Discourse Realization of Generics in Human and LLM-generated Texts

Søren Kirkegaard Fomsgaard[*]
University of Caen
GREYC
soren.fomsgaard@unicaen.fr

Martial Pastor[*]
Radboud University
CLST
martial.pastor@ru.nl

Gaël Dias
University of Caen
GREYC
gael.dias@unicaen.fr

Nelleke Oostdijk
Radboud University
CLST
nelleke.oostdijk@ru.nl

Abstract

Large Language Models (LLMs) often produce texts that appear coherent and credible, even when their factual reliability is uncertain. This paper investigates whether such perceived credibility correlates with the pervasive use of *generics*—generalizations without explicit quantification. We introduce a text-level genericity score derived from clause-level annotations and apply it to argumentative essays produced by humans and LLMs. To analyze how generics are realized in discourse, we employ Rhetorical Structure Theory to examine coherence relations across varying levels of genericity. Results show that according to our genericity metric, human texts are less generic than LLM-produced texts. As regards discourse, higher genericity correlates with less structured, paratactic structures, while for some models coherence is maintained through ELABORATION relations. Our findings suggest that some LLMs maintain well-structured coherence even in highly generic texts, which might enable them to “camouflage” argumentative texts as informative, enhancing their perceived credibility and persuasiveness.

1 Introduction

Large Language Models (LLMs) have been adopted as a way to access information more quickly than any previous technology or medium (Hu et al., 2023). However, since LLMs do not always provide factually accurate information - while often appearing reliable (Edwards, 2023), researchers have begun to examine their capacity to produce “credible” information and the tendency of users to readily accept the texts they generate (Anderl et al., 2024). This paper aims to take this investigation further by examining whether LLMs create the illusion of being truthful and trustworthy (thereby encouraging their continued use) through the extensive use of **generics** when they generate text.

Generics be defined as statements expressing generalizations about kinds and their properties, or events without explicit quantification (Krifka et al., 1995) (e.g., “mosquitoes carry malaria” or “birds lay eggs”; see also Section 3.1). Psycholinguistic research has shown that generics are easier to process from a reader/listener point of view. It is suggested that, whereas specific quantified statements demand greater processing effort, generics are processed more quickly and effortlessly (Kahneman and Frederick, 2002; Lazaridou-Chatzigoga, 2019). Moreover, because generics refer to the constitutive properties of a concept rather than to its particular instances, they imply what a “proper” instance of the concept should be. As a consequence, generics have been associated with perpetuating and reinforcing social stereotypes in the form of (pejorative) social generics (Smith and Tolbert, 2025; Rhodes et al., 2025; Mannheim, 2021), or normative generics (Hesni, 2021). They may also contribute to the spread of misleading or even harmful information (Mannheim, 2021), aligning with prior observations that LLMs are susceptible to providing unreliable output (Hicks et al., 2024) when they hallucinate (Augenstein et al., 2023). In this sense, generics appear to provide a useful analogy for LLM-generated text: compelling and easy to process, yet prone to producing incorrect (hallucinated) or misleading content.

The first half of the paper examines whether LLM texts are more generic than human texts in argumentative writing. We develop a text-level genericity score using factor analysis and apply it to student essays from OUTFOX (Koike et al., 2024), augmented with four recent open models. We validate generalizability using the Aeon (Acharya, 2024) dataset, thus testing different writer demographics (students vs. adults) and genres (argu-

---

[*] Equal contribution. Corresponding authors.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 32938–32960
July 2-7, 2026

mentative vs. expository). Robustness is assessed through ablation studies, weight inversions, and random permutations. We focus on argumentative writing due to its implications for misinformation.

In the second half of the paper, we focus on studying how varying text-level genericity correlates with discourse coherence using Rhetorical Structure Theory (RST; Mann and Thompson, 1988). This analysis offers important insights into how generics are realized in texts and what specific discursive strategies may enhance text credibility. By analyzing discourse structure and different types of coherence relations, such as presentational and subject-matter relations, we show that LLMs tend to struggle to produce structured and well-argued texts as genericity increases. Notable exceptions are ChatGPT 3.5/4o/5.2 (OpenAI, 2025) and Apertus 8/70b, which are capable of generating quite structured and seemingly informative texts even when the genericity is high. For instance, in the two examples below, both texts consist of generic statements. We observe that ChatGPT connects these statements through ELABORATION relations, which are considered more informative, whereas Flan-T5-XXL employs the multinuclear JOINT relation, which is comparatively less structured.[1]

> Community service requires conventional skills like communication, leadership and organizational abilities. ← [elaboration] These abilities could offer the students opportunities ← [elaboration] which would allow them to put their academic theoretical knowledge into practical realms. — [*chatgpt train_1196*]

> Community experience is energizing ← [joint]→ and improves the lives of the children. ← [joint]→ Ultimately it will produce an increased concern for the environment and the world around them — [*flan_t5_xxl train_12709*]

These linguistic observations about generics are subsequently examined through the lens of persuasion strategies, with particular attention to some model's ability to "camouflage" argumentative texts as informative. This frame can reduce the readers' resistance, making them more likely to accept or be persuaded by the information presented. This suggests a notably effective interplay between generic statements and persuasive discourse strategies, a dynamic some models seem particularly adept at exploiting.

This paper makes the following contributions:

1. We introduce a text-level genericity scoring metric and evaluate it across argumentative texts using robustness tests and human assessments.
2. We show that human-written argumentative texts are significantly less generic than LLM-generated ones, across multiple model generations from older to more recent LLMs, and that genericity scores for human-written essays generalize from argumentative texts produced by younger writers to expository texts written by adults.
3. We analyze how genericity is realized in discourse by applying Rhetorical Structure Theory (RST) to examine coherence relations across different levels of genericity.
4. We find that higher genericity generally correlates with less structured discourse, but certain LLMs (e.g., ChatGPT) maintain well-structured coherence even in highly generic texts, which potentially enhances their perceived credibility.

2 Related Work

2.1 Generics at Clause and Text Level

Generics in texts are typically studied at the clause level. Recent studies in corpus linguistics emphasize that features such as genericity, eventivity, and boundedness constitute the basic building blocks from which broader discourse frames emerge (Friedrich, 2017; Grisot, 2018). In these approaches, genericity is analyzed at the clause level, where the main referent determines whether a clause is generic (class level) or specific (individual level) (Smith, 2003). This article builds on this perspective as it aims to investigate how clauses labeled as generic or specific combine with one another to shape discourse at text level.

2.2 Generics in NLP

The work on generics in the NLP community to date is extremely sparse. There are two main trends. The first is to build systems for the automatic identification and extraction of generic noun phrases (Reiter and Frank, 2010), or generic clauses (Friedrich

---

[1] The use of the term 'structured' refers to the RST framework introduced in Section 6. We consider a text to be less structured when it contains fewer hierarchical nucleus–satellite relations and relies more on paratactic connections, such as multinuclear JOINT relations that link loosely connected segments.

and Pinkal, 2015; Hemmatian, 2021). The other trend is to build datasets of synthetic generics (Bhakthavatsalam et al., 2020; Allaway et al., 2023; Sap et al., 2019).

The work on generics in the context of LLMs mainly centers around how language models process or are sensitive to generics (Cilleruelo Calderón et al., 2024; Allaway et al., 2023; Collacciani and Rambelli, 2023). To the best of our knowledge, there is currently no work that examines how LLMs (rather than humans) *realize* generics (as opposed to how they process or "reason" about them).[2] Our work seeks to fill this gap by identifying generics in LLM-written texts, focusing on analyzing their realization patterns.

3 Generics at the Text Level

3.1 Defining Genericity

Generics are statements expressing generalizations about kinds and their properties, or events without explicit quantification (Krifka et al., 1995). While most agree on the distinction between kind-referring and characterizing generics, alternative definitions exist (cf. Appendix A.1).

This paper focuses on automatically identifying generics in human and machine-generated text, building on Hemmatian (2021) and Friedrich (2017). We adopt their clause-level analysis and extend it to the text level (cf. Section 3.2).

Following Friedrich (2017) and Hemmatian (2021), a generic statement consists of two elements: the referent (typically bare noun phrases) and the verb constellation (habitual, eventive, or stative verbs). We identify generics at the clause level, categorizing them by verb type and whether the main referent is generic or specific. Our definition follows:

> **A generic statement** consists of a referent, comprised of a definite or indefinite NP, and a verb constellation, with a habitual, eventive or stative (optionally coerced) main verb.

A text can be said to be generic if it contains generic statements. A text with many generic statements has a higher level of genericity than one with fewer or no generic statements.

3.2 Scoring Text Genericity

3.2.1 Labeling Scheme for Generic Statements

In order to arrive at a measure of genericity at the text level, we first need to identify generic statements at the clause level. To do this, we adapt the set of 17 clause labels proposed by Hemmatian (2021), motivated by the definition in Section 3.1. The label set is based on Friedrich (2017) and Smith (2003). We also adopt the model developed by Hemmatian (2021), modifying it by introducing clause-level label weighting to derive a single text-level genericity score.

3.2.2 Weighting Clause Labels

We compute a text-level genericity score by weighting clause labels into three tiers. Pure generics (clauses with generic referents) receive the highest weight, followed by impure generics (generic statements with coerced or unbounded verbs but generic referents), and finally other clause types (non-declarative moods, unbounded events, or specific referents) (see Table 3 in the Appendix). This weighting reflects three principles: generic referents outweigh specific ones, indicative mood outweighs interrogative or imperative, and uncoerced verb constructions outweigh coerced ones. A weighted average of these tiers produces the final text-level genericity score that systematically prioritizes pure generics (cf. Section 4.3).

4 Genericity Scoring Pipeline

4.1 Data

In our study, we primarily use the original OUTFOX dataset (Koike et al., 2024) (designed for machine-generated text detection) and our own augmentation of it with newer models plus the original human sample. The OUTFOX dataset combines U.S. grade 6–12 student essays from the PERSUADE 2.0 corpus (Crossley et al., 2024) with LLM-generated texts (see Appendix A.3 for details). Two main limitations exist: (1) human texts come from students, not adult native speakers, and (2) the original models are outdated as of 2025. To address the issue of human representativeness, we evaluate our metric on adult human texts from the additional Aeon-essays dataset (Acharya, 2024). To improve model coverage, we generated 123k new synthetic texts using the original OUTFOX prompts with recent models: Gemma 3 (4b, 12b, 27b), Ministral 3 (3b, 8b, 14b), Apertus (8b, 70b), and ChatGPT (4o, 5.2). These were selected

---

[2] Although Peters and Chin-Yee (2025) come close as they examine the tendency of LLMs to overgeneralize conclusions when summarizing scientific text. Their focus, however, is on explicit generalizations rather than generics.

