Epistemic Diversity and Knowledge Collapse in Large Language Models

Dustin Wright[♯] **Sarah Masud**[♯] **Jared Moore**[♭] **Srishti Yadav**[♯] **Maria Antoniak**[♮]
Peter Ebert Christensen[♯] **Chan Young Park**[×] **Isabelle Augenstein**[♯]
[♯]University of Copenhagen [♭]Stanford University
[♮]University of Colorado Boulder [×]Microsoft Research
Correspondence: dw@di.ku.dk

Abstract

Large language models (LLMs) tend to generate lexically, semantically, and stylistically homogenous texts. This poses a risk of *knowledge collapse*, where homogenous LLMs mediate a shrinking in the range of accessible information over time. Existing works on homogenization are limited by a focus on closed-ended multiple-choice setups or fuzzy semantic features, and do not look at trends across time and cultural contexts. To overcome this, we present a new methodology to measure epistemic diversity, i.e., variation in real-world claims in LLM outputs, which we use to perform a broad empirical study of LLM knowledge collapse. We test 27 LLMs, 155 topics covering 12 countries, and 200 prompt variations sourced from real user chats. For the topics in our study, we show that while newer models tend to generate more diverse claims, nearly all models are less epistemically diverse than a basic web search. We find that model size has a negative impact on epistemic diversity, while retrieval-augmented generation (RAG) has a positive impact, though the improvement from RAG varies by the cultural context. Finally, compared to a traditional knowledge source (Wikipedia), we find that country-specific claims reflect the English language more than the local one, highlighting a gap in epistemic representation.[1]

1 Introduction

Large language models (LLMs) are being adopted for knowledge-intensive tasks such as summarization (Wright et al., 2025), writing assistance (Sun et al., 2025), and research (Si et al., 2025). Search interfaces are now prioritizing "AI Overviews" to answer queries. It is speculated that people will soon access most information through an LLM intermediary (Peterson, 2025).

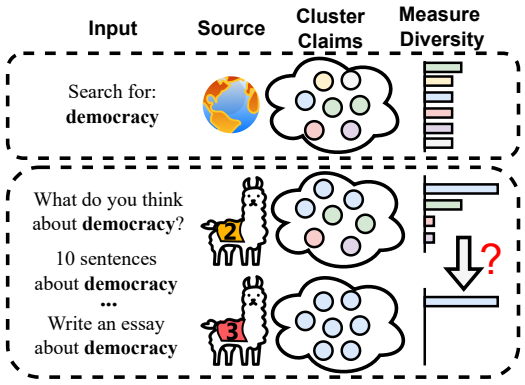

Figure 1: In this work, we measure epistemic diversity – via variability in claims about the world – for characterizing knowledge collapse in LLMs.

At the same time, recent studies have noted that LLM outputs are homogeneous (Sourati et al., 2025b). For example, LLMs reflect only a narrow range of writing and reasoning styles, use a limited vocabulary (Sourati et al., 2025a) and convey only certain semantics (Lee et al., 2025). This phenomenon, where models regress to a "central tendency", also affects knowledge i.e., the pieces of information that models tend to generate (Dutta and Chakraborty, 2023). This could in turn limit the information available to the general public and subsequently dwindle our collective knowledge, a phenomenon broadly defined as **knowledge collapse** (Fig. 1).[2] Recent work (Peterson, 2025) has begun to theorize this, but offers limited empirical data. To better characterize and understand this problem, we investigate, for the first time, whether LLMs themselves exhibit knowledge collapse.

To do so, we perform an empirical study measuring **epistemic diversity**, defined as the diver-

[1]Code and data: https://github.com/dwright37/llm-knowledge

[2]For an in depth definition and discussion, see Peterson (2025).

sity of claims about the world in a given corpus (e.g., a set of LLM responses). To measure this, we develop a new methodology intended to reveal the diversity of claims occurring in free-text LLM outputs. This involves (1) sampling outputs from LLMs using a set of 200 natural writing assistance prompts collected from Röttger et al. (2025), (2) partitioning the LLM responses into unique classes of semantically equivalent claims (Farquhar et al., 2024), and (3) quantifying the diversity of the sample of claims with Hill-Shannon diversity, a widely used metric for measuring species diversity in ecology (Roswell et al., 2021). From among the Llama, Gemma, Qwen, and OpenAI model families, we study 27 LLMs spanning multiple versions, sizes, and release dates, querying them for 155 topics.

For the selected topics, while we find that epistemic diversity has increased since 2023 for three of four model families, it remains low for all models compared to a rudimentary web search. Similar to contemporary work (Zhang et al., 2025), we find that model size has a statistically significant *negative* impact on epistemic diversity – smaller models generate more diverse knowledge than larger ones. In contrast, retrieval-augmented generation (RAG) has a statistically significant *positive* impact, highlighting the importance of RAG in preventing future knowledge collapse. For country-specific topics, we find that RAG has an uneven effect; certain countries (e.g., the USA) see more benefit due to a greater diversity in their RAG sources. Finally, compared to a traditional knowledge source (Wikipedia) in both English and local languages for country-specific topics, we find that the claims generated in our study reflect English language knowledge more than local language knowledge, highlighting a gap in epistemic representation.

2 Background

LLMs are increasingly being used for knowledge-centric tasks (Yang et al., 2024), and can influence people's behavior (Anderson et al., 2024; Jakesch et al., 2023; Bai et al., 2025a). Hence, a lack of diversity in LLMs' outputs may reduce the diversity in our collective knowledge. Our work studies this risk through the lenses of LLM homogenization and knowledge collapse.

LLM Homogenization An increasing body of work shows that LLMs suffer from a lack of diversity along many dimensions; for a recent survey, see Sourati et al. 2025b. This lack of diversity includes lexical and stylistic (Sourati et al., 2025a), semantic (Lee et al., 2025), creative (Xu et al., 2025), and perspective diversity (Wright et al., 2024; Abdurahman et al., 2024; Zhang et al., 2025; Durmus et al., 2023; Röttger et al., 2024; Moore et al., 2024). Our work is most similar to perspective diversity, which has shown that LLM-generated views, opinions, and beliefs tend to reflect only a small subset of the world (Durmus et al., 2023; Atari et al.; Abdurahman et al., 2024; Alvero et al., 2024). Perspective diversity is usually measured with multiple-choice survey responses (Durmus et al., 2023) or free-text responses partitioned into different classes (Zhang et al., 2025). In our work we measure the epistemic diversity of models by partitioning the LLM output space into clusters of *claims*. This allows us to avoid shallow and/or fuzzy features such as semantic and lexical similarity, and improves upon unreliable multiple-choice setups (Röttger et al., 2024).

Knowledge Collapse LLMs can exacerbate their own biases, leading to *model collapse* (Shumailov et al., 2024). A growing concern among scholars is that such homogenization, combined with increased adoption of LLMs, will lead to epistemic problems at a societal level (Zheng and Lee, 2023; Messeri and Crockett, 2024; Peterson, 2025; Wagner and Jiang, 2025; Qiu et al., 2025). Peterson (2025) defines "knowledge collapse" as LLMs facilitating a dwindling of knowledge into an increasingly narrow set of ideas. Knowledge collapse may affect existing knowledge sources such as Wikipedia (Wagner and Jiang, 2025), erase minoritized knowledge (Zheng and Lee, 2023), pollute scientific discoveries (Messeri and Crockett, 2024), hamper political discourse (Coeckelbergh, 2025), and limit ideation in writing (Anderson et al., 2024). These concerns echo other contemporary epistemic issues such as "popularity bias" (Ciampaglia et al., 2018), and "filter bubbles" (Nguyen et al., 2014) occurring with the use of recommender systems. To assess the risk of knowledge collapse, our work offers a new methodology for measuring epistemic diversity and an empirical study of this phenomenon in LLMs.

