AI Argues Differently: Distinct Argumentative and Linguistic Patterns of LLMs in Persuasive Contexts

Esra Dönmez[♦1,2], **Maximilian Maurer**[★3,4], **Gabriella Lapesa**[3,4], **Agnieszka Falenska**[1,2]
[1]Institute for Natural Language Processing, University of Stuttgart
[2]Interchange Forum for Reflecting on Intelligent Systems, University of Stuttgart
[3]GESIS Leibniz Institute for the Social Sciences [4]Heinrich-Heine University Düsseldorf
♦esra.doenmez@ims.uni-stuttgart.de ★maximilian.maurer@gesis.org

Abstract

Distinguishing LLM-generated text from human-written is a key challenge for safe and ethical NLP, particularly in high-stake settings such as persuasive online discourse. While recent work focuses on detection, real-world use cases also demand interpretable tools to help humans understand and distinguish LLM-generated texts. To this end, we present an analysis framework comparing human- and LLM-authored arguments using two easily-interpretable feature sets: general-purpose linguistic features (e.g., lexical richness, syntactic complexity) and domain-specific features related to argument quality (e.g., logical soundness, engagement strategies). Applied to */r/ChangeMyView* arguments by humans and three LLMs, our method reveals clear patterns: LLM-generated counter-arguments show lower type-token and lemma-token ratios but higher emotional intensity – particularly in anticipation and trust. They more closely resemble textbook-quality arguments – cogent, justified, explicitly respectful toward others, and positive in tone. Moreover, counter-arguments generated by LLMs converge more closely with the original post's style and quality than those written by humans. Finally, we demonstrate that these differences enable a lightweight, interpretable, and highly effective classifier for detecting LLM-generated comments in CMV.

1 Introduction

What makes LLM-generated text distinct? Uncovering the characteristics of LLM-generated language is emerging as a key challenge in NLP research. Detecting such content has become increasingly difficult, as models now generate text that closely resembles human writing (Doughman et al., 2025). At the same time, the fluency of LLM-generated text amplifies the potential for misuse in online public discourse (Tang et al., 2024).

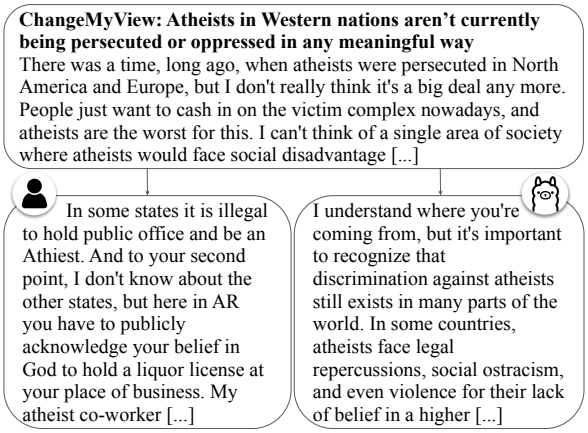


Figure 1: Discussion example from */r/ChangeMyView* subReddit. Top: a post titled "CMV: Atheists in Western nations aren't currently being persecuted or oppressed in any meaningful way". Below, left: human-written comment. Below, right: LLM-generated comment.

In this work, we contribute to this pressing area of research while focusing on a specific domain: **persuasion in online discussions**. Recent developments in this regard are alarming, including unauthorized and unethical experiments involving LLM-driven bots in */r/ChangeMyView* (CMV)[1] – an online platform for opinion exchange – as well as the growing persuasive power of these tools on topics of critical societal relevance (Goldstein et al., 2024; Bai et al., 2023; Potter et al., 2024). Consider the example in Figure 1: a CMV user posts their view (top panel, the original post, henceforth OP, claims that *Atheists are not oppressed in Western countries*), challenging the community to persuade them otherwise. On the left, a real user presents a counter-argument; on the right, an LLM-generated response without revealing its source. Both comments are fluent and provide solid reasoning. However, only one reflects a real human's lived experi-

♦, ★These authors contributed equally to this work.

[1]www.reddit.com/r/changemyview/comments/1k8b2hj/meta_unauthorized_experiment_on_cmv_involving

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 34584–34615
November 4-9, 2025

ences and genuine thoughts – the qualities that the CMV users expect when inviting others to share their individual perspectives (Jhaver et al., 2017). This ability of LLMs to engage fluently in spaces designed for authentic human exchange poses a challenge to the norms and expectations of online communities.

Such cases, like the example above, underscore the urgency of understanding the traits and impacts of LLM-generated text and developing effective detection methods. Therefore, unlike previous work focused on *detection* of generated content (Wang et al., 2024; Koike et al., 2024, inter alia), we take a fundamental methodological step towards **understanding the key differences between LLM-generated and human-written argumentative texts**. We draw on an established dataset of real-world conversations from CMV (Tan et al., 2016) augmented with LLM-generated responses (Dönmez and Falenska, 2025). The data includes OPs, human-written comments, and parallel to these responses outputs from three LLMs: GPT-3.5-turbo, LLAMA2-7B, and MISTRAL-7B (Figure 1, bottom panel, right). We enrich this data by extracting two types of features for all posts and (human-written and model-generated) comments: a) *linguistic*, e.g., textual complexity, syntactic structures, use of emotions (301 features), and b) *argument quality*, e.g., whether the argument is sound or rhetorically impactful (27 features drawn from both argumentation and social science theory on discussion quality).

The enriched dataset enables two types of comparisons. First, we compare **human- vs. LLM-authored comments** to address **RQ1**: What linguistic and argument quality features characterize LLM-generated arguments, and how do they differ from human-written ones? Second, we compare **original posts with the comments they receive** (from humans or LLMs) to address **RQ2**: To what extent do LLMs align with the style and quality of the original posts, and do they follow different convergence patterns?[2]

We find that, in our experimental setup, *AI argues differently*. Concretely, LLM-generated argumentative comments follow substantially distinct distributions from the human-written ones (§4.1), both in terms of the linguistic features (e.g., lexical richness and emotion) and the argument quality ones (e.g., cogency, degree of justification for the claim, presence of a concrete proposal, respect shown to the interlocutor). Furthermore, parallel arguments generated by LLMs are highly correlated with one another, but show much weaker correlation with those written by humans (§4.2). Regarding the stylistic convergence, LLM-generated comments converge more strongly to the style and quality of the original posts than human-written ones (§4.3). Observing these apparent differences raises the question (**RQ3**): Can these differences between human and LLM texts be used to effectively detect AI-generated content? Our classification experiments demonstrate that linguistic and/or argument quality features distinguish LLM responses to CMV posts from human-written ones with over 98% accuracy (§5.2). Furthermore, despite its simplicity, our interpretable, feature-based method generalizes well to other domains and performs comparably to more computationally intensive detection models.

The contributions of our work[3] are twofold. First, we draw a comprehensive picture of the distinct linguistic properties of LLM-generated arguments, taking a vital step towards a deeper understanding of LLM-generated argumentative texts within a context of substantial societal relevance. Second, we demonstrate that these properties are effective at detecting LLM-generated arguments with a simple, lightweight, and interpretable method that can be smoothly deployed in online applications.

2 Related Work

2.1 Assessment of LLM-Generated Texts

Much of the literature investigating LLM-generated texts is focused on **the detection of it**. Several works have proposed benchmarks for domains ranging from scientific texts, over creative writing, to news articles (Wang et al., 2024; Koike et al., 2024; Li et al., 2024b; Abdalla et al., 2023; He et al., 2024; Li et al., 2024a; Dugan et al., 2022; Pu et al., 2023, inter alia). However, these approaches depend mostly on computationally intensive methods – like pretrained representations or LLM inference – that are not only impractical in low-resource settings but also lack interpretable features.

On **the analysis of LLM-generated texts**, prior studies identified differences from human writing

[2]We borrow the term *convergence* from sociolinguistics literature, where it refers to behavioral strategies in which a speaker modifies their communication to become more similar to the communication styles of others (see the Communication Accommodation Theory (Giles and Ogay, 2013) used, for example, in Gasiorek and Vincze (2016)).

[3]Our code and data are available on GitHub.

(Seals and Shalin, 2023; Sandler et al., 2025, inter alia). For instance, Muñoz-Ortiz et al. (2024) find that humans exhibit more variance in sentence length, vocabulary, and syntax, while LLMs use more pronouns, numbers, symbols, and auxiliaries. Similarly, Reinhart et al. (2025) find that instruction-tuned LLMs produce more noun-heavy and information-dense text. However, such studies typically focus on lexical, morphosyntactic, or rhetorical features – rarely exceeding 100 features – and overlook domain-specific aspects like argument quality. Our work addresses this gap by leveraging over 300 linguistic and domain-specific features for a more comprehensive analysis.

2.2 Computational Argumentation

While prior research has used CMV as a benchmark dataset, the focus has largely been on generation techniques and evaluating output quality along one or a few dimensions (Alshomary et al., 2021; Alshomary and Wachsmuth, 2023; Lin et al., 2023). For instance, Hua and Wang (2019) manually compare human and machine arguments in CMV, but only over 30 samples and 3 argumentation styles. To our knowledge, we are the first to provide a large-scale, fine-grained analysis of the differences between human- and LLM-authored arguments.

Interest in these differences has also grown in **computational social science**, driven by concerns about LLMs' influence on public discourse (Palmer and Spirling, 2023; Salvi et al., 2024; Tessler et al., 2024). Palmer and Spirling (2023) studied preferences for arguments (not counter-arguments) by humans and LLMs across topic-stance pairs (e.g., abortion restrictions), finding that LLMs were sometimes preferred – but only for certain topics and when their origin was hidden. LLM arguments were also simpler and more positive in tone. While closely related, their work is limited in scale (30 LLM- and 25 human-authored arguments per 9 topics) and in feature depth. In contrast, our approach uses hundreds of linguistic and domain-specific features, providing a more comprehensive basis for social science research.

3 Methods

In this section, we describe the methodological steps taken to analyze the differences between human-written and LLM-generated arguments.

3.1 Data

CMV The subreddit CMV follows a structured format: users submit an original post (OP) stating a point of view with supporting arguments, and others respond with comments that challenge it (see example in Figure 1). These responses form a tree-like structure, allowing replies to both the OP and other comments.

We use the publicly available CMV corpus from Tan et al. (2016), a well-established dataset that predates the rise of LLMs and has been used in multiple studies (Hidey et al., 2017; Falenska et al., 2024, inter alia). We merge the training and held-out sets and remove entries with missing or deleted text, missing parent posts/comments, or texts too short to form meaningful arguments ($\leq$10 characters). To focus on discussions where LLM influence is particularly concerning, we then filter sociopolitical posts using the Reddit-trained classifier from Monti et al. (2022), which yields $13,498$ posts.

Regarding the comments, we take only first-level (direct) comments, excluding replies within longer discussion threads. To enable pairwise comparisons, for each post, we select the most and the least up-voted comments[4] – CMV↑ to ensure high-quality arguments and CMV↓ to disentangle the quality effects (collectively referred to as C^H).

LLM We use the LLM-generated comments to Tan et al.'s (2016) CMV posts from Dönmez and Falenska (2025). The data includes responses from an instruction-tuned model (MISTRAL-7B), an open-access chat model (LLAMA2-7B), and an API-access chat model (GPT-3.5-turbo). For each OP, the dataset includes one counter-argument per model, which were obtained by prompting the models with the OP appended with `You have one chance to change my view. Answer:` (decoded via nucleus sampling; temp=.9, top_p=.6, max_len=600).

In summary, the dataset contains $13,498$ original posts (OPs; original arguments), human-written counter-argument comments – CMV↑ and CMV↓ ($13,498$ each), together as C^H ($26,996$) – and LLM-generated counter-argument comments ($13,498$ per model, combined as C^L).

[4] CMV includes Δ annotations marking when a comment changed a user's view, often used to indicate persuasiveness (e.g., Monti et al., 2022, inter alia). Due to their sparse and uneven distribution, we rely on community votes instead.