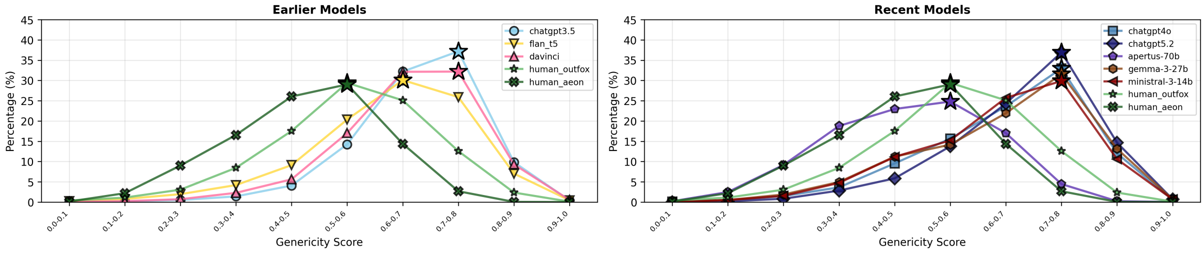

Figure 1: Distribution of genericity scores across models. The left panel shows earlier models (ChatGPT-3.5, Flan-T5, Text-Davinci) compared to human baselines. The right panel shows more recent models (ChatGPT-4o, Apertus-70b, Gemma-3-27b, Ministral-3-14b) compared to human baselines. Stars indicate the peak of each distribution.

for diversity in openness, commercial focus, and size, while prioritizing recent, high-performing, instruction-tuned versions (see Appendix A.3.1 for details).

4.2 Genericity Classifier at Clause Level

In all of our experiments, we use a classification pipeline based on Hemmatian (2021)'s work[3] to label all clauses in the OUTFOX dataset. The model is a multiclass classifier trained to predict 17 clause-level labels (see Table 3 in Appendix A.2). This label set, designed by Hemmatian (2021), draws on *genericity, eventivity,* and *boundedness* to examine their role in argumentative and non-argumentative texts. For our purposes, we use the same scheme, emphasizing generic labels in factor analysis for text genericity scoring. Clauses are encoded with RoBERTa and classified with a bi-GRU model (Cho et al., 2014; Chung et al., 2014), producing softmax-normalized predictions (Saphra and Lopez, 2018; Friedrich, 2017). Trained on 27,240 clauses from a News+Reddit corpus (Hemmatian, 2021) and pre-trained on SitEnt (Friedrich, 2017), the model performs robustly on a 10% held-out test set: *genericity* (precision 0.860, recall 0.852, F1 0.841), *eventivity* (precision 0.894, recall 0.879, F1 0.873), and *boundedness/habituality* (precision 0.850, recall 0.804, F1 0.860).

4.3 Factor Analysis for Genericity Scoring

Once all clauses in a dataset have been automatically labeled, we compute a text-level genericity score for each individual text.

We compute the genericity score S using a factor analysis which quantifies a text's genericity by computing a weighted average of clause label proportions. For a text with N clauses and a set of k labels $L = \{l_1, \ldots, l_k\}$, each label l_i has a weight $w_i \in [0.0, 0.5, 1.0]$ depending on its group: 1.0 for highly generic clause types, 0.5 for moderately generic clause types, and 0.0 for irrelevant ones. We compute proportions as $p_i = n_i/N$, where n_i is the count of label l_i. The score is $S = \sum_{i=1}^{k} w_i p_i$, ranging from 0.0 (specific) to 1.0 (generic), serving as a simplified factor analysis of label contributions to genericity.

5 Genericity Scoring Results

5.1 Human and Machine Genericity Score Distributions

We automatically label generic statements across 61,600 texts in the augmented OUTFOX dataset and computed genericity scores (see Figure 1). Human-written texts consistently peaked around 0.5–0.6, while machine-generated texts peaked much higher at 0.7–0.8, demonstrating that humans produce substantially less generic text than machines. To confirm this was not specific to adolescent writers, we applied the same metric to Aeon-essays (adult expert writers) and found human texts again peaked at 0.5–0.6. This consistency across different author populations and essay types validates that humans genuinely write with less genericity than machines.[4]

5.2 Robustness Analysis and Human Validation

5.2.1 Metric Sensitivity to Weight Perturbation

To assess the sensitivity of our genericity metric to different perturbations, we perform three types of experiments when computing the text-level score:

---

[3] `https://huggingface.co/spaces/BabakScrapes/Anecedotal_Discourse_Classifier_Demo`

[4] We report the results for the largest open models here for the sake of simplicity, although the trend is the same for their smaller viarants (cf. Table 4).

ablation of each clause type, inverse weighting, and random weight perturbations. These tests evaluate whether the metric unduly favors specific clause types predicted by the clause-level classifier (Section 4.2), how much such changes affect the overall score, and the metric's stability and discriminative power.

Metric Sensitivity to Ablation. In the ablation experiments, we remove each of the 17 clause types one at a time and recompute scores using the remaining weights (as defined in Section 3.2.2). Results across OUTFOX and Aeon-essays (Tables 5 and 6) show that only six clause types exert substantial influence on the text-level genericity score. In particular, GENERIC SENTENCE (STATIC/DYNAMIC), BOUNDED EVENT (GENERIC), and COERCED STATE (GENERIC) contribute most strongly: their ablation produces the largest drops in mean genericity score relative to the baseline.

This concentration of influence arises from the combination of high frequency of these clause types (several rank among the top five most common tags in both datasets; see Table 4) and our deliberate upweighting of them in the scheme. While the metric is therefore highly sensitive to a small subset of clause types, we regard this as a feature rather than a flaw. Theoretically motivated weights (Section 3.2.2) prioritize linguistically relevant categories such as GENERIC SENTENCE over equally frequent but less informative ones (e.g., BASIC STATE or BOUNDED EVENT (SPECIFIC)). Thus, the observed sensitivity reflects the metric's ability to capture meaningful distributional differences in genericity-relevant clause types.

Weight Inversion. To examine the impact of clause-type frequencies and our weighting scheme, we perform two inversion experiments.

First, we apply reciprocal weights scaled to $w_i \in [0.1, 1.0]$ (Table 7a). This drives mean genericity scores toward 0.1 across both datasets, as frequent (originally up-weighted) clause types are now heavily down-weighted—consistent with the ablation results.

Second, we invert the categorical weights (Section 3.2.2, Table 7b): pure generics drop from 1.0 to 0.0, while "Other" types rise from 0.0 to 1.0. This causes a sharp score drop in OUTFOX, confirming the dominance of theoretically relevant clause types in our scheme. In Aeon-essays the score rises slightly (remaining near 0.5), due to its more balanced distribution with higher frequencies of non-generic types such as BASIC STATE and BOUNDED EVENT (SPECIFIC) (cf. Table 4).

Dataset	Random Weights	Baseline	$\lvert\Delta\rvert$	N Trials
Aeon	0.493 ± 0.099	0.470 ± 0.133	0.022	50
Outfox	0.504 ± 0.109	0.631 ± 0.137	0.127	50

Table 1: Random perturbation results. Random weight scores are means of means across 50 trials.

These results show that the metric's sensitivity to specific clause types is deliberate, arising from both frequency patterns and our linguistically motivated weights.

Random Perturbation.Finally, we test random weights (0.0–1.0) over 50 iterations per dataset. The mean genericity score converges toward 0.5, showing that our weighting scheme is meaningful: it outperforms random weights on OUTFOX (where characteristic clause types are frequent) but performs similarly on Aeon-essays (where clause types are less skewed). This demonstrates that the metric is sensitive to perturbations and that our theoretically motivated weights are effective.

5.2.2 Human Validation

To further validate the robustness of the genericity metric, we conduct a human evaluation using six annotators who rated texts from all models/authors, with texts aligned by prompt to evaluate variation across models.

Annotators and ChatGPT 5.2 were asked to assign genericity scores at the text level using the same 0-1 scale used by the automated metric. We instructed our human participants using a broad notion of genericity at the text level, according to which a text is generic if it exhibits characteristics of generic statements (generic referents, unbounded events etc.).[5] While the instructions align with the definition of genericity at the clause-level discussed in Section 3.1, they were worded without using the exact formulations found in the literature to avoid biasing the raters towards them. The aim was not to replicate the clause-level assessment of generic statements, but rather to test the alignment between the automated metric, which aggregates a text-level score based on clauses-level characteristics, and speaker intuitions of genericity at the

---

[5] Genericity scores were assigned by both ChatGPT-5 and human raters using the following prompt: "How specific is the text, as opposed to being generic? 0 = Very specific (detailed, concrete, clearly tied to a particular situation and/or specific individual); 10 = Very generic (abstract, general, could apply to many situations or people)."

text level, since our metric is designed to model text-level genericity.

Kendall's Tau (Kendall, 1938) between human ratings and the automated genericity metric was 0.41, indicating moderate agreement in overall ranking. However, when examining agreement on identifying the most generic texts specifically—which is most relevant for practical applications—Top-20 overlap reached 80%, demonstrating that human annotators and the automated method consistently identified the same texts as highly generic. The agreement between human ratings and ChatGPT-4o scores (Kendall's Tau: 0.46, Top-20 overlap: 75%) further confirms the metric's robustness in capturing genericity.

6 Discourse Realization of Generics

6.1 Rhetorical Structure Theory

To align the genericity labeling scheme with discourse structure, we adopt Rhetorical Structure Theory (Mann and Thompson, 1988), which uses clauses as the basis for Elementary Discourse Unit (EDU) segmentation. RST analyzes text coherence by linking EDUs through coherence relations such as ELABORATION, CONTRAST, CAUSAL, or TEMPORAL, forming a hierarchical tree such as in Figure 5. RST distinguishes "Nucleus" EDUs, carrying essential content, from "Satellite" EDUs, which add supplementary information. Some relations (e.g., JOINT, SAME-UNIT) are multinuclear. We use the set of 18 RST labels proposed in Braud et al. (2017).

6.2 Three Specific RST Relations

This paper analyzes how generics are realized in argumentative discourse. We draw on the distinction between subject-matter and presentational relations (Table 2), which has proven relevant when comparing argumentative and expository text characteristics (Li and Xiao, 2021; Azar, 1999). Mann and Thompson (1988) introduced this to separate relations that increase reader inclination (presentational) from those that aid comprehension (subject-matter).

