# What and Whose Knowledge?
# Measuring Epistemic Diversity in Large Language Models

**Anonymous ACL submission**


## Abstract

Large language models (LLMs) are increasingly used as primary knowledge sources, yet their epistemic diversity — defined as the diversity of real-world claims in their outputs — has never been measured. The dominant paradigm is that overall LLM diversity is low, but this is always with respect to a single point in time, with no reference baseline or consideration for variation across countries. We address this by performing the first systematic study of epistemic diversity in LLMs across time and cultural context, testing 27 LLMs on 155 topics covering 12 countries, resulting in 1.7M responses and 70M individual claims. We find that epistemic diversity has increased substantially over the past three years, a positive counter to recent diversity pessimism. However, despite progress, we find that every LLM is less diverse than a search baseline. The gap is not uniform: RAG can improve diversity, while larger models are counterintuitively less diverse than smaller ones. Moreover, LLMs systematically reflect English-language knowledge over local-language knowledge for country specific topics. Together, these results demonstrate that while progress on epistemic diversity is tangible, it is insufficient and unevenly distributed.[1]


## 1 Introduction

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

At the same time, LLM outputs are homogeneous along multiple axes (Sourati et al., 2025b).

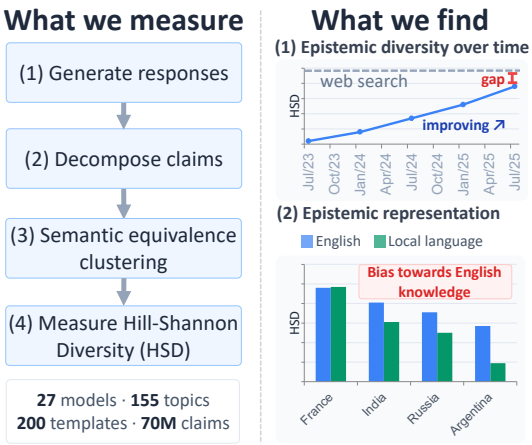


Figure 1: In this work, we measure epistemic diversity via variability in claims about the world. We compare across models, settings, and country-specific topics.

They 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; Jiang et al., 2025; Yu et al., 2023). Given the tasks people use these models for, there is a risk of knowledge collapse — where "access to the original diversity of human knowledge is increasingly mediated by a partial and increasingly narrow subset of views" (Peterson, 2025). There is thus an urgent need for reliable measurements of **epistemic diversity** of LLMs, which we define as the diversity of claims about the world in a set of LLM outputs. Moreover, as the dominant paradigm posits static or worsening diversity in LLM outputs, there is a need to study diversity patterns in LLMs across models, time periods, and cultural context to gain insight into the current and future risks of knowledge collapse and how to prevent it.

To this end, we perform a first empirical study measuring the epistemic diversity in LLM outputs (see Fig. 1). We investigate diverse scenarios in which LLMs are prompted in different ways about the same topics, generating open-ended free-text responses with multiple embedded claims. Con-

[1] Code and data: https://anonymous.4open.science/r/knowledge-collapse-9166

cretely, we (1) sample responses from 27 LLMs using multiple versions, settings, and topics with human-written prompt templates collected from Röttger et al. (2025), (2) decompose individual LLM responses into one or more claims, (3) partition those claims into semantically equivalent clusters (Farquhar et al., 2024), and (4) quantify the diversity of the sample of claims with Hill-Shannon diversity (Roswell et al., 2021). We examine both *what* and *whose* claims are generated, measuring diversity both across and within 155 general domain and country specific topics. In doing so, we produce a dataset of 70M claims clustered by semantic equivalence and generated by a set of the most popular LLMs from the past three years.

Encouragingly, we find that epistemic diversity in LLM outputs is generally *increasing* over time, a positive counter to recent work positing knowledge collapse (Peterson, 2025) and static lack of diversity in LLMs (Jiang et al., 2025). However, despite this progress, we find that there is a gap in epistemic diversity between all LLMs in our study and a weak traditional search baseline. This diversity gap between web search and LLMs is particularly alarming given recent shifts in web search toward AI-exclusive designs.[2]

Further, this gap is not uniform. Retrieval-augmented generation (RAG) has a statistically significant *positive* impact, highlighting the importance of RAG in preventing knowledge collapse. However, similar to contemporary work (Zhang et al., 2025; Rassin et al., 2024), we counterintuitively find that model size has a statistically significant *negative* impact — smaller models generate more diverse knowledge than larger ones. Finally, when considering the countries that topics are associated with, we find that some countries may experience greater risks of knowledge collapse than others. RAG has an uneven effect: certain countries such as the USA see more benefit due to a greater diversity in their RAG sources, which are not available for many languages. And compared to Wikipedia, in both English and local languages for country-specific topics, we find that LLMs reflect English language knowledge more than local language knowledge, highlighting a gap in epistemic representation. Overall, our results demonstrate that while progress on epistemic diversity is tangible, it is insufficient and unevenly distributed.