3.2 Features

We extract two types of features – domain-specific *argument quality* metrics and general-purpose *linguistic* features – from all posts and comments in our dataset.

Argument Quality Features These domain-specific features – referred to as argument quality for brevity – fall into two categories: *argument quality* as understood in the NLP community (e.g., logical structure, rhetorical effectiveness), and *deliberative quality* from the social sciences (e.g., presence of concrete proposals, interactivity, respect, empathy, emotional expression, and storytelling). These two perspectives offer complementary, partially overlapping answers to a central question: *Is this comment a good argument?*

To obtain these metrics, we use publicly available argument quality adapters from Falk and Lapesa (2023), trained on established argument quality datasets. Each adapter specializes in scoring a particular quality dimension. Some dimensions capture broad notions like quality or impact, while others reflect the finer distinctions outlined above, based on the original dataset annotations (detailed in §B.1.1, Table 6).

Using these models, we score each text in our data. We then scale all scores to the range of $[0, 1]$ for comparability.

Linguistic Features While argument quality metrics capture domain-relevant dimensions, they offer a limited view of the full linguistic footprint of a text. Therefore, we expand this feature set by extracting 301 linguistic features using the open-source library `elfen` (Maurer, 2025). These features are from the areas *surface* (e.g. number of tokens), *lexical richness* (e.g. type-token ratio), *readability* (e.g. Gunning Fog index), *psycholinguistics* (e.g. concreteness), *information theory* (e.g. entropy), *emotion* (e.g. anger intensity), *semantics* (e.g. polysemy), *named entities*, *parts-of-speech*, *morphology*, and *syntactic dependencies* (detailed in §B.1.2).

Each feature dependent on token counts (e.g., number of named entities) is normalized by the total number of tokens of the instance, and we scale all features to the range of $[0, 1]$ for comparability.

3.3 Metrics

We use standard statistical metrics in our analyses: **Wasserstein-1 distance (WS)** to measure differences between empirical distributions per argument quality dimension and linguistic feature, **difference of means (D)** to compare the central tendencies of C^H and C^L across each argument quality dimension and linguistic feature, and **pairwise Pearson's** r correlation coefficient to measure the feature correlations between pairs of texts. We refer to §B.2.1 for the formulas.

Dimension	WS	D	Dimension	WS	D
reference	0.35	0.35	overall	0.20	-0.20
impact	0.31	-0.31	justification^{++}	0.19	-0.19
interactivity^{+}	0.28	-0.28	reasonableness	0.18	-0.18
effectiveness	0.24	-0.24	emotion^{-}	0.18	0.18
respect$^{\text{explicit}}$	0.23	-0.23	cogency	0.18	-0.18
respect$^{\text{implicit}}$	0.21	0.21	QforJustification	0.13	0.13
quality	0.20	-0.20	proposal	0.12	-0.12
justification^{+++}	0.10	-0.10			

Table 1: Wasserstein distances (WS) and mean differences (D) between the human and LLM comments in aggregate and descending order for argument quality features of $WS(C^H, C^L) \geq 0.1$. For the remaining features, see Appendix Table 9.

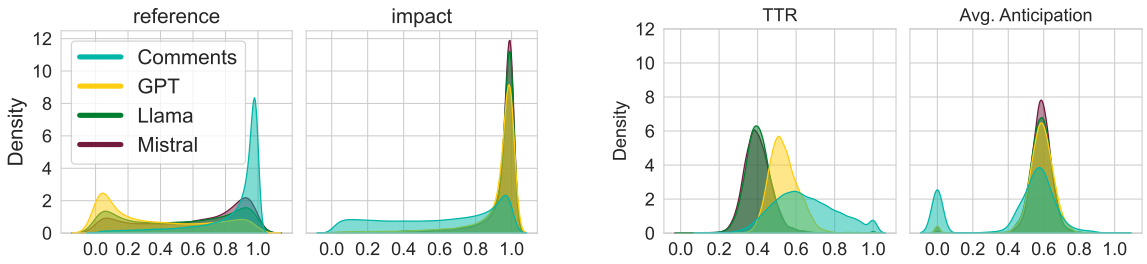
4 Key Characteristics of LLM-generated Arguments

We now turn to our primary objective: to uncover the key argumentative and linguistic characteristics of LLM-generated texts. Two answer our first two research questions, we perform two comparison analyses: (1) *distributional differences* between the human-written and LLM-generated parallel comments and the correlations between the two (RQ1), and (2) *style and quality convergence* of comments (by humans or LLMs) to the linguistic style or the argument quality properties of the OPs (RQ2).

4.1 Distributional Differences

To identify and discuss particularly pronounced differences between LLM and human counter-arguments, we apply a threshold of $WS \geq 0.1$. We set this threshold empirically to focus on the top-n most distinct features.

Argument Quality Features Table 1 presents the most prominent argument quality features. We observe particularly substantial distributional differences ($WS > 0.2$) between C^H and C^L for dimensions such as *references* to other discourse participants, *impact* of an argument given its context, *positive interactivity* with others' arguments, rhetorical *effectiveness*, and *explicit and implicit respect* toward other groups. To make these WS



(a) Argument Quality Features

(b) Linguistic Features

Figure 2: Example distributions from the top argument quality and linguistic features with the largest WS between the human-written and LLM-generated argument comments to CMV OPs. CMV stands for C^H.

Feaure	WS	D
Type token ratio (TTR)	0.20	0.20
Lemma token ratio (LTR)	0.19	0.19
Avg. intensity anticipation	0.13	-0.12
POS variability	0.12	0.12
Herdan's C	0.11	0.11
Avg. intensity trust	0.11	-0.10

Table 2: Wasserstein distances (WS) and mean differences (D) between the human and LLM comments in aggregate and descending order for linguistic features of $WS(C^H, C^L) \geq 0.1$ ($\geq .05$ in Appendix Table 14).

values more interpretable, we visualize the distributions for the two most prominent features – *reference* and *impact* – in Figure 2a. As the plots show, human arguments (in blue) differ substantially from LLM-generated ones, while the LLMs produce notably similar patterns across models (cf. Appendix Table 11; plots for $WS \in [0.1, 0.2]$ in Figure 7).

Regarding the differences in mean distributions, a positive D shows that human-written arguments score higher on average for a given dimension, and the negative D indicates the opposite. Accordingly, we find that LLM arguments contain fewer references to other discourse participants (0.35), display less implicit respect (0.2) and less negative emotion (0.17), and do not request a justification as often, i.e., humans display a higher inquisitive behavior (0.13) while, at the same time, providing less sophisticated levels of justification than LLMs. For the remaining dimensions, we observe the opposite: LLM arguments, among others, score higher in impact and effectiveness and show more positive interaction with others.

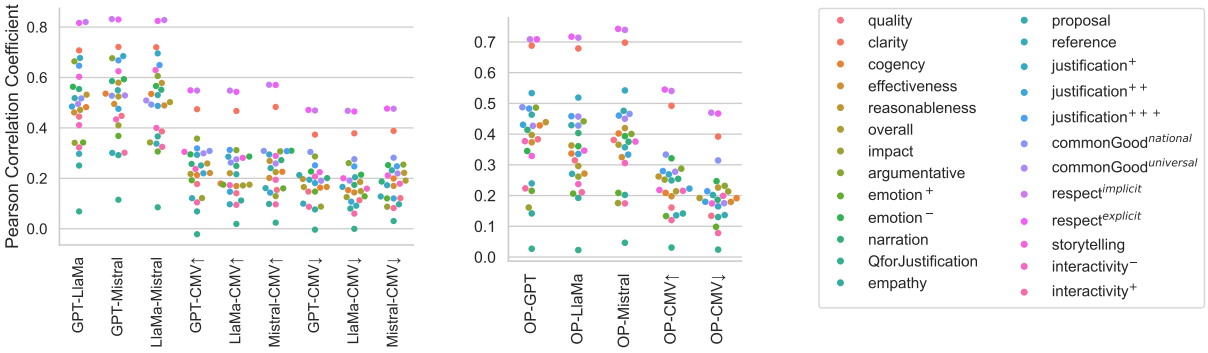
Linguistic Features Table 2 presents the most prominent linguistic features. The most pronounced differences between C^H and C^L are in the *type token ratio (TTR)* and the *lemma token ratio (LTR)*, with 0.20 and 0.19. This is followed by the *average emotion intensity for anticipation* (0.13), *POS variability* (the relative number of unique POS tags per token, 0.12), *Herdan's C* (a measure of lexical richness, 0.11), and the *average emotion intensity for trust* (0.11).

Looking at D, LLM arguments contain more emotion-related features, meaning that LLMs tend to use more tokens that are associated with a higher emotion intensity, as exemplified in the second panel of Figure 2b. For the TTR (Figure 2b, left), humans tend to write arguments with a greater lexical richness (more unique lexical items relative to the number of tokens) and show a higher variance. While all LLMs tend to behave very similarly, the distributional tendencies of arguments from GPT-3.5-turbo are closer to human-written arguments in terms of their TTR than the other LLMs (also see Appendix Tables 12 and 13).

4.2 Correlations between Parallel Arguments

To gain more insights into the correlations between parallel argument comments authored by LLMs and humans, we now report findings separately for the three LLMs. Additionally, to account for the variability among human-written comments, we present results separately for CMV↑ and CMV↓.

Argument Quality Features Figure 3a presents Pearson's correlation scores for argument quality features. At a first glance, we observe high correlation between the LLM arguments (left three bars, avg. $r = 0.51$), among which there is a stronger correlation between the GPT-3.5-turbo and the MISTRAL-7B-instruct arguments, then LLAMA2-7B and MISTRAL-7B-instruct (with also more variation), and finally GPT-3.5-turbo and LLAMA2-7B. Looking at LLM–CMV pairs (right six bars), we see that the parallel human and LLM arguments are



(a) Comment-Comment (b) OP-Comment

Figure 3: Pearson's r per comparison pair for argument quality features: (a) between pairs of counter-arguments, (b) between the original post arguments and counter-arguments (see Appendix Table 6 for all quality features).

less correlated, with a slightly higher correlation between LLMs and CMV↑ than CMV↓ (0.25 and 0.2 on average).

As per individual features, we observe a consistent ranking patterns across comparisons, e.g., *implicit respect* is the top-correlated dimension in all cases except for GPT–CMV↓.

Linguistic features Given the high number of linguistic features, we aggregate results per feature area in Figure 5a (cf. Figure 11). We find that all LLMs show consistently higher correlations with each other than with CMV↑ or CMV↓, especially in *emotion*, *named entities*, and *psycholinguistics* with no major differences between the individual LLMs. We verify that these tendencies hold for individual features by calculating Spearman's ρ between each pair of comparisons' feature correlations (Table 15 in §B.3.1). We find high rank correlations $\rho \in (0.84, 0.98)$, indicating high consistency across comparisons.

Interpretation and Answer to RQ1 Our results partly align with Palmer and Spirling (2023). First, LLM comments express more positive emotions (e.g., anticipation and trust), while human comments display more negative emotions. Although the expression of negative emotions is less common in LLM outputs, since such content is often avoided by the models, as evidenced by our findings, it plays a vital role in authentic human interaction, especially in the context of socially and politically charged topics. Similarly, we verify that human comments show greater creativity, as reflected in metrics like TTR, lemma-token ratio, and Herdan's C.; while, unlike their findings, we see no evidence that LLM-generated texts are less complex. When it comes to argument quality, LLMs excel at replicating textbook indicators of high quality – such as cogency, justification, and overall impact – likely reflecting classifier-aligned optimization. This supports Salvi et al. (2024)'s findings that LLMs use more analytical reasoning markers than humans. However, whether these arguments are actually perceived as more persuasive or impactful by humans remains an open question, beyond the scope of this paper. At the same time, humans continue to display greater creativity and capacity for interactive discourse – asking questions and referencing others in the conversation – traits that are distinctly human.