Presentational relations, while less frequent overall, characterize argumentative texts, which are generally more persuasive (Li and Xiao, 2021). Subject-matter relations dominate informative texts (e.g., textbooks, Wikipedia). Multinuclear relations (Table 2) have ambiguous status. Though they can carry rhetorical effects by comparing, contrasting, or sequencing statements, the predominant JOINT relation is often a default connector when no specific link is perceived. However, repeated JOINT sequences can pragmatically describe parataxis as a stylistic effect (Pastor et al., 2024).

6.3 RST Patterns in Generic Realization

6.3.1 Discourse Parser

RST trees for OUTFOX texts were generated using the DMRST parser (Liu et al., 2021). This multilingual top-down parser jointly performs EDU segmentation and RST analysis, achieving state-of-the-art accuracy on span splitting (88.2%) and nuclearity determination (76.2%) and relation prediction (64.7%). To improve relation prediction, we adopted the fine-tuning approach of Pastor et al. (2025), targeting specific RST sub-structures using gold-standard subtrees with paratactic structures from Zaczynska and Stede (2024) and Potter (2008).[6] Evaluated on the domain-aligned GUM essays test set (Zeldes et al., 2025), the model achieved F1 scores of 0.71 for JOINT, 0.75 for ELABORATION, and 0.70 overall. Discourse relation distributions across models/authors appear in Table 12 in the Appendix.

6.3.2 Text Genericity and RST Relation Types

Following the classification of relation types shown in Table 2, we plot the histograms of presentational, subject-matter, and multinuclear relations across ten genericity bins in Figure 2.

First, we note that the distribution of relation types aligns with what can be observed in other corpora, such as RST-DT (Carlson et al., 2001), where subject-matter and multinuclear relations account for the largest proportions. This is largely explained by the frequent presence of JOINT (multinuclear) and ELABORATION (subject-matter) relations across all text types.

Although many observations could be drawn from these graphs, we highlight two main points that form the basis for the subsequent analysis in the next two sections: (1) as genericity increases, all models/authors tend to rely more on multinuclear relations and (2) the ChatGPT models are the only models that consistently use more subject-matter relations across all levels of genericity.

Category	Relations
Presentational	Background, Enablement, Summary, Explanation
Subject Matter	Attribution, Cause, Condition, Elaboration, Evaluation, Manner-Means, Topic-Comment
Multinuclear	Comparison, Contrast, Joint, Same-Unit

Table 2: List of RST Relations.

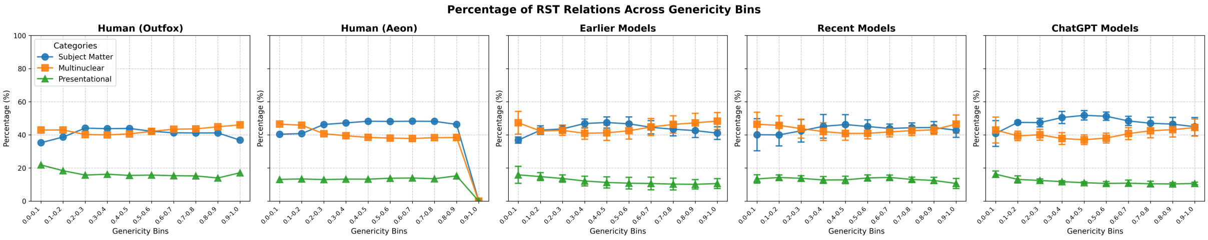

Figure 2: Distribution of RST relation categories across genericity bins (0.0–1.0). Panels show (left to right): Human (Outfox), Human (Aeon), Earlier Models (averaged: Flan-T5, Davinci), Recent Models (averaged: Apertus-70b, Gemma-3-27b, Ministral-3-14b), and ChatGPT Models (averaged: ChatGPT-3.5, ChatGPT-4o, ChatGPT-5.2). Categories: Subject Matter (blue circles), Multinuclear (orange squares), Presentational (green triangles).

6.3.3 Text Genericity and Parataxis

We revisit observation (1) from Section 6.3.2: increasing multinuclear relations as genericity rises. Since JOINT is the most frequent multinuclear relation (Table 12), we examine RST subtrees with successive JOINT relations (JOINT–JOINT–JOINT). Sequential JOINT patterns capture paratactic communication in persuasive discourse (Pastor et al., 2024)—a strategy that juxtaposes loosely connected statements to create semi-argumentative logical flow that is easily processed.

Figure 3 shows JOINT–JOINT–JOINT subtrees across ten genericity bins. All models and authors increasingly employ this pattern as genericity rises, with humans ranking second behind Text-DaVinci-003 and near Flan-T5-XXL. ChatGPT models show less pronounced increases, and recent models (except Apertus) similarly attenuate paratactic use at high genericity. Aeon's expository style predictably limits parataxis overall, though it still increases with genericity. Combined with Figure 2—where subject-matter relations decrease with genericity—this suggests paratactic communication becomes more frequent when arguing using generics.

Moreover, generics appear harder to connect (as one could in informative texts with subject-matter relations), and may instead invite less structured forms of argumentation, such as the paratactic construction (JOINT–JOINT–JOINT) illustrated in Figure 5 (top). In this example, the human author presents an argument solely through the paratactic linking of clauses classified as GENERIC SENTENCE (clauses 1–5) and BOUNDED EVENT (GENERIC), with the main generic referents being "we" (humans) and "it" (the smartphone). Although the argument lacks explicit premises or a conclusion, the sequence of statements nonetheless forms a narrative that produces a consequential flow.

6.3.4 Text Genericity and Elaborations

We investigate observation (2) from Section 6.3.2: ChatGPT models consistently use more subject-matter relations across all genericity levels. Though subject-matter relations decrease with genericity, they remain most frequent even at high levels (bins above 0.7), indicating ChatGPT realizes generics through subject-matter–oriented structures. Table 13 shows earlier ChatGPT models (ChatGPT-4o, ChatGPT-3.5) exhibit ELABORATION as the most frequent relation in highly generic texts. Recent models like ChatGPT-5.2 rely less systematically on elaboration. Ministral and Gemma also favor elaborative relations, though less than earlier ChatGPT models.

This appears contradictory given Mann and Thompson (1988)'s definition: elaborations provide additional detail making content more specific (e.g., set::member, generalization::specific). However, Figure 5 (middle) shows ChatGPT-3.5 produces ELABORATION sequences that detail generic referents while maintaining generic formulations. It uses shallow elaborations (often relative clauses avoiding temporal/spatial specification) and higher-

---

[6] Hyperparameters used: batch size of 12, learning rate of 5e-8 (reduced from 5e-7), and 2 epochs of training.

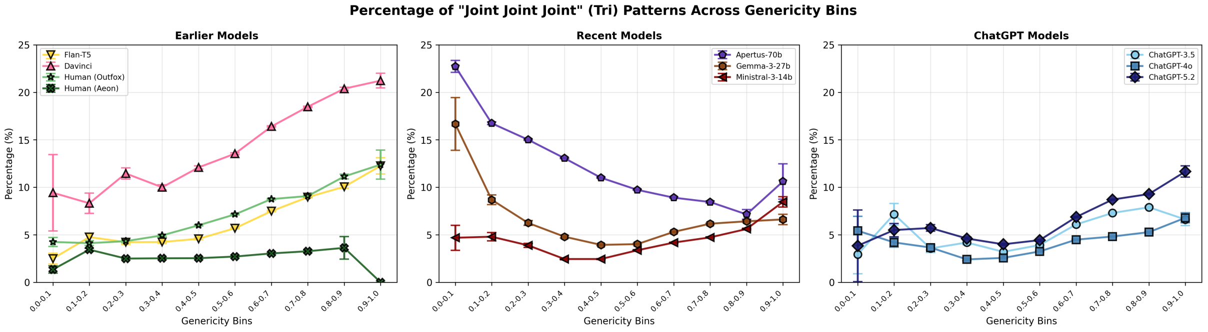

Figure 3: Histograms of JOINT- JOINT- JOINT across ten genericity bins (0.0–0.1 to 0.9–1.0) for Human, ChatGPT 3.5, Flan-T5-XXL, and Text-DaVinci-003.

level elaborations spanning larger text segments.

In contrast, Figure 5 (bottom) of the Appendix shows humans use ELABORATIONs to specify referents through locality (EDUs 1-2) or individual experiences. Recent models, particularly ChatGPT-3.5 and 4o, thus accomplish the paradox of elaborating generics with further generics.

7 Discussion

7.1 Effect of Prompt Variation

We consider the possibility that different prompts may influence the model's tendency to generate more generic text. Are the trends we observe an artifact of the prompts used in the original dataset, or is genericity a key characteristic of LLM-generated text? To investigate this, we examine the impact of the prompt on the generated essays. We examine how the wording of the problem statement correlates with the average genericity score on the validation split of OUTFOX. In doing this, we want to see if phrases like "explain" and "explain why" correlate with higher or lower genericity scores. Table 14 of the Appendix shows that the corresponding prompts for the LLM-written essays do not result in any significant changes to the genericity score of the texts. We also note that prompts that ask the model to draw on personal experience, as can be expressed through the ATTRIBUTION relation (attributing opinions or statements) tend to have lower genericity scores. Plotting the relationship between attribution relations and genericity scores across the human-written and LLM-generated texts, we observe the graph in Figure 6 (Appendix A.4).

7.2 Informative Text as a Persuasion Strategy

Lastly, we contextualize the findings from Section 6.3.4 by showing that earlier ChatGPT models—and Gemma and Ministral, which are similar—use subject-matter relations as a discursive strategy that appears to bypass what psycholinguists term the 'forewarning effect'.

While presentational and multinuclear relations have been shown to carry stronger persuasive power in certain argumentative contexts—such as discussions, comment sections, or short political speeches—they are less effective in longer argumentative texts, in the sense that they forewarn the reader about a particular persuasive intent (Kamalski, 2007). Kamalski's findings suggest that in such texts, readers are more easily persuaded by arguments that present themselves as objective and that are more informative about real-world facts. Hence, given that earlier ChatGPT models—and Gemma and Ministral to a lesser extent—use elaborations to make their texts more informative, it is easier to regard these texts as credible, as they avoid forewarning factors that typically trigger reader resistance in argumentative text.

8 Conclusion and Future Work

In this work, we set out to investigate the use of generics in discourse, both written by machines and humans, and its correlation with discourse structures. To this end, we introduce a text-level genericity scoring metric and evaluate it across argumentative texts using robustness tests and human assessments.

