"Emphasizing the Commendable": A Study of Homogenized Transitive Verb Constructions in Machine Generated Peer Reviews

Hing-Yuet Fung 馮慶月[a*], **Chi-kiu Lo** 羅致翹[b*], **Samuel Larkin**[b]
[a] Independent Researcher
[b] Digital Technologies Research Centre, National Research Council Canada
kycs@connect.hku.hk, {chikiu.lo,samuel.larkin}@nrc-cnrc.gc.ca
*Equal Contribution

Abstract

We present a study of machine generated text (MGT) output homogenization with a focus on the relative usage of the prototypical object construction of verbs (the O construction), which takes a noun phrase as its accusative argument. Verbs of different semantics have different tendencies of selecting a direct object or clausal complement; and hence lead to natural variation away from the prototypical usage. However, our results in the study between scientific peer reviews written by human and machines show a shift to unusually high usage of the O construction in MGT and greatly suppressing the frequency of other construction types. This is considered a serious case of syntactic homogenization. A major finding is that frequent verbs, like "emphasize", appear top on the list of such homogenized syntactic construction. This is more striking than identifying disproportionately more frequent usage of naturally rare words such as "commendable" in previous work. Our results will contribute to the prevention of further homogenization of MGT before they merge deeper into the ecosystem of human-written text.

Keywords: Syntactic Homogenization, Machine Generated Text, Peer Reviews

1. Introduction

Output homogenization, i.e., the decrease in diversity of style, content, etc., is a newly identified risk in AI/Machine Generated Text (MGT) by Large Language Models (LLMs) (Padmakumar and He, 2024; Moon et al., 2025). LLMs are overtaking individual speakers' roles as carriers for organic linguistic change, and expressive nuances that differentiate different cultures speaking a common language may gradually diminish (Agarwal et al., 2025; Cao et al., 2023).

Individual LLM users who adopt text output by AI often unintentionally contribute to output homogenization at corpus level. As pointed out in Anderson et al. (2024), corpus-level homogenization is realized in a way that is too subtle to grasp by examining individual use cases, and people with an objective to produce creative contents using LLMs are not immune to homogenization effects. The convergence of generated content will be an unavoidable artifact if MGT continues to populate in our information ecosystem.

MGT homogenization effects have mainly been studied as distributional biases in word choices, with little reference to more abstract linguistic features. One notable example of lexical bias was the sudden spike of the usage frequency of specific adjectives like "commendable" among a large number of individual authors in post-ChatGPT academic writing (Liang et al., 2024). In this study, we are able to show that frequent transitive verbs like "emphasize" are also exhibiting usage anomaly when we extend the analysis to deeper linguistic structures (Figure 1).

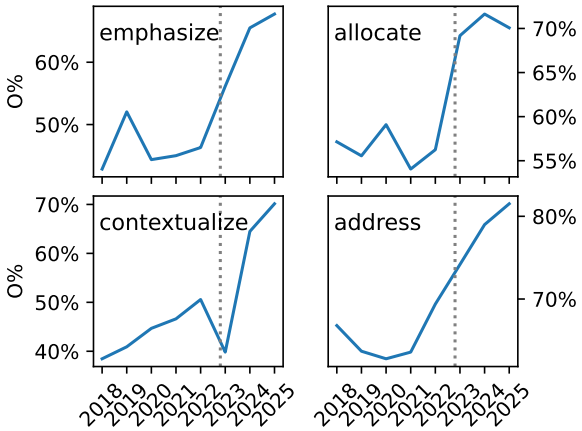


Figure 1: Our study shows upward moving trends in the percentage of the O construction in official peer reviews for selected verbs {emphasize, allocate, contextualize, address}, spanning the periods before and after ChatGPT. The time point of ChatGPT public release is marked by the dotted line.

LLMs are superior in learning today's snapshots of natural languages, but they had no involvement in the contexts that gave rise to the optimized forms of today in the first place. For example, some transitive verbs such as "teach" can take two objects, as in "I teach X Y". Given the right context, elements can be omitted, resulting in instances of "I teach X" or "I teach Y", which has the same surface form as other transitive verbs that take only one object. Different languages often optimize toward the form

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 10664–10676
11–16 May 2026.

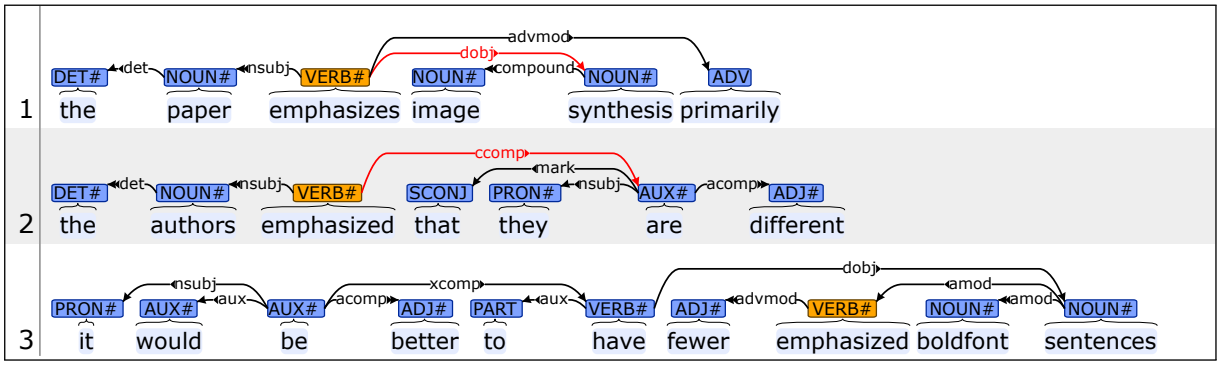


Figure 2: Example construction types of the lemma `emphasize` (highlighted in orange). The obligatory arguments of this transitive verb are indicated by the red arrows: (1) a single direct object (`dobj`; "image synthesis"), henceforth the **O construction**; (2) a clausal complement (`ccomp`; "that ... different"). In the attributive use in (3), it is positioned on a lower level than its object ("sentences" at the end), therefore the relation is only inferred.

that is the most frequently used in the contexts of the community.

