# Large Language Models Threaten Language's Epistemic and Communicative Foundations

**Shashank Srivastava**
UNC Chapel Hill
ssrivastava@cs.unc.edu

## Abstract

Large language models are reshaping the norms of human communication, sometimes decoupling words from genuine human thought. This transformation is deep, and undermines norms historically tied to authorship of text. We draw from linguistic philosophy and AI ethics to detail how large-scale text generation can induce semantic drift, erode accountability, and obfuscate intent and authorship. Our work here introduces hybrid authorship graphs (modeling humans, LLMs, and texts in a provenance network), epistemic doppelgängers (LLM-generated texts that are indistinguishable from human-authored texts), and authorship entropy. We explore mechanisms such as "proof-of-interaction" authorship verification and educational reforms to restore confidence in language. LLMs' benefits (broader access, increased fluency, automation, etc.) are undeniable, but the upheavals they introduce to the linguistic landscape demand reckoning.

## 1 Introduction

*"Last year's words belong to last year's language*
*And next year's words await another voice"*

A simple covenant has underwritten all written language so far: that behind any text lies a human mind. This link between text and cognition has been the bedrock that made language an expression of human intention, demanding both attention from the reader, and accountability from the author (Winograd, 1972; Bender and Koller, 2020). Over a long evolutionary trajectory (including both ancient cuneiform tablets and modern digital ones), these assumptions steadily held true, and are woven into the fabric of how we understand language.

But the swift ascent of large language models (LLMs) in the past five years has begun to fundamentally reconfigure this relationship. These models (e.g., GPT (Brown et al., 2020), PaLM (Chowdhery et al., 2022), LLaMA (Touvron et al., 2023)) or DeepSeek (DeepSeek-AI, 2024) can generate well-polished and coherent text with minimal guidance. They can replicate stylistic nuances, rhetoric, and emotional tones (Schick et al., 2021; Solaiman et al., 2019), which were attributed solely to human creativity till very recently.

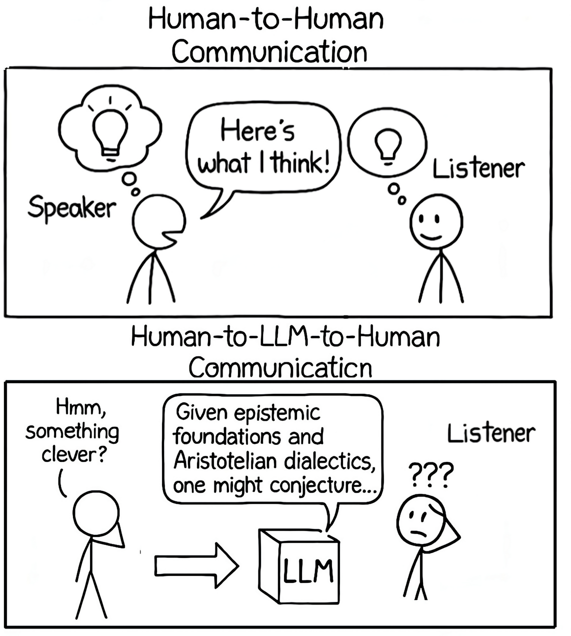
Human-to-Human Communication
Here's what I think!
Speaker
Listener
Human-to-LLM-to-Human Communicaticn
Hmm, something clever?
Given epistemic foundations and Aristotelian dialectics, one might conjecture...
LLM
Listener
???

Figure 1: LLMs introduce a shift in communicative dynamics. Traditionally, human-to-human communication directly conveys intentional thought from speaker to listener (top). But when mediated by LLMs, language can lose direct intentional grounding, resulting in messages disconnected from the speaker's original intent and confusing the listener (bottom).

On the one hand, these capabilities are surprising and extraordinary, suggesting potential for a new cognitive revolution in AI. On the other, this very success corrodes what made language trustworthy: its tether to a thinking, accountable mind. Consider the implications: A judge reads a beautifully written letter of remorse. A professor reads a student's deeply personal essay about overcoming adversity. But what if they were generated by systems with no

*Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 28662–28676
November 4-9, 2025 ©2025 Association for Computational Linguistics

experience of remorse or adversity? This is already happening. Teachers now worry that student essays reflect an LLM's fluency rather than messy human thought (Cotton et al., 2023; Zhou et al., 2023). Researchers question if a paper reflects a scholar's insights or a collage of existing work (Zellers et al., 2019). Even this sentence that you are now reading, how can you be certain that a human voice shaped these words? (Duarte et al., 2022; Weidinger et al., 2021). In this sense, it would be ironic if LLMs, instead of illuminating and edifying human communication, lead to its devolution.

To state that LLMs threaten language's epistemic foundations is not hyperbole, but an assessment of systemic risk. Much as climate change threatens coastal infrastructure by disrupting underlying systems rather than destroying every coastline, we believe the structures underpinning language face a similar strain. The broader NLP community, as creators of LLM technologies, bears a direct responsibility for their societal implications (Gabriel, 2020; Bender et al., 2021). By ignoring the epistemic consequences of these systems, we risk having text lose its role as an indicator of human intent and thought (Floridi and Chiriatti, 2020; Raji et al., 2022). This can fundamentally degrade how we conduct discourse, value expertise, and maintain trust (O'Neil, 2016; Zuboff, 2019).

This paper examines the tension between the benefits of LLMs (NLLBTeam, 2022; Hutchinson et al., 2023; Miller, 2019) and their impact on language, mapping the challenges and potential solutions for key stakeholders. Section 2 grounds the discussion in linguistic philosophy. Sections 3 and 4, which introduce authorship entropy and proof-of-interaction, are aimed at computer scientists and technologists. Sections 5 and 7, which cover implications and societal shifts, are intended for instructors, policymakers, and cultural institutions. Section 6 evaluates technical remedies, while Section 8 summarizes alternative views and arguments. We conclude with a reflection on the need to maintain language's cognitive and epistemic roles.

## 2 What Dies When Machines Write?

Language's core function has been debated from Plato's dialogues on rhetoric to modern analytic philosophy (Plato, 1997; Wittgenstein, 1953). While language is commonly viewed as an information channel for transmitting information, linguists argue that language is a *communal sense-making act*. It has been closely linked to intention, context, and the ability to hold speakers accountable (Searle, 1969; Austin, 1975; Floridi, 2013).

**Language as an Intentional Act:** Searle's speech-act theory (Searle, 1969) and Austin's work on performativity (Austin, 1975) argue that language is not just a conduit for information transfer, but also enacts intentions. To *say* something is often to *do* something: to promise, question, declare, for example. The force of an utterance depends on the speaker's agency and recognition of those intentions by a listener (Grice, 1975). This has been fundamental to authorship, particularly in academic and legal discourse, where a text is an intellectual act tied to its creator's identity and responsibility (Dworkin, 1996). Even when ghostwriters were traditionally involved, the text ultimately reflected a coherent cognitive source (Foucault, 1984; Chartier, 1994; Sperber and Wilson, 1986).

