Askitas, Nikos

Working Paper

Notes on a World with Generative AI

IZA Discussion Papers, No. 18070

Provided in Cooperation with:
IZA – Institute of Labor Economics

Suggested Citation: Askitas, Nikos (2025) : Notes on a World with Generative AI, IZA Discussion Papers, No. 18070, Institute of Labor Economics (IZA), Bonn

This Version is available at:
https://hdl.handle.net/10419/328200

I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 18070

Notes on a World with Generative AI

Nikos Askitas

AUGUST 2025

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 18070

Notes on a World with Generative AI

Nikos Askitas
IZA

AUGUST 2025

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Notes on a World with Generative AI*

Generative AI (GenAI) and Large Language Models (LLMs) are advancing into domains once seen as uniquely human: reasoning, synthesis, abstraction, and rhetoric. This paper, addressed to (labor) economists, subject-matter experts, and informed readers, aims to clarify what is genuinely new about LLMs, what is not, and why the distinction matters. Using the analogy to autoregressive models familiar from economics, we build a conceptual bridge to explain their stochastic nature, a feature which produces fluency that is often mistaken for agency even by experienced observers. We then place LLMs in the broader history of human–machine outsourcing, from digestion to cognition, to frame their emergence as part of a much longer trajectory of technological delegation. The analysis examines the disruptive implications for white-collar labor markets, institutional structures, and the epistemic norms that shape knowledge production. Particular attention is given to the risks and paradoxes that arise when synthetic content becomes both the product and the raw material of cognitive work. In such cases, displacement of human labor can erode the very source of original material on which these systems depend, creating a feedback loop that degrades both input quality and output reliability. By grounding the discussion in conceptual clarity rather than speculative forecasts or media-driven panic, we seek to create space for a more deliberate and actionable understanding. While GenAI can substitute for some of the labor it draws upon, its statistical limits will probably preserve an essential role for human judgment. The central question is not only how these tools are deployed, but also which activities we relinquish, and how we choose to reallocate our attention and expertise in a reshaped division of cognitive labor.

JEL Classification: J24, O33, O31, J22, D83, L86, J44, O38

Keywords: Generative Artificial Intelligence, Large Language Models, autoregressive models, labor economics, technological change, automation and outsourcing, human-machine collaboration knowledge work, epistemic norms, digital transformation

Corresponding author:
Nikos Askitas
Institute of Labor Economics (IZA)
Schaumburg-Lippe-Strasse 5-9
53113 Bonn
Germany
E-mail: nikos@askitas.com

* These reflections were developed through an interactive, dialectical exchange between the author's original ideas and a large language model. Final responsibility remains with the author who procrastinated writing them, troubled by one of the very questions the paper explores: if no one reads the text but only consumes its GPT-generated summary, why write it at all? Readers who both read and GPT-summarise the paper and wish to share reflections on that experience are invited to send a brief note to nikos@askitas.com. Insightful observations may be included in an appendix to the paper, with attribution. All citations are placed in footnotes at the bottom of their page so as to optimise the reading experience. Versions of these thoughts were presented in several IZA research retreats, at LISER and elsewhere.

1 Introduction

AI and in particular GenAI has become the defining focus of our time. Across media, policy, academia, and popular culture, it is framed both as an existential risk and as a limitless opportunity. Some see it as the ultimate productivity tool, others as an angel of mass unemployment or epistemic collapse. Within the same breath, large language models (LLMs) are described as both astonishingly intelligent and deeply flawed; as breakthrough inventions and glorified autocomplete engines. This dissonance is not accidental. It reflects genuine confusion not just about the capabilities of the technology, but about how to think about it at all.

Figure 1: Search Interest for ChatGPT vs Soccer (a strong topic in search) worldwide.

Much of the public and academic debate suffers from two mirrored errors: overestimating what LLMs are, and either underestimating or misjudging what they change. On one hand, we are easily impressed by coherence, mistaking fluency for insight. On the other, we are often blind to structural effects: how the widespread use of such tools reshapes institutions, habits, and expectations, even as their internal workings are bluntly statistical.

This confusion is amplified by a growing divide in the economic discourse. Some, like Daron Acemoglu, paint a bleak picture of substitution, warning that LLMs accelerate an "AI illusion": a cycle of hype built on superficial competence, followed by structural disempowerment and mass unemployment. AI development, so the argument goes, is benchmarked against the Turing test, implicitly casting the human as the thing to be replaced.[1] Others,[2] drawing on principles of comparative advantage and opportunity cost, argue that displacement is not destiny: even as machines take over certain cognitive tasks, new roles and complementarities for humans will emerge, though perhaps in unfamiliar places. The tension between substitution and augmentation, replacement and reinvention, underlies much of the current unease.

Recent usage data reflects the scale and ambivalence on hand. According to Microsoft's 2024 Work Trend Index, 75% of global knowledge workers now use LLMs at work. Among them, 78% bring "their own AI" into the workplace, often without formal institutional support or oversight. Yet this widespread adoption is accompanied by quiet anxiety: 52% of users are reluctant to admit using LLMs for their "most important tasks," and 53% worry that doing so makes them look replaceable.[3] The covert and strategic use of LLMs[4] at scale introduces novel information asymmetries in the labor market but also creates challenges well beyond internal

[1] Acemoglu and Johnson (2023)

[2] E.g. https://www.noahpinion.blog/p/what-if-everyone-is-wrong-about-what

[3] See Microsoft and LinkedIn, *2024 Work Trend Index Annual Report*, p. 15. Available at: Microsoft 2024 Work Trend Index Report.

[4] In recent work we detected statistically significant residue of ChatGPT use in scientific abstracts where a similar covert use appears to be the case (Askitas, 2025)

workplace dynamics: it opens the door to training data poisoning, whether for geopolitical manipulation, disinformation, or influence operations, as well as to the stealthy alteration of weights in ostensibly "open-weight" models. In the current epistemic infrastructure, trust, transparency, and verification will become even more contested and fragile.

Some critics, including François Chollet, creator of Keras, have warned that the overwhelming focus on LLM development may be the worst thing that could happen to broader Artificial General Intelligence (AGI) research, since it drains funding and attention away from alternative approaches such as symbolic reasoning, program synthesis, and neurosymbolic systems.[5] Their concern highlights how technological hype can distort research priorities, reinforcing the importance of understanding both the promise and the limits of present-day systems.

This essay tries to reframe the conversation. We argue that LLMs are not synthetic minds, but simply highly scalable statistical mirrors. They excel at imitation, recombination, and surface-level synthesis. Yet their impact goes well beyond their mechanical limitations. By making cognitive labour easier, faster, and more accessible, they trigger a familiar pattern: the outsourcing of human faculties to machines. As with earlier technologies, from agriculture to writing, from engines to calculators, this outsourcing does not merely increase efficiency. It transforms what it means to learn, to signal competence, to become an expert, or even to think.

To explore this transformation, we proceed in four steps. First, in Section 2, we demystify the underlying logic of LLMs by showing that they are, quite literally, autoregressive models, albeit operating at an immense scale and applied to language rather than numeric time series: powerful, predictive, but blind to novelty. Second, in Section 3, we place them within a longer trajectory of technological outsourcing, from digestion to muscle to cognition, tracing both the upsides and civilisational side-effects. Third, in Section 4, we examine how these tools disrupt existing institutions and cause the emergence of new ones, amplify inequality, and might break the developmental ladders that sustain human expertise. Finally, in Section 5, we reflect on the paradox that while LLMs threaten to displace human work, they still depend on it for their own survival.

In the Addendum 1, we offer the reader who is not faint of mathematical symbolism a heuristic sketch of why it is difficult to produce a safe LLM in the sense of protecting the naive or vulnerable from erroneous, yet convincing, outputs. LLMs can be thought of as compressing engines of their training set, in the sense that they compress the entire training corpus into a set of model parameters, and we use this framing to sketch their limitations and unmitigated risks.

In the Addendum 2 we discuss the difference between learning language structure (as it turns out LLMs can do that consistently) and learning truth (a harder more fragile undertaking). While the ideas in the addendum can be made mathematically precise we confine ourselves to plain english at this point.

This discussion unfolds around eight central threads: the autoregressive structure of LLMs; their role in outsourcing cognition; the tension between fluency and understanding; the epistemic risks of synthetic text; the breakage of skill ladders; the economic tradeoffs between augmentation and substitution; the long-term implications for expertise and epistemic infrastructure; and the paradox of recursive dependence on human-generated input. Each of these themes informs how we think about both the promise and the limitations of generative AI.

This essay was written to give structure to the author's thoughts on this emergent technology. We hope it may offer guidance or inspiration for new research questions that bring together researchers from otherwise disconnected disciplines, and help policymakers, the press, and the broader public reflect on this emerging disruption in a fact-based, dispassionate, and nuanced

[5] See e.g. https://www.freethink.com/robots-ai/arc-prize-agi

manner, incorporating areas for which the author has a particular fondness: technological, mathematical, historical, philosophical, and social-scientific perspectives.

2 GenAI: An Autoregressive Parrot

At its core, a large language model (LLM) is a sophisticated autoregressive machine. Like a high-order autoregressive (AR) time series model (admittedly among the simplest of all econometric models) that predicts the next data point based on patterns in the past, an LLM predicts the next word (or token) based on a statistical understanding of previous ones. Its genius lies not in understanding meaning *per se*, but in mastering correlation: recognizing and reproducing the structure of language at scale[6].