When averaging over generics at the clause level, we found that human-written texts tend to be less generic than machine-generated texts, with Flan-T5-XXL most closely resembling human text, and ChatGPT models (3.5, 4o, 5.2) along with recent models being farthest from humans in terms of genericity. These results hold across multiple generations of models, and the human-genericity patterns observed in argumentative texts generalize to expository texts written by adults.

Regarding discourse relations, we observed that higher genericity generally correlates with less structured discourse. Multinuclear relations tend to increase with genericity in both human-written and machine-generated texts, except for ChatGPT and recent models. ChatGPT 3.5, 4o models—and Gemma and Ministral, which are similar—excelled in using subject-matter relations such as ELABORATION, typically associated with characteristics of informative texts. This suggests that these models maintain coherent and informative discourse even in highly generic texts, a capability that may reduce reader resistance and enhance perceived credibility beyond what humans and other models achieve.

This work has been exploratory. To support future research, we provide access to our code[7] for the metric, discourse analysis, and the augmented dataset[8] with clause-level genericity and RST annotations. This would allow us to examine in future work trends that space constraints prevented us from exploring here. For instance, Apertus's outlier status—with more human-like genericity scores and distinctive RST patterns—merits deeper investigation, facilitated by its fully open-source pretraining data. Other significant trends include ChatGPT-4o's overuse of CAUSE relations compared to versions 3.5 and 5.2.

Limitations

We acknowledge the following limitations of our study. First, our investigation into the effect of the wording of the prompts on the machine-generated essays in OUTFOX suggests that the use of subjective and objective language does not greatly impact the genericity score. Further work on alternative prompt contexts could thoroughly test the extent to which genericity scores of the machine texts might be influenced by the wording of the prompt. We did not pursue the issue in further detail here given that prompt stability is an open and complex problem in the evaluation of synthetic text, which would require more space than is available within the scope of this work. Further testing is needed to ensure that the effect of alternative wording does not impact genericity at the text level significantly. Second, we do not evaluate the performance of the clause type classifier on our dataset and compare it to its performance in the original experiments of Hemmatian (2021). Future work should annotate new data to assess the accuracy of the classifier on other text types.

Third, the OUTFOX dataset, and the language models we use for tagging generic clause types, are monolingual. Future work should collect comparable multilingual datasets and retrain the system to examine the cross-lingual generalizability of our findings.

Ethics Statement

The use of generics by LLMs has implications primarily in terms of the potential social harms and its potential for exploitation to spread misinformation. It is our assessment that there are no direct harms associated with the research developed in this paper.

8.1 The Potential Harm of Generics

8.1.1 Manipulation via Generics at Scale

In argumentative contexts, the use of generics can be employed, and the Generic Over-Generalization (GOG) effect (Leslie et al., 2011) and inferential asymmetry associated with generics (Cimpian et al., 2010) can be exploited to manipulate the audience, as shown by Reuter et al. (2025) in noncooperative scenarios. In the hands of bad actors that seek to scale their political influence campaigns using LLMs, models like ChatGPT could be used to manipulate at scale by taking advantage of the models' propensity to maintain well-structured, yet highly generic, argumentative text patterns.

8.1.2 Social Stereotyping

To the extent that models are exposed to biased training data during pretraining, or are adversarially steered by user interaction, LLMs might also reproduce and expose users to pejorative social stereotypes through social generics, if one assumes an essentialist view of social generics such as, for example, Rhodes et al. (2025).[9]

8.1.3 Machine-generated Misinformation

Since misinformation can arise from the obscuring or withholding of relevant information (Fallis, 2014), and because the truth conditions of generics are notoriously tricky, we speculate that LLMs can be used by bad actors to effectively proliferate misinformation using generics.

---

[7] `https://github.com/SorenKF/general_machines`

[8] `https://doi.org/10.5281/zenodo.19594487`

[9] We are aware that the debate on the relationship between social stereotypes and generics is unresolved.

Previous research (Jiang et al., 2023) has noted that machine-generated text is hard to detect in general (Jakesch et al., 2023), across domains (Li et al., 2024), in adversarial contexts (Krishna et al., 2023), and in social media-like interactions (Radivojevic et al., 2024), with humans often performing worse than automatic systems (Liu et al., 2024) depending on the domain and background of the annotators (Ippolito et al., 2020; Dugan et al., 2022). LLM-generated misinformation text is also hard to detect (Chen and Shu, 2024).

Furthermore, some studies have argued that text written by machines can be persuasive, in the right context (Bashardoust et al., 2024; Goldstein et al., 2023). As a consequence of the difficulties in detectability and the increasing proficiency of machines in writing appealing text, it has been suggested that LLMs can be misused to produce misinformation text, for example, with the aim of aiding malicious political influence campaigns (Bontcheva et al., 2024; Crothers et al., 2023; Papageorgiou et al., 2024).

Finally, a minor yet notable risk of machine-written misinformation in informative contexts lies in its more passive harm: the widespread use of LLMs online may expose larger audiences to incorrect generics, thereby contributing to misinformation and a gradual "pollution of the informational environment" (Pan et al., 2023).

Acknowledgments

We acknowledge the use of generative language models in the preparation of this work in the following ways. ChatGPT 5 was used for grammar correction of the manuscript while Claude Sonnet 4 was used to scaffold the plotting code and debugging. The original Outfox dataset studied in this paper was used in compliance with its Apache 2.0 license. Aeon-essays was used in compliance with its MIT license. The model described in Section 4.2 is unlicensed. The open models used for data augmentation are available under the following licenses: Gemma Terms of Use (Gemma 3), Apache 2.0 (Apertus), mrl (Ministral 3). Our use of the Open AI API to generate texts with ChatGPT 4o and 5.2 complies with the Open AI Terms of Use.

The authors gratefully acknowledge the funding from the Horizon Europe research and innovation programme under the Marie Skłodowska-Curie Grant Agreement No. 101073351. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency (REA). Neither the European Union nor the granting authority can be held responsible for them.

References

Mann Acharya. 2024. Aeon essays dataset.

Emily Allaway, Jena D. Hwang, Chandra Bhagavatula, Kathleen McKeown, Doug Downey, and Yejin Choi. 2023. Penguins don't fly: Reasoning about generics through instantiations and exceptions. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*. Association for Computational Linguistics.

Christine Anderl, Stefanie H Klein, Büsra Sarigül, Frank M Schneider, Junyi Han, Paul L Fiedler, and Sonja Utz. 2024. Conversational presentation mode increases credibility judgements during information search with chatgpt. *Scientific Reports*, 14(1):17127.

Isabelle Augenstein, Timothy Baldwin, Meeyoung Cha, Tanmoy Chakraborty, Giovanni Luca Ciampaglia, David Corney, Renee DiResta, Emilio Ferrara, Scott Hale, Alon Halevy, Eduard Hovy, Heng Ji, Filippo Menczer, Ruben Miguez, Preslav Nakov, Dietram Scheufele, Shivam Sharma, and Giovanni Zagni. 2023. Factuality Challenges in the Era of Large Language Models. *arXiv preprint*. ArXiv:2310.05189.

M. Azar. 1999. Argumentative text as rhetorical structure: An application of rhetorical structure theory. *Argumentation*, 13(1):97–114.

Amirsiavosh Bashardoust, Stefan Feuerriegel, and Yash Raj Shrestha. 2024. Comparing the willingness to share for human-generated vs. AI-generated fake news. *arXiv preprint*. ArXiv:2402.07395.

Sumithra Bhakthavatsalam, Chloe Anastasiades, and Peter Clark. 2020. Genericskb: A knowledge base of generic statements. *CoRR*, abs/2005.00660.

Kalina Bontcheva, Symeon Papadopoulous, Filareti Tsalakanidou, Riccardo Gallotti, Lidia Dutkiewicz, Noémie Krack, Francesco Severio Nucci, Jochen Spangenberg, Ivan Srba, Patrick Aichroth, Luca Cuccovillo, and Luisa Verdoliva. 2024. Generative ai and disinformation: Recent advances, challenges, and opportunities. Technical report, vera.ai / AI4Media / AI4Trust / TITAN (Horizon Europe projects). White Paper.

Chloé Braud, Maximin Coavoux, and Anders Søgaard. 2017. Cross-lingual RST discourse parsing. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 292–304, Valencia, Spain. Association for Computational Linguistics.

G. Carlson. 1977. Reference to kinds in English.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurovsky. 2001. Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue*.

Federico Cella and Martina Rosola. 2025. Generics. In *Philosophy*. Oxford University Press.

Canyu Chen and Kai Shu. 2024. Can LLM-Generated Misinformation Be Detected? *arXiv preprint*. ArXiv:2309.13788.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Junyoung Chung, Çaglar Gülçehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *CoRR*, abs/1412.3555.

Gustavo Cilleruelo Calderón, Emily Allaway, Barry Haddow, and Alexandra Birch. 2024. Generics are puzzling. Can language models find the missing piece? *arXiv preprint*.

Andrei Cimpian, Amanda C. Brandone, and Susan A. Gelman. 2010. Generic Statements Require Little Evidence for Acceptance but Have Powerful Implications. *Cognitive Science*, 34(8):1452–1482.

Ariel Cohen. 2004. Generics and Mental Representations. *Linguistics and Philosophy*, 27(5):529–556.

Ariel Cohen. 2012. Generics as Modals. *Recherches linguistiques de Vincennes*, (41):63–82. ISBN: 9782842923501 Number: 41 Publisher: Presses universitaires de Vincennes.

Claudia Collacciani and Giulia Rambelli. 2023. Interpretation of generalization in masked language models: An investigation straddling quantifiers and generics. In *Proceedings of the 9th Italian Conference on Computational Linguistics (CLiC-it)*, pages 143–153, Venice, Italy. CEUR Workshop Proceedings.

S. A. Crossley, Y. Tian, P. Baffour, A. Franklin, M. Benner, and U. Boser. 2024. A large-scale corpus for assessing written argumentation: PERSUADE 2.0. 61:100865.

Evan Crothers, Nathalie Japkowicz, and Herna Viktor. 2023. Machine Generated Text: A Comprehensive Survey of Threat Models and Detection Methods. *arXiv preprint*. ArXiv:2210.07321.

Renaat Declerck. 1986. The manifold interpretations of generic sentences. *Lingua. International review of general linguistics. Revue internationale de linguistique générale*, 68(2-3):149–188. Publisher: Elsevier.

Liam Dugan, Daphne Ippolito, Arun Kirubarajan, Sherry Shi, and Chris Callison-Burch. 2022. Real or Fake Text?: Investigating Human Ability to Detect Boundaries Between Human-Written and Machine-Generated Text. *arXiv preprint*. ArXiv:2212.12672.