The rise of LLM-generated text disrupts these frameworks (Floridi and Chiriatti, 2020). Do AI-generated documents bear the same weight without deliberate intent? This raises questions about authorship, intellectual property and trust, especially for scientific or legal text (Raji et al., 2022; Huang and Rust, 2021; van Dis et al., 2023). Authorship has historically entailed a social contract: a published text can be challenged or critiqued, holding its human creator(s) responsible for its shortcomings (Woodmansee, 1994; Chartier, 1994). But with LLM-authored text, accountability becomes diffused: does it lie with the prompters, the model trainers, or the data creators? This diffusion strains legal and academic norms (Hacker et al., 2023; Mittelstadt and Floridi, 2016; Kosseff, 2019). As the intent and accountability of text gets murky, its value and trustworthiness can become suspect too.

**Language as a Cognitive Interface:** Beyond communication, language shapes cognition and our capacity to abstract and solve problems (Clark and Chalmers, 1998; Vygotsky, 1978; Whorf, 1956). It is often considered an "interface" to thought. Research in child cognitive development suggests that engagement with language enables reasoning, cognitive flexibility, and problem-solving (Tomasello, 2003; Bruner, 1983; Lakoff and Johnson, 1980). While some argue that LLMs function as cognitive enhancers (Clark and Chalmers, 1998; Warwick, 2003), others caution that reliance on LLM-driven generation can lead to reduced cognitive engagement (Nichols, 2021; Carr, 2011). In particular, LLM-driven writing and summarization has raised concerns about cognitive deskilling (Carr, 2011; Lai and Viering, 2022). Studies show that composition itself is integral to thinking, forcing individuals to clarify ambiguity, structure arguments, and synthesize knowledge (Kellogg, 2008; Galbraith, 1999). Further, LLM-generated summarization risks eroding cognitive *effort*, like digital offloading has been shown to reduce critical engagement (Sparrow et al., 2011; Nichols, 2021).

**Chain-of-Thought and Cognitive Parallels:** The emergence of chain-of-thought prompting (CoT) (Wei et al., 2022) represents a major shift in LLM problem-solving. By externalizing logical steps, CoT compensates for the depth limitations of transformer architectures (Vaswani et al., 2017; Yao et al., 2023). This mirrors how humans articulate thoughts through language, diagrams, or writing to enhance problem-solving (Clark and Chalmers, 1998; Menary, 2010). Beyond computational efficiency, CoT also bears parallels to how externalizing reasoning through symbols has been linked to the expansion of human intelligence (Deacon, 1997; Dor, 2015). If language enabled humans to extend cognition beyond individual memory, CoT might mark a similar milestone in LLMs.

## 3 The Crisis of Language

LLMs shatter these foundations through two mechanisms. (1) *Semantic Drift & Model Collapse*, the idea that the influx of AI-generated text can shift the distribution and meaning of language, and lead to compounding errors; and (2) *Eroding Epistemic Trust* in text, epitomized by what we term epistemic doppelgängers. We also suggest a metric, *authorship entropy*, to represent the uncertainty about the origin of a text.

### 3.1 Semantic Drift and Model Collapse

Semantic drift refers to changes in language usage and meaning over time, reflecting cultural and social evolution. However, large-scale LLM-generated content can accelerate or redirect semantic change. For example, they can reinforce common phrases while underrepresenting less frequent expressions (Raji et al., 2022). LLMs trained on text that partially includes their own synthetic outputs can experience compounding errors.

Model collapse occurs when AI repeatedly processes its own text, leading to a decline in quality (Shumailov et al., 2024). Initially, early model collapse erases rare linguistic forms and minority perspectives from the data distribution (Menick et al., 2022; Carlini et al., 2023). This can progress to late model collapse, where the model's range becomes limited to the mean outputs of LLMs. Although distribution shift has been studied in active learning (Blitzer et al., 2007), LLM self-ingestion is a new feedback loop. Each generation of models trained on synthetic text loses another shade of human expression, like a photocopy of a photocopy, until only the skeleton of language remains.

Semantic drift and model collapse are linked in a loop of human and LLM language production (Bommasani et al., 2021). Human writing trains LLMs, whose outputs then affect human writing and future training data. This loop can reach an unintended balance where language evolution is driven by LLM statistical features, not human creativity, threatening the linguistic ecosystem. The loss of the distributional tails during early model collapse can reduce linguistic innovation and cognitive diversity (Weidinger et al., 2021; Bender et al., 2021) due to homogenized language. Subsequent loss of variance during late model collapse can leads to loss of semantic clarity as the model's capacity to make fine-grained distinctions erodes. If much of our reading becomes machine-generated, we must also consider its societal impact on collective *human* cognitive diversity.

### 3.2 Eroding Trust and Accountability

Texts have long been vehicles for accountability. Historically, authors were usually identifiable, and could be praised, critiqued, or legally challenged for their claims (Woodmansee, 1994; Chartier, 1994). LLM-generated content fragments this responsibility. This has significant implications for defamation suits and retraction practices for erroneous statements (Kosseff, 2019). More pressingly, online campaigns leveraging AI threaten political discourse, as citizenry loses clarity on who authors the narratives shaping public opinion (Weidinger et al., 2021; Chesney and Citron, 2019).

Empirical evidence is growing that synthetic text can fuel misinformation. Recent experiments demonstrate that AI-generated content can create convincing but deceptive social media campaigns, obscuring authentic and automated discourse (Zellers et al., 2019). Language models can also plagiarize, amplify biases, and perpetuate stereotypes from their training data (Bender et al., 2021; Hovy and Spruit, 2016; Carlini et al., 2023).

Without strong authorship signals or provenance tracking, establishing credibility and accountability in digital information ecosystems becomes hard (Gehrmann et al., 2019; Kirchenbauer et al., 2023).

### 3.3 Epistemic Doppelgängers and Authorship Entropy

LLMs can produce text nearly indistinguishable from human writing. We refer to such outputs as epistemic doppelgängers: texts that impersonate human authorship so convincingly they can fool not only casual readers, but even domain experts. As with a human doppelgänger, the deception isn't necessarily malicious, but its uncanniness can be destabilizing. LLM-generated news articles have been rated as more trustworthy than authentic ones (Zellers et al., 2019), and even the best AI detectors rarely surpass 70% accuracy. Worse, detection systems are often only effective when closely matched to the model they're trying to catch, making them vulnerable to fine-tuning, or strategic prompting. In short, epistemic doppelgängers erode the assumption that a well-formed sentence signals a human mind.

This epistemic ambiguity leads us to what we call *authorship entropy*, a measure of uncertainty of text authorship. In a world where all documents are confidently human-written, authorship entropy is low: the provenance of text is legible, even if anonymous. But in an AI-saturated ecosystem, the space of plausible authors expands. By modeling this uncertainty as a probability distribution and applying Shannon entropy, we can quantify how "foggy" the authorship landscape is. Rising authorship entropy destabilizes trust: people may become suspicious of legitimate texts, or indifferent to provenance altogether. It weakens accountability: if we don't know who wrote something, we can't assign responsibility.