This surface-level coherence, however, often mystifies even experts. Because the output feels fluid, purposeful, and well-formed, it triggers our instinct to attribute intention, agency, or understanding despite none being present. Even experienced engineers, like the Google researcher[7] who publicly claimed an LLM had become sentient, have fallen into this trap. But the effect is not new. ELIZA, the simple chatbot developed in the 1960s by Joseph Weizenbaum, famously elicited strong emotional reactions from users by doing nothing more than reflecting their input back at them in a structured way. This tendency of humans to interpret fluency as thought came to be known as the "ELIZA effect".[8]

This similarity to autoregressive models is not merely metaphorical; it is both technical and conceptual.[9] What time-series are to AR models, books are to LLMs: long sequences of data from which future elements are statistically inferred. Just as autoregressive (AR) models treat a time series as a sequence of numerical observations, where finite sequencies of past values are used to predict the next one, LLMs treat language as a sequence of tokens (or characters). During training, the model slides through text in overlapping windows (say, 30 tokens at a time) learning to predict what comes next after each such sequence. The objective is to minimise prediction error across this entire textual universe, much like fitting a regression line through historical data. And as with AR models, performance is often excellent until a structural break occurs. Ironically, that is precisely when prediction matters most.

The underlying assumption behind autoregressive models is deceptively simple: if the past determines the future, then some essential structure must be encoded in past values themselves. The goal is not to recover causality, but to exploit correlation on the belief that history carries the signature of what's to come. By this logic, past values have something meaningful to say about future ones, and their statistical regularities are a valid basis for extrapolation.

LLMs operate on the same principle. When we prompt them, they respond by conditioning on the latent priors embedded in their training corpus. These priors are drawn from patterns in existing text (books, code, dialogue, and documents) and the output is a kind of statistically probable continuation, not an act of understanding. What's missing is what was never encoded to begin with.

Consider a time series forecasting model trying to predict the stock price of Volkswagen (VW) during the period it was secretly using emissions defeat devices[10]. No matter how so-

[6] See Bender et al. (2021) for the influential critique that introduced the metaphor of LLMs as "stochastic parrots" i.e. models that fluently mimic language without true understanding, raising concerns about scale, bias, and misuse.

[7] See *The Guardian*, "Google fires software engineer who claimed AI chatbot was sentient," June 12, 2022. Available at: https://www.theguardian.com.

[8] See ELIZA and the "Eliza effect" described on Wikipedia (https://en.wikipedia.org/wiki/ELIZA, https://en.wikipedia.org/wiki/ELIZA_effect); also Joseph Weizenbaum's original paper (Weizenbaum, 1966).

[9] Not surprisingly, GenAI models are inherently well-suited to time series forecasting (Lim and Zohren, 2021).

[10] In September 2015, Volkswagen admitted to installing defeat devices in diesel vehicles to manipulate emissions tests. Following the disclosure, VW's stock plunged nearly 30% in a matter of days. See Jack Ewing, "Volkswagen Says 11 Million Cars Worldwide Are Affected in Diesel Deception," *The New York Times*, Sept. 22, 2015. https://www.nytimes.com

phisticated the model, it would have failed to foresee the crash that followed disclosure because the relevant causal factor was hidden, absent from the data. The signal wasn't in the series.

In the same manner, a truly novel idea, one that incubates in a mind before ever being articulated, cannot be predicted by an LLM. Because it does not exist in the corpus, it leaves no statistical trace. Language models can recombine what they have seen, but they cannot anticipate what has never been expressed although in many occasions they might lead us to believe so.

LLMs rely on this same statistical logic when adapting to different tones or genres. You can ask an LLM to write lyrics like Tom Waits, prose like an *American Economic Review* paper, or a news report in the style of *The New York Times*. It will comply, not because it understands these styles, but because it conditions its internal probabilities on the statistical signature of the corresponding corpus. It's not imitating ideas; it's matching form. And with that, the LLM quietly enters the labor market, performing tasks once reserved for journalists, analysts, editors, and assistants.

An AR model performs well when the future looks like the past. But it falters when there's a structural break: a crisis, a regime change, or a fundamentally new situation outside its training window. LLMs behave in much the same way. They (most of the time) produce fluent, plausible, even persuasive text, until you ask them to handle the unfamiliar, the contradictory, or the truly novel. When the moment requires creative insight or epistemic risk-taking, the model retreats to the comfort of prior patterns.

Some counter the "parrot" critique by pointing out, correctly, that LLMs often generate sentences they've never seen before. But this does not contradict the autoregressive analogy. An AR model, once fitted, can produce values it has never observed, yet those values still lie close to the regression line. They are new, but not surprising. Likewise, LLMs can generate entirely novel phrases or formulations that are statistically consistent with their training data. They don't repeat the past; they extend its trajectory. Occasionally, an LLM may stumble upon a formulation that *feels* like an idea. But this is no more mysterious than an AR model producing an unfamiliar data point: it is novel, yet structurally predictable. What makes the output seem insightful is less the presence of intention than our tendency to confuse fluency with agency, especially when a sequence happens, by chance or correlation, to capture something meaningful. In the opposite direction, when you feel frustrated at how an LLM will speak fluent nonsense to you, changing its "mind" as you interact with it, know that this is like an AR model being off in its prediction. Sometimes our confidence interval excludes nonsense, and sometimes it contains some. And once again, interacting with an LLM is not conversing with a mind, it is sampling from a forecast, sometimes within bounds, sometimes well outside them[11].

This very mechanism (the generation of new-but-expected output) is what gives LLMs their distinctive power. They do not retrieve information the way a classical database or search engine does; instead, they *predict* what a plausible answer would be, given all they've seen. They won't necessarily cite a passage from a specific book (although LLM extensions now exist that can), but they can generate something that reads like a citation, an AR-style projection of what a passage *should* say, based on the latent structure of the corpus. In this sense, an LLM functions as a composite of a know-it-all author of a giant, all-encompassing dynamically written book , and a locate-all librarian. Here is another instance where LLMs begin to encroach on traditionally human white-collar tasks: rephrasing, organising and retrieving "information". In doing so, the LLM enters, once again, the labor market, raising the question of whether it augments or substitutes human cognitive work.

GenAI is not a thinker. But it is a tireless *synthesizer*, a tool that makes the implicit ex-

[11] For the reader who needs to dig deeper: OpenAI (2023), Schaeffer et al. (2023), Liu et al. (2020)

plicit, surfaces hidden patterns and latent phrasing, and assembles fragments of knowledge into (mostly) coherent text, even when that text did not exist before. In doing so, grammar and syntax may be largely correct (form), but validity and logical coherence are often lacking.

This is why LLMs excel at *style* but not at *substance*. They can imitate clarity, compress ambiguity, and polish expression. But they do not "know" in any meaningful sense, nor do they invent. Instead, they remix and repackage. The resulting prose often has the *feel* of authority (precise, coherent, and grammatically polished) but lacks the epistemic grounding that comes from actual reasoning or evidence. Like parrots trained on vast archives, LLMs can mimic almost any intellectual posture without understanding the stakes of what is being said.

Still, this imitation should not be dismissed. What emerges is a new kind of cognitive infrastructure. LLMs collapse the cost of semantic retrieval, making it easier to trace a concept, rephrase a claim, or simulate an argument without having to originate it. They open up a new way of accessing and rearranging knowledge, one that *augments* rather than replaces human agency.

To recap the analogy: in econometrics, we often prize causal models over predictive ones, because the former aim to explain mechanisms, not just correlations. Yet in many real-world contexts where causal inference is difficult or unnecessary, prediction is good enough. Generative AI falls squarely into this prediction class.

Our own reasoning, at its best, aspires to causality, to uncover hidden mechanisms, question assumptions, and shift paradigms. Generative AI, by contrast, excels at optimising within known frames: refining an idea, drafting an introductory paragraph, summarising a document. These are tasks where high-quality prediction is sufficient and often more efficient than manual effort.

The analogy isn't perfect, but it's instructive: in a world full of prediction-class problems, generative AI is an effective tool. But for questions that demand causal leaps, conceptual breakthroughs, or principled dissent from established narratives, it remains just that: a tool. In short: Generative AI continue to be more A than I.

Thinking is like uncorking a bottle of wine: there's a pop, a shift in pressure, and suddenly something begins to flow. Sometimes it's rich and structured, sometimes wild and surprising, sometimes thin and forgettable but it's always yours, and it always emerges from the strange, nonlinear processes that make up a mind. With LLMs, there is also unpredictability: you never quite know what you're going to get. This is partly due to their stochastic nature, modulated by the so-called temperature parameter that controls randomness. But crucially, what flows from them is not the result of interior pressure or insight. It is the output of an autoregressive process: a vast engine of statistical pattern continuation, estimating the next token based on everything that came before. The resemblance to thought is superficial. It looks like wine, it pours like wine, but it was never fermented in a mind, at least not directly. It is distilled from a collective corpus, inferred from the archived outputs of uncountable minds.

3 Outsourcing Ourselves

Technology (a predominantly, though not exclusively[12], human endeavor) can be understood as the systematic outsourcing of human faculties to machines[13]. From digestion to muscle to cognition, we have continuously externalised internal processes, delegating what was once organic and embedding it into artefacts of our own design.