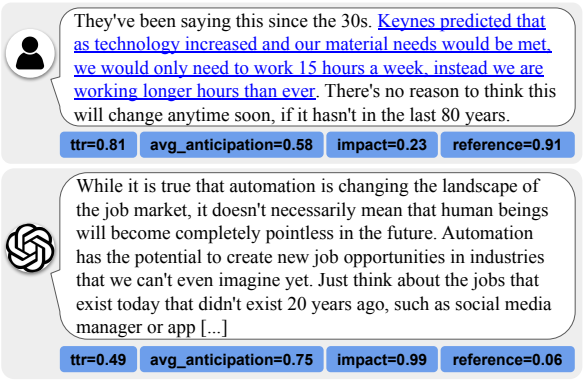
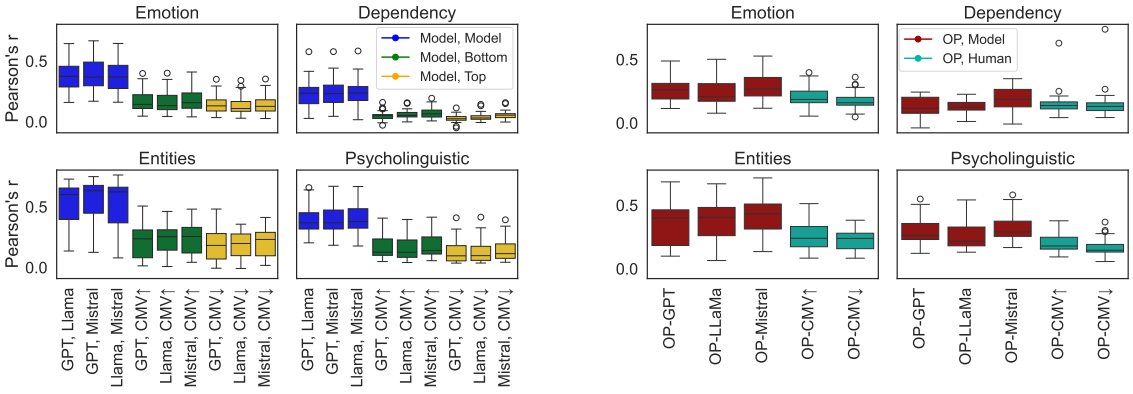


Figure 4: Example parallel texts from a CMV user and GPT-3.5-turbo with feature values. Underlined blue text indicates a hyperlink included in the text (cf. Appendix Figure 8).

These tendencies are illustrated in Figure 4 with a parallel pair of counter-arguments: The human-written argument is shorter yet more lexically rich, as evidenced by its high TTR, and references to other discourse participants (Keynes, they). In contrast, GPT-3.5 produces a more formulaic argument, repeating terms such as *automation*, leading to a lower TTR, and uses words connoted with antici-



(a) Comment-Comment

(b) OP-Comment

Figure 5: Aggregated Pearson's r for each pair of linguistic features per feature area, for the areas with the largest differences: (a) between pairs of counter-arguments, and (b) between original posts and their counter-arguments.

pation such as *potential* or *opportunities*. Lastly, the GPT-3.5 argument scores higher on impact, indicating greater perceived persuasiveness.

4.3 Style and Quality Convergence

Ananthasubramaniam et al. (2023) show that Reddit users often adapt their linguistic style to that of their interlocutors. This naturally raises the question: Do LLMs exhibit similar convergence patterns? To explore this, we turn to **RQ2** and examine whether comments – authored by either humans or LLMs – align with the linguistic style or argument quality of the OPs. To measure the convergence, we use the same method described in §4.2, this time comparing each OP with the corresponding comment, whether written by a human (CMV↑ or CMV↓) or generated by a particular LLM.

Argument Quality Features Figure 3b demonstrates that the LLM arguments are much more correlated with the OP than the human arguments (CMV↑ and CMV↓), with OP–CMV↓ as the least correlated pair. For individual dimensions, the correlations for *empathy* is consistently weak across all comparisons (0.16 on avg.). The correlations for *implicit* and *explicit respect* toward others and *clarity* are the highest across comparisons – strong for the OP–LLM (0.72, 0.72, 0.69, resp.) and moderate for OP–Human (0.51, 0.47, 0.44, resp.). The remaining dimensions also behave consistently across pairs with the Pearson's r being the largest to smallest for OP–LLM, OP-CMV↑ and OP-CMV↓. Lastly, we observe a cluster around the moderate to strong association between the OP–LLM pairs and weak for the OP–Human pairs.

Features	A	P	R	$F1$
Argument Quality	.9808	.9846	.9834	.9840
Linguistic	.9885	.9927	.9881	.9904
All Features	**.9927**	**.9952**	**.9927**	**.9939**

Table 3: Performance of logistic regression over features averaged across 5 folds cross validation based on argument quality, linguistic, and both sets of features.

Linguistic Features Figure 5b shows that, for most feature areas, Pearson's r is in the weak ($0.0 \leq r < 0.3$) to moderate range ($0.3 \leq r < 0.5$) across comparison pairs. The correlations are highest for OP–LLM comparisons for *emotion*, *named entity*, and *psycholinguistics* features. Specifically, LLM arguments show higher correlations with OP than human arguments across feature areas. The highest correlations for OP–CMV comparisons, in contrast, can be found in *lexical richness*, *dependency*, and *POS* features (cf. Appendix Figure 12). Examination of the individual features reveals that human argumentative comments align with OPs particularly well for *uncommon words and constructions*. These tendencies hold for individual features – we find high Spearman's rank correlations ($\rho \in (0.76, 0.95)$) between feature correlations for each pair (cf. Table 16 in §B.3.1).

Interpretation and Answer to RQ2 Overall, LLM-generated comments exhibit stronger convergence with OPs, particularly in the use of named entities, emotion-associated words, and psycholinguistic features. Interestingly, these are precisely the areas where we observe the highest inter-model correlations (cf. §4.2). In contrast, human-written

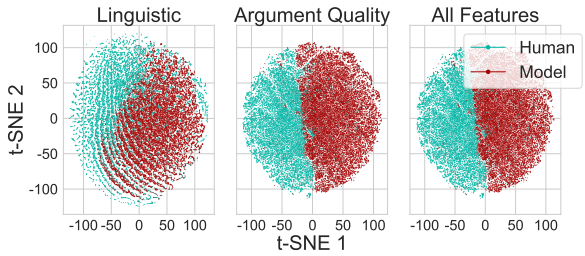


Figure 6: t-SNE projections of linguistic features, argument quality dimensions, and both of them concatenated.

Domain	A	P	R	$F1$
CMV$^{\diamond\bullet}$	0.879	0.809	0.961	0.878
Yelp$^{\diamond}$	0.791	0.875	0.667	0.757
ELI5$^{\bullet}$	0.844	0.815	0.847	0.831
WP$^{\bullet}$	0.940	0.907	0.965	0.935
Avg.	0.809	0.761	0.900	0.811

Table 4: Classification results for Testbed 3 (fixed domain & arbitrary model) from MAGE. Argumentative texts are marked with $\diamond$ and the Reddit data with $\bullet$. *Avg.* reports the average across all datasets in MAGE (cf. Appendix Table 33).

comments show less linguistic convergence to the OPs overall but tend to align more in their use of uncommon vocabulary and syntactic constructions.

The observed patterns suggest that LLMs may rely on default, model-internal strategies for alignment, leading them to converge with the OPs and, thus, with one another. In contrast, human convergence appears more influenced by individual preference and/or community or group norms. The consistent use of uncommon words and syntactic structures likely reflects forum-specific stylistic conventions (Danescu-Niculescu-Mizil et al., 2013; Nguyen and P. Rosé, 2011), rather than direct adaptation to individual posts.

5 Linear Classification over Features

So far, we have shown that LLMs differ from humans in both linguistic style and several dimensions of argument quality. This raises a natural question: Can they be used to successfully recognize generated content (**RQ3**)?

To get an intuition about the data distribution and select an appropriate method, we first projected the features into two-dimensional latent spaces using t-SNE. As displayed in Figure 6, the resulting visualizations revealed well-structured clusters, with clear separation boundaries between the human- and LLM-authored texts in the embedding space, suggesting that the linguistic and argument quality features of these texts capture meaningful distinctions among the underlying categories.

5.1 Method

To answer RQ3, we train and evaluate a linear classifier with interpretable features introduced in §3.2, choosing linear classification since the projections suggested clear separation.

Logistic Regression We use a simple logistic regression classifier to distinguish the two types of parallel texts (C^H and C^L; details in §C.1.1). We represent each argument comment by a vector of its argument quality metrics, linguistic features, and the concatenation of both. This allows for inspection of the most predictive features via the regression coefficients, making the classifier's outputs interpretable.

Evaluation We evaluate the classifier via stratified k-fold cross validation on balanced CMV data (80% train and 20% test at each iteration; 5 folds, stratified for class). For metrics, we follow the common practice and use Precision (P), Recall (R), F_1 (F_1), and Accuracy (A) scores.

5.2 Results

The results in Table 3 demonstrate almost perfect classification performance (cf. Appendix Tables 17 to 19). This is true for using only the argument quality scores, only the linguistic features, or both, while the overall performance is better when both sets of features are used. For the argument quality metrics, the most salient features in distinguishing between the two are *quality*, *clarity*, *positive interactivity*, *cogency*, and *negative interactivity*. The most salient linguistic features are *Maas' TTR* (a log-normalized formulation of TTR), *the number of hapax dislegomena* (tokens occurring exactly twice in the text), *the number of nominal subjects, types, and pronouns* (cf. Appendix Figure 18 and Table 21).

5.3 Generalizability

To test the generalizability of our findings beyond our prompt settings, the social media domain, or text type (e.g., argumentative vs. informative), we evaluate our method on an external benchmark.[5]

[5]Here, we only use linguistic features given the limited applicability of argument quality dimensions to other domains. For results including argument quality dimensions on applica-

Method	R_{C^H}	R_{C^L}	R_{avg}	*AUROC*
FastText	0.894	0.739	0.817	0.89
GLTR	0.373	0.889	0.631	0.80
Longformer	**0.898**	**0.972**	**0.935**	**0.99**
DetectGPT	0.869	0.341	0.605	0.57
LogReg+Features	0.753	0.900	0.827	0.83

Table 5: Comparison of linguistic feature-based logistic regression to methods reported in Li et al. (2024a). Following the original papers' evaluation scheme, we report the human and LLM recall and the average of the two (R_{C^H}, R_{C^L}, and R_{avg}) as well as *AUROC*. Best per metric bolded, second-best underlined.

Benchmark MAGE (Li et al., 2024a) is a benchmark suite for detecting LLM-generated text, where each instance pairs a human-written passage with continuations generated by 27 models (listed in §C.4.2). These continuations are prompted with the first 30 words of the human-written text, using data from diverse online sources (see Appendix Table 25), including CMV. It includes eight testbeds (see Appendix Table 26). We focus on Testbed 3 (fixed domain & arbitrary models), which aligns most closely with our setup, and report additional results in the Appendix (§C.4.3 and §C.4.4). We train a logistic regression classifier on linguistic features extracted from MAGE data (same procedure in Section 5.1) and evaluate using their metrics – *AUROC* (higher is better), recalls for human (R_{C^H}) and LLM (R_{C^L}), and average (R_{avg}) – as well as P, R, F_1, and A.

5.3.1 Results

Table 4 shows classification results for opinion (CMV, Yelp; ⋄) and Reddit (CMV, ELI5, WP; •) subsets. Overall performance is lower than on our CMV dataset (avg. $F_1 = 0.81$; Yelp lowest at 0.76, WP highest at 0.94, CMV at 0.88). Compared to Li et al. (2024a) (on their evaluation scheme; see Table 5), our simple logistic regression over linguistic features performs comparably to more complex, resource-intensive detection methods.