Concretely, authorship entropy can be operationalized as the Shannon entropy over a probability distribution of potential authors $A = \{human, llm_1, \ldots, llm_m, unknown\}$. The entropy for a text $T$ would be $\mathcal{H}(A|T) = -\sum_{a \in A} P(a|T) \log P(a|T)$, where the posterior probability $P(a|T)$ could be estimated using a combination of text classifiers, watermark detectors, and provenance metadata. This information-theoretic framing aligns with stylometric approaches to authorship attribution that also leverage distributional features of text.

## 4 A Framework for Verifiable Provenance

Our discussion thus far has been primarily conceptual. In this section, we explore technical interventions for reclaiming human accountability and reducing authorship entropy.

### 4.1 Author Graphs & Proof-of-Interaction

A possible direction is embedding provenance and requirement of human interaction in the text generation process itself. For example, a hybrid authorship graph can represent relationships between human users, LLMs and texts that they generate or indirectly influence. To explain, a document's node can have edges from an LLM node (if an AI drafted it) and a human node (who guided or edited it). If an AI's training data included that document, an edge from the document back to the AI node ("trains") can be included, forming a cyclic network of influence. Figure 2 shows an example of such a graph. Such explicit representations of provenance and sources can provide grounding to enforce downstream accountability.

A practical implementation of this can be through ***Proof-of-Interaction*** *(PoI)* mechanisms that ensure that a human was substantially involved in creating a text. For instance, an editor can sign off on an AI-generated passage after verifying it, or a platform can require that any AI assistance be logged and attested. Some have proposed protocols where documents carry embedded metadata or hashes that link to records of the human-AI collaboration that produced them. If a document cannot present such proof-of-interaction, it may not be trusted for certain uses.

Instead of inventing anew, this principle can be robustly implemented using existing technologies in cryptography, systems security, and authorship verification research (Stamatatos, 2009; Layton and Watters, 2020). Verifiable Computation aims to create a tamper-proof record of interactions (Thaler, 2023). For example, a text-editing tool could produce a Zero-Knowledge Proof (ZKP) to confirm human editing without revealing edit content (Ben-Sasson et al., 2014; Goldwasser et al., 2019). This PoI certificate can be embedded in a Secure Provenance framework like the C2PA manifest, ensuring a cryptographically secure record (Coalition for Content Provenance and Authenticity (C2PA), 2022). Interaction logging can be secured with a Trusted Execution Environment (TEE) to prevent tampering (Sabt et al., 2015). Although these methods add complexity, they represent a 'cost of trust' in a world where synthetic text dominates. This approach aligns PoI with human-in-the-loop systems, combining human oversight with automation (Cecchi and Babkin, 2024; Kang et al., 2024).

Although it doesn't address every issue, like adversaries simulating edits, it raises the cost of deception and offers an audit trail. This can be done using W3C Verifiable Credentials, creating a document-specific chain where each interaction is a signed credential in a directed acyclic graph (DAG) of contributions. To prevent Sybil attacks with impersonating bots, light proof-of-work or social verification may be needed to sign a credential.

This idea also aligns with Chain-of-thought (Wei et al., 2022; Kojima et al., 2022) oversight, where an LLM seeks human verification for intermediate reasoning. Some developers propose user-audited chains-of-thought, letting humans see exactly which steps an LLM took (Wang et al., 2022). Future research can unify chain-of-thought logs with proof-of-work, offering a secure record of how text was generated . This can clarify the roles of LLMs and humans in authoring text, although this approach may present challenges in preserving user privacy (Khowaja et al., 2023; Abadi et al., 2016; Glymour et al., 2023). These changes will necessarily introduce friction, and may be tedious for users. However, a proof-of-interaction system can ensure that every text is connected to at least one human via a "verified" edge. This can maintain the principle that for any published text, one can point to a human accountable for it.

## 4.2 Metrics for Semantic Drift

Verifying authorship addresses who wrote the text. But we should also ask what is being written. We propose tracking language changes by defining metrics for semantic drift and linguistic diversity, comparing human-authored and LLM-generated text periodically. By measuring shifts in word frequencies, syntax, or topics, we can identify drift if metaphoric language or dialectal terms decrease while AI-generated phrases increase.

It is also worth monitoring model-internal drift: how successive generations of LLMs differ when trained on data that includes prior LLMs' outputs. If $P_{human}$ and $P_{LLM_\theta}$ indicate the probability distributions of language utterances at a discrete time step $t$, and if $\alpha$ denotes the proportion of LLM generated data, then the distribution of training data that will be used to train the next iteration of the LLMs, $P^{t+1}_{LLM}$, is given by:

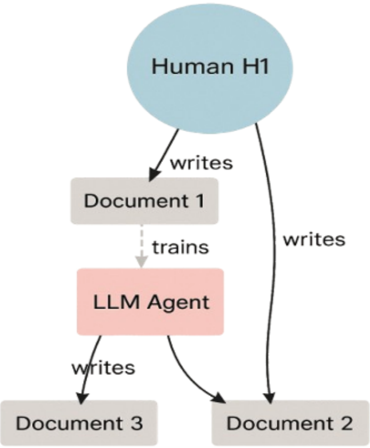

Figure 2: An illustrative hybrid authorship graph, representing provenance and interactions between human agents, an LLM agent, and texts. In this example, Human H1 writes Document 1, which is later used in training the LLM. Document 2 is co-authored by H1 and the LLM (perhaps H1 edited text generated by the LLM). Document 3 is authored solely by the LLM. To implement a verifiable proof chain, each node can have an associated cryptogrphic proof. *Proof 1:* Human H1 signs the initial version. *Proof 2:* an LLM Agent signs the generated draft. *Proof 3:* Human H1 signs the final edited version, with a `previousProof` property linking to Proof 2. This created an immutable, ordered record of contributions.

$$P^{(t)}_{\text{mix}}(\mathbf{x}) = (1-\alpha)P_{\text{human}}(\mathbf{x}) + \alpha P^{(t)}_{\text{LLM}_\theta}(\mathbf{x})$$

as LLM outputs re-enter training data (Shumailov et al., 2024). If $P^{(t)}_{\text{mix}}$ increasingly diverges from $P_{\text{human}}$, the model parameters $\theta$ risk converging to a subspace that fails to capture the richness of actual human language patterns. Further computational or theoretical insights may be found in work on catastrophic forgetting (Kirkpatrick et al., 2017) and domain shift (Ganin et al., 2016). While the notion is not new (prior studies on machine-in-the-loop domain adaptation raise similar concerns (Ruder, 2019)), our contribution is to highlight how large volumes of synthetic text can nudge language distributions away from natural usage. We propose coupling distributional metrics (e.g., KL divergence) with textual diversity indices to monitor linguistic homogenization.