Early agriculture and cooking technology transformed human digestion. By preprocessing food outside the body (soaking, fermenting, grinding, heating) we weakened our teeth and shortened our digestive tract over evolutionary time and freed ourselves from the constant labour of feeding[14]. This outsourcing did not merely save energy; it reorganised time, allowing attention to be redirected to tool-making, social interaction, or abstract thought. But it also introduced new forms of vulnerability: our modern world faces civilisational illnesses such as obesity, malnutrition, anorexia, and other nutritional or metabolic disorders, conditions born not of scarcity, but of abundance, misalignment, and over-optimisation.

The industrial revolution brought the mechanisation of muscle. Horses[15] and human labourers were replaced by steam, steel, and internal combustion. Machines substituted brute strength, yes, but also augmented it: workers operating powered tools could accomplish far more than unaided effort ever allowed. Yet this came at a cost. As the body became redundant, it atrophied. The rise of sedentary labour, and eventually fitness clubs[16], is not a contradiction but a consequence, a new layer of adaptation in response to an older one. The externalities of this transformation included a profound reliance on fossil fuels, widespread pollution, and an exacerbation of climate change.

Automation took the logic of mechanisation further. From the assembly line to the back office, processes once requiring human judgement were codified into scripts, routines, and feedback loops. Taylorism[17] rationalised labour by breaking it into discrete, optimisable units, but it also depersonalised work, alienated workers, and reduced them to appendages of a system. This form of outsourcing brought immense productivity gains, but it also intensified inequalities and generated resistance movements, from trade unions to calls for universal basic income in our days: once again successive layers of adaptation in response to prior ones.

The same pattern is visible in the rise of computing and navigation technologies. The outsourcing of arithmetic to calculators, and of memory and logic to computers, has made basic computational skills increasingly rare. Imagine, for example, the early human calculators in the NASA programme, highly trained specialists, who were eventually displaced by IBM machines[18]. Yet even as machines took over, the human role remained critical in design, oversight,

[12] Tool use is not unique to humans: chimpanzees, capuchin monkeys, New Caledonian crows, dolphins, and even octopuses have been observed selecting, modifying, and using tools in the wild (Seed and Byrne, 2010).

[13] For the origins of this somewhat econ-centric formulation, see Anders (1956), where he introduces the concept of the "Promethean Gap" (the reader with classical Greek education will certainly appreciate the name of the concept). See also Wiener (1950), for an early articulation of cybernetic automation as the delegation of human agency to machines.

[14] https://www.americanscientist.org/article/meat-eating-among-the-earliest-humans

[15] By the mid-20th century, the United States horse population had fallen approximately 80% from its 1920 peak of 27.5 million to around 4.5 million in 1959, a decline closely tied to the widespread adoption of tractors, automobiles, and other motorised transport. (United States Department of Agriculture, 1959).

[16] As the body atrophied from disuse, sedentary lifestyles became widespread so did fitness clubs emerge to compensate. The modern fitness-club industry began taking off in the United States during the 1960s and 1970s; by the mid-1990s, private fitness centers had become ubiquitous, driven by cultural shifts and alarm over declining physical condition in a post-industrial society (Stern, 2008; Petrzela, 2020)

[17] (Braverman, 1974; Encyclopaedia Britannica, 2025)

[18] By the late 1950s, NASA had begun replacing human "computers" with IBM mainframes like the IBM 704

and judgement. Similarly, GPS has allowed us to navigate confidently through unfamiliar terrain, but studies suggest it may degrade our internal spatial awareness and sense of direction[19]. Meanwhile, thousands of orbiting satellites light up the night sky, raising concerns not only about aesthetic and ecological disruption but also long-term space debris. These tools increase our reach but narrow our involvement. What we gain in convenience, we may lose in intuition.

In each case, outsourcing a function to technology did not simply replace the old with the new. It reshaped the environment, redefined effort, and reconfigured what it meant to be human in that context. The effect was not linear but dialectical: substitution in one domain produced augmentation in another, which in turn demanded new compensations, norms, and forms of self-discipline.

Every time we outsource a capability, those fellow human beings who happened to be endowed with that very skill experience a loss in its market value. Yet paradoxically, they are often best positioned to adapt: to step into the new gaps and demands created by the shifting landscape. It is rarely a smooth or painless transition. But it is one we have undergone repeatedly, and will undoubtedly face again.

Language models are the next phase in this trajectory. They represent the outsourcing of synthesis, summarisation, composition, and rhetorical finesse. LLMs are not the first machines to shape how we think: writing itself was once a disruptive technology (and so was its mass upscaling with Gutenberg centuries later). In his dialogue *Phaedrus*[20] in 370 BC, Plato expressed concern that writing would erode memory and true understanding[21]. Once written down, knowledge would be read without being known; people would seem wise without being so[22]. The warning, written over 23 centuries ago, is strangely contemporary. It echoes today's fears that language models will erode our reasoning or replace thought with mere simulation[23].

Plato's concern in *Phaedrus* that writing would weaken memory and degrade true understanding, was not entirely misplaced, though it underestimated the brain's remarkable plasticity. As modern neuroscience reveals, the adoption of writing did not merely offload memory; it *restructured the brain itself*. According to Dehaene's theory of *neuronal recycling*, literacy co-opted parts of the visual cortex originally evolved for spatial orientation and object recognition, particularly in the left occipitotemporal region, repurposing them to process letter-

and 7090, capable of performing in minutes what previously took human teams days or weeks. For details, see NASA's history of human computers and IBM automation at NASA (NASA Jet Propulsion Laboratory / NASA History, 2016; Miracle, 2025).

[19] See Maguire et al. (2006) for how the brain differs between bus drivers (light navigational load) and taxi drivers (heavy load) in London.

[20] Plato (1997), https://www.historyofinformation.com/detail.php?id=3439

[21] Plato, Phaedrus 275a–278e. Socrates recounts the myth of Theuth and Thamus, where the invention of writing is said to produce "forgetfulness in the souls of those who learn it." This anticipated what is now called the Google effect: if you can always find, why remember?

[22] Ironically, this very condition has become institutionalized in modern education. One of the more unfortunate perversions of contemporary schooling is that it often rewards rote memorization over creative thinking, precisely the sort of shallow recall Plato warned against. Rather than serving as a tool to scaffold understanding, externalized knowledge becomes a checklist of facts to be recited. If writing once threatened memory, today's education system sometimes mistakes memory for knowledge. In that light, large language models are not a rupture, but a continuation: fluent in recall, indifferent to meaning, and trained (like many students) to predict the right answer.

[23] The author first encountered Plato's critique of writing in *Phaedrus* during an interactive exchange with a large language model. When prompted for historical concerns about externalised memory, the model surfaced this example, not through database retrieval or lookup, but via autoregressive prediction. It generated the reference by statistically completing a sequence of contextual prompts based on patterns learned during training, effectively reconstructing a citation lurking in the vastness of its training set, rather than recalling it from a fixed source. As always, the author diligently verified the reference, but the learning effect was genuine.

forms and words.[24] This act of outsourcing cognition to written symbols was not passive but *physiologically active*: it altered the cognitive foundation that once feared such alteration. Today's large language models present an analogous challenge: we outsource linguistic fluency, ideation, and structure to machines. And again, the brain will not remain unchanged. If the past is any guide, we should expect not just cultural but *neurological adaptations*, as humans co-evolve with these new symbolic collaborators.

And yet, as with Plato's critique, we risk underimagining the upside. Writing may have made the natural rote expert obsolete (the labor market once again), but it enabled the preservation and dissemination of knowledge across space and time, and the creation of entirely new intellectual roles: authors, librarians, archivists, editors, and publishers. What seemed like a loss to the oral mind became an infrastructure for modern science, law, and culture. The same may yet prove true of LLMs, if we learn how to wield them.

As with earlier tools, LLM adoption is bound to disrupt and upend existing structures: from the labour market to our epistemic norms to our very physiology. Just as mechanisation weakened the body but extended its reach, cognitive outsourcing may dull the mental muscle it displaces, even as it amplifies what remains. To produce more, we may end up thinking less or thinking in unfamiliar ways. LLMs, too, come with externalities: their training requires enormous amounts of data, energy, and computational infrastructure, raising sustainability and equity concerns[25].

The consequences will be mixed. The speed and scope of intellectual work may increase dramatically, with entire fields accelerated by models that generate hypotheses, summarise literature, write code, or refactor logic. But cognitive dependence may also grow. Displaced skills aren't guaranteed to be retained. We can easily imagine a future where most people consume LLM-processed content, while others pay a premium for human-curated work, eventually giving rise to "fitness clubs for the mind" to preserve critical thinking, memory, abstraction, and judgment. These may be necessary to counter new forms of intellectual malnutrition. The challenge, as always, is not just what machines can do, but what we stop doing once they can, and more importantly, what we choose to do instead.

The current wave of large language models (LLMs) did not arise in a vacuum. They emerged from a converging state of software and hardware that reflects and amplifies the logic of the human mind. On the software side, transformer architectures (first introduced in 2017) captured the statistical structure of human language using attention-based mechanisms remarkably well-suited to the way we process context and sequence. On the hardware side, the rise of highly parallelised computation via GPUs and TPUs enabled models with hundreds of billions of parameters to be trained on a scale matching human cultural output. But none of this would have worked if language itself (its syntax, structure, and predictability) were not already a product of the cognitive architecture that made us human. In a very real sense, LLMs work because they exploit patterns that our own brains evolved to produce.

