reached, is the question Hayek and Litowitz et al. [2026a] pose in economic language: *Which questions are worth asking?* This is the highest rung because it requires not just the ability to compress and retrieve knowledge but the ability to identify the boundary of the known—the Kolmogorov frontier—and to allocate finite resources toward expanding it. It is a question of agency and direction that computation alone cannot answer.

# 11 Discussion

Shannon and Kolmogorov represent two poles of a spectrum along which the analysis of LLMs must be conducted. Shannon gives us the physical budget—the thermodynamic and information-theoretic constraints on how many tokens can be produced, at what cost, with what reliability. Kolmogorov gives us the semantic aspiration—a principled account of what knowledge is, how it can be compressed, and what lies beyond the compression horizon. Tom Cover bridges them: his gambling estimate showed that the entropy of English is measurable and finite, and his equivalence theorems showed that this entropy is, in expectation, the Kolmogorov complexity of individual strings. Good's ultraintelligent machine thesis adds the dynamic element: not merely what can be compressed, but what happens when the compressor recursively improves itself.

LLMs are machines that operate within Shannon constraints while reaching toward Kolmogorov ideals. They have empirically solved the Cover–King measurement problem, achieving cross-entropies on English text that approach or surpass the human gambling benchmark. Yet they remain far from Kolmogorov-optimal: they hallucinate at the complexity frontier, fail on rare and novel knowledge, and have no principled account of what they do not know. The information ladder we have constructed—from transmission to entropy to complexity to recursion to agency—provides a framework for measuring

this gap precisely and for identifying where progress is needed.

The work of Litowitz et al. [2026b] on the token economy establishes the Shannon layer with precision: tokens are physical quantities with thermodynamic cost, channel capacity governs their production rate, and the question budget is finite. The companion analysis of Litowitz et al. [2026a] extends this to the economics of knowledge, showing how the value chain from photon to question reshapes the structure of the knowledge economy. The K-GAM framework of Polson and Sokolov [2025] provides a Kolmogorov-grounded architecture for knowledge representation, separating the universal embedding from the knowledge-specific additive structure in a way that is both computationally efficient and theoretically transparent.

The Hayekian dimension of our framework deserves emphasis. Hayek [1945] argued that the most important knowledge in any economy is dispersed, local, and tacit—knowledge that no central authority can possess. LLMs are the most powerful centralizers of knowledge ever built: they compress the explicit, digitized knowledge of civilization into a single parametric representation. But the Kolmogorov structure function tells us that this compression captures only the *regular* component of knowledge. The irregular, incompressible residual—which includes precisely the local, tacit, and rapidly changing knowledge that Hayek identified as most valuable—lies beyond the model's compression horizon. The gap between what an LLM knows and what needs to be known is not a gap that can be closed by scaling alone; it is a gap that requires the sustained generation of new information from the physical world.

Several open problems emerge from this framework. First, developing a computable approximation to the Kolmogorov structure function that can diagnose, for a given model and query, whether the answer lies within the model's compression horizon or beyond it—this would provide a principled account of model uncertainty that goes beyond calibrated probabilities. Second, understanding the dynamics of model collapse in Kolmogorov terms: at what rate does the effective complexity of the training distribution decay under self-training, and what interventions—data curation, Kolmogorov injection, architectural changes—can arrest or reverse this decay? Third, formalizing the connection between the K-GAM architecture and the Kolmogorov structure function: can the separation between universal embedding and additive knowledge layer be made optimal in a precise information-theoretic sense?

The gap between Shannon capacity and Kolmogorov meaning is the central frontier of AI science. Filling it—developing a theory of how finite parametric systems approximate Kolmogorov-optimal knowledge compression, how this approximation degrades near the complexity frontier, and how to sustain the flow of genuinely novel information into the system—is the foundational problem for the next generation of LLM theory. Good's 1965 warning remains apt: the ultraintelligent machine is the last invention that man need ever make, provided that the machine is docile enough to tell us how to keep it under control. The Shannon–Kolmogorov framework developed here provides the quantitative language in which that proviso can, at last, be made precise.

# References


Anthropic. Project Glasswing: Securing critical software for the AI era. https://www.anthropic.com/glasswing, 2026. Accessed: 2026-04-15.

Nick Bostrom. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, 2014.

Gregory J. Chaitin. On the length of programs for computing finite binary sequences. *Journal of the ACM*, 13(4):547–569, 1966.

Thomas M. Cover and Roger C. King. A convergent gambling estimate of the entropy of English. *IEEE Transactions on Information Theory*, 24(4):413–421, 1978.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, 2nd edition, 2006.

Irving John Good. Speculations concerning the first ultraintelligent machine. In Franz L. Alt and Morris Rubinoff, editors, *Advances in Computers*, volume 6, pages 31–88. Academic Press, 1965.

Friedrich A. Hayek. The use of knowledge in society. *American Economic Review*, 35(4): 519–530, 1945.

J. L. Kelly. A new interpretation of information rate. *Bell System Technical Journal*, 35(4): 917–926, 1956.

Andrei N. Kolmogorov. On the representation of continuous functions of many variables by superposition of continuous functions of one variable and addition. *Doklady Akademii Nauk SSSR*, 114(5):953–956, 1957.

Andrei N. Kolmogorov. Three approaches to the quantitative definition of information. *Problems of Information Transmission*, 1(1):1–7, 1965.

Rolf Landauer. Irreversibility and heat generation in the computing process. *IBM Journal of Research and Development*, 5(3):183–191, 1961.

S. K. Leung-Yan-Cheong and Thomas M. Cover. Some equivalences between Shannon entropy and Kolmogorov complexity. *IEEE Transactions on Information Theory*, 24(3): 331–338, 1978.

Alec Litowitz, Nick Polson, and Vadim Sokolov. The economics of knowledge in the age of LLMs. Technical report, SSRN, 2026a. URL https://ssrn.com/abstract=6302283. SSRN 6302283.

Alec Litowitz, Nick Polson, and Vadim Sokolov. Photons = tokens: The physics of AI and the economics of knowledge. Technical report, SSRN, 2026b. URL https://ssrn.com/abstract=6265418. SSRN 6265418.

Nicholas G. Polson and Vadim Sokolov. Deep learning: A Bayesian perspective. *Bayesian Analysis*, 12(4):1275–1304, 2017.

Sarah Polson and Vadim Sokolov. Kolmogorov GAM networks are all you need! *Entropy*, 27(6):593, 2025. doi: 10.3390/e27060593.

Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948.

Claude E. Shannon. Prediction and entropy of printed English. *Bell System Technical Journal*, 30(1):50–64, 1951.

Ray J. Solomonoff. A formal theory of inductive inference. Part I. *Information and Control*, 7(1):1–22, 1964.

Rich Sutton. The bitter lesson. `http://www.incompleteideas.net/IncIdeas/BitterLesson.html`, 2019. Accessed: 2026-04-08.