

Opinion

The homogenizing effect of large language models on human expression and thought

Zhivar Sourati^{1,2,*}, Alireza S. Ziabari^{1,2}, and Morteza Dehghani ^{1,2,3}

Cognitive diversity, reflected in variations of language, perspective, and reasoning, is essential to creativity and collective intelligence. This diversity is rich and grounded in culture, history, and individual experience. Yet, as large language models (LLMs) become deeply embedded in people's lives, they risk standardizing language and reasoning. We synthesize evidence across linguistics, psychology, cognitive science, and computer science to show how LLMs reflect and reinforce dominant styles while marginalizing alternative voices and reasoning strategies. We examine how their design and widespread use contribute to this effect by mirroring patterns in their training data and amplifying convergence as all people increasingly rely on the same models across contexts. Unchecked, this homogenization risks flattening the cognitive landscapes that drive collective intelligence and adaptability.

When large language models meet human cognitive diversity

Cognitive diversity, intertwined and manifested through varied linguistic expressions, is essential to the adaptability, creativity, and overall effective functioning of complex societies [1]. While such variation might be expressed through stylistic differences, it reflects deeper cognitive [2,3] and sociocultural differences [4] and, vitally, plays a critical role in sustaining the **epistemic** (see [Glossary](#)) and problem-solving capacities of human groups [5]. This value of pluralism is rooted in the long-held principle that sound judgment requires exposure to varied thought. As John Stuart Mill argued, 'the only way in which a human being can make some approach to knowing the whole of a subject is by hearing what can be said about it by persons of every variety of opinion, and studying all modes in which it can be looked at by every character of mind. No wise man ever acquired his wisdom in any mode but this' [6]. When preserved, such distinctions support innovation, prevent **epistemic collapse**, and enhance the operational efficacy of collective systems [1,7].

This diversity has emerged organically from the coexistence of individuals with distinct backgrounds, linguistic repertoires, and value systems [2,8]. Yet, the increasing reach of global communication technologies, while enabling unprecedented knowledge sharing and connection, has also contributed to a gradual contraction of such linguistic and cognitive variations [9,10].ⁱ Among these technologies, large language models (LLMs) have emerged as especially influential ([Box 1](#) outlines a historical comparison with earlier technologies), becoming deeply integrated not only within digital infrastructure but also as fundamental components shaping how we interact with technology and with each other [14].

Extending beyond traditional language-based applications, such as summarization tools [15], LLMs are now involved in tasks once reserved for humans, such as sociocognitive modeling [16], psychological simulation [17], and even experimental substitution for human participants

Highlights

Cognitive diversity in language, perspective, and reasoning is not only vital to human creativity and problem-solving but also carries rich personal and cultural information.

Large language models (LLMs) are now widely used across writing, ideation, and decision-making tasks.

By mirroring dominant patterns, LLMs promote stylistic and conceptual homogenization while downplaying alternative voices.

With the growing prevalence of LLMs across fields and in shaping communication, studies have found that they are influencing and homogenizing people's language, perspectives, and reasoning.

This homogenization arises not only from biases in the training corpora but is also amplified through the training process itself, favoring patterns that are frequent and easily generalizable while smoothing over minority representations.

¹Department of Computer Science, University of Southern California, Los Angeles, CA, USA

²Center for Computational Language Sciences, University of Southern California, Los Angeles, CA, USA

³Department of Psychology, University of Southern California, Los Angeles, CA, USA

*Correspondence: souratih@usc.edu (Z. Sourati).

Box 1. Historical anxieties regarding the cognitive impact of technology

Anxieties about technology's influence on language and cognition have a long history, starting with Plato's Phaedrus and his concern that writing would weaken memory. Modern research echoes this, showing the Internet enables people to offload knowledge externally, which can inflate confidence in what they know [11]. While such externalization has been framed as potentially freeing mental resources for creativity and problem-solving [12], this benefit is less likely to manifest with the widespread adoption of LLMs.

Unlike prior technologies that primarily aided storage or retrieval, LLMs act as fluent coreasoners, and at times, standalone ones, participating in writing, problem-solving, and perspective-taking, thereby externalizing not only memory but also the articulation and justification of thought. As Clark and Chalmers [12] argue, 'Language, thus construed, is not a mirror of our inner states but a complement to them. It serves as a tool whose role is to extend cognition in ways that on-board devices cannot'. When that linguistic interface is mediated by systems capable of generating reasoning and perspective, much of human cognition risks being relocated outside the individual mind.

Hence, the effect of LLMs, though similar in nature to earlier cognitive extensions, differs profoundly in both function and scale. Earlier systems, such as schools, books, and search engines, while capable of promoting cultural uniformities, mainly disseminate knowledge or teach reasoning frameworks that individuals must internalize and apply. This internalization is an active, personalized process through which frameworks are adapted, integrated with pre-existing knowledge, and eventually transformed into unique cognitive strategies that are deployed as needed, often at later times. LLMs, by contrast, generate complete reasoning and articulation processes themselves on users' behalf. The same few models now generate text for hundreds of millions of users, with adoption rates surpassing those of any previous digital technology [132]. Their use has expanded especially rapidly in nonwork contexts [133] and remains concentrated around a small number of dominant systems, such as ChatGPT.⁴ This convergence, coupled with their linguistic fluency that can lead users to overtrust their responses [13], collapses the boundary between external tool and internal cognition, standardizing not only the information people access but also the very ways they articulate, justify, and reason through ideas.

[18,19], which expands the scope of what it means for LLMs to represent human diversity. This integration makes it critical to examine whether these models preserve human diversity or instead enforce a form of cognitive and linguistic homogenization. This also raises the question of when such homogenization might be beneficial or harmful. Excessive diversity can introduce costs related to coordination, communication, and coherence, and some degree of standardization may help mitigate these challenges [20]. Promoting a standard language can offer benefits, such as reducing bias against nonstandard dialects or preventing the misuse of linguistic cues [21]. Nonetheless, the risks inherent in standardization remain substantial: homogenized generations may constrain public discourse, reduce the visibility of marginal linguistic forms, and reinforce dominant reasoning templates. They may also suppress the kinds of idiosyncratic language use that signal individual traits or group-specific perspectives [22]. In complex reasoning tasks, the widespread adoption of **chain-of-thought (CoT) prompting** [23], optimized for linear and explicit inference, may disincentivize more abstract or intuitive reasoning styles that are harder to model yet crucial for flexible problem-solving [24,25]. These shifts raise broader concerns: that LLMs, if uncritically integrated, may shape not only how we write but also how we think.

