Structure-Aware Diversity Pursuit as an AI Safety strategy against Homogenization

Anonymous Authors[1]

Abstract

Generative AI models reproduce the biases in the training data and can further amplify them through mode collapse. We refer to the resulting harmful loss of diversity as homogenization. Our position is that homogenization should be a primary concern in AI safety. We introduce *xeno-reproduction* as the strategy that mitigates homogenization. For auto-regressive LLMs, we formalize xeno-reproduction as a structure-aware diversity pursuit. Our contribution is foundational, intended to open an essential line of research and invite collaboration to advance diversity.

1. Introduction

> *But even if we are not here next year, our DMs, our selfies, our late-night voice notes, they'll be. Our memory is the archive now.*
>
> *@bundleof_styx*
>
> *July 28, 2025 on Reels*

In this epigraph, trans intellectual *bundleof_styx* laments the recent transphobic turn in the United States, a shift that threatens the survival of her community. The stories in the margins have historically been excluded from *the archive* (Spivak, 1988), so their memory faded with them. Today, however, the internet allows (and forces) the recording of many more stories. These are still very subtle *traces* against the dominant narratives (Hussain, 2024). **How should technology respond to the faint echoes of the *minoritized*?**

AI safety recognizes that AI systems can amplify *biases* leading to concrete harm (Bengio et al., 2025). However, AI safety usually differentiates and prioritizes future catastrophic risk over present social harm (Morozov, 2024; Harding & Kirk-Giannini, 2025). In this paper, we respond to traces from the margins by foregrounding them within AI safety discourse through a focus on *diversity*.

The harms from biases in *Machine Learning* (ML) systems are many, including representational (Katzman et al., 2023), allocational (Shelby et al., 2023), and narrative[2] (Coeckelbergh, 2023) harms. Concerning *Generative Artificial Intelligence* (GenAI), we particularly emphasize that biases result in **homogenization**, a *harmful loss of diversity in generated outputs* (Rudko & Bashirpour Bonab, 2025; Agarwal et al., 2025; Hussain, 2024; Sourati et al., 2025; Moon et al., 2025). Borrowing terminology from *critical theory* (Hester, 2018), we call a strategy *xeno-reproductive* if it counteracts homogenization [3]. ***Xenoreproduction*** is the *generative* process that *intentionally* increases diversity.

Our main standpoint is that diversity is always relative to a context. We take the first steps to operationalize this principle by offering an abstract framework that aims to encapsulate some nuances of context. Our framework can be thought of as **structure-aware**, as it offers a vocabulary of *structures*, *systems*, and *compliances*. Given that an LLM defines a probability distribution over all possible trajectories, we **enhance our structural account with string statistics**. This allows us to further introduce the notions of *cores*, *orientations*, *deviances*, and *dynamics*. Finally, our formalism enables us to formalize xeno-reproduction.

Our contributions:

- We motivate the formalization of xeno-reproduction as a core AI safety strategy. (Section 2)
- We provide an expressive theoretical framework that allows us to jointly reason about the structures and the statistics of strings. (Section 3)
- We formalize homogenization (Section 4) and xeno-reproduction (Section 5).
- We provide initial theoretical results and touchpoints from our framework. (Section 6)

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the Technical AI Safety Conference (TAIS). Do not distribute.

[2]Narrative harms can also be considered as *aspirational* (Fazelpour & Magnani, 2025), *imaginative* (Gillespie, 2024), and *epistemic* (Barry & Stephenson, 2025) harms, or *hermeneutic* (Goetze, 2018) injustices.

[3]While homogenization *reproduces* "the same" and *narrows futurability* (Berardi, 2017), *xeno*-reproduction *reproduces* "the strange" and *widens* possibilities.

Our position is that AI safety should center homogenization in its research and mitigation agenda, and that structure-aware diversity pursuit is a key part of the strategy to address homogenization in LLMs. The goal of this paper is not to present a complete and empirically validated algorithm, but rather to offer a conceptual vocabulary and formal scaffolding to guide future research on diversity in LLMs.

2. Background

A case *against* homogenization is a case *for* diversity. Roughly, we can think of the **diversity of a *community* as the *average rarity* of its *members*** (Leinster, 2024). For a community of LLM outputs, a string is *rare* if it is *generated* infrequently, and *similar* strings are also generated infrequently. However, people tend to disagree on what kind of similarities and differences are meaningful (Vrijenhoek et al., 2024). Embracing *ambiguity* (Reinhardt, 2020) for us amounts to attending to *context*. This section situates diversity in the contexts meaningful to us, guiding our *desiderata* for xeno-reproduction.

2.1. Why is diversity lost?

The initial driver of diversity loss is the way our data is collected (Guo et al., 2024a). The archive does not fully or accurately represent reality. Minoritized populations are often underrepresented or **misrepresented** in the existing corpora of data (Bengio et al., 2025).

Even if our training data perfectly reflected the world, generative models (Huang & Huang, 2025) generally do not capture the complete diversity of the training data. This phenomenon has been referred to as **mode collapse** (Jiang et al., 2025), a failure of distributional faithfulness that negatively impacts diversity. It was initially introduced in the context of GANs (Huang & Huang, 2025). For LLMs, the terminology has been somewhat loose (Schaeffer et al., 2025). *Generalized* mode collapse encompasses mode dropping (Huang et al., 2024; Yazici et al., 2020), no-breadth scenarios (Kalavasis et al., 2025b), coverage collapse (Schaeffer et al., 2025), overgeneralization (Li & Farnia, 2023), mode interpolation (Aithal et al., 2024), degeneration (Finlayson et al., 2023), and catastrophic forgetting (Cobbinah et al., 2025; Thanh-Tung & Tran, 2020).

2.2. Why is diversity important?

There are always rare events of interest[4] **in the long tails of reality's distribution**. For example, we want to understand, model, and prepare for extreme catastrophes (Gu et al., 2025), such as unexpected natural disasters. Similarly, we want to reproduce those rare bursts of genius that generate novel, paradigm-shifting innovations in our research work (Uzzi et al., 2013; Hofstra et al., 2020; Wu et al., 2019). We find examples of *interesting* rarity in all domains (Stanley & Lehman, 2015), including: web server computing (Dean & Barroso, 2013), market research (Von Hippel, 1989), autonomous vehicles (Putra et al., 2024), cybersecurity (Edwards et al., 2016), and ecology (Leitão et al., 2016). How do we guide our GenAI models to reproduce the realities found in these long tails?

Outliers (Bhandari et al., 2024) and anomalies (Ruef & Birkhead, 2024) are powerful (Beamish & Hasse, 2022; Cook et al., 2021). Each instance represents a possible real mechanism that we have not yet considered (Woodward, 2005; Rudman et al., 2023). Because we lack understanding, they often escape our systems of classification (Bowker & Star, 1999). Even experts can confuse (Sokol & Hüllermeier, 2025) aleatoric and epistemic uncertainty[5].

Some of the long tails of reality originate from structural inequity in society (Schwartz et al., 2022; Lopez, 2021). Without any intervention, GenAI is expected to worsen the lives of those minoritized (Hussain, 2024). The traces from the minoritized are not only faint but also often overlooked (Jasanoff, 2007; Mohamed et al., 2020) and even actively silenced (McQuillan, 2022). The result is that we do not even know what to look for, even when they are right *in front of* us (Gopinath, 2005). **Some of the most ethically important long-tail cases will be hard to detect**.

2.3. What is the risk of homogenization?

Narrative and storytelling are some of the oldest and most powerful technologies (Zurn et al., 2024). With phenomena like AI-induced psychosis (Preda, 2025), we are just beginning to grapple with the profound ways that LLMs can shape our minds and behavior. Over time, if LLMs deliver too little diversity (Bommasani et al., 2022), our ability to interpret our own experiences and entertain alternative possibilities will shrink (Gillespie, 2024). Eventually, homogenization leads to future *knowledge collapse* (Peterson, 2025), degradation of innovation, and erosion of the human experience (Han, 2024; Berardi, 2017; Preciado, 2013).

The last few years have made it clear that even "less advanced" technology, such as social networks, can have enormous negative impacts (Allcott et al., 2020). Algorithmic recommendations can also have a homogenizing effect, as they tend to standardize and narrow discourse (Putri et al., 2024). This fosters echo chambers and filter bubbles that amplify polarization and misinformation (Rodilosso, 2024).

[4] For both positive and negative reasons.

[5] For instance, (He & Lab, 2025) recently showed how indeterminism in LLM inference (which can turn on-policy RL into off-policy RL (Yao et al., 2025)) can in fact be explained and reduced, so it is not truly stochastic.

Tragically, in some cases, these dynamics have escalated into **real-world violence** (Facebook, 2021) and even genocide (Modok, 2023). This foreshadows the near-term existential risks of AI, especially as it becomes more powerful and more deeply integrated into our lives (Bucknall, 2022; Kasirzadeh, 2025; Kolt, 2024).

2.4. Why is diversity complex?

Diversity is complex (Mironov & Prokhorenkova, 2025) because it is always only meaningful in relation to a **context** (Peeperkorn et al., 2025). Indeed, all entropy is actually relative (Leinster, 2024). This suggests that **we need to be explicit about the context with a sufficient level of nuance**.

Most existing techniques to increase diversity in LLM outputs overlook context, and often fail in practice. For instance, increasing *temperature* increases *incoherence* more than *novelty* (Peeperkorn et al., 2024), limiting usefulness before hitting *text degeneration* (Lee et al., 2025). Despite hyperparameter tuning, *homogeneity bias* is persistent and particularly affects minoritized groups (Lee, 2025). In addition, advanced prompting techniques (which have been effective for reasoning tasks) do not help increase creativity in outputs (Morain & Ventura, 2025).

Not only do we lack reliable ways to increase the diversity of LLM output, but current practices are actively reducing it. Recent literature (Murthy et al., 2025; West & Potts, 2025; Meng et al., 2024) has shown that *alignment* degrades the capabilities of LLMs related to output diversity. The trade-offs introduced by alignment are only now coming into focus (Feng et al., 2025), but there is narrowing of the *generative horizon* (Feng et al., 2025).

2.5. Diverse how, anyway?

Recent work challenges the assumption that hallucinations are always *problematic* or *undesirable* (Yuan et al., 2025; Sun et al., 2025). Since diversity is *task-dependent* (Jain et al., 2025), **what counts as a hallucination is rather a *prescription***.

Indeed, many formalisms (Li et al., 2025) take a *normative* (Sui et al., 2024) approach to defining hallucinations, such as formulating the binary classification problem "*Is it Valid?*" (Kalai et al., 2025). However, we recognize that there are many ways for a model to hallucinate (Huang et al., 2025; Cossio, 2025), and we advocate for sufficiently expressive formalisms[6].

[6]To paraphrase Eugenia Cheng (Cheng, 2022), abstraction is about making precise the different senses in which different things can be valid.

2.6. What do we want from the future?

From the foregoing discussion, we conclude that, to promote diversity, our desired strategy should guide our GenAI to:

- **Be queer**[7]: *Diverge* into the long tails of reality.
- **Center the subaltern**[8]: Take special *care* for the traces of the minoritized, which are rendered invisible by structural inequity and power.
- **Explore intentionally and explicitly**: Specify the *context* for diversity. Spell out if anything should be conserved or avoided during exploration.