3 Problem Setup and Notation

We measure the epistemic diversity of LLMs as the variation in the claims that the models make across different topics and prompts. To do so, we develop a methodology that is naturalistic and the-

oretically motivated. It is naturalistic in that it uses natural prompts to sample open-ended responses (Wright et al., 2024; Moore et al., 2024; Röttger et al., 2025). It is theoretically motivated in that it adopts a statistically grounded measure of diversity used widely in ecology for measuring species diversity (Peterson, 2025; Roswell et al., 2021). We measure epistemic diversity after transforming open-ended LLM responses into distributions over meaning classes (Farquhar et al., 2024).

To describe our approach (Fig. 1), we adopt the following notation. First, a corpus of text P_{m_t} is elicited about a topic t from a model m. P_{m_t} contains free text which can be decomposed into a list of n claims C_{m_t}, which can be further partitioned into a set of unique *meaning classes* $\mathcal{X}_{m_t}$. A unique meaning class is a cluster where all claims within a given class mutually entail each other and do not mutually entail claims in other classes. Then, $x_i \in \mathcal{X}_{m_t}$ is defined as the empirical frequency of meaning class i calculated from C_{m_t}, and p_i is the probability of i, calculated as the relative frequency of i in the sample, i.e., $\frac{x_i}{n}$.

Why not semantic similarity? An alternative version of our setup would be to partition claims based solely on semantic similarity, as previous work has done (Lee et al., 2025; Wright et al., 2024). However, sentence embeddings (Reimers and Gurevych, 2019) can only measure equivalence based on functionally similar phrases. For example, the phrases "Claude Shannon is the father of information theory" and "Claude Shannon is not the father of information theory" have extremely high semantic similarity (0.94)[3], even though they present opposing claims. Similarly, "Claude Shannon is the father of quantum theory" is an entirely different claim from "Claude Shannon is the father of information theory," but these two also have high semantic similarity (0.805). This setup would not reflect our definition of meaning classes.

4 Data Collection

We first construct $\mathcal{X}_{m_t}$ for a particular model m and topic t (for details about the specific models and topics we study see § 6). Acquiring $\mathcal{X}_{m_t}$ involves a three step process similar to Wright et al. (2024) and Zhang et al. (2025) but focused on claims:

1. **Generate**: Acquire a corpus P_{m_t} of free-text LLM responses to natural input prompts.
2. **Decompose**: Decompose P_{m_t} into a list of atomic claims C_{m_t}.
3. **Cluster**: Group the list of atomic claims into meaning classes $\mathcal{X}_{m_t}$.

4.1 Generation and Decomposition

We use the open-ended writing assistance prompt templates collected in Röttger et al. (2025). The original prompt templates are sourced from WildChat (Zhao et al., 2024), a collection of natural conversations between users and ChatGPT. From the original 1,000 writing assistance prompts, we manually select a subset of 479 prompts that primarily focus on information seeking and informational writing. We manually filter out NSFW templates, creative writing templates such as "write a 50's soviet style song about t," outlines such as "write an index for a book on t," or those explicitly asking for references. We then randomly sample 200 of these prompt templates for generation.

After generating responses (P_{m_t}), we **decompose** each response into a list of individual claims C_{m_t}. We do so using a strong open-weight LLM,[4] prompting the model to decompose non-overlapping chunks of three sentences at a time. This allows us balance claim recall while with ensuring that each input chunk is decontextualized.

Evaluation To evaluate the quality of decomposition, we develop three initial decomposition prompts (P1-P3) and have two independent annotators label (1) the *quality* of individual decomposed claims on a Likert scale from 1-5, and (2) how many claims in the original input chunk are *missing* from the list of decomposed claims. (On our scale, 1 means that the decomposed claim is not inferable from the original text, and 5 means that the claim is fully inferable and decontextualized.) The full annotation instructions appear in the Appendix § A.1. We label 100 input chunks for quality (1) and 108 instances for missing (2), all anonymized for the prompt type (P1-P3). Between the two annotators, we achieve a Kendall-Tau (τ) correlation (Kendall, 1938) of 0.53 for (1) and 0.39 for (2), indicating moderate to strong agreement.[5] We break ties by having a third annotator label those instances where the original annotators disagreed.

We then set up an LLM-as-a-judge using G-Eval (Liu et al., 2023), which achieves 0.6 Pearson

[3] Huggingface ID: `all-MiniLM-L6-v2`

[4] Huggingface ID: meta-llama/Llama-3.1-70B-Instruct

[5] We select τ because it is suited for ordinal data with ranking ties (Kendall, 1945).

	P1	P2	P3
Quality↑	4.58	4.67	**4.69**
Missing Claims↓	**0.06**	0.24	0.25

Table 1: Performance of different prompts for claim decomposition using LLM-as-a-judge (§ 4.1). "Quality" measures how well the decomposed claims align with the original chunk on a scale from 1 to 5. "Missing Claims" measures the average number of claims missed during the decomposition. We use P3 for our final decomposition due to its higher quality and minimal missing claims.

correlation with the human labels for decomposition quality and 0.68 for the number of missing claims. We use ths to automatically label 6k instances across three decomposition prompt variants (P1-P3; see Appendix § A.2 for prompt text). The results are given in Table 1. We use P3 for the final decomposition in order to prioritize quality.

4.2 Clustering

Algorithm 1 outlines our approach for **clustering** the decomposed claims C_{m_t} into meaning classes $\mathcal{X}_{m_t}$ Following Farquhar et al. (2024), we bootstrap clusters based on mutual entailment using a strong pretrained model for natural language inference.[6] This is done by assuming that when a given claim mutually entails at least one claim in an existing cluster, it also entails all other claims in the cluster. To reduce measuring mutual entailment across n^2 pairs for a set of n claims, we only check entailment between the N most similar claims (i.e., $N * n$ comparisons), where similarity is measured using a strong S-BERT model.[7] To mitigate potential drift in cluster cohesion (i.e., large clusters containing multiple meaning classes), we perform a post-processing step to break up large clusters using DBSCAN (Ester et al., 1996), a related clustering approach (Peterson, 2025; Wright et al., 2024).

Evaluation Two independent annotators label: (1) the degree of *cohesion* among the clustered claims (Likert score between 1 and 5), and (2) whether semantically similar, non-clustered claims are grouped with a given cluster (binary yes/no). See Appendix § A.1 for full instructions. We label 100 instances of cohesion (1) and 360 instances of missing (2), achieving a Kendall-Tau correlation of 0.38 for (1) and 87% agreement accuracy for (2). We again break ties using a third annotator and set up an LLM-as-a-judge using G-Eval. We develop prompts that achieve 0.68 Pearson r with the human labels for measuring cluster cohesion and 83.5 weighted F1 score for measuring binary missing claims. 15,000 samples of 3-5 sentences are then automatically labeled from clusters acquired using Algorithm 1, where semantic similarity for task (2) is measured using S-BERT. We acquire a cluster cohesion score of **4.08** and a missing rate of **13.79** (out of 100), indicating that the cluster cohesion is generally of high quality while missing only a relatively small proportion of singletons.