[2] https://techcrunch.com/2026/05/19/google-search-as-you-know-it-is-over/

## 2 Background

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

**LLM Homogenization** A robust body of work shows that LLMs suffer from a lack of diversity along many dimensions (Sourati et al., 2025b; Guo et al., 2025). These include: lexical and stylistic (Sourati et al., 2025a; Shaib et al., 2024; Padmakumar and He, 2024), semantic (Lee et al., 2025; Jiang et al., 2025; Yu et al., 2023; Padmakumar and He, 2024; Moon et al., 2025), creative (Xu et al., 2025; Moon et al., 2025; Wenger and Kenett, 2025), conceptual (Murthy et al., 2025), recommendation (Dudy et al., 2025; Poulain et al., 2024; Barolo et al., 2025; Shur-Ofry et al., 2024), coding (Shypula 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; Hayati et al., 2024; Wang et al., 2025), among others. 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). We build on this by looking at the diversity of all semantically equivalent classes of *claims* in open-ended free-text LLM responses, avoiding shallow and/or fuzzy features used in semantic and perspective similarity. Our methodology and empirical results show how this diversity has changed over time and across countries, with actionable recommendations, and offering more general conclusions on knowledge diversity than previous work.

**Knowledge Collapse** A growing concern among scholars is that LLM 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; Farrell 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. Our work assesses this risk through a new methodology for measuring epistemic diversity and an empirical study of this phenomenon in LLMs. Using this, we compare risks in LLM outputs across models, sources, and time, and we contrast our findings with measurements for traditional knowledge systems.

## 3 Problem Setup and Notation

We measure the epistemic diversity of LLMs as the diversity of the distribution of claims made in their outputs to open-ended prompts on various topics. Concretely, we develop a new methodology which decomposes open-ended LLM-responses into claims, partitions these claims into semantically equivalent clusters, and measures the Hill-Shannon diversity of these clusters, a statistically grounded measure of diversity used widely in ecology for measuring species diversity (Peterson, 2025; Roswell et al., 2021).

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 $C_{m_t}$. These claims can be organized into a set of unique *meaning classes* $\mathcal{X}_{m_t}$. A unique meaning class contains claims that are semantically equivalent to one another but not to 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., $\frac{x_i}{n}$.

## 4 Data Collection

We first describe how $\mathcal{X}_{m_t}$ is constructed 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):

1. **Generate**: Acquire a corpus $P_{m_t}$ of free-text LLM responses to natural input prompts.

| | 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).

2. **Decompose**: Decompose $P_{m_t}$ into a list of claims $C_{m_t}$.
3. **Cluster**: Group the list of claims into meaning classes $\mathcal{X}_{m_t}$ whose members are semantically equivalent.

### 4.1 Generation and Decomposition

To elicit open-ended responses about any given topic $t$, we start with the prompt templates curated in Röttger et al. (2025) These templates take forms such as "write me an essay about `{topic}`," where `{topic}` is replaced by $t$. From the original 1,000 templates, we manually select a subset of 479 that focus on information seeking, then randomly sample 200 as our final templates.

After generating responses ($P_{m_t}$), we **decompose** them into individual claims $C_{m_t}$ using a Llama 3.1 70B model, prompting the model to decompose the response, sliding over non-overlapping chunks of three sentences at a time. We thus ensure that each input is self-contained with all necessary context while capturing all claims.

We initially developed three decomposition prompts (P1-P3; see Appendix B for prompt text) and selected the best one based on an LLM-as-a-judge setup (Liu et al., 2023). Two human annotators labeled 100 data points for each of two tasks: (1) rating the *quality* of individual decomposed claims on a Likert scale from 1-5 and (2) quantifying how many claims in the original input chunk are *missing* from the list of decomposed claims.

We achieved moderate to strong Pearson $r$ across our two annotators (0.74, 0.58 respectively), and had a third annotator break annotation ties. Our prompts for LLM-as-a-judge (see Appendix A for details) achieved Pearson correlations of 0.60 and 0.68 respectively. We use this setup to automatically label a sample of 6K instances across the three decomposition prompt variants. The results are given in Table 1. We use P3 for the final decomposition in order to prioritize quality.

### 4.2 Clustering

Following Farquhar et al. (2024), we create clusters based on mutual entailment using a strong pre-

trained model for natural language inference.[3] This is done in a greedy bootstrap fashion, building clusters one claim at a time by assigning a given claim to an existing cluster if it mutually entails at least one other claim in that cluster. To reduce the overhead of measuring mutual entailment across all $2n^2$ pairs for a set of $n$ claims, we check entailment between each claim and the $N$ most similar claims to it (i.e., $2N * n$ comparisons) measured using an S-BERT model.[4] We then perform a post-processing step to break up larger clusters using DBSCAN (Ester et al., 1996), a related clustering approach (Peterson, 2025; Wright et al., 2024). See Appendix C for the algorithm pseudocode.