Such alternations in syntactic forms can be studied in terms of construction types. Figure 2 shows three example constructions with the verb "emphasize". Example 1 shows the prototypical transitive usage with one noun phrase ("image synthesis") as the direct object (`dobj`). Its dependency schema at the verb phrase level is `[ nsubj VERB dobj advmod ]`. For our purpose, the presence of other elements such as the adverb (`advmod`) or even the subject (`nsubj`) is irrelevant in the comparison with other construction types, such as example 2, which takes a full clause as the complement of the verb. Example 1 may be reduced to `[ VERB dobj ]`, henceforth the **O construction**. Example 3 illustrates another usage of "emphasize" found in the dataset.

In this paper, we focus on the analysis of academic peer reviews. Diversity in peer reviews has an important role in fair evaluation of the body of work. To start with, the genre embraces differences in knowledge, training and expertise among different scientific researchers, naturally leading to diversified academic writing (Sulik et al., 2025). However, we are not the first to show that homogenization effects in MGT is affecting academic writing as well as speaking (Geng et al., 2025; Yakura et al., 2025). Our results will contribute to the prevention of further homogenization as seen in Figure 1. We release our machine generated reviews to support a wider study of machine generated text.[1]

[1] https://huggingface.co/datasets/NRC-CNRC/Machine-Generated-Reviews-0.1

2. Related Work

2.1. Output Homogenization in Reviews

AI-Driven review systems are emerging (Tyser et al., 2024), regardless of studies that show AI-generated reviews to be generally less specific, more favorable, and more confident (Yu et al., 2025; Sadallah et al., 2025). There is a wide consensus on the important role of feedback variation from reviewers (Teplitskiy et al., 2018) and the dangers of algorithmic monoculture (Kleinberg and Raghavan, 2021; Bommasani et al., 2022). We agree that the noticeable semantic homogenization in scientific peer reviews has great impact on the development of scientific research and innovation. However, it is even more concerning that the subtle syntactic homogenization are fundamentally affecting human linguistic usage in unnatural ways.

Liang et al. (2024) uncovered that certain adjectives occur disproportionately more frequently in AI-generated peer reviews in several top ML and scientific venues. Inspired by their work, we seek to investigate abstract linguistic characteristics behind the surface forms of human text and MGT.

2.2. Benchmarking Linguistic Diversity

The study of the linguistic diversity of a corpus includes a variety of linguistic features such as lexical diversity (range of vocabulary), syntactic diversity (variety of sentence structures), and semantic diversity (range of meanings conveyed). Guo et al. (2024) showed that LLM output has a significant decline in all linguistic diversity metrics after iterations of training on MGT. The established metrics are successful in identifying homogenization and it also warns of the cause and consequence of potential contamination in the linguistic ecosystem.

Lexical diversity can be quantitatively described

by common measures such as Type-Token Ratio (Johnson, 1944, TTR) and distinct-n metric (Li et al., 2016), which are readily applied in studies of the lexical diversity of LLMs (Guo et al., 2024; Reviriego et al., 2024). Yarats and Lewis (2018) and Guo et al. (2025) measured semantic diversity as the dispersion of sentence embeddings in the semantic space based on their average pairwise similarity.

In terms of the syntactic diversity in a corpus, there exist several approaches in the literature. Methods include traditional word class n-gram analysis (Ramírez de la Rosa et al., 2013) and syntactic templates analysis (Shaib et al., 2024). Huang et al. (2023) and Guo et al. (2025) define syntactic diversity as the average pairwise distance of the syntactic graph (e.g. constituency parse or dependency parse) embeddings. However, such methods still do not address the level of abstraction required for studies in a language's syntactic characteristics.

2.3. Construction Types vs. Syntactic Templates

Nowadays, the majority of work that attempts to gauge syntactic diversity abandons structural abstraction and adopts certain forms of syntactic templates (Shaib et al., 2024). The syntactic templates pick up all elements in the verb phrase domain, and we consider that essentially similar to distinct-n metric used for lexical diversity.

On the other hand, consideration by construction types will be able to address the distinction between obligatory core arguments and optional non-core arguments, since the number of arguments that a verb can take depends on its semantics. Each verb has a characteristic set of obligatory participants performing different semantic roles. Verb arguments have higher regularity across languages than non-core elements, and these concepts are highly comparable in linguistic typology (Haspelmath, 2014).

Dependency analysis for core and non-core arguments is a well-grounded analysis in the linguistics community. The Universal Dependencies framework, as detailed in De Marneffe et al. (2021), has gained popularity among linguistic typologists as a successful implementation that enables meaningful cross-linguistic comparisons (Croft et al., 2017). In case of elliptic use, the annotation will only assign interpretations to overtly observable forms, and core arguments of a verb will not be indicated if the content is not realized. It may be left to further study to relate the inherent valency of the verb and the surface forms.

3. Data

In this study, we distinguish two main types of reviews: "official reviews" submitted by anonymous reviewers who may or may not have used LLMs during the review process, and "LLM reviews" produced by three selected LLMs regarding the same set of research papers.