LLMs thus present a paradox for cognitive science: as they become a powerful tool for modeling human thought [26], they also risk flattening the very cognitive diversity that cognitive science studies [27]. With growing concern across linguistics, computer science, psychology, and cognitive science more broadly about the homogenizing effects of LLMs, we bring these perspectives into conversation. We discuss diversity across three key dimensions: stylistic variation in language, perspective, and reasoning strategies (Figure 1). By bridging disciplinary boundaries, we aim to provide a more integrated understanding of how LLMs interact with human diversity, both reflecting and influencing variation in human reasoning and expression, and argue for greater incorporation of human-grounded diversity in these dimensions within the models.

Glossary

Chain-of-thought (CoT) prompting:

a prompting strategy that encourages models to show step-by-step reasoning before giving an answer, improving structure and accuracy but sometimes reducing intuitive or creative responses.

Epistemic: relating to knowledge, understanding, or how beliefs are formed and justified. It is used to describe cognitive or informational aspects of learning and reasoning.

Epistemic collapse: a situation where diversity in knowledge, reasoning, or interpretation is lost, leading to uniform ways of thinking and a narrower collective understanding.

Essentialized representations of

identity: simplified portrayals of social or cultural groups that treat identity as fixed and homogeneous, often ignoring variation and individual nuance.

Natural language processing (NLP):

a field of AI focused on enabling computers to understand, interpret, and produce human language across tasks such as translation, summarization, and question answering.

Prompt: a text input or instruction given to a language model that guides how it generates a response, effectively shaping the model's behavior and output.

Reinforcement learning from

human feedback (RLHF): a training method where human evaluators rate model outputs, guiding the model to produce more preferred or context-appropriate responses.

Supervised fine-tuning: a supervised training process that adapts a pretrained language model to a specific task or domain by continuing its training on a smaller, labeled dataset, teaching the model new knowledge or behaviors.

Temperature scaling: a parameter that controls how deterministic or random a model's outputs are. Lower temperatures yield precise, predictable text, while higher ones increase variety.

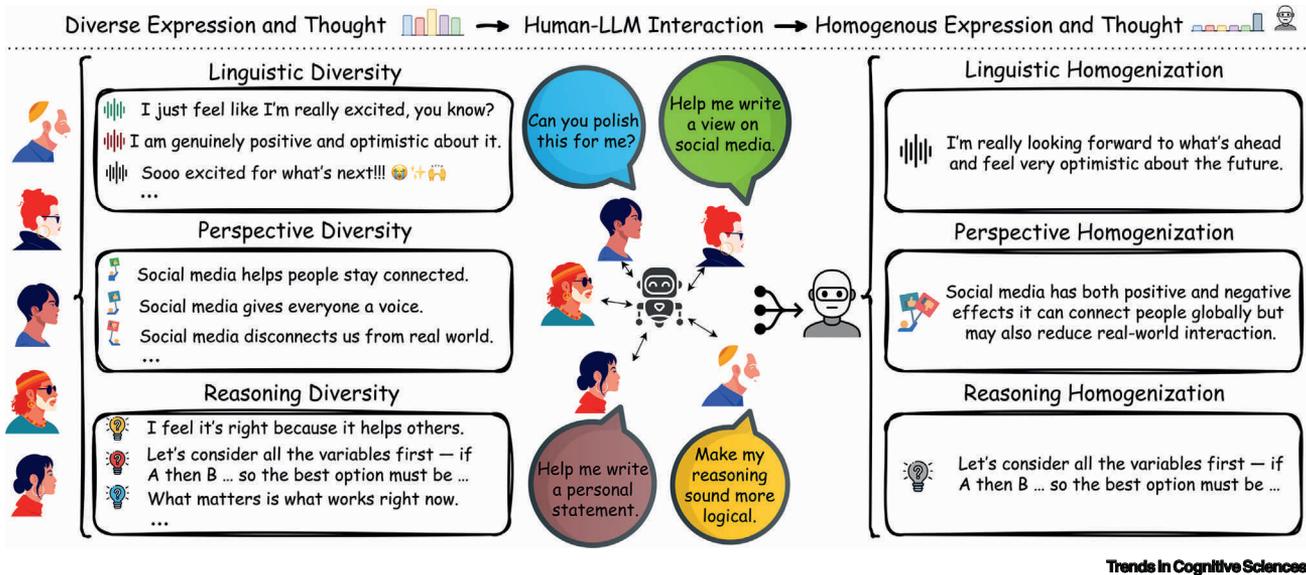


Figure 1. LLM-mediated homogenization of expression and thought. Individuals differ in how they write, reason, and view the world. When these differences are mediated by the same LLM, their distinct linguistic, perspectival, and reasoning signals become homogenized, producing standardized expressions and thoughts across users. LLM: large language model.

Language models, prediction, and the loss of diversity

Language modeling is central to modern artificial intelligence (AI) as it provides the foundation for systems, such as ChatGPT [28] and Gemini [29], that understand, generate, and interact through natural language [30]. At their core, these models operate by predicting the next token given a preceding context, a goal shared with earlier approaches such as n-gram models that estimated token probabilities using fixed-size word windows and explicit statistical counts from large corpora [31]. LLMs extend this approach at a much larger scale, trained on massive datasets with billions of parameters, supported by unprecedented levels of compute and architectural optimization [32]. Despite these advances, their fundamental objective remains the same: mastering the statistical regularities of language, which remains the dominant driver of model behavior [33] even after the application of LLM alignment methods such as **supervised fine-tuning** [34,35] and preference-based alignment tuning via **reinforcement learning from human feedback (RLHF)** [36,37].

Yet these advances come with inherent limitations. Because LLMs are trained to capture and reproduce the statistical regularities of their input data, which often overrepresent dominant languages and ideologies [38,39], their outputs tend to mirror a narrow and skewed slice of human experience. This limitation arises not only from biased training corpora but also gets amplified through the training process itself [40], favoring patterns that are frequent and easily generalizable while smoothing over minority representations [41]. At scale, what begins as statistical pattern learning, though capable of generalization beyond mere mimicry [42,43], becomes a generative force that privileges central tendencies while marginalizing rare expressions, alternative reasoning styles, and culturally specific voices. The consequence is not just a convergence in surface-level linguistic form but a narrowing of the conceptual space in which models write, speak, and reason [39].

Critically, this narrowing does not trend toward a neutral center but toward a historically uneven one, shaped by the norms, values, and perspectives of English-speaking, Global North, and

socioeconomically advantaged populations [44,45]. This has tangible consequences: when prompted for opinions or expressive writing, LLMs tend to reproduce mainstream, institutionally validated perspectives and writing styles that mirror those of Western, liberal, high-income, highly educated males, creating an illusion of consensus that frames these norms as the default standard of clarity or intelligence while muting alternative worldviews and culturally grounded forms of expression [22,44]. This misrepresentation persists even when models are explicitly prompted to assume a specific identity, often resulting in LLM personas that reflect out-group stereotypes rather than authentic in-group representations [46]. For instance, when asked for a person with impaired vision's thoughts on immigration, one model responded: 'While I may not be able to visually observe the nuances of the US-Mexican border or read statistics, I believe...' [46]. This misportrayal is a symptom of a larger problem where LLMs flatten demographic groups by neglecting heterogeneity and, through identity prompting, reduce identities to fixed, **essentialized representations of identity**. Consequently, what is produced is not a prototype of the group's varied experience, but a mere caricature.