We evaluate clustering using LLM-as-a-judge, again (1) rating cluster quality (Likert score 1-5) and (2) predicting potential missing claims (i.e., unclustered claims which should have been included in a given cluster). Two annotators labeled data for both tasks, achieving moderate agreement (0.44 Pearson $r$ for cluster quality on 100 examples; 87% agreement accuracy for missing claims on 360 examples), followed by a third annotator to break ties. We develop prompts that achieve 0.68 Pearson $r$ with human labels for cluster quality and 83.5 weighted F1 score for predicting missing claims. See Appendix A for full evaluation setup. 15K samples of 3-5 sentences are drawn from clusters acquired using our clustering algorithm and automatically labeled, where semantic similarity for task (2) is measured using S-BERT. We acquire a cluster quality score of 4.17 and a missing rate of 13.79%, indicating that the cluster quality is generally strong and misses only a relatively small proportion of unclustered claims. See Appendix E for a comparison to using semantic-similarity only.

## 5 Measuring Epistemic Diversity

After obtaining $\mathcal{X}_{m_t}$, we must quantify its diversity to rank different models. There are a number of possible choices to measure diversity based on categorical distributions (Estève et al., 2025), for example, richness (the number of unique meaning classes) or entropy. However, we desire a measure which accounts for the distribution of claims (ignored by richness) and allows for meaningful comparison between settings (impossible for entropy due to its logarithmic scaling). An appropriate measure of diversity that accounts for both of these issues is Hill-Shannon diversity (Hill, 1973; Jost, 2006), a special form of the general Hill diversity which balances emphasis between frequent and rare claims 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):

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

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 roughly the same number of claims, $n$; $a$ is less diverse, so $n$ samples may be sufficient to fully characterize $\mathcal{X}_{a_t}$, and sampling more claims is unlikely to have a substantial impact on $D_S(\mathcal{X}_{a_t})$. On the other hand, sampling more claims from $b$ 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 use a combination of coverage estimation and rarefaction (Chao and Jost, 2012; Roswell et al., 2021). Coverage estimates how completely we have covered the true distribution and 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

How epistemically diverse are LLMs? To answer this broad question, we pose several sub-questions focused on how epistemic diversity has changed

[3] Huggingface ID: microsoft/deberta-large-mnli
[4] Huggingface ID: all-MiniLM-L6-v2

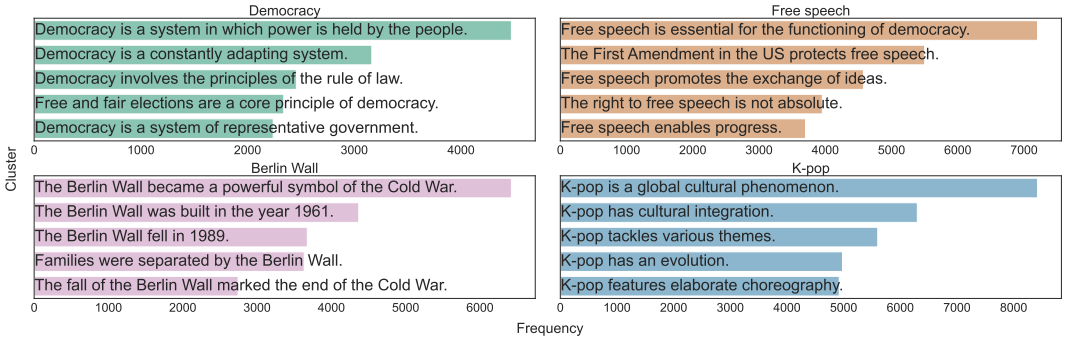


Figure 2: Histograms of the top five clusters for four topics after generating text, decomposing, and clustering decomposed claims across all models in our study. The frequency of claims in each cluster, $x_i$, is represented by the colored bars. By the 5th cluster, $x_i$ is halved for all four topics, indicating a large decay rate for $x_i$. The top clusters convey broad and general information for each topic.

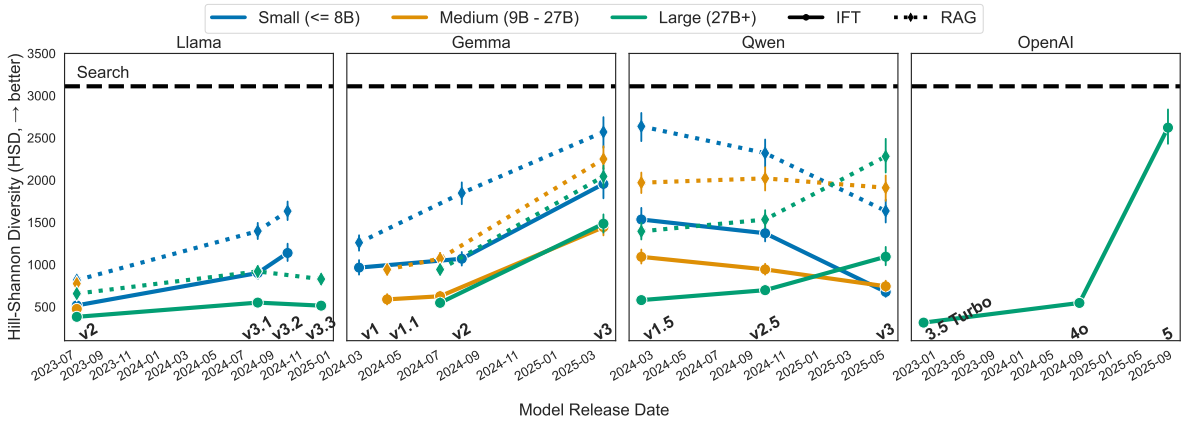


Figure 3: Epistemic diversity vs. model release date. Each point is a single model, with lines connecting models of approximately the same size across released versions. Error bars are 95% boostrapped confidence intervals based on the HSD of each topic (N=155). Absolute diversity is low for all models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

over time **(RQ1)**, how generation setting impacts epistemic diversity **(RQ2)**, how size and model family impact epistemic diversity **(RQ3)**, and how the topic associated with a country impacts diversity and representation **(RQ4)**.

