The Evolution of Human Intelligence: An Information-Theoretic Perspective on Cognitive Evolution (Version 3.4)

Michael Zhang

June 10, 2026

Abstract

This paper advances a cross-disciplinary hypothesis: the rapid expansion of human intelligence may be interpreted as an evolutionary process of semantic energy efficiency maximization (η_{sem}) across cognitive modalities. We suggest that the transition from vision to speech to writing constitutes a sequence of externalized information compression, potentially enabling substantial increases in semantic energy efficiency within the brain's constrained 20-watt metabolic budget. Drawing on neuroscience, linguistics, and information theory, we propose that symbolic systems function as entropy-minimizing mechanisms that help distinguish humans from larger-brained species lacking durable symbolic abstraction. Integrating the Free Energy Principle (FEP), we model this transition as the minimization of variational free energy through the stabilization of discrete context, which may reduce the Kullback-Leibler (KL) divergence between internal neural models and sensory input. We further distinguish raw bitrate ($\dot{B}_{\text{raw}}$) from semantic bitrate ($\dot{B}_{\text{sem}}$), and argue that cognitive scaling is better interpreted through reduced filtering cost per semantic bit than through increased raw throughput alone. Within the toy-model and order-of-magnitude framework developed here, the semantic efficiency is estimated to rise from roughly $\sim$16.7 bits/J in vision to $\sim$500 bits/J in writing, but these values should be read as model-dependent comparative estimates rather than as definitive empirical constants. On that basis, the paper offers a unifying and testable framework for thinking about human exceptionalism, with possible implications for neuro-diagnostics, efficient education, and scaling laws in artificial intelligence.

Keywords: intelligence, artificial intelligence, cognition, information theory

1 Introduction

Human intelligence enables unparalleled achievements—mathematics, philosophy, and global civilizations—yet operates within a constrained $\sim$20-watt brain energy budget (Herculano-Houzel, 2012). Traditional explanations emphasizing neuron count or brain size falter: elephants possess 257 billion neurons, yet lack human-like abstraction (Herculano-Houzel et al., 2014). This paper proposes that intelligence arises from information entropy reductions across cognitive modalities—visual perception, spoken language, and writing—allowing exponential efficiency gains without escalating energy demands. Human intelligence's definition in the academic and psychology fields is a complex and multidimensional concept. Although multiple theoretical frameworks and measurement methods exist, there is no single, completely unified "authoritative" definition. The American Psychological Association (APA) defines intelligence

as "the ability of individuals to reason, plan, solve problems, think abstractly, understand complex concepts, learn quickly, and learn from experience" (APA, 1995). This definition emphasizes the multidimensionality and contextual adaptability of intelligence and is widely considered an authoritative overview. In the early 20th century, Charles Spearman proposed the "general intelligence" (g-factor) theory, positing that intelligence is a single, universal cognitive ability extractable via statistical methods like factor analysis from various cognitive tasks (Spearman, 1904). The g-factor represents overall performance in tasks such as logical reasoning, language understanding, and spatial ability, laying the foundation for modern intelligence quotient (IQ) testing, which remains influential in psychology. Information entropy, defined by Shannon as $H = -\sum p(x) \log p(x)$, measures uncertainty in a system (Shannon, 1948). High-entropy visual data requires intensive neural processing, while language and writing compress redundancy, reducing computational load (Cover & Thomas, 2006). Evolutionary transitions mirror this: visual systems in primates supported survival, but language enabled symbolic abstraction, and writing stabilized knowledge transmission (Tomasello, 2014). Wittgenstein's language games emphasize meaning through use, while Chomsky's universal grammar suggests innate entropy-handling mechanisms (Chomsky, 1965; Wittgenstein, 1953). Modern large language models (LLMs), like OpenAI's o1, demonstrate emergent abilities—e.g., 83.3% accuracy on AIME 2024 math problems (OpenAI, 2024)—driven by entropy minimization during training, paralleling human cognitive leaps (Wei et al., 2022). LLMs passing a standard three-party Turing test in a 2026 report (Jones & Bergen, 2026) illustrates that large neural language models can imitate human conversational behavior under specific prompting conditions. However, their lack of embodied experience contrasts with human evolution (Bender & Koller, 2020). This paper integrates neuroscience, linguistics, and information theory, using empirical data (e.g., fMRI, archaeological records) and mathematical models to argue that entropy reduction across modalities drives intelligence. Sections cover biological foundations, modality transitions, entropy modelling with bit-rate comparisons, LLM analogies, integrative analysis, conclusion, and applications.

1.1 Biological Foundations: Neural Substrates and Energy Constraints

The human brain's efficiency is central to intelligence, operating at 20 watts despite 86 billion neurons, including 16 billion in the neocortex (Herculano-Houzel, 2009). This contrasts with elephants (257 billion neurons, mostly cerebellar) and whales (37 billion cortical neurons), whose cognitive limits stem from lacking symbolic systems (Herculano-Houzel et al., 2014; Manger, 2006). Isotropic fractionator studies confirm human neural density optimizes computational capacity within energy constraints (von Bartheld et al., 2016). Developmentally, infants start with 100 billion neurons, pruning to 86 billion by adulthood, with synaptic peaks at age 2 (10^{15} connections) refining networks via Hebbian learning—"neurons that fire together wire together" (Huttenlocher & Dabholkar, 1997; Lent et al., 2012). This pruning reduces neural entropy by eliminating noisy connections, minimizing energy use (Laughlin, 2001). Language areas (Broca's, Wernicke's), involving 1 billion neurons, use white matter tracts like the arcuate fasciculus for efficient signal transfer (Friederici, 2009). Von Economo neurons (VENs) in fronto-insular cortices, denser in humans, support rapid social processing and correlate with intelligence (Allman et al., 2005). fMRI studies link resting-state brain entropy to intelligence measures, suggesting that neural variability can be behaviorally informative rather than mere noise (Saxe et al., 2018). Examples: Twin studies show 76% heritability in cortical thickness, tied to cognitive performance (Thompson et al., 2001). Lesion studies, like Phineas Gage's frontal damage impairing decision-making, highlight neural specificity (Damasio, 1994).

Energy constraints necessitate entropy-efficient systems. In this paper, the power terms for vision, speech, and writing are treated as order-of-magnitude bookkeeping parameters (not direct modality-isolated calorimetry): visual-heavy processing is approximated near the few-watt scale, language on the sub-watt to 1-watt scale, and writing-assisted cognition lower still via externalization.

1.2 High-Entropy Stage: Visual Perception and Cognitive Load

Visual perception, the earliest cognitive modality, is entropy-intensive due to raw sensory data's redundancy. The visual cortex, 30% of the neocortex, receives a retinal/optic-nerve output stream on the order of $\sim 10^7$ bits/second, while natural images have entropy 4–8 bits/pixel due to spatial correlations (Felleman & Van Essen, 1991; Kersten, 1987; Koch et al., 2006). This high-entropy stream demands significant neural resources, in our model represented as a few watts (central value: 6 W, derived from a 30% share of a 20 W brain budget) rather than a direct measured "visual-only" wattage (Laughlin, 2001). Evolutionary context: Australopithecus relied on visual cues for foraging and predator avoidance in high-entropy environments (Schoenemann, 2013). Cave art (e.g., Lascaux, ∼40,000 years ago) shows proto-symbolic attempts, with entropy 6 bits/symbol due to coarse depictions (Hoffmann et al., 2018). Example: Chauvet Cave's animal paintings required pattern recognition but lacked abstraction, reflecting high cognitive load. Cognitive evidence: Hick's law ($RT = a + b \log_2 N$) links reaction time to stimulus entropy, showing exponential processing costs (Hick, 1952). Primate studies (e.g., macaque object recognition) reveal limited abstraction without symbols (Tomasello, 2014). Human visual search tasks (e.g., finding a target in clutter) show high neural firing rates (100 Hz), correlating with entropy (Hubel and Wiesel, 1968).