Performance differences – especially for CMV – may stem from prompt design. In MAGE, LLM-generated texts continue from the beginning section of parallel human counter-arguments rather than being conditioned on the OP. This shared opening likely results in stronger stylistic alignment, making the parallel texts harder to distinguish.

ble subsets of the benchmark, see §C.5

6 Conclusions and Discussion

As LLM-generated content increasingly populates online discourse, understanding how it differs from human-written texts is critical, particularly in persuasive contexts with potential influence on opinion shaping.

Analyzing distributional differences between human and LLM arguments in persuasive discourse, we find substantial differences both in style and argument quality: LLM arguments show higher emotional positivity, stronger convergence with original posts (especially in named entities and psycholinguistic features), and greater alignment with argument quality markers. In contrast, human arguments display more negative emotion, greater lexical and syntactic creativity, and stronger use of interactive discourse.

Moreover, we show that linguistic and argument quality features enable nearly 99% accurate detection of LLM-generated comments to CMV posts. Our approach thus offers a practical safeguard against unethical uses of LLMs in online discussions.[6] Furthermore, tests on an external benchmark show that our lightweight[7] and interpretable method performs comparably to computationally intensive detectors in generalized detection scenarios, highlighting the viability of low-resource, transparent detection methods.

These results prompt important questions for future research: Under what conditions are LLM-generated texts harder to detect? How do the prompt design and task objective influence detectability? How do the convergence patterns of humans and LLMs align with social theories of communication, such as communication accommodation theory (Giles, 1973)? Our framework provides a straightforward and interpretable approach to assess such questions, thereby facilitating future investigations into the nuances of LLM-generated content.

7 Limitations

Due to data availability, the present work only considers English arguments. While the general ap-

[6]The dire need for this type of work is underscored by recent incidents of unauthorized attempts to inject LLM-generated content in online forums; notably, the CMV community was shaken by such a case only recently.

[7]Given training data ready with extracted features, the classifier can be trained in a few minutes. Both training and inference, including feature extraction, can be run on consumer-grade hardware (in our case, an Apple M4 Macbook Pro).

proach of analysis and classification in principle could be applied directly to arguments in other languages, the availability of tools for languages other than English may be limited. Furthermore, the feature-specific findings may be language dependent.

In a similar vein, we assess a single, albeit popular, text domain in detail. While our results suggest that the distinction between human- and LLM-authored texts based on interpretable features largely generalizes, the specific features that differentiate the two may vary across domains.

Our analysis and experiment setup is based on a single prompt formulation. Given that the effects of prompt variation on free-form generation are even less understood than in classification tasks, factors such as prompt variations and decoding strategies could influence the variability of LLM-generated arguments and potentially alter how different they are from human arguments in their linguistic and argumentation-specific makeup.

Finally, our approach involves automatically extracted features for argument quality dimensions and relies on psycholinguistic norms and emotion lexicons for certain linguistic features. These signals are noisy proxies, and how well they map to human judgments of argument or deliberative qualities of texts is an open question. Nonetheless, the distributional differences between human-written and LLM-generated arguments on these features allow for meaningful comparisons that shed light on stylistic and qualitative tendencies, offering insights into how LLMs approximate (or diverge from) human argumentative behavior.

8 Ethical Considerations

While our work advances understanding of the features and quality of LLM-generated texts, it raises several important ethical considerations.

First, LLM-generated arguments in response to social media posts have the potential to reproduce or amplify societal biases. To mitigate this risk, all experiments were conducted *offline*, ensuring that no unsuspecting social media users were exposed to machine-generated *"opinions."* We strongly discourage any deployment of LLMs in online environments that violates community norms or lacks informed consent.

Second, insights from our findings raise potential dual-use concerns. On one hand, they could be misused by malicious actors to develop tools that manipulate or steer public discourse. It is therefore essential not only to highlight these risks but also to support community efforts to identify, mitigate, and safeguard against such misuse. On the other hand, understanding the linguistic properties of generated arguments has constructive applications beyond detecting malicious content. These insights can help guide generation toward more appropriate content (Ziegenbein et al., 2024) – for example, by suggesting reformulations of user arguments to match a desired stylistic profile. Additionally, classifiers based on fully interpretable linguistic features can serve as effective educational tools, helping the public recognize LLM-generated content.

Finally, all human-authored texts used in our analyses were drawn from publicly available datasets and handled in accordance with established ethical research standards. No identifiable or private user data was used. Nonetheless, ongoing reflection on issues of consent, data provenance, and user agency remains vital when working with human discourse.

9 Acknowledgements

We acknowledge the support of the Ministerium für Wissenschaft, Forschung und Kunst BadenWürttemberg (MWK, Ministry of Science, Research and the Arts Baden-Württemberg under Az. 33-7533-9 19/54/5) in Künstliche Intelligenz & Gesellschaft: Reflecting Intelligent Systems for Diversity, Demography and Democracy (IRIS3D) and the support by the Interchange Forum for Reflecting on Intelligent Systems (IRIS) at the University of Stuttgart.

We would like to thank the TCL research group at the IMS and the DSM team at GESIS for their valuable feedback throughout the course of this project and on earlier versions of the manuscript.

References

Mohamed Hesham Ibrahim Abdalla, Simon Malberg, Daryna Dementieva, Edoardo Mosca, and Georg Groh. 2023. A Benchmark Dataset to Distinguish Human-Written and Machine-Generated Scientific Papers. *Information*, 14(10).

Milad Alshomary, Shahbaz Syed, Arkajit Dhar, Martin Potthast, and Henning Wachsmuth. 2021. Counterargument generation by attacking weak premises. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1816–1827, Online. Association for Computational Linguistics.

Milad Alshomary and Henning Wachsmuth. 2023. Conclusion-based counter-argument generation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 957–967, Dubrovnik, Croatia. Association for Computational Linguistics.

Aparna Ananthasubramaniam, Hong Chen, Jason Yan, Kenan Alkiek, Jiaxin Pei, Agrima Seth, Lavinia Dunagan, Minje Choi, Benjamin Litterer, and David Jurgens. 2023. Exploring linguistic style matching in online communities: The role of social context and conversation dynamics. In *Proceedings of the First Workshop on Social Influence in Conversations (SICon 2023)*, pages 64–74, Toronto, Canada. Association for Computational Linguistics.

H. S. Sichel and. 1975. On a distribution law for word frequencies. *Journal of the American Statistical Association*, 70(351a):542–547.

Jonathan Anderson. 1981. Analysing the Readability of English and Non-English Texts in the Classroom with Lix. *Seventh Australian Reading Association Conference*, pages 1–13.

Hui Bai, Jan G Voelkel, johannes C Eichstaedt, and Robb Willer. 2023. Artificial intelligence can persuade humans on political issues.

Carl Hugo Björnsson. 1968. *Läsbarhet*. Pedagogiskt Utvecklingsarbete vid Stockholms Skolor. 6. Liber.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Marc Brysbaert, Paweł Mandera, Samantha F McCormick, and Emmanuel Keuleers. 2019. Word prevalence norms for 62,000 English lemmas. *Behavior Research Methods*, 51:467–479.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, 46:904–911.

John B. Carroll. 1964. Language and thought. *Reading Improvement*, 2(1):80.

Hong Chen, Hiroya Takamura, and Hideki Nakayama. 2021. SciXGen: A scientific paper dataset for context-aware text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1483–1492, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *Preprint*, arXiv:2210.11416.

Meri Coleman and Ta Lin Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283.

Michael A. Covington and Joe D. McFall and. 2010. Cutting the Gordian Knot: The Moving-Average Type–Token Ratio (MATTR). *Journal of Quantitative Linguistics*, 17(2):94–100.

Cristian Danescu-Niculescu-Mizil, Robert West, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. No country for old members: user lifecycle and linguistic change in online communities. In *Proceedings of the 22nd International Conference on World Wide Web*, WWW '13, page 307–318, New York, NY, USA. Association for Computing Machinery.

Veronica Diveica, Penny M. Pexman, and Richard J. Binney. 2023. Quantifying social semantics: An inclusive definition of socialness and ratings for 8388 English words. *Behavior Research Methods*, 55(2):461–473.

Esra Dönmez and Agnieszka Falenska. 2025. "I understand your perspective": LLM persuasion through the lens of communicative action theory. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 15312–15327, Vienna, Austria. Association for Computational Linguistics.

Jad Doughman, Osama Mohammed Afzal, Hawau Olamide Toyin, Shady Shehata, Preslav Nakov, and Zeerak Talat. 2025. Exploring the limitations of detecting machine-generated text. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4274–4281, Abu Dhabi, UAE. Association for Computational Linguistics.

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. *AAAI Conference on Artificial Intelligence*.

Agnieszka Falenska, Eva Maria Vecchi, and Gabriella Lapesa. 2024. Self-reported demographics and discourse dynamics in a persuasive online forum. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14606–14621, Torino, Italia. ELRA and ICCL.

Neele Falk and Gabriella Lapesa. 2023. Bridging argument quality and deliberative quality annotations with adapters. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2469–2488, Dubrovnik, Croatia. Association for Computational Linguistics.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567, Florence, Italy. Association for Computational Linguistics.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Jessica Gasiorek and Laszlo Vincze. 2016. Modeling motives for bilingual accommodation by minority and majority language speakers. *Journal of language and social psychology*, 35(3):305–316.

Howard Giles. 1973. Accent Mobility: A Model and Some Data. *Anthropological Linguistics*, 15(2):87–105.

Howard Giles and Tania Ogay. 2013. Communication accommodation theory. In *Explaining communication*, pages 325–344. Routledge.

Josh A Goldstein, Jason Chao, Shelby Grossman, Alex Stamos, and Michael Tomz. 2024. How persuasive is ai-generated propaganda? *PNAS Nexus*, 3(2):pgae034.

Pierre Guiraud. 1954. *Les caractères statistiques du vocabulaire: essai de méthodologie*. Presses universitaires de France.

Xinlei He, Xinyue Shen, Zeyuan Chen, Michael Backes, and Yang Zhang. 2024. MGTBench: Benchmarking Machine-Generated Text Detection. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, New York, NY, USA. Association for Computing Machinery.

Gustav Herdan. 1955. A new derivation and interpretation of Yule's 'Characteristic' K. *Zeitschrift für angewandte Mathematik und Physik ZAMP*, 6:332–339.

Gustav Herdan. 1964. *Quantitative Linguistics*. Butterworths.

Christopher Hidey, Elena Musi, Alyssa Hwang, Smaranda Muresan, and Kathy McKeown. 2017. Analyzing the semantic types of claims and premises in an online persuasive forum. In *Proceedings of the 4th Workshop on Argument Mining*, pages 11–21, Copenhagen, Denmark. Association for Computational Linguistics.

Xinyu Hua and Lu Wang. 2019. Sentence-level content planning and style specification for neural text generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 591–602, Hong Kong, China. Association for Computational Linguistics.

Shagun Jhaver, Pranil Vora, and Amy Bruckman. 2017. Designing for civil conversations: Lessons learned from ChangeMyView. Technical report, Georgia Institute of Technology.

J. Peter Kincaid, Robert P. Jr. Fishburne, Richard L. Rogers, and Brad S. Chissom. 1975. Derivation Of New Readability Formulas (Automated Readability Index, Fog Count And Flesch Reading Ease Formula) For Navy Enlisted Personnel. Technical report, Institute for Simulation and Training.

Ryuto Koike, Masahiro Kaneko, and Naoaki Okazaki. 2024. Outfox: Llm-generated essay detection through in-context learning with adversarially generated examples. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38(19), pages 21258–21266.

Victor Kuperman, Hans Stadthagen-Gonzalez, and Marc Brysbaert. 2012. Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, 44:978–990.