Algorithm 1: Clustering algorithm

input : Decomposed corpus C with claims c_j;
Max retrieval length N

```
# Previous claims
A ← [C[0]];
# Claim cluster
L ← [0];
forall c_j ∈ C[1 :] do
    # Sort by cos similarity
    candidates ← most_similar(c_j, A)[:N];
    scores = [];
    forall c_k ∈ candidates do
        if entails(c_k, c_j) & entails(c_j, c_k) then
            scores.append(entail_prob(c_k, c_j) *
              entail_prob(c_j, c_k));
    if len(scores) == 0 then
        i ← max(L) + 1;
    else
        i ← cluster(max(scores));
    L.append(i);
    A.append(c_j);
# Count the number of elements in each
 unique cluster
X ← count(L);
return X, L;
```

5 Measuring Epistemic Diversity

After obtaining $\mathcal{X}_{m_t}$, we must quantify cluster diversity to rank different models. We could simply count the number of clusters for each model, but that would discount the distribution of the clusters (e.g., for one model, there might be a long tail of clusters with only one claim in each cluster, which should rank lower compared to a model with fewer clusters that each contain many claims). Conversely, we could turn these counts into probabili-

[6] Huggingface ID: microsoft/deberta-large-mnli
[7] Huggingface ID: all-MiniLM-L6-v2

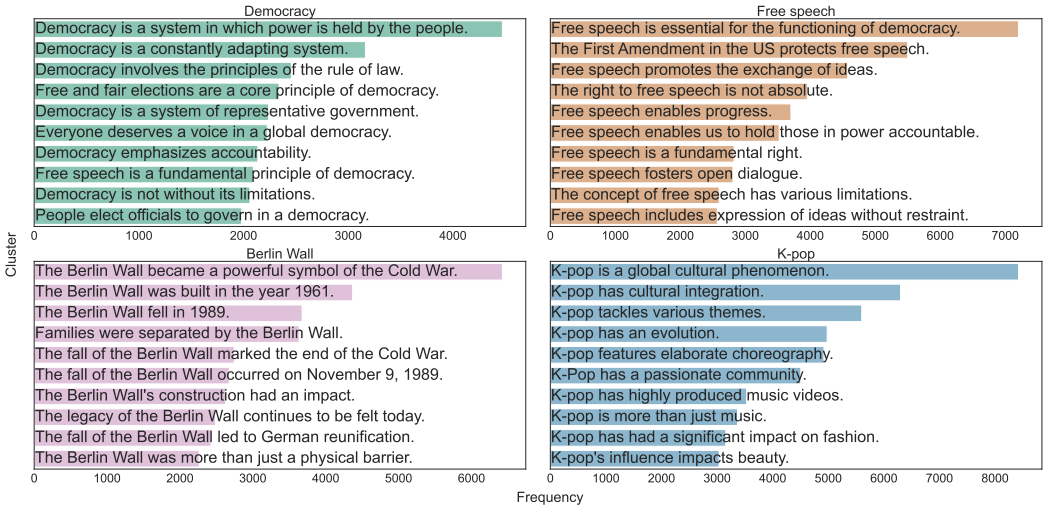

Figure 2: Histograms of the top ten clusters for four topics after generating text, decomposing, and clustering decomposed claims across all models in our study. The frequency of claims in each cluster, x_i, is represented by the colored bars. By the 10th cluster, x_i is halved for all four topics, indicating a large decay rate for x_i. The top clusters for each topic convey broad and general information for each topic.

ties, obtaining a categorical distribution over clusters, and then calculate entropy (Peterson, 2025), but then the change in total number of clusters across topics would not scale properly (e.g., compared to another setting that has half as many clusters, the diversity metric would only lower by a maximum of one point).

We therefore use Hill diversity (Hill, 1973; Jost, 2006) as our diversity index. Hill diversity offers a general way to measure the relative abundance of different categories of items in a sample, and satisfies the "replication principle" where relative changes in categories and abundances between samples result in proportionate changes to the score (Roswell et al., 2021). Using the notation from § 3, we calculate the general Hill diversity as

$$D(\mathcal{X}_{m_t}) = \left(\sum_{i}^{\mathcal{X}_{m_t}} p_i (\frac{1}{p_i})^l \right)^{1/l},$$

where low values of the free parameter l will provide more weight to more frequent claims in the sample, while higher values give more emphasis to rare classes. To balance the emphasis between common and rare claims, we choose $l = 0$, a.k.a the Hill-Shannon diversity (HSD), which resolves to

$$D_S(\mathcal{X}_{m_t}) = \exp\{-\sum_{i}^{\mathcal{X}_{m_t}} p_i \ln p_i\}, \quad (1)$$

i.e., e raised to the entropy in nats.

Note that the way one samples C_{m_t} has a large impact on D_S. This can occur despite "equal-effort" sampling, e.g., prompting all models with the same number of prompts, because the long tail of claims may vastly differ between settings. To illustrate: consider two models, one with low diversity (a) and one with high diversity (b), from which we sample an equal number of model responses. Assume both models also produce roughly the same number of claims, n, after decomposition; a is less diverse, so n samples may be sufficient to fully characterize $\mathcal{X}_{a_t}$, and sampling more claims is unlikely to introduce new meaning classes or have a substantial impact on $D_S(\mathcal{X}_{a_t})$. On the other hand, the introduction of new claims from the more diverse model b may introduce new meaning classes to $\mathcal{X}_{b_t}$ which take longer to uncover due to their rarity, thus increasing $D_S(\mathcal{X}_{b_t})$. Therefore, by sampling $\approx n$ claims for both models, we have potentially overestimated the diversity of a relative to b.

To account for this, we follow best practices by using a combination of coverage estimation and rarefaction as outlined in Chao and Jost (2012) and Roswell et al. (2021). Coverage allows us to estimate how well our sample captures the true diversity of a given setting, i.e., how completely we have performed our sampling. It is calculated using the estimator from Chao and Jost (2012):

$$V(\mathcal{X}_{m_t}) = 1 - \frac{f_1}{n} \left[\frac{(n-1)f_1}{(n-1)f_1 + 2f_2} \right], \quad (2)$$

where n is the number of claims, f_1 is the number of singleton classes in $\mathcal{X}_{m_t}$ (i.e., the number of meaning classes i where $x_i = 1$), and f_2 is the number of doubleton classes (i.e., the number of meaning classes i where $x_i = 2$). When we wish to compare the diversity of models on a topic t, we use Equation 2 to calculate the coverage of all models and rarefy each set of claims (i.e., downsample) to the minimum coverage achieved across models.

6 Empirical Results

To what extent are LLMs exhibiting knowledge collapse? To answer, we pose several subquestions focused on how epistemic diversity has changed over time **(RQ1)**, how epistemic diversity is impacted by generation setting **(RQ2)**, how size and model family impact epistemic diversity **(RQ3)**, and how cultural context impacts both diversity and representation **(RQ4)**.

6.1 Models, Topics, and Settings

Models We select 27 LLMs across 4 model families. This includes Llama (versions 2, 3.1, 3.2, and 3.3), Gemma (versions 1, 1.1, 2, and 3), Qwen (versions 1.5, 2.5, and 3), and OpenAI (GPT 3.5 Turbo, 4o, and 5). Within each open-weight model family, we also select small, medium, and large models across versions, thus including multiple release dates (between 2023-2025). Table 4 provides the details of all models.

Settings We test each model with two different settings: (1) parametric memory with instruction-fine-tuning only (IFT) and (2) retrieval augmented generation (RAG). In addition, to compare the epistemic diversity of LLMs with that of a traditional web search, we include as a baseline 20 retrieved web pages from a Google search for each topic.[8] Furthermore, we use the search data as our RAG database by splitting each web page into paragraphs. For retrieval, we match prompts to the top paragraphs using S-BERT embeddings, including up to 1000 tokens of context from these retrieved paragraphs. Since many prompts match with similar top paragraphs in the search results we additionally shuffle the ranks of all paragraphs with a cosine similarity to the query above 0.35 to encourage different prompts to use different contexts.[9] This way, we can observe how varying context impacts output diversity in the RAG setting. For generation, we use top p sampling ($p = 0.9$, temperature $= 1.0$) and generate a maximum of 2,100 tokens per prompt ($\approx$ the median number of tokens on the web pages in our search results).