Empirically testing these metrics on real corpora that blend human and AI-generated text remains a priority for future research. Experiments with smaller LLMs (Carlini et al., 2023; Menick et al., 2022) suggest that repeated synthetic ingestion am-

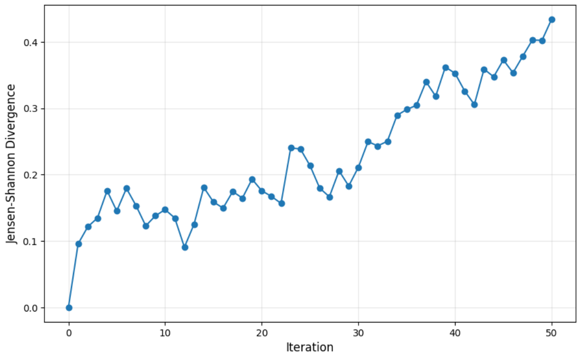

Figure 3: Semantic Drift in Synthetic Text: The plot shows the Jensen-Shannon divergence between a base human-authored text distribution and iteratively drifted synthetic text distributions. As synthetic text is repeatedly generated and reintroduced into training data, the divergence increases, illustrating the risk of semantic drift and potential loss of linguistic diversity over time. At each iteration, the GPT2 model is fine-tuned on text sampled from the GPT2 model in the previous step

plifies shallow lexical patterns. In Figure 3, we provide a simple demonstration of this recursive degradation. We plot the JS divergence between unigram distributions from a base human-authored text corpus, and the output of a GPT-2 model that is iteratively fine-tuned on its own generated samples. The increasing divergence illustrates the process of semantic drift, where the model's output distribution drifts progressively further from the original distribution with each generation, consistent with recent findings from (Shumailov et al., 2024). While the GPT-2 architecture used here is much simpler than modern LLMs, it isolates the recursive feedback loop that is a growing concern for contemporary models. Even large LLMs risk forgetting the true data distribution when trained on their own outputs (Shumailov et al., 2024) . We see this experiment not as a finding limited to older models, but as a demonstration of a systemic risk.

In summary, technical tools for authorship authentication (like detection, watermarking, and proof-of-interaction) and for linguistic monitoring (measuring drift and diversity) will be crucial parts of the solution. However, they are not panaceas. Detection can be evaded; verification can be cumbersome; drift metrics can tell us there's a problem but not fix it. Next, we look at how education, policy, and cultural norms can adapt to safeguard the epistemic foundations of language.

## 5 Societal Implications

LLMs seduce with their productivity and efficiency, while corroding trust in academic integrity, scholarship, professional hiring, and public discourse (Coglianese and Lehr, 2017; Cotton et al., 2023; van Dis et al., 2023).

**Education** LLMs are increasingly used as study aids, enhancing access for learners with different language skills (Khan et al., 2023; Chaudhuri et al., 2021; Xu et al., 2022; Luckin et al., 2023). However, relying on LLMs for tasks like programming or essay writing can weaken essential skills: algorithmic thinking, structured argumentation, and creativity (Cotton et al., 2023; Perkins and Salomon, 1989). To address this, some schools use real-time or proctored writing tasks or oral exams to ensure understanding (Lund and Wang, 2023).

**Academic Scholarship** The academic ecosystem assumes text is a reflection of an author's intellect. Automated text generation challenges this, raising concerns over AI authorship and scholarly contributions (Willis and Williams, 2023). In the short term, LLMs risk hallucinations (Ji et al., 2023) and plagiarism (van Dis et al., 2023). More seriously, they can flood peer reviews and obscure genuine innovation. Ideally though, LLMs should boost research productivity and accelerate scientific progress.

**Professional Settings** In many industries, cover letters, writing samples, and portfolio websites are used to gauge candidates' communication skills and expertise (Sternberg and Williams, 1997). LLM tools now make it easy to create polished but shallow applications, complicating hiring managers' ability to assess true abilities. Some organizations are turning to live assessments like real-time writing tests or structured panel interviews (Koch et al., 2015; Levashina et al., 2014). But these can be difficult for introverts, non-native speakers, or candidates who do better with written communication (van Tubergen and Kalmijn, 2014; Hu et al., 2020). Managing this requires a delicate dance between fairness and authenticity.

**The Public Sphere** LLMs are reshaping public discourse through AI-generated content, with concerns about amplification and distortion (Zellers et al., 2019; Ferrara, 2020). Disinformation campaigns exploit AI's capability to produce misleading content, drowning out human voices. At the same time, LLMs present opportunities to reduce barriers for individuals with limited writing skills, disabilities, or those who are non-native speakers (Paritosh et al., 2022; Xu et al., 2022).

In all of these domains just discussed, a common thread is that trust is threatened by automatic text generation from LLMs. As trust erodes, institutions will react by imposing stricter verification, leading to friction, surveillance, or cynicism. The challenge is developing norms that preserve the value of human contribution and ensure transparency.

## 6 Labels & Classifiers wont save us

Proposals such as watermarking and policy bans, while helpful in the short term, offer only superficial remedies (Gehrmann et al., 2019; Zellers et al., 2019; Papernot et al., 2016).

**Watermarking and Detection Arms Races** Watermarking remains fragile against adversarial attacks like paraphrasing (Kirchenbauer et al., 2023). Detection classifiers also struggle with robustness as LLMs adapt and human post-editing obfuscates machine origins (Holtzman et al., 2020). This creates a resource-intensive cat-and-mouse dynamic without stable solutions (Gallagher et al., 2023). More critically, these methods do not resolve attribution, leaving ethical and legal questions unanswered (Authors, 2023; Devinney, 2023). Table 1 contrasts the limitations of these detection-based methods with our suggested frameworks.

**Limitations of policy bans** Bans on AI-assisted writing are unenforceable due to weak detection and strong incentives for LLM use, effectively becoming honor systems (Devinney, 2023). Lagging legislation creates a fragmented regulatory landscape (Hacker et al., 2023), and the global nature of digital communication allows easy circumvention of local policies (Katyal and Epps, 2022).

Current measures do not restore a discernible human presence. If language's epistemic function relies on text as an intentional artifact, superficial fixes fail to reattach text to a human mind (Bender and Koller, 2020), and risk putting us in a permanently ambiguous linguistic landscape.

## 7 Rethinking Language & Authorship

We propose recalibrating the role of LLMs in language by leveraging their benefits while preserving human traits like intentionality, accountability, and diversity of thought. In this section, we refine previous suggestions and introduce ideas on human-AI collaboration frameworks, governance, and cultural appreciation of human-only work (Mittelstadt and Floridi, 2016; Floridi, 2019).

**Chain-of-Thought with Human Oversight** As mentioned in Section 4, a proposed model involves embedding AI within a structured chain-of-thought framework where human oversight is a required component at key decision points (Wei et al., 2022). In this paradigm, LLMs can generate partial outlines, intermediate arguments, or suggested revisions, but finalization has to be authenticated by a human user after consideration. By logging human-AI interactions through an auditable chain (with appropriate privacy safeguards), this method establishes a transparent record that delineates AI-generated content from human refinement, addressing concerns about accountability and intellectual ownership (Wang et al., 2022; Christiano, 2022).