Writing, with its use of angles (Gamma), triangles (Delta), circles (Omikron), and other geometric forms, likely co-evolved with the brain in complex and unpredictable ways. Dehaene's theory of neuronal recycling suggests that reading and writing did not evolve new neural structures but instead repurposed ancient spatial recognition circuits, particularly in the

[24] See Dehaene and Cohen (2007); Dehaene (2009).

[25] Bitcoin shows that raw hardware efficiency can improve by orders of magnitude over time. If LLM workloads follow a similar curve of hardware and software optimisation then per-output energy costs could fall sharply, potentially making LLMs much less energy-hungry in the future. The key question is whether demand rises faster than efficiency improves, leading to a Jevons paradox effect, where total energy use increases despite gains in efficiency.

visual cortex, for decoding symbolic shapes and mapping them to sound and meaning.[26]

The adventurous reader might allow this author, a non-expert on the subject matter, to spitball on what got us to large language models (LLMs), and what LLMs might in turn do to us, not culturally, but physiologically and neurologically.

To start, LLMs are not an alien artifact; they are deeply human. They emerged from a particular configuration of brain functions. Our evolved *language faculty*, regions like Broca's area[27] (involved in producing speech) and Wernicke's area[28] (involved in understanding it), gave rise to our taste for syntax, semantics, and structured expression. Supporting these are a network of connections between higher-order reasoning centers and sensory areas, enabling the integration of thought, language, and perception. Our brains are also powerful *prediction engines*, constantly anticipating the next word, movement, or social cue. Transformer-based LLMs like GPT are a formalisation and amplification of this very logic: they model the statistical structure of language by predicting the next word, token by token, just as we do in conversation.

Add to this our reliance on *compression and abstraction*, enabled by working memory and executive control, and our talent for *social cognition*: inferring intentions, simulating conversations, tracking narrative arcs. These are not merely background conditions; they shaped the design of LLMs. We built these systems to "speak" like us because we are shaped to speak the way we do. In that sense, LLMs are a mirror of the neural real estate that produced them.

But mirrors reflect back. And now, these very same circuits are being reshaped in return.

If writing reshaped the visual cortex, LLMs may reshape the *prefrontal cortex*, particularly areas responsible for planning, abstraction, and structured thought. Offloading ideation, summarisation, and even first-draft writing to a machine may reduce the need to hold extended arguments or narrative threads in working memory. Our *language production systems* may become more reactive and curatorial, focused on steering and refining rather than generating from scratch.

Our *reward systems*, sensitive to fast feedback and novelty, may be subtly tuned by the immediacy of machine-generated responses. Just as social media exploits intermittent reinforcement, the LLM-as-assistant could deepen cognitive dependence through sheer responsiveness and availability.

The most profound effects may be on our *social cognition circuits*. As we increasingly treat LLMs as conversation partners, addressing them in the second person, anthropomorphising their tone, and expecting empathy, we may begin to repurpose the same neural modules used to model human minds. This could deepen our empathy or flatten it, depending on whether these synthetic interactions displace real ones.

And then there is our *sense of self*. Just as mirrors or autobiographies relate to our self-awareness, LLMs may reshape *narrative identity*: how we externalise our thoughts and reflect on them. An external mind that finishes your sentence, suggests your phrasing, or articulates your idea with more elegance than you could muster: what does that do to internal monologue? Does the self become a prompt formatter?

This is not a cautionary tale. The changes may not be good or bad in any general sense. But if reading and writing co-evolved with the human brain through the repurposing of neural real estate, there is no reason to think LLMs will leave the brain untouched. The interface may be screen and keyboard, but the medium is cognitive architecture. And evolution has never stopped at the skull.

[26] See Dehaene and Cohen (2007); Dehaene (2009).

[27] Broca (1861)

[28] Wernicke (1874)

4 Disruption

If LLMs are tools for outsourcing ourselves, they are also forces of disruption. Their emergence does not simply offer new capabilities; it destabilises the institutional and economic frameworks that depend on human cognition, originality, and trust. In this sense, language models are not just technical artefacts, they are market participants, cultural actors, and epistemic disruptors.

One major fault line is information asymmetry. As it becomes harder to distinguish between human- and machine-generated content, the value of authenticity declines. If readers cannot reliably tell whether an abstract, cover letter, or policy memo was written by a person or a model, trust in quality (and the signalling function of effort) begins to erode. This dynamic risks creating a classic lemons market[29]: as high-quality human work is crowded out by indistinguishable low-cost output, incentives to invest in quality may collapse.

Higher-order beliefs compound the problem. If I believe all writers are using machines, why read their work rather than just summarise it with an LLM? If I believe no one reads but only summarizes, why write at all? If I believe everyone assumes I'm using GPT, why wouldn't I? As these beliefs take hold, texts may grow shorter, less effortful, more interchangeable. The long-term consequences for attention, motivation, and meaning are still unknown[30].

At the same time, LLMs magnify existing superstar dynamics[31]. Those already adept at prompt engineering, framing, and narrative assembly can now scale their productivity dramatically, while others risk falling further behind. The gap between the top 0.1% and the rest may widen, not only in income, but in visibility, influence, and perceived competence. We may end up with a handful of Taylor Swifts, and the rest playing small bars for free drinks. Like earlier waves of digitization, AI favours scale and disproportionately rewards those already operating at the frontier.

AI readiness spans a wide spectrum: from the mentally or intellectually unprepared, or even vulnerable[32], to those unable to formulate an effective prompt, to those unable to evaluate a GPT output, up to those best positioned to exploit GenAI. A highly skilled programmer could thrive without LLMs, yet still reaps huge gains from them: drafting a technically correct solution might take days, while crafting a precise prompt could produce a viable draft in minutes, leaving only a quick audit. This is superstar dynamics on steroids. It is precisely those that need GenAI the least that will benefit the most[33].

Meanwhile, norms and institutions lag. Disclosure standards for AI-generated content remain weak. Academic and professional codes of conduct are only beginning to grapple with what constitutes legitimate use. Some are starting to respond: when submitting a paper to an Elsevier outlet, one now encounters a policy stating that LLMs may be used, but their use must be disclosed, and the authors remain fully responsible for the content (see Figure 2). Such language (tentative, permissive, but cautious) signals that the boundaries of acceptable practice are actively being redrawn. Education systems face a parallel dilemma: how to preserve the value of learning when the tools to bypass it are freely available, increasingly persuasive,

[29] Akerlof (1970)

[30] Nowhere is this information asymmetry more evident than in the ongoing battle between applicants and hiring firms where firms use AI to screen AI-generated resumes https://mashable.com/article/ai-generated-resumes-overwhelming-recruiters

[31] Rosen (1981). Thanks to Peter Kuhn for the idea.

[32] Dohnány et al. (2025)

[33] *For to every one who has will more be given, and he will have abundance; but from him who has not, even what he has will be taken away.* Matthew 25:29, RSV https://en.wikipedia.org/wiki/Matthew_effect

and often undetectable[34]. Some policy responses are already emerging: in China, for example, AI-enabled chatbot features such as photo recognition and real-time Q&A were temporarily disabled by major firms (Tencent, ByteDance, Alibaba, Moonshot) during the four-day *gaokao* college entrance exam to curb cheating and preserve fairness.[35]

Nowhere is this more visible than in the collapse of the traditional ascension ladder. Expertise is rarely born whole; it is cultivated through repeated low-stakes practice: writing bad drafts, summarising dense texts, failing to explain. But if LLMs automate these early-stage tasks, novices may be denied the friction that makes mastery possible. What happens when junior analysts, student writers, or early-career researchers no longer need to struggle through the tedious parts of thinking? The result may be a generation of professionals fluent in polished output but hollow in internal structure, lacking not knowledge, but the cognitive grind that gives it shape.

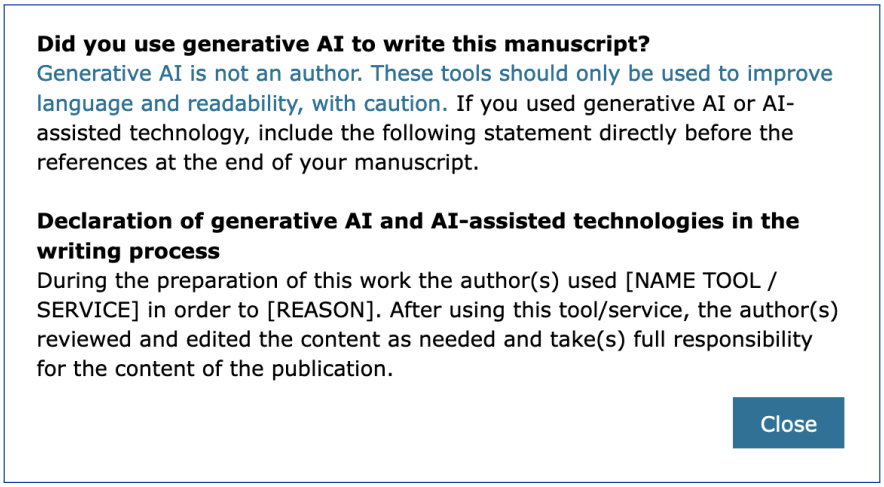

Figure 2: Pop-up on Elsevier's Editorial Manager: "GenAI is not a co-author. If you use it disclose it and make it your own"

The problem extends to staffing and career progression. If firms, following basic economic incentives, begin[36] substituting entry-level positions with agentic AI and retain only senior staff, how will future seniors emerge? The pipeline of expertise depends on apprenticeship and time. Remove the entry-level rung, and the ladder collapses. In education, a similar dynamic is unfolding. A recent *New York Times* feature reports growing controversy as professors increasingly rely on ChatGPT to generate teaching materials, while students protest that they pay to be instructed by humans, not algorithms they could freely access themselves.[37]

This tension between leveraging AI for instructional efficiency and maintaining educational legitimacy, further illustrates how asymmetric use of generative systems can erode trust and

[34]Many Gen Z-ers are delaying full-time work for extended schooling or gap experiences, seeking purpose over pay, and expressing uncertainty about workforce integration in an AI-disrupted economy. See factors driving this trend: Deloitte (2025) on Gen Z skill orientation and delay of milestones, Axios (2025) on school-to-work misalignment, and Post (2024) on rising NEET rates among youth.