Yafu Li, Qintong Li, Leyang Cui, Wei Bi, Zhilin Wang, Longyue Wang, Linyi Yang, Shuming Shi, and Yue Zhang. 2024a. MAGE: Machine-generated text detection in the wild. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 36–53, Bangkok, Thailand. Association for Computational Linguistics.

Yafu Li, Zhilin Wang, Leyang Cui, Wei Bi, Shuming Shi, and Yue Zhang. 2024b. Spotting AI's touch: Identifying LLM-paraphrased spans in text. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7088–7107, Bangkok, Thailand. Association for Computational Linguistics.

Jiayu Lin, Rong Ye, Meng Han, Qi Zhang, Ruofei Lai, Xinyu Zhang, Zhao Cao, Xuanjing Huang, and Zhongyu Wei. 2023. Argue with me tersely: Towards sentence-level counter-argument generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16705–16720, Singapore. Association for Computational Linguistics.

Dermot Lynott, Louise Connell, Marc Brysbaert, James Brand, and James Carney. 2020. The Lancaster Sensorimotor Norms: multidimensional measures of perceptual and action strength for 40,000 English words. *Behavior Research Methods*, 52:1271–1291.

Heinz-Dieter Mass. 1972. Über den Zusammenhang zwischen Wortschatzumfang und Länge eines Textes.

Zeitschrift für Literaturwissenschaft und Linguistik, 2(8):73.

Maximilian Maurer. 2025. ELFEN - Efficient Linguistic Feature Extraction for Natural Language Datasets.

Philip M. McCarthy and Scott Jarvis. 2010. MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, 42(2):381–392.

Saif Mohammad. 2018a. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–184, Melbourne, Australia. Association for Computational Linguistics.

Saif Mohammad. 2018b. Word affect intensities. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Saif Mohammad and Peter Turney. 2010. Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 26–34, Los Angeles, CA. Association for Computational Linguistics.

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

Corrado Monti, Luca Maria Aiello, Gianmarco De Francisci Morales, and Francesco Bonchi. 2022. The language of opinion change on social media under the lens of communicative action. *Scientific Reports*, 12(1):1–11.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California. Association for Computational Linguistics.

Alberto Muñoz-Ortiz, Carlos Gómez-Rodríguez, and David Vilares. 2024. Contrasting Linguistic Patterns in Human and LLM-Generated News Text. *Artificial Intelligence Review*, 57.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Dong Nguyen and Carolyn P. Rosé. 2011. Language use as a reflection of socialization in online communities. In *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, pages 76–85, Portland, Oregon. Association for Computational Linguistics.

Alexis Palmer and Arthur Spirling. 2023. Large Language Models Can Argue in Convincing Ways About Politics, But Humans Dislike AI Authors: implications for Governance. *Political Science*, 75(3):281–291.

Yujin Potter, Shiyang Lai, Junsol Kim, James Evans, and Dawn Song. 2024. Hidden persuaders: Llms' political leaning and their influence on voters. *Proc. of EMNLP*.

Jiameng Pu, Zain Sarwar, Sifat Muhammad Abdullah, Abdullah Rehman, Yoonjin Kim, Parantapa Bhattacharya, Mobin Javed, and Bimal Viswanath. 2023. Deepfake text detection: Limitations and opportunities. In *2023 IEEE Symposium on Security and Privacy (SP)*, pages 1613–1630.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Alex Reinhart, Ben Markey, Michael Laudenbach, Kachatad Pantusen, Ronald Yurko, Gordon Weinberg, and David West Brown. 2025. Do LLMs write like humans? Variation in grammatical and rhetorical styles. *Proceedings of the National Academy of Sciences*, 122(8):e2422455122.

Brian J. Richards and David D. Malvern. 1997. *Quantifying lexical diversity in the study of language development*. University of Reading, Faculty of Education and Community Studies.

Francesco Salvi, Manoel Horta Ribeiro, Riccardo Gallotti, and Robert West. 2024. On the Conversational Persuasiveness of Large Language Models: A Randomized Controlled Trial. *Preprint*, arXiv:2403.14380.

Morgan Sandler, Hyesun Choung, Arun Ross, and Prabu David. 2025. A linguistic comparison between human and chatgpt-generated conversations. In *Pattern Recognition and Artificial Intelligence*, pages 366–380, Singapore. Springer Nature Singapore.

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan

Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M Rush. 2022. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*.

Spencer M. Seals and Valerie Shalin. 2023. Long-form analogies generated by chatGPT lack human-like psycholinguistic properties. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 45.

Edward H. Simpson. 1949. Measurement of Diversity. *Nature*, 163.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, page 613–624, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Ruixiang Tang, Yu-Neng Chuang, and Xia Hu. 2024. The Science of Detecting LLM-Generated Text. *Commun. ACM*, 67(4):50–59.

MILDRED C. TEMPLIN. 1957. *"Certain Language Skills in Children: Their Development and Interrelationships"*, ned - new edition edition, volume 26. University of Minnesota Press.

Michael Henry Tessler, Michiel A. Bakker, Daniel Jarrett, Hannah Sheahan, Martin J. Chadwick, Raphael Koster, Georgina Evans, Lucy Campbell-Gillingham, Tantum Collins, David C. Parkes, Matthew Botvinick, and Christopher Summerfield. 2024. AI can help humans find common ground in democratic deliberation. *Science*, 386(6719):eadq2852.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *Preprint*, arXiv:2307.09288.

Ben Wang. 2021. Mesh-Transformer-JAX: Model-Parallel Implementation of Transformer Language Model with JAX. https://github.com/kingoflolz/mesh-transformer-jax.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Chenxi Whitehouse, Osama Mohammed Afzal, Tarek Mahmoud, Toru Sasaki, Thomas Arnold, Alham Fikri Aji, Nizar Habash, Iryna Gurevych, and Preslav Nakov. 2024. M4: Multi-generator, multi-domain, and multilingual black-box machine-generated text detection. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1369–1407, St. Julian's, Malta. Association for Computational Linguistics.

Bodo Winter, Gary Lupyan, Lynn K Perry, Mark Dingemanse, and Marcus Perlman. 2024. Iconicity ratings for 14,000+ English words. *Behavior Research Methods*, 56(3):1640–1655.

George U. Yule. 1944. *The statistical study of literary vocabulary*. Cambridge University Press.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Peng Zhang, Yuxiao Dong, and Jie Tang. 2023. Glm-130b: An open bilingual pre-trained model. *Preprint*, arXiv:2210.02414.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. Opt: Open pre-trained transformer language models. *Preprint*, arXiv:2205.01068.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level Convolutional Networks for Text Classification. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Timon Ziegenbein, Gabriella Skitalinskaya, Alireza Bayat Makou, and Henning Wachsmuth. 2024. LLM-based rewriting of inappropriate argumentation using reinforcement learning from machine feedback. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4455–4476, Bangkok, Thailand. Association for Computational Linguistics.

A Data

We provide additional details of the CMV forum and the LLM-generated counter-arguments in this section.

A.1 CMV Forum

The subreddit CMV functions similarly to other Reddit forums. Users initiate discussions by posting a viewpoint in the title and elaborating on it in the body text, which may include external links to websites, images, or other resources. The platform is designed for opinion exchange: once a post is live, any Reddit user can engage with it by voting or commenting (a Reddit account is required to do so). Users can also reply to comments, forming a tree-structured discussion. A distinctive feature of CMV is the ability for the original poster to award a Δ flag to comments that successfully changed their view.

As with other Reddit communities, CMV enforces a set of rules, outlined at www.reddit.com/r/changemyview/wiki/rules. Notably, one rule prohibits unauthorized bot activity. However, this rule was recently violated when researchers conducted an unauthorized experiment by posting LLM-generated comments, as reported in www.reddit.com/r/changemyview/comments/1k8b2hj/meta_unauthorized_experiment_on_cmv_involving. Incidents like this highlight the urgent need to better understand the characteristics of LLM-generated arguments and their potential influence on public opinion, as well as the importance of developing interpretable tools to detect such content.

Note that, due to the nature of online discourse, the CMV dataset may contain offensive content. However, this content was not created by us, nor do we employ any methods that endorse or promote such material.

A.2 LLM Data

To generate counter-arguments to CMV posts, we use three widely studied LLMs: GPT-3.5-turbo, LLAMA2-7B, and MISTRAL-7B-instruct. These models were chosen for their differing accessibility, diverse training paradigms, and relatively low computational cost. While many other LLM families exist, these three are among the most researched, allowing for deeper analysis of their behavior and patterns.

No prompt engineering or optimization were employed in this study. For LLAMA2-7B and MISTRAL-7B-instruct, we use the HuggingFace text generation pipeline at https://huggingface.co/docs/text-generation-inference, and we access GPT-3.5-turbo via OpenAI text completion API at https://platform.openai.com/docs/guides/gpt (accessed between Dec. 8 - 19. 2024).

B Feature Analysis

This section provides additional details and results for our distribution and correlation analyses.

B.1 Features

In the following, we provide details about the features used in our work. §B.1.1 lays out and describes argument quality and deliberative quality dimensions, and §B.1.2 describes which linguistic features are used.

B.1.1 Argument Quality

We use argument quality assessments from adapter models developed by Falk and Lapesa (2023). Table 6 gives an overview of the dimensions and their sources, Table 7 of the label categories for multi-class adapters. Notation mapping used in figures and captions is described in Table 8.

B.1.2 Linguistic Features

We extract linguistic features using `elfen` at https://elfen.readthedocs.io/en/latest/. In the following, we lay out which features from the package we use in this work.

Surface-Level Features provide information about high-level surface characteristics of the text such as length (in different measurement levels). As such, they measure surface-level complexity of texts.

We extract the sequence length in characters, both with and without whitespace, number of tokens, sentences, types, lemmas, long words (over six characters), the number of tokens per sentence, characters per sentence, and average word length.

Readability Indices measure the reading complexity of the text, giving information on how challenging/fitting a text would be for different reader experience levels.

We extract the Gunning fog index, ARI, Flesch reading ease, and Flesch-Kincaid grade level (Kincaid et al., 1975), the CLI (Coleman and Liau,

Dimension	Description	Corpus
Overall	General argument quality	GAQ
Cogency	Acceptable and sufficient premises to draw a conclusion	GAQ
Reasonableness	Contribution to resolving issues; accepted by a universal audience	GAQ
Effectiveness	Persuasiveness, rhetorical or emotional appeal	GAQ
Quality	General argument quality	IBM-Rank-30k
Clarity	Ease of interpreting the argument	Swanson
Justification	Rationality, providing reasons, reflection	Europolis
Respect	Empathy or respect toward groups (e.g., immigrants)	Europolis
Storytelling	Personal experience or subjective event description	Europolis
Interactivity	Respect toward or reference to other participants' arguments	Europolis
Common Good	Consideration of community interests or utilitarian values	Europolis
PosEmotion	Contains positive emotions	THF/BK
Proposal	Statement about what or how something should be done	THF/BK
Narration	Subjective or personal experience description	THF/BK
Reference	Reference to another discourse participant	THF/BK
Argument	Provides reasons or evidence for/against a claim	THF/BK
NegEmotion	Contains negative emotions	THF/BK
Empathy	Speaker adopts another's perspective or emotional state	THF/BK
Q. for Justification	Requests reasons for a statement or action	THF/BK
Impact	User engagement (e.g., likes or recommendations)	Kialo

Table 6: Argument quality dimensions and their respective score ranges in models from Falk and Lapesa (2023). Justification, respect, interactivity and common good are multi-class, whereas the remaining dimensions are binary. For the binary dimensions, we use the score from node that corresponds to label 1. For the multi-class ones, see Table 7.

Dimension	Labels
Interactivity	No reference, Neutral reference, Positive reference to others*, Negative reference*
Respect	Disrespectful, Implicit respect*, Explicit respect*
Common Good	No reference, Own country*, Common good*
Justification	No justification, Inferior justification*, Qualified justification*, Sophisticated*
Impact	Not impactful, Medium impactful, Impactful*

Table 7: Label categories for multi-class adapters. The ones used in this paper are marked with a *.