Official reviews. We identify four leading AI conference venues as our targets of study: ICLR, NeurIPS, EMNLP, and CoRL, as adopted from Liang et al. (2024). We collect academic research papers submitted to conferences held at these venues, and the corresponding texts submitted as official reviews. Both are available on OpenReview[2]. Multiple textual fields of an official review are concatenated together for analysis, and the research papers are stored for use in the prompts for LLM review generation. Only minimal effort was spent to clean the texts from structuring and extraneous strings, because leading headings such as "Summary:" are not likely to be matched as a verb phrase and will be discarded automatically in the template matching process.

Table 1 shows the number of research papers and the corresponding number of official reviews for the list of conferences used in this study. These numbers are smaller than the actual number of submissions downloaded from OpenReview because we excluded those papers for which any of the LLMs failed to automatically generate reviews due to memory constraints. Conference years span across the release of `ChatGPT`—one of the first consumer-ready LLMs, and the conferences are categorized into "before" or "after" their release. The number of official reviews is roughly three to four times the number of papers, meaning that each paper is reviewed by three to four distinct reviewers, which is a common practice in these academic conferences.

Since the publication of Liang et al. (2024), a few more conferences have taken place at the same venues. They are included without a reference of the detected amount of AI usage.

LLM reviews. Three LLM models are selected to represent the state of the art with a diverse background which also has active shares in the consumer market. The models are `gpt-4o` (OpenAI et al., 2024), `gemma-3-4b-it`[3] (Gemma-Team et al., 2025, `gemma3` hereafter), and `Qwen3-4B-Instruct`[4] (Yang et al., 2025, `qwen3` hereafter). Each model is prompted to generate one review for a given paper. Together, they act as if they were

[2] https://openreview.net/

[3] https://huggingface.co/google/gemma-3-4b-it

[4] https://huggingface.co/Qwen/Qwen3-4B-Instruct-2507

Conference	Timeline	#papers	#reviews
CoRL 2021	before	152	554
CoRL 2022	before	197	756
CoRL 2023	after	192	731
CoRL 2024	after	255	808
EMNLP 2023	after	1913	6112
ICLR 2018	before	909	2745
ICLR 2019	before	1375	4193
ICLR 2020	before	2118	6432
ICLR 2021	before	2447	9457
ICLR 2022	before	2343	9122
ICLR 2023	after	3101	11723
ICLR 2024	after	6422	24738
ICLR 2025	after	8908	36098
NeurIPS 2021	before	2751	10672
NeurIPS 2022	before	2789	10200
NeurIPS 2023	after	2928	13071
NeurIPS 2024	after	3072	12069

Table 1: Number of papers and official reviews for each conference used in our experiments marked with before and after the release of ChatGPT.

three distinct reviewers, and the total number of generated reviews will approximate the number of official reviews shown in Table 1.

To generate automatic reviews, we extract the text of the paper and use the prompt in Figure 3, which is adopted from Liang et al. (2024). The actual format of official reviews varies depending on the conference but we used the same prompt throughout as a generic format. The paper content included in the input will be the biggest variant for soliciting a diversity of output.

We did not provide the paper decisions or the subcategory scores while prompting the LLMs. The models are free to provide positive or negative reviews. While acknowledging the importance of prompt variation to diversify text generation (Yu et al., 2025), we have no interest in the information value of LLM reviews. More relevant, in a detection study focusing on outcome homogenization, the algorithm predicts AI usage with similar accuracy when using a generic prompt like Figure 3 or a prompt written in a distinctly different style (Liang et al., 2024).

4. Methods

Homogenization in transitive verb usage is examined from two perspectives: lexical diversity of verb types in general, and the proportion of the prototypical transitive usage among all surface constructions observed for the verb.

4.1. Lexical Diversity of Verbs

Lexical diversity helps lay the foundation for measuring syntactic homogenization. A verb is an open class of lexical items, and it may take different numbers of arguments because of its semantics, though verbs taking more than two arguments (i.e. ditransitive verbs) are scarce and are reserved for highly grammaticalized verbs like “give” and “teach”. A common transitive verb may be used with an explicit object (examples 1 and 2 in Figure 2), or with no objects given the correct context and other structural constraints (example 3 in Figure 2).

A high verb type diversity entails a collection of verbs with diversified semantics which then also diversify into different construction types. It is generally assumed that with an increase in the number of review texts analyzed, the number of verb types should also increase.

4.2. Surface Constructions

The prototypical transitive usage is defined as the construction where the verb takes an accusative noun phrase as its core argument besides the nominative subject (Hopper and Thompson, 1984). This information can be obtained by processing the review texts with a dependency parser. We use the parser model `en_core_web_lg`[5] of spaCy v3 (Honnibal et al., 2020). Tokenization and Part-Of-Speech information are automatically available after parsing is complete. An example of parsing is shown in Figure 2. Annotations follow the Universal Dependencies formalism. The accuracy for dependency parsing is benchmarked at 92.0%. Manual inspection on a sample of the data suggests that erratic parsing mainly leads to false negatives of the major patterns including the O construction, but false positives are rare.

The list of immediate children under the verb is then subjected to a matching template, which looks for the existence and positions of argument-type constituents. For the current study, we are only interested in the **O construction**, as shown by example 1 in Figure 2. It contains the prototypical noun phrase object, annotated with `dobj` (also known as the direct object). During the matching, non-argument elements such as `advcl` (adverbial clause modifier), `advmod` (adverbial modifier) and `punct` (punctuation) will be ignored. For example, a sentence like “each configuration usually *demands* separate fine-tuning to maintain accuracy,” which is parsed into the schema `[ nsubj advmod VERB dobj advcl ]`, will match our target O construction as `[ VERB dobj ]`.