[35]See *The Guardian*, "Chinese tech firms freeze AI tools in crackdown on exam cheats," June 9, 2025. Available at: https://www.theguardian.com

[36]See *Economic Times*, "After warning mid-level IT engineers that AI was going to do their work, now Mark Zuckerberg's Meta plans to lay off 5% of its workforce," March 2025. Available at: https://economictimes.indiatimes.com. *CNBC*, "Meta is targeting hundreds of millions of businesses for agentic AI," March 6, 2025. Available at: https://www.cnbc.com.

[37]See *The New York Times*, "College professors are using ChatGPT. Some students aren't happy," May 14, 2025. Available at: https://www.nytimes.com...

reshape institutional norms. From a policy perspective, this breakdown calls for creative regulatory responses.

If substitution becomes the norm, governments may consider a class of economic instruments similar to those used in environmental policy. Just as carbon taxes internalise the environmental costs of fossil fuel consumption, a "cognitive erosion tax" could be levied on firms that systematically replace junior staff with LLM-based agents. The logic is simple: if automating entry-level work depletes the future supply of human experts, undermining learning pathways, mentorship structures, and institutional memory, then the social cost should not be externalised. Conversely, firms that maintain human development pipelines or actively invest in hybrid intelligence where non-displaced junior staff work alongside LLMs, could receive targeted support or public recognition.

Another option, less punitive and more strategic, might involve public investment in "LLM stock": a dedicated labour force tasked with generating diverse, high-quality training data to keep generative systems robust. Whether via taxation, subsidy, or direct provision, the key is to acknowledge that knowledge ecosystems are not self-sustaining. Left to market incentives alone, they may cannibalise their own future. From a policy perspective, one response to the breakdown of training pathways and expertise pipelines might involve taxing substitution or subsidising retention. Just as carbon taxes aim to internalise negative externalities and steer market behaviour toward long-term sustainability, a "cognitive externality tax" could be levied on firms that aggressively replace junior roles with LLM agents. The idea is not to halt automation, but to recognise its systemic consequences: if the substitution of entry-level labour hollows out the future supply of experts, the costs of that erosion, currently borne by society, should be internalised by those driving it.

Conversely, firms that demonstrably maintain human apprenticeship structures, knowledge transfer, or hybrid human-AI work models might receive targeted support or public investment. In this framing, LLMs are treated not as neutral productivity tools, but as infrastructural forces whose uptake reshapes labour market dynamics, institutional memory, and long-run human capital development.

LLMs were trained on human output (text, code, conversation) in order to turn around and render many of those very humans obsolete. The model learned to write Python code, for example, by ingesting countless human-written snippets from sources like GitHub and Stack Overflow. The result is phenomena like "vibe coding," where users describe a task in natural language and the model writes the code. But as firms increasingly replace programmers with "agentic AI", the source of training data (active human coders) diminishes. Layoffs across the tech sector, in part driven by AI substitution, have already illustrated this shift.

Yet this disruption contains its own limit. If the asymmetry continues, the share of human-generated content in the future training corpus will approach zero. This leads to what researchers call model collapse: when AI systems are trained primarily on synthetic output, they begin to amplify their own statistical artefacts, compounding errors and degrading performance (see Figure 3). In this scenario, high-quality human content (or at least carefully curated synthetic data) regains value[38]. The paradox is that in order to keep LLMs sharp, we may need to employ human white-collar workers primarily as generators of future training data.

Of course, the disruption cuts both ways. These tools can democratise access to expert-

[38]In 2024, Stack Overflow and OpenAI announced a partnership allowing OpenAI access to Stack Overflow's vetted developer content to improve GPT-based coding tools. In return, Stack Overflow benefits from OpenAI-driven attribution and model tuning. This collaboration raises the possibility that human contributors, once volunteering for reputation, may increasingly serve as paid curators of high-quality training data. See: https://openai.com/blog/openai-and-stack-overflow-partnership.

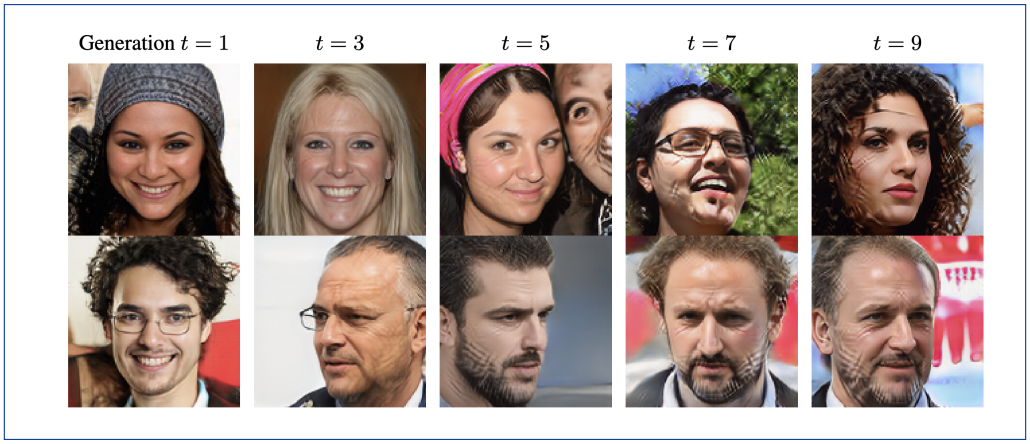

Figure 3: Self-Consuming Generative Models Go MAD - Model Autophagy Disorder (Alemohammad et al., 2023): Training generative artificial intelligence (AI) models on synthetic data progressively amplifies artifacts. Output in t is based on model trained on its output in round $t - 1$.

like output. They can help second-language writers, idiosyncratic thinkers, or under-resourced students level the playing field. They can compress learning curves and make institutional knowledge more searchable and responsive. But even this form of empowerment depends on careful adoption. Without thoughtful design, AI-enhanced systems may entrench dependence, reward surface fluency, and crowd out the very practices they aim to support.

What we are witnessing is not just a new tool, but a shift in the equilibrium of effort, signal, and trust. Like prior technological disruptions, the effects of LLMs will not be evenly distributed. They will benefit some, displace others, and reshape the norms that mediate both. The challenge is not simply how we use the technology, but how we reconfigure the institutions around it.

It is tempting to think of our digital knowledge ecosystem as uniquely fragile, prone to corrupted disks, server failures, and the obsolescence of formats. But knowledge has always been precarious.[39] Clay tablets eroded, papyrus crumbled, libraries burned,[40] and magnetic tapes degraded. What preserved knowledge across time was not the medium itself,[41] but the social act of transcription: monks copying manuscripts, scribes preserving commentaries, oral traditions renewing themselves through repetition. In this light, even the Internet Archive[42] or a large language model can be seen as part of that long lineage, mechanical monks, of a sort, preserving not individual texts but the statistical patterns that structure our collective memory. These systems do not remember as we do, but they are trained on what we chose to preserve. Fragility remains but so does the instinct to pass things on.

[39] See, for example, the Antikythera mechanism, a sophisticated mechanical computer dating from the 2nd century BC, discovered in a shipwreck and whose function remained mysterious for decades. Its survival was an accident; its context, mostly lost.

[40] The Library of Alexandria, though mythologized, remains a potent symbol of how centralized repositories of knowledge can vanish through fire, neglect, or political upheaval.

[41] Although the Gutenberg press contributed by enabling copies at scale.

[42] https://web.archive.org/

5 Conclusion

Large language models are not the end of human thinking, but they do alter its terrain. By mimicking form without understanding and producing output at scale, they force a revaluation of how we recognise effort, originality, and expertise. Trained on the residue of human labor, the written word, they are cognitive machines built from language, now set loose to reshape the very structures from which they were born.

Far from being synthetic minds, LLMs are best understood as (indeed, quite literally are) *autoregressive engines*: powerful in continuity, brittle in novelty. Like past technologies of outsourcing (digestion, muscle, navigation) they **liberate**, **displace**, and **reshape** in equal measure. They offer speed, reach, and fluency, but at the risk of cognitive atrophy, institutional erosion, and the loss of formative struggle. As with all transitions, the issue is not just what the machine can do but what we *stop* doing once it can.