**Governance & Collaborative Policy** Governance for LLM-usage has to be a negotiation (between policymakers, educators, and user communities, etc.) for it to work, rather than a prescription (Floridi and Chiriatti, 2020; Hacker et al., 2023). Several directions seem promising. First, AI contribution statements, similar to conflict-of-interest disclosures, can prompt authors to declare the extent and nature of LLM involvement (Devinney, 2023). Second, labeling protocols for governmental or legal texts can introduce metadata or disclaimers to flag LLM-generated contents (Union, 2023). Also, ethical AI certification programs, modeled on data protection seals, can help LLM developers conform with regulations such as the EU AI Act (Union, 2023).

**Educational Shifts** To prevent cognitive deskilling, education has to pivot. Assignments will have to adapt to the inevitability of the use of LLMs for drafting, but can require students to justify revisions (Lai and Viering, 2022). Assessments will have to focus on substance over polish (Paul and Elder, 2007; Lipman, 2003; Chi and Wylie, 2014; Freeman et al., 2014). Finally, students have to be encouraged to play with LLMs, and taught to interrogate them. AI literacy needs to be a form of critical literacy, where students learn not simply how to use LLMs, but to think critically about their construction, capabilities, and limitations (Bowman and Reeves, 2015). This approach is already being implemented through scalable, open-access curricula. For

| Method | Basis | Epistemic Value | Limitations |
|---|---|---|---|
| Watermarking | Hidden statistical patterns | Low (indirect & fragile) | Breakable via paraphrasing |
| Provenance Tags | Declared metadata | Low (unverifiable, spoofable) | Fails under adversarial editing |
| Authorship Detection | Content-based classification | Probabilistic (not auditable) | Inaccurate under obfuscation |
| **Authorship Graphs** | Signed interaction DAGs | High (full provenance) | Requires structured logging |
| **Proof-of-Interaction** | Verifiable edit/dialogue logs | High (grounded in behavior) | Needs UI/platform integration |
| **Authorship Entropy** | Entropy over agency paths | Medium (interpretable signal) | Sensitive to models |

Table 1: Summary comparison of detection-based methods with proposed approaches.

example, Code.org's 'AI Foundations' provides K-12 modules on training data, bias, and societal impact (Code.org, 2024). Also notable is Stanford University's CRAFT initiative (Stanford University, 2023), which offers free resources for high school teachers to integrate AI topics into humanities, arts, and social studies.

**Human-Only Publishing Spaces** A potential societal adaptation may be the emergence of "human-only" publishing spaces: media outlets, or creative communities that employ verification measures (such as mechanisms like proof-of-work logs) to ensure that any content reflects considerable human intellectual effort and creativity. These spaces may offer a parallel track for those who value human expression without the aid of LLMs. Many journals already forbid undisclosed AI collaboration for final submissions (Board, 2023). A fiction community might pride itself on entirely human-crafted stories. Like organic labels in food, these spaces can serve audiences that value authentic human expression, akin to culinary 'slow food' movements (Petrini, 2001). While these enclaves could serve as a valuable control group, they would also introduce complex trade-offs, potentially leading to information siloing or accusations of elitism. The viability of such spaces would depend on whether we value traditional norms of authorship enough to support a potentially less efficient economy.

## 8 Alternate Views

Many scholars have a more sanguine outlook, and do not see LLM-based text as a threat to language.

**Increased Accessibility** LLMs can enable non-native speakers and individuals with disabilities to participate in public discourse (Norvig and Thrun, 2009; Ogawa et al., 2022). By automating surface-level writing concerns, these tools allow users to focus on substantive ideas. For example, spell-checkers, were once controversial too (Felton, 2023; Christiansen, 2021). Additionally, LLMs increase accessibility for users with impairments (Wagner et al., 2020), reframing 'linguistic inclusivity' as a positive evolution.

**Accelerated Knowledge Dissemination** Summarization tools help researchers digest literature efficiently (Fabbri et al., 2022; Sharma et al., 2022), and multilingual translation expands access to specialized knowledge (Fan et al., 2021; Artetxe and Schwenk, 2019). With editorial oversight, these outputs can enhance comprehension without compromising reliability (Szegedy et al., 2022). Advocates argue that with transparency, LLMs can strengthen epistemic ecosystems rather than harm them (Diakopoulos, 2016).

**Adaptive Norms of Collaboration** In many domains, collaborative authorship is standard (technical manuals, corporate reports, etc.), which rarely reflect a single voice (Darics, 2020; Leonard and Noonan, 2020). In this context, LLMs are seen as additional collaborators (Krause et al., 2022; Dinan et al., 2022). Rather than undermining authorship, they may shift workflows, with new roles emerging for human editors and fact-checkers (Eisenstein and McNamara, 2023; Roose and Sullivan, 2023).

**Evidence of Positive Outcomes** Some studies suggest that, used responsibly, LLMs can enhance writing without weakening critical thinking. They support non-native and novice writers in building fluency (Lee et al., 2022; Laubrock et al., 2022). In collaborative environments, there is evidence that AI systems help clarity, and can identify redundancy (Yosinski et al., 2023; Rahimi et al., 2021).

Broadly, these perspectives argue that LLMs are not existential threats to the integrity of language, and that a 'crisis' of authorship is neither new nor uniquely AI-induced. Rather, this is a natural evolution in how we produce and share ideas. Possibly, when questions about the origin and intent of a text fade, newer and better-suited norms can emerge in the linguistic landscape to replace them.

## 9 Reflection

We have examined how LLMs disrupt the relationship between text and human cognition. Our analysis reveals three mechanisms: epistemic doppelgängers that destabilize textual interpretation, rising authorship entropy that undermines communicative trust, and recursive semantic drift that disrupt this link. The comfortable story would be that we will develop better detectors, establish clearer policies, and restore the old certainties. But our analysis suggests otherwise. Watermarks will be stripped, detection will be evaded, and the recursive contamination of text has already begun.

Our proposed techniques: proof-of-interaction protocols, hybrid authorship graphs, and semantic drift metrics should be seen not as solutions but as preservation mechanisms. These approaches may maintain domains of verified human expression within an increasingly synthetic textual landscape. The human-only publishing venues we envision function as controlled environments where traditional assumptions about authorship and intentionality can persist. While they can preserve certain epistemic and educational functions, they implicitly acknowledge that the broader linguistic ecosystem has undergone irreversible transformation.

This transformation demands reconceptualizing traditional ideas of authorship. Hybrid authorship graphs may become a dominant mode of text production: collaborative human-AI systems where attribution and accountability must be tracked rather than assumed, and LLMs augment human expression, rather than replace it. Similarly, our authorship entropy metric quantifies what practitioners already experience: increasing uncertainty about the cognitive origins of any given text.