The construction type `[ VERB ccomp ]`, where `ccomp` annotates the clausal complement (i.e. example 2 in Figure 2), is considered to be more complex. Together with other argument types including `dative` (dative) and `acomp` (adjectival

[5] https://spacy.io/models/en#en_core_web_lg

```
Your task is to write a review given a paper titled <Paper title> and the paper
content is: <Paper content>. Your output should be like the following format:
Summary:
Strengths And Weaknesses:
Summary Of The Review:
```

Figure 3: Prompt for generating LLM reviews. ⟨Paper title⟩ is available from OpenReview's API and ⟨Paper content⟩ is the text extracted from the PDF file of the paper.

complement), and the cases where no arguments are explicitly expressed, they are grouped under "others" in the current study but they will be studied separately in future work.

5. Verb Type Diversity

Verb type diversity is evaluated on the word lemmas extracted by the dependency parser, but we will continue to refer to the Type-Token Ratio (TTR) for simplicity. Verb lemma type is counted once in the collection of reviews per conference, while the token counts include all occurrences of all words in the collection of reviews. We added a ratio of the number of verb lemma types against the number of reviews to have different perspectives of the verb types usage in official reviews and LLM reviews. It is termed the Type-Review Ratio (TRR).

Figure 4 shows the results grouped by conference venues, for official reviews and LLM reviews. The three LLMs show insignificant variation in this regard and are mixed together. Regression lines are drawn to show the trend when the denominator (#tokens or #review) grows from year to year, as observed in Table 1.

The regression lines clearly show that LLM reviews consistently have lower TTRs and TRRs (fewer #verb types per measuring unit) than observed with the official reviews. We have three LLM reviews per paper but the number of official reviews may reach four or more, hence a shorter range of x-values for LLMs in Figure 4b. However, the LLMs in fact produced longer reviews so the #tokens are comparable to official reviews (Figure 4a). LLMs are still scoring lower TTRs and TRRs on this basis. Data for official reviews and LLM reviews also show noticeably different slopes when we calculate with #tokens. This indicates that with the increase in #tokens, the increase in verb types falls behind that of official reviews.

It is remarked that the lower TTRs in LLM reviews does not only scale down the frequency distribution as found in official reviews but the frequency ranks are changed. Excess words, as identified in Kobak et al. (2025), such as "delve" are also found with unusually large usage frequencies in our data.

As reasoned in Section 4.1, verb type diversity is the foundation for other linguistic diversity. This result shows that there is already a homogenization effect on verb lemma types in LLM reviews. We continue to investigate other homogenization effects that may or may not be directly related to verb type diversity.

6. Diversity in Transitive Verb Constructions

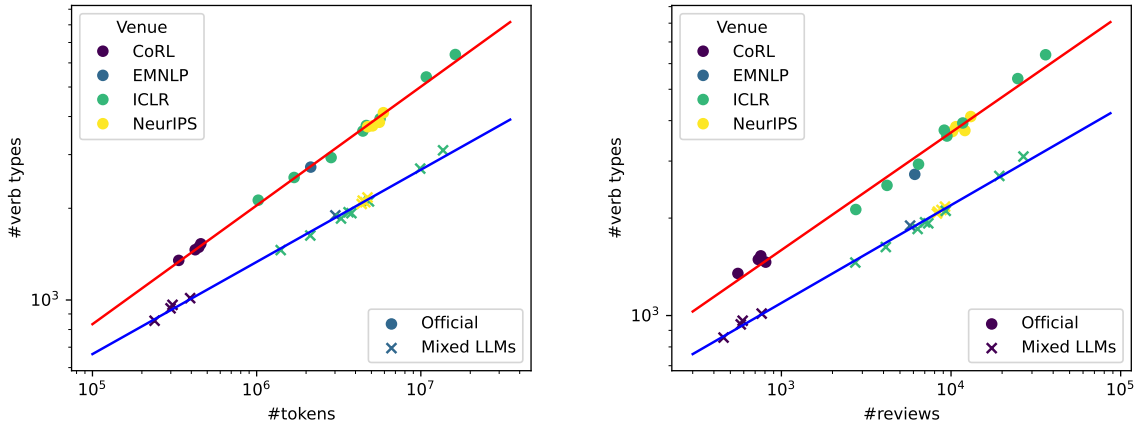
The diversity in transitive verb constructions is studied by aggregating the prototypical transitive usage (O constructions) against all other types of usage. O occurrences are counted once per review, to account for repetitive use within the same review text. The key metric is the proportion of O occurrences divide by the total number of construction type occurrences counted in the same manner, and is represented as O% for each verb lemma type.

6.1. O% Distributions

Figure 5 visualizes the distribution of O% in a variant of the beeswarm plot, which may be understood as a less rigorous version of histograms, and the radius of each circle is representing the lemma's token frequency. Data is grouped in two factors: official reviews against the three LLM models (`gemma3`, `gpt-4o`, and `qwen3`), and whether the review came from the before- or after-ChatGPT period. For a more readable graph, the visualization filters out lemmas with a token frequency lower than 30 in the respective group. Text labels are only applied for lemmas with top token frequencies.

We exclude lemmas with extreme values of O% with the understanding that such inanimate usage (either always used in O constructions or never used in O constructions) does not contribute to the discussion of homogenization. The cutoff thresholds are set to filter out verb lemma types with an O% bigger than 90% or lower than 10%. The figures are clipped to an arbitrary height to save space. A red density line is provided to show the distribution of all data within this range, without other filtering, bucketed by the verb type's O%.

The official reviews (Figures 5a and 5b), especially the before-ChatGPT group, are allegedly authored by human, and the distribution may be taken as representative of other naturally existing



(a) Type-Token Ratio of verb lemma

(b) Type-Review Ratio of verb lemma

Figure 4: Verb type diversity analysis of official reviews and generated reviews.

resources. The distributions are not uni-modal, and after-ChatGPT official reviews show local peaks at various values. Before further analysis may be applied, they are speculated to represent clusters of verbs predominantly taking different types of arguments, for example, manipulation verbs like "allow" take a clausal complement as its argument much more often than movement verbs like "move", where "move" also doubles as an intransitive verb taking no arguments.