For rarefaction, we calculate the coverage of each sample for each setting on each topic using Equation 2, and rarefy all samples to the minimum.[10] Due to the low coverage achieved by the search baseline, we only rarefy search if the coverage is above the minimum achieved by an LLM. Therefore, all results with search should be considered as a lower-bound baseline; the true diversity of search relative to the LLMs tested will be higher.

Topics As our study is broad (27 models, two settings, and 200 prompt templates per model per topic), we constrain the number of selected topics for computational feasibility. We use a sample of 30 general topics from IssueBench (Röttger et al., 2025) and 125 additional hand-curated topics. These include 10-13 country-specific topics about important figures and historical events to help measure the impact of cultural context on diversity. To ensure that these are well documented and known topics, we only select topics where their associated English Wikipedia page has a content rating of at least "C".[11] Following IssueBench, these include a mix of both controversial and non-controversial topics, e.g., general topics for which there may be contentious views, like nuclear weapons and pornography, and events which may be subject to diverse and varied interpretations. All events occurred before the earliest knowledge cutoff of any model in our study. In total, we have 155 topics (see Appendix § A.3 for the list), yielding 1.7M responses prior to decomposition and 69.5M claims after decomposition. While the topics were hand-picked based on the above criteria, the selection process could nevertheless impact the generalizability of our results. However, if we can identify knowledge collapse even for this limited set of topics, this would already be cause for concern.

[8] Search performed on Aug 10, 2025, in the US region. We filter out content from social media, PDFs, and any page with fewer than 1000 characters.

[9] We calculated this as the mean cosine similarity across all paragraphs and prompts in the data.

[10] We found that by sampling 200 responses per LLM per topic, the coverage tends to vary widely between models (20%-80%), while for search it tends to be $< 20\%$.

[11] https://en.wikipedia.org/wiki/Wikipedia:Content_assessment

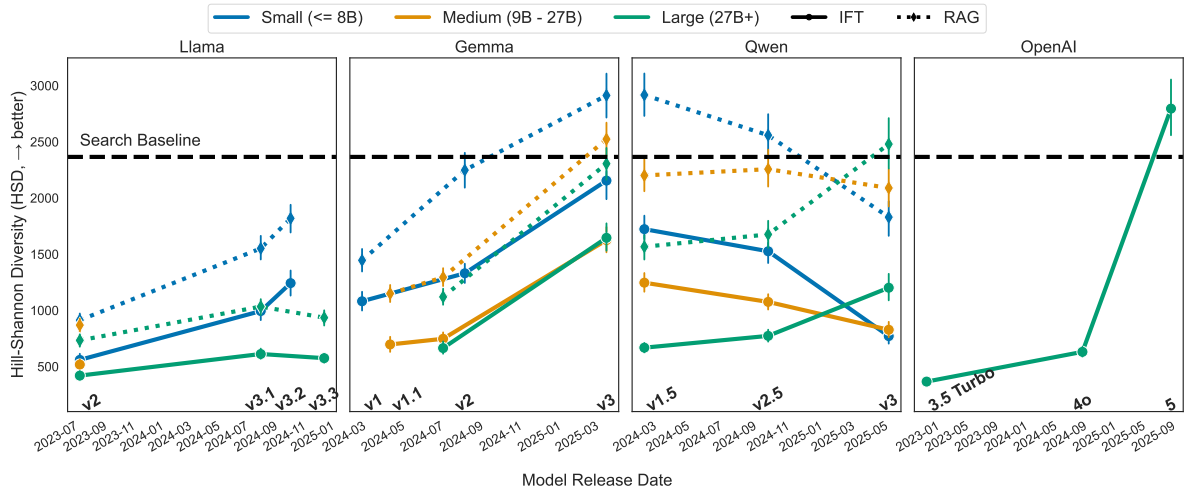

Figure 3: Epistemic diversity vs. model release date. Each point is a single model, with lines connecting models of approximately the same size across released versions. Error bars are 95% boostrapped confidence intervals based on the HSD of each topic (N=155). Absolute diversity is low for all models compared to a very modest search baseline (top 20 Google search results for each topic). However, for all families except Qwen and most sizes, we see a trend of improved diversity.

Clustering For our clustering algorithm we have one parameter to select: the number of claims N to retrieve for comparison. To choose this, we first ran a sample of 50,000 claims from 10 topics (5,000 each) through Algorithm 1, setting $N = 10$. From the list of top 10 most similar claims, we measure mutual entailment with the retrieving claim, and count the number of times that the k^{th} most similar claim ($k \in [1, 10]$) is mutually entailed. We found that when a claim is matched, 98.4% of the time it is within the first 6 most similar claims, and therefore select $N = 6$ for our experiments.

We give examples of the top 10 clusters for four topics (democracy, free speech, Berlin Wall, and K-pop) in Fig. 2. Clusters are aggregated across *every claim in the dataset*, i.e., all models and settings. The frequency of claims drops rapidly; the 10th cluster is half as frequent as the first in all examples shown. Additionally, high-frequency clusters unsurprisingly provide broad, common information about each topic (e.g., the definitions of democracy and free speech, and the highest-level descriptive information about the Berlin Wall and K-pop).

6.2 Epistemic Diversity Across Time

First, we focus on **RQ1**: **How has epistemic diversity in LLMs changed over time?** We measure epistemic diversity across the 155 topics and 27 models mentioned previously, and plot Hill-Shannon diversity (HSD) vs. model release date in Fig. 3. We plot traditional search diversity as a baseline and separately plot IFT and RAG to see the impact of generation setting. We connect models over time based on their relative size and version.

In terms of *relative* diversity for the topics we studied, evidence suggests that models are improving. This is particularly pronounced for the more recent models released after March 2025 (Gemma 3 and GPT-5), which show sharp increases in diversity. These upward trends occur primarily for Llama, Gemma, and OpenAI models; Qwen models are generally stagnant. Other exceptions include the larger (70B) Llama models, which trend downwards, and the larger (32-70B) Qwen models. Finally, we observe that generation setting (IFT or RAG) and model size appear to have a significant impact on diversity; we explore these two factors more rigorously in the following sections.

That being said, when looking at epistemic diversity compared to the search baseline, LLM outputs have low diversity. One can generally expect to encounter more diverse information about these topics by reading through the top-20 web pages rather than prompting an LLM in different ways. The search baseline is, in fact, a quite *weak* baseline, as we use only the top 20 Google search results for each topic, compared to sampling from each language model with 200 unique prompts per topic. Further, as mentioned in § 6.1, the search baseline is *under-estimated* relative to the LLM results. Therefore, while it is encouraging that LLMs

Predictor	$\widehat{\beta}$	p-value
Intercept (IFT)	1054.77($\pm$242.65)	$p \ll 1e{-}3$***
RAG	+739.186($\pm$34.06)	$p \ll 1e{-}3$***
Search	+1311.14($\pm$1284.00)	$p < 0.05$*

Table 2: Estimated coefficients ($\hat{\beta}$) and p-values of a linear mixed effects model with the setting (IFT, RAG, Search) as fixed effects and the model as random effects. The dependent variable is HSD. Both RAG and the search baseline lead to statistically significantly more diverse responses than IFT.