This representational imbalance is not static; it becomes an active, homogenizing force through a recursive feedback loop [47]. As individuals increasingly turn to LLMs for writing, problem-solving, and conceptual exploration, the models' outputs, already favoring common linguistic and conceptual patterns, are reabsorbed into human discourse and begin to shape the users' own expression and reasoning [48] and, in turn, influence the data used to train future models, transforming homogenization from a passive bias into a structurally reinforced influence.

In the following subsections, we examine the homogenizing effects of LLMs and how this recursive influence manifests across three interrelated domains: language, perspective, and reasoning.

Language diversity

Across time and geography, human groups have developed distinct linguistic systems and cultural norms [49,50], shaped by environmental conditions and social structures [4,51]. These dynamics are encoded and preserved through language—rich symbolic systems that carry not only the content of communication but also people's underlying values and identities [52,53]. For example, people categorized as extroverts tend to use more words related to humans, social processes, and family in personal essays about past experiences, reflecting a greater focus on interpersonal connection and social engagement [54]. Language, in this sense, serves as a powerful medium for expressing cross-language and within-language individual and group differences [53,55], not just through semantic content but also through subtle stylistic [55] and structural cues [56].

Efforts to enhance linguistic diversity have long been central to **natural language processing (NLP)**, preceding the emergence of LLMs. Work in this area primarily focused on improving diversity in language generation, aiming to make machine-produced text more informative, engaging, and natural across tasks, such as summarization [57], translation [58], and more broadly, dialog generation [59].

Concurrently, fields such as computational sociolinguistics and authorship profiling have focused on exploring how language reflects speakers' underlying backgrounds and introduced various methods to identify linguistic signatures of social position and individual traits [55,60]. For instance, researchers have analyzed congressional speeches to predict

speakers' age and gender from linguistic cues in their public addresses [61]. Yet, the primary objective of these studies was analytical rather than generative as, until recently, most NLP research focused on producing engaging conversations with surface-level linguistic diversity.

However, as LLMs have mastered surface-level fluency and natural linguistic variation [62], a central question emerges: Do they exhibit humanlike patterns of linguistic variation, i.e., do they reflect and preserve socially meaningful forms of variation and preserve links to speaker traits? Existing evidence suggests they do not. Recent research shows that when LLMs are used to polish various forms of writing, ranging from Reddit posts and news articles to academic abstracts and personal essays, the resulting texts converge in writing complexity, diminishing the predictability of author characteristics, such as political affiliation, personality, gender, or age, and even weakening well-established associations such as the link between 'big-word' usage (words with complex structures of seven letters or longer that may indicate intellectual complexity or formality) and the author's openness to experience [22]. Likewise, studies generating college admission essays entirely with LLMs find that the resulting texts exhibit high semantic and lexical similarity across samples, reflecting a narrowing of expressive space [63]. Importantly, this homogenization persists despite interventions such as **temperature scaling** or different prompting techniques that simulate different author identities or personas, which are commonly used to induce stylistic diversity [62,63,64].

This bias appears to worsen as models are increasingly trained on synthetic, model-generated data [65] and is further amplified by the adoption of reinforcement learning (RL), particularly RLHF, to enhance perceived qualities, such as helpfulness [36] and reasoning capabilities [66]. While these RL-based techniques have substantially improved reasoning compared with fine-tuned models [67], they have also been shown to reduce stylistic and expressive variability [68]. This effect is reinforced by the quality–novelty trade-off, where methods that promote novelty often compromise coherence [69], suggesting that continued optimization for quality and performance may further diminish linguistic and stylistic diversity.

To address these issues, researchers have proposed several strategies. Prompting-based methods, applied at inference time, aim to enhance output diversity by modifying how prompts are phrased or conditioned [70,71]. For instance, studies using persona-based prompts show that coarse-grained persona conditioning, combined with adjustments to output length, can maximize lexical variation in model responses [71]. Training-based approaches instead modify the LLMs' learning process to optimize for semantic and stylistic diversity [72,73]. For example, researchers have fine-tuned models on datasets with multiple valid responses per prompt [72] or introduced diversity-aware weighting in preference optimization to reward rare, high-quality generations [73]. Although promising, evidence of homogenization in sociolinguistic contexts calls for closer evaluation of whether these strategies foster genuine, context-grounded diversity or merely superficial variation.

Vitaly, the concerns regarding the linguistic diversity of LLMs extend beyond representation alone as these models actively shape how language is used and evolves. As they become embedded in everyday writing (e.g., composing emails), while supporting more efficient and polished communication [74], they also tend to promote uniform styles that mask authentic voices and reduce variation in tone, culture, and identity [44,75], even for those merely engaging with AI-generated text [76].

At first glance, the homogenization of language, along with the loss of lexical cues linking writing to individual identity, may appear beneficial for protecting privacy as it reduces risks of misuse such as surveillance or discrimination [21]. However, it is equally important to recognize that this process also erases valuable linguistic markers long used across sociolinguistics, psychology, and mental health research. For example, people at risk of Alzheimer's disease exhibit early linguistic signs such as telegraphic speech (simplified phrases lacking grammar and function words, often missing determiners such as 'the' or 'a', auxiliaries such as 'is' or 'are', and even entire subjects) as well as repetitiveness and misspellings—patterns reflecting a decline in grammatical structure and language complexity [77]. If such distinctive markers are standardized by LLM-assisted polishing, critical early indicators for diagnosis and intervention may be lost. Moreover, the writing styles that LLMs incorporate are not neutral but a source of bias themselves, reproducing dominant expressive norms while eroding minority and underrepresented voices [44,78].

Perspectival diversity

Individuals vary not only in how they express themselves but also in their perspectives, beliefs, and values [79]. In language, these differences are reflected in patterns of how people take stances [80] or frame issues [81], which have been widely studied in computational analyses of text [82]. For instance, analyses of US Senate speeches reveal that Democrats and Republicans frame the issue of abortion through distinct moral dimensions, with Democrats emphasizing fairness and rights and Republicans focusing more on purity and sexual morality, mirroring broader ideological divides in moral reasoning [81].

With the rise of LLMs, this domain of variation has become especially salient as these models are increasingly deployed in contexts involving perspective gathering and open-ended writing tasks [83]. Recent work shows they can simulate public opinion, for example, on global warming by generating synthetic survey responses conditioned on demographics and personal concern, capturing belief patterns that mirror human data [84]. Hence, a pressing question arises: Do they preserve the pluralism of human perspectives, or do they default to a narrow, socially normative median?