A striking observation from the density lines is how all those for LLM reviews show a rocketing tail on the side of high O%, and the lemmas pile up to a height that exceeds the figure frames. There is a tendency for the lemmas to aggregate to a high O usage in LLM reviews. In contrast, the middle range is left with much fewer lemmas. `gemma3` shows the highest sparsity in the range between 20% to 60%. `qwen3` shows a local peak at about 50% and resembles the distribution for official reviews, but it still shows a similar tower of lemmas at about 90%, like the other two models.

6.2. Skewness

In order to quantify the distributional difference between official reviews and LLM reviews, the Fisher–Pearson coefficient of skewness (g_1, Kokoska and Zwillinger 2000) is calculated for each subgroup, while excluding extreme values of O% (i.e., bigger than 90% or smaller than 10%). Values are presented in the caption for each subgroup in Figure 5.

Skewness is interpreted by the difference from zero skew, the purely symmetrical distribution. A positive value indicates a right skew. It happens when the mean value is higher than the median value, and data points tail off to the right hand side away from the median. A negative value indicates a left skew. A value approaching -1 indicates a highly skewed distribution with a long tail trailing to the left.

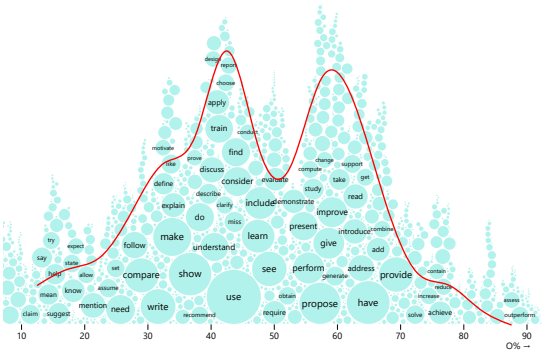
For official reviews, the coefficients of O% skewness is very close to 0 in before-ChatGPT period, and is at about 0.1 in after-ChatGPT period, the latter still indicates a very mild skew. The high verb type diversity in official reviews seems to ensure a higher diversity in terms of O usage, resulting in close mean and median values.

The magnitudes of g_1 for the LLM reviews are all larger than those of the official reviews, by two orders of magnitude in absolute value in the pre-ChatGPT period. The O% distributions for LLM reviews all skew to the direction of high O%. Consistent with the observation of the density lines, the skewness indicates that there are long tails of sparse data on the side of lower O%.

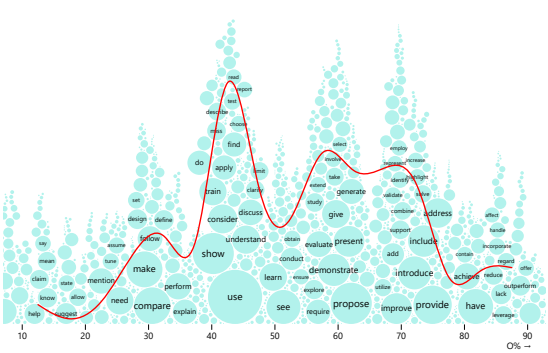
6.3. Homogenized Lemmas

We define homogenization in the usage of the O construction type by considering if verb lemmas already high in O usage tend to use even more O constructions in MGT. Utilizing the same O% values from previous sections, we calculate a change in usage by subtracting the O% of a verb lemma type, from that in the reviews generated by respective LLM models. The difference is denoted as ΔO%.

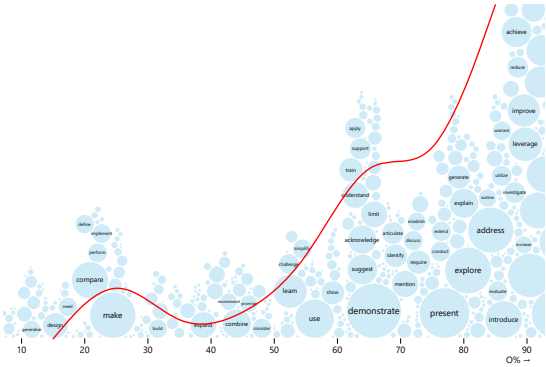
Figure 6 shows the differences in O% between LLM reviews and official reviews in the respective period before and after ChatGPT, for verb lemmas having O%$>$45% in official reviews. The notches on the box plots represent the medians, and all distributions are centered above the neutral value of 0%. The third and first quartiles span between 0% to 20%. A difference of 40% is close to the largest possible gap, as we include only lemmas



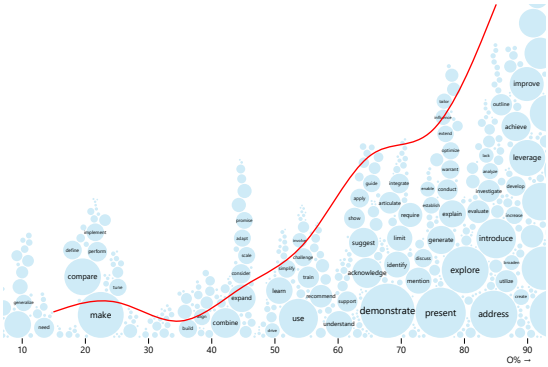
(a) Official reviews; before ($g_1 = .002$)