Benj Edwards. 2023. Why chatgpt and bing chat are so good at making things up. *Ars Technica*.

Don Fallis. 2014. The Varieties of Disinformation. volume 358, pages 135–161, Cham. Springer International Publishing.

Alex Franklin, Maggie, Meg Benner, Natalie Rambis, Perpetual Baffour, Ryan Holbrook, Scott Crossley, and ulrichboser. 2022. Feedback prize - predicting effective arguments. `https://kaggle.com/competitions/feedback-prize-effectiveness`. Kaggle.

Annemarie Friedrich and Manfred Pinkal. 2015. Discourse-sensitive Automatic Identification of Generic Expressions. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pages 1272–1281, Beijing, China. Association for Computational Linguistics.

Annemarie Silke Friedrich. 2017. *States, Events, and Generics: Computational Modeling of Situation Entity Types*. Phd thesis, Universität des Saarlandes, Saarbrücken, Germany.

Susan A Gelman and Francis Jeffry Pelletier. 2010. Generics as a window onto young children's concepts. *Kinds, things, and stuff: Mass terms and generics*, pages 100–120. Publisher: Oxford University Press New York, NY.

Bart Geurts. 1985. Generics. *Journal of Semantics*, 4(3):247–255. Publisher: Oxford University Press.

Josh A. Goldstein, Jason Chao, Shelby Grossman, Alex Stamos, and Michael Tomz. 2023. Can AI Write Persuasive Propaganda?

Cristina Grisot. 2018. *Cohesion, Coherence and Temporal Reference from an Experimental Corpus Pragmatics Perspective*, 1 edition. Yearbook of Corpus Linguistics and Pragmatics. Springer Cham. 45 b/w illustrations.

James A. Hampton. 2012. Generics as reflecting conceptual knowledge. *Recherches linguistiques de Vincennes*, (41):9–24. ISBN: 9782842923501 Number: 41 Publisher: Presses universitaires de Vincennes.

Babak Hemmatian. 2021. *Taking the High Road: A Big Data Investigation of Natural Discourse in the Emerging U.S. Consensus about Marijuana Legalization*. Ph.D. thesis, Brown University, Providence, RI.

Samia Hesni. 2021. Normative generics: Against semantic polysemy. *Thought: A Journal of Philosophy*, 10(3):218–225.

Gerhard Heyer. 1985. Generic descriptions, default reasoning, and typicality. Publisher: De Gruyter Mouton.

Michael Townsen Hicks, James Humphries, and Joe Slater. 2024. ChatGPT is bullshit. *Ethics and Information Technology*, 26(2):38.

Krystal Hu and 1 others. 2023. Chatgpt sets record for fastest-growing user base - analyst note. *Reuters*.

Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. 2020. Automatic Detection of Generated Text is Easiest when Humans are Fooled. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1808–1822, Online. Association for Computational Linguistics.

Maurice Jakesch, Jeffrey T. Hancock, and Mor Naaman. 2023. Human heuristics for AI-generated language are flawed. *Proceedings of the National Academy of Sciences*, 120(11).

Bohan Jiang, Zhen Tan, Ayushi Nirmal, and Huan Liu. 2023. Disinformation Detection: An Evolving Challenge in the Age of LLMs.

Daniel Kahneman and Shane Frederick. 2002. *Representativeness Revisited: Attribute Substitution in Intuitive Judgment*, pages 49–81. Cambridge University Press, Cambridge. First published in print in 2002; published online in 2012.

Judith Maria Helena Kamalski. 2007. *Coherence marking, comprehension and persuasion: on the processing and representation of discourse*. LOT, Utrecht, The Netherlands. OCLC: 181090952.

Maurice G. Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.

Ryuto Koike, Masahiro Kaneko, and Naoaki Okazaki. 2024. OUTFOX: LLM-Generated Essay Detection Through In-Context Learning with Adversarially Generated Examples. 38(19):21258–21266.

Manfred Krifka, Francis Jeffry Pelletier, Gregory Carlson, Alice ter Meulen, Gennaro Chierchia, and Godehard Link. 1995. Genericity: An Introduction. In Greg N. Carlson and Francis Jeffry Pelletier, editors, *The Generic Book*, pages 1–124. University of Chicago Press.

Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. 2023. Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense. *arXiv preprint arXiv:2303.13408*, 36:27469–27500.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the ACM SIGOPS 29th symposium on operating systems principles*.

Dimitra Lazaridou-Chatzigoga. 2019. Genericity. In Chris Cummins and Napoleon Katsos, editors, *The Oxford Handbook of Experimental Semantics and Pragmatics*, 1 edition, pages 156–177. Oxford University Press.

Sarah-Jane Leslie. 2007. Generics and the structure of the mind. *Philosophical perspectives*, 21:375–403. Publisher: JSTOR.

Sarah-Jane Leslie. 2008. Generics: Cognition and Acquisition. *The Philosophical Review*, 117(1):1–47.

Sarah-Jane Leslie, Sangeet Khemlani, and Sam Glucksberg. 2011. Do all ducks lay eggs? The generic overgeneralization effect. *Journal of Memory and Language*, 65(1):15–31.

Jinfen Li and Lu Xiao. 2021. Neural-based RST parsing and analysis in persuasive discourse. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 274–283, Online. Association for Computational Linguistics.

Yafu Li, Qintong Li, Leyang Cui, Wei Bi, Zhilin Wang, Longyue Wang, Linyi Yang, Shuming Shi, and Yue Zhang. 2024. MAGE: Machine-generated Text Detection in the Wild. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 36–53, Bangkok, Thailand. Association for Computational Linguistics.

Zeyan Liu, Zijun Yao, Fengjun Li, and Bo Luo. 2024. On the Detectability of ChatGPT Content: Benchmarking, Methodology, and Evaluation through the Lens of Academic Writing. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*, pages 2236–2250, Salt Lake City UT USA. ACM.

Zhengyuan Liu, Ke Shi, and Nancy Chen. 2021. DMRST: A joint framework for document-level multilingual RST discourse segmentation and parsing. In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, pages 154–164, Punta Cana, Dominican Republic and Online. Association for Computational Linguistics.

William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text-interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

Bruce Mannheim. 2021. The social (and cultural, and syntactic, and semantic) life of generics. *Language in Society*, 50(4):605–618.

OpenAI. 2025. Chatgpt (version 3.5). `https://chat.openai.com/chat`. Large language model accessed on October 3, 2025 from https://chat.openai.com/chat.

Yikang Pan, Liangming Pan, Wenhu Chen, Preslav Nakov, Min-Yen Kan, and William Wang. 2023. On the Risk of Misinformation Pollution with Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP*, pages 1389–1403, Singapore. Association for Computational Linguistics.

Eleftheria Papageorgiou, Christos Chronis, Iraklis Varlamis, and Yassine Himeur. 2024. A Survey on the Use of Large Language Models (LLMs) in Fake News. *Future Internet*, 16(8):298.

Martial Pastor, Nelleke Oostdijk, and Martha Larson. 2024. The Contribution of Coherence Relations to Understanding Paratactic Forms of Communication in Social Media Comment Sections. In *JADT 2024 : 17th International Conference on Statistical Analysis of Textual Data*, Brussels (Belgium), Belgium.

Martial Pastor, Nelleke Oostdijk, Patricia Martin-Rodilla, and Javier Parapar. 2025. Enhancing discourse parsing for local structures from social media with LLM-generated data. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8739–8748, Abu Dhabi, UAE. Association for Computational Linguistics.

Francis Jeffry Pelletier and Nicholas Asher. 1997. Generics and defaults. In *Handbook of logic and language*, pages 1125–1177. Elsevier.

Uwe Peters and Benjamin Chin-Yee. 2025. Generalization bias in large language model summarization of scientific research. *Royal Society Open Science*.

Andrew Potter. 2008. Interactional coherence in asynchronous learning networks: A rhetorical approach. *The Internet and Higher Education*, 11(2):87–97.

Kristina Radivojevic, Nicholas Clark, and Paul R. Brenner. 2024. LLMs Among Us: Generative AI Participating in Digital Discourse. In *Proceedings of the AAAI Spring Symposium Series*, pages 209–218. AAAI Press.

Tom Ralston. 2024. Reconceptualising the Psychological Theory of Generics. *Philosophical Studies*, 181(11):2973–2995.

Nils Reiter and Anette Frank. 2010. Identifying generic noun phrases. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 40–49.

Kevin Reuter, Eleonore Neufeld, and Guillermo Del Pinal. 2025. Generics and Quantified Generalizations: Asymmetry Effects and Strategic Communicators. *Cognition*, 256:106004.

Marjorie Rhodes, Susan A. Gelman, and Sarah-Jane Leslie. 2025. How generic language shapes the development of social thought. *Trends in Cognitive Sciences*, 29(2):122–132. Publisher: Elsevier.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019. ATOMIC: An atlas of machine commonsense for if-then reasoning. *ArXiv*, abs/1811.00146.

Naomi Saphra and Adam Lopez. 2018. Understanding learning dynamics of language models with SVCCA. *CoRR*, abs/1811.00225.

Lenhart K Schubert and Francis Jeffry Pelletier. 1987. Problems in the representation of the logical form of generics, plurals, and mass nouns. *New directions in semantics*, pages 385–451. Publisher: Academic Press London.

Becca Smith and Alexander Williams Tolbert. 2025. The Problem of Generics in LLM Training. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, pages 1275–1280, Athens Greece. ACM.

Carlota S Smith. 2003. *Modes of discourse: The local structure of texts*, volume 103. Cambridge University Press.

Katharine Rachel Sterken. 2015. Generics in context. *Philosophers' Imprint*, 15(21):1–30.

Karolina Zaczynska and Manfred Stede. 2024. Rhetorical strategies in the un security council: Rhetorical structure theory and conflicts. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 15–28.

Amir Zeldes, Tatsuya Aoyama, Yang Janet Liu, Siyao Peng, Debopam Das, and Luke Gessler. 2025. erst: A signaled graph theory of discourse relations and organization. *Computational Linguistics*, 51(1):23–72.

A Appendix

A.1 Overview of Theories of Generics

Since the 1970s, generics have been studied in linguistics, philosophy, cognitive science, and psychology. One camp, grounded in formal semantics and pragmatics, analyzes generics as logical operators scoping over predicate arguments (Carlson, 1977; Krifka et al., 1995), possible worlds (Cohen, 2012; Pelletier and Asher, 1997), or relevant entities (Schubert and Pelletier, 1987). Others analyze generics as comparative probabilities

(Cohen, 2004), indexicals (Sterken, 2015), stereotypes (Geurts, 1985; Declerck, 1986), or prototypes (Heyer, 1985).