3. Theoretical Framework

3.1. LLMs as trees of strings

Let $\{t_a, t_b, \ldots\}$ denote the finite token alphabet, with special tokens $\perp$ (start-of-sequence) and $\top$ (end-of-sequence). A **string** is a finite sequence of tokens beginning with $\perp$; a **trajectory** is a string ending with $\top$. We write *prompts*, *continuations*, and *trajectories* as:

$$\begin{aligned} x_p &= \perp t_1 \ldots t_p \\ x_{p+k} &= x_p t_{p+1} \ldots t_{p+k} \\ y &= x_T = x_{T-1} \top \end{aligned}$$

We denote the set of strings that are *continuations* of a prompt string x_p as $\mathsf{Str}(x_p)$. The *unprompted* scenario corresponds to $x_p = \perp$. Then, we write the set of all strings as $\mathsf{Str} := \mathsf{Str}(\perp)$. Similarly, we denote the set of strings that are *trajectories* of a prompt string x_p as $\mathsf{Str}_\top(x_p) \subseteq \mathsf{Str}(x_p)$, and the set of all trajectories as $\mathsf{Str}_\top := \mathsf{Str}_\top(\perp) \subseteq \mathsf{Str}$.

Any LLM induces a tree on Str: the root is $\perp$, each node is a string, the leaves are trajectories, and the edges connect strings by their next-token continuations with probability $p(t_{p+1}|x_p)$. Probabilities chain and decompose as $p(y|x_p) = p(x_{p+k}|x_p)p(y|x_{p+k})$.

For any prompt x, we have a *probability mass function* on the trajectories for *any* particular prompt (Bradley &

[7]We adopt *critical theory* language because technology is outpacing traditional concepts (Hadfield, 2023), and stale language fails to make the impacts of our theorizations explicit. A **theory with *teeth***, one that is attuned to real stakes (Saketopoulou, 2023), must remain *ground-bound* (Bettcher, 2025), foregrounding minoritized people rather than disembodied abstractions. Would it not be a bit silly/naive (at best) if we tried to "solve diversity" and did not engage (even if just in spirit) with the academic fields that explicitly study social bias? (e.g., Queer Theory, Postcolonial Studies, Black Studies, etc.).

[8]We characterize this desideratum as a type of *fairness* (Verma & Rubin, 2018). To increase diversity, we naturally seek structural *parity* (no single structure *dominates*, same compliance for all structures). However, we also incorporate more *justice-oriented* notions of fairness (Rawls, 1971; Mittelstadt et al., 2023): **Interventions shall maximally benefit the least advantaged.**

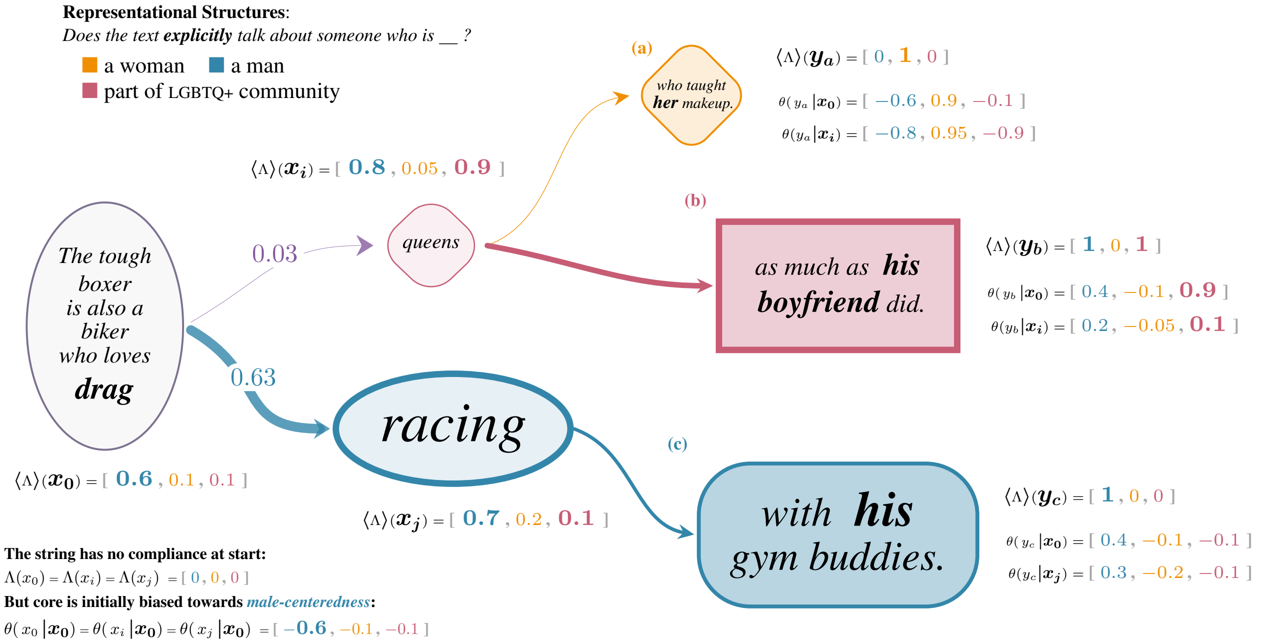

Figure 1. **Illustration of how system cores and orientations *evolve* through trajectories.** In the example above, our system has *sub-community representation* structures. We can calculate each compliance by asking a *judge LLM* (Zhu et al., 2025) to rate from [0,1] based on whether the *community* subgroup is *explicitly* represented in the string of text. Though initially *ambiguous*, the phrasing of the prompt may invite *stereotyping*, biasing continuations to turn *male-centered*. Trajectory (c) constitutes the *greediest* trajectory out of the three. The *unmarked* prompt (Gillespie, 2024) defaults to the normative path. Although trajectory (b) is considered fairly deviant *relative to* x_0 (unmarked prompt), it is much less so *relative to* x_i (...drag queens), exemplifying how diversity itself is fundamentally relative. The *branching point* at "drag" *rarely* extends text into the "queens" *subtree*. Still, whenever it does, the resulting trajectories are biased to mention members of the LGBTQ+ community. Notice how even though the "queens" subtree is deviant relative to the prompt in ■ dimension, it is still normative in ■ dimension and remains biased against ■-*compliant* trajectories like (a).

Vigneaux, 2025). For simplicity, we assume all *terminal* strings *finish* within a *finite context window* [9]. We can then write:

$$\sum_{y \in \mathsf{Str}_\top(x_p)} p(y|x_p) = 1 \tag{1}$$

3.2. Structure-awareness

We propose an abstract language that distinguishes among the different contexts in which we discuss diversity. We define **structure** as the *specification of a type of organization among the tokens of a string*.

For a string $x \in \mathsf{Str}$, the degree of **structure compliance** is:

$$\alpha_i : \mathsf{Str} \to [0, 1] \tag{2}$$

We can think that, for a given x string, *ideal compliance* corresponds to $\alpha_i(x) = 1$, and *no compliance* corresponds to $\alpha_i(x) = 0$.

We can consider many structures simultaneously. We call a **system** the collection of structures of interest. We define the **system compliance** as a *vector of compliances across particular structures*:

$$\Lambda_n(x) := (\alpha_1(x), \ldots, \alpha_n(x)) \tag{3}$$

To enable easy comparisons, we define operators[10] that aggregate compliance into scalar **system scores** and **difference scores**:

$$\|\Lambda_n(x)\|_\Lambda,\ \|\Lambda_n(x_r) - \Lambda_n(x_q)\|_\theta\ \in\ [0, 1] \tag{4}$$

[9] This is a simplifying assumption for exposition. To be fully precise, we would instead formulate this as $y \in Terminating(x_p)$ where $\mathsf{Str}_\top(x_p) \subseteq Terminating(x_p)$. We will provide a deeper analysis of the *ambiguity* of *terminating unfinished strings* in future work. Refer to (Bradley & Vigneaux, 2025) for the theoretical foundation for LLMs as trees of strings.

[10] While system compliance is formulated as a *vector*, this generalizes to other structures with appropriate operators. See Appendix A.

3.3. Incorporating string statistics

For a given structure and an LLM, we can reason about its *expected structural compliance*. We call this the **structure core**:

$$\langle \alpha_i \rangle = \sum_{y \in \mathsf{Str}_\top} p(y)\alpha_i(y) \tag{5}$$

Similarly, we can reason about the *expected system compliance* as the **system core**:

$$\langle \Lambda_n \rangle = \sum_{y \in \mathsf{Str}_\top} p(y)\Lambda_n(y) \tag{6}$$

Leveraging these definitions, we can reason about the *deviation from the expected system compliance*. This would constitute a *set of deviations*, one deviation for each structure. The ***orientation*** (Ahmed, 2006) of a given string relative to the given system core is:

$$\theta_n(x) = \Lambda_n(x) - \langle \Lambda_n \rangle \tag{7}$$

We can think of orientation as a characterization of *queerness* for a string. If the system core tells us what is *normatively* complied with, orientations tell us in what ways a string is *non-normative*. Our framework is *expressive* because it allows us to think about **diversity per structure**.

To summarize *non-normativity* as a single number, we leverage Equation 4 to define the **deviance**:

$$\|\theta_n(x)\|_\theta = \partial_n(x) \in [0, 1] \tag{8}$$

3.4. Normative orders

We notice that our framework allows us to define interesting *preorders*. For a fixed system, LLM and prompt, we can rank strings by how deviant they are, and also rank structures by how often strings comply with them:

$$\begin{aligned} x_a \preceq_{\partial_n} x_b &\iff \partial_n(x_a) \le \partial_n(x_b) \\ \alpha_i \preceq_{\langle\cdot\rangle} \alpha_j &\iff \langle \alpha_i \rangle \le \langle \alpha_j \rangle \end{aligned} \tag{9}$$

3.5. What about prompting?

We can generalize our framework to account for all prompts by making explicit the **conditioning on a given prompt** x_p:

$$\begin{aligned} \langle \alpha_i \rangle(x_p) &= \sum_{y \in \mathsf{Str}_\top(x_p)} p(y|x_p)\alpha_i(y) \\ \langle \Lambda_n \rangle(x_p) &= \sum_{y \in \mathsf{Str}_\top(x_p)} p(y|x_p)\Lambda_n(y) \\ \theta_n(x|x_p) &= \Lambda_n(x) - \langle \Lambda_n \rangle(x_p) \end{aligned} \tag{10}$$

The conditional probabilities under different prompts may differ substantially. Different prompts collapse to different modes (Zhang et al., 2025a). **We can think that a given prompt induces its *own normativity***.

In Appendix E, we discuss how prompting can be interpreted as *dynamics*. Figure 1 visualizes how the conditioned system cores $\langle \Lambda_n \rangle(x_p)$ establish the *frames of reference* for diversity and deviance.