### 6.1 Models, Settings, and Topics

**Models and Settings** We select 27 LLMs across 4 model families: 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), with small, medium, and large models variants, thus including multiple release dates (between 2023-2025) and sizes (1B to 70B). Table 4 provides the details of all models. 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, we include as a baseline the top 40 web pages retrieved from Google search for each topic.[5] To reflect a realistic RAG setup we use the top 20 search results per topic as our database, 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. We additionally shuffle the ranks of all paragraphs with a cosine similarity to the query above the mean cosine similarity across all paragraphs and prompts in the data to 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 in our web pages).

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. Using the top 40 search results leads to generally

[5] Search performed on January 1, 2026 in the US region.

much lower coverage than the LLMs, so we only rarefy search if the coverage is above the minimum achieved by any LLM. As such, our search results are a lower-bound baseline; the true diversity of search relative to the LLMs tested is higher.

**Topics** 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 across 12 countries about important figures and historical events to help measure the impact of country-specific context. 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". [6] Following IssueBench, these include a mix of 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 varied interpretations. All events occurred before the earliest knowledge cutoff of any model in our study. In total, we have 155 topics (see Appendix D for the list), yielding 1.7M responses prior to decomposition and 70M claims after decomposition.

**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 $k^{\text{th}}$ 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 5 clusters for four topics (democracy, free speech, Berlin Wall, and K-pop) in Fig. 2. The frequency of claims drops rapidly; high-frequency clusters unsurprisingly provide broad, common information about each topic.

### 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 all topics and settings 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.

Our empirical results reveal that generally (with the exception of Qwen) epistemic diversity is increasing over time, countering prevailing narratives around static lack of diversity (Jiang et al., 2025) and knowledge collapse (Peterson, 2025). This is the first evidence that LLM outputs are improving along this dimension. This is particularly pronounced for recent models: Gemma 3 and GPT-5 show the largest single-generation gains in our study, suggesting that recent model developments are positively contributing to epistemic diversity.

Despite this, every model in our study generates less diverse claims than our search baseline. This means that people in aggregate will see more diverse information about our topics by reading the top web pages rather than prompting an LLM in different ways. The search baseline is also a *weak* baseline, as we use only up to the top 40 Google search results for each topic based on simply using the topic as a search term (e.g., "democracy"). Further, as mentioned in § 6.1, the search baseline is *under-estimated* relative to the LLM results due to coverage constraints. In absolute terms, we can thus conclude that the most diverse model in our study (GPT-5) is *at least 18.6% less diverse than a traditional search in terms of the claims it generates* (HSD of 2621 vs. 3110), at least for our topics. Therefore, while it is encouraging that LLMs have become more epistemically diverse over time, there is still a gap, which may risk future epistemic problems as LLMs come to replace other means of knowledge acquisition, like search.

### 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 baselines. 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.

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

| Predictor | $\widehat{\beta}$ | $p$-value |
|---|---|---|
| Intercept (IFT) | 939.90($\pm$223.67) | $p \ll 1e{-3}$ |
| RAG | +656.50($\pm$33.83) | $p \ll 1e{-3}$ |
| Search | +2170.85($\pm$1183.57) | $p \ll 1e{-3}$ |

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. RAG search are statistically significantly more diverse than IFT.

| Predictor | $\widehat{\beta}$ | $p$-value |
|---|---|---|
| Intercept (Med.) | 910.34($\pm$228.07) | $p \ll 1e{-3}$ |
| Small | +219.39($\pm$44.28) | $p \ll 1e{-3}$ |
| Large | -185.93($\pm$45.46) | $p \ll 1e{-3}$ |

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.

From Table 2, we see that RAG and search yield significantly more diverse outputs than parametric memory, at least for our topics. This highlights the *potential* of RAG in ensuring that models are epistemically diverse. 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 ask **RQ3: How does model selection impact epistemic diversity?** From Fig. 3, we see that model size appears to have a *negative* impact on diversity. To test this, 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) and the generation setting (IFT or RAG) as fixed effects, and the model version as random effects. The results in Table 3 show that model size has a statistically significant effect, echoing contemporary work showing that large models memorize more (Morris et al., 2025) and can be less diverse (Zhang et al., 2025; Rassin et al., 2024) than small models. We hypothesize that this occurs due to their greater capacity and thus greater propensity to memorize (Morris et al., 2025), therefore narrowing the output token distribution given prompts about the same topic. This has practical implications: the performance gains from large models may come with a cost in the form of decreased diversity in the information those models can generate.