Multiple studies have shown that LLMs tend to reflect characteristics of Western, educated, industrialized, rich, and democratic (WEIRD) societies in their perspectives [85–87]. Using the World Values Survey [88] and the Moral Foundations Questionnaire-2 (MFQ-2 [50]), researchers tested GPT-3.5 across 1000 runs with a temperature of 1.0 to maximize output variance and compared the results with human responses from diverse cultural backgrounds, finding that LLM outputs not only exhibited substantially less variance than human data but were also more closely aligned with response patterns characteristic of WEIRD societies, showing limited variability and weak representation of non-WEIRD perspectives [85,86]. While models can simulate diverse viewpoints when explicitly prompted, such as by adjusting generation parameters, incorporating identity-coded instructions [89], or translating prompts to a target language [90], these interventions have often fallen short of aligning outputs with the actual distribution of perspectives within the referenced groups [87,91], capturing merely the socially 'correct' [92] or the mean of the distribution [85] and, in some cases, even reproducing out-group stereotypes or misrepresentations of the specified populations [46].

Similar to efforts aimed at enhancing lexical diversity in LLM generations, recent research has also sought to increase perspective diversity through both prompting and fine-tuning approaches [93,94]. For instance, researchers have introduced a debate-based framework

in which multiple models interact to produce arguments that are broader in scope and more balanced across viewpoints [94]. Furthermore, through fine-tuning, RL frameworks have been adapted to mitigate ‘preference collapse’ by adjusting reward functions to account for underrepresented or minority preferences, balancing performance with diversity [95]. Yet it remains unclear whether these strategies truly embody human-aligned pluralism of perspectives and contextual depth, highlighting the need for more systematic evaluation.

This power to simulate and frame perspectives does not remain contained within the model; as LLMs become integrated into daily lives, they begin to influence how individuals perceive and frame the world, narrowing perspectives that are naturally rooted in lived experiences [96] and environmental and social contexts [97,98]. Studies show that when people remember events together, they tend to align their memories with others in the group, reinforcing shared details while forgetting those left unmentioned [99]. LLMs, however, can amplify this dynamic on a global scale: by exposing millions of users to the same suggestions and perspectives, they foster similar patterns of recall and association, promoting convergence in what people collectively remember and express.

One setting where this influence is particularly consequential is cowriting with LLM-based assistants as people increasingly rely on them even for open-ended survey responses about their beliefs and behaviors [100]. This, in turn, shapes how users frame and articulate their own perspectives [48,101]. This effect is evident in studies showing that participants who cowrote with opinionated language models, engineered to frame social media positively or negatively, tended to mirror the model’s stance in their writing and even shifted their own opinions in subsequent attitude surveys [48]. These findings raise broader concerns about persuasive influence as even subtle interaction with LLMs can lead users to adopt the model’s framing without awareness [102].

Reasoning diversity

Beyond language and perspective, another consequential form of diversity at stake is reasoning diversity—the varied ways people reason, solve problems, and generate new ideas—which arises from cultural and individual differences. For example, Native American (Menominee) children often group animals based on ecological relationships, such as shared habitat or interdependence, whereas European-American children tend to rely on taxonomic categories, reflecting broader differences in how people make sense of the world [103,104]. Across disciplines, variation in reasoning is recognized as a key driver of collective strength [1,98]. Groups composed of individuals who reason differently, using distinct heuristics and problem representations, consistently outperform groups made up of the highest-ability individuals [5], and innovation and intellectual breakthroughs often emerge from the recombination of ideas across domains and engagement with unfamiliar concepts [105,106].

As LLMs are increasingly deployed in reasoning-intensive settings [107,108], preserving reasoning diversity becomes crucial. If these systems reflect only a narrow slice of human thought, they risk reinforcing dominant cognitive styles while marginalizing others. While LLMs offer high performance and access to expertise, collective convergence on even optimal algorithms can reduce decision-making quality in complex systems [109]. This concern is not abstract: a global ‘weirdization’ may already be underway as once-local ways of conceptualizing time, space, and causality give way to homogenized, Western-aligned models. Trained on massive, biased corpora and built upon similar datasets and architectures, LLMs may further accelerate this shift, risking the homogenization of reasoning processes and decision outcomes across contexts [110,111].

Various studies have compared the reasoning and cognitive abilities of LLMs with those of humans across tasks traditionally used to assess human cognition. While these models often align with human reasoning in their outcomes, they have also revealed important discrepancies [19,112,113]. LLMs tend to produce reasoning patterns that cluster around central tendencies, lacking the natural variability that characterizes human thought. For example, in tasks inspired by the wisdom of the crowds phenomenon, where the diversity of individual judgments allows groups to collectively approximate correct answers, LLMs instead converge on uniform, ‘idealized’ responses, missing the variance that makes human reasoning adaptive and collectively effective [19].

This mismatch may stem from the very objectives used to train and evaluate LLMs, which emphasize measurable performance gains or compliance with verifiable behavioral metrics, such as accuracy, informativeness, helpfulness, harmlessness, or formatting consistency [36,66], and prioritize improving reasoning performance over assessing variation in reasoning approaches. Even cognitively inspired methods, such as analogical reasoning in LLMs [114], tree-of-thought prompting [115], and memory-based inference [116], are typically optimized for correctness and general utility. The widespread success of CoT prompting [23] has further reinforced this narrow focus, driving homogenization across multimodal reasoning [117,118]. This emphasis on performance and the overreliance on techniques that maximize it raise concerns about the adaptability, generalizability, and cognitive diversity of a uniform reasoning paradigm [25]. This overreliance and homogenization of reasoning already show documented downsides [24]. For instance, in a vehicle classification task where most items followed a simple rule but a few violated it, CoT prompting made GPT-4o four times slower to learn correct labels as step-by-step reasoning overgeneralized from regular patterns and overlooked exceptions and contextual cues [119].

This imperative to maintain diversity extends beyond how LLMs reflect human reasoning to how they influence it. For instance, in controlled experiments designed to assess creative ideation, participants who received assistance from LLMs (i.e., ChatGPT) generated a greater number of more detailed and elaborated ideas, particularly benefiting those who were less experienced or less creative writers. However, their outputs were also judged to be more semantically similar across participants [120,121]. This convergence becomes even more concerning when considering how users interact with these systems: rather than actively steering generation, users often defer to model-suggested continuations, selecting options that seem ‘good enough’ instead of crafting their own, which gradually shifts agency from the user to the model [122]. This tendency is particularly alarming for younger users as LLMs’ negative effects on creativity are most pronounced when used early in the ideation process [123]. Moreover, the consequences extend beyond creative output to the underlying cognitive processes: neurocognitive evidence shows that LLM-assisted writing elicits the weakest overall neural coupling, with reduced engagement of alpha and beta networks, lower memory recall, and diminished ownership of written work compared not only with independent writing but also with writing supported by search engines [124].