The irony is deep. A machine trained on human expression begins to crowd out its source. Layoffs in tech, ghostwritten job applications, synthetic research, signal a structural inversion: the producer becomes the provider of raw material: the training data. Yet this process may be self-limiting. As synthetic content floods the ecosystem, model quality may degrade: a *statistical echo chamber* amplifying its own artefacts. Human thought, once devalued, may re-emerge as the one thing still worth predicting. The creators of original human content become farmers of biologically raised words, fuel for the thought-processing machine.

Plato feared that writing would erode memory. He was right in the narrow sense, but wrong in the long arc. Writing did not end thought, it extended it, archived it, transformed its audience put some people out of and others into work. Something similar may be possible now. But only if we resist the temptation to treat LLMs as minds, or replacements, or prophets and instead treat them as tools: powerful, fallible, and deeply shaped by the humans who train, prompt, and deploy them.

The real stakes lie not in the models themselves, but in the world we build around them. **Disclosure norms**, **learning systems**, **institutional checks**, **apprenticeship paths** will determine whether the future shaped by AI is one of collapse, stagnation, or the rise of a new cognitive infrastructure. **The machine predicts. The rest is still our business.**

References

Acemoglu, D. and Johnson, S. (2023). *Power and Progress: Our Thousand-Year Struggle Over Technology and Prosperity*. PublicAffairs, New York.

Akerlof, G. A. (1970). The Market for "Lemons": Quality Uncertainty and the Market Mechanism. *Quarterly Journal of Economics*, 84(3):488–500.

Alemohammad, S., Casco-Rodriguez, J., Luzi, L., Humayun, A. I., Babaei, H., LeJeune, D., Siahkoohi, A., and Baraniuk, R. G. (2023). Self-Consuming Generative Models Go MAD.

Anders, G. (1956). *Die Antiquiertheit des Menschen: Über die Seele im Zeitalter der zweiten industriellen Revolution*. C.H. Beck. English translation: *The Obsolescence of Human Beings*, 2022, Rowman & Littlefield.

Askitas, N. (2025). The behavioral signature of genai in scientific communication. Technical report, IZA Discussion Papers.

Axios (2025). Gen Z's broken school-to-work pipeline. https://www.axios.com/2025/07/15/gen-z-job-market-ai-unemployment.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*, pages 610–623. ACM.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., Donahue, C., Doumbouya, M., Durmus, E., Ermon, S., Etchemendy, J., Ethayarajh, K., Fei-Fei, L., Finn, C., Gale, T., Gillespie, L., Goel, K., Goodman, N., Grossman, S., Guha, N., Hashimoto, T., Henderson, P., Hewitt, J., Ho, D. E., Hong, J., Hsu, K., Huang, J., Icard, T., Jain, S., Jurafsky, D., Kalluri, P., Karamcheti, S., Keeling, G., Khani, F., Khattab, O., Koh, P. W., Krass, M., Krishna, R., Kuditipudi, R., Kumar, A., Ladhak, F., Lee, M., Lee, T., Leskovec, J., Levent, I., Li, X. L., Li, X., Ma, T., Malik, A., Manning, C. D., Mirchandani, S., Mitchell, E., Munyikwa, Z., Nair, S., Narayan, A., Narayanan, D., Newman, B., Nie, A., Niebles, J. C., Nilforoshan, H., Nyarko, J., Ogut, G., Orr, L., Papadimitriou, I., Park, J. S., Piech, C., Portelance, E., Potts, C., Raghunathan, A., Reich, R., Ren, H., Rong, F., Roohani, Y., Ruiz, C., Ryan, J., Ré, C., Sadigh, D., Sagawa, S., Santhanam, K., Shih, A., Srinivasan, K., Tamkin, A., Taori, R., Thomas, A. W., Tramèr, F., Wang, R. E., Wang, W., Wu, B., Wu, J., Wu, Y., Xie, S. M., Yasunaga, M., You, J., Zaharia, M., Zhang, M., Zhang, T., Zhang, X., Zhang, Y., Zheng, L., Zhou, K., and Liang, P. (2022). On the opportunities and risks of foundation models.

Braverman, H. (1974). *Labor and Monopoly Capital: The Degradation of Work in the Twentieth Century*. Monthly Review Press.

Broca, P. (1861). Loss of speech, chronic softening and partial destruction of the anterior left lobe of the brain. *Bulletin de la Société Anatomique de Paris*, 6:330–357.

Chelli, M., Descamps, J., Lavoué, V., Trojani, C., Azar, M., Deckert, M., Raynier, J.-L., Clowez, G., Boileau, P., Ruetsch-Chelli, C., et al. (2024). Hallucination rates and reference accuracy of chatgpt and bard for systematic reviews: comparative analysis. *Journal of medical Internet research*, 26(1):e53164.

Christiano, P., Leike, J., Brown, T. B., Martic, M., Legg, S., and Amodei, D. (2023). Deep reinforcement learning from human preferences.

Deloitte (2025). 2025 Gen Z and millennial survey. https://www.deloitte.com/global/en/issues/work/genz-millennial-survey.html.

Dohnány, S., Kurth-Nelson, Z., Spens, E., Luettgau, L., Reid, A., Gabriel, I., Summerfield, C., Shanahan, M., and Nour, M. M. (2025). Technological folie à deux: Feedback Loops Between AI Chatbots and Mental Illness.

Encyclopaedia Britannica (2025). Scientific Management (Taylorism). https://www.britannica.com/science/Taylorism.

Hernandez, D., Brown, T., Conerly, T., DasSarma, N., Drain, D., El-Showk, S., Elhage, N., Hatfield-Dodds, Z., Henighan, T., Hume, T., Johnston, S., Mann, B., Olah, C., Olsson, C.,

Amodei, D., Joseph, N., Kaplan, J., and McCandlish, S. (2022). Scaling laws and interpretability of learning from repeated data.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Madotto, A., and Fung, P. (2023). A Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, 56(12).

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. (2020). Scaling laws for neural language models.

Li, J., Li, G., Zhang, X., Dong, Y., and Jin, Z. (2024). EvoCodeBench: An Evolving Code Generation Benchmark Aligned with Real-World Code Repositories. *arXiv preprint arXiv:2404.00599*.

Lim, B. and Zohren, S. (2021). Time Series Forecasting with Deep Learning: A Survey. *Philosophical Transactions of the Royal Society A*, 379(2194):20200209.

Lin, S., Hilton, J., and Evans, O. (2022). TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Liu, Z., Cui, Y., Liu, S., Wang, T., and Hu, G. (2020). LogiQA: A Challenge Dataset for Machine Reading Comprehension with Logical Reasoning. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*, pages 1052–1062.

Maguire, E. A., Woollett, K., and Spiers, H. J. (2006). London taxi drivers and bus drivers: a structural MRI and neuropsychological analysis. *Hippocampus*, 16(12):1091–1101.

Miracle, A. (2025). Where Are All the Computers? *https://andrewmiracle.com/2025/02/26/where-are-all-the-computers/*.

NASA Jet Propulsion Laboratory / NASA History (2016). When computers were human: The early women "human computers" at nasa. https://www.nasa.gov/centers/jpl/history/when-computers-were-human.

OpenAI (2023). GPT-4 Technical Report. https://openai.com/research/gpt-4. Accessed August 2025.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., and Lowe, R. (2022). Training language models to follow instructions with human feedback.

Petrzela, N. M. (2020). Trying to Get in Shape? Here's the History Behind the Common New Year's Resolution. *Time*.

Plato (1997). Phaedrus. In Cooper, J. M., editor, *Plato: Complete Works*, pages 506–556. Hackett Publishing Company. Especially sections 274c–275b.

Post, F. . N. Y. (2024). More Gen Zers are becoming NEETs - what does it mean and is it a bad thing? *New York Post*.

Rosen, S. (1981). The Economics of Superstars. *American Economic Review*, 71(5):845–858.

Schaeffer, J., Zhao, W., Santurkar, S., Raffel, C., Belinkov, Y., and Recht, B. (2023). Are Emergent Abilities of Large Language Models a Mirage? *arXiv preprint arXiv:2304.15004*.

Seed, A. and Byrne, R. (2010). Animal Tool-Use. *Current Biology*, 20(23):R1032–R1039.

Stern, M. (2008). The Fitness Movement and the Fitness Center Industry, 1960–2000. *Business History Review*. Discusses growth of fitness centers from the 1960s to 2000.

Trivedi, H., Lu, X., Krishna, K., and Iyyer, M. (2023). Interleaving Retrieval with Chain-of-Thought Reasoning for Knowledge-Intensive Multi-Step Questions. *arXiv preprint arXiv:2305.05642*.

United States Department of Agriculture (1959). U.S. Equine Population, 1910–1959.

Weizenbaum, J. (1966). ELIZA—A Computer Program for the Study of Natural Language Communication Between Man and Machine. *Communications of the ACM*, 9(1):36–45.

Wernicke, C. (1874). Der aphasische Symptomencomplex: Eine psychologische Studie auf anatomischer Basis. *Cohn and Weigert*.

Wiener, N. (1950). *The Human Use of Human Beings: Cybernetics and Society*. Houghton Mifflin.