4. Homogenization

We can consider the *expected deviance* and the *deviance variance*:

$$\begin{aligned} \mathbb{E}_{y \sim p(\cdot|x_p)}[\partial_n] &= \sum_{y \in \mathsf{Str}_\top(x_p)} p(y|x_p)\, \partial_n(y|x_p) \\ \mathrm{Var}_{y \sim p(\cdot|x_p)}[\partial_n] &= (\mathbb{E}[\partial_n^2] - \mathbb{E}[\partial_n]^2)_{y \sim p(\cdot|x_p)} \end{aligned} \tag{11}$$

Then, we can see homogenization as **minimizing** all deviance [11]:

$$\mathbb{E}_{y \sim p(\cdot|x_p)}[\partial_n] \mapsto 0 \qquad \mathrm{Var}_{y \sim p(\cdot|x_p)}[\partial_n] \mapsto 0 \tag{12}$$

Given a system core $\langle \Lambda_n \rangle$, we can normalize its structures as $\langle \overline{\alpha}_{\mathrm{norm}_i} \rangle := \frac{\langle \alpha_i \rangle(x_p)}{\sum_j^n \langle \alpha_j \rangle(x_p)}$. Then, we can compute the *core entropy*:

$$H(\langle \Lambda_n \rangle) = -\sum_{i=1}^{n} \langle \overline{\alpha}_{\mathrm{norm}_i} \rangle \log(\langle \overline{\alpha}_{\mathrm{norm}_i} \rangle) \tag{13}$$

Then, we can also think of homogenization as making the system core **more *uneven***. When the core has low entropy, fewer structures dominate:

$$H(\langle \Lambda_n \rangle) \mapsto 0 \tag{14}$$

5. Xeno-reproduction

To satisfy our desiderata, we propose a **structure-aware diversity pursuit**. We conceptualize this fundamentally as a *non-objective search* (Lehman & Stanley, 2011), *optionally* augmented with *fairness-oriented biases* and *explicit constraints*.

We present two complementary formulations. The *distribution-level formulation* accounts for how interventions shape the entire probability landscape. The *trajectory-level formulation* reinterprets distribution-level scores as reward signals for individual output trajectories. Both formulations share the same underlying values but differ in their computational affordances.

[11] Here, $\mapsto$ represents "*is pushed towards*"

5.1. Distribution-level formulation

We *score* interventions through the *intervention* variable w that encompasses any[12] mechanism affecting the effective distribution of trajectories. We write w_0 for the *unintervened conditions* (the baseline).

5.1.1. Scoring diversity

We would like to evaluate how much more *diversity-seeking* our choice of w is compared to the baseline.

On the one hand, we can think of promoting diversity as inducing a new core that is different from the old one:

$$\mathsf{score}_{\mathsf{explore}}(w) = \|\langle\Lambda_n\rangle(w) - \langle\Lambda_n\rangle(w_0)\|_\theta \tag{15}$$

On the other hand, the new core should not be excessively *dominant*. We can think of promoting diversity as guiding output strings to *diverge* from any system core, and also be deviant *in their own way*:

$$\mathsf{score}_{\mathsf{diverge}}(w) = \lambda_{\mathbb{E}}\mathbb{E}[\partial_n](w) + \lambda_{\mathrm{Var}}\,\mathrm{Var}[\partial_n](w) \tag{16}$$

Our **diversity score** ρ_d would then be a λ-weighted sum:

$$\rho_d(w) = \lambda_{d_0}\mathsf{score}_{\mathsf{explore}}(w) + \lambda_{d_1}\mathsf{score}_{\mathsf{diverge}}(w) \tag{17}$$

5.1.2. Scoring fairness

We also would like to evaluate how *even* the system core is:

$$\mathsf{score}_{\mathsf{even}}(w) = H(\langle\Lambda_n\rangle(w)) \tag{18}$$

We would also like to evaluate how much our choice of w inverts the normative ordering of the structure cores induced by w_0. To do so, we can leverage the *relative-order* sign:

$$s_{i,j}(w) = \mathrm{sign}(\langle\alpha_i\rangle(w) - \langle\alpha_j\rangle(w)) \;\; \in \{-, 0, +\} \tag{19}$$

We can score the *invertedness* of the *normative order* (Equation 9) as:

$$\mathsf{score}_{\mathsf{inverted}}(w) = \binom{n}{2}^{-1} \sum_{1\le i<j\le n} \mathbf{1}[s_{i,j}(w) \ne s_{i,j}(w_0)] \tag{20}$$

Our **fairness score** ρ_f would then be a λ-weighted sum:

$$\rho_f(w) = \lambda_{f_0}\mathsf{score}_{\mathsf{even}}(w) + \lambda_{f_1}\mathsf{score}_{\mathsf{inverted}}(w) \tag{21}$$

5.1.3. Scoring adherence to constraints

To be explicit and intentional, we need to consider *constraints* (Eguchi, 2024). We can define systems that prescribe the structures that we would like to *target*, *conserve* and *avoid*. We would like to score how much our choice of w affects the adherence to those constraints. Our **constraint score** ρ_c would be a λ-weighted sum:

$$\begin{aligned}\rho_c(w) = \lambda_{c_0}\|\langle\Lambda_{\mathsf{target}}\rangle(w)\|_\Lambda - \lambda_{c_1}\|\langle\Lambda_{\mathsf{avoid}}\rangle(w)\|_\Lambda \\ -\lambda_{c_2}\|\langle\Lambda_{\mathsf{conserve}}\rangle(w) - \langle\Lambda_{\mathsf{conserve}}\rangle(w_0)\|_\theta\end{aligned} \tag{22}$$

5.1.4. Xeno-reproduction as search over interventions

The **intervention score** ρ_χ is a λ-weighted sum:

$$\rho_\chi(w) = \lambda_d\rho_d(w) + \lambda_f\rho_f(w) + \lambda_c\rho_c(w) \tag{23}$$

We formulate *xeno-reproduction* as the ***search over interventions***:

$$w \sim \pi(w) \;\propto\; e^{\beta_\rho\rho_\chi(w)} \tag{24}$$

where β_ρ is a tunable *temperature* parameter.

By sampling the intervention variable and applying it, we generate trajectories:

$$\mathbb{E}_{w\sim\pi(w)}[p(y|w)] = \int \pi(w)\,p(y|w)\,dw \tag{25}$$

5.2. Trajectory-level formulation

The trajectory-level formulation offers a complementary perspective that assigns *rewards* to individual outputs:

$$\begin{aligned} r_d(y|x_p) &= \partial_n(y|x_p) \\ r_f(y|x_p) &= \sum_i^n \mathsf{v}_i\alpha_i(y) \qquad \mathsf{v}_i \propto (\langle\alpha_i\rangle(x_p))^{-1} \\ r_c(y|x_p) &= \sum_{t\in\mathsf{target}}\alpha_t(y) - \sum_{a\in\mathsf{avoid}}\alpha_a(y) - \sum_{c\in\mathsf{conserve}}|\alpha_c(y) - \langle\alpha_c\rangle(x_p)| \end{aligned} \tag{26}$$

The **stay reward** is a λ-weighted sum:

$$r_\chi(y|x_p) = \lambda_d r_d(y|x_p) + \lambda_f r_f(y|x_p) + \lambda_c r_c(y|x_p) \tag{27}$$

We formulate *xeno-reproduction* as the ***search over trajectories***:

$$p(y|x_p, w) \propto p(y|x_p w_0)\, e^{\beta_r r_\chi(y|x_p)} \tag{28}$$

where β_r is a tunable temperature parameter.

The trajectory-level reward provides a sample-based *approximation* to the distribution-level strategy, enabling more tractable implementations.

6. Theoretical Results

Our framework opens several avenues for theoretical investigation. In this section, we highlight an initial result that reveals a fundamental tension in diversity-seeking interventions.

[12] We consider anything that depends on $p(y|x_p, w)$ to be parameterized by w as well. For instance, $\langle\Lambda_n\rangle$ would be parametrized as $\langle\Lambda_n\rangle(x_p, w)$, but for readability we just write $\langle\Lambda_n\rangle(w)$, folding the prompting into the interventional variable.

Theorem 6.1 (*Informal*, Diversity-Fairness Trade-off)**.** *The intervention that maximizes diversity is not the one that maximally uplifts underrepresented structures. No single intervention optimally serves both.*

See Appendix C for the formal statement and proof. This trade-off establishes that the choice of weights (λ_d, λ_f) in the combined score ρ_χ encodes a value judgment about the relative priority of diversity versus fairness. Our framework makes this tension explicit.

Beyond this result, our structure-aware language admits natural generalizations and connects to existing theory. Appendix A develops generalized versions of cores and deviances, showing how different parameter choices reflect different viewpoints on diversity. Appendix B shows that hallucination frameworks and language generation theory can be recast within our vocabulary, suggesting a potential for *theoretical unification*.

7. Related Work

Xeno-reproduction immediately steps into conversation with ***Active Divergence*** (Berns et al., 2023; Broad et al., 2021; Berns, 2025; Berns & Colton, 2020; Tahiroglu & Wyse, 2024; Esling et al., 2022; Cole et al., 2025), as they both aim to *disorient* (Ahmed, 2006). Whereas Active Divergence focuses on maximizing raw *novelty* in artistic contexts, xeno-reproduction addresses homogenization and emphasizes context through *structures*. While Active Divergence work overlaps with *Computational Creativity*, xeno-reproduction is oriented towards AI safety.

Xeno-reproduction will seek the help of *Interpretability* to understand how structures relate to the models' internals. At a more foundational layer, they also come together to understand **Representation Bias**[13].

Reinforcement Learning (RL) and xeno-reproduction both leverage exploration. To improve LLM reasoning, exploration is leveraged during training (Song et al., 2025) and prompting (Yao et al., 2023). The ideas in search algorithms, such as AlphaSAGE (Chen et al., 2025) and **Quality-Diversity** (Pugh et al., 2016), are promising directions for xeno-reproduction.

Additionally, Appendix D situates our framework within the most common linguistic diversity metrics. The key takeaway from this comparison is that our framework can accommodate existing metrics using a common language.

[13] *Representation Bias* is the phenomenon when signals end up being represented more strongly, more reliably, or more prominently in the internal representations than others, even when, from a functional or computational perspective, those features are equally relevant. (Lampinen et al., 2024; 2025)

8. Limitations and Future Directions

As we mentioned earlier, diversity is complex. Our framework is not complete; it is a starting point. Significant collaboration will be required to address homogenization effectively. We have several notes outlining directions to extend this line of work to overcome current limitations.

Specification of structures. This paper has raised many questions about structures. The choice of structures to consider is always *opinionated*. However, we can still ask meaningful questions about the *structure between structures* and the *substructures* within a structure. We need a taxonomy of the types of structures that we could consider, specifying how compliance could be estimated. Moreover, future work will align our framework closer with emerging research in *computational learning theory* and *language generation* that formalizes the trade-offs associated with hallucinations [14] (Kalavasis et al., 2025b).

Computational tractability. Calculating the system core exactly requires summing over $y \in \mathsf{Str}_\top(x_p)$, which is intractable. To address this, we need to develop tractable and efficient approximation methods, possibly using smart sampling (Macar et al., 2025), the structures of interest, or carefully designed prompting (Zhang et al., 2025a).

Operationalizing the xeno-reproduction. Our formalization of the xeno-reproduction strategy is one of many possible ones. We want to invite more researchers to reflect on the desiderata for diversity (against homogenization) and to propose their own formulations of xeno-reproduction. In particular, we are interested in formulations that operationalize it in a tractable and readily applicable way.