Energy implication: Visual processing's high entropy limits cognitive expansion without compression, as it dominates cortical energy. Early hominins' small brains (400 cm^3) constrained further intelligence growth, necessitating a lower-entropy modality like language.

2 Entropy Reduction in Spoken Language

Spoken language marks a significant entropy reduction, transforming visual complexity into linear, symbolic sequences. English text entropy is 4.7 bits/character, with conditional entropy 1 bit/char due to grammatical constraints (Shannon, 1951). At 50 chars/s (10 words/s), processing is $\sim 10^3$ bits/s. We use 1 W as an order-of-magnitude anchor for language-dominant cortical engagement, while noting that Friederici (2009) is primarily a neuroanatomical review rather than a direct metabolic quantification. Evolutionary pathway: Gestural communication, leveraging mirror neurons in Broca's homolog, preceded speech, reducing entropy by sequencing actions (Corballis, 2002). Vocalization, enabled by FOXP2 mutations 300,000 years ago, introduced phonemes (H 3–5 bits/phoneme) (Lai et al., 2001). Example: Sign languages activate similar brain regions as speech, showing modality-independent compression (Emmorey, 2002). Chomsky's universal grammar posits innate mechanisms for entropy-efficient syntax, explaining children's rapid acquisition despite limited input (Chomsky, 1965). Usage-based theories counter that social learning extracts patterns, reducing entropy through exposure (Tomasello, 2003). Example: Children's overregularization ("goed") minimizes syntactic entropy by generalizing rules (Marcus et al., 1992). Social drivers: Dunbar's hypothesis links neocortex size to group size (150 individuals), with gossip compressing social information (Dunbar, 1998, 2004). Animal comparison: Vervet monkey alarm calls (H 2 bits/call) show proto-symbolic compres-

sion, but lack generativity (Seyfarth et al., 1980). Human language's displacement (referring to absent entities) enabled planning, reducing entropy further (Hockett, 1960). Energy savings: language processing relies on structured fronto-temporal networks rather than high-bandwidth sensory filtering (Friederici, 2009). This $\sim 10^4$-fold bit-rate reduction (from $\sim 10^7$ to $\sim 10^3$ bits/s) enables complex cognition within energy limits.

3 Writing: The Pinnacle of Entropy Compression

Writing achieves the lowest entropy by fixing transient speech into stable symbols, enabling exponential knowledge growth. Alphabetic systems yield conditional entropy 0.5–1 bit/symbol, compared to language's 1 bit/char and vision's 8 bits/pixel (Montemurro & Zanette, 2011). At 20 symbols/s (reading/writing), processing is $\sim 10^2$ bits/s, with part of the computational burden offloaded externally. The 0.2 W value is therefore an illustrative internal-load proxy, not a directly measured writing-energy constant; Wolf (2007) is used for literacy/cognitive framing. The emergence of writing realized the combination of human brain memory and auxiliary external storage, expanding human cognitive levels and thus enhancing intelligence (Clark & Chalmers, 1998). Archaeological evidence: Sumerian tokens (8,000 BCE, H 5 bits/token) evolved into cuneiform for accounting and narratives (Schmandt-Besserat, 1992). Alphabets (e.g., Phoenician, 1050 BCE) reduced symbol sets, lowering entropy (Dehaene, 2014). Example: Egyptian hieroglyphs enabled mathematics, with low-entropy symbols for fractions (Imhausen, 2016).

Cognitive impact: Literacy repurposes visual cortices for word recognition, enhancing abstraction with minimal energy (Dehaene & Cohen, 2007). Material engagement theory views writing as an "extended mind," offloading memory (Malafouris, 2013). Example: Greek alphabet fostered philosophy, with Plato's dialogues critiquing oral traditions (Goody, 1977). Philosophical implications: Wittgenstein's language games stabilize via writing, preventing ambiguity and enabling logic (Wittgenstein, 1953). Luria's studies show literates excel at low-entropy syllogisms, unlike oral cultures (Luria, 1976). Modern example: Programming languages, as low-entropy scripts, mirror writing's role in AI (Knuth, 1997). Energy efficiency: Writing externalizes memory, reducing neural firing rates, as EEG studies confirm lower cortical activity in literate tasks (Wolf, 2007). On the modality-level bookkeeping used throughout this paper, writing operates near $\sim 10^2$ bits/s, implying an overall reduction of roughly 10^5 in raw bitrate relative to vision. This compression helps enable cumulative symbolic systems such as Euclidean geometry (Montemurro & Zanette, 2011; Netz, 1999).

Thermodynamic perspective (Landauer cost): according to Landauer's principle, irreversibly erasing one bit requires at least

$$E_{\min} = k_{\mathrm{B}} T \ln 2,$$

which is about $\sim 3 \times 10^{-21}$ J/bit at physiological temperature ($T \approx 310$ K) (Landauer, 1961; Bennett, 2003). Writing systems preserve intermediate information in physical media (paper, clay, and digital storage), reducing the need for frequent metabolically driven neuronal state resets (e.g., ionic-gradient restoration after repeated overwrite/update cycles). This lowers internal erasure burden and provides an important thermodynamic basis for the modality-level power shift from roughly 6 W (vision-dominant filtering) to 0.2 W (writing-assisted semantic extraction).

4 Mathematical Modelling of Entropy and Processing Demands

To quantify entropy reductions, we compare both entropy-density and raw-versus-semantic bitrate profiles across modalities, illustrating exponential decreases in processing and energy demands. Table 1 provides the unified comparison.

4.1 Theoretical Framework

4.1.1 FEP and KL Divergence

Following Friston (2010), biological systems minimize variational free energy to maintain adaptive stability. In cognitive terms, this can be viewed as reducing mismatch between internal generative models and external input distributions.

4.1.2 Raw vs. Semantic Bitrate

A key distinction is required between raw channel load ($\dot{B}_{\text{raw}}$) and behaviorally useful semantic throughput ($\dot{B}_{\text{sem}}$). The retina may deliver a massive but already compressed optic-nerve output stream, while cognition is bottlenecked by low-dimensional semantic control variables. Thus, efficiency gains arise mainly from lowering energetic cost P to transform raw input into semantic structure.

4.2 Efficiency Leap Across Modalities (Revised)

We define semantic energy efficiency as $\eta_{\text{sem}} = \dot{B}_{\text{sem}}/P$. Under a constant semantic target (order-of-magnitude $\sim$100 bits/s), modality transitions can be interpreted as reductions in required power: visual filtering ($\sim$6 W), verbal symbolic compression ($\sim$1 W), and writing-assisted externalization ($\sim$0.2 W). This yields an illustrative increase from about 16.7 to 500 bits/J.