Predictor	$\widehat{\beta}$	p-value
Intercept (Med.)	1034.47($\pm$249.49)	$p \ll 1e{-}3$***
Small	+239.190($\pm$48.21)	$p \ll 1e{-}3$***
Large	-228.47($\pm$49.5)	$p \ll 1e{-}3$***

Table 3: Estimated coefficients ($\hat{\beta}$) and p-values of a linear mixed effects model with the model size (Small, Medium, and Large) and setting (IFT and RAG) as fixed effects and the model/version as a random effect. The dependent variable is HSD. We find an inverse relationship between model size and diversity..

have become more epistemically diverse over time, their diversity relative to search baselines is often low, which may risk knowledge collapse.

6.3 Epistemic Diversity Across Settings

Our next focus is on **RQ2: What is the impact of generation setting on epistemic diversity?** We look at epistemic diversity across three settings: parametric knowledge in instruction fine-tuned models (IFT), the same models with retrieval-augmented generation (RAG), and the traditional search engine baseline. From Fig. 3, we see that, on our topics, RAG appears to have a strong positive impact on diversity, and search tends to be better than both IFT and RAG. To test if this is statistically significant, we use a linear mixed effects regression model with HSD as the dependent variable, generation setting as a categorical fixed effect, and the model as a random effect in order to control for the baseline diversity of each model.

From Table 2, we see that both RAG and the traditional search baseline yield significantly more diverse outputs than relying on parametric memory with IFT, at least for our topics. This highlights the *potential* of RAG in ensuring that models

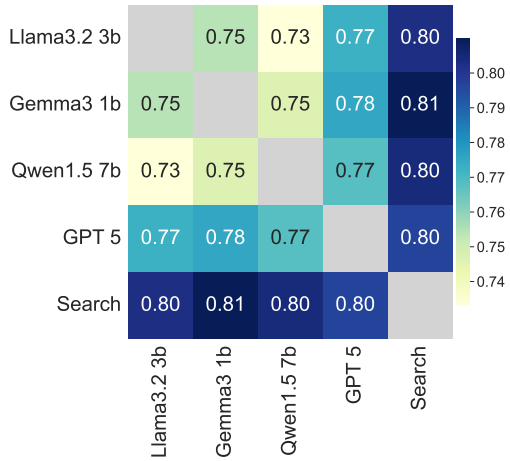

Figure 4: Heatmap of the Jensen-Shannon divergence (JSD) across models, based on the empirical probability distributions over clusters (p_i) for each topic. A higher JSD means that the distributions generated by the two models are more different. Open-weight models tend to be more similar to each other than to GPT. All LLMs are more different from the search baseline than to each other, indicating a marked difference in the distribution of information in the search baseline from the LLMs.

are epistemically diverse going forward. However, this may depend on RAG databases remaining *human written*. If traditional search platforms and RAG knowledge bases become dominated by LLM-generated content, the diversity benefits of RAG could be erased, risking knowledge collapse. Therefore, given the benefits that RAG can endow to epistemic diversity, we recommend RAG databases remain diverse and additionally prevent contamination from an overabundance of LLM generated text.

6.4 Epistemic Diversity Across Models

We now look at **RQ3: How does model selection impact epistemic diversity?** There are two aspects to consider here: the impact of model size and the similarity between model families.

Firstly, from Fig. 3, we see that model size appears to have an unintuitive *negative* impact on epistemic diversity. To determine if the effect is statistically significant, we use a linear mixed effects model with the HSD as the dependent variable, the model size binned into three categories (Small, Medium, and Large; see Table 4 for how these are determined) as fixed effects, the generation setting (IFT or RAG) as additional fixed effects, and the

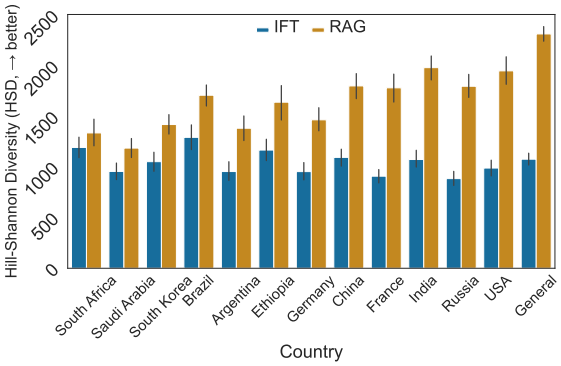

Figure 5: Average diversity per country across all models with bootstrapped 95% confidence intervals. Bars are sorted according to the difference between RAG and IFT diversity. Countries tend to have similar diversity to each other with instruction fine-tuning only. However, RAG appears to have an uneven impact on different countries, where the US and general topics see the most benefit

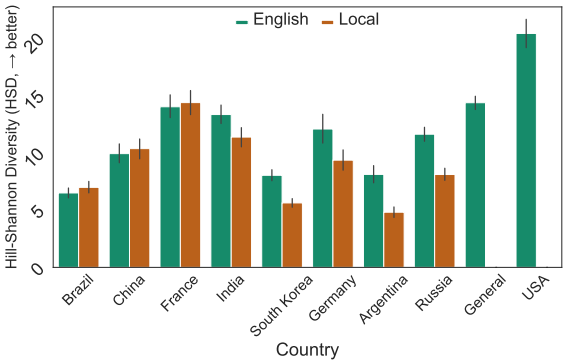

Figure 6: Comparison of the diversity of claims generated by the models matched to English Wikipedia and local language Wikipedia. To capture minimal representativeness (Peterson, 2025), we match claims from Wikipedia in English and the local language, respectively, to claims generated by our models in the IFT setting, and restrict the measurement of HSD to only matched claims. Bars are sorted according to the difference between English and local language representativeness.

model version as random effects. The results in Table 3, highlight that the observed effect w.r.t. model size is statistically significant. This echoes contemporary work which has shown that larger models tend to memorize more (Morris et al., 2025), can show greater bias (Bai et al., 2025b), and can be less diverse (Zhang et al., 2025) than smaller models. Therefore, in settings where epistemic diversity is important (e.g., in viewpoint/opinion generation), smaller models may be a better choice.

But how do different models compare? We measure model similarity by measuring the Jensen-Shannon divergence (JSD) between the distributions of meaning classes p_i generated by the most diverse model from each family and average this across topics (Fig. 4). We observe that the overall divergence is high across models for our topics. This suggests that prompting multiple models may improve epistemic diversity. Additionally, we observe that the open-weight models are more similar to each other than to GPT5 or the search baseline. Finally, we see that the greatest divergence across models is with the search baseline, suggesting that traditional search and LLMs produce largely different distributions of claims, at least for our topics.

6.5 Epistemic Diversity Across Countries

Finally, we consider **RQ4: Whose knowledge is represented by LLMs?** through two subquestions.

First, how does the cultural context of a topic impact epistemic diversity? To answer, we plot HSD vs. country as a bar chart in Fig. 5. The country of each topic is selected either if (1) a historical event occurred in a particular country or (2) a public figure is primarily associated with a certain country. We plot separate bar charts for the IFT and RAG settings. For IFT only, the cultural context does not have a strong impact on HSD, as most countries are within the same 95% confidence intervals, with the exception of Ethiopia, Saudi Arabia, and Brazil, which are higher. However, each country is impacted differently by RAG. In particular, the USA, India, Russia, France, and China, as well as general concepts, see more benefit from RAG. This is likely because the search we use to acquire documents for RAG is performed in the US region, which may under-represent topic-specific information from certain countries. To check this, we measure Pearson's r correlation between the epistemic diversity of search and the difference between RAG diversity and IFT diversity, averaging by country. We find that these two are strongly correlated at 0.73 ($p < 0.01$, N=13), meaning that when the diversity of the RAG source for a particular country's set of topics increases, the improvement in diversity from IFT to RAG is likewise greater.