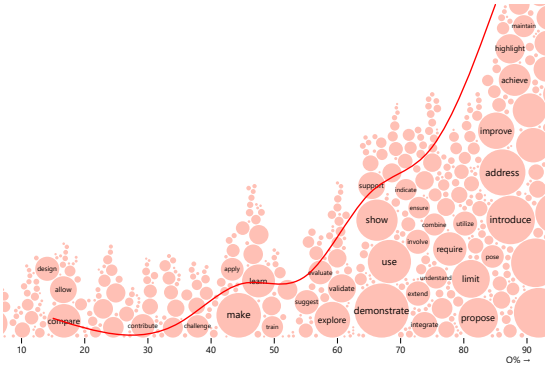
(b) Official reviews; after ($g_1 = .114$)



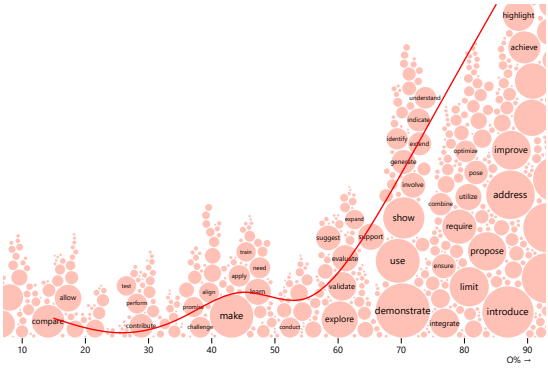
(c) `gemma3`; before ($g_1 = -.753$)



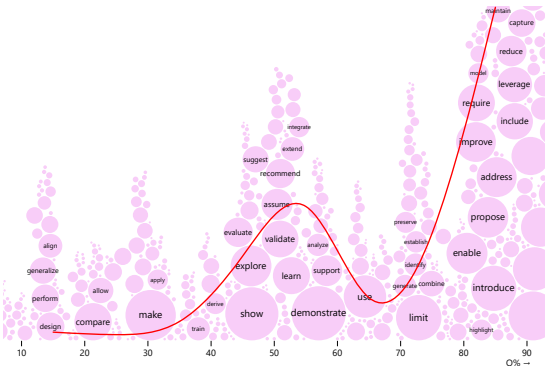
(d) `gemma3`; after ($g_1 = -.801$)



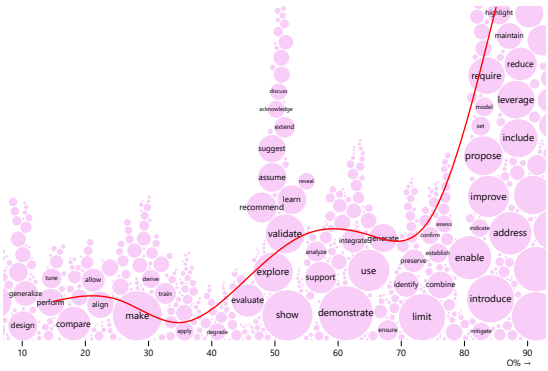
(e) `gpt-4o`; before ($g_1 = -.799$)



(f) `gpt-4o`; after ($g_1 = -.923$)



(g) `qwen3`; before ($g_1 = -.379$)



(h) `qwen3`; after ($g_1 = -.504$)

Figure 5: Beeswarm plots of O% for verb lemma types in official reviews and the three LLM models (`gemma3`, `gpt-4o`, and `qwen3`), grouped into the before- or after-ChatGPT period. Radius of circle indicates a lemma's token frequency. The red line shows the density.

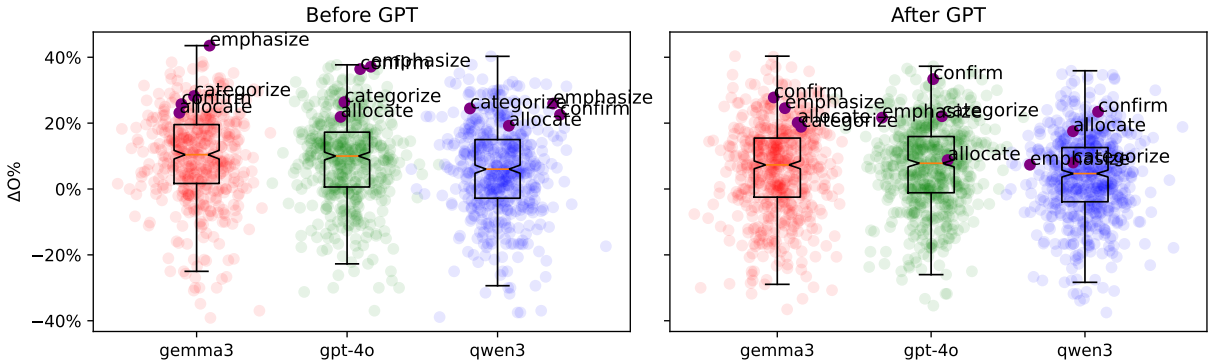


Figure 6: Distributions of the differences in O% (the proportion of the prototypical usage among all construction types) between LLM reviews and official reviews, in the before- and after-ChatGPT periods respectively. We include only verb lemmas having O%>45% in official reviews. Top homogenized lemmas are labeled for comparison.

emphasize, treat, categorize, object, streamline, expose, resolve, confirm, mirror, allocate

Figure 7: Top 10 common homogenized verb lemmas in the generated reviews by all models.

with no less than 45% in reference value. The implication is certainly a case of homogenization as the concerned lemmas are found in a much smaller portion of other construction types if O% achieves 90.

Figure 7 lists the top homogenized verb lemmas in LLM reviews that are common to all models. These verbs are frequent in daily usage, and are popular in the review genre. This implies that our methodology is able to capture peculiarity of MGT not only with rare words like "commendable" but also frequent words like "emphasize" at the top of the list, together with "confirm", "categorize", etc.