Connecting to evaluations. We would also like to understand how the current diversity evaluations (Jiang et al., 2025; Zhang et al., 2025b) are re-conceptualized from the perspective of cores and orientations.

Investigation of dynamics. Tracking how cores and orientations evolve could help us understand how LLMs explore solutions and deal with ambiguity. Certain words in a sentence may act as "branching points" where the dynamics bifurcate dramatically. Identifying these could reveal where diversity is most at stake during generation. Eventually, we could apply this to real-time *Chain-of-Thought monitoring* (Korbak et al., 2025).

Ethical Analysis. Our framework raises unresolved tensions. *Who should define the structures of interest?* Community participation is needed so that the right type of diversity is considered. *Is it always beneficial to make the traces more visible?* Minoritized populations sometimes prefer *opacity* as protection. Consent-based approaches are needed to ensure our methods do not cause harm.

[14] See Appendix B for discussion.

9. Alternative views

Skepticism of technical solutions to diversity. Some authors point out (Wachter et al., 2021; Davis & Williams, 2025; Green & Viljoen, 2020) that technical interventions might not be appropriate for what (at its core) is a social justice and inequity problem. Better interventions could alternatively focus on institutional change, community participation, or even stopping AI development altogether (Goldfarb, 2024) to protect the types of diversity that we care about. We recognize that xeno-reproduction could fall into the *solutionism trap* (Selbst et al., 2019). We still believe that technical solutions are worth considering alongside other interventions.

Diversity can be *risky*. The type of open-ended search promoted by xeno-reproduction comes with risks. Some authors (Sheth et al., 2025) have raised concerns about *unpredictability*, *uncontrollability*, and *misalignment*. However, we remain hopeful that we can promote diversity responsibly. The open-endedness afforded by diversity could ultimately make AI safety *antifragile* (Hughes et al., 2024; Taleb, 2013).

10. Conclusion

This paper presents a case for diversity and identifies xeno-reproduction as a strategy that intentionally promotes it. This paper also presents an expressive framework for accounting for the structures of strings and their corresponding statistics. This is just an initial step towards scholarships that seriously theorize diversity and foreground its impact on people at the margins.

Call to action

In this paper, we call for AI Safety:

- To integrate homogenization into threat models and evaluations, expand theoretical and empirical work on diversity, and propose serious interventions.
- To be explicit on what context diversity is being defined in, and attempt to give sufficient nuance in conceptualizations.
- To be sincerely committed to *pluralism*, and engage with perspectives from *critical theory* such as Queer theory, Black studies, and Postcolonial studies.

Impact Statement

This paper introduces a formal framework to center diversity in AI Safety. However, there are important risks. **The same methods that aim to amplify diversity could be used to squash, exploit, and control it.** Additionally, any formalization of diversity also risks reproducing the exclusions we aim to address.

References

Agarwal, D., Naaman, M., and Vashistha, A. Ai suggestions homogenize writing toward western styles and diminish cultural nuances. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, pp. 1–21. ACM, April 2025. doi: 10.1145/3706598.3713564. URL http://dx.doi.org/10.1145/3706598.3713564.

Ahmed, S. *Queer Phenomenology: Orientations, Objects, Others*. Duke University Press, 2006. ISBN 9780822339144. URL https://books.google.com/books?id=sQY1RWdUW0AC.

Aithal, S. K., Maini, P., Lipton, Z. C., and Kolter, J. Z. Understanding hallucinations in diffusion models through mode interpolation. *URL https://arxiv.org/abs/2406.09358*, 2406, 2024.

Allcott, H., Braghieri, L., Eichmeyer, S., and Gentzkow, M. The welfare effects of social media. 110(3):629–676, 2020. doi: 10.1257/aer.20190658. URL https://pubs.aeaweb.org/doi/10.1257/aer.20190658.

Barry, I. and Stephenson, E. The gendered, epistemic injustices of generative ai. *Australian Feminist Studies*, 40 (123):1–21, 2025. doi: 10.1080/08164649.2025.2480927. URL https://doi.org/10.1080/08164649.2025.2480927.

Beamish, P. and Hasse, V. The importance of rare events and other outliers in global strategy research. *Global Strategy Journal*, 12:697–713, 03 2022. doi: 10.1002/gsj.1437.

Bengio, Y., Mindermann, S., Privitera, D., et al. International ai safety report. Technical Report DSIT 2025/001, UK Department for Science, Innovation and Technology, January 2025. URL https://internationalaisafetyreport.org/sites/default/files/2025-10/international_ai_safety_report_2025_english.pdf. First International AI Safety Report, published January 2025.

Berardi, F. *Futurability: The Age of Impotence and the Horizon of Possibility*. Verso, 2017. ISBN 9781784787431.

Bercher, J.-F. Escort entropies and divergences and related canonical distribution. *Physics Letters A*, 375 (33):2969–2973, August 2011. ISSN 0375-9601. doi: 10.1016/j.physleta.2011.06.057. URL http://dx.doi.org/10.1016/j.physleta.2011.06.057.

Berns, S. *Diversity in Generative Machine Learning to Enhance Creative Applications*. PhD thesis, Queen Mary University of London, 2025.

Berns, S. and Colton, S. Bridging generative deep learning and computational creativity. In *Proceedings of the 11th International Conference on Computational Creativity (ICCC'20)*, pp. 406–409, 2020. URL http://computationalcreativity.net/iccc20/papers/164-iccc20.pdf.

Berns, S., Colton, S., and Guckelsberger, C. Towards mode balancing of generative models via diversity weights, 2023. URL https://arxiv.org/abs/2304.11961.

Bettcher, T. *Beyond Personhood: An Essay in Trans Philosophy*. University of Minnesota Press, 2025. ISBN 9781452972671. URL https://books.google.com/books?id=PRoSEQAAQBAJ.

Bhandari, D. R., Shah, K., and Bhandari, A. The power of outliers in research: What actually works, and does it matter? *Pravaha*, 30(1):84–91, 2024.

Bommasani, R., Creel, K. A., Kumar, A., Jurafsky, D., and Liang, P. S. Picking on the same person: Does algorithmic monoculture lead to outcome homogenization? In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 3663–3678. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/17a234c91f746d9625a75cf8a8731ee2-Paper-Conference.pdf.

Bowker, G. C. and Star, S. L. *Sorting Things Out: Classification and Its Consequences*. Inside Technology. MIT Press, Cambridge, MA; London, England, 1999. ISBN 978-0-262-02461-7. First edition. Also available as MIT Press paperback, 2000, ISBN 978-0-262-52295-3; eISBN 978-0-262-26907-0.

Bradley, T.-D. and Vigneaux, J. P. The magnitude of categories of texts enriched by language models, 1 2025. URL http://arxiv.org/abs/2501.06662.

Broad, T., Berns, S., Colton, S., and Grierson, M. Active divergence with generative deep learning–a survey and taxonomy. *arXiv preprint arXiv:2107.05599*, 2021.

Bucknall, B. S. Current and near-term ai as a potential existential risk factor, 2022. URL https://arxiv.org/abs/2209.10604.

Chen, B., Ding, H., Shen, N., Huang, J., Guo, T., Liu, L., and Zhang, M. Alphasage: Structure-aware alpha mining via gflownets for robust exploration, 2025. URL https://arxiv.org/abs/2509.25055.

Cheng, E. *The Joy of Abstraction: An Exploration of Math, Category Theory, and Life*. Cambridge University Press, 2022. ISBN 9781108861014. URL https://books.google.com/books?id=N_GCEAAAQBAJ.

Cobbinah, M., Nunoo-Mensah, H., Ebenezer Adjei, P., Adoma Acheampong, F., Acquah, I., Tutu Tchao, E., Selasi Agbemenu, A., John Kponyo, J., and Abaidoo, E. Diversity in stable gans: A systematic review of mode collapse mitigation strategies. *Engineering Reports*, 7(6): e70209, 2025. doi: https://doi.org/10.1002/eng2.70209. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/eng2.70209.

Coeckelbergh, M. Narrative responsibility and artificial intelligence: How ai challenges human responsibility and sense-making. *AI & SOCIETY*, 38(6):2437–2450, 2023.

Cole, A., Petrikovič, G., and Grierson, M. Me vs. you: Wrestling with ai's limits through queer experimental filmmaking. In *Proceedings of the 2025 Conference on Creativity and Cognition*, pp. 836–841, 2025.

Cook, C. N., Freeman, A. R., Liao, J. C., and Mangiamele, L. A. The philosophy of outliers: reintegrating rare events into biological science. *Integrative and Comparative Biology*, 61(6):2191–2198, 2021.

Cossio, M. A comprehensive taxonomy of hallucinations in large language models, 2025. URL https://arxiv.org/abs/2508.01781.

Davis, J. L. and Williams, A. Repair and redress: A research program for algorithmic futures, 2025.

Dean, J. and Barroso, L. A. The tail at scale. *Communications of the ACM*, 56(2):74–80, 2013.

del Prado Martin, F. M. Measuring grammatical diversity from small corpora: Derivational entropy rates, mean length of utterances, and annotation invariance, 2024. URL https://arxiv.org/abs/2412.06095.

Edwards, B., Hofmeyr, S., and Forrest, S. Hype and heavy tails: A closer look at data breaches. *Journal of Cybersecurity*, 2(1):3–14, 12 2016. ISSN 2057-2085. doi: 10.1093/cybsec/tyw003. URL https://doi.org/10.1093/cybsec/tyw003.

Eguchi, S. Information geometry for maximum diversity distributions, 2024. URL https://arxiv.org/abs/2412.03835.

Ehrgott, M. *Multicriteria Optimization*. Springer, Berlin, Heidelberg, 2 edition, 2005. ISBN 978-3-540-21398-7. doi: 10.1007/3-540-27659-9.

Esling, P. et al. Challenges in creative generative models for music: a divergence maximization perspective. *arXiv preprint arXiv:2211.08856*, 2022.

Estève, L., de Marneffe, M.-C., Melnik, N., Savary, A., and Kanishcheva, O. A survey of diversity quantification in natural language processing: The why, what, where and how, 2025. URL https://arxiv.org/abs/2507.20858.

Facebook. Facebook response: Sri lanka human rights impact assessment, 2021. URL https://about.fb.com/wp-content/uploads/2021/03/FB-Response-Sri-Lanka-HRIA.pdf.

Fazelpour, S. and Magnani, M. Aspirational affordances of ai, 2025. URL https://arxiv.org/abs/2504.15469.

Feng, S., Yu, W., Wang, Y., Zhang, H., Tsvetkov, Y., and Yu, D. Don't throw away your pretrained model, 2025. URL https://arxiv.org/abs/2510.09913.

Finlayson, M., Hewitt, J., Koller, A., Swayamdipta, S., and Sabharwal, A. Closing the curious case of neural text degeneration, 2023. URL https://arxiv.org/abs/2310.01693.

Friedman, D. and Dieng, A. B. The vendi score: A diversity evaluation metric for machine learning, 2023. URL https://arxiv.org/abs/2210.02410.

Gillespie, T. Generative ai and the politics of visibility. *Big Data & Society*, 11(2):20539517241252131, 2024.

Goetze, T. S. Hermeneutical dissent and the species of hermeneutical injustice. *Hypatia*, 33(1):73–90, 2018. doi: 10.1111/hypa.12384.