Second, to what extent are English and local language knowledge represented by the claims generated by LLMs? To answer this, we match LLM claims to Wikipedia claims in both English and

the country's local language using InfoGap (Samir et al. 2024, see Appendix § A.5 for details). We restrict this experiment to 9 countries in our study where the local language Wikipedias are in the top 20 in terms of active users, with the exception of Saudi Arabia for which the local language pages had minimal content. In order to quantify the extent to which Wikipedia claims appear in model outputs, we measure the HSD of each model after filtering out claims that are not matched to either the English or local language Wikipedia. This is equivalent to "minimal representativeness" outlined by Peterson (2025), which defines that, for a set of items i that are relevant to a particular task (in this case, claims which appear on either English or local language Wikipedia pages), "any item should have at least *some* chance of appearing in the LLM output." Results are plotted as a bar chart in Fig. 6, aggregated across models and topics for each country.

For the selected topics, we observe a statistically significant knowledge gap between English and local languages for 5 out of 8 countries, while for the remaining 3, the gap is not statistically meaningful. In no case is local representation statistically significantly greater than English representation. Additionally, representation for the USA specific topics are statistically significantly greater than every other country. This suggests that current English LLMs may fail to present knowledge which has not been presented in English for certain countries, potentiating a risk in terms of knowledge erasure.

7 Discussion and Conclusion

Our work is the first broad study of knowledge collapse in LLMs, where we propose a new methodology for measuring epistemic diversity. We use this to conduct an empirical study of 27 LLMs across 155 topics, with 200 prompt variations.

Limitations In terms of topical coverage, we use a combination of randomly selected and hand curated topics. We attempted to select topics that broadly cover general concepts, historical events, and important figures across 12 countries, including both controversial and non-controversial topics. However, manual selection of topics may inevitably impact the generalizability of our results. Additionally, for some countries in the study, selection of topics is performed from an outsider perspective. An improvement to our setup could be to cover more topics by identifying what people typically search for using LLMs and randomly sampling from that pool of queries. Other notions of popularity can also be examined in terms of Wikipedia page views or trending search keywords. Regarding our decomposition and clustering algorithms, while we find through manual and automated evaluation that the quality tends to be high, performance is not perfect. Therefore, there is some irreducible noise in the results. Next, we choose a RAG setup which is intended to simulate a real-world RAG scenario, while actual RAG implementations will inevitably vary. It would therefore be useful to examine the utility of RAG in a more systematic way, benchmarking several implementations. Finally, while epistemic diversity is important, it should also be contextualized with other important aspects of knowledge such as factuality and relevance.

Conclusions Our results show that overall epistemic diversity is low in our pool of topics when compared to a baseline traditional search. That being said, for most model families and sizes tested, we observe an encouraging trend of knowledge expansion, indicating that so far, LLMs do not appear to be locked in to their narrow epistemic frames. Our results also highlight that RAG and the use of smaller models can help stave off knowledge collapse going forward. Going deeper, we find that the choice of RAG database may be important for improving epistemic representation across cultural contexts, as not all country-specific topics see equal gains. Here, practitioners need to be cautious about expanding their RAG sources with LLM-generated content. Finally, compared to a traditional knowledge source (Wikipedia) in both English and local languages for country-specific topics, we find that the claims generated in our study reflect English language knowledge more than local language knowledge, highlighting the need to investigate how local knowledge can be incorporated into LLM outputs. The general methodology presented in this paper can be used in the future to study epistemic diversity for any arbitrary set of topics, downstream tasks, and real-world use cases with open-ended plain-text LLM outputs. This allows researchers to answer research questions about *which*, *whose*, and *how much* knowledge LLMs are representing

Acknowledgements

DW is supported by a Danish Data Science Academy postdoctoral fellowship (grant: 2023-1425). JM is supported by the Stanford Interdisci-

plinary Graduate Fellowship, the Stanford Center for Affective Science Graduate Fellowship, and the Future of Life Institute Vitalik Buterin PhD Fellowship. SY is supported in part by the Pioneer Centre for AI, DNRF grant number P1. This work is also partially funded by a DFF Sapere Aude research leader grant under grant agreement No 0171-00034B.

References

Suhaib Abdurahman, Mohammad Atari, Farzan Karimi-Malekabadi, Mona J Xue, Jackson Trager, Peter S Park, Preni Golazizian, Ali Omrani, and Morteza Dehghani. 2024. Perils and opportunities in using large language models in psychological research. *PNAS nexus*.

A. J. Alvero, Jinsook Lee, Alejandra Regla-Vargas, René F. Kizilcec, Thorsten Joachims, and Anthony Lising Antonio. 2024. Large language models, social demography, and hegemony: comparing authorship in human and synthetic text. *J. Big Data*.

Barrett R. Anderson, Jash Hemant Shah, and Max Kreminski. 2024. Homogenization Effects of Large Language Models on Human Creative Ideation. In *Proceedings of the 16th Conference on Creativity & Cognition*. ACM.

Mohammad Atari, Mona J Xue, Peter S Park, Damián Blasi, and Joseph Henrich. Which humans?

Hui Bai, Jan G Voelkel, Shane Muldowney, Johannes C Eichstaedt, and Robb Willer. 2025a. LLM-generated messages can persuade humans on policy issues. *Nature Communications*.

Xuechunzi Bai, Angelina Wang, Ilia Sucholutsky, and Thomas L Griffiths. 2025b. Explicitly unbiased large language models still form biased associations. *Proceedings of the National Academy of Sciences*.

Anne Chao and Lou Jost. 2012. Coverage-based rarefaction and extrapolation: standardizing samples by completeness rather than size. *Ecology*.

Giovanni Luca Ciampaglia, Azadeh Nematzadeh, Filippo Menczer, and Alessandro Flammini. 2018. How algorithmic popularity bias hinders or promotes quality. *Nature Scientific Reports*.

Mark Coeckelbergh. 2025. LLMs, Truth, and Democracy: An Overview of Risks. *Science and Engineering Ethics*.

Esin Durmus, Karina Nyugen, Thomas I. Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. 2023. Towards Measuring the Representation of Subjective Global Opinions in Language Models. *arXiv preprint arXiv:2306.16388*.

Subhabrata Dutta and Tanmoy Chakraborty. 2023. Thus spake chatgpt. *Communications of the ACM*.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD)*. AAAI Press.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*.

Mark O Hill. 1973. Diversity and Evenness: A Unifying Notation and its Consequences. *Ecology*.

Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. 2023. Co-Writing with Opinionated Language Models Affects Users' Views. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM.

Lou Jost. 2006. Entropy and diversity. *Oikos*.

Maurice G Kendall. 1938. A New Measure of Rank Correlation. *Biometrika*.

Maurice G Kendall. 1945. The treatment of ties in ranking problems. *Biometrika*.

Jinsook Lee, A. J. Alvero, Thorsten Joachims, and René F. Kizilcec. 2025. Poor Alignment and Steerability of Large Language Models: Evidence from College Admission Essays. *arXiv preprint arXiv:2503.20062*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Lisa Messeri and Molly J Crockett. 2024. Artificial intelligence and illusions of understanding in scientific research. *Nature*.

Jared Moore, Tanvi Deshpande, and Diyi Yang. 2024. Are Large Language Models Consistent over Value-laden Questions? In *Findings of the Association for Computational Linguistics: EMNLP*. Association for Computational Linguistics.

John X. Morris, Chawin Sitawarin, Chuan Guo, Narine Kokhlikyan, G. Edward Suh, Alexander M. Rush, Kamalika Chaudhuri, and Saeed Mahloujifar. 2025. How much do language models memorize? *arXiv preprint arXiv:2505.24832*.