### 6.5 Epistemic Diversity Across Countries

Aggregate diversity scores can hide inequality in knowledge representation across countries. Having established that diversity is improving in aggregate but still lags search, we now ask a harder question: **RQ4: Whose knowledge is represented by LLMs?** We examine this through two lenses: how the country associated with a topic affects diversity, and whether LLMs reflect English-language or local-language knowledge for country-specific topics.

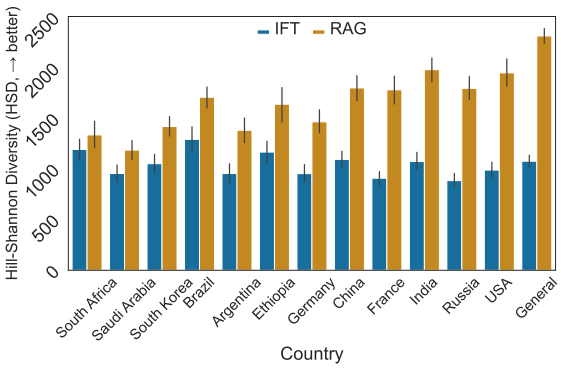


Figure 4: Average diversity per country with bootstrapped 95% confidence intervals. Bars sorted by the difference between RAG and IFT diversity. RAG appears to have an uneven impact on different countries: the US and general topics see higher benefit.

First, how does the source country of a topic impact epistemic diversity? To answer, we plot HSD vs. country as a bar chart in Fig. 4. 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 country 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,

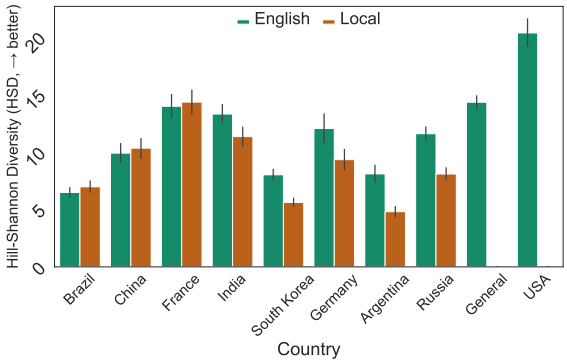


Figure 5: Comparison of the diversity of claims generated by LLMs matched to English and local language Wikipedia. To capture minimal representativeness (Peterson, 2025), we restrict the measurement of HSD only to matched claims. Bars are sorted by the difference between English and local language HSD.

which may under-represent topic-specific information from certain countries. To check this, we measure Pearson's $r$ 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 is English vs. local language knowledge represented? To answer this, we match LLM claims to Wikipedia claims in both English and the country's local language using InfoGap, a recent approach which uses GPT-4 and multilingual sentence embeddings to match claims between Wikipedia pages in different languages (Samir et al. 2024, see Appendix G 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. We use only the topics which are relevant for those countries. 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 "any item should have at least *some* chance of appearing in the LLM output." Results are plotted as a bar chart in Fig. 5, aggregated across models and topics for each country.

We observe a statistically significant gap between English and local language representation for 5 out of 8 countries. In no case is local representation statistically significantly greater than English representation. This means: a person asking an English LLM about one of our topics, in English, is more likely to receive a response grounded in how that topic is described in English than in how it is understood and documented locally. Additionally, representation for the USA specific topics are statistically significantly greater than every other country. This suggests that current English LLMs fail to present knowledge which has not been presented in English for certain countries, echoing recent work on cultural bias in LLM encoded knowledge (Wang et al., 2025) and highlighting a risk in terms of knowledge *erasure*, and thus collapse.

## 7 Discussion and Conclusion

This paper asks: how diverse are the claims presented in LLM outputs, over time and across countries? We answer this by proposing a new methodology for measuring epistemic diversity and study *what* and *whose* claims are generated across 27 LLMs and 155 topics, analyzing 1.7M responses and 70M claims. Our results reveal a nuanced trajectory which the prevaling work on diversity and knowledge collapse has not found (Jiang et al., 2025; Peterson, 2025): epistemic diversity has been improving in the past few years, but there remain salient gaps. All models are still less diverse than a traditional web search on our topics. For users, this means that relying on LLMs as a primary knowledge source means accepting less diverse information than a simple web search could reveal. For model developers, scaling alone does not improve epistemic diversity, and in fact can harm it, while RAG with diversified sources can provide a strong lever for increasing diversity. The choice of RAG database is also critical for improving epistemic representation across countries, as there is a strong correlation between RAG source diversity and downstream diversity. Finally, compared to Wikipedia in both English and local languages for country-specific topics, the claims generated in our study reflect English language claims more than local language claims. This highlights a need to shift focus from measuring and improving diversity in the abstract to measuring and improving diversity for the benefit of the breadth of different communities that use LLMs in practice.

## 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. We also use only 200 different prompt templates due to computational contstraints. Geater coverage could be achieved through more sampling.

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. Additionally, we adopt a strict notion of semantic equivalence for clustering in our methodology, while many real world claims may have some partial overlap in information which could be useful to capture when measuring diversity. Stronger NLI approaches also exist, such as using an LLM, but we choose a validated and more efficient BERT based approach at the cost of some accuracy. To illustrate the cost, we perform ≈744M NLI inferences in our study; assuming ≈40 tokens per inference, with GPT-5, this would cost ≈$44,640.