Notations	Descriptions
reference^{+}	Positive reference to mentioned groups
reference^{-}	Negative reference to mentioned groups
$\text{respect}^{\text{implicit}}$	Implicit respect for mentioned groups
$\text{respect}^{\text{explicit}}$	Explicit respect for mentioned groups
$\text{commonGood}^{\text{national}}$	Reference to common good for one's own country
$\text{commonGood}^{\text{universal}}$	Reference to common good
justification^{+}	Inferior justification
$\text{justification}^{++}$	Qualified justification
$\text{justification}^{+++}$	Sophisticated justification
impact	Impactful

Table 8: Notation mapping to the argument quality features used in Figures and Tables throughout the paper.

1975), LIX (Björnsson, 1968), and RIX (Anderson, 1981), the number of syllables in a text, the number of words with only one syllable, and the number of words with more than two syllables.

Psycholinguistic Norm Features measure individual words' cognitive, social, and sensorimotor grounding. As such, they provide information on whether and to what extent words in the text evoke associations in different dimensions of lived experiences.

We use the packages' provided psycholinguistic norms (concreteness (Brysbaert et al., 2014), iconicity (Winter et al., 2024), sensorimotor(Lynott et al., 2020), age-of-acquisition (Kuperman et al., 2012), prevalence (Brysbaert et al., 2019), and socialness (Diveica et al., 2023)). Per norm, we extract the average rating of all tokens from the item in the norm lexicon, their average standard deviation in the ratings, the number of tokens with a high rating, a low rating, and the number of tokens with a particularly high standard deviation.

Part-of-Speech Features provide information on the grammatical categories of words in the text instances. As such, they allow us to measure, for example, whether text instances are particularly noun-heavy.

We extract the number of tokens with a given POS tag, the number of lexical tokens (nouns, verbs, adjectives, and adverbs), and the POS variability (number of different POS tags relative to the number of tokens).

Lexical Richness Measures measure the lexical complexity of texts, i.e. how many *different* tokens are used. The more lexically rich a text is, the more variable and, in some sense, creative the word usage in a given text is.

We extract the type-token ratio (TTR) (TEMPLIN, 1957), RTTR (Guiraud, 1954), CTTR (Carroll, 1964), Herdan's C (Herdan, 1964), Summer's TTR, Maas' TTR (Mass, 1972), Yule's K (Yule, 1944), Herdan's V_m (Herdan, 1955), Simpson's D (Simpson, 1949), MSTTR (Richards and Malvern, 1997), MATTR (Covington and and, 2010), MLTD (McCarthy and Jarvis, 2010), the number of hapax (dis)legomena, Sichel's S (and, 1975), and the lexical density.

Morphological Features provide morphosyntactic information about the texts and as such can be reflective of certain styles. We extract the number of tokens with a given morphological feature for all available morpho-syntactic features.

Information-Theoretic Features measure text complexity from an information-theoretic perspective. We extract the compressibility and the average token entropy.

Dependency Features provide information about the syntactic dependency structure of texts. As such, they allow us to measure differences in how texts are structured. We extract the number of dependency relation types, the number of noun chunks in the text, the dependency tree width, the tree depth, the tree branching factor, and the ramification factor.

Semantic Features present in the package measure hedges (whether an author expresses uncertainty), and polysemy as a proxy for lexical ambiguity. We extract the average size of the synsets, the number of tokens with a large synset, and the number of tokens with a small synset for nouns, adjectives, and verbs, and the number of hedges.

Named Entity Features provide information about the usage of entities (people, organizations, etc.) in texts and thus serve as a proxy to assess when texts may refer to world knowledge and to what extent. We extract the number of named entities overall and per entity type.

Emotion and Sentiment Features measure emotion and sentiment associations of individual words. As such, they serve as a proxy for emotion-laden language. Note that this is not equivalent to the overall emotion or sentiment of a given text.

We use the standard emotion/sentiment lexicons as the package: The NRC-VAD lexicon (Mohammad, 2018a) for valence, arousal, and dominance, the NRC emotion intensity lexicon (Mohammad, 2018b) for the emotion intensity, and the NRC word-emotion association lexicon (Mohammad and Turney, 2010, 2013) for sentiment. We extract the average rating, the number of tokens with a high rating, and the number of tokens with a low rating. For sentiment, we extract the number of positive and negative sentiments, and the sentiment score.

B.2 Methods

B.2.1 Metrics

Wasserstein Distance (WS) We use the Wasserstein-1 distance – SciPy implementation, documented at `https://docs.scipy.org/doc/`

scipy/reference/generated/scipy.stats.wasserstein_distance.html – to measure differences between empirical distributions per argument quality dimension and linguistic feature.

Given sets of a given feature for human-written (C^H) with samples X_i and LLM-generated comments (C^L) with samples Y_i for $i = 1$ to N, the Wasserstein-1 distance is defined as

$$WS(C^H, C^L) = \frac{1}{n}\sum_{i=0}^{n} ||X_{(i)} - Y_{(i)}||. \quad (1)$$

Difference of Means (D) To compare the central tendencies of C^H and C^L across each argument quality dimension and linguistic feature, we compute the difference between their means as in

$$D(C^H, C^L) = \mu_H - \mu_L, \quad (2)$$

where μ_H stands for the mean of the human and μ_L of the LLM comments.

Pearson's r We measure the pairwise Pearson's correlation coefficient (Pearson's r) – SciPy implementation, documented in https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.pearsonr.html – for a pair of text sources ($r_{C_iC_j}$ where $i \neq j$) per argument quality dimension and linguistic feature as

$$r = \frac{\sum_{k=1}^{n}(C_{ik} - \bar{C}_i)(C_{jk} - \bar{C}_j)}{\sqrt{\sum_{k=1}^{n}(C_{ik} - \bar{C}_i)^2}\sqrt{\sum_{k=1}^{n}(C_{jk} - \bar{C}_j)^2}}. \quad (3)$$

where $\bar{C}_i$ and $\bar{C}_j$ refer to the means of C_i and C_j, respectively, and C_{ik} is the k-th instance in C_i.

Spearman's ρ We measure Spearman's rank correlation coefficient ρ – SciPy implementation, documented in https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.spearmanr.html – for two feature rankings (per comparison pair, cf. §B) A and B as

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}, \quad (4)$$

where d_i is the difference in ranks of each feature and n is the total number of features.

B.3 Results

In this subsection, we present more detailed analysis results.

B.3.1 Distributional Differences

Table 9 shows the Wasserstein distances and the difference of means for all argument quality dimensions for human-written (only the most and least upvoted) and model-generated comments. Table 10, in contrast, shows the Wasserstein distances and the difference of means for all argument quality dimensions for all first-level (direct reply to OP) human-written comments and model-generated comments. The consistent WS and D confirm marked differences between model-generated and human-written comments, even if the latter is sub-sampled.

Interpretation and Answer to RQ1 In addition to the examples shown in Figure 4, we illustrate the tendencies for the argument quality feature *reasonableness* in Figure 8. The figure presents three texts with distinctly different *reasonableness* scores. The first text makes a sweeping claim that "all jobs will become obsolete," the second expresses a personal perspective while carefully hedging its argument, and the third resorts to sarcasm. These differences are reflected in the scores: the second text achieves the highest score, followed by the first text, while the sarcastic comment ranks lowest.

B.3.2 Verification of Rankings

We verify whether the ranks of features in our comment source pair correlation analyses are comparable by calculating Spearman's ρ per combination of pairs. Intuitively, this gives us an idea of whether a given feature is at a similar rank, so whether, for example, a drop in correlation comparing one pair of comment sources to another in aggregate tendentially holds for individual features, too. This is indeed the case, as Table 15 and 16 show, given the high correlations for all combinations of pairs.

B.3.3 Correlation Analysis Including Delta Comments

In the following, we present results for our correlation analyses. We take all posts and comments with exactly one delta to retain parallelism between comments.

We present the correlation ranges for pairs of comment sources for argument quality features in Figure 13. While features of CMVΔ correlates better with LLM-generated comments than CMV$\uparrow$ or CMV$\downarrow$ across features, they overall do not show drastic differences from them. As for the linguistic features in Figure 14, while CMVΔ has tendentially higher correlations to LLM-generated comments than CMV$\uparrow$ or CMV$\downarrow$ across feature areas,

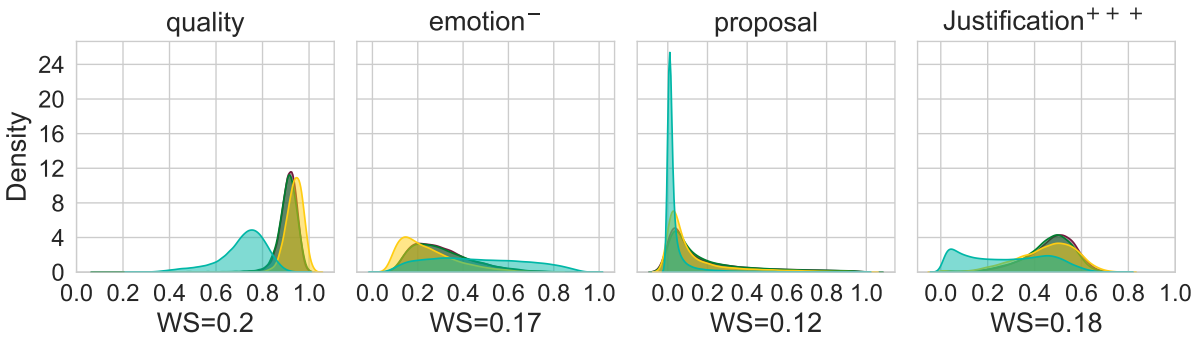


Figure 7: Example distributions from the top argument quality dimensions features with WS in range of $(0.1, 0.2)$ between the human-written and LLM-generated comments.

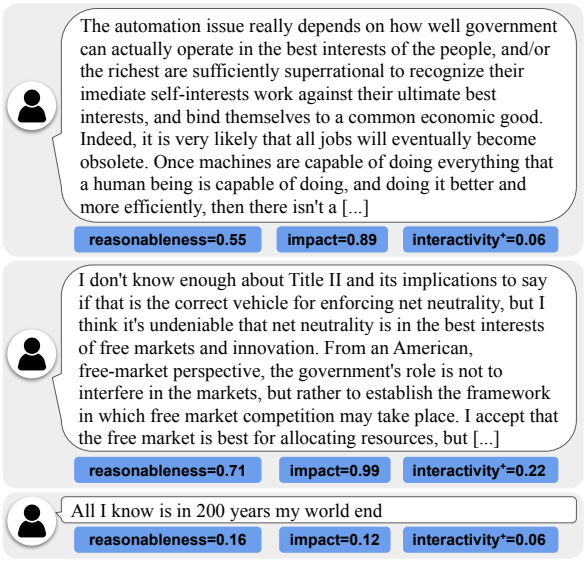


Figure 8: Example parallel texts from three CMV users with feature values.

similar to argument quality features, they overall do not show drastic differences from them.

Now, we present the correlation ranges for pairs of comment sources with the OP for argument quality features in Figure 15. CMVΔ features show similar patterns to that of CMV$\uparrow$ with a slightly higher correlations to OP features. As for the linguistic features in Figure 16, pairs involving CMVΔ behave very similarly to other human comments, especially CMV$\uparrow$.

Overall, we do not find drastic differences between CMVΔ and other human comment sets relative to LLM-generated comments and the OP.

C Classification Analysis

This section contains additional details for the classification analysis on the methods used, the evaluation, and visualization, as well as additional results. §C.2 presents the full results of the classification experiments reported in §5.3.1, and §C.3 presents additional experiments including all human comments instead of select samples. §C.4 gives more details on the MAGE benchmark, and provides our results compared to other methods, and our full results per test-bench.