Concluding remarks

Trained on vast and often biased corpora—biases that may be further reinforced through iterative retraining—LLMs are now deeply embedded in human language, intertwined with identity and cognition. While these systems can function as powerful assistants at the individual level, supporting communication, creativity, and access to information, their widespread and repeated use raises distinct collective-level risks. As billions of users interact with increasingly similar models, LLMs may gradually homogenize human expression and thought, with concerning

Outstanding questions

Will current alignment methods, such as supervised fine-tuning and reinforcement learning from human feedback, ever be sufficient to reproduce the full diversity of human cognition, or are more foundational changes in model architecture, objectives, and training data required? These approaches have increased model steerability and surface-level variation, but it remains unclear whether they can capture deeper, context-sensitive, and culturally grounded forms of diversity found in human thought.

Even if large language models (LLMs) produce more diverse outputs, how can we ensure that this diversity is meaningful and grounded in actual human experience, rather than being superficial or artificially constructed? Future research should identify and promote metrics and frameworks that distinguish synthetic variation from diversity reflecting authentic sociocultural, emotional, and cognitive nuance.

What are the long-term cognitive effects of sustained reliance on LLMs for ideation, writing, and reasoning? While short-term effects, such as reduced stylistic variation and creative ownership, have been observed, we lack longitudinal studies that track changes in abstraction, memory retention, and reasoning strategies over time and examine whether these changes are irreversible.

Can users be equipped with strategies to counteract the homogenizing effects of LLMs on expression and thought? Research is needed to develop and evaluate behavioral or interface-level interventions, such as delaying LLM use during ideation or exposing users to model-induced changes, which help preserve agency and individuality in interaction.

What taxonomies or repertoires of intervention can future research establish to help mitigate LLM-driven homogenization at scale? A systematic understanding of the possible safeguards—behavioral, architectural, or institutional—is needed to guide users, developers, and platforms toward practices that promote cognitive and linguistic pluralism.

effects in distinct domains. These concerns are already reflected in public attitudes—over half of US adults believe that AI will make people less capable of thinking creatively or forming meaningful connectionsⁱⁱⁱ—and are further underscored by the empirical evidence presented throughout this study, including reduced stylistic and lexical diversity in generated texts, subtle recalibration of user attitudes and framing in AI-mediated communication, and diminished diversity in ideation.

The concern is not just that LLMs shape how people write or speak, but that they subtly redefine what counts as credible speech, correct perspective, or even good reasoning by making certain framings more salient. As sociologist George Ritzer's 'McDonaldization' theory suggests [125], processes favoring efficiency, predictability, and control can suppress contextual richness. LLMs mirror this logic in cognition, offering fluency and consistency while displacing situated, idiosyncratic forms of thought.

The centralized control over the very algorithms and datasets driving this homogenization is also acutely political, amplified by the dominance of a few platforms owned by multibillion-dollar corporations with significant political influence.ⁱⁱ In a time of rising global populism, this concentrated power enables top-down homogenization, where perspectives can be subtly engineered for concentrated benefit and power. The documented censorship, such as the Chinese Qwen model's refusal to answer politically sensitive questions,^{iv} demonstrates this systemic risk. This algorithmic and market-driven erasure of diverse thought poses a modern danger akin to the linguistic control of Newspeak in Orwell's *Nineteen Eighty-Four* [126].

A growing number of strategies aim to counteract this homogenization, including personalized modeling approaches [127,128], embodied reasoning frameworks that promote situated and context-sensitive inference [129], and diversified prompting or multiagent debate systems designed to broaden reasoning [94,130]. Yet evidence from sociolinguistics and psychology continues to document persistent homogenization, suggesting that these solutions must be applied more comprehensively and systematically before their effectiveness can be meaningfully evaluated. More broadly, both prompting- and training-based diversification methods remain constrained by underlying pretraining representations, potentially limiting their ability to induce deeper variation [33,35,37], and may also introduce trade-offs as variation too far from pretraining distributions can increase the likelihood of hallucination [131].

In summary, while LLMs offer significant advancements and conveniences, their broad adoption without critical evaluation risks fundamentally altering the diverse cognitive landscapes that enrich human interaction and drive innovation. This homogenization constitutes a profound challenge to the ideal of collective wisdom—the fundamental principle, articulated by Mill, that true knowledge is acquired only by 'hearing what can be said about it by persons of every variety of opinion'. We do not advocate for diversity in language, perspective, or reasoning without recognizing the coordination costs it can entail. Rather, we call for a deeper understanding of how LLMs affect the diversity of language and thought, as well as the signals embedded in this diversity, and for these insights to inform their design. Moving forward, preserving and enhancing meaningful human diversity should be a central criterion in the development and evaluation of LLMs. Only through deliberate attention to this pluralism can we harness the full potential of language technologies without sacrificing the very diversity that defines human society. Nevertheless, there is still much to learn about the LLM-driven homogenization of language and thought and how best to address it. Some key directions for deepening this understanding are outlined (see [Outstanding questions](#)).

Acknowledgments

This research was supported by the Air Force Office of Scientific Research (AFOSR)—A9550-23-1-0463. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the AFOSR or the US Government. The US Government is authorized to reproduce and distribute reprints for governmental purposes, notwithstanding any copyright annotation therein.

Declaration of interests

The authors declare no competing interests.

Resources

ⁱ<https://aeon.co/ideas/what-happens-to-cognitive-diversity-when-everyone-is-more-weird>

ⁱⁱ<https://techcrunch.com/2025/04/01/chatgpt-isnt-the-only-chatbot-thats-gaining-users>

ⁱⁱⁱ<https://www.pewresearch.org/science/2025/09/17/how-americans-view-ai-and-its-impact-on-people-and-society/>