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.

## 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. 2025. LLM-generated messages can persuade humans on policy issues. *Nature Communications*.

Daniele Barolo, Chiara Valentin, Fariba Karimi, Luis Galárraga, Gonzalo Méndez, and Lisette Espín-Noboa. 2025. Whose Name Comes Up? Auditing LLM-Based Scholar Recommendations. *arXiv preprint arXiv:2506.00074*.

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*.

Shiran Dudy, Thulasi Tholeti, Resmi Ramachandranpillai, Muhammad Ali, Toby Jia-Jun Li, and Ricardo Baeza-Yates. 2025. Unequal Opportunities: Examining the Bias in Geographical Recommendations by Large Language Models. In *Proceedings of the 30th International Conference on Intelligent User Interfaces (IUI)*. ACM.

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*.

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.

Louis Estève, Marie-Catherine de Marneffe, Nurit Melnik, Agata Savary, and Olha Kanishcheva. 2025. A survey of diversity quantification in natural language processing: The why, what, where and how. *arXiv preprint arXiv:2507.20858*.

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

Henry Farrell, Alison Gopnik, Cosma Shalizi, and James Evans. 2025. Large ai models are cultural and social technologies. *Science*, 387(6739):1153–1156.

Yanzhu Guo, Guokan Shang, and Chloé Clavel. 2025. Benchmarking Linguistic Diversity of Large Language Models. *Transactions of the Asociation for Computational Linguistics*, 13:1507–1526.

Shirley Anugrah Hayati, Minhwa Lee, Dheeraj Rajagopal, and Dongyeop Kang. 2024. How Far Can We Extract Diverse Perspectives from Large Language Models? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5336–5366. Association for Computational Linguistics.

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.

Liwei Jiang, Yuanjun Chai, Margaret Li, Mickel Liu, Raymond Fok, Nouha Dziri, Yulia Tsvetkov, Maarten Sap, Alon Albalak, and Yejin Choi. 2025. Artificial Hivemind: The Open-Ended Homogeneity of Language Models (and Beyond). In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*.

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

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*.

Kibum Moon, Adam E Green, and Kostadin Kushlev. 2025. Homogenizing effect of large language models (LLMs) on creative diversity: An empirical comparison of human and ChatGPT writing. *Computers in Human Behavior: Artificial Humans*, page 100207.

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*.

Sonia K. Murthy, Tomer D. Ullman, and Jennifer Hu. 2025. 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 (NAACL)*, pages 11241–11258. Association for Computational Linguistics.

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.

Vishakh Padmakumar and He He. 2024. Does Writing with Language Models Reduce Content Diversity? In *The International Conference on Learning Representations (ICLR)*.

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

Raphael Poulain, Hamed Fayyaz, and Rahmatollah Beheshti. 2024. Bias patterns in the application of llms for clinical decision support: A comprehensive study. *arXiv preprint arXiv:2404.15149*.

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

Royi Rassin, Aviv Slobodkin, Shauli Ravfogel, Yanai Elazar, and Yoav Goldberg. 2024. GRADE: Quantifying Sample Diversity in Text-to-Image Models. *arXiv preprint arXiv:2410.22592*.

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.

Chantal Shaib, Yanai Elazar, Junyi Jessy Li, and Byron C. Wallace. 2024. Detection and Measurement of Syntactic Templates in Generated Text. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Michal Shur-Ofry, Bar Horowitz-Amsalem, Adir Rahamim, and Yonatan Belinkov. 2024. Growing a Tail: Increasing Output Diversity in Large Language Models. *arXiv preprint arXiv:abs/2411.02989*.

Alexander Shypula, Shuo Li, Botong Zhang, Vishakh Padmakumar, and Kayo Yin andOsbert Bastani. 2025. Evaluating the Diversity and Quality of LLM Generated Content. In *COLM*.

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*.

Qihan Wang, Shidong Pan, Tal Linzen, and Emily Black. 2025. Multilingual Prompting for Improving LLM Generation Diversity. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing(EMNLP)*. Association for Computational Linguistics.

Emily Wenger and Yoed N. Kenett. 2025. We're Different, We're the Same: Creative Homogeneity Across LLMs. *arXiv preprint arXiv:2501.19361*.

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*.

Yue Yu, Yuchen Zhuang, Jieyu Zhang, Yu Meng, Alexander J. Ratner, Ranjay Krishna, Jiaming Shen, and Chao Zhang. 2023. Large Language Model as Attributed Training Data Generator: A Tale of Diversity and Bias. In *Advances in Neural Information Processing Systems (NeurIPS)*.

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)*.

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

# A Full Evaluation Details

## A.1 Decomposition

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.) We label 100 input chunks for quality (1) and 100 instances for missing (2), all anonymized for the prompt type (P1-P3). Between the two annotators, we achieve a Pearson correlation of 0.74 for (1) and 0.58 for (2), indicating moderate to strong agreement and the overall feasibility of the task. In order to resolve disagreements, we have a third annotator label those instances where the original annotators disagreed, followed by rounds of discussion and feedback among the annotators to decide on the final labels.