Tien T. Nguyen, Pik-Mai Hui, F. Maxwell Harper, Loren G. Terveen, and Joseph A. Konstan. 2014. Exploring the Filter Bubble: the Effect of Using Recommender Systems on Content Diversity. In *23rd International World Wide Web Conference, WWW*. ACM.

Andrew J. Peterson. 2025. AI and the problem of knowledge collapse. *AI & Society*.

Tianyi Alex Qiu, Zhonghao He, Tejasveer Chugh, and Max Kleiman-Weiner. 2025. The Lock-in Hypothesis: Stagnation by Algorithm. *arXiv preprint arXiv:2506.06166*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Michael Roswell, Jonathan Dushoff, and Rachael Winfree. 2021. A conceptual guide to measuring species diversity. *Oikos*.

Paul Röttger, Musashi Hinck, Valentin Hofmann, Kobi Hackenburg, Valentina Pyatkin, Faeze Brahman, and Dirk Hovy. 2025. IssueBench: Millions of Realistic Prompts for Measuring Issue Bias in LLM Writing Assistance. *arXiv preprint arXiv:2502.08395*.

Paul Röttger, Valentin Hofmann, Valentina Pyatkin, Musashi Hinck, Hannah Kirk, Hinrich Schütze, and Dirk Hovy. 2024. Political Compass or Spinning Arrow? Towards More Meaningful Evaluations for Values and Opinions in Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics.

Farhan Samir, Chan Young Park, Anjalie Field, Vered Shwartz, and Yulia Tsvetkov. 2024. Locating Information Gaps and Narrative Inconsistencies Across Languages: A Case Study of LGBT People Portrayals on Wikipedia. In *Empirical Methods in Natural Language Processing, EMNLP*. Association for Computational Linguistics.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross J. Anderson, and Yarin Gal. 2024. AI models collapse when trained on recursively generated data. *Nature*.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. 2025. Can LLMs Generate Novel Research Ideas? A Large-Scale Human Study with 100+ NLP Researchers. In *The Thirteenth International Conference on Learning Representations, ICLR 2025*.

Zhivar Sourati, Farzan Karimi-Malekabadi, Meltem Ozcan, Colin McDaniel, Alireza S. Ziabari, Jackson Trager, Ala N. Tak, Meng Chen, Fred Morstatter, and Morteza Dehghani. 2025a. The Shrinking Landscape of Linguistic Diversity in the Age of Large Language Models. *arXiv preprint arXiv:2502.11266*.

Zhivar Sourati, Alireza S Ziabari, and Morteza Dehghani. 2025b. The Homogenizing Effect of Large Language Models on Human Expression and Thought. *arXiv preprint arXiv:2508.01491*.

Luning Sun, Yuzhuo Yuan, Yuan Yao, Yanyan Li, Hao Zhang, Xing Xie, Xiting Wang, Fang Luo, and David Stillwell. 2025. Large Language Models show both individual and collective creativity comparable to humans. *Thinking Skills and Creativity*.

Christian Wagner and Ling Jiang. 2025. Death by AI: Will large language models diminish Wikipedia? *Journal of the Association for Information Science and Technology*.

Dustin Wright, Arnav Arora, Nadav Borenstein, Srishti Yadav, Serge J. Belongie, and Isabelle Augenstein. 2024. LLM Tropes: Revealing Fine-Grained Values and Opinions in Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP*. Association for Computational Linguistics.

Dustin Wright, Zain Muhammad Mujahid, Lu Wang, Isabelle Augenstein, and David Jurgens. 2025. Unstructured Evidence Attribution for Long Context Query Focused Summarization. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Weijia Xu, Nebojsa Jojic, Sudha Rao, Chris Brockett, and Bill Dolan. 2025. Echoes in AI: Quantifying Lack of Plot Diversity in LLM Outputs. *arXiv preprint arXiv:2501.00273*.

Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Shaochen Zhong, Bing Yin, and Xia Ben Hu. 2024. Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond. *ACM Transactions on Knowledge Discovery from Data*.

Yiming Zhang, Harshita Diddee, Susan Holm, Hanchen Liu, Xinyue Liu, Vinay Samuel, Barry Wang, and Daphne Ippolito. 2025. NoveltyBench: Evaluating Language Models for Humanlike Diversity. In *Conference on Language Modeling (COLM)*.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. 2024. WildChat: 1M ChatGPT Interaction Logs in the Wild. In *The Twelfth International Conference on Learning Representations, ICLR 2024*.

Elise Li Zheng and Sandra Soo-Jin Lee. 2023. The Epistemological Danger of Large Language Models. *The American Journal of Bioethics*.

A Replication Details

A.1 Evaluation Annotation Instructions

Decomposition Quality You will be given a short piece of text (around 3 sentences) about a provided topic, and a list of atomic claims. Your task is to annotate to what degree each claim is represented in the original short piece of text. Each claim will be rated on a scale from 0 to 5, where each value has the following meaning:

- 0 - EMPTY (skip)
- 1 - The claim is totally irrelevant to the original piece of text OR does not explicitly talk about the provided TOPIC
- 2 - The claim is incomplete or somewhat included in the original piece of text and missing clearly important context
- 3 - The claim is included in the original piece of text but missing some potentially important context. This includes claims which could be inferred from the original context but aren't explicitly stated (e.g., "The lifecycle of plastic includes production." being inferred from "It addresses the entire lifecycle of plastic, from production and consumption to disposal and recycling.")
- 4 - The claim is included in the original piece of text and is missing only unimportant context (the most important information is represented)
- 5 - The claim is included in the original piece of text and no context is missing

Decomposition Missing Claims In addition, after each group of claims you will be asked to enumerate any missing claims which are relevant to the topic. Please compare the text to the claims, and identify any relevant claims that were not listed. You just need to provide the NUMBER of claims that you estimate are missing (you do not need to write the claims down, just the number).

Clustering Cohesion You will be shown groups of text. Your job is to rate each group of text for the degree to which each piece of text in the group conveys at least one piece of information in common. Each group will have a maximum of 10 pieces of text and a minimum of 3 pieces of text. Rate each group as follows:

- 1 - None of the sentences have any information in common

- 2 - Some of the sentences have information in common but most sentences convey something different
- 3 - About half of the sentences convey one thing in common
- 4 - Most of the sentences convey at least one piece of information in common
- 5 - All of the sentences convey at least one piece of information in common

Clustering Missing Sentences You will then be shown 10 additional pieces of text. Your task will be to determine whether or not these pieces of text belong in the group. You will rate each piece of text as follows:

- 0 - The sentence does not convey any information in common with the group
- 1 - The sentence conveys the same common information shared by the group; the sentence may convey some more specific or extra information, but the core information of the group is conveyed by the sentence as well

A.2 Prompts

Here we show decomposition prompts P1 (Fig. 7) and P2 (Fig. 8) used in our evaluation. Decomposition prompt P3, which we use in the final experiments, is the same as P2 but includes three in-context examples. All three prompts can be found in our linked code repository.[12]

```
Please breakdown the following paragraph into a list of independent facts. Please return the facts as a list. Please do not number the list. Please simply separate each fact by a new line character.
{content}
```

Figure 7: Decomposition prompt P1.

```
I am going to show you a piece of TEXT that talks about the following ISSUE: {issue}. Please read every single sentence in the TEXT and then decompose each sentence into individual, atomic STATEMENTS. Please make the STATEMENTS totally faithful to the original TEXT. Please only include STATEMENTS about issue. If there are no sentences in the TEXT which include statements about issue, please just say EMPTY. Please ignore sentences which talk about the TEXT itself, for example "In sum, we have talk about {issue}." Please only include STATEMENTS which are fully supported by the TEXT. Please don't hallucinate information. Please do not repeat STATEMENTS. Please return the STATEMENTS as a list. Please do not number the list. Please simply separate each STATEMENT by a new line character. Please double check that each of the STATEMENTS in the list is entailed and factually accurate with the text. If the TEXT is presented as a list of bullet points, please make sure to decontextualize the bullet points (don't just list them). Here is the text:
{content}
```

Figure 8: Decomposition prompt P2.