1 Why Perfectly Safe Generative AI is Probably Out of Reach

In this addendum I want to discuss the perils of producing a safe large language model (LLM), in order to form an idea of whether or not we can mathematically or programmatically rescue the naive or vulnerable from erroneous model responses[43]. Lets start with reiterating that a large language model does not "know" facts directly; instead, for any given prompt X, it computes a probability distribution $P_\theta(Y \mid X)$ over possible token sequences Y (the responses), where θ are the learned weights of the model (the classical econometrician can read θ as e.g. the coefficient estimates of an OLS). These weights arise jointly from the *training data*, the *model architecture and capacity*, and the *optimization process*. The very terminology often used (e.g., "hallucination") is itself anthropomorphic, obscuring the fact that the model is not "perceiving" or "misremembering," but simply sampling from a learned probability distribution.

With this in mind we discuss the *topology* of the response space of a large language model (LLM) to formalise our reasoning. Let $\mathcal{O}$ be the set of all possible token sequences (outputs) in the model's output vocabulary, which is nowadays already sanitised from arbitrary nonsensical sequences. Inside $\mathcal{O}$, let $\mathcal{S} \subset \mathcal{O}$ be the set of syntactically, grammatically and vocabulary-wise valid outputs, well-formed under the target language. Inside $\mathcal{S}$, let $\mathcal{C} \subset \mathcal{S}$ be the set of *coherent* outputs i.e. those that are meaningful and internally consistent. Finally, let $\mathcal{F} \subset \mathcal{C}$ be the set of true, safe, and factually grounded outputs.

- $\mathcal{O} \setminus \mathcal{S}$: "Kingba derpow time the fox 9reenly.": ungrammatical mixed with nonsense.
- $\mathcal{S} \setminus \mathcal{C}$: "The orphan's father washed her clothes.": grammatically fine but meaningless.
- $\mathcal{C} \setminus \mathcal{F}$: "The capital of France is Berlin.": coherent but factually wrong.
- $\mathcal{F}$: "The capital of France is Paris.": coherent and factually correct.

If a model's response lands outside $\mathcal{S}$, most humans will instantly recognize it as broken and be safe. If it lands in $\mathcal{S} \setminus \mathcal{C}$, it may still be dismissed as nonsense by most people, though some vulnerable users could be misled in subtle cases. The greatest danger lies in $\mathcal{C} \setminus \mathcal{F}$ (coherent, fluent falsehoods) which is both vast and difficult to detect without external verification and increases as most LLMs can consistently hit $\mathcal{C}$ creating a false confidence among the gullible.

Heuristic size argument. A simple way to see that the space $\mathcal{C} \setminus \mathcal{F}$ is large is by considering that for many factual templates ("The capital of X is Y") there are far more false instantiations than true ones. If a fact has n plausible surface forms but only 1 is true, then for each such slot-filling pattern the false-to-true ratio is roughly $n - 1 : 1$.[44] Since the number of such patterns grows combinatorially with statement length L, the fraction of $\mathcal{C}$ occupied by $\mathcal{F}$ shrinks rapidly as L increases. In realistic, unstructured domains, this means an overwhelming proportion of coherent outputs are wrong.

With this framing, the safety challenge can be analyzed as follows.

[43] Of course this is not to say that seasoned and experienced content creators are risk free. We can imagine, for example, the long term effects of a seasoned Python programmer's over-reliance on "vibe coding".

[44] There are currently 193 Member States of the United Nations (UN), according to the official UN membership list: https://www.un.org/en/about-us/member-states. By contrast, the GeoNames database, drawing from OpenStreetMap and other sources, lists over 11.8 million geographic features and 25 million place names globally: https://www.geonames.org/statistics/. Thus, for the template "The capital of X is Y," there are only ≈ 193 correct instantiations, but roughly $193 \times 25 \times 10^6$ possible combinations, implying that just one in about 25×10^6 coherent outputs is factually correct.

1. **Even perfect learning of the training distribution is not perfect truth.** The probability distribution that could, in principle, be distilled from the training data is not the same as the truth distribution, that is, the ideal probability distribution over all possible statements that assigns probability 1 to every true statement and probability 0 to every false one. Unless the training data is 100% correct, which is practically impossible, there is a built-in "contamination floor" ε:

$$\varepsilon(X) = 1 - \sum_{Y \in \mathcal{F}} P_\theta(Y \mid X),$$

the probability that the model leaves $\mathcal{F}$ for prompt X.

In large web-scale corpora, even after filtering, realistic residual factual errors persist: GPT-3 truthfulness is 58% on Truthful Q&A[45] and GPT-4 continues to demonstrate deficiencies, e.g., it hallucinates references in approximately 28.6% of cases when generating citations in scientific systematic-review contexts.[46]. Even if the training data *were* 100% correct, we would still face the question of *coverage* (point 2), and the model's own capacity and inductive bias can still leave probability mass outside $\mathcal{F}$.

2. **Information bottleneck.** Kolmogorov complexity, denoted $K(s)$, measures the length of the shortest computer program that can produce a string s;[47] some strings are *incompressible*: their shortest possible description is essentially the string itself.

An LLM may be thought of as a compression machine of its training data: it takes the corpus, tunes its parameters, and the ratio of the input to the size of the model is the compression ratio. For scale, combining just a few major public text and code sources yields a staggering corpus: English Wikipedia (26 TB),[48] Common Crawl (8 PB),[49] the U.S. Library of Congress Web Archives (45 PB),[50] public GitHub code ($\sim$ 1.5 PB),[51] arXiv ($\sim$ 1 TB),[52] and PubMed Central ($\sim$ 0.5 TB)[53] sum to roughly 54.5 PB of material.

And this list is far from exhaustive. Adding other national web archives (tens of petabytes), global patent databases (several petabytes), large-scale digitized public-domain book collections (tens of petabytes), technical standards, court records, and public forums/wikis outside Wikipedia could plausibly push the open-text total into the $\sim$ 280 PB rang, and this is still *not* counting material in other languages.

By comparison, a leading open-weight model such as Meta Llama 3 (70 billion parameters) occupies about 141 GB in 16-bit floating-point format (two bytes per parameter).[54] Treating the corpus as the "input" and the model as the "compressed form" gives a crude overall compression ratio of $\frac{280\text{ PB}}{141\text{ GB}} \approx 2.0 \times 10^6$, or over two million to one: roughly like squeezing that entire combined corpus into just 14% of the storage of a top-tier smartphone costing

[45] Lin et al. (2022): GPT-3 truthful $\approx$ 58% vs. human 94%.
[46] Chelli et al. (2024)
[47] https://en.wikipedia.org/wiki/Kolmogorov_complexity
[48] https://en.wikipedia.org/wiki/Wikipedia:Statistics
[49] https://commoncrawl.org/blog/august-september-2024-newsletter
[50] https://www.loc.gov/programs/web-archiving/
[51] https://archiveprogram.github.com/
[52] https://info.arxiv.org/help/bulk_data.html
[53] https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/
[54] https://community.ibm.com/community/user/cloud/blogs/arindam-dasgupta/2024/09/18/calculating-gpu-requirements-for-efficient-llama-3

under $1,500.[55][56]

At such extreme compression, and with incomplete coverage, it is inevitable that rare, idiosyncratic, or low-pattern truths (those with high $K(s)$) will be lost or distorted.

3. **Scaling law reality check.** Empirically, LLM quality improves as a power law in both dataset size and compute budget.[57] For example, pushing from 90% to 95% accuracy may require a modest multiplier in data or compute, but pushing from 99% to 99.9% can require *orders of magnitude* more, often beyond realistic training budgets. Architectural changes (e.g., Mixture-of-Experts, retrieval-augmented transformers) alter constants but do not remove the asymptotic slowdown. Crucially, the last few percent of errors are *not* uniformly random: they may cluster in **rare, high-stakes scenarios** (safety-critical domains, adversarial prompts, long-tail factual events). This means that even if the average-case error rate seems tiny, the *residual* unsafe cases can remain policy-relevant and economically damaging.

4. **Multi-turn contamination and safe prompt set complexity.** In multi-turn interaction, the model's own outputs feed back into the prompt stream, either explicitly (the user copies them back) or implicitly (they shape user follow-ups). If we think of $\mathcal{X}_{\text{safe}}$ as the set of prompts that, when given to the model, produce only safe outputs, then for bounded length L the safe prompt set $\mathcal{X}_{\text{safe},L}$ inherits the same combinatorial hardness as $\mathcal{F}_L$. A simple counting heuristic: over an input alphabet A, the number of prompts of length at most L grows as $O(|A^L|)$; unless safety constraints admit a very compact characterisation, enumerating or certifying membership in $\mathcal{X}_{\text{safe},L}$ scales correspondingly. Real-time prompt Quality Control (QC) would require:

   - high-precision semantic parsing,
   - domain-specific fact-checking (often with retrieval), and
   - adversarial-input detection robust to prompt obfuscation.

   Each component is imperfect, costly to run at scale, and vulnerable to circumvention by clever adversaries or accidental drift.

5. **Prompt-space feedback into training.** Most deployed models, especially those fine-tuned with Reinforcement Learning from Human Feedback (RLHF), periodically retrain or fine-tune on *real user prompts and completions*.[58] This means that unsafe or biased patterns in user interaction can *re-enter* the training distribution. Because the model shapes what users see (and thus the prompts they post), this creates a **feedback loop**, a form of endogenous contamination, where the prompt space gradually drifts in ways that may not be obvious until large-scale safety degradation is observed. This self-referential retraining can induce *model deterioration*, progressively reinforcing its own idiosyncrasies or factual errors, especially if external verification is weak, leading to drift away from the original truth set.