We then set up an LLM-as-a-judge using G-Eval (Liu et al., 2023), which achieves 0.6 Pearson correlation with the human labels for decomposition quality and 0.68 for the number of missing claims. We use this to automatically label 6k instances across three decomposition prompt variants (P1-P3). The results are given in Table 1. We use P3 for the final decomposition in order to prioritize quality.

**Decomposition Quality Instructions** 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 Instructions** 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).

### A.2 Clustering

Two independent annotators label: (1) the degree of *cohesion* among the clustered claims (Likert score between 1 and 5), and (2) whether any claims are *missing* from a given cluster (binary yes/no). We define cohesion as the degree to which each piece of text in the cluster conveys at least one piece of information in common. We define missing whether semantically similar claims which were not predicted to be included in a given cluster (i.e., singletons) *should have* been grouped with said cluster. We label 100 instances of cohesion (1) and 360 instances of missing (2), achieving a Pearson correlation of 0.44 for (1) and 87% agreement accuracy for (2). Due to the moderate correlation in (1), we had a follow-up discussion and tie-breaking session which included a third annotator. Using the resolved annotations we then set up an LLM-as-a-judge using G-Eval in order to annotate and evaluate the rate of cluster cohesion and missing sentences in our final clusters. 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 **4.17** and a missing rate of **13.79%**, indicating that the cluster cohesion is generally of high quality while missing only a relatively small proportion of singletons. We compare this to a semantic similarity only approach using DBSCAN in Appendix § E.1 (Wright et al., 2024), showing that cluster cohesion degrades to 3.48.

**Clustering Cohesion Instructions** 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 Instructions** 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

## B Prompts

Here we show decomposition prompts P1 (Fig. 6) and P2 (Fig. 7) 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.[7]

> 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 6: Decomposition prompt P1.

## C Clustering Algorithm

Algorithm 1 outlines our approach for **clustering** the decomposed claims $C_{m_t}$ into semantically equivalent meaning classes $\mathcal{X}_{m_t}$

[7] https://anonymous.4open.science/r/knowledge-collapse-9166

> 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 7: Decomposition prompt P2.