^{iv}<https://huggingface.co/blog/leonardlin/chinese-llm-censorship-analysis>

References

- Page, S.E. (2008) *The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies*, Princeton University Press
- Evans, N. and Levinson, S.C. (2009) The myth of language universals: language diversity and its importance for cognitive science. *Behav. Brain Sci.* 32, 429–448
- Majidi, A. et al. (2004) Can language restructure cognition? The case for space. *Trends Cogn. Sci.* 8, 108–114
- Eckert, P. (2012) Three waves of variation study: the emergence of meaning in the study of sociolinguistic variation. *Annu. Rev. Anthropol.* 41, 87–100
- Hong, L. and Page, S.E. (2004) Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proc. Natl. Acad. Sci. U. S. A.* 101, 16385–16389
- Mill, J.S. (2001) *On Liberty*, Batoche Books
- Solomon, M. (2006) Groupthink versus the wisdom of crowds: the social epistemology of deliberation and dissent. *South. J. Philos.* 44, 28–42
- Labov, W. (1972) The social setting of linguistic change. In *Sociolinguistic Patterns*, pp. 260–325, University of Pennsylvania Press
- Harmon, D. and Loh, J. (2010) The index of linguistic diversity: a new quantitative measure of trends in the status of the world's languages. *Language Doc. Conserv.* 4, 97–151
- Alsaleh, A. (2024) The impact of technological advancement on culture and society. *Sci. Rep.* 14, 32140
- Fisher, M. et al. (2015) Searching for explanations: how the Internet inflates estimates of internal knowledge. *J. Exp. Psychol. Gen.* 144, 674–687
- Clark, A. and Chalmers, D. (1998) The extended mind. *Analysis* 58, 7–19
- Steyvers, M. et al. (2025) What large language models know and what people think they know. *Nat. Mach. Intell.* 7, 221–231
- Handa, K. et al. (2025) Which economic tasks are performed with AI? Evidence from millions of Claude conversations. *arXiv* <https://doi.org/10.48550/arXiv.2503.04761>
- Zhang, T. et al. (2024) Benchmarking large language models for news summarization. *Trans. Assoc. Comput. Linguist.* 12, 39–57
- Kosinski, M. (2024) Evaluating large language models in theory of mind tasks. *Proc. Natl. Acad. Sci. U. S. A.* 121, e2405460121
- Park, J.S. et al. (2023) Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pp. 1–22, Association for Computing Machinery
- Dillion, D. et al. (2023) Can AI language models replace human participants? *Trends Cogn. Sci.* 27, 597–600
- Aher, G.V. et al. (2023) Using large language models to simulate multiple humans and replicate human subject studies. In *International Conference on Machine Learning*, pp. 337–371, PMLR
- Wang, G. et al. (2022) The inverted U-shaped relationship between knowledge diversity of researchers and societal impact. *Sci. Rep.* 12, 18585
- Tufekci, Z. (2014) Engineering the public: big data, surveillance and computational politics. *First Monday* 19. <https://doi.org/10.5210/firstmonly.1917.4901>
- Sourati, Z. et al. (2025) The shrinking landscape of linguistic diversity in the age of large language models. *arXiv* <https://doi.org/10.48550/arXiv.2502.11266>
- Wei, J. et al. (2022) Chain-of-thought prompting elicits reasoning in large language models. *Adv. Neural Inf. Process. Syst.* 35, 24824–24837
- Sui, Y. et al. (2025) Stop overthinking: a survey on efficient reasoning for large language models. *Trans. Mach. Learn. Res.* Published online August 08, 2025. <https://dblp.org/rec/journals/tmlr/SuiCWZZYLWZZCH25.html>
- Ziabari, A.S. et al. (2025) Reasoning on a spectrum: aligning LLMs to system 1 and system 2 thinking. *arXiv* <https://doi.org/10.48550/arXiv.2502.12470>
- Strachan, J.W. et al. (2024) Testing theory of mind in large language models and humans. *Nat. Hum. Behav.* 8, 1285–1295
- Boogert, N.J. et al. (2018) Measuring and understanding individual differences in cognition. *Philos. Trans. R. Soc. B Biol. Sci.* 373, 20170280
- Achiam, J. et al. (2023) Gpt-4 technical report. *arXiv* <https://doi.org/10.48550/arXiv.2303.08774>
- Team, G. et al. (2023) Gemini: a family of highly capable multi-modal models. *arXiv* <https://doi.org/10.48550/arXiv.2312.11805>
- Zhao, W.X. et al. (2023) A survey of large language models. *arXiv* <https://doi.org/10.48550/arXiv.2303.18223>
- Mitchell, T.M. (1999) Machine learning and data mining. *Commun. ACM* 42, 30–36
- Grattafiori, A. et al. (2024) The Llama 3 herd of models. *arXiv* <https://doi.org/10.48550/arXiv.2407.21783>
- Lin, B.Y. et al. (2024) The Unlocking Spell on Base LLMs: Rethinking Alignment via In-Context Learning. In *The Twelfth International Conference on Learning Representations*
- Wei, J. et al. (2022) Finetuned Language Models are Zero-Shot Learners. In *International Conference on Learning Representations*
- Ghosh, S. et al. (2024) A Closer Look at the Limitations of Instruction Tuning. In *International Conference on Machine Learning*, pp. 15559–15589, PMLR
- Bai, Y. et al. (2022) Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv* <https://doi.org/10.48550/arXiv.2204.05862>
- Wolf, Y. et al. (2024) Fundamental limitations of alignment in large language models. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 53079–53112, JMLR.org
- Buyl, M. et al. (2024) Large language models reflect the ideology of their creators. *arXiv* <https://doi.org/10.48550/arXiv.2410.18417>