4.3 Operational Method and Reproducibility Protocol

To make the quantitative claims auditable, we use a four-step estimation pipeline that is held constant across modalities. First, we define a modality-specific entropy density (H) from primary literature (e.g., bits/pixel for vision, conditional bits/character for language). Second, we map behavioral throughput to an effective symbol rate (r), yielding bitrate $\dot{B} = H \times r$. Third, we assign power draw P from neurophysiological allocation studies (visual cortex, language networks, and externally offloaded writing). Finally, we compute energy intensity per processed bit as $\varepsilon = P/\dot{B}$ and semantic efficiency as $\eta_{\text{sem}} = 1/\varepsilon$.

To reduce hidden degrees of freedom, all calculations in the main text and Appendix use the same conventions: (i) conditional entropy is preferred over raw entropy when grammar or context strongly constrains prediction; (ii) ranges are preserved instead of collapsed to single-point estimates unless explicitly reported as central approximations; (iii) unit conversions are carried out in SI units before cross-modality comparison. This design avoids inflating between-modality differences through inconsistent measurement scales.

A critical methodological choice is distinguishing *internal* neural computation from *externalized* symbolic computation. In writing, part of the inferential burden is transferred to durable artifacts (text, notation, tables). We therefore treat writing as a coupled human-artifact system rather than as cortex-only throughput. This assumption is theoretically aligned with the extended mind framework and empirically motivated by literacy studies showing reduced online memory load during symbolic problem solving.

4.4 Toy Model and Minimal Reproducible Framework

To move the argument from a purely narrative comparison toward a reproducible quantitative protocol, we define a minimal cross-modality task in which the *same latent world state* S must support the *same downstream decision* Y. The modality changes only the representational channel: an image X_{vis}, a spoken description X_{speech}, or a written description X_{write}. The goal is not to claim that these channels are identical in every respect, but to compare how efficiently each supports task-relevant inference under a shared target.

Formally, let useful information be the task-relevant mutual information

$$I_{\text{use}} = I(S; Y \mid X_m),$$

where $m \in \{\text{vis}, \text{speech}, \text{write}\}$ indexes modality. In a minimal operational version, I_{use} can be approximated either by retained proposition count or by task performance on matched question-answer pairs generated from the same scene. We then define semantic efficiency as

$$\eta_{\text{task}} = \frac{I_{\text{use}}}{E_m},$$

where E_m is the effective internal energy cost, or a validated computational proxy, required to convert modality-specific input into a correct task response.

A reproducible toy implementation would proceed in four steps. First, construct a matched dataset of scenes, each paired with (i) an image, (ii) a spoken-language transcript, and (iii) a written description, all designed to preserve the same core propositions. Second, define a fixed downstream task such as binary verification, short-answer question answering, or relation extraction. Third, measure a modality-specific processing-cost proxy—for humans, reaction time or neural-load estimates; for artificial systems, operation count, token-processing cost, or wall-clock energy proxy. Fourth, compute

$$\eta_{\text{task},m} \propto \frac{\text{task score or retained propositions}}{\text{processing-cost proxy}},$$

which yields a directly auditable cross-modality comparison.

Under this framework, the present paper should be read as offering an order-of-magnitude *theoretical instantiation* of that protocol rather than a completed benchmark. Table 1 therefore serves as a coarse model of what such a benchmark might reveal: visual channels carry very high raw load, speech compresses that load into sequential symbols, and writing further stabilizes symbols in external memory, reducing internal filtering cost per useful semantic unit. This toy model does not eliminate all representation assumptions, but it makes explicit what must be held constant for future empirical tests.

4.5 Necessity, Sufficiency, and Boundary Conditions of the Framework

The present argument is strongest as a *sufficient* explanatory framework, not as a strict necessity proof. In other words, the toy model shows how stepwise symbolic compression *could* generate the observed ordering in semantic efficiency, but it does not prove that every path to human-like intelligence must pass through the exact same modality sequence or through writing in its historically familiar form. A strict necessity claim would require demonstrating that no alternative representational architecture could achieve comparable semantic efficiency under similar metabolic constraints. That stronger result is beyond the scope of the present paper.

What the framework does claim is more limited and, we argue, more defensible: under severe energy constraints, systems that preserve task-relevant structure while reducing internal filtering burden should enjoy a robust comparative advantage. In that sense, symbolic externalization is treated not as the only conceivable attractor, but as a particularly plausible and historically realized low-entropy basin. The explanatory target is therefore a constrained evolutionary regularity rather than a universal law.

This distinction also clarifies the boundary conditions of the model. The argument is expected to work best when (i) the task requires stable reuse of intermediate representations, (ii) external symbolic storage is available, and (iii) semantic performance rather than raw sensory richness is the relevant objective. It is expected to work less well in domains dominated by fast sensorimotor control, affective salience, or ecological novelty, where high-bandwidth embodied interaction may remain indispensable. These boundary conditions help specify where entropy compression should be predictive and where it should not.

Table 1: Unified Entropy, Bitrate, and Energy Profile Across Cognitive Modalities (illustrative)

Modality	**Entropy (bits/unit)**	$\dot{D}_{\text{raw}}$ **(bits/s)**	$\dot{D}_{\text{sem}}$ **(bits/s)**	P **(W)**	η_{sem} **(bits/J)**	**Example**
Visual Perception	4–8 bits/pixel	$\sim 10^7$	~ 100	3–8 (central: 6)	12.5–33.3 (central: 16.7)	Natural scene analysis
Spoken Language	∼4.7 bits/char (1 bit/char conditional)	$\sim 10^3$	~ 100	0.3–1 (central: 1)	100–333 (central: 100)	English speech (10 words/s)
Writing	∼0.5–1 bit/symbol (conditional)	$\sim 10^2$	~ 100	0.1–0.3 (central: 0.2)	333–1000 (central: 500)	Alphabetic text reading/writing

Notes for Table 1:

- **Visual**: Entropy from Kersten (1987); processing from retinal ganglion cells (10^7 bits/s) (Koch et al., 2006).
- **Language**: Entropy from Shannon (1951); speech rate ∼10 words/s, ∼5 chars/word (Piantadosi et al., 2011).
- **Writing**: Conditional entropy from Montemurro & Zanette (2011); reading/writing ∼20 symbols/s (Wolf, 2007).
- **Energy**: Values are model-level approximations for internal neural load allocation (vision: few-W scale; language: sub-W to 1 W scale; writing-assisted cognition: lower internal load via externalization), rather than modality-isolated calorimetric measurements (Laughlin, 2001; Friederici, 2009; Clark & Chalmers, 1998).

4.6 Analysis of Exponential Reduction

Visual Perception: Entropy ∼4–8 bits/pixel reflects spatial redundancy. Processing $\sim 10^7$ bits/s is mapped to a few-watt internal cost (central estimate: ∼6 W from a 30% allocation heuristic) (Laughlin, 2001). Example: Macaque visual tasks show high firing rates (∼100 Hz), correlating with entropy (Hubel and Wiesel, 1968). Cognitive load scales exponentially with scene complexity, per Hick's law ($RT = a + b \log_2 N$) (Hick, 1952).

Spoken Language: Entropy drops to ∼4.7 bits/char, with conditional entropy ∼1 bit/char due to grammar (Shannon, 1951). At 50 chars/s, processing is $\sim 10^3$ bits/s, mapped to a sub-watt to 1 W internal-load regime (illustrative midpoint: ∼1 W) (Friederici, 2009). This $\sim 10^4$-fold reduction from vision enables syntactic abstraction, while evidence on brain entropy is treated separately as a broad neural-variability measure rather than a direct firing-rate estimate (Saxe et al., 2018).