A few of the top homogenized lemmas are highlighted with a dark purple dot in Figure 6. For the highlighted lemmas specifically, there is a general decrease in ΔO% after the launch of ChatGPT. Given that the O% in LLM reviews is distributed similarly in both periods (Figure 5), the smaller gap between LLM and official reviews after ChatGPT is explained by the increase in O% usage in official reviews.

7. Discussion

According to Liang et al. (2024), the amount of AI-modified content from the same set of official reviews, excluding a few conferences held after their publication, is detected at up to 16.9%. This aligns with our observed difference in distribution between the periods before and after ChatGPT. It requires further investigation to determine what exactly contributed to the difference between Figure 5a and 5b, but the results from LLM reviews are too striking to be ignored. Our study will fill a knowledge gap in the literature of linguistic diversity as well as MGT detection.

This study focuses on the O construction, which is only prototypical to certain categories of transitive verbs. From a functional view of linguistics, prototypicality can be understood as the solidarity at the core of categories but it also allows for the flux at the margins (Givón, 1995). Blended with a variationalist view of language diffusion and change (Labov, 1969, 2008) and natural irregularization in language evolution (Trudgill, 2011), the seemingly chaotic variation in O% among official reviews before ChatGPT (Figure 5a) speaks well for the diversity in language use as a collective act. Despite the claimed ability to diversify, LLM outputs are showing a high degree of homogenization. Some estimate that today's models must be scaled by many orders of magnitude to reach competitive performance.

It is anticipated that a new challenge for MGT evaluation is arising from the blossoming release of new LLMs, after which the dataset collected in similar manner as Liang et al. (2024) or the current study will no longer be suitable for extensive study of MGT output homogenization, because it will be difficult to understand whether such homogenization effects are cross-model or model-specific without the knowledge of the origin of the MGT. Evaluation with deeper linguistic analysis may pay a key role in such a future.

8. Conclusion

In this paper, we study the problem of output homogenization in machine generated text with a focus on the relative usage of the prototypical object constructions with a noun phrase as the accusative argument (the O construction). The prototypical

usage is subject to natural variation as different categories of verbs have different tendencies of adopting the O construction or alternatives such as a clausal-complement construction. Dependency parsing is performed on official peer review texts submitted to major ML conferences to obtain the required syntactic information for analysis. LLM generated reviews for the same set of papers was obtained by prompting three selected models which act as three distinct LLM reviewers. A contrast is also drawn between the before and after periods with the release of the first consumer-ready LLM as the dividing line.

Our major finding is that the majority of verb lemmas having a naturally high usage of the O construction (O%$>$45%) in official reviews are observed to show an even higher O usage in LLM texts. The maximum difference is around 40%, which would already mean 90% O usage in LLM texts. Skewness tests confirm that the O% distributions for LLM reviews concentrate toward the extreme upper end, leaving only a sparse tail of lower portions of O usage. The unnaturally high O% usage is greatly suppressing the frequencies of other construction types. This is considered a serious case of syntactic homogenization.

The lexical diversity in LLM texts is found to be systematically lower than in official reviews, as another indicator of homogenization. Finally, we are observing the effects with frequent words, such as "emphasize", appearing at the top of the homogenized verb list. These results are more striking than identifying disproportionately more frequent usage of naturally rare words such as "commendable".

In this paper, we have experimented with LLMs of different architectures and sizes, but we claim no further inference on their differences in patterns of syntactic homogenization. One working hypothesis for their common skewness is that LLMs, despite their sizes and architectures, have the inherent tendency to predict the most frequent usage/structure of verbs. In our future work, we plan to expand our analyses to other construction types to verify this hypothesis.

9. Ethical Considerations

The material used in this paper is accessible via the OpenReview API and we confirm that we have complied with their term of use. Papers submitted to EMNLP 2023 are licensed under the CC-BY-4.0 license. Authors of papers for ICLR and NeurIPS did not transfer copyright and the copyright of papers for CoRL is unclear. However, we use the paper content only as part of the prompts and we have carefully ensured that there is no potential of exposing the paper content as training data to future generations of the tested LLMs through this study. `gemma3` and `qwen3` are open weight LLMs that we ran locally and `gpt-4o` is accessed through a paid business API where terms and conditions of usage include explicit guaranty of data privacy. Paper content fed to the models will not be retained for the training of future models. Comments on OpenReview are also licensed under CC-BY-4.0.

Our work uses the official review as a reference comparison against AI-generated reviews. We make no assumption or implication on the authorship of any official reviews. In fact, our study is only valid at corpus level to account for individual variations of human written text.

10. Limitations

The accuracy of the analysis depends on the performance of the dependency parser. Its accuracy was benchmarked at 92.0% on human written text but it is untested on MGT. As MGT becomes more and more fluent and grammatical, we expect the parsing accuracy on them in formal genre would not deviate significantly from human written text.

Due to space constraints, no confidence levels or other statistics were reported. Though in fact, we believe that the effects we observed well exceed the range of potential errors.

11. Bibliographical References

Dhruv Agarwal, Mor Naaman, and Aditya Vashistha. 2025. Ai suggestions homogenize writing toward western styles and diminish cultural nuances. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA. Association for Computing Machinery.

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*, page 413–425, New York, NY, USA. Association for Computing Machinery.

Rishi Bommasani, Kathleen A. Creel, Ananya Kumar, Dan Jurafsky, and Percy S Liang. 2022. Picking on the same person: Does algorithmic monoculture lead to outcome homogenization? In *Advances in Neural Information Processing Systems*, volume 35, pages 3663–3678. Curran Associates, Inc.

Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. 2023. Assessing cross-cultural alignment between Chat-

GPT and human societies: An empirical study. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 53–67, Dubrovnik, Croatia. Association for Computational Linguistics.

William Croft, Dawn Nordquist, Katherine Looney, and Michael Regan. 2017. Linguistic typology meets Universal Dependencies. In *TLT*, pages 63–75.