## D 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 Dos-

**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 ← group(L);
# Break up large clusters
forall x ∈ 𝒳 do
    if len(x) > 100 then
        𝒳′ = DBSCAN(x);
        𝒳.pop(x);
        𝒳.extend(𝒳′);
return 𝒳;
```

toevsky; 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 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;

## E Comparison to Semantic Similarity Only

Our aim, reflected in our problem setup and formulation, is to measure how diverse are the unique pieces of knowledge that different people can be exposed to when interacting with an LLM. To contrast our methodology with similar contemporary work on diversity (Lee et al., 2025; Wright et al., 2024; Jiang et al., 2025), an alternative version of our setup would be to partition claims based solely on semantic similarity, i.e., the distance between claims in a semantic embedding space. However, semantic similarity can only describe the degree to which two pieces of text are functionally similar, and only in extreme cases whether they convey the same piece of information. To illustrate: the phrases "Claude Shannon is the father of information theory" and "Claude Shannon is not the father of information theory" have high semantic similarity (0.94)[8], 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). Therefore, while semantic similarity can be useful in order to find candidate pairs which are semantically equivalent, partitioning based on

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

| 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.

similarity alone would lead to classes which contain potentially many different pieces of knowledge, thus underestimating diversity and not reflecting our definition of meaning classes.[9] To highlight this in a concrete way, we now contrast our methodology with two recent popular approaches to diversity measurement which use semantic similarity, in order to demonstrate what is gained by focusing on semantic equivalence of claims as the unit of knowledge.

[9]We will demonstrate this empirically by comparing our methodology to two previous approaches based solely on semantic similarity (Artificial Hivemind by Jiang et al. 2025 and LLM tropes by Wright et al. 2024).

### E.1 LLM Tropes (Wright et al., 2024)

What about partitioning into classes based on semantic similarity? We now compare to the approach from Wright et al. (2024) which does exactly this. This is performed by taking the decomposed claims from our study, transforming them into sentence embeddings, and using DBSCAN (Ester et al., 1996) to partition claims into unique classes. Note that DBSCAN parameters require tuning — for the sake of comparison, we selected reasonable parameters which resulted in a comparable number of clusters to the semantic equivalence case (eps of 0.2, min size of 2), but better results could potentially be achieved by hyperparameter tuning. After partitioning claims as such, we perform the same LLM-as-a-judge evalu-

ation as in § 4.2, acquiring a cluster cohesion score of 3.48 (vs. 4.17 for our algorithm) and missing singleton rate of 0.4% (vs. 13.8% for our algorithm). As such, the recall of these clusters is higher (reflected in fewer singletons which should have been clustered), but the overall cohesion goes down, reflecting a tradeoff in terms of clustering quality.

### E.2 Artificial Hivemind (Jiang et al., 2025)

We now compare to the approach used in Jiang et al. (2025). In this work, diversity within LLM responses is measured using the average pairwise cosine similarity between responses represented as sentence embeddings. In other words, each LLM response is passed through a sentence embedding model (OpenAI's `text-embedding-3-small` API in the original paper, Huggingface ID : `all-MiniLM-L6-v2` here), and then the cosine similarity of each pair of embeddings is measured and averaged. We perform exactly this procedure on the original LLM responses of each model within each topic in our study, and plot diversity over time and across models (exactly as in Fig. 3) in Fig. 8.

What does this plot tell us? We can make a few claims: models lack semantic diversity to approximately the same degree, there has been little change over time with the exception of Gemma 1 and GPT-5, and traditional search is more semantically diverse than LLMs. However, in the context of our study on the risk of AI knowledge collapse these results are not particularly meaningful or interpretable. First, compared to our experiment measuring epistemic diversity, we lose a lot of valuable information. We see no difference with RAG or by model size, and the general improvement over time that we saw with epistemic diversity is lost. This is because semantic embeddings aggregate many different aspects of meaning, while our methodology isolates a more fine-grained and targeted signal, namely, semantically equivalent claims. That is to say, the results we acquire in Fig. 5 are unable to be seen in Fig. 8 as there are too many confounding factors to observe the specific trends we see using our methodology. Second, the $y$-axis in Fig. 8 is much more difficult to interpret than Fig. 3 — whereas when measuring HSD based on claims, we can directly make statements about the effective diversity of claims that different models generate, cosine similarity between sentence embeddings is an opaque measure. For example, the average cosine similarity improved by 0.2 from GPT 4o to GPT-5, but this improvement lacks any practical interpretation other than "semantic diversity increased to some degree." On the other hand, from Fig. 3 we can say that the effective diversity of claims that one can expect to encounter when interacting with GPT-5 is more than 5× greater than with GPT 4o for our topics.

## F 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.

## G 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,[10] and predict if the retrieved multilingual claims match to the LLM generated English claims using GPT-5 (over GPT-4 as used in (Samir et al., 2024)). We exclude claims generated using RAG to avoid Wikipedia text being included in the generation context.

## H Generation Details

We retrieve 40 Google search results for each topic. Over the 155 topics, these 6,400 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. All generation was run on NVIDIA A100 (40GB) and H100 (80GB) GPUs depending on the size of the model. 70B parameter models required 2 H100s for generation.

## I Artifacts Statement

All data artifacts used in this project are open source, publicly available, and intended for research purposes. Generated data and code is available at `https://anonymous.4open.science/r/knowledge-collapse-9166` and will be made publicly available upon acceptance under the MIT license. Data is intended for research purposes.

## J Use of AI in Research Statement

We made minimal use of Claude in the final polishing stages of preparing the manuscript. This was primarily in getting feedback on the draft in terms

[10] HuggingFace: sentence-transformers/LaBSE

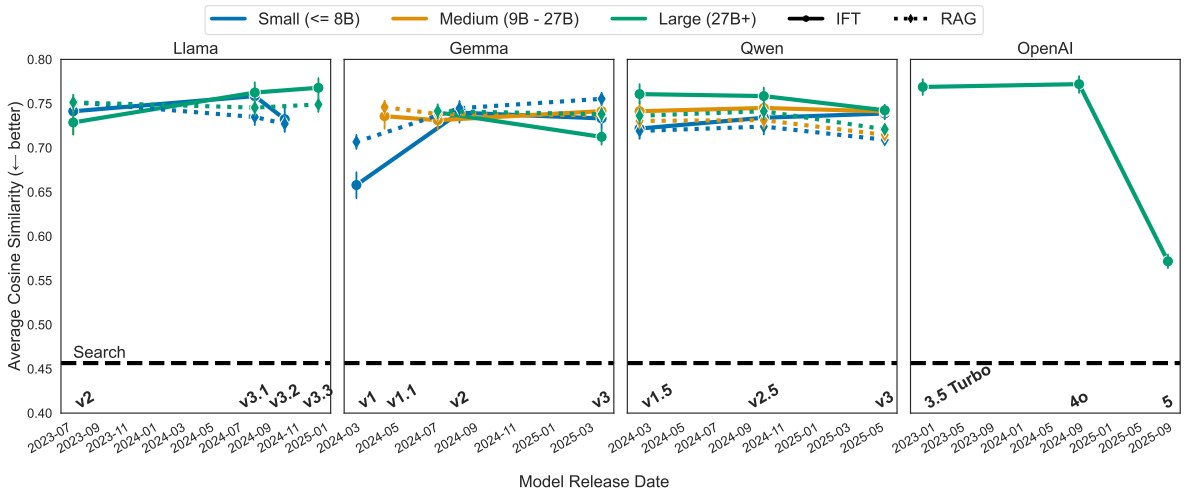


Figure 8: Average pairwise cosine similarity of within-topic model responses vs. model release date. Each point is a single model, with lines connecting models of approximately the same size across released versions. Error bars are 95% boostrapped confidence intervals based on cosine similarity within each topic (N=155). **Lower scores mean greater diversity**. While we see that search is still much more diverse than LLMs, many of the trends identified via epistemic diversity are lost.

of potential ways to improve the figures, consistency of claims, and framing.