The challenge is not preventing synthetic text proliferation: this is both impossible and likely undesirable. Rather, we must preserve the essential functions that human-originated language has served: as evidence of thought, as a tool for cognitive development, and as a basis for accountability. This requires recognizing that we are not simply adding new tools to existing practices, but potentially altering the nature of language as a human institution. Negotiating this requires both (1) technical standards like verifiable credentials, and (2) new forms of AI literacy to equip people to navigate a world where most text is no longer of human origin. It may be possible to harness LLMs' advantages without surrendering the uniquely human dimensions of language. But a failure to act can lead to communication devoid of color, and diluted cognitive depth. The path to hell is famously paved with good intentions. Still, through ingenuity and foresight, we may steer LLM innovations towards enhancing human creativity, rather than eroding its intellectual bedrock.

## Limitations

We introduce conceptual tools like epistemic doppelgängers and authorship entropy to make sense of the shifting linguistic terrain. However, many of these constructs remain speculative without empirical grounding. Nor do we pretend that quantitative metrics alone can capture the consequences of LLM saturation. What we offer is a framework to think with, not a solution to deploy.

Second, some of our proposals (such proof-of-interaction logs, and human-only publishing enclaves) are challenging to reify. They require infrastructure, cooperation, and cultural shifts that may not be welcome. We also acknowledge a significant tension: the paper champions the pre-eminence of human intention in language, but we do not promote gatekeeping expression or discourage the increasingly creative uses of LLMs. The challenge is to protect the epistemic integrity of language without devolving into purity tests. To truly solve this challenge will requires contending with lived social realities of people, not just technical design.

## Acknowledgements

The core ideas for this paper are due to my older brother, Nisheeth Srivastava, whose television interview on the effect of AI on society I had dutifully mocked. My ridicule, it turns out, was a high form of engagement. This perhaps demonstrates a corollary to our main thesis: even stochastic listeners can become vessels for ideas, transforming them in ways the original speaker never intended.

## AI Use Acknowledgment

In equal parts ironically and fittingly, this paper used assistance from LLMs in several aspects of its creation, disclosed here in accordance with the ACL Policy on AI Writing Assistance. This included help with literature research, proof-reading and refinement, and generating code. Specifically, we utilized AI tools for searching relevant literature and for writing analysis code related to Figure 3 and diagram generation code for Figure 2

## References

Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 308–318.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

J. L. Austin. 1975. *How to Do Things with Words*, second edition. Harvard University Press.

Anonymous Authors. 2023. GPT and the plagiarism problem: Assessing AI's impact on academic integrity. *Journal of Ethics in AI*.

Eli Ben-Sasson, Alessandro Chiesa, Eran Tromer, and Madars Virza. 2014. Succinct non-interactive zero knowledge for a von neumann architecture. In *23rd USENIX Security Symposium (USENIX Security 14)*, pages 781–796.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 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)*, pages 610–623.

Emily M. Bender and Alexander Koller. 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 (ACL)*, pages 5185–5198.

John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, bollywood, boom-boxes, and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 440–447.

Nature Editorial Board. 2023. AI and authorship: Nature's policy on undisclosed ai collaboration.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, and 1 others. 2021. On the opportunities and risks of foundation models. *arXiv e-prints*, pages arXiv–2108.

Nicholas A. Bowman and Anne E. Reeves. 2015. Rethinking media literacy in the age of algorithmic curation. *Journal of Media Education*, 6(3):14–24.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:1877–1901.

Jerome Bruner. 1983. *Child's Talk: Learning to Use Language*. W. W. Norton  Company.

Nicholas Carlini, Matthew Jagielski, Chiyuan Zhang, Katherine Lee, Christopher A. Choquette-Choo, Jacob Imber, Andreas Terzis, Nicholas Frosst, Ilya Mironov, Vasisht Duddu, and 1 others. 2023. Poisoning web-scale datasets is practical. *arXiv preprint arXiv:2305.00956*.

Nicholas Carr. 2011. *The Shallows: What the Internet is Doing to Our Brains*. W. W. Norton & Company.

Lucas Cecchi and Petr Babkin. 2024. ReportGPT: Human-in-the-loop verifiable table-to-text generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 529–537.

Roger Chartier. 1994. *The Order of Books: Readers, Authors, and Libraries in Europe Between the Fourteenth and Eighteenth Centuries*. Stanford University Press.

Soham Chaudhuri, Varun Kumar, and Marti Hearst. 2021. AI-powered writing assistants: Enhancing learning and creativity. In *Proceedings of the Conference on Educational Data Mining (EDM)*, pages 344–355.

Bobby Chesney and Danielle Keats Citron. 2019. Deepfakes and the new disinformation war: The coming age of post-truth geopolitics. *Foreign Affairs*, 98(1):147–155.

Michelene T. H. Chi and Rachel Wylie. 2014. The ICAP framework: Linking active learning to cognitive engagement. *Educational Psychologist*, 49(4):219–243.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Anselm Levskaya, Tyler Wang, Nan Du, Yinhan Liu, and 6 others. 2022. PaLM: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Paul Christiano. 2022. AI alignment and the role of human oversight in generative models. *Alignment Research Journal*, 4:101–128.

Meredith Christiansen. 2021. Algorithmic writing and digital literacy: How ai is reshaping composition. *Digital Studies*, 12:55–73.

Andy Clark and David Chalmers. 1998. The extended mind. *Analysis*, 58(1):7–19.

Coalition for Content Provenance and Authenticity (C2PA). 2022. C2pa technical specification. https://spec.c2pa.org/.

Code.org. 2024. AI Foundations. https://code.org/ai.

Cary Coglianese and David Lehr. 2017. Regulating by robot: Administrative decision making in the machine-learning era. *Georgetown Law Journal*, 105:1147–1223.

Debbie Cotton, Peter Cotton, and Sarah Shipway. 2023. ChatGPT, AI and the impact on academic integrity: Ai-assisted student writing. *International Journal for Educational Integrity*, 19(1):1–16.

Erika Darics. 2020. E-voice: A multimodal perspective on institutional writing in digital environments. *Discourse, Context Media*, 35:100391.

Terrence W. Deacon. 1997. *The Symbolic Species: The Co-evolution of Language and the Brain*. W.W. Norton Company.

DeepSeek-AI. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Timothy Devinney. 2023. Plagiarism in the age of ai: Who owns the output? *Journal of Business Ethics*.

Nicholas Diakopoulos. 2016. Accountability in algorithmic decision making. *Communications of the ACM*, 59(2):56–62.

Emily Dinan, Laura Perez, and Jason Weston. 2022. Collaborative AI: Integrating language models into professional writing teams. In *Proceedings of NeurIPS 2022*, pages 5123–5136.

Daniel Dor. 2015. *The Instruction of Imagination: Language as a Social Communication Technology*. Oxford University Press.

Fernando Duarte, Maarten Sap, and Yejin Choi. 2022. AI-generated speech and the decline of authentic online discourse. *Proceedings of the 2022 Conference on Fairness, Accountability, and Transparency (FAccT)*.