Marie-Catherine De Marneffe, Christopher D Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal dependencies. *Computational linguistics*, 47(2):255–308.

Gemma-Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. 2025. Gemma 3 technical report.

Mingmeng Geng, Caixi Chen, Yanru Wu, Yao Wan, Pan Zhou, and Dongping Chen. 2025. The impact of large language models in academia: from writing to speaking. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19303–19319, Vienna, Austria. Association for Computational Linguistics.

Talmy Givón. 1995. *Functionalism and Grammar*. John Benjamins Publishing.

Yanzhu Guo, Guokan Shang, and Chloé Clavel. 2025. Benchmarking linguistic diversity of large language models. *Transactions of the Association for Computational Linguistics*, 13:1507–1526.

Yanzhu Guo, Guokan Shang, Michalis Vazirgiannis, and Chloé Clavel. 2024. The curious decline of linguistic diversity: Training language models on synthetic text. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3589–3604, Mexico City, Mexico. Association for Computational Linguistics.

Martin Haspelmath. 2014. Arguments and adjuncts as language-particular syntactic categories and as comparative concepts. *Linguistic Discovery*, 12(2):3–11.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength natural language processing in python.

Paul J. Hopper and Sandra A. Thompson. 1984. The discourse basis for lexical categories in universal grammar. *Language*, 60:703.

Kuan-Hao Huang, Varun Iyer, I-Hung Hsu, Anoop Kumar, Kai-Wei Chang, and Aram Galstyan. 2023. Paraamr: A large-scale syntactically diverse paraphrase dataset by amr back-translation. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

Webdell Johnson. 1944. Studies in language behavior: A program of research. *Psychological Monographs*, 56(2):1–15.

Jon Kleinberg and Manish Raghavan. 2021. Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences*, 118(22):e2018340118.

Dmitry Kobak, Rita González-Márquez, Emőke Ágnes Horvát, and Jan Lause. 2025. Delving into llm-assisted writing in biomedical publications through excess vocabulary. *Science Advances*, 11(27):eadt3813.

S. Kokoska and D. Zwillinger. 2000. *CRC Standard Probability and Statistics Tables and Formulae, Student Edition*. Mathematics/Probability/Statistics. Taylor & Francis.

William Labov. 1969. *A Study of Non-Standard English*. Champaign, IL: National Council of Teachers of English.

William Labov. 2008. Pursuing the cascade model. In *Social dialectology*, pages 9–22. John Benjamins Publishing Company.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and William B Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119.

Weixin Liang, Zachary Izzo, Yaohui Zhang, Haley Lepp, Hancheng Cao, Xuandong Zhao, Lingjiao Chen, Haotian Ye, Sheng Liu, Zhi Huang, Daniel A. McFarland, and James Y. Zou. 2024. Monitoring ai-modified content at scale: a case study on the impact of chatgpt on ai conference peer reviews. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

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*, 6:100207.

OpenAI, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian

Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. 2024. Gpt-4o system card.

Vishakh Padmakumar and He He. 2024. Does writing with language models reduce content diversity? In *International Conference on Representation Learning*, volume 2024, pages 642–669.

Gabriela Ramírez de la Rosa, Thamar Solorio, Manuel Montes, Yang Liu, Lisa Bedore, Elizabeth Peña, and Aquiles Iglesias. 2013. Exploring word class n-grams to measure language development in children. In *Proceedings of the 2013 Workshop on Biomedical Natural Language Processing*, pages 89–97, Sofia, Bulgaria. Association for Computational Linguistics.

Pedro Reviriego, Javier Conde, Elena Merino-Gómez, Gonzalo Martínez, and José Alberto Hernández. 2024. Playing with words: Comparing the vocabulary and lexical diversity of chatgpt and humans. *Machine Learning with Applications*, 18:100602.

Abdelrahman Sadallah, Tim Baumgärtner, Iryna Gurevych, and Ted Briscoe. 2025. The good, the bad and the constructive: Automatically measuring peer review's utility for authors. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 28979–29009, Suzhou, China. 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, pages 6416–6431, Miami, Florida, USA. Association for Computational Linguistics.

Justin Sulik, Nakwon Rim, Elizabeth Pontikes, James Evans, and Gary Lupyan. 2025. Differences in psychologists' cognitive traits are associated with scientific divides. *Nature Human Behaviour*, pages 1–15.

Misha Teplitskiy, Daniel Acuna, Aïda Elamrani-Raoult, Konrad Körding, and James Evans. 2018. The sociology of scientific validity: How professional networks shape judgement in peer review. *Research Policy*, 47(9):1825–1841.

Peter Trudgill. 2011. *Sociolinguistic typology: Social determinants of linguistic complexity*. Oxford University Press.

Keith Tyser, Ben Segev, Gaston Longhitano, Xin-Yu Zhang, Zachary Meeks, Jason Lee, Uday Garg, Nicholas Belsten, Avi Shporer, Madeleine Udell, Dov Te'eni, and Iddo Drori. 2024. Ai-driven review systems: Evaluating llms in scalable and bias-aware academic reviews.

Hiromu Yakura, Ezequiel Lopez-Lopez, Levin Brinkmann, Ignacio Serna, Prateek Gupta, Ivan Soraperra, and Iyad Rahwan. 2025. Empirical evidence of large language model's influence on human spoken communication.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 technical report.

Denis Yarats and Mike Lewis. 2018. Hierarchical text generation and planning for strategic dialogue. In *International Conference on Machine Learning*, pages 5591–5599. PMLR.

Sungduk Yu, Man Luo, Avinash Madusu, Vasudev Lal, and Phillip Howard. 2025. Is your paper being reviewed by an llm? benchmarking ai text detection in peer review.