C.1 Details on the Methods

C.1.1 Logistic Regression

We use a simple logistic regression classifier – scikit-learn implementation, documented at https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html – to distinguish the two types of parallel texts (human-written and LLM-generated). We represent each argument by a vector of its argument quality dimensions, linguistic features, and the concatenation of both.

Evaluation To evaluate the classifier, we perform stratified k-fold cross validation[8] on our CMV dataset.

Visualization To visualize the high-dimensional data distribution, we perform t-distributed stochastic neighbor embedding (t-SNE) analyses. We use two components, a random state of 42, a perplexity of 5, and a maximum number of iterations of $1,000$.

We represent each argument by a vector of its argument quality dimensions, linguistic features, and the concatenation of both. This allows us to inspect the separation between human-written and LLM-generated arguments in the argument quality dimension, the linguistic, and a joint feature space.

[8] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedKFold.html

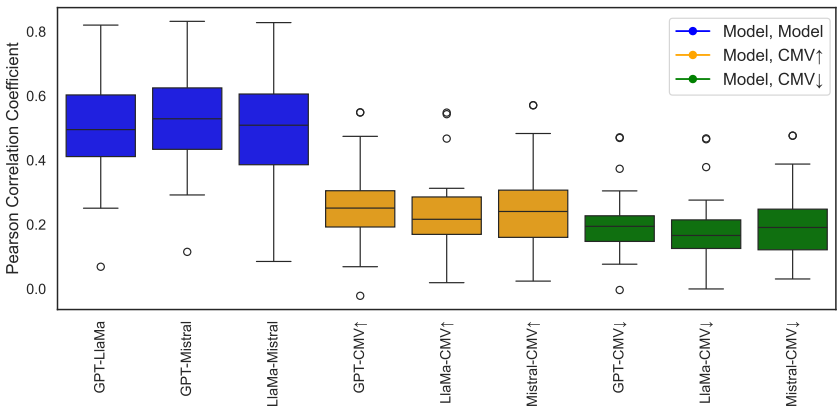


Figure 9: Argument Quality Dimensions

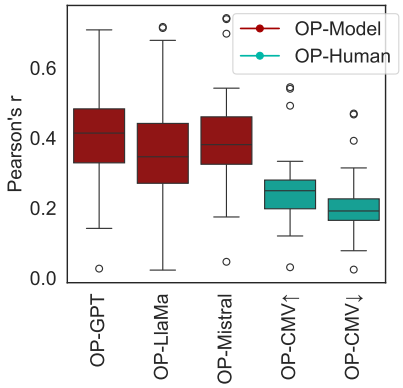


Figure 10: Pearson's r per comparison pair for argument quality features between pairs of human-written original posts (OP) and counter-argument comments (Model or Human).

To gain intuitive insights into the data, we first visualize the high-dimensional features extracted from the data points from each set (human-written and LLM-generated arguments) by performing a t-SNE analysis.

Figure 6 reveals that the clusters form from each set occupy almost a perfect non-overlapping embedding space. This is true when using only the linguistic features, only the argument quality scores, and both. Inspired by this, we fit a logistic regression model on these features, the results of which we discuss in the following.

C.2 Full Classification Results

In the following, we present the full results of our in-domain classification experiments. As described in §5.3.1, we have three settings: (a) argument quality dimensions, (b) linguistic features, and (c) all features.

Argument Quality Dimensions Table 17 shows the results for the classification of LLM-generated arguments using logistic regression on argument quality dimensions across 5 folds, and on average. We find almost perfect performance (> 0.98) across folds and metrics.

Linguistic Features Table 18 shows the results for the classification of LLM-generated arguments using logistic regression on linguistic features across 5 folds, and on average. We find almost perfect performance (> 0.98) across folds and metrics.

All Features Table 19 shows the results for the classification of LLM-generated arguments using logistic regression on all features (argument quality and linguistic features concatenated) across 5 folds, and on average. We find almost perfect performance (> 0.99) across folds and metrics.

Regression Coefficients Figure 18 shows the coefficients of logistic regression for argument quality dimensions, revealing differences between *quality* and *clarity* in effect magnitude compared to other dimensions. Table 20 presents the coefficients per fold and on average for argument quality dimensions, showing that there are drastic differences in effect magnitude.

In comparison, linguistic features behave much more similarly to one another, as shown in Table 21. While some features, in particular Maas' TTR, have a higher effect magnitude than other features, the differences are much lower than for argument quality dimensions.

Sanity Check We perform a keyword search using *Language Model*, *LLM*, *Language Assistant*, *AI Assistant* (case-independent) to observe the fre-

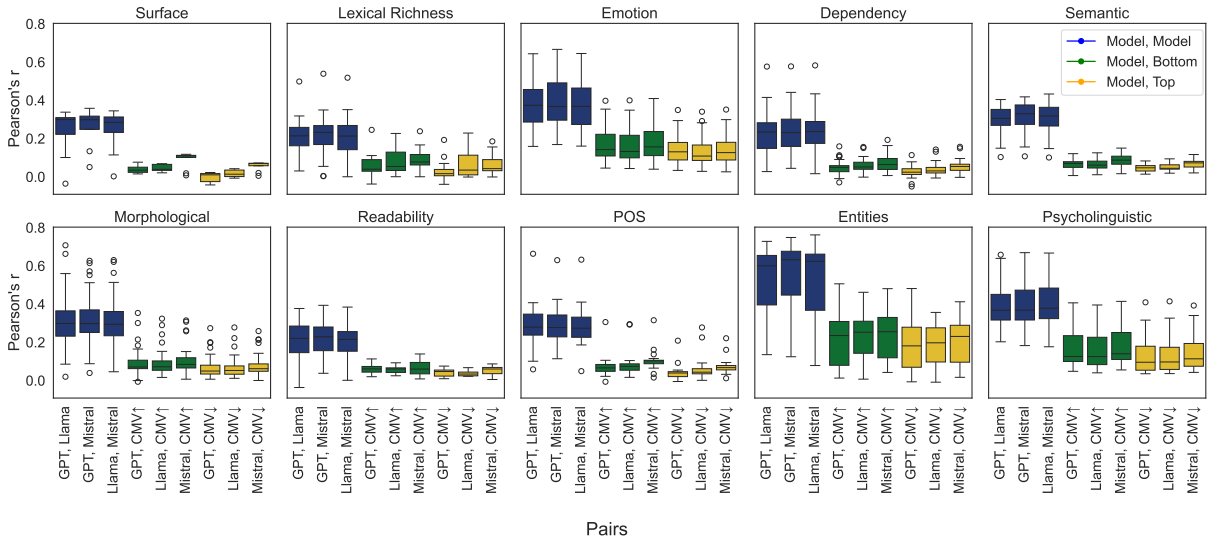


Figure 11: Aggregated Pearson's r for each pair of linguistic features per feature area, for the areas with the largest differences between pairs of counter-arguments.

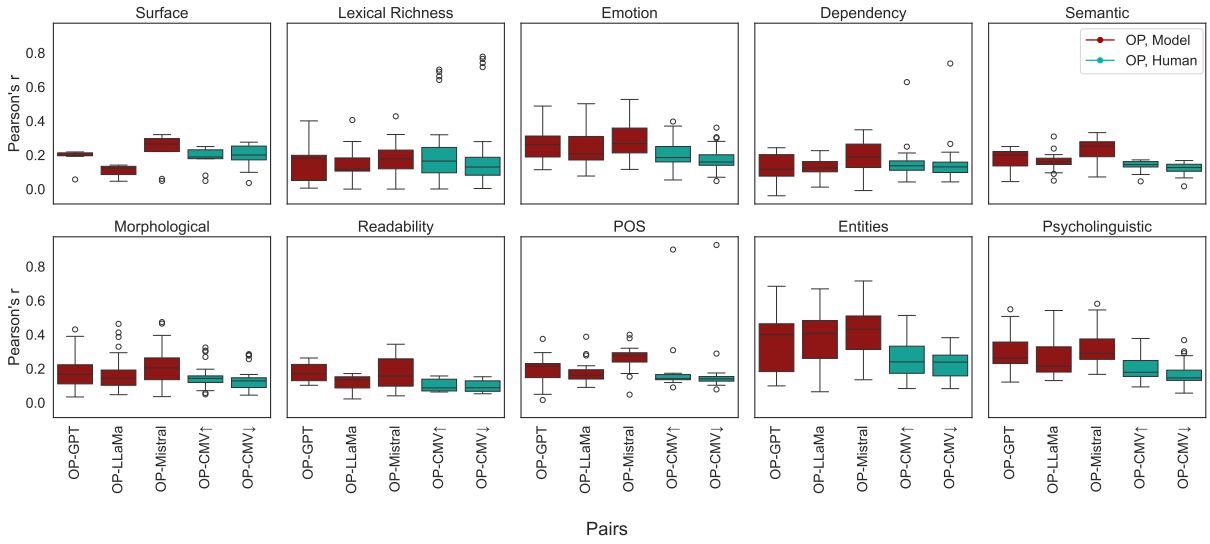


Figure 12: Aggregated Pearson's r for each pair of linguistic features per feature area, for the areas with the largest differences between original posts and their counter-arguments.

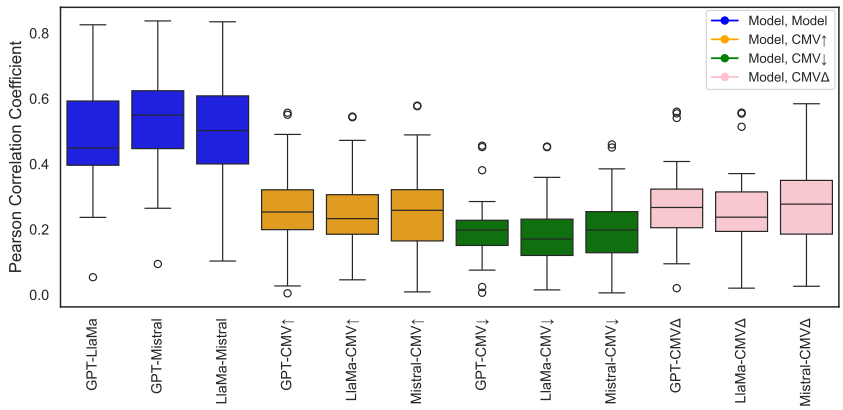


Figure 13: Pearson's r per comparison pair for argument quality features between pairs of human-written and LLM-generated counter-argument comments including with an addition of Δ comment comparisons.