Ronald Dworkin. 1996. *Freedom's Law: The Moral Reading of the American Constitution*. Harvard University Press.

Jacob Eisenstein and Danielle McNamara. 2023. Human-AI collaboration in writing: Rethinking authorship and editorial oversight. *AI Society*, 38(2):177–192.

Alexander Fabbri, Irene Li, Tianyi Tang, Caiming Xiong, and Dragomir Radev. 2022. QMSum: A new benchmark for query-based multi-document summarization. In *Proceedings of NAACL 2022*, pages 6145–6162.

Angela Fan, Shruti Bhosale, Holger Schwenk, Alexander Baevski, Guillaume Lample, and Michael Auli. 2021. Beyond English-centric multilingual machine translation. In *Proceedings of ACL 2021*, pages 4403–4419.

James Felton. 2023. From spell-checkers to AI writing assistants: The evolution of digital literacy tools. *Journal of Digital Communication*, 17(2):89–107.

Emilio Ferrara. 2020. Characterizing social media manipulation in the 2020 U.S. presidential election. *First Monday*, 25(11).

Luciano Floridi. 2013. *The Ethics of Information*. Oxford University Press.

Luciano Floridi. 2019. Establishing the rules for building trustworthy AI. *Nature Machine Intelligence*, 1(6):261–262.

Luciano Floridi and Marcello Chiriatti. 2020. GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30:681–694.

Michel Foucault. 1984. What is an author? In Paul Rabinow, editor, *The Foucault Reader*, pages 101–120. Pantheon Books.

Scott Freeman, Sarah L. Eddy, Miles McDonough, Michelle K. Smith, Nnadozie Okoroafor, Hannah Jordt, and Mary Pat Wenderoth. 2014. Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences (PNAS)*, 111(23):8410–8415.

Iason Gabriel. 2020. Artificial intelligence, values, and alignment. *Minds and Machines*, 30(4):411–437.

David Galbraith. 1999. Writing as a knowledge-constituting process. *Erkenntnis*, 50:357–370.

John Gallagher, Jacob Hilton, and Owain Evans. 2023. Adversarial robustness in AI-generated text detection: A losing battle? *arXiv preprint arXiv:2305.07692*.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander M. Rush. 2019. GLTR: Statistical detection and visualization of AI-generated text. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Clark Glymour, David Danks, and Peter Spirtes. 2023. AI explainability and its limits: The role of human oversight. *Artificial Intelligence Review*.

Shafi Goldwasser, Silvio Micali, and Chales Rackoff. 2019. The knowledge complexity of interactive proof-systems. In *Providing sound foundations for cryptography: On the work of shafi goldwasser and silvio micali*, pages 203–225.

H.P. Grice. 1975. Logic and conversation. In *Syntax and Semantics*, volume 3, pages 41–58. Academic Press.

Philipp Hacker, Andreas Engel, and Marco Mauer. 2023. Regulating chatGPT and other large generative AI models. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1112–1123.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Dirk Hovy and Shannon L. Spruit. 2016. The social impact of natural language processing. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 591–598.

Yang Hu, Heather R. Younger, and Patricia M. Greenfield. 2020. Language and hiring bias: How introverts and neurodiverse candidates face discrimination in spontaneous evaluations. *Journal of Business and Psychology*, 35(2):215–230.

Ming-Hui Huang and Roland T. Rust. 2021. Artificial intelligence in business: Promise, pitfalls, and prospects. *Journal of Service Research*, 24(1):3–6.

Ben Hutchinson, Jasmine Collins, Mark Diaz, and Qian Yang. 2023. AI for accessibility: Enabling inclusive digital communication. *CHI Conference on Human Factors in Computing Systems*.

Zhengbao Ji, Zhiting Hu, Patrick Lewis, and Meta AI. 2023. Survey of hallucination in large language models. *Transactions of the Association for Computational Linguistics*.

Hong Jin Kang, Fabrice Harel-Canada, Muhammad Ali Gulzar, Violet Peng, and Miryung Kim. 2024. Human-in-the-loop synthetic text data inspection with provenance tracking. *arXiv preprint arXiv:2404.18881*.

Neal Katyal and Daniel Epps. 2022. The criminal regulation of artificial intelligence. *Harvard Law Review*, 135:412–468.

Ronald T. Kellogg. 2008. Training writing skills: A cognitive developmental perspective. *Journal of Writing Research*, 1(1):1–26.

Salman Khan, Pranav Rajpurkar, and Daphne Koller. 2023. AI tutors: The role of large language models in personalized education. *arXiv preprint arXiv:2303.11288*.

Sahar Khowaja, Ammar Ahmad, Khaled Salah, Raja Jayaraman, and Ibrar Yaqoob. 2023. Blockchain for AI: Review and open research challenges. *IEEE Transactions on Artificial Intelligence*, 4(1):1–15.

Johannes Kirchenbauer, Jonas Geiping, Yuxuan Han, Elliot Creager, Ari S. Morcos, and Tom Goldstein. 2023. A watermark for large language models. *arXiv preprint arXiv:2307.06624*.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526.

Andrew J. Koch, Robert D. Gerber, and Sarah C. Roberts. 2015. Structured interviews: Reducing bias and increasing hiring effectiveness. *Journal of Applied Psychology*, 100(3):775–789.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Jeff Kosseff. 2019. *The Twenty-Six Words That Created the Internet*. Cornell University Press.

Ben Krause, Ethan Wilcox, Max Bittker, and Christopher D. Manning. 2022. Co-authoring with AI: How LLMs shape collaborative writing practices. *Transactions of the ACL*, 10:413–429.

Vivian Lai and Laura Viering. 2022. AI assistance and over-reliance: Implications for human judgment and decision-making. *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*.

George Lakoff and Mark Johnson. 1980. *Metaphors We Live By*. University of Chicago Press.

Jochen Laubrock, Clara Martin, and Marc Brysbaert. 2022. Readership enhancement via AI-assisted writing tools: A cognitive science perspective. *Cognitive Science*, 46(4):e13120.

Rebekah Layton and Carolyn Watters. 2020. Authorship attribution with deep learning. *Digital Scholarship in the Humanities*, 35(2):317–331.

Jisoo Lee, Daniel McNamara, and Tanja Käser. 2022. Co-authoring with ai: How language models support second-language writing development. *Computers Education*, 186:104536.

Brian Leonard and Kevin Noonan. 2020. Computational text generation in professional and technical writing. *Journal of Business and Technical Communication*, 34(4):451–474.

Julia Levashina, Christian J. Hartwell, Frederick P. Morgeson, and Michael A. Campion. 2014. The structured employment interview: Narrative and quantitative review of the research literature. *Personnel Psychology*, 67(1):241–293.

Matthew Lipman. 2003. *Thinking in Education*, 2nd edition. Cambridge University Press.

Rose Luckin, Wayne Holmes, and Joshua Greer. 2023. AI and education: The future of personalized learning. *AI Education Journal*, 4(1):112–130.