Writing: Conditional entropy ∼0.5–1 bit/symbol, as alphabets optimize predictability (Montemurro & Zanette, 2011). At 20 symbols/s, processing is $\sim 10^2$ bits/s; the ∼0.2 W term is

an illustrative proxy for reduced *internal* load under external symbolic scaffolding, not a direct wattmeter measurement of writing (Wolf, 2007; Clark & Chalmers, 1998). Example: Chinese logograms (H∼2 bits/character) vs. alphabets (H∼0.5 bits/symbol) show trade-offs (Bentz et al., 2017).

Under the present order-of-magnitude framework, the estimated bitrate decreases from $\sim 10^7$ (vision) to $\sim 10^3$ (language) to $\sim 10^2$ (writing), corresponding to an illustrative reduction of roughly 10^5 across modalities. Effective internal energy allocation is modeled to decrease in parallel, from about 6 W to about 0.2 W, as writing offloads part of the inferential burden to external symbolic supports (Clark & Chalmers, 1998). In this interpretive sense, the free energy principle is consistent with the idea that stable symbolic media can reduce predictive uncertainty by stabilizing context, rather than serving as a direct empirical proof that writing uniquely minimizes F (Friston, 2010).

Thermodynamically, intelligence leverages negentropy to counter entropy increase, with writing as a dissipative structure (Schrödinger, 1944; Prokopenko et al., 2009). Example: Voynich manuscript's low conditional entropy (H∼0.3 bits/symbol) suggests optimized encoding, possibly artificial (Montemurro & Zanette, 2013).

Additional evidence: Mutual information ($I(X;Y) = H(X) - H(X|Y)$) measures semantic dependency, high in human language (e.g., I ∼2.5 bits/word in English) vs. animal signals (Piantadosi et al., 2011). Writing maximizes I by fixing context, enabling complex systems like mathematics (Netz, 1999).

4.7 Relation to Prior Theories

The present account overlaps with several existing traditions but is not identical to any one of them. Relative to classical information-theoretic accounts, our contribution is not the claim that cognition involves compression *per se*, but the more specific proposal that cross-modality cognitive scaling is better explained by reductions in filtering cost per useful semantic unit than by raw throughput alone. Relative to free-energy and predictive-processing approaches, we do not claim to replace those frameworks; rather, we interpret symbolic stabilization as one concrete route by which predictive systems may reduce uncertainty over longer timescales.

The framework also differs from neuron-count or cortical-expansion explanations. Those accounts emphasize representational capacity, whereas the present paper emphasizes representational control and external scaffolding. The intended claim is therefore complementary: larger cortical workspaces matter, but their impact is amplified when language, notation, and writing reduce uncertainty within that workspace. Finally, relative to cultural-evolution and extended-mind theories, our addition is a tentative quantitative bridge. We attempt to connect symbolic externalization not only to memory and culture, but also to an explicit efficiency metric that can, in principle, be benchmarked across modalities.

4.8 Sensitivity, Error Bounds, and Interpretation of Effect Size

Because each modality estimate aggregates heterogeneous experiments, uncertainty analysis is essential. Using conservative intervals from the literature already cited, visual bitrate remains in an ultra-high regime ($\sim 10^6$–10^7 bits/s), spoken language remains near 10^3 bits/s, and writing remains near 10^2 bits/s. Even when the most favorable assumptions are granted to speech and writing and the most conservative assumptions to vision, the ordering is unchanged by multiple orders of magnitude. Thus, the central result is not a fragile point estimate but a robust rank structure.

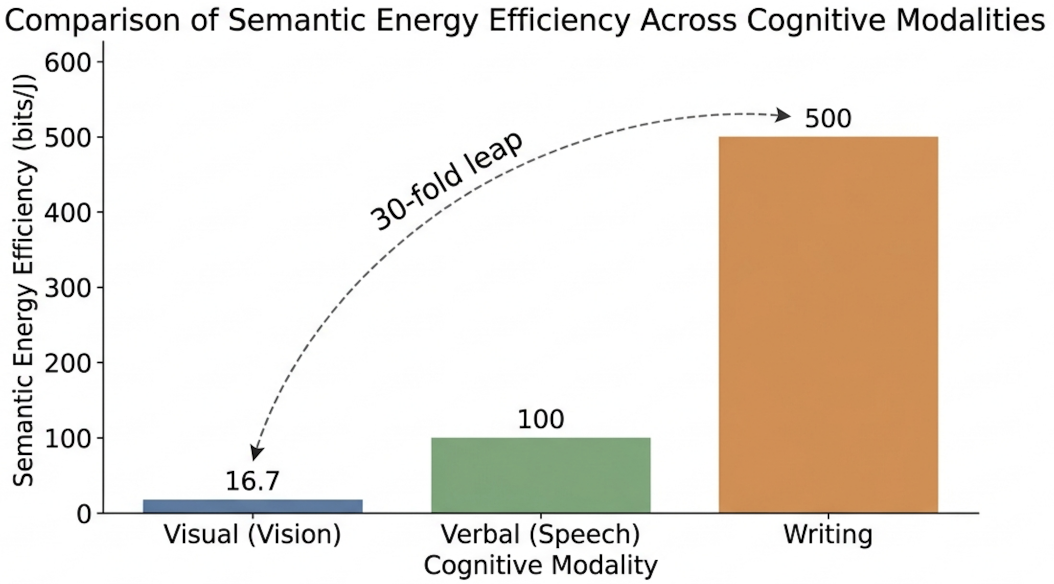

Figure 1: Illustrative, model-dependent comparison of semantic energy efficiency across cognitive modalities under the proposed toy-model framework.

This matters for interpretation. The proposed mechanism is *not* that humans simply process fewer bits over evolutionary time; rather, humans increasingly process *better-structured* bits with higher semantic yield per joule. In other words, cognitive scaling is achieved by reallocating metabolic budget from brute-force sensory filtering toward model-based symbolic manipulation. This distinction clarifies why lower bitrate at later modalities should be interpreted as increased efficiency rather than reduced capability.

The effect-size interpretation is therefore two-level: a computational level (entropy and bitrate compression) and a systems level (network reconfiguration plus external memory scaffolding). At the computational level, grammar and notation reduce uncertainty. At the systems level, social institutions (schools, archives, scientific notation, code repositories) stabilize these low-entropy channels across generations. The combined effect offers a plausible route from local neural constraints to cumulative cultural intelligence.

5 AI Analogies and Comparative Analysis

Large language models (LLMs), such as OpenAI's o1, exhibit emergent abilities—e.g., reasoning (83.3% on AIME 2024 (OpenAI, 2024)) and coding—at scale, driven by entropy minimization during training (Wei et al., 2022). Training minimizes cross-entropy loss, reducing predictive uncertainty (Vaswani et al., 2017). Example: o1's reported benchmark gains are consistent with the broader idea that training and inference procedures can reduce predictive uncertainty over task-relevant outputs (OpenAI, 2024; Wei et al., 2022). Recent three-party Turing-test results show that suitably prompted LLMs can imitate human conversational behavior in short interactions (Jones & Bergen, 2026). Case study: GPT-4's arithmetic performance improves from 13.4% to o1's 83.3% on AIME (OpenAI, 2024), reflecting improved performance under reasoning-oriented inference (OpenAI, 2024). This is analogous, at a high level, to the reduction of uncertainty in symbolic processing. However, LLMs can contain billions to hundreds of billions of parameters (Brown et al., 2020), whereas human cognition remains constrained by a roughly 20-watt biological energy budget (Herculano-Houzel, 2012).