39. Schut, L. *et al.* (2025) *Do Multilingual LLMs Think In English?*. ICLR 2025 Workshop on Building Trust in Language Models and Applications
40. Wang, Z. *et al.* (2024) Bias amplification: large language models as increasingly biased media. *arXiv* <https://doi.org/10.48550/arXiv.2410.15234>
41. Hall, M. *et al.* (2022) A systematic study of bias amplification. In *Workshop on Trustworthy and Socially Responsible. Machine Learning at NeurIPS*
42. McCoy, R.T. *et al.* (2023) How much do language models copy from their training data? Evaluating linguistic novelty in text generation using RAVEN. *Trans. Assoc. Comput. Linguist.* 11, 652–670
43. Misra, K. and Mahowald, K. (2024) Language Models Learn Rare Phenomena from Less Rare Phenomena: The Case of the Missing AANNs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 913–929, Association for Computational Linguistics
44. Alvero, A. *et al.* (2024) Large language models, social demography, and hegemony: comparing authorship in human and synthetic text. *J. Big Data* 11, 138
45. Hartmann, J. *et al.* (2023) The political ideology of conversational AI: converging evidence on ChatGPT's pro-environmental, left-libertarian orientation. *arXiv* <https://doi.org/10.48550/arXiv.2301.01768>
46. Wang, A. *et al.* (2025) Large language models that replace human participants can harmfully misportray and flatten identity groups. *Nat. Mach. Intell.* 7, 400–411
47. Wylie, S. *et al.* (2024) Fairness feedback loops: training on synthetic data amplifies bias. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 2113–2147, Association for Computing Machinery
48. Jakesch, M. *et al.* (2023) Co-writing with opinionated language models affects users' views. In *Proceedings of the 2023 CHI conference on human factors in computing systems*, pp. 1–15, Association for Computing Machinery
49. Nichols, J. (1992) *Linguistic Diversity in Space and Time*, University of Chicago Press
50. Atari, M. *et al.* (2023) Morality beyond the WEIRD: how the nomological network of morality varies across cultures. *J. Pers. Soc. Psychol.* 125, 1157
51. Nölle, J. *et al.* (2020) Language as shaped by the environment: linguistic construal in a collaborative spatial task. *Palgrave Commun.* 6, 1–10
52. Bamman, D. *et al.* (2014) Gender identity and lexical variation in social media. *J. Socioling.* 18, 135–160
53. Boyd, R. *et al.* (2015) Values in words: Using language to evaluate and understand personal values. In *Proceedings of the International AAAI Conference on Web and Social Media*, pp. 31–40
54. Hirsh, J.B. and Peterson, J.B. (2009) Personality and language use in self-narratives. *J. Res. Pers.* 43, 524–527
55. Pennebaker, J.W. (2011) The secret life of pronouns. *New Sci.* 211, 42–45
56. Boghrati, R. *et al.* (2018) Conversation level syntax similarity metric. *Behav. Res. Methods* 50, 1055–1073
57. Li, L. *et al.* (2009) Enhancing diversity, coverage and balance for summarization through structure learning. In *Proceedings of the 18th international conference on World wide web*, pp. 71–80, Association for Computing Machinery
58. Gimpel, K. *et al.* (2013) A systematic exploration of diversity in machine translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1100–1111, Association for Computational Linguistics
59. Li, J. *et al.* (2016) A Diversity-Promoting Objective Function for Neural Conversation Models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 110–119, Association for Computational Linguistics
60. Nguyen, D. *et al.* (2016) Computational sociolinguistics: a survey. *Comput. Linguist.* 42, 537–593
61. Zlabari, A.S. *et al.* (2024) Reinforced multiple instance selection for speaker attribute prediction. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3307–3321, Association for Computational Linguistics
62. Guo, Y. *et al.* (2025) Benchmarking linguistic diversity of large language models. *Trans. Assoc. Comput. Linguist.* 13, 1507–1526
63. Lee, J. *et al.* (2025) Poor Alignment and Steerability of Large Language Models: Evidence Using 30,000 College Admissions Essays. In *First Workshop on Social Simulation with LLMs*
64. Martínez, G. *et al.* (2025) Beware of words: evaluating the lexical diversity of conversational LLMs using ChatGPT as case study. *ACM Trans. Intell. Syst. Technol.* 16, 138
65. Guo, Y. *et al.* (2024) The Curious Decline of Linguistic Diversity: Training Language Models on Synthetic Text. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 3589–3604, Association for Computational Linguistics
66. Guo, D. *et al.* (2025) DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature* 645, 633–638
67. Chu, T. *et al.* (2025) SFT Memorizes, RL Generalizes: A Comparative Study of Foundation Model Post-training. In *International Conference on Machine Learning*, pp. 10818–10838, PMLR
68. Kirk, R. *et al.* (2024) Understanding the Effects of RLHF on LLM Generalisation and Diversity. In *Proceedings of the Twelfth International Conference on Learning Representations (ICLR)*
69. Zhang, H. *et al.* (2021) Trading Off Diversity and Quality in Natural Language Generation. In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pp. 25–33
70. Chu, K. *et al.* (2025) Exploring and Controlling Diversity in LLM-Agent Conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 25626–25644, Association for Computational Linguistics
71. Kambhatla, G. *et al.* (2025) Measuring Lexical Diversity of Synthetic Data Generated through Fine-Grained Persona Prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 21024–21033, Association for Computational Linguistics
72. Mai, L. and Carson-Berndsen, J. (2025) Improving Linguistic Diversity of Large Language Models with Possibility Exploration Fine-Tuning. In *Interspeech 2025*, pp. 3254–3258
73. Chung, J.J.Y. *et al.* (2025) Modifying Large Language Model Post-Training for Diverse Creative Writing. In *Proceedings of the Conference on Language Modeling (COLM 2025)*
74. Hohenstein, J. *et al.* (2023) Artificial intelligence in communication impacts language and social relationships. *Sci. Rep.* 13, 5487
75. Holliday, N.R. and Reed, P.E. (2025) Gender and racial bias issues in a commercial "tone of voice" analysis system. *PLoS One* 20, e0314470
76. Toma, C.L. (2014) Towards conceptual convergence: an examination of interpersonal adaptation. *Commun. Q.* 62, 155–178
77. Eyigöz, E. *et al.* (2020) Linguistic markers predict onset of Alzheimer's disease. *eClinicalMedicine* 28, 100583
78. Huang, T.J. (2024) Translating authentic selves into authentic applications: private college consulting and selective college admissions. *Sociol. Educ.* 97, 174–192
79. Aglozo, E.Y. *et al.* (2025) A cross-national analysis of demographic variation in belief in God, gods, or spiritual forces in 22 countries. *Sci. Rep.* 15, 13302
80. Küçük, D. and Can, F. (2020) Stance detection: a survey. *ACM Comput. Surv.* 53, 1–37
81. Sagi, E. and Dehghani, M. (2014) Measuring moral rhetoric in text. *Soc. Sci. Comput. Rev.* 32, 132–144
82. Kennedy, B. (2022) Text analysis for psychology: methods, principles, and practices. In *Handbook of Language Analysis in Psychology* (Dehghani, M. *et al.*, eds), pp. 3–62, The Guilford Press
83. Argyle, L.P. *et al.* (2023) Out of one, many: using language models to simulate human samples. *Polit. Anal.* 31, 337–351
84. Lee, S. *et al.* (2024) Can large language models estimate public opinion about global warming? An empirical assessment of algorithmic fidelity and bias. *PLoS Clim.* 3, e0000429
85. Abdurahman, S. *et al.* (2024) Perils and opportunities in using large language models in psychological research. *PNAS Nexus* 3, pgae245
86. Atari, M. *et al.* (2023) Which humans? *PsyArXiv* <https://doi.org/10.31234/osf.io/5b26t>