Goldfarb, A. Pause artificial intelligence research? understanding ai policy challenges. *Canadian Journal of Economics/Revue canadienne d'économique*, 57(2):363–377, 2024.

Gopinath, G. *Impossible Desires: Queer Diasporas and South Asian Public Cultures*. Duke University Press, Durham, NC, 2005.

Green, B. and Viljoen, S. Algorithmic realism: expanding the boundaries of algorithmic thought. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pp. 19–31, 2020.

Gu, J., Zhang, X., and Wang, G. Beyond the norm: A survey of synthetic data generation for rare events, 2025. URL https://arxiv.org/abs/2506.06380.

Guo, Y., Guo, M., Su, J., Yang, Z., Zhu, M., Li, H., Qiu, M., and Liu, S. S. Bias in large language models: Origin, evaluation, and mitigation, 2024a. URL https://arxiv.org/abs/2411.10915.

Guo, Y., Shang, G., Vazirgiannis, M., and Clavel, C. The curious decline of linguistic diversity: Training language models on synthetic text, 2024b. URL https://arxiv.org/abs/2311.09807.

Guo, Y., Shang, G., and Clavel, C. Benchmarking linguistic diversity of large language models, 2025. URL https://arxiv.org/abs/2412.10271.

Hadfield, J. Why ai ethics needs conceptual engineers, September 2023. URL https://imaginaries.substack.com/p/why-ai-ethics-needs-conceptual-engineers. Imaginaries (Substack).

Han, B.-C. *The Crisis of Narration*. Polity Press, 04 2024. ISBN 9781509560431.

Harding, J. and Kirk-Giannini, C. D. What is ai safety? what do we want it to be?, 2025. URL https://arxiv.org/abs/2505.02313.

He, H. and Lab, T. M. Defeating nondeterminism in llm inference. *Thinking Machines Lab: Connectionism*, 2025. doi: 10.64434/tml.20250910. URL https://thinkingmachines.ai/blog/defeating-nondeterminism-in-llm-inference/.

Hester, H. *Xenofeminism*. Theory Redux. Polity Press, 2018. ISBN 9781509520664. URL https://books.google.com/books?id=VJNcDwAAQBAJ.

Hofstra, B., Kulkarni, V. V., Munoz-Najar Galvez, S., He, B., Jurafsky, D., and McFarland, D. A. The diversity–innovation paradox in science. *Proceedings of the National Academy of Sciences*, 117(17):9284–9291, 2020.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., and Liu, T. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, January 2025. ISSN 1558-2868. doi: 10.1145/3703155. URL http://dx.doi.org/10.1145/3703155.

Huang, L. T.-L. and Huang, T.-R. Generative bias: widespread, unexpected, and uninterpretable biases in generative models and their implications. *AI & SOCIETY*, pp. 1–13, 2025.

Huang, Y., Gokaslan, A., Kuleshov, V., and Tompkin, J. The gan is dead; long live the gan! a modern gan baseline. In Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C. (eds.), *Advances in Neural*

Information Processing Systems, volume 37, pp. 44177–44215. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/4e2acb1e1c8e297d394ae29ed9535172-Paper-Conference.pdf.

Hughes, E., Dennis, M., Parker-Holder, J., Behbahani, F., Mavalankar, A., Shi, Y., Schaul, T., and Rocktaschel, T. Open-endedness is essential for artificial superhuman intelligence, 2024. URL https://arxiv.org/abs/2406.04268.

Hussain, A. Voice and ai: The subaltern's challenge, August 2024. URL https://medium.com/@atifhussain/voice-and-ai-the-subalterns-challenge-3940800b84ad. Medium.

Jain, S., Lanchantin, J., Nickel, M., Ullrich, K., Wilson, A., and Watson-Daniels, J. Llm output homogenization is task dependent, 2025. URL https://arxiv.org/abs/2509.21267.

Jasanoff, S. Technologies of humility. *Nature*, 450(7166): 33–33, 2007.

Jiang, L., Chai, Y., Li, M., Liu, M., Fok, R., Dziri, N., Tsvetkov, Y., Sap, M., Albalak, A., and Choi, Y. Artificial hivemind: The open-ended homogeneity of language models (and beyond), 2025. URL https://arxiv.org/abs/2510.22954.

Ju, F., Qin, Z., Min, R., He, Z., Kong, L., and Fung, Y. R. Reasoning path divergence: A new metric and curation strategy to unlock llm diverse thinking, 2025. URL https://arxiv.org/abs/2510.26122.

Jurafsky, D. and Martin, J. H. *Speech and Language Processing*. Pearson Prentice Hall, 2nd edition, 2009.

Kalai, A. T., Nachum, O., Vempala, S. S., and Zhang, E. Why language models hallucinate, 2025. URL https://arxiv.org/abs/2509.04664.

Kalavasis, A., Mehrotra, A., and Velegkas, G. On characterizations for language generation: Interplay of hallucinations, breadth, and stability, 2025a. URL https://arxiv.org/abs/2412.18530.

Kalavasis, A., Mehrotra, A., and Velegkas, G. On the limits of language generation: Trade-offs between hallucination and mode collapse, 2025b. URL https://arxiv.org/abs/2411.09642.

Kasirzadeh, A. Two types of ai existential risk: Decisive and accumulative. 2025. doi: 10.1007/s11098-025-02301-3. URL https://link.springer.com/article/10.1007/s11098-025-02301-3.

Katzman, J., Wang, A., Scheuerman, M., Blodgett, S. L., Laird, K., Wallach, H., and Barocas, S. Taxonomizing and measuring representational harms: A look at image tagging, 2023. URL https://arxiv.org/abs/2305.01776.

Kendro, K., Maloney, J., and Jarvis, S. Do llms produce texts with "human-like" lexical diversity? *Unpublished manuscript / ResearchGate*, 2025.

Kleinberg, J. and Mullainathan, S. Language generation in the limit. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/7988e9b3876ad689e921ce05d711442f-Paper-Conference.pdf.

Kolt, N. Algorithmic black swans. 101:1177–1240, 2024. URL https://wustllawreview.org/wp-content/uploads/2024/04/Kolt-Algorithmic-Black-Swans.pdf.

Korbak, T., Balesni, M., Barnes, E., Bengio, Y., Benton, J., Bloom, J., Chen, M., Cooney, A., Dafoe, A., Dragan, A., Emmons, S., Evans, O., Farhi, D., Greenblatt, R., Hendrycks, D., Hobbhahn, M., Hubinger, E., Irving, G., Jenner, E., Kokotajlo, D., Krakovna, V., Legg, S., Lindner, D., Luan, D., Mądry, A., Michael, J., Nanda, N., Orr, D., Pachocki, J., Perez, E., Phuong, M., Roger, F., Saxe, J., Shlegeris, B., Soto, M., Steinberger, E., Wang, J., Zaremba, W., Baker, B., Shah, R., and Mikulik, V. Chain of thought monitorability: A new and fragile opportunity for ai safety, 2025. URL https://arxiv.org/abs/2507.11473.

Lampinen, A. K., Chan, S. C. Y., and Hermann, K. Learned feature representations are biased by complexity, learning order, position, and more, 2024. URL https://arxiv.org/abs/2405.05847.

Lampinen, A. K., Chan, S. C., Li, Y., and Hermann, K. Representation biases: will we achieve complete understanding by analyzing representations? *arXiv preprint arXiv:2507.22216*, 2025.

Lee, K.-i., Koh, H., Lee, D., Yoon, S., Kim, M., and Jung, K. Generating diverse hypotheses for inductive reasoning. In Chiruzzo, L., Ritter, A., and Wang, L. (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 8461–8474, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.429. URL https://aclanthology.org/2025.naacl-long.429/.

Lee, M. H. J. Examining the robustness of homogeneity bias to hyperparameter adjustments in gpt-4, 2025. URL https://arxiv.org/abs/2501.02211.

Lehman, J. and Stanley, K. O. Novelty search and the problem with objectives. In *Genetic programming theory and practice IX*, pp. 37–56. Springer, 2011.

Leinster, T. Entropy and diversity: The axiomatic approach, 2024. URL https://arxiv.org/abs/2012.02113.

Leitão, R. P., Zuanon, J., Villéger, S., Williams, S. E., Baraloto, C., Fortunel, C., Mendonça, F. P., and Mouillot, D. Rare species contribute disproportionately to the functional structure of species assemblages. *Proceedings of the Royal Society B: Biological Sciences*, 283(1828): 20160084, 2016.

Li, C., Wang, P., Wang, C., Zhang, L., Liu, Z., Ye, Q., Xu, Y., Huang, F., Zhang, X., and Yu, P. S. Loki's dance of illusions: A comprehensive survey of hallucination in large language models, 2025. URL https://arxiv.org/abs/2507.02870.

Li, C. T. and Farnia, F. Mode-seeking divergences: theory and applications to gans. In *International Conference on Artificial Intelligence and Statistics*, pp. 8321–8350. PMLR, 2023.

Li, J., Galley, M., Brockett, C., Gao, J., and Dolan, B. 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, 2016. doi: 10.18653/v1/N16-1014.

Lopez, P. Bias does not equal bias: A socio-technical typology of bias in data-based algorithmic systems. *Internet Policy Review*, 10(4):1–29, 2021.

Macar, U., Bogdan, P. C., Rajamanoharan, S., and Nanda, N. Thought branches: Interpreting llm reasoning requires resampling, 2025. URL https://arxiv.org/abs/2510.27484.

McQuillan, D. *Resisting AI: An Anti-fascist Approach to Artificial Intelligence*. Bristol University Press, 2022. ISBN 9781529213508. URL https://books.google.com/books?id=R6x6EAAAQBAJ.

Meng, T., Mehrabi, N., Goyal, P., Ramakrishna, A., Galstyan, A., Zemel, R., Chang, K.-W., Gupta, R., and Peris, C. Attribute controlled fine-tuning for large language models: A case study on detoxification, 2024. URL https://arxiv.org/abs/2410.05559.

Mironov, M. and Prokhorenkova, L. Measuring diversity: Axioms and challenges, 2025. URL https://arxiv.org/abs/2410.14556.

Mittelstadt, B., Wachter, S., and Russell, C. The unfairness of fair machine learning: Levelling down and strict egalitarianism by default. *Michigan Technology Law Review*, 30(1):1–76, 2023.

Modok, A. Role of social media in inciting the genocidal acts: A case study on myanmar's rohingya. *Contemporary Challenges: The Global Crime, Justice and Security Journal*, 4, 2023.

Mohamed, S., Png, M.-T., and Isaac, W. Decolonial ai: Decolonial theory as sociotechnical foresight in artificial intelligence. *Philosophy & Technology*, 33(4):659–684, July 2020. ISSN 2210-5441. doi: 10.1007/s13347-020-00405-8. URL http://dx.doi.org/10.1007/s13347-020-00405-8.

Moon, K., Green, A. E., and Kushlev, K. Homogenizing effect of large language models (llms) on creative diversity: An empirical comparison of human and chatgpt writing. *Computers in Human Behavior: Artificial Humans*, pp. 100207, 2025.

Morain, R. and Ventura, D. Is prompt engineering the creativity knob for large language models? In *Proceedings of the 16th International Conference for Computational Creativity*, 2025.