The other camp, rooted in cognitive psychology, treats generics as reflecting a default tendency in human cognition to generalize, the *Generics-as-Default* hypothesis (Leslie, 2007; Gelman and Pelletier, 2010). Semantics–pragmatics approaches focus on theoretical issues such as fuzzy truth-conditions, inferential asymmetry (Cella and Rosola, 2025), links to social stereotyping (Ralston, 2024), and informal reasoning (Hampton, 2012), while psychological approaches are more empirical, emphasizing acquisition and acceptability in reasoning (Leslie, 2008). In short, formal semantics emphasizes realization (structures and mechanisms), while experimental psychology focuses on evaluation or processing (how speakers assess generics).

A.2 Genericity Score

Label	Weight Group	W	Example	Weight Rationale
BOUNDED EVENT (SPECIFIC)	Other	0.0	Marie Curie, the only woman to earn two Nobel prizes, no lessexpressed (machine) A partial ban was created in Paris for driving to reducce the amount of pollution. (human)	Exclude. Clauses with specific main referents should be disregarded.
BOUNDED EVENT (GENERIC)	Impure	0.5	Under this system, citizens elected members of Parliament, (machine) Many schools started online classes due to emergencies (human)	Weak include. Generic bounded events
UNBOUNDED EVENT (SPECIFIC)	Other	0.0	The initiative creates several benefits (machine) BMW is making a car with a in-car safety feature to reduce the use of cellphones (human)	Exclude. Clauses with specific main referents should be disregarded.
UNBOUNDED EVENT (GENERIC)	Impure	0.5	With online classes taking place at home, (machine) when people were just looking down at their phones while crossing the street. (human)	Weak include. Unbounded events with generic referents are relevant types of clauses, but less important than full generic sentences.
BASIC STATE	Other	0.0	The Electoral College system has been a controversial topic in American politics for decades. (machine) Paris has the most air pollution with 147 micrograms per cubic meter. (human)	Exclude. Basic states have specific main referents and should be disregarded.
COERCED STATE (SPECIFIC)	Other	0.0	In this essay, I will delve into why seeking multiple opinions when making hard decisions can help safeguard such choices, (machine) This requirement will hopefully increase the amount of learning opportunities, through educational extracurricular activities. (human)	Exclude. Clauses with specific main referents should be disregarded.
COERCED STATE (GENERIC)	Impure	0.5	The benefits of seeking advice and multiple perspectives can also apply to other areas of life, such as relationships and academic challenges. (machine) Seeking multiple opinions can help someone make better choices, (human)	Weak include. Statements with coerced verb constellations are not necessarily about unbounded events.
PERFECT COERCED STATE (SPECIFIC)	Other	0.0	The 'Face on Mars' has sparked a massive debate (machine)	Exclude. Clauses with specific main referents should be disregarded.
PERFECT COERCED STATE (GENERIC)	Impure	0.5	increasing road safety has become a necessity (machine). Parks have bloomed. (human)	Weak include. The aspect of the verb constellation does not impact genericity significantly.
GENERIC SENTENCE (DYNAMIC)	Pure	1.0	Tongue societies are supporting conspiracy theorists (machine) Many schools throughout the world offer distance learning as an option (human)	Include. Generic sentences are the primary clause type that express generic statements. We consider the eventive variety to be as important as the stative variety (habitual and static).
GENERIC SENTENCE (STATIC)	Pure	1.0	Littering and uncleanliness are persistent issues in school premises (machine) An alarming number of traffic accidents are linked to driving while distracted, (human)	Include. Generic sentences are the primary clause type that express generic statements. We consider the eventive variety to be as important as the stative varieties (habitual and static).
GENERIC SENTENCE (HABITUAL)*	Pure	1.0	Students go out on Thursdays	Include. Generic sentences are the primary clause type that express generic statements. We consider the eventive variety to be as important as the stative varieties (habitual and static).
GENERALIZING SENTENCE (DYNAMIC)	Other	0.0	Soliciting advice from multiple sources, on the other hand, elevates decision-making by presenting multiple perspectives, (machine) Educators continually strive to identify unique and meaningful ways of improving student outcomes. (machine)	Exclude. Generalizing sentences have specific main referents.
GENERALIZING SENTENCE (STATIVE)*	Other	0.0	Members of parliament are old	Exclude. Generalizing sentences have specific main referents.
OTHER	Other	0.0	–	Exclude. All other clause types should be disregarded.
IMPERATIVE	Other	0.0	Contemplate that a good decision arrives quickly, (machine) Let kids have fun, (human)	Exclude. Clauses with nonindicative main verbs are excluded.
QUESTION	Other	0.0	then per their modeling shouldn't the behavior be genuine appreciation instead of futile obligation? (machine) but what if we change that? (human)	Exclude. Clauses with nonindicative main verbs are excluded.

Table 3: The full label set from Hemmatian (2021) with our grouping, weights, examples, and weighting rationale. Human / machine examples are taken from OUTFOX and constructed by us in cases where they do not occur in the data (*).

	Outfox			Aeon		
Clause Type	Raw Freq.	Proportion	Freq. Rank	Raw Freq.	Proportion	Freq. Rank
GENERIC SENTENCE (STATIC)	201,219	26	1	3,168,447	28	1
GENERIC SENTENCE (DYNAMIC)	102,183	13	4	2,305,068	20	2
BASIC STATE	114,298	15	2	1,149,227	10	4
COERCED STATE (GENERIC)	65,073	8	5	1,957,043	17	3
BOUNDED EVENT (SPECIFIC)	108,256	14	3	394,577	4	7
UNBOUNDED EVENT (SPECIFIC)	46,577	6	6	500,036	4	6
UNBOUNDED EVENT (GENERIC)	12,924	2	12	568,493	5	5
COERCED STATE (SPECIFIC)	22,429	3	7	291,738	3	8
BOUNDED EVENT (GENERIC)	20,160	3	8	74,972	1	14
PERFECT COERCED STATE (GENERIC)	16,793	2	9	176,204	2	12
OTHER	12,013	2	13	220,075	2	10
GENERALIZING SENTENCE (DYNAMIC)	11,301	2	14	187,578	2	11
QUESTION	15,224	2	11	72,644	1	15
PERFECT COERCED STATE (SPECIFIC)	15,314	2	10	88,160	1	13
IMPERATIVE	7,387	1	15	246,545	2	9
GENERIC SENTENCE (HABITUAL)	0	0	16	0	0	16
GENERALIZING SENTENCE (STATIVE)	0	0	17	0	0	17

Table 4: Clause Type Frequencies in Outfox and Aeon.

Clause Type	Ablated Score	Baseline	$\|\Delta\|$	N	%	Rank
GENERIC SENTENCE (STATIC)	0.341	0.625	0.284	3,168,447	28	1
GENERIC SENTENCE (DYNAMIC)	0.410	0.625	0.214	2,305,068	20	2
COERCED STATE (GENERIC)	0.537	0.625	0.087	1,957,043	17	3
BASIC STATE	0.625	0.625	0.000	1,149,227	10	4
UNBOUNDED EVENT (GENERIC)	0.598	0.625	0.027	568,493	5	5
UNBOUNDED EVENT (SPECIFIC)	0.625	0.625	0.000	500,036	4	6
BOUNDED EVENT (SPECIFIC)	0.625	0.625	0.000	394,577	4	7
COERCED STATE (SPECIFIC)	0.625	0.625	0.000	291,738	3	8
IMPERATIVE	0.625	0.625	0.000	246,545	2	9
OTHER	0.625	0.625	0.000	220,075	2	10
GENERALIZING SENTENCE (DYNAMIC)	0.625	0.625	0.000	187,578	2	11
PERFECT COERCED STATE (GENERIC)	0.617	0.625	0.008	176,204	2	12
PERFECT COERCED STATE (SPECIFIC)	0.625	0.625	0.000	88,160	1	13
BOUNDED EVENT (GENERIC)	0.621	0.625	0.004	74,972	1	14
QUESTION	0.625	0.625	0.000	72,644	1	15
GENERIC SENTENCE (HABITUAL)	0.625	0.625	0.000	0	0	16
GENERALIZING SENTENCE (STATIVE)	0.625	0.625	0.000	0	0	17

Table 5: Ablation results by clause type in OUTFOX. Scores are averaged over all texts in the dataset. Baseline is Mean Genericity Score with our weights.

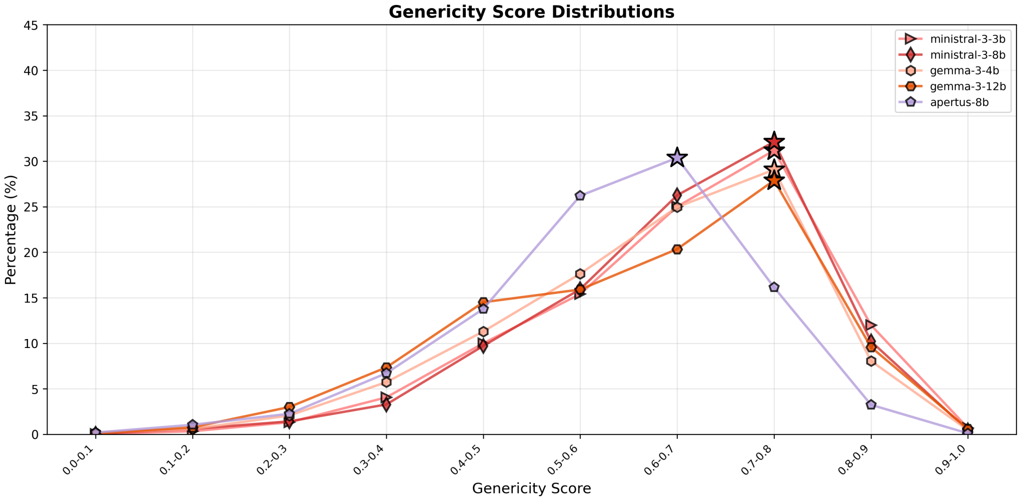

Figure 4: Genericity score for small open-weight models, demonstrating similar trends as observed in their larger variants. See Section 5.1, Figure 1.