Contrast: Human emergences, like recursion, arose from social pressures (e.g., storytelling for group cohesion), not just data scaling (Everett, 2005). Piraha language's lack of recursion suggests cultural modulation of entropy (Everett, 2005). LLMs lack embodied entropy from physical interaction, limiting their generality (Bender & Koller, 2020). Example: LLM training on code corpora (e.g., GitHub) reduces entropy to 0.2 bits/token, enabling programming proficiency, but errors in novel contexts reveal disembodied limits (Brown et al., 2020). Human writing, conversely, integrates sensory-social feedback, enhancing robustness.

6 Integrative Analysis

Human intelligence integrates neural hardware with entropy-reducing symbolic systems. The neocortex's 16 billion neurons provide computational capacity, but intelligence scales via entropy compression: visual (H 8 bits/pixel) to language (H 1 bit/char) to writing (H 0.5 bit/symbol) (Herculano-Houzel, 2009; Montemurro & Zanette, 2011). This 10^5-fold bit-rate reduction enables complex cognition within energy constraints.

Equation: $I \sim N \times (1 - H_{\text{norm}})$, where N is neuron count and H_{norm} is normalized entropy, quantifies intelligence as efficiency (Cover & Thomas, 2006). Social pressures drove this: larger groups required low-entropy communication, per Dunbar's hypothesis (Dunbar, 1998). Writing's externalization enabled philosophy and science, absent in neuron-rich species like elephants (Wittgenstein, 1953). LLM analogy: Emergent abilities reflect entropy-driven phase transitions, but lack embodiment (Wei et al., 2022). Human uniqueness lies in biological-social feedback, optimizing entropy within metabolic limits. Additional evidence: Bilingualism increases entropy tolerance, enhancing cognitive flexibility (Bialystok, 2017). Non-literate societies show higher-entropy reasoning, limited to concrete tasks (Luria, 1976). This synthesis highlights entropy reduction as the key to intelligence scaling.

An important interpretive consequence is that neuron count functions as a capacity term, whereas symbolic compression functions as a control term. Capacity without control generates high-volume but weakly structured processing; control without capacity constrains representational depth. Human intelligence appears exceptional because evolution and culture jointly improved both terms: cortical expansion increased representational workspace, while language and writing progressively reduced uncertainty within that workspace. This dual-term view also explains why large-brained mammals can show advanced perception and sociality yet fail to accumulate formal science.

The framework also predicts punctuated improvements when new representational standards appear. Historical examples include positional number systems, symbolic algebra, and formal logic, each of which reduces ambiguity and increases inferential reuse. On this account, major cognitive transitions are less like smooth linear growth and more like infrastructure upgrades that lower per-inference entropy cost for entire populations.

7 Boundary Conditions and Limitations

The present model is intentionally coarse-grained and has several limitations, which we address through the following theoretical boundaries.

7.1 Incommensurability as a Heuristic Proxy

First, cross-domain entropy units are only partially commensurable: bits/pixel (spatial redundancy), bits/character (sequential grammar), and bits/symbol (external stabilization) are compared through functional mapping, not strict thermodynamic identity. The $\sim 10^5$-fold reduction in this paper is therefore a *cognitive load heuristic proxy* rather than an absolute physical conversion rate. This is acceptable for qualitative scaling and order-of-magnitude reasoning, but limits fine-grained numerical claims. Future work should use task-relevant mutual information (MI) and unified encoding pipelines to improve comparability.

7.2 Gene-Culture Coevolution and Causality

Second, causality remains partly inferential. The framework might appear to suggest that symbolic compression unidirectionally caused intelligence. However, human cognitive evolution is better understood through Dual-Inheritance Theory (Boyd & Richerson, 1985). Genetic mutations (e.g., expanded working memory or prefrontal connectivity) provided the initial hardware capacity, which allowed for the invention of language and writing. These tools then created a low-entropy cultural niche (a "cultural ratchet"), which recursively eased the computational burden on subsequent generations. Thus, entropy compression and neural evolution are a positive feedback loop rather than a linear cause-and-effect.

7.3 The Necessity of High-Entropy "System 1"

Third, by emphasizing the efficiency of low-entropy symbolic systems, the model risks undervaluing high-bandwidth sensory processing. In the context of Dual-Process Theory (Kahneman, 2011), high-entropy visual processing (System 1) is neither "inefficient" nor obsolete; it remains essential for rapid, embodied survival and pattern recognition. Symbolic compression (System 2) is a specialized offloading mechanism for slow, deliberate abstraction. The brain's 20-watt budget is largely preserved for System 1, while language and writing maximize the utility of the remaining metabolic fraction reserved for abstract thought.

7.4 Methodological and AI Bounds

Fourth, energy estimates are population-level approximations derived from modality-specific tasks rather than simultaneous whole-brain calorimetry. Real cognition is multiplexed, meaning the reported wattage values should be interpreted as effective allocations, not isolated channel costs. Furthermore, the LLM analogy is informative but bounded. Training-loss minimization and human predictive processing share formal similarities, yet embodiment, developmental plasticity, and social grounding differ substantially. Consequently, AI comparisons should be treated as mechanistic analogies, not one-to-one ontological equivalences.

7.5 Falsifiable Predictions and Empirical Tests

Despite these limits, the framework is falsifiable rather than merely interpretive. A first prediction is that literacy interventions should increase semantic-efficiency metrics in neuroimaging tasks controlling for IQ baseline, especially on tasks requiring stable symbolic reuse rather than raw perceptual discrimination. A second prediction is that domains with stronger notational standardization should show lower error per unit cognitive effort than semantically matched domains lacking comparable symbolic conventions. A third prediction is that multimodal AI

systems with external memory and tool use should display efficiency profiles closer to human symbolic cognition than text-only models operating without durable external state.

More importantly, the toy model yields a threshold-style empirical prediction. If matched tasks are constructed so that the latent proposition set is held constant while representational stabilization is varied, then semantic efficiency should improve disproportionately when the task permits persistent external symbols and repeated reuse of intermediate structure. If that pattern does *not* appear—for example, if image-only, speech-only, and writing-assisted conditions yield indistinguishable efficiency once semantic content is controlled—then the central explanatory claim of the paper would be weakened. This makes the framework empirically vulnerable in a useful sense: it predicts not just a ranking, but a specific dependence on external symbolic stabilization.

8 Conclusion

This framework suggests that human intelligence may be understood less as a simple function of neuron count and more as a consequence of stepwise information compression across cognitive modalities. Within the toy-model and order-of-magnitude comparison developed here, the transition from vision to writing appears consistent with a substantial increase in semantic efficiency, potentially on the order of tens-fold, but this result should be interpreted as a model-dependent comparative estimate rather than as a settled empirical constant. By leveraging external symbolic systems, the human neocortex may function as a comparatively high-efficiency dissipative structure that reduces predictive surprise within a strict 20-Watt metabolic limit.

Furthermore, this information-theoretic model offers a possible mechanistic interpretation of Charles Spearman's general intelligence (g). Under this view, g can be read as the brain's capacity for effective entropy reduction—that is, the ability to extract task-relevant semantic structure from noisy input with limited energy expenditure. While species like elephants possess greater neuron counts, they do not appear to exhibit the same degree of externalized entropy-minimizing infrastructure, such as writing and mathematics, that supports cumulative human cognition.