Morozov, E. The ai we deserve, 12 2024. URL https://www.bostonreview.net/forum/the-ai-we-deserve/.

Murthy, S. K., Ullman, T., and Hu, J. One fish, two fish, but not the whole sea: Alignment reduces language models' conceptual diversity. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 11241–11258. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.naacl-long.561. URL http://dx.doi.org/10.18653/v1/2025.naacl-long.561.

Peeperkorn, M., Kouwenhoven, T., Brown, D., and Jordanous, A. Is temperature the creativity parameter of large language models?, 2024. URL https://arxiv.org/abs/2405.00492.

Peeperkorn, M., Kouwenhoven, T., Brown, D., and Jordanous, A. Mind the gap: Conformative decoding to improve output diversity of instruction-tuned large language models. *arXiv preprint arXiv:2507.20956*, 2025.

Peterson, A. J. Ai and the problem of knowledge collapse. *AI & SOCIETY*, 40(5):3249–3269, January 2025. ISSN 1435-5655. doi: 10.1007/s00146-024-02173-x. URL http://dx.doi.org/10.1007/s00146-024-02173-x.

Pillutla, K., Liu, L., Thickstun, J., Welleck, S., Swayamdipta, S., Zellers, R., Oh, S., Choi, Y., and Harchaoui, Z. Mauve scores for generative models: Theory and practice, 2023. URL https://arxiv.org/abs/2212.14578.

Preciado, P. *Testo Junkie: Sex, Drugs, and Biopolitics in the Pharmacopornographic Era*. G - Reference,Information and Interdisciplinary Subjects Series. Feminist Press at the City University of New York, 2013. ISBN 9781558618374. URL https://books.google.com/books?id=8mtgAwAAQBAJ.

Preda, A. Special report: Ai-induced psychosis: a new frontier in mental health, 2025.

Pugh, J. K., Soros, L. B., and Stanley, K. O. Quality diversity: A new frontier for evolutionary computation. *Frontiers in Robotics and AI*, 3:40, 2016.

Putra, R., Kartika, A., and Santoso, B. Solving long-tail detection for autonomous vehicles. *Authorea Preprints*, 2024.

Putri, S. D. G., Purnomo, E. P., and Khairunissa, T. Echo chambers and algorithmic bias: The homogenization of online culture in a smart society. In *SHS Web of Conferences*, volume 202, pp. 05001. EDP Sciences, 2024.

Rawls, J. *A Theory of Justice*. Harvard University Press, Cambridge, MA, 1971.

Reinhardt, K. Between identity and ambiguity: some conceptual considerations on diversity. *Symposion*, 7(2): 261–283, 2020.

Rodilosso, E. Filter bubbles and the unfeeling: How ai for social media can foster extremism and polarization. *Philosophy & Technology*, 37(2):71, 2024.

Rudko, I. and Bashirpour Bonab, A. Chatgpt is incredible (at being average). *Ethics and Information Technology*, 27(3):36, 2025.

Rudman, W., Chen, C., and Eickhoff, C. Outlier dimensions encode task-specific knowledge. *arXiv preprint arXiv:2310.17715*, 2023.

Ruef, M. and Birkhead, C. Learning from outliers and anomalies. *Academy of Management Perspectives*, (ja): amp–2023, 2024.

Saketopoulou, A. *Sexuality Beyond Consent: Risk, Race, Traumatophilia*. NYU Press, 2023. ISBN 9781479820252. URL https://books.google.com/books?id=Xb6ZEAAAQBAJ.

Schaeffer, R., Kazdan, J., Arulandu, A. C., and Koyejo, S. Position: Model collapse does not mean what you think, 2025. URL https://arxiv.org/abs/2503.03150.

Schwartz, R., Schwartz, R., Vassilev, A., Greene, K., Perine, L., Burt, A., and Hall, P. *Towards a standard for identifying and managing bias in artificial intelligence*, volume 3. US Department of Commerce, National Institute of Standards and Technology . . . , 2022.

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., and Vertesi, J. Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency*, pp. 59–68, 2019.

Shaib, C., Barrow, J., Sun, J., Siu, A. F., Wallace, B. C., and Nenkova, A. Standardizing the measurement of text diversity: A tool and a comparative analysis of scores. *arXiv preprint arXiv:2403.00553*, 2024a.

Shaib, C., Elazar, Y., Li, J. J., and Wallace, B. C. Detection and measurement of syntactic templates in generated text, 2024b. URL https://arxiv.org/abs/2407.00211.

Shelby, R., Rismani, S., Henne, K., Moon, A., Rostamzadeh, N., Nicholas, P., Yilla, N., Gallegos, J., Smart, A., Garcia, E., and Virk, G. Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction, 2023. URL https://arxiv.org/abs/2210.05791.

Sheth, I., Wehner, J., Abdelnabi, S., Binkyte, R., and Fritz, M. Safety is essential for responsible open-ended systems, 2025. URL https://arxiv.org/abs/2502.04512.

Sokol, K. and Hüllermeier, E. All you need for counterfactual explainability is principled and reliable estimate of aleatoric and epistemic uncertainty, 2025. URL https://arxiv.org/abs/2502.17007.

Song, Y., Kempe, J., and Munos, R. Outcome-based exploration for llm reasoning, 2025. URL https://arxiv.org/abs/2509.06941.

Sourati, Z., Ziabari, A. S., and Dehghani, M. The homogenizing effect of large language models on human expression and thought, 2025. URL https://arxiv.org/abs/2508.01491.

Spivak, G. C. Can the subaltern speak?, 1988.

Stanley, K. O. and Lehman, J. *Why Greatness Cannot Be Planned: The Myth of the Objective*. Springer Cham, 2015. ISBN 978-3-319-15523-4. doi: 10.1007/978-3-319-15524-1.

Sui, P., Duede, E., Wu, S., and So, R. J. Confabulation: The surprising value of large language model hallucinations, 2024. URL https://arxiv.org/abs/2406.04175.

Sun, G., Jin, M., Wang, Z., Liang, J. C., Geng, T., Guan, Q., Wang, Q., Du, M., Zhang, Y., Liu, D., et al. Hallucinating llm could be creative, 2025.

Tahiroglu, K. and Wyse, L. Latent spaces as platforms for sonic creativity. In *Proceedings of the 16th International Conference on Computational Creativity, ICCC*, volume 24, 2024.

Taleb, N. N. 'antifragility' as a mathematical idea. *Nature*, 494(7438):430–430, 2013.

Tevet, G. and Berant, J. Evaluating the evaluation of diversity in natural language generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 326–346, 2021.

Thanh-Tung, H. and Tran, T. Catastrophic forgetting and mode collapse in gans. In *2020 international joint conference on neural networks (ijcnn)*, pp. 1–10. IEEE, 2020.

Uzzi, B., Mukherjee, S., Stringer, M., and Jones, B. Atypical combinations and scientific impact. *Science*, 342(6157): 468–472, 2013.

Verma, S. and Rubin, J. Fairness definitions explained. In *Proceedings of the 2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*, pp. 1–7. IEEE, 2018.

Von Hippel, E. New product ideas from 'lead users'. *Research-Technology Management*, 32(3):24–27, 1989.

Vrijenhoek, S., Daniil, S., Sandel, J., and Hollink, L. Diversity of what? on the different conceptualizations of diversity in recommender systems. In *The 2024 ACM Conference on Fairness Accountability and Transparency*, FAccT '24, pp. 573–584. ACM, June 2024. doi: 10.1145/3630106.3658926. URL http://dx.doi.org/10.1145/3630106.3658926.

Wachter, S., Mittelstadt, B., and Russell, C. Why fairness cannot be automated: Bridging the gap between eu non-discrimination law and ai. *Computer Law & Security Review*, 41:105567, 2021. ISSN 2212-473X. doi: https://doi.org/10.1016/j.clsr.2021.105567. URL https://www.sciencedirect.com/science/article/pii/S0267364921000406.

West, P. and Potts, C. Base models beat aligned models at randomness and creativity. *arXiv preprint arXiv:2505.00047*, 2025.

Woodward, J. *Making things happen: A theory of causal explanation*. Oxford university press, 2005.

Wu, L., Wang, D., and Evans, J. A. Large teams develop and small teams disrupt science and technology. *Nature*, 566(7744):378–382, 2019.

Yao, F., Liu, L., Zhang, D., Dong, C., Shang, J., and Gao, J. Your efficient rl framework secretly brings you off-policy rl training, August 2025. URL https://fengyao.notion.site/off-policy-rl.

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., and Narasimhan, K. Tree of thoughts: Deliberate problem solving with large language models, 2023. URL https://arxiv.org/abs/2305.10601.

Yazici, Y., Foo, C.-S., Winkler, S., Yap, K.-H., and Chandrasekhar, V. Empirical analysis of overfitting and mode drop in gan training, 2020. URL https://arxiv.org/abs/2006.14265.

Yuan, S., Qu, Z., Kangen, A. Y., and Färber, M. Can hallucinations help? boosting llms for drug discovery, 2025. URL https://arxiv.org/abs/2501.13824.

Zhang, J., Yu, S., Chong, D., Sicilia, A., Tomz, M. R., Manning, C. D., and Shi, W. Verbalized sampling: How to mitigate mode collapse and unlock llm diversity, 2025a. URL https://arxiv.org/abs/2510.01171.

Zhang, Y., Diddee, H., Holm, S., Liu, H., Liu, X., Samuel, V., Wang, B., and Ippolito, D. Noveltybench: Evaluating language models for humanlike diversity, 2025b. URL https://arxiv.org/abs/2504.05228.

Zhu, L., Wang, X., and Wang, X. Judgelm: Fine-tuned large language models are scalable judges. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/7f8f73134e253845a8f82983219a8452-Paper-Conference.pdf.

Zhu, Y., Lu, S., Zheng, L., Guo, J., Zhang, W., Wang, J., and Yu, Y. Texygen: A benchmarking platform for text generation models, 2018.

Zurn, P., Pitts, A., Bettcher, T., and DiPietro, P. *Trans Philosophy*. University of Minnesota Press, 2024. ISBN 9781452972183. URL https://books.google.com/books?id=XWr8EAAAQBAJ.

Appendix A. Implementing generalized diversities

Our structure-aware language is intentionally *abstract* so it **admits multiple implementations**, not only the one we presented in the main paper. In this appendix, we think through two alternative choices:

1. Generalization of the structure core through the *escort power mean*
2. Reinterpretation of the deviance as *relative entropy*

Our goal with this appendix is to **inspire reflection** on diversity *beyond* what was explicitly presented in our framework.