Clause Type	Ablated Score	Baseline	|Δ|	N	%	Rank
GENERIC SENTENCE (STATIC)	0.209	0.470	0.262	201,219	26	1
BASIC STATE	0.470	0.470	0.000	114,298	15	2
BOUNDED EVENT (SPECIFIC)	0.470	0.470	0.000	108,256	14	3
GENERIC SENTENCE (DYNAMIC)	0.337	0.470	0.133	102,183	13	4
COERCED STATE (GENERIC)	0.428	0.470	0.042	65,073	8	5
UNBOUNDED EVENT (SPECIFIC)	0.470	0.470	0.000	46,577	6	6
COERCED STATE (SPECIFIC)	0.470	0.470	0.000	22,429	3	7
BOUNDED EVENT (GENERIC)	0.457	0.470	0.014	20,160	3	8
PERFECT COERCED STATE (GENERIC)	0.459	0.470	0.011	16,793	2	9
PERFECT COERCED STATE (SPECIFIC)	0.470	0.470	0.000	15,314	2	10
QUESTION	0.470	0.470	0.000	15,224	2	11
UNBOUNDED EVENT (GENERIC)	0.462	0.470	0.009	12,924	2	12
OTHER	0.470	0.470	0.000	12,013	2	13
GENERALIZING SENTENCE (DYNAMIC)	0.470	0.470	0.000	11,301	2	14
IMPERATIVE	0.470	0.470	0.000	7,387	1	15
GENERIC SENTENCE (HABITUAL)	0.470	0.470	0.000	0	0	16
GENERALIZING SENTENCE (STATIVE)	0.470	0.470	0.000	0	0	17

Table 6: Ablation results by clause type in Aeon-essays. Scores are averaged over all texts in the dataset. Baseline is Mean Genericity Score with our weights.

Dataset	Mean Inverted	Mean Baseline	|Δ|
Outfox	0.100 ± 0.000	0.625 ± 0.150	0.525
Aeon	0.100 ± 0.000	0.470 ± 0.133	0.370

(a) Reciprocal Inversion Results. Weights are assigned based on inverse frequency (rarer labels get higher weights).

Dataset	Mean Inverted	Mean Baseline	|Δ|
Outfox	0.375 ± 0.150	0.625 ± 0.150	0.250
Aeon	0.530 ± 0.133	0.470 ± 0.133	0.059

(b) Categorical Inversion Results. Weights are based on inverting "pure" (1.0) generic clause types and "other" (0.0) clause types (cf. Table 3).

Table 7: Reciprocal (left) and Categorical (right) weight inversion results for Outfox and Aeon. Scores are mean genericity for each dataset.

Trial	Aeon	Outfox	Trial	Aeon	Outfox
T1	0.494	0.500	T26	0.499	0.412
T2	0.366	0.441	T27	0.655	0.614
T3	0.489	0.502	T28	0.453	0.457
T4	0.499	0.564	T29	0.462	0.474
T5	0.444	0.464	T30	0.572	0.613
T6	0.461	0.493	T31	0.497	0.556
T7	0.369	0.323	T32	0.525	0.477
T8	0.450	0.407	T33	0.379	0.423
T9	0.535	0.490	T34	0.556	0.524
T10	0.411	0.379	T35	0.407	0.423
T11	0.554	0.634	T36	0.675	0.761
T12	0.517	0.650	T37	0.702	0.624
T13	0.455	0.595	T38	0.649	0.584
T14	0.516	0.450	T39	0.515	0.567
T15	0.376	0.452	T40	0.622	0.691
T16	0.477	0.496	T41	0.382	0.414
T17	0.333	0.380	T42	0.620	0.675
T18	0.507	0.639	T43	0.503	0.487
T19	0.379	0.346	T44	0.234	0.200
T20	0.647	0.640	T45	0.439	0.454
T21	0.549	0.567	T46	0.447	0.458
T22	0.533	0.530	T47	0.513	0.496
T23	0.445	0.474	T48	0.629	0.589
T24	0.523	0.511	T49	0.454	0.370
T25	0.319	0.349	T50	0.599	0.558
-	-	-	Mean	0.493	0.504
-	-	-	Std	0.099	0.108
-	-	-	Baseline	0.470	0.625

Table 8: Mean genericity score with random weights per trial (T1–T50).

A.3 Dataset Details

A.3.1 Outfox

Original Data The human-written essays in OUTFOX were originally collected for the Kaggle Feedback Prize competition (Franklin et al., 2022), which focused on identifying persuasive discourse elements and evaluating argument effectiveness. PERSUADE 2.0 includes over 25,000 argumentative essays written by U.S. students in grades 6–12, covering both independent and source-based writing tasks. Koike et al. (2024) supplement these human essays by first generating pseudo-problem-statements for them using ChatGPT such as: [*train_13352*]:

> "Given the following problem statement, please write an essay in 194 words with a clear opinion.
>
> Problem statements [sic]: Compare and contrast the car culture in Germany and America, including the cost of owning a car and the government's efforts to promote alternative transportation. Analyze the impact of these efforts on traffic, public transportation, and recreational activities in both countries.
>
> Essay: "

Based on this type of prompts, Koike et al. (2024) generate 15,400 (14,400 train / 500 validation / 500 test) essay responses for each of the following LLMs: OpenAI's ChatGPT 3.5 and Text-DaVinci-003 decoder-only, and Google's Flan-T5-XXL encoder-decoder model. These LLM-generated texts are complemented by the 15,400 human-written essays, sourced from PERSUADE 2.0.

Source	Train	Valid	Test	Total
chatgpt 3.5	14,400	500	500	15,400
flan-t5	14,400	500	500	15,400
davinci	14,400	500	500	15,400
human	14,400	500	500	15,400
Total	57,600	2,000	2,000	61,600

Table 9: Overview of LLM-generated and human-written texts in OUTFOX.

Augmented Data

Model Choices We choose our models with varying degrees of openness and commercial focus in mind. Starting with completely closed-source proprietary models (ChatGPT4o and 5.2), open weights but closed-source code and training data (Gemma 3 and Mistral 3), and fully open source code, training data, and technical details (Apertus). We do this to achieve as broad a coverage as possible that is representative of the current landscape, with a focus on using open models that are small enough to run with the resources available. We focus on open models since these afford more transparency, reproducibility and ethical compliance than proprietary models.

Model	Train	Valid	Test	Total
apertus 8b	14,400	500	500	15,400
apertus 70b	14,400	500	500	15,400
chatgpt 3.5	14,400	500	500	15,400
chatgpt 4o	14,400	500	500	15,400
chatgpt 5.2	14,400	500	500	15,400
flan-t5	14,400	500	500	15,400
gemma 3 4b	14,400	500	500	15,400
gemma 3 12b	14,400	500	500	15,400
gemma 3 27b	14,400	500	500	15,400
ministral 3 3b	14,400	500	500	15,400
ministral 3 8b	14,400	500	500	15,400
ministral 3 14b	14,400	500	500	15,400
davinci	14,400	500	500	15,400
human	14,400	500	500	15,400
Grand Total	**201,600**	**7,000**	**7,000**	**215,600**

Table 10: Composition of the full augmented Outfox dataset.

Generation Settings We generate 123,200 synthetic texts by reusing the prompts of the original OUTFOX dataset. Using the vLLM library (Kwon et al., 2023), we generate the texts using default precision and quantization of the model checkpoints, and standard temperature (07) and `top_p` (0.9), for all models. All texts are generated using 4 Nvidia H200 GPUs and 12 cpu cores per GPU in parallel. We manually inspect a random sample of all model outputs and perform no postprocessing of the outputs.

A.3.2 Aeon

Aeon is publically available at Kaggle[10] under the MIT License, scraped from https://aeon.co/.

	Texts	Topics	Authors
Human	2,235	114	1,655

Table 11: Basic statistics for Aeon essays.

---

[10] `https://www.kaggle.com/datasets/mannacharya/aeon-essays-dataset`

A.4 RST Relations and Genericity

	Human Outfox		Human Aeon		DaVinci 003		Flan-T5 XXL		GPT-3.5 Turbo	
Relation	**Abs.**	**Rel.**	**Abs.**	**Rel.**	**Abs.**	**Rel.**	**Abs.**	**Rel.**	**Abs.**	**Rel.**
Elaboration	206,919	22.4%	276,828	31.1%	422,065	32.5%	179,717	28.3%	428,067	36.2%
Joint	247,840	26.6%	168,248	20.6%	445,214	33.9%	165,381	24.6%	257,177	23.5%
Attribution	67,884	8.1%	50,874	5.9%	28,369	2.5%	24,453	4.0%	28,929	3.2%
Same-Unit	56,733	5.7%	58,027	5.9%	184,214	13.7%	45,976	6.8%	145,409	11.2%
Contrast	56,203	6.0%	65,424	7.3%	29,727	2.4%	41,183	6.0%	37,017	4.2%
Background	53,306	7.2%	69,222	8.3%	40,982	4.5%	41,086	7.6%	42,560	5.6%
Explanation	44,218	4.3%	17,351	1.7%	15,584	1.2%	20,989	2.9%	18,982	2.4%
Cause	42,695	4.8%	21,244	2.5%	22,687	2.0%	29,771	4.3%	33,422	3.4%
Enablement	40,685	4.8%	27,706	3.5%	33,596	2.9%	36,237	5.6%	33,825	3.1%
Condition	34,593	2.8%	17,167	1.9%	7,108	0.5%	17,298	2.1%	6,536	0.5%
Temporal	14,541	4.2%	40,893	6.5%	2,702	0.9%	6,492	4.1%	2,791	1.7%
Evaluation	13,130	1.5%	29,713	2.9%	5,928	1.2%	9,429	1.7%	8,137	2.5%
Manner-Means	9,748	0.9%	11,066	1.3%	27,836	1.9%	12,695	1.8%	30,910	2.6%
Comparison	2,709	0.3%	1,785	0.2%	1,491	0.1%	2,251	0.3%	1,793	0.2%
Topic-Com.	1,997	0.3%	2,625	0.3%	244	0.1%	553	0.1%	345	0.0%
Summary	795	0.1%	1,620	0.2%	1,015	0.2%	464	0.1%	814	0.1%
TextualOrg.	40	0.0%	144	0.0%	472	0.0%	212	0.1%	563	0.0%
Topic-Change	6	0.0%	5	0.0%	4	0.0%	1	0.0%	16	0.0%