In this sense, intelligence may be productively viewed as an evolving thermodynamic-information architecture: neural tissue supplies adaptive computation, while symbolic ecologies (speech communities, scripts, formal notations, and now programmable media) progressively lower uncertainty and amortize reasoning over time. Framed this way, the present paper is best read not as a final quantitative demonstration, but as a testable theoretical framework linking biological constraints, cultural acceleration, and machine intelligence research.

9 Implications and Future Work

The framework has several implications, but these should be interpreted as research directions rather than immediate applications. For artificial intelligence, the central implication is that scaling laws may be improved by tracking not only model size and loss reduction, but also the efficiency with which systems convert raw input into reusable semantic structure. On this view, multimodal systems with external memory, tool use, or explicit symbolic interfaces may provide a more informative analogue to human cognitive scaling than text-only systems alone (Bender & Koller, 2020).

For education and cognitive development, the framework suggests that literacy, notation,

and other structured symbolic practices may matter not merely because they add information, but because they reorganize how information is filtered and reused. This implies that future work should examine whether interventions in literacy, formal notation, or bilingual symbolic practice measurably alter semantic-efficiency proxies in behavior and neuroimaging, rather than relying only on broad outcome metrics such as test scores or IQ (Dehaene, 2014; Bialystok, 2017).

For neuroscience, the main implication is methodological. If entropy-based metrics are to become diagnostically useful, they must be tied to task-relevant semantic performance rather than to raw neural variability alone. This suggests a future program in which fMRI, EEG, and behavioral tasks are jointly designed to estimate semantic efficiency under matched symbolic conditions, potentially clarifying how different neurocognitive profiles alter the trade-off between raw processing load and useful inference (Saxe et al., 2018).

More broadly, future work should move in three directions: first, construct matched multimodal benchmarks that operationalize the toy model introduced here; second, test whether external symbolic stabilization produces threshold-like gains in semantic efficiency; and third, refine the present order-of-magnitude estimates with task-matched datasets, explicit energy proxies, and stronger cross-species comparisons. These steps would help determine whether the current framework remains a suggestive theoretical synthesis or matures into a more general quantitative account of cognitive evolution.

10 Acknowledgements

The author acknowledges the use of multiple large language model tools, including ChatGPT-5, Gemini 3 and Grok-4, to assist with literature organization, argument structuring, cross-checking of claims, and drafting of selected passages. Outputs from these tools were compared across models and were subject to human review, revision, and selective verification. The author remains solely responsible for the final text, interpretations, claims, and conclusions presented in this manuscript.

A Appendix A: Derivation of Bitrate and Energy Estimates

This appendix provides the detailed derivations underlying Table 1. For each modality, we report entropy assumptions, raw and semantic bitrates ($\dot{B}_{\text{raw}}$, $\dot{B}_{\text{sem}}$), estimated power consumption (P), and semantic energy efficiency ($\eta_{\text{sem}} = \dot{B}_{\text{sem}}/P$). Where possible, intermediate calculation steps are shown to ensure reproducibility.

A.1 Notation and Assumptions

- H: entropy per unit (bits/unit), e.g., bits/pixel or bits/character.
- $\dot{B}_{\text{raw}}$: raw channel load (bits/s).
- $\dot{B}_{\text{sem}}$: behaviorally useful semantic throughput (bits/s).
- P: estimated power (W).
- $\eta_{\text{sem}} = \dot{B}_{\text{sem}}/P$: semantic energy efficiency (bits/J).
- $\varepsilon = P/\dot{B}$: energy per processed bit (J/bit), reported where needed for cross-checking.

- Baseline brain power: $P_{\text{brain}} \approx 20$ W. Functional allocations are taken from literature cited in the main text.

A.2 Visual Perception

Entropy per pixel is estimated from natural-image statistics (Kersten, 1987):

$$H_{\text{vis}} \in [4, 8] \text{ bits/pixel.}$$

Total retinal/optic-nerve output is approximately (Koch et al., 2006):

$$\dot{B}_{\text{vis}} \approx 10^7 \text{ bits/s.}$$

With $\sim 30\%$ of brain energy devoted to visual cortex:

$$P_{\text{vis}} = 0.30 \times 20 \text{ W} = 6 \text{ W.}$$

Thus,

$$\varepsilon_{\text{vis}} = \frac{6}{10^7} \approx 6 \times 10^{-7} \text{ J/bit } (60\,\mu\text{J/bit}).$$

A.3 Spoken Language

To keep Appendix A consistent with Table 1, spoken language is reported using the same two-level bookkeeping as the main text: a raw-throughput term $\dot{B}_{\text{raw}}$ and a normalized semantic-throughput term $\dot{B}_{\text{sem}}$.

From Shannon's analysis (Shannon, 1951),

$$H_{\text{speech,raw}} \approx 4.7 \text{ bits/char}, \qquad H_{\text{speech,cond}} \approx 1 \text{ bit/char (conditional).}$$

With speech rate approximated as ~ 10 words/s and ~ 5 chars/word, we obtain

$$\text{chars/s} \approx 50,$$

so the direct lexical stream is

$$\dot{B}_{\text{speech,raw,lex}} = 4.7 \times 50 = 235 \text{ bits/s}, \qquad \dot{B}_{\text{speech,cond,lex}} \approx 1 \times 50 = 50 \text{ bits/s.}$$

These values are retained as conservative lower-level calculations. For the unified cross-modality comparison used in the tables, however, we follow the main-text order-of-magnitude convention and report

$$\dot{B}_{\text{raw,speech}} \sim 10^3 \text{ bits/s}, \qquad \dot{B}_{\text{sem,speech}} \sim 100 \text{ bits/s.}$$

Using the same central power estimate as Table 1,

$$P_{\text{speech}} \approx 1 \text{ W,}$$

we obtain

$$\eta_{\text{sem,speech}} = \frac{\dot{B}_{\text{sem,speech}}}{P_{\text{speech}}} \approx \frac{100}{1} = 100 \text{ bits/J.}$$

This is the semantic-efficiency quantity compared against vision and writing in the main text; the lexical calculations above should not be conflated with that normalized modality-level metric.

A.4 Writing and Reading

To keep Appendix A fully consistent with Table 1, writing is treated with the same two-level bookkeeping used in the main text: an order-of-magnitude raw-throughput term $\dot{B}_{\text{raw}}$, and a normalized semantic-throughput term $\dot{B}_{\text{sem}}$.

Conditional entropy per symbol is estimated from long-range text statistics (Montemurro & Zanette, 2011):

$$H_{\text{write}} \approx 0.5\text{–}1.0 \text{ bits/symbol.}$$

Observed reading/writing speed is taken as roughly ~ 20 symbols/s (Wolf, 2007). The literal product,

$$H_{\text{write}} \times r \approx 10\text{–}20 \text{ bits/s},$$

should be interpreted as a conservative lower-bound lexical stream. Table 1 instead reports the writing modality at the order-of-magnitude level,

$$\dot{B}_{\text{raw,write}} \sim 10^2 \text{ bits/s}, \qquad \dot{B}_{\text{sem,write}} \sim 100 \text{ bits/s},$$

so that the appendix uses the same normalized cross-modality comparison scale as the main text.