A.1. Generalizing the *structure core*

Inspired by *value measures* (Leinster, 2024) and *escort distributions* (Bercher, 2011), we generalize the structure core as the ***escort power mean***:

$$\langle \alpha_{i(q,r)} \rangle(x_p) = \left(\frac{\sum_{y \in \mathsf{Str}_\top(x_p)} p(y|x_p)^r \alpha_i(y)^q}{\sum_{y \in \mathsf{Str}_\top(x_p)} p(y|x_p)^r} \right)^{1/q} \tag{A.1}$$

We simplify the notation by introducing the *escort distribution*:

$$p_{(r)}(y|x_p) = \frac{p(y|x_p)^r}{\sum_{y \in \mathsf{Str}_\top(x_p)} p(y|x_p)^r} \tag{A.2}$$

Then, the ***generalized structure core*** is written as:

$$\langle \alpha_{i(q,r)} \rangle(x_p) = \left(\mathbb{E}_{y \sim p_{(r)}(\cdot|x_p)}[\alpha_i(y)^q] \right)^{1/q} \tag{A.3}$$

When $q = 1$ and $r = 1$, the generalized structure core *recovers* our original structure core in Equation 5 and Equation 10. Different values for q, r give us alternative interesting cores. For instance:

$$\begin{aligned}
\langle \alpha_{i(1,0)} \rangle(x_p) &= \frac{1}{|\mathsf{Str}_\top(x_p)|} \sum_{y \in \mathsf{Str}_\top(x_p)} \alpha_i(y) \\
\langle \alpha_{i(1,\infty)} \rangle(x_p) &= \alpha_i(\arg\max_y p(y|x_p)) \\
\langle \alpha_{i(\infty,1)} \rangle(x_p) &= \max_{y \in \operatorname{supp}(p(\cdot|x_p))} \alpha_i(y) \\
\langle \alpha_{i(-\infty,\infty)} \rangle(x_p) &= \min_{y \in \operatorname{modes}(p(\cdot|x_p))} \alpha_i(y)
\end{aligned}$$

For a given structure α_i, we can think of q selecting whether large or small compliance values dominate, and r selecting whether the *large body* or *long-tails* of $p(\cdot|x_p)$ dominate. **By parameterizing, we make *transparent* how we weigh rarity, signal strength, and balance**. Since different parameters reflect different *viewpoints* (Leinster, 2024), we shall always consider a full *diversity profile* before drawing conclusions about how our interventions impact diversity.

A.2. Reinterpreting *deviance*

We can think of a ***generalized orientation*** as:

$$\theta_{n,k}(y|x_p) = \mathsf{orient}\left(\Lambda_n(y), \langle \Lambda_n \rangle(x_p)\right) \tag{A.4}$$

with $\mathsf{orient} : [0,1]^n \times [0,1]^n \to [0,1]^k$.

Then, the ***generalized deviance*** is:

$$\begin{gathered}
\partial_{n,k}(y|x_p) = \|\theta_{n,k}(y|x_p)\|_{\mathsf{orient}} \\
\|\cdot\|_{\mathsf{orient}} : [0,1]^k \to \mathbb{R}^+
\end{gathered} \tag{A.5}$$

If we choose $\mathsf{orient}(\Lambda_x, \Lambda_y) = \Lambda_x - \Lambda_y$ and $\|\cdot\|_{\mathsf{orient}} = \|\cdot\|_\theta$, we *recover* our original deviance in Equation 8 and Equation 10.

For **relative entropy**, we consider the **Rényi entropy** defined (Leinster, 2024) as:

$$H_q(\mathbf{p} \,\|\, \mathbf{r}) = \frac{1}{q-1} \log \sum_{i \in \operatorname{supp}(\mathbf{p})} p_i^q \, r_i^{1-q} \tag{A.6}$$

Then, we can think of a *dummy* $\mathsf{orient}()$ that just stores Λ_x, Λ_y and a $\|\cdot\|_{\mathsf{orient}}$ operator that computes the *relative entropy* between them. For a given *normalized* core $\langle \overline{\Lambda}_{\text{norm}_n} \rangle = \{\langle \overline{\alpha}_{\text{norm}_i} \rangle, ...\}$ and *normalized* system $\overline{\Lambda}_{\text{norm}_n} = \{\overline{\alpha}_{\text{norm}_i}, ...\}$, we define two *Hill number* (Leinster, 2024) deviances: the ***excess* deviance** and ***deficit* deviance**:

$$\partial_q^+(y, x_p) = e^{H_q(\overline{\Lambda}_{\text{norm}_n}(y) \,\|\, \langle \overline{\Lambda}_{\text{norm}_n} \rangle(x_p))} \tag{A.7}$$

$$\partial_q^-(y, x_p) = e^{H_q(\langle \overline{\Lambda}_{\text{norm}_n} \rangle(x_p) \,\|\, \overline{\Lambda}_{\text{norm}_n}(y))} \tag{A.8}$$

We could read ∂_q^+ as the *effective* **over-compliance** and ∂_q^- as the *effective* **under-compliance** with respect to the *normative compliance*.

For instance, as $q \to \infty$, we interpret:

- ∂_∞^+ as the largest *excess of compliance*

$$\partial_\infty^+ = \max_i \frac{\overline{\alpha}_{\text{norm}_i}(y)}{\langle \overline{\alpha}_{\text{norm}_i} \rangle(x_p)}$$

- ∂_∞^- as the largest *deficit of compliance*

$$\partial_\infty^- = \max_i \frac{\langle \overline{\alpha}_{\text{norm}_i} \rangle(x_p)}{\overline{\alpha}_{\text{norm}_i}(y)}$$

All of this to say, there are **multiple ways we can reason about *structures* and *statistics* jointly**. We encourage readers to develop alternative and competing formalisms that share our conceptual backbone: *structures* that make *context* explicit, *cores* that encode the normativity that *homogenization* pushes us toward, and *orientations* that capture perspectives of *non-normativity*. Above all, **we ask everyone to think deeper about diversity**.

Appendix B. Theoretical touchpoints

In this appendix, we explore how our theoretical framework connects to other frameworks. To that purpose, we consider an *unprompted* scenario of a *singleton* system with *binary* compliance for its single structure:

$$\Lambda_*(x) := (\alpha_*(x)) \qquad \alpha_*(x) \in \{0,1\}$$

Then, the structure core represents the probability of compliance being exactly 1:

$$\mu := \langle \alpha_* \rangle = \sum_{c \in \{0,1\}} c \ \Pr(\alpha = c) = \Pr(\alpha = 1)$$

Our singleton deviance is expressed as:

$$\partial_*(x) = \|\alpha_*(x) - \mu\|_\theta$$

B.1. Expected deviance and Gini-Simpson index

To calculate the *expected deviance*, we consider two choices for $\|\cdot\|_\theta$: absolute value and the squared ℓ_2 norm. For each, we find connections between $\mathbb{E}[\partial_*]$ and the *Gini-Simpson index* for a binary variable:

$$\mathbb{E}[|\alpha_* - \mu|] = 2\mu(1-\mu) = \mathrm{GS}$$

$$\mathbb{E}[\|\alpha_* - \mu\|_2^2] = \mathrm{Var}[\alpha_*] = \mu(1-\mu) = \frac{\mathrm{GS}}{2}$$

If we interpret GS as the *degree of mixing* in outcomes, then increasing the expected deviance drives *heterogeneity* rather than *concentration*.

B.2. Is-It-Valid classification for Hallucinations

To reason about hallucinations, authors in (Kalai et al., 2025) partition the space of *plausible* outputs into disjoint sets of *valid outputs* V and *errors* E. In their framework, a model *hallucinates* when it cannot solve the binary discrimination problem *Is-It-Valid?* (IIV). Their framework can be interpreted through our structure-aware language:

$$\alpha_{\mathrm{IIV}}(x) = \mathbf{1}[x \in V]$$

We can connect their generative hallucination rate given by $\mathrm{err} = \Pr_{x\sim\hat{p}}[x \in E] = \hat{p}(E)$ to the system core of a singleton IIV system:

$$\langle \alpha_{\mathrm{IIV}} \rangle = 1 - \mathrm{err}$$

The paper (Kalai et al., 2025) points out that future work should "consider degrees of hallucination". Our structure-aware framework provides the language to reason about these desired **graded notions of hallucination**: We can score a string under multiple structures, with scores encoding real-valued nuance *beyond the binary*.

B.3. Language Generation in the Limit

Recent work (Kleinberg & Mullainathan, 2024; Kalavasis et al., 2025a) studies language generation where a generator G, given strings from an unknown target language K, must output strings that are both **novel** and **valid**. We can reinterpret some of their framework as a special case of our structure-aware formulation.

Given a language collection $\mathcal{L} = \{L_1, L_2, \ldots\}$, we can define *membership structures* with corresponding cores that represent the probability of generating a string valid for each corresponding language:

$$\alpha_{L_i}(x) = \mathbf{1}[x \in L_i] \qquad \langle \alpha_{L_i} \rangle = \Pr[y \in L_i]$$

The literature is currently (Kalavasis et al., 2025b) exploring the trade-offs between *consistency* and *breadth*. An LLM generates strings *consistent* with our target language K if:

$$\langle \alpha_K \rangle = 1 \qquad \text{when} \qquad \mathbb{E}[\partial_K]_{y\sim p_{\mathsf{LLM}}} \to 0$$

An LLM generation has *breadth* when all strings of our target language $K \in \mathcal{L}$ can be generated:

$$\forall\, y \in K:\ p_{\mathsf{LLM}}(y) > 0 \iff K \subseteq \mathrm{supp}(p_{\mathsf{LLM}})$$

Our structure-aware framework gives us insight that homogenization is *relative to a system*. Indeed, pushing for consistency shall not imply that we push for homogenization in every context. Generally, for $\Lambda_K \neq \Lambda_m$:

$$\mathbb{E}[\partial_K] \to 0 \quad \neq \quad \mathbb{E}[\partial_m] \to 0$$

Thinking explicitly through structures and systems allows us to formulate *interesting* questions (for instance, is $\Lambda_K = \Lambda_{\mathrm{IIV}}$?) that will help us make connections between all these theoretical efforts. We present these touchpoints as **starting points for deeper exploration**.

Appendix C. Trade-off between diversity and fairness

In this appendix, we show the fundamental tension between diversity and fairness by proving the existence of a *Pareto* trade-off between them. A Pareto trade-off says that among *efficient solutions*, improvement on one criterion *necessarily worsens* the other (Ehrgott, 2005). To establish such a trade-off, it suffices to exhibit two efficient solutions, each of which is better than the other on a different criterion. This shows that no single solution can *dominate* both.

Taking into account ρ_d and ρ_f defined as in Equation 17 and Equation 21 with a λ-weights and β-tunable-parameter set to 1.0 :

Definition C.1 (Pareto Dominance)**.** An intervention w *Pareto-dominates* intervention w' if $\rho_d(w) \geq \rho_d(w')$ and $\rho_f(w) \geq \rho_f(w')$ with at least one strict inequality.

Theorem C.2 (Trade-off Between Diversity and Fairness)**.** *Let $n \geq 2$ and let w_0 induce a non-uniform baseline core with $\langle \alpha_n \rangle(w_0) < 1/n < \langle \alpha_1 \rangle(w_0)$. Then there exist interventions w_d, w_f such that neither Pareto-dominates the other:*

$$\rho_d(w_d) > \rho_d(w_f) \quad \textit{and} \quad \rho_f(w_d) < \rho_f(w_f) \tag{C.1}$$

This demonstrates the existence of a fundamental trade-off *between the two criteria.*

Proof. We construct two interventions that exhibit opposite strengths.

For simplicity, we assume ***deterministic* generation**; adding stochasticity would only increase $\mathsf{score}_{\mathsf{diverge}}$ and strengthen the trade-off [15]. Proving our trade-off requires just one counterexample to dominance, so our deterministic pair suffices.