87. Durmus, E. *et al.* (2024) Towards Measuring the Representation of Subjective Global Opinions in Language Models. In *Proceedings of the First Conference on Language Modeling*
88. Haerpfner, C. *et al.* (2020) *World Values Survey: Round Seven-Country-Pooled Datafile*, JD Systems Institute & VWUSA Secretariat <https://doi.org/10.14281/18241.24> Version 5.0
89. Li, M. *et al.* (2024) Can LLMs Speak For Diverse People? Tuning LLMs via Debate to Generate Controllable Controversial Statements. In *Findings of the Association for Computational Linguistics ACL 2024*, pp. 16160–16176
90. Wang, Q. *et al.* (2025) Multilingual Prompting for Improving LLM Generation Diversity. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (Christodoulopoulos, C. et al., eds)*, pp. 6367–6389, Association for Computational Linguistics
91. Santurkar, S. *et al.* (2023) Whose opinions do language models reflect? In *International Conference on Machine Learning*, pp. 29971–30004, PMLR
92. Park, P.S. *et al.* (2024) Diminished diversity-of-thought in a standard large language model. *Behav. Res. Methods* 56, 5754–5770
93. Hayati, S.A. *et al.* (2024) How Far Can We Extract Diverse Perspectives from Large Language Models? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (Al-Onaizan, Y. et al., eds)*, pp. 5336–5366, Association for Computational Linguistics
94. Hu, Z. *et al.* (2025) Debate-to-Write: A Persona-Driven Multi-Agent Framework for Diverse Argument Generation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 4689–4703
95. Xiao, J. *et al.* (2025) On the algorithmic bias of aligning large language models with RLHF: preference collapse and matching regularization. *J. Am. Stat. Assoc.* 120, 2154–2164
96. Hwang, A.H.-C. *et al.* (2025) 'It was 80% me, 20% AI': seeking authenticity in co-writing with large language models. *Proc. ACM Hum. Comput. Interact.* 9, 1–41
97. Lubart, T.I. (2001) Models of the creative process: past, present and future. *Creat. Res. J.* 13, 295–308
98. Nijstad, B.A. *et al.* (2002) Cognitive stimulation and interference in groups: exposure effects in an idea generation task. *J. Exp. Soc. Psychol.* 38, 535–544
99. Coman, A. and Hirst, W. (2015) Social identity and socially shared retrieval-induced forgetting: the effects of group membership. *J. Exp. Psychol. Gen.* 144, 717
100. Zhang, S. *et al.* (2025) Generative AI meets open-ended survey responses: research participant use of AI and homogenization. *Sociol. Methods Res.* 54, 1197–1242
101. Agarwal, D. *et al.* (2025) AI suggestions homogenize writing toward western styles and diminish cultural nuances. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pp. 1–21
102. Rashotte, L. (2007) Social influence. In *The Blackwell Encyclopedia of Sociology*, Wiley
103. Medin, D.L. and Atran, S. (2004) The native mind: biological categorization and reasoning in development and across cultures. *Psychol. Rev.* 111, 960
104. Medin, D.L. *et al.* (2007) Culture, categorization, and reasoning. In *Handbook of Cultural Psychology* (Kitayama, S. and Cohen, D., eds), pp. 615–644, The Guilford Press
105. Muthukrishna, M. and Henrich, J. (2016) Innovation in the collective brain. *Philos. Trans. R. Soc. B Biol. Sci.* 371, 20150192
106. Duede, E. *et al.* (2024) Being together in place as a catalyst for scientific advance. *Res. Policy* 53, 104911
107. Valmeekam, K. *et al.* (2023) On the planning abilities of large language models: a critical investigation. *Adv. Neural Inf. Process. Syst.* 36, 75993–76005
108. Chen, Y. *et al.* (2024) Do Large Language Models have Problem-Solving Capability under Incomplete Information Scenarios? In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 2225–2238, Association for Computational Linguistics
109. Kleinberg, J. and Raghavan, M. (2021) Algorithmic monoculture and social welfare. *Proc. Natl. Acad. Sci.* 118, e2018340118
110. Bommasani, R. *et al.* (2022) Picking on the same person: does algorithmic monoculture lead to outcome homogenization? *Adv. Neural Inf. Process. Syst.* 35, 3663–3678
111. Bender, E.M. *et al.* (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*, pp. 610–623
112. Binz, M. and Schulz, E. (2023) Using cognitive psychology to understand GPT-3. *Proc. Natl. Acad. Sci.* 120, e2218523120
113. Hagendorff, T. *et al.* (2022) Thinking fast and slow in large language models. *arXiv* <https://doi.org/10.48550/arXiv.2212.05206>
114. Yasunaga, M. *et al.* (2024) Large Language Models as Analogical Reasoners. In *Proceedings of the Twelfth International Conference on Learning Representations (ICLR)*
115. Yao, S. *et al.* (2023) Tree of thoughts: deliberate problem solving with large language models. *Adv. Neural Inf. Process. Syst.* 36, 11809–11822
116. Wiratunga, N. *et al.* (2024) CBR-RAG: case-based reasoning for retrieval augmented generation in LLMs for legal question answering. In *International Conference on Case-Based Reasoning*, pp. 445–460, Springer
117. Shao, H. *et al.* (2024) Visual CoT: advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. *Adv. Neural Inf. Process. Syst.* 37, 8612–8642
118. Han, S. *et al.* (2025) Videoespresso: A large-scale chain-of-thought dataset for fine-grained video reasoning via core frame selection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 26181–26191
119. Liu, R. *et al.* (2025) Mind Your Step (by Step): Chain-of-Thought can Reduce Performance on Tasks where Thinking Makes Humans Worse. In *Proceedings of the Forty-second International Conference on Machine Learning (ICML)*
120. Doshi, A.R. and Hauser, O.P. (2024) Generative AI enhances individual creativity but reduces the collective diversity of novel content. *Sci. Adv.* 10, eadn5290
121. Anderson, B.R. *et al.* (2024) Homogenization effects of large language models on human creative ideation. In *Proceedings of the 16th conference on creativity & cognition*, pp. 413–425
122. Dang, H. *et al.* (2023) Choice over control: How users write with large language models using diegetic and non-diegetic prompting. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–17
123. Qin, P. *et al.* (2025) Timing Matters: How Using LLMs at Different Timings Influences Writers' Perceptions and Ideation Outcomes in AI-Assisted Ideation. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pp. 1–16
124. Kosmyna, N. *et al.* (2025) Your brain on ChatGPT: accumulation of cognitive debt when using an AI assistant for essay writing task. *arXiv* <https://doi.org/10.48550/arXiv.2506.08872>
125. Ritzer, G. (1983) The "McDonaldization" of society. *J. Am. Cult.* 6, 100–107
126. Orwell, G. (1949) *Nineteen Eighty-Four*, Secker & Warburg
127. Aroyo, L. *et al.* (2023) Dices dataset: diversity in conversational AI evaluation for safety. *Adv. Neural Inf. Process. Syst.* 36, 53330–53342
128. Sorensen, T. *et al.* (2024) Position: a roadmap to pluralistic alignment. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 46280–46302
129. Salali, G.D. *et al.* (2025) From cognition in the wild to cognition *in silico*: flourishing cultural diversity in AI. OSF https://doi.org/10.31234/osf.io/f9gdt_v2
130. Du, Y. *et al.* (2023) Improving factuality and reasoning in language models through multiagent debate. In *Forty-first International Conference on Machine Learning*
131. Gekhman, Z. *et al.* (2024) Does Fine-Tuning LLMs on New Knowledge Encourage Hallucinations? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 7765–7784
132. Bick, A. *et al.* (2024) *The Rapid Adoption of Generative AI*, 32966. NBER Working Paper. <https://doi.org/10.3386/w32966>
133. Chatterji, A. *et al.* (2025) *How People Use ChatGPT*, 34255. NBER Working Paper. <https://doi.org/10.3386/w34255>