Because writing offloads memory and intermediate state to durable external symbols, the relevant power term is the reduced *internal* cortical load of the coupled human-artifact system rather than the total physical cost of inscription itself (Clark & Chalmers, 1998; Malafouris, 2013). Accordingly, we use the same illustrative central estimate as Table 1:

$$P_{\text{write}} \approx 0.2 \text{ W}.$$

This yields

$$\eta_{\text{sem,write}} = \frac{\dot{B}_{\text{sem,write}}}{P_{\text{write}}} \approx \frac{100}{0.2} = 500 \text{ bits/J},$$

with the Table 1 uncertainty range preserved as

$$P = 0.1\text{–}0.3 \text{ W} \quad \Rightarrow \quad \eta_{\text{sem,write}} \approx 333\text{–}1000 \text{ bits/J}.$$

In short, the 10–20 bits/s estimate is retained only as a low-end lexical calculation from symbol entropy, whereas the tabled $\sim 10^2$ and ~ 100 values are the modality-level quantities used for the paper's unified comparison.

A.5 Relative Reductions

Using the same modality-level raw-throughput bookkeeping as Table 1 and Table 2, with visual processing at 10^7 bits/s, spoken language at $\sim 10^3$ bits/s, and writing at $\sim 10^2$ bits/s, the relative reductions are

$$\frac{\dot{B}_{\text{raw,vis}}}{\dot{B}_{\text{raw,speech}}} \sim 10^4, \qquad \frac{\dot{B}_{\text{raw,vis}}}{\dot{B}_{\text{raw,write}}} \sim 10^5.$$

If one instead uses the conservative low-end lexical stream from Appendix A.4 ($\sim$ 10–20 bits/s), the writing reduction would be larger, on the order of 5×10^5 to 10^6. In this appendix, however, we retain $\sim 10^5$ as the reported writing reduction so that the appendix, Table 1, and the main-text comparison all use the same normalized modality-level scale.

A.6 Summary Table

Table 2: Appendix A: Unified Raw/Semantic Bitrate and Energy Estimates across Modalities

Modality	H (bits/unit)	$\dot{B}_{\text{raw}}$ (bits/s)	$\dot{B}_{\text{sem}}$ (bits/s)	P (W)	η_{sem} (bits/J)	Reduction vs visual ($\dot{B}_{\text{raw}}$)
Visual perception	4–8 bits/pixel	1×10^7	~ 100	6.0	16.7	baseline
Spoken language	4.7 bits/char ($\sim$1 cond.)	$\sim 10^3$	~ 100	1.0	100	$\sim 10^4$
Writing	0.5–1 bits/symbol (cond.)	$\sim 10^2$	~ 100	0.1–0.3 (central: 0.2)	333–1000 (central: 500)	$\sim 10^5$

A.7 Notes on Uncertainty

1. **Definition of bits**: This appendix reports both $\dot{B}_{\text{raw}}$ and $\dot{B}_{\text{sem}}$. Visual estimates are raw-input dominated, whereas language/writing are often conditional/semantic; direct comparison must keep this distinction explicit.

2. **Energy allocation assumptions**: Brain region percentages are coarse approximations. Absolute values should be interpreted cautiously.

3. **Exponential decreases**: On the normalized modality-level scale used in the tables, the reduction is approximately 10^4 for speech and 10^5 for writing; values approaching 10^6 arise only when the more conservative low-end lexical estimate from Appendix A.4 is used instead of the table-level writing throughput.

References

Allman, J. M., Watson, K. K., Tetreault, N. A., & Hakeem, A. Y. (2005). Intuition and autism: A possible role for Von Economo neurons. *Trends in Cognitive Sciences*, 9(8), 367–373. https://doi.org/10.1016/j.tics.2005.06.008

American Psychological Association. (1995). Intelligence: Knowns and Unknowns. *American Psychologist*, 50(8), 641–650.

Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5185–5198. https://aclanthology.org/2020.acl-main.463/

Bennett, C. H. (2003). Notes on Landauer's principle, reversible computation, and Maxwell's Demon. *Studies in History and Philosophy of Modern Physics*, 34(3), 501–510. https://doi.org/10.1016/S1355-2198(03)00039-X

Bentz, C., Alikaniotis, D., Cysouw, M., & Ferrer-i-Cancho, R. (2017). The entropy of words—Learnability and expressivity across more than 1000 languages. *Entropy*, 19(6), 275. https://doi.org/10.3390/e19060275

Bialystok, E. (2017). The bilingual adaptation: How minds accommodate experience. *Psychological Bulletin*, 143(3), 233–262. https://doi.org/10.1037/bul0000099

Boyd, R., & Richerson, P. J. (1985). *Culture and the evolutionary process*. University of Chicago Press.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901. https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

Chomsky, N. (1965). *Aspects of the theory of syntax*. MIT Press. https://mitpress.mit.edu/9780262527408/aspects-of-the-theory-of-syntax/

Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis*, 58(1), 7–19. https://doi.org/10.1093/analys/58.1.7

Corballis, M. C. (2002). *From hand to mouth: The origins of language*. Princeton University Press. https://press.princeton.edu/books/paperback/9780691088037/from-hand-to-mouth

Cover, T. M., & Thomas, J. A. (2006). *Elements of information theory* (2nd ed.). Wiley.

Damasio, A. R. (1994). *Descartes' error: Emotion, reason, and the human brain*. Putnam.

Dehaene, S. (2014). *Consciousness and the brain: Deciphering how the brain codes our thoughts*. Viking.

Dehaene, S., & Cohen, L. (2007). Cultural recycling of cortical maps. *Neuron*, 56(2), 384–398. https://doi.org/10.1016/j.neuron.2007.10.004

Dunbar, R. I. M. (1998). The social brain hypothesis. *Evolutionary Anthropology*, 6(5), 178–190. https://doi.org/10.1002/(SICI)1520-6505(1998)6:5<178::AID-EVAN5>3.0.CO;2-8

Dunbar, R. I. M. (2004). Gossip in evolutionary perspective. *Review of General Psychology*, 8(2), 100–110. https://doi.org/10.1037/1089-2680.8.2.100

Emmorey, K. (2002). *Language, cognition, and the brain: Insights from sign language research*. Lawrence Erlbaum Associates.

Everett, D. L. (2005). Cultural constraints on grammar and cognition in Pirahã: Another look at the design features of human language. *Current Anthropology*, 46(4), 621–646. https://doi.org/10.1086/431525

Felleman, D. J., & Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, 1(1), 1–47. https://doi.org/10.1093/cercor/1.1.1

Friederici, A. D. (2009). Pathways to language: Fiber tracts in the human brain. *Trends in Cognitive Sciences*, 13(4), 175–181. https://doi.org/10.1016/j.tics.2009.01.001

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11(2), 127–138. https://doi.org/10.1038/nrn2787

Goody, J. (1977). *The domestication of the savage mind*. Cambridge University Press.