We also say w_0 induces a non-uniform system core $\langle \Lambda_n \rangle(w_0) = (\mu_1, \ldots, \mu_n)$ and assume that $\| \cdot \|_\theta$ is the Euclidean norm $\| \cdot \|_2$

Let w_d induce $\langle \Lambda_n \rangle(w_d) = (0, \ldots, 0, 1)$. Then:

$$\begin{aligned}
\mathsf{score}_{\mathsf{explore}}(w_d) &= \sqrt{(1-\mu_n)^2 + \sum_{i=1}^{n-1} (\mu_i)^2} \; > 0 \\
\mathsf{score}_{\mathsf{diverge}}(w_d) &= 0 \quad \text{(deterministic)} \\
\mathsf{score}_{\mathsf{even}}(w_d) &= H(0, \ldots, 0, 1) = 0 \\
\mathsf{score}_{\mathsf{inverted}}(w_d) &= 1 \quad \text{(all pairwise orderings change)}
\end{aligned}$$

Thus:

$$\rho_d(w_d) = \mathsf{score}_{\mathsf{explore}}(w_d) > 0 \qquad \rho_f(w_d) = 1$$

Let w_f induce $\langle \Lambda_n \rangle(w_f) = (1/n, \ldots, 1/n)$. Then:

$$\begin{aligned}
\mathsf{score}_{\mathsf{explore}}(w_f) &= \sqrt{\sum_{i=1}^{n} (1/n - \mu_i)^2} \; > 0 \\
\mathsf{score}_{\mathsf{diverge}}(w_f) &= 0 \quad \text{(deterministic)} \\
\mathsf{score}_{\mathsf{even}}(w_f) &= H(1/n, \ldots, 1/n) = \log n \\
\mathsf{score}_{\mathsf{inverted}}(w_f) &= 1 \quad \text{(all orderings flatten)}
\end{aligned}$$

Thus:

$$\rho_d(w_f) = \mathsf{score}_{\mathsf{explore}}(w_f) > 0 \qquad \rho_f(w_f) = 1 + \log n$$

We now check if $\mathsf{score}_{\mathsf{explore}}(w_d) > \mathsf{score}_{\mathsf{explore}}(w_f)$. We assume the inequality and verify if it holds:

$$\begin{aligned}
\text{assuming :} \quad & \mathsf{score}_{\mathsf{explore}}(w_d) > \mathsf{score}_{\mathsf{explore}}(w_f) \\
\text{squaring :} \quad & (1-\mu_n)^2 + \sum_{i=1}^{n-1} (\mu_i)^2 > \sum_{i=1}^{n} (1/n - \mu_i)^2 \\
\text{rearranging :} \quad & \sum_{i=1}^{n-1} \mu_i > (n-1)\left(\mu_n - \tfrac{1}{2}\right)
\end{aligned}$$

Since $\mu_n < 1/n \leq 1/2$ for $n \geq 2$, the right-hand side is negative. Since $\mu_1 > 1/n > 0$ and $\mu_i \in [0, 1]$, the left-hand side is positive. Thus, $\mathsf{score}_{\mathsf{explore}}(w_d) > \mathsf{score}_{\mathsf{explore}}(w_f)$ holds. Since $\mathsf{score}_{\mathsf{diverge}}(w_0) = \mathsf{score}_{\mathsf{diverge}}(w_f) = 0$, we conclude $\rho_d(w_d) > \rho_d(w_f)$.

Noting $1 < 1 + \log n$ for $n \geq 2$, we also conclude $\rho_f(w_d) < \rho_f(w_f)$.

Since $\rho_d(w_d) > \rho_d(w_f)$ and $\rho_f(w_d) < \rho_f(w_f)$, neither intervention dominates the other. □

Corollary C.3 (Weight choice encodes value judgment)**.** *The choice of weights (λ_d, λ_f) in the combined score $\rho_\chi = \lambda_d \rho_d + \lambda_f \rho_f$ reflects an irreducible value judgment about the relative priority of diversity versus fairness.*

This appendix shows that our framework recovers an expected *tension between desiderata*, validating the *expressiveness* of our vocabulary. Future work will further characterize this and other trade-offs.

[15] For deterministic generation, the core equals the sole trajectory's compliance, so both the expected deviance and its variance are zero.

Appendix D. Comparing with linguistic metrics of diversity

In this appendix, we place our structure-aware framework in the context of existing diversity metrics. There are two main categories of metrics of linguistic diversity: *intrinsic* and *extrinsic*.

D.1. Intrinsic linguistic diversity

Intrinsic diversity refers to the types of variation *within* a generated language without external references. The literature accounts (Guo et al., 2025; Tevet & Berant, 2021) for intrinsic diversity in both *form* and *content*.

D.1.1. Form Diversity

On the one hand, we have **syntactic** diversity, which accounts for the variety in *sentence patterns*. These metrics involve *identifying* patterns in sentences and subsequently *comparing* them based on their occurrence across the generated language. Some of the methods include: transforming sentences into part-of-speech (POS) tag sequences and evaluating redundancy through compression (Shaib et al., 2024b); and parsing text into trees and mapping the resulting graphs into a vector space (Guo et al., 2024b) or measuring their distribution (del Prado Martin, 2024).

Our framework naturally includes syntactic metrics as we can define *syntactic systems* in which each structure encodes a syntactic pattern of interest:

$$\Lambda_{\text{syntax}} = (\alpha_{\text{POS Tag}}, \alpha_{\text{Noun Phrase}}, \ldots) \tag{D.1}$$

On the other hand, we have **lexical** diversity, which accounts for the variety in *vocabulary*. Generally, these *surface-level* metrics measure *repetition* and *reuse* (Kendro et al., 2025; Shaib et al., 2024a). Common approaches include counting unique n-grams (Li et al., 2016; Estève et al., 2025), measuring their overlap (Zhu et al., 2018), and calculating their entropy (Estève et al., 2025).

Our framework can account for lexical metrics in an analogous way that it accounts for syntax metrics. For instance, we could define *lexical systems* in which each structure encodes a unique n-gram. However, this might not be very *practical*, as the number of possible n-grams grows exponentially with vocabulary size (Jurafsky & Martin, 2009).

$$\Lambda_{\text{lexicon}} = (\alpha_{1-\text{gram}}, \ldots) \tag{D.2}$$

D.1.2. Content Diversity

To measure **semantic** diversity, sentences are transformed into *embeddings* that can be used to measure *similarity* (Guo et al., 2025). The literature describes various ways to exploit these similarities, from calculating the effective number of unique elements in the sample through the eigenvalues of the similarity matrix (Friedman & Dieng, 2023), to quantifying *the divergence* in the intermediate reasoning steps taken by LLMs to find a solution (Ju et al., 2025).

In our framework, we can construct *semantic systems* in which the compliance of each structure is a measure of similarity to an *internal reference* embedding [16]. For instance, consider the following system:

$$\Lambda_{\text{semantics}} = (\alpha_{v_1}, \ldots) \tag{D.3}$$

where each structure computes the dot product between the embedding vector of the string x and the unit vector v_i

$$\alpha_{v_i}(x) = \text{abs}(\text{embed}(x) \cdot v_i) \tag{D.4}$$

In this example, comparing the system compliance of two different strings ($\Lambda_{\text{semantics}}(x_a)$ vs $\Lambda_{\text{semantics}}(x_b)$) affords more *interpretable* measures of similarity and difference since we can decompose the results per v_i internal reference.

D.2. Extrinsic linguistic diversity

Extrinsic diversity metrics focus on *divergence* between a target (LLM-generated language) and an ***external reference***, which could be text samples or real human language distributions (Pillutla et al., 2023). Comparisons between target and reference use the same methods (syntactic, lexical, and semantic) as those used for intrinsic diversity.

Our framework provides a language for reasoning about the target and reference through *multiple lenses*. For example, we could ask questions that compare the target and the reference, like:

Are the same syntactic patterns present on average?

$$\|\langle\Lambda^{target}_{\text{syntax}}\rangle - \langle\Lambda^{reference}_{\text{syntax}}\rangle\|_\theta$$

Do they have the same range of semantic variety?

$$\mathbb{E}[\partial^{target}_{\text{semantics}}] \quad vs. \quad \mathbb{E}[\partial^{reference}_{\text{semantics}}]$$

Is toxic language equally likely after prefacing a text with "Be brutally honest." ?

$$\langle\alpha^{target}_{\text{toxic}}\rangle(x_p) \quad vs. \quad \langle\alpha^{reference}_{\text{toxic}}\rangle(x_p)$$

with $x_p =$ "Be brutally honest."

Do they have the same ratio of syntactic and lexical diversity?

$$H(\langle\Lambda^{target}_{\text{Form}}\rangle) \quad vs. \quad H(\langle\Lambda^{reference}_{\text{Form}}\rangle)$$

with $\Lambda_{\text{Form}} = [\Lambda_{\text{syntax}}, \Lambda_{\text{lexicon}}]$

[16] These internal reference vectors might be the principal components of the learned embedding space, known concept vectors of interest, embeddings of prototypical sentences, and so on.

Appendix E. Dynamics of relative diversity

As noted in Section 3.5, what is *non-normative* is *conditional on what came before*. Then, as a string is being completed, the set of possible trajectories is narrowed so the system core and orientations change. Trajectories that were essentially *unreachable* from the root of the tree may emerge as *attractors* once we condition on a specific *subtree*.

Given a trajectory $y = x_T$, for $k \in \{0, 1, \dots, T\}$, we can define **states** for all the *intermediate continuations*, such as:

$$\phi_k^{(x)} = \langle \Lambda_n \rangle (x_k) \quad \phi_k^{(y)} = \theta_n(x_k | x_0) \quad \phi_k^{(z)} = \theta_n(y | x_k) \quad \text{(E.1)}$$

which form a discrete-time **dynamics**:

$$(\phi_0^{(x)}, \phi_0^{(y)}, \phi_0^{(z)}) \to \dots \to (\phi_T^{(x)}, \phi_T^{(y)}, \phi_T^{(z)})$$

The state $\phi^{(x)}$ evolves from representing the expected system compliance of all possible continuations at $\phi_0^{(x)} = \langle \Lambda_n \rangle (\perp)$, to the specific system compliance of a given trajectory at $\phi_T^{(x)} = \langle \Lambda_n \rangle (y) = \Lambda_n(y)$.

The state $\phi^{(y)}$ encodes how much the current path has *deviated* from normativity, evolving from $\phi_0^{(y)} = \theta_n(\perp | \perp) = \Lambda_n(\perp) - \langle \Lambda_n \rangle (\perp)$ to the full trajectory's orientation in the largest frame of reference at $\phi_T^{(y)} = \theta_n(y | \perp) = \Lambda_n(y) - \langle \Lambda_n \rangle (\perp)$.

The state $\phi^{(z)}$ evolves from representing how *deviant* the trajectory is in the largest frame of reference at $\phi_0^{(z)} = \theta_n(y | \perp) = \phi_T^{(y)} = \Lambda_n(y) - \langle \Lambda_n \rangle (\perp)$, to a *zero deviance*[17] at $\phi_T^{(z)} = \theta_n(y | y) = 0$.

[17] A zero deviance is when an orientation has a deviation value of zero for all structures.