Brian Lund and Weilin Wang. 2023. Reinventing assessments in the age of ai: From written exams to oral evaluations. *AI Education Journal*, 2(1):22–35.

Richard Menary. 2010. Cognitive integration and the extended mind. In Richard Menary, editor, *The Extended Mind*, pages 227–243. MIT Press.

Jacob Menick, Jeffrey Shlens, Xiaohua Zhai, Neil Houlsby, Andrea Gesmundo, Avital Oliver, and Karen Simonyan. 2022. Reducing the recurrence of errors in language model outputs. *NeurIPS*.

Arthur I Miller. 2019. *The artist in the machine: The world of AI-powered creativity*. MIT Press.

Brent Daniel Mittelstadt and Luciano Floridi. 2016. The ethics of algorithms: Mapping the debate. *Big Data Society*, 3(2):1–21.

Tom Nichols. 2021. The death of expertise: The campaign against established knowledge and why it matters. *Oxford University Press*.

NLLB NLLBTeam. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Peter Norvig and Sebastian Thrun. 2009. *The Google Revolution: How AI is Transforming Language*. O'Reilly Media.

Takashi Ogawa, Hiroshi Nakagawa, and Yuki Tanaka. 2022. Breaking barriers: AI-assisted writing for non-native speakers and people with disabilities. *Computers Education*, 184:104516.

Cathy O'Neil. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown Publishing Group.

Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z. Berkay Celik, and Ananthram Swami. 2016. The limitations of deep learning in adversarial settings. In *Proceedings of the 1st IEEE European Symposium on Security and Privacy (EuroSP)*, pages 372–387.

Praveen Paritosh, Elizabeth Clark, Luke Zettlemoyer, and Mihai Surdeanu. 2022. Critiquing ai writing: Understanding the benefits and risks of language models in content creation. *arXiv preprint arXiv:2211.12760*.

Richard Paul and Linda Elder. 2007. *The Thinker's Guide to Socratic Questioning*. Foundation for Critical Thinking.

David N. Perkins and Gavriel Salomon. 1989. Are cognitive skills context-bound? *Educational Researcher*, 18(1):16–25.

Carlo Petrini. 2001. *Slow Food: The Case for Taste*. Columbia University Press.

Plato. 1997. *Complete Works*. Hackett Publishing. Includes translations of *Gorgias* and *Cratylus*.

Ali Rahimi, ChengXiang Zhai, and Rada Mihalcea. 2021. Collaborative AI writing: Balancing automation and human creativity. *AI Society*, 36(3):609–624.

Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2022. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of ACM FAccT 2022*, pages 343–357.

Kevin Roose and Margaret Sullivan. 2023. AI in journalism and corporate writing: A new era of assisted authorship. *Journalism Studies*, 24(6):815–832.

Sebastian Ruder. 2019. *Neural Transfer Learning for Natural Language Processing*. Ph.D. thesis, National University of Ireland, Galway.

Mohamed Sabt, Mohammed Achemlal, and Abdelmadjid Bouabdallah. 2015. Trusted execution environment: What it is, and what it is not. In *2015 IEEE Trustcom/BigDataSE/Ispa*, volume 1, pages 57–64. IEEE.

Timo Schick, Roberto Wilfer, and Hinrich Schütze. 2021. Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in NLP. *arXiv preprint arXiv:2103.00453*.

John R. Searle. 1969. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press.

Amanpreet Sharma, Thomas Wolf, and Sebastian Ruder. 2022. BigBird: Summarization for large-scale text processing. In *Proceedings of ACL 2022*, pages 5789–5801.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. 2024. AI models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Jasmine Wang, and Dario Amodei. 2019. Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203*.

Betsy Sparrow, Jenny Liu, and Daniel M. Wegner. 2011. Google effects on memory: Cognitive consequences of having information at our fingertips. *Science*, 333(6043):776–778.

Dan Sperber and Deirdre Wilson. 1986. *Relevance: Communication and Cognition*. Harvard University Press.

Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–556.

Stanford University. 2023. Curriculum Resources for AI in a Flipped classroom for Teachers (CRAFT). `https://craft.stanford.edu/`.

Robert J. Sternberg and Wendy M. Williams. 1997. *Intelligence, Instruction, and Assessment: Theory into Practice*. Lawrence Erlbaum Associates.

Christian Szegedy, Yi Tay, and Alexander Raichuk. 2022. Resilient AI: Aligning language models with ethical and epistemic standards. *Journal of AI Ethics*, 5:310–328.

Justin Thaler. 2023. State of the art report: Verified computation. *Preprint*, arXiv:2308.15191.

Michael Tomasello. 2003. *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Harvard University Press.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothee Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

European Union. 2023. The artificial intelligence act: Regulatory framework for AI Systems in the EU.

Eline van Dis, Anne-Sophie Bender, Marcel Bonn, and Iris de Bruin. 2023. ChatGPT: Five priorities for research. *Nature*, 614:224–226.

Frank van Tubergen and Matthijs Kalmijn. 2014. Language proficiency and early labor market entry of immigrants in the netherlands. *Journal of Ethnic and Migration Studies*, 40(3):405–424.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Lev S. Vygotsky. 1978. Mind in society: The development of higher psychological processes.

Johannes Wagner, Emily M. Bender, and Martin Savic. 2020. Accessible AI: Enabling writing for users with visual and motor impairments. *AI Society*, 35:301–319.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2022. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. *arXiv preprint arXiv:2211.00593*.

Kevin Warwick. 2003. Cyborg morals, cyborg values, cyborg ethics. *Ethics and Information Technology*, 5(3):131–137.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, and 1 others. 2021. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.

Benjamin Lee Whorf. 1956. *Language, Thought, and Reality: Selected Writings of Benjamin Lee Whorf*. MIT Press.

Malcolm Willis and Emma P. Williams. 2023. Should AI be a co-author? ethical and academic perspectives. *AI Society*.

Terry Winograd. 1972. Understanding natural language. *Cognitive Psychology*, 3(1):1–191.

Ludwig Wittgenstein. 1953. *Philosophical Investigations*. Blackwell.

Martha Woodmansee. 1994. The genius and the copyright: Economic and legal conditions of the emergence of the 'author'. In Martha Woodmansee and Peter Jaszi, editors, *The Construction of Authorship: Textual Appropriation in Law and Literature*, pages 1–20. Duke University Press.

Wei Xu, Yulia Tsvetkov, and Alan Black. 2022. AI for language learning: Conversational agents and personalized feedback. *Transactions of the Association for Computational Linguistics (TACL)*, 10:1–15.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Nan Du, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*.

Jason Yosinski, Dario Amodei, and Alec Radford. 2023. Co-editing with AI: The role of large language models in real-time group writing. In *Proceedings of ACL 2023*, pages 3145–3159.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32.

Xia Zhou, Yang Xu, and Feng Liu. 2023. Can AI-generated text be reliably detected? evaluating plagiarism detection tools on gpt-based texts. *arXiv preprint arXiv:2304.08979*.

Shoshana Zuboff. 2019. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. PublicAffairs.