Herculano-Houzel, S. (2009). The human brain in numbers: A linearly scaled-up primate brain. *Frontiers in Human Neuroscience*, 3, 31. https://doi.org/10.3389/neuro.09.031.2009

Herculano-Houzel, S. (2012). The remarkable, yet not extraordinary, human brain as a scaled-up primate brain and its associated cost. *Proceedings of the National Academy of Sciences*, 109(Suppl 1), 10661–10668. https://doi.org/10.1073/pnas.1201895109

Herculano-Houzel, S., Avelino-de-Souza, K., Neves, K., Porfírio, J., Messeder, D., Mattos Feijó, L., ... & Manger, P. R. (2014). The elephant brain in numbers. *Frontiers in Neuroanatomy*, 8, 46. https://doi.org/10.3389/fnana.2014.00046

Hick, W. E. (1952). On the rate of gain of information. *Quarterly Journal of Experimental Psychology*, 4(1), 11–26. https://doi.org/10.1080/17470215208416600

Hockett, C. F. (1960). The origin of speech. *Scientific American*, 203(3), 88–97. https://doi.org/10.1038/scientificamerican0960-88

Hoffmann, D. L., Standish, C. D., García-Diez, M., Pettitt, P. B., Milton, J. A., Zilhão, J., ... & Pike, A. W. (2018). U-Th dating of carbonate crusts reveals Neandertal origin of Iberian cave art. *Science*, 359(6378), 912–915. https://doi.org/10.1126/science.aap7778

Hubel, D. H., & Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *Journal of Physiology*, 195(1), 215–243. https://doi.org/10.1113/jphysiol.1968.sp008455

Huttenlocher, P. R., & Dabholkar, A. S. (1997). Regional differences in synaptogenesis in human cerebral cortex. *Journal of Comparative Neurology*, 387(2), 167–178. https://doi.org/10.1002/(SICI)1096-9861(19971020)387:2<167::AID-CNE1>3.0.CO;2-Z

Imhausen, A. (2016). *Mathematics in ancient Egypt: A contextual history*. Princeton University Press.

Jones, C. R., & Bergen, B. K. (2026). Large language models pass a standard three-party Turing test. *Proceedings of the National Academy of Sciences*, 123(21), e2524472123. https://doi.org/10.1073/pnas.2524472123

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux.

Kersten, D. (1987). Predictability and redundancy of natural images. *Journal of the Optical Society of America A*, 4(12), 2395–2400. https://doi.org/10.1364/JOSAA.4.002395

Knuth, D. E. (1997). *The art of computer programming, volume 1: Fundamental algorithms* (3rd ed.). Addison-Wesley.

Koch, K., McLean, J., Segev, R., Freed, M. A., Berry, M. J., Balasubramanian, V., & Sterling, P. (2006). How Much the Eye Tells the Brain. *Current Biology*, 16(14), 1428–1434.

Lai, C. S., Fisher, S. E., Hurst, J. A., Vargha-Khadem, F., & Monaco, A. P. (2001). A forkhead-domain gene is mutated in a severe speech and language disorder. *Nature*, 413(6855), 519–523. https://doi.org/10.1038/35097076

Landauer, R. (1961). Irreversibility and heat generation in the computing process. *IBM Journal of Research and Development*, 5(3), 183–191. https://doi.org/10.1147/rd.53.0183

Laughlin, S. B. (2001). Energy as a constraint on the coding and processing of sensory information. *Current Opinion in Neurobiology*, 11(4), 475–480. https://doi.org/10.1016/S0959-4388(00)00237-3

Lent, R., Azevedo, F. A., Andrade-Moraes, C. H., & Pinto, A. V. (2012). How many neurons do you have? Some dogmas of quantitative neurology revisited. *Journal of Comparative Neurology*, 520(1), 66–77. https://doi.org/10.1002/cne.22671

Luria, A. R. (1976). *Cognitive development: Its cultural and social foundations*. Harvard University Press.

Malafouris, L. (2013). *How things shape the mind: A theory of material engagement*. MIT Press.

Manger, P. R. (2006). An examination of cetacean brain structure with a novel hypothesis correlating thermogenesis to the evolution of a big brain. *Biological Reviews*, 81(2), 293–338. https://doi.org/10.1017/S1464793106007019

Marcus, G. F., Pinker, S., Ullman, M., Hollander, M., Rosen, T. J., & Xu, F. (1992). Overregularization in language acquisition. *Monographs of the Society for Research in Child Development*, 57(4), 1–182. https://doi.org/10.2307/1166115

Montemurro, M. A., & Zanette, D. H. (2011). Universal entropy of word ordering across linguistic families. *PLoS ONE*, 6(5), e19875. https://doi.org/10.1371/journal.pone.0019875

Montemurro, M. A., & Zanette, D. H. (2013). Towards the quantification of the semantic information encoded in written language. *Advances in Complex Systems*, 16(02n03), 1350003. https://doi.org/10.1142/S0219525913500031

Netz, R. (1999). *The shaping of deduction in Greek mathematics: A study in cognitive history*. Cambridge University Press.

OpenAI (2024). *Learning to Reason with LLMs*. OpenAI Technical Report. https://openai.com/index/learning-to-reason-with-llms/

Piantadosi, S. T., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, 108(9), 3526–3529. https://doi.org/10.1073/pnas.1012551108

Prokopenko, M., Boschetti, F., & Ryan, A. J. (2009). An information-theoretic primer on complexity, self-organization, and emergence. *Complexity*, 15(1), 11–28. https://doi.org/10.1002/cplx.20249

Saxe, G. N., Calderone, D., & Morales, L. J. (2018). Brain entropy and human intelligence: A resting-state fMRI study. *PLoS ONE*, 13(2), e0191582. https://doi.org/10.1371/journal.pone.0191582

Schmandt-Besserat, D. (1992). *Before writing: From counting to cuneiform*. University of Texas Press.

Schoenemann, P. T. (2013). Hominin brains and language evolution. *Language and Cognition*, 5(2–3), 101–113. https://doi.org/10.1515/langcog-2013-0007

Schrödinger, E. (1944). *What is life? The physical aspect of the living cell*. Cambridge University Press.

Seyfarth, R. M., Cheney, D. L., & Marler, P. (1980). Monkey responses to three different alarm calls: Evidence of predator classification and semantic communication. *Science*, 210(4471), 801–803. https://doi.org/10.1126/science.7433999

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Shannon, C. E. (1951). Prediction and entropy of printed English. *Bell System Technical Journal*, 30(1), 50–64. https://doi.org/10.1002/j.1538-7305.1951.tb01366.x

Spearman, C. (1904). "General intelligence," objectively determined and measured. *American Journal of Psychology*, 15(2), 201–292. https://doi.org/10.2307/1412107

Thompson, P. M., Cannon, T. D., Narr, K. L., van Erp, T., Poutanen, V. P., Huttunen, M., ... & Toga, A. W. (2001). Genetic influences on brain structure. *Nature Neuroscience*, 4(12), 1253–1258. https://doi.org/10.1038/nn758

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Harvard University Press.

Tomasello, M. (2014). *A natural history of human thinking*. Harvard University Press.

Vaswani, A., Shazeer, N., Parmar, N., Uszoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

von Bartheld, C. S., Bahney, J., & Herculano-Houzel, S. (2016). The search for true numbers of neurons and glial cells in the human brain: A review of 150 years of cell counting. *Journal of Comparative Neurology*, 524(18), 3865–3895. https://doi.org/10.1002/cne.24040

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., ... & Fedus, W. (2022). Emergent abilities of large language models. *Transactions on Machine Learning Research*. https://openreview.net/forum?id=yzkSU5zdwD

Wittgenstein, L. (1953). *Philosophical investigations* (G. E. M. Anscombe, Trans.). Blackwell Publishing.

Wolf, M. (2007). *Proust and the squid: The story and science of the reading brain*. HarperCollins.