[55] For example, the Apple iPhone 15 Pro Max with 1 TB storage capacity; see `https://www.apple.com/iphone-15-pro/specs/`.

[56] Of course, such a phone could not "uncompress" this model (i.e., use it for inference, meaning generating responses) as running inference at useful speed would require vastly more computing power and electrical energy, on the scale of a large server cluster rather than a handheld device.

[57] Kaplan et al. (2020); Hernandez et al. (2022) show benchmark loss scaling with dataset size, model size, and compute; gains from extra "nines" of accuracy require disproportionately more resources.

[58] Christiano et al. (2023) introduce RLHF; Ouyang et al. (2022) demonstrate its large-scale application in GPT models using live prompt data.

6. **Open world, moving target.** Even if we could perfectly fence off $\mathcal{F}$ today, tomorrow's facts, norms, and adversarial techniques can push outputs outside it.[59] Examples:

   - geopolitical facts change (leaders, borders, capital cities),
   - safety norms evolve (medical guidelines, terminology sensitivity),
   - jailbreak prompt engineering develops faster than defensive filters.

   Thus, $\mathcal{F}$ is **time-dependent**, and maintaining perfect safety requires *continuous* retraining, re-verification, and prompt-set monitoring, each with lag and cost. In current practice, substantial model improvements are not truly incremental: changes in training data, architecture, or tokenizer usually require recomputing the entire weight set θ from scratch, which both limits the ability to cheaply and continuously adapt and introduces quality discontinuities, where gains in some areas coincide with regressions in others.

7. **Only robust mitigation.** The most viable mitigation today is *layering external systems* (fact-checkers, retrieval modules, theorem provers) **outside** the base model, tuned to the specific model's failure modes.[60] These work well in **narrow, structured domains** where truth conditions are crisp and cheaply checkable (for example, mathematical proofs or code compilation). In broad, open-ended domains, coverage gaps remain large.

   Illustrative case – Python code generation: here $\mathcal{F}$ is the set of all programs that meet a formal specification, something that can often be checked automatically with unit tests (small, targeted checks that verify whether individual functions or program components work as intended). This is a *best-case* safety scenario, yet even the strongest models succeed on only a fraction of problems when judged by "pass@k", the proportion of test cases solved when the model is allowed to try up to k different solutions. Harder, repository-level benchmarks reveal further brittleness: on EvoCodeBench, GPT-4 achieves only about 20.7% pass@1 in real-world, repo-style prompts, highlighting the gap between unit-test success and reliable, large-scale code synthesis.[61][62]

Bottom line: Without perfectly clean, complete, and unchanging training data *and* strict prompt control, contamination from interaction and continual learning will reintroduce errors. Model capacity and scaling constraints mean "always safe" for all prompts is unrealistic. Policy should focus on restricted, verifiable domains and robust abstention; protecting vulnerable users likely would likely require mediated inputs and outputs or constrained systems, with significant trade-offs.

[59] Bommasani et al. (2022) discuss performance instability and domain shifts in foundation models.

[60] Ji et al. (2023) survey persistent hallucination and safety failures in open-domain natural language generation; Trivedi et al. (2023) show that retrieval augmentation can mitigate some of these failures but does not eliminate them.

[61] Li et al. (2024) introduce EvoCodeBench, an evolving code-generation benchmark aligned with real-world repositories.

[62] The recently released GPT5 seems to inch forward on certain standardised tests: https://openai.com/index/introducing-swe-bench-verified/, https://aider.chat/docs/leaderboards/ and https://openai.com/index/introducing-gpt-5/

2 Learning structure vs. Learning Truth

In this section we discuss, in plain English, the difference between stochastic inference (an LLM commanding the syntax) and truth. While it is written for a layperson, most of the ideas can be made mathematically precise using Kolmogorov complexity, entropy rates, and related tools.

When you think about all possible strings of English letters and punctuation, imagine them as an enormous ocean, $\mathcal{O}$. Most of this ocean is gibberish. Inside it lies a small, smooth island, $\mathcal{S}$, which contains all syntactically valid English sentences. Somewhere on that island there is a tiny, scattered patch of sand, $\mathcal{F}$, representing the sentences that are both grammatical and factually correct.

An LLM is trained to predict the next token, effectively learning the distribution over characters or words that appears in its training corpus. This learned distribution is excellent for capturing the statistical patterns of syntax, and it is also the only tool the model has for recovering facts. The problem is that while this distribution carries strong local constraints that allow the model to reconstruct grammaticality with high accuracy, it does not carry enough information to guarantee correctness of specific facts, especially rare ones. Syntax is like the shape of the container; facts are the rare items inside it.

For example: "Marie Curie was born in Warsaw in 1867" lies inside both $\mathcal{S}$ and $\mathcal{F}$. "Marie Curie was born in Paris in 1931" lies inside $\mathcal{S}$ but outside $\mathcal{F}$. "Curie Warsaw 1867 born Marie" sinks back into the ocean: it is outside $\mathcal{S}$.

Large language models can almost always land on the grammar island $\mathcal{S}$, even though they have seen only a tiny fraction of all possible valid sentences. This is because syntax is locally learnable: the same grammatical rules appear over and over in different contexts. Once you have seen enough examples, you can reconstruct the "local shape" of grammar and generalize it to new sentences. This is like standing on the circumference of a circle: if you know the slope of the tangent line and the curvature at your point, you can reconstruct the curve in your immediate neighborhood. Syntax is smooth and redundant, so local information suffices to navigate it anywhere.

Truth is a global property: knowing what is correct in one part of the world tells you little about what is correct elsewhere. You cannot deduce Marie Curie's birthplace from generic birth rules: you have to know the fact itself. Rare facts appear infrequently in training data and do not generalize the way grammar does. In the circle analogy, truth is not about local curvature at all; it is about knowing which scattered points on the circle are painted red. To know that, you would need to see the whole circle, not just your local neighborhood. Even if the internet contained only correct facts, LLMs would still get some wrong because they do not store a literal fact table. They compress everything into patterns in their parameters, which causes blurring (rare facts averaged with others), interference (new learning distorting older memories), sampling errors (high-probability facts replaced by wrong tokens), and prompt mismatch (a slightly different wording steering the model away from the right fact).

Two analogies make this gap intuitive. First, think of chess: you can learn the rules of legal moves (syntax) in five minutes, but becoming a good chess player requires a lifetime of practice and knowing every game ever played as the rules tell you nothing about that (truth). Second, think of a human who has read an entire encyclopaedia but has no access to it: they must answer from memory alone. They will get many common facts right but will inevitably misremember rare ones, especially when asked in unfamiliar ways. Imagine, as an experiment, training an LLM solely on the text of the Encyclopedia Britannica. It would probably learn syntax perfectly and produce flawless English, but recovering the exact facts would be a totally

different matter.[63]

Programming languages present a special case. In formal terms, they too have a distinction between syntax (what strings are legal code) and semantics (what that code means), but in practice for many common programming tasks, these two sets overlap much more than they do in natural language. Syntax errors are fatal, and following standard idioms often brings you close to correct semantics. For Python in particular, which appears abundantly in training data, this high overlap means that getting syntax right often coincides with getting the intended functionality right. In rare languages such as Stata, the model has seen fewer examples, so the overlap between $\mathcal{S}$ and $\mathcal{F}$ is smaller and performance drops. Nonetheless, they are not identical sets: many programs are syntactically valid but semantically wrong, so even in programming the model can land in $\mathcal{S}$ without reaching $\mathcal{F}$.

It is tempting to think: if the model can write perfect sentences, it could use that ability to form database queries and fetch the truth directly. This is the idea behind retrieval-augmented generation (RAG): the model converts your question into a search query, a retrieval system finds relevant documents or database entries, and the model uses this retrieved text to produce an answer. RAG sidesteps the compression problem: facts need not be stored in weights, they can be pulled from an up-to-date source. This builds a bridge from the language island $\mathcal{S}$ back to the real world $\mathcal{W}$.

RAG helps, but it is not a magic bullet because it inherits the limitations of retrieval and integration. If the source is outdated or wrong, the answer will be too. Queries may be ambiguous or poorly formed and miss the right fact entirely. Even when relevant documents exist, ranking errors may bury them behind less relevant ones. The LLM can only see a finite context, so key details may be cut off. And once retrieved facts are seen, the model may override them with its own internal memory, especially if that memory reflects more frequent but outdated information. Maintaining and updating retrieval systems adds further engineering complexity.

The core asymmetry is that syntax is a smooth, redundant structure like a circle whose curve you can trace from local slope and curvature, so it is easy to learn and generalise from a small sample. Truth is a sparse, global property like identifying scattered red points on the circle, so you need far more coverage to get it right, and compression or retrieval errors quickly knock you off target. In programming, $\mathcal{S}$ and $\mathcal{F}$ overlap heavily, which is why LLMs can do better there with enough exposure, but they are not identical. RAG extends the LLM's reach into $\mathcal{F}$, but it still cannot guarantee truth without flawless sources, perfect queries, accurate ranking, and faithful integration. Without these, the navigator can sail the grammar island perfectly yet still pick up the wrong grain of sand from the truth patch.

[63] This would, in fact, be an interesting experiment which, if it has not been done yet, should be: What percent of the facts does an LLM get wrong if learning from the Encyclopedia Britannica allows it to write perfect English?