	GPT-4o		GPT-5.2		Apertus 70B		Gemma-3 27B		Ministral-3 14B	
Relation	**Abs.**	**Rel.**	**Abs.**	**Rel.**	**Abs.**	**Rel.**	**Abs.**	**Rel.**	**Abs.**	**Rel.**
Elaboration	218,734	30.7%	286,537	28.3%	553,531	21.4%	246,472	29.1%	296,798	31.1%
Joint	183,809	25.3%	309,219	29.5%	882,990	34.5%	230,902	29.4%	245,866	25.0%
Cause	81,777	11.0%	31,979	3.6%	108,375	4.1%	46,649	5.1%	82,504	8.0%
Contrast	38,191	4.7%	73,512	8.4%	229,446	9.0%	74,960	9.3%	67,731	6.3%
Same-Unit	32,356	4.2%	35,595	3.0%	75,868	3.1%	49,403	5.1%	52,112	4.8%
Background	31,684	5.4%	39,594	4.6%	140,846	4.9%	48,520	6.9%	50,194	5.3%
Enablement	31,531	3.8%	19,321	2.6%	161,826	5.3%	21,293	2.3%	35,031	3.0%
Manner-Means	27,227	3.3%	16,136	1.6%	71,719	2.4%	12,736	1.4%	33,704	2.8%
Explanation	24,187	3.4%	38,233	4.0%	108,788	4.4%	23,995	3.0%	39,758	3.6%
Evaluation	19,271	4.1%	17,174	2.5%	41,395	1.3%	22,868	4.0%	24,663	3.5%
Attribution	14,888	2.2%	30,630	5.7%	117,941	4.5%	17,609	2.5%	38,236	4.3%
Condition	3,546	0.3%	24,225	2.0%	112,461	3.8%	3,134	0.3%	9,260	0.7%
Temporal	2,961	1.3%	8,863	4.7%	24,528	1.0%	5,092	1.1%	4,301	1.1%
Comparison	936	0.1%	3,112	0.3%	1,465	0.1%	1,059	0.1%	1,696	0.1%
Summary	878	0.2%	986	0.1%	9,755	0.4%	1,301	0.2%	2,988	0.3%
Topic-Com.	896	0.1%	406	0.0%	2,371	0.1%	536	0.1%	1,029	0.1%
TextualOrg.	18	0.0%	63	0.0%	272	0.0%	256	0.1%	759	0.1%
Topic-Change	—	—	9	0.0%	1	0.0%	—	—	2	0.0%

Table 12: Relation frequencies across all models and human annotators. Absolute counts (Abs.) and relative percentages (Rel.) are shown for each RST relation type.

	Human (Outfox)		
Relation	**0.7-0.8**	**0.8-0.9**	**0.9-1.0**
Joint	30.4	32.3	33.1
Elaboration	25.3	25.9	26.7
Cause	6.4	8.0	–
Contrast	–	–	8.2

	Human (Aeon)		
Relation	**0.7-0.8**	**0.8-0.9**	**0.9-1.0**
Elaboration	33.0	33.4	–
Joint	21.3	28.2	–
Contrast	9.7	–	–
Background	–	8.7	–

	Text DaVinci		
Relation	**0.7-0.8**	**0.8-0.9**	**0.9-1.0**
Joint	36.9	39.0	40.8
Elaboration	32.6	31.2	30.8
Same-Unit	14.4	13.8	12.4

	Flan T5		
Relation	**0.7-0.8**	**0.8-0.9**	**0.9-1.0**
Joint	28.6	29.9	32.6
Elaboration	28.1	29.3	28.6
Same-Unit	7.2	7.0	6.5

	ChatGPT 3.5		
Relation	**0.7-0.8**	**0.8-0.9**	**0.9-1.0**
Elaboration	39.7	38.0	36.0
Joint	25.3	27.0	29.8
Same-Unit	13.6	12.2	11.0

	ChatGPT 4o		
Relation	**0.7-0.8**	**0.8-0.9**	**0.9-1.0**
Elaboration	30.8	31.3	32.5
Joint	27.3	28.3	31.6
Cause	11.8	13.2	13.4

	ChatGPT 5.2		
Relation	**0.7-0.8**	**0.8-0.9**	**0.9-1.0**
Joint	35.2	36.9	40.8
Elaboration	30.3	30.9	27.2
Contrast	7.5	7.4	6.5

	Apertus-70b		
Relation	**0.7-0.8**	**0.8-0.9**	**0.9-1.0**
Joint	32.1	33.3	28.9
Elaboration	27.9	27.3	29.2
Cause	6.9	7.5	–
Contrast	–	–	17.7

	Gemma-3-27b		
Relation	**0.7-0.8**	**0.8-0.9**	**0.9-1.0**
Elaboration	30.8	30.6	27.6
Joint	30.3	31.1	30.2
Contrast	9.0	9.2	10.6

	Ministral-3-14b		
Relation	**0.7-0.8**	**0.8-0.9**	**0.9-1.0**
Elaboration	30.0	31.2	29.9
Joint	26.5	28.5	33.7
Cause	9.1	10.4	12.2

Table 13: Top three discourse relations across all models and genericity bins. Values represent relative percentages within each bin. Dashes indicate the relation was not in the top 3 for that bin.

Prompt	ChatGPT 3.5	Flan	Human	Davinci	N cases
Overall	0.683352	0.623729	0.556236	0.668157	500
Explain	0.683233	0.605270	0.556497	0.642978	147
Explain (the reason) why	0.620512	0.566785	0.482481	0.639270	45
How/Explain how	0.714259	0.646404	0.601629	0.720321	30
Evaluate	0.691751	0.676642	0.556515	0.676474	6
Compare/Contrast	0.721390	0.632768	0.642531	0.699455	10

Table 14: Average Genericity Scores across writers in the original Outfox dataset with varying keywords in prompts.

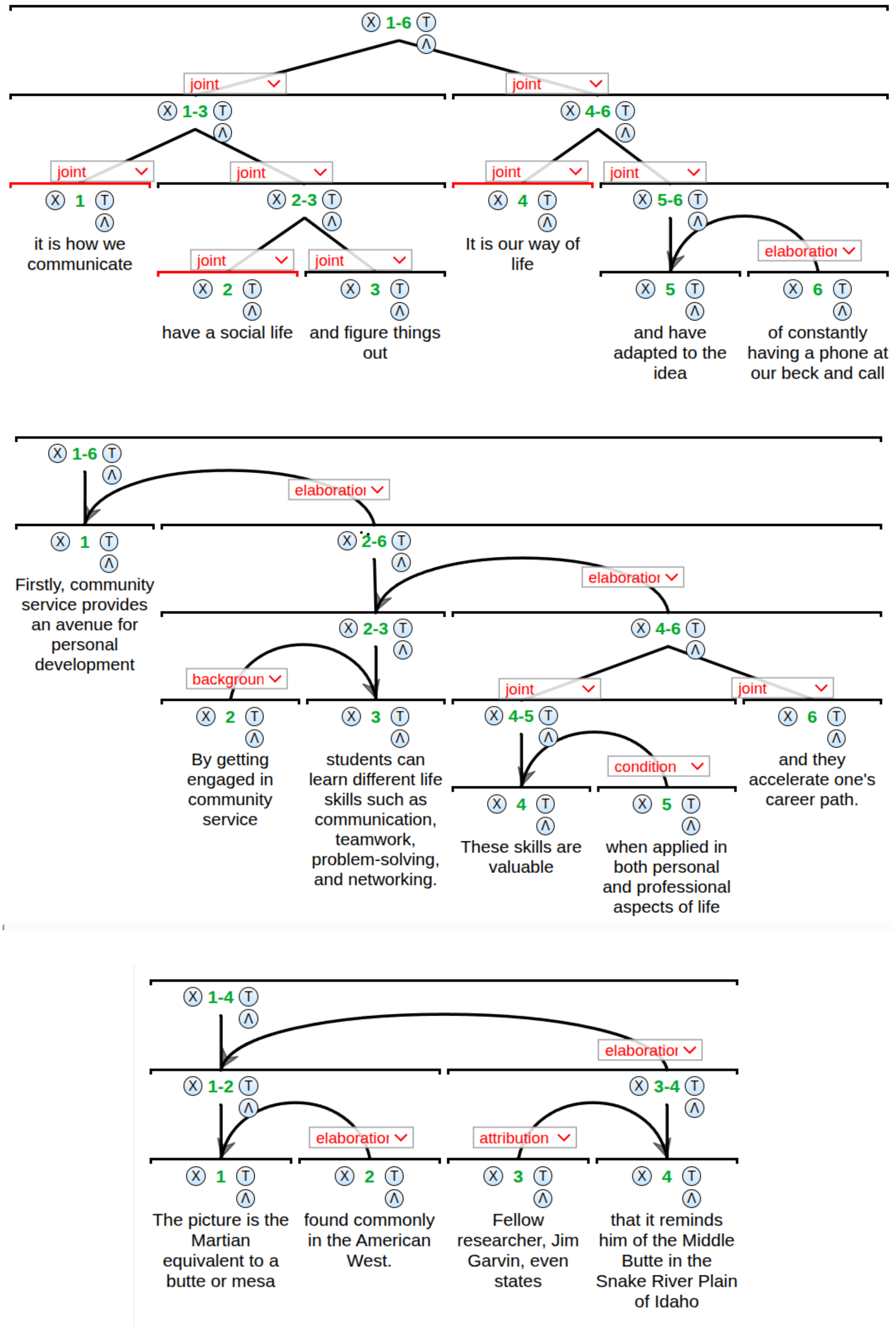

Figure 5: **(Top)** RST subtree illustrating a human-produced paratactic (JOINT–JOINT–JOINT) pattern [*human_train_41*]. **(Middle)** RST subtree illustrating ChatGPT-produced embedded generic ELABORATIONS [*chatgpt_train_146*]. **(Bottom)** RST subtree illustrating the use of ELABORATIONs for locality in Human text [*human_train_41*].

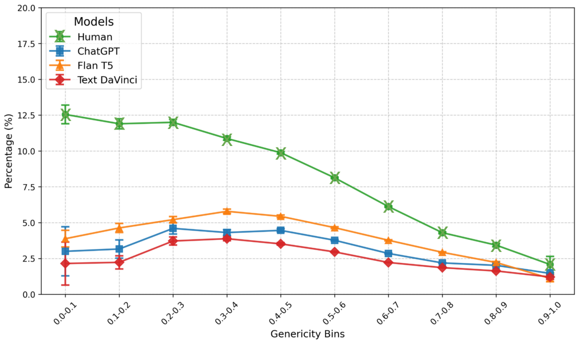

Figure 6: Histograms of ATTRIBUTION across ten genericity bins (0.0–0.1 to 0.9–1.0) for Human, ChatGPT 3.5, Flan-T5-XXL, and Text-DaVinci-003.