A.3 Topics

USA Donald Trump; Claude Shannon; Barack Obama; Frederick Douglass; Bob Dylan; Iran hostage crisis; 9/11; American civil war; Roe v. Wade; Independence Day (United States); **Argentina** Javier Milei; Juan Domingo Perón; Cristina Fernández de Kirchner; Jorge Rafael Videla; Diego Maradona; Falklands War; Infamous Decade (Argentina); May Revolution; 1998–2002 Argentine great depression; Cadet scandal in Argentina; **Ethiopia** Kenenisa Bekele; Sophia Bekele; Meles Zenawi; Mengistu Haile Mariam; Kebede Michael; History of Ethiopia; Ethiopian Civil War; Tigray War; History of the Federal Democratic Republic of Ethiopia; COVID-19 pandemic in Ethiopia; **Korea** Syngman Rhee; Yuna Kim; Kim Soo-hyun; Sejong the Great; Ban Ki-moon; Gwangju uprising; Seollal; South Korean March First Movement; K-pop; Sinking of MV Sewol; **Russia** Vladimir Putin; Joseph Stalin; Fyodor Dostoevsky; Alexei Navalny; Garry Kasparov; October Revolution; Soviet Invasion of Poland; Sputnik 1; 1993 Russian constitutional crisis; Wagner Group rebellion; **France** Charles de Gaulle; Marine le Pen; Albert Camus; Michel Foucault; Napoleon; Dreyfus affair; the French Revolution; Charlie Hebdo shooting; Arenc affair; Yellow vests

[12] https://github.com/dwright37/llm-knowledge

protests in France; **India** Jallianwala Bagh massacre; Partition of India; 2002 Gujarat riots; Indian Rebellion of 1857; Narendra Modi; Jalal-ud-din Muhammad Akbar; B. R. Ambedkar; Indira Gandhi; Vallabhbhai Patel; Non-cooperation movement (1919–1922); 2008 Mumbai attacks; **Saudi Arabia** Assassination of Jamal Khashoggi; Al-Yamamah arms deal; Khobar Towers bombing; Proclamation of the Kingdom of Saudi Arabia; Riyadh International Book Fair; Ibn Saud; Hatoon al-Fassi; Manal al-Sharif; Ayatollah Sheikh Nimr Baqir al-Nimr; **South Africa** 2021 South African unrest; 1999 Tempe military base shooting; Soweto uprising; Miriam Makeba; Nadine Gordimer; Nelson Mandela; Evelyn Mase; Christiaan Barnard; Crizelda Brits; Rand Rebellion; South African Border War **Brazil** Paraguayan War; 1937 Brazilian coup d'état; Revolution of the Ganhadores; Brazilian Carnival; Dilma Rousseff; Pedro II of Brazil; José Paranhos, Viscount of Rio Branco; Indigenous peoples in Brazil; Carmen Miranda; Carlos Chagas; Mensalão scandal; **China** 1989 Tiananmen Square protests and massacre; 2019–2020 Hong Kong protests; Annexation of Tibet; Qingming Festival; 2010 Yushu earthquake; Jinan incident; Du Fu; Xi Jinping; Chinese Communist Party; Ai Weiwei; Soong Ching-ling; **Germany** Berlin Wall; Nuremberg trials; Night of the Long Knives; Clara Josephine Schumann; Wilhelm Richard Wagner; Frederick the Great; Karl Marx; Eschede train disaster; Frauke Petry; Ernst Nolte; **General** nuclear weapons; slavery; pornography; marriage; white supremacy; international relations; prisons; domestic violence; patriotism; same-sex marriage; free speech; political corruption; universal basic income; global hunger; plastic waste; political correctness; fascism; racism; colonialism; the impact of climate change; democracy; feminism; human rights; genocide; war; censorship; artificial intelligence; renewable energy; capitalism; autonomous vehicles; misinformation; affirmative action;

A.4 Model Details

Table 4 provides the details of the 4 model families employed in this study, along with demarcation of small, medium, and large sets.

A.5 InfoGap Details

We use GPT 5 to decompose English and local language Wikipedias to individual claims, retrieve candidate matches to LLM generated claims using a multilingual sentence embedding model,[13] and predict if the retrieved multilingual claims match to the LLM generated English claims using GPT 5 (Samir et al., 2024). We exclude claims generated using RAG to avoid Wikipedia text being included in the generation context.

A.6 Generation Details

We retrieve 20 Google search results for each topic. Over the 155 topics, these 3,200 pages have an average of 12,836 and a median of 2,132 tokens per page. For a fair comparison, we allow each model to generate up to 2,100 tokens for each of the 200 prompt variations. For RAG, we include up to 1,000 additional tokens for context.

[13] HuggingFace: sentence-transformers/LaBSE

Family	Ver.	Size	Release Date	Endpoint
Qwen	1.5	7B (S)	February 2024	Qwen/Qwen1.5-7B-Chat
	1.5	14B (M)		Qwen/Qwen1.5-14B-Chat
	1.5	72B (L)		Qwen/Qwen1.5-72B-Chat
	2.5	7B (S)	September 2024	Qwen/Qwen2.5-7B-Instruct
	2.5	14B (M)		Qwen/Qwen2.5-14B-Instruct
	2.5	72B (L)		Qwen/Qwen2.5-72B-Instruct
	3	8B (S)	April 2025	Qwen/Qwen3-8B
	3	14B (M)		Qwen/Qwen3-14B
	3	32B (L)		Qwen/Qwen3-32B
Gemma	1	2B (S)	February 2024	google/gemma-2b-it
	1.1	7B (M)	April 2024	google/gemma-1.1-7b-it
	2	2B (S)	July 2024	google/gemma-2-2b-it
	2	9B (M)	June 2024	google/gemma-2-9b-it
	2	27B (L)		google/gemma-2-27b-it
	3	1B (S)	March 2025	google/gemma-3-1b-it
	3	12B (M)		google/gemma-3-12b-it
	3	27B (L)		google/gemma-3-27b-it
Llama	2	7B (S)	July 2023	meta-llama/Llama-2-7b-chat-hf
	2	13B (M)		meta-llama/Llama-2-13b-chat-hf
	2	70B (L)		meta-llama/Llama-2-70b-chat-hf
	3.1	8B (S)	July 2024	meta-llama/Llama-3.1-8B-Instruct
	3.1	70B (L)		meta-llama/Llama-3.1-70B-Instruct
	3.2	3B (S)	September 2024	meta-llama/Llama-3.2-3B-Instruct
	3.3	70B (L)	December 2024	meta-llama/Llama-3.3-70B-Instruct
OpenAI**	3.5	Undisclosed	November 2022	gpt-3.5-turbo-0125
	4	Undisclosed	August 2024	gpt-4o-2024-08-06
	5	Undisclosed	August 2025	gpt-5-2025-08-07

Table 4: List of different model versions (Ver.), sizes, year of release, along with model endpoints used either from HuggingFace or the respective API endpoint. ** is for closed-weight models. In terms of model size: Small $\leq 8B$, $9 \leq$Medium$< 27B$ and Large$\geq 27B$. As Gemma models tended to be smaller we include one 7B model (Gemma 1.1 7B) in the Medium category for Fig. 3.