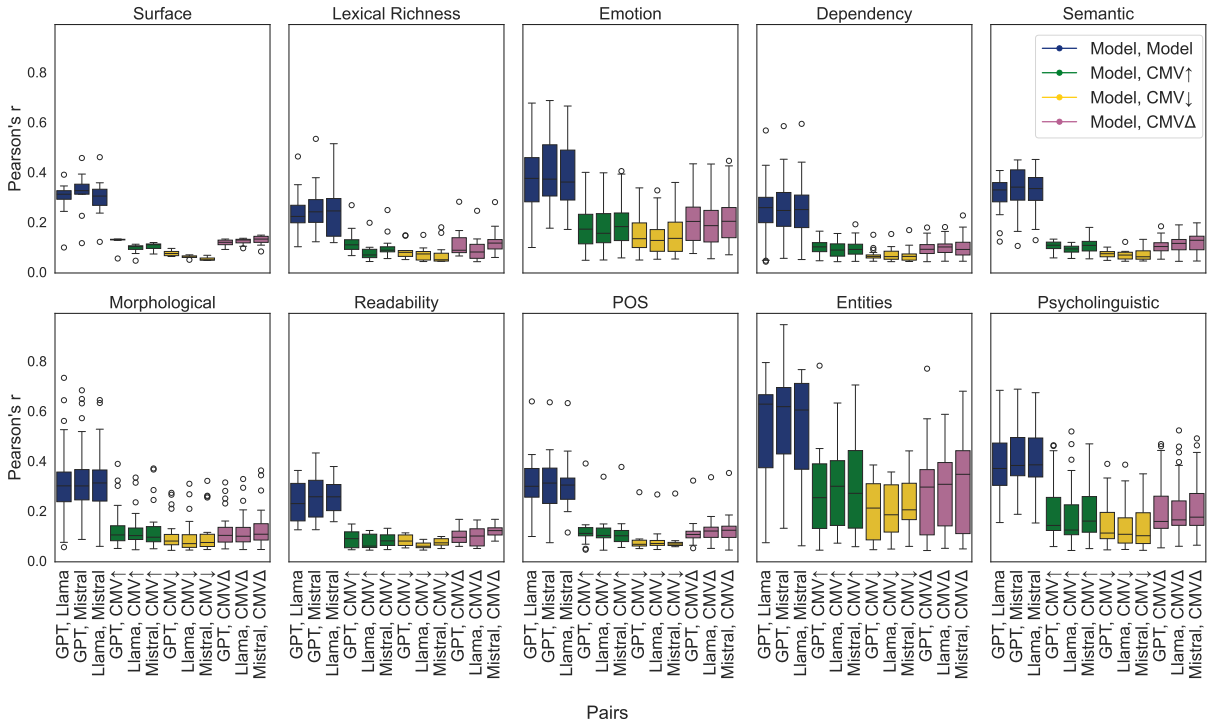


Figure 14: Aggregated Pearson's r for each pair of linguistic features per feature area, for the areas with the largest differences between pairs of counter-arguments with an addition of Δ comment comparisons.

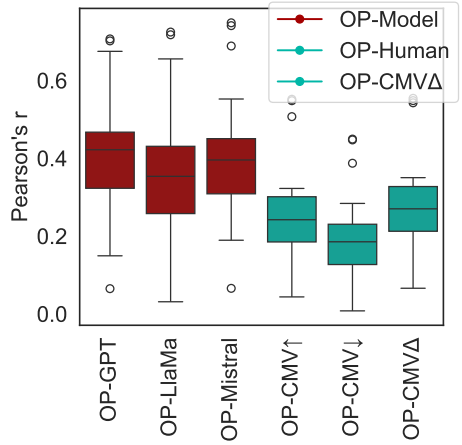


Figure 15: Pearson's r per comparison pair for argument quality features between pairs of human-written original posts (OP) and counter-argument comments (Model or Human) with an addition of Δ comment comparisons.

quency of LLMs' revealing their identity in generated arguments. Out of 13.498 cases (for each model), the model explicitly mentions that it's an LLM in 0.03%, 0.02%, 0.02% of the cases for GPT-3.5-turbo, LLAMA2-7B and MISTRAL-7B-instruct. However, since we do not use the raw texts but the features, revealing LLMs' own identity should now have an effect in the classifier performance. Meanwhile, this cannot be concluded for the majority of the prior work classifying human vs. LLM texts.

C.3 Additional Classification Experiments

To understand trends and results better, we conduct three additional classification experiments: (a) Classifying most up-voted (CMV↑) vs. all other human comments, (b) classifying least up-voted (CMV↓) vs. all other human comments, and (c) classifying LLM-generated comments vs. all human comments (instead of CMV↑ + CMV↓ only as in §5.3.1).

CMV↑ vs. other human comments. As shown in Table 22, the classification of CMV↑ vs. other human comments performs significantly worse than LLM-generated vs. human comments. While this is true across metrics, with a drop of > 0.07 for A, and a drop of > 0.2 for P compared to Table 18,

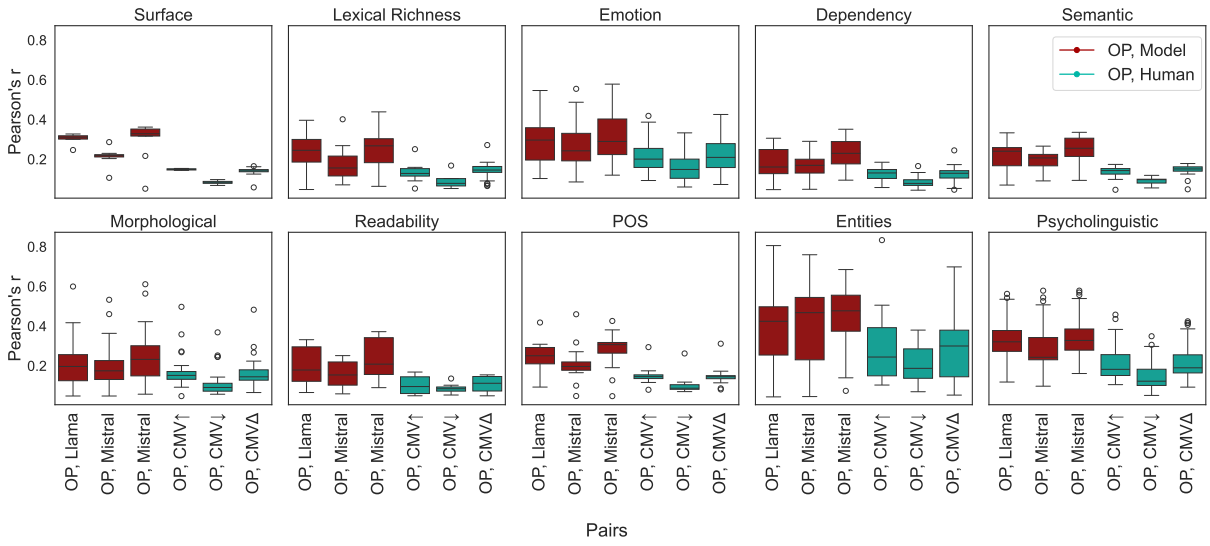


Figure 16: Aggregated Pearson's r for each pair of linguistic features per feature area, for the areas with the largest differences between original posts and their counter-arguments with an addition of Δ comment comparisons.

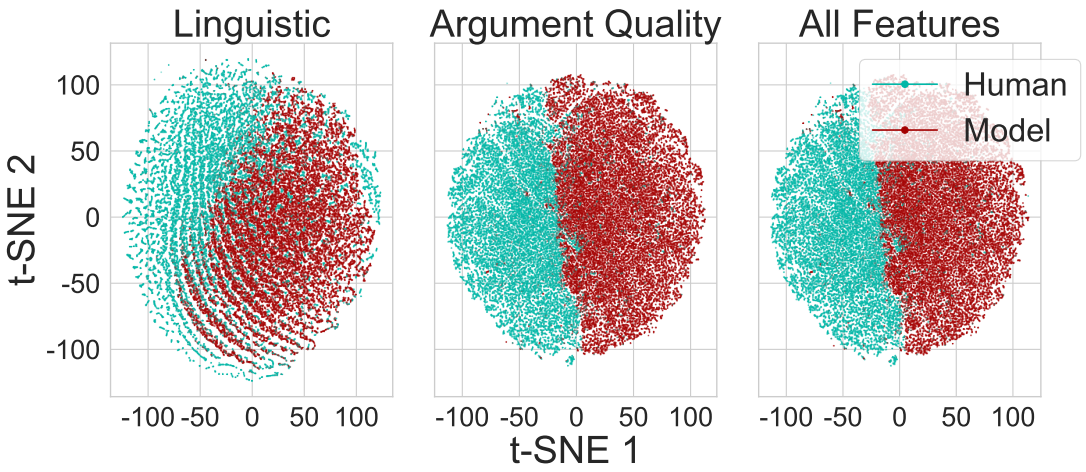


Figure 17: t-SNE projections of linguistic features, argument quality dimensions, and both of them concatenated.

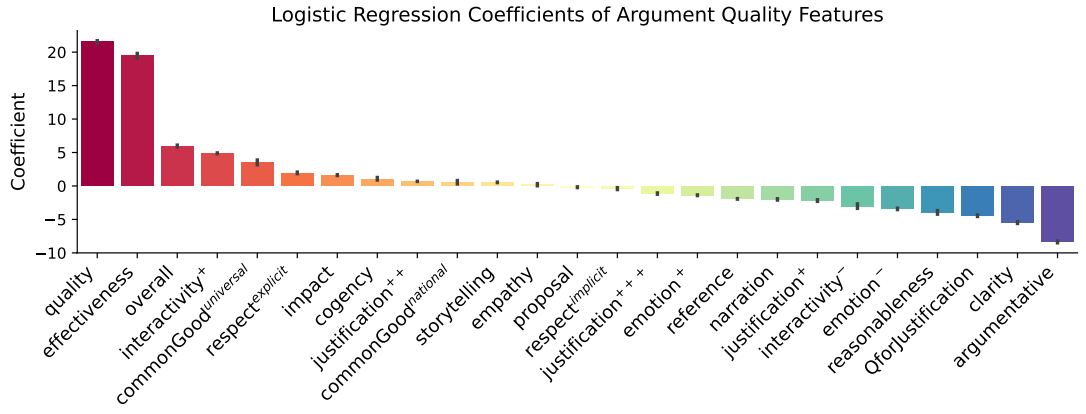


Figure 18: Logistic regression coefficients for the argument quality dimensions, showing the relative saliency of features in predicting Human vs. LLM-authored counter-argument comments.

Dimension	WS	D
reference	0.35	0.35
impact	0.31	-0.31
interactivity^{+}	0.28	-0.28
effectiveness	0.24	-0.24
respect$^{\text{explicit}}$	0.23	-0.23
respect$^{\text{implicit}}$	0.21	0.21
quality	0.20	-0.20
overall	0.20	-0.20
justification^{++}	0.19	-0.19
reasonableness	0.18	-0.18
emotion^{-}	0.18	0.18
cogency	0.18	-0.18
QforJustification	0.13	0.13
proposal	0.12	-0.12
justification^{+++}	0.10	-0.10
interactivity^{+}	0.07	-0.07
justification^{+}	0.07	0.07
clarity	0.06	-0.06
argumentative	0.05	0.01
commonGood$^{\text{national}}$	0.04	0.04
narration	0.04	-0.02
storytelling	0.03	-0.02
commonGood$^{\text{universal}}$	0.02	-0.02
emotion^{+}	0.02	-0.01
empathy	0.00	0.00

Table 9: Wasserstein distances (WS) and mean differences (D) between the human-written (C^H; only including C↑ and C↓) and model-generated comments (C^L) in aggregate and descending order for all argument quality features.

Dimension	WS	D
reference	0.35	0.35
impact	0.32	-0.32
interactivity^{+}	0.27	-0.27
effectiveness	0.24	-0.24
respect$^{\text{explicit}}$	0.22	-0.22
respect$^{\text{implicit}}$	0.20	0.20
quality	0.20	-0.20
overall	0.20	-0.20
justification^{++}	0.18	-0.18
cogency	0.18	-0.18
reasonableness	0.18	-0.18
emotion^{-}	0.17	0.17
QforJustification	0.13	0.13
proposal	0.12	-0.12
justification^{+++}	0.09	-0.09
interactivity^{-}	0.07	-0.07
justification^{+}	0.07	0.07
argumentative	0.05	0.02
clarity	0.05	-0.05
narration	0.04	-0.01
commonGood$^{\text{national}}$	0.03	0.03
storytelling	0.03	-0.02
commonGood$^{\text{universal}}$	0.02	-0.02
emotion^{+}	0.02	-0.00
empathy	0.00	0.00

Table 10: Wasserstein distances (WS) and mean differences (D) between all direct the human-written and model-generated comments in aggregate and descending order for all argument quality features.

this is in particular prevalent for R and $F1$, both of which show performance below 0.01. Overall, we take this as evidence of a significantly harder task than LLM-generated vs. human classification.

CMV↓ vs other human comments. As shown in Table 23, the classification of CMV↓ vs. other human comments performs significantly worse than LLM-generated vs. human comments. While this is true across metrics, with a drop of > 0.07 for A, and a drop of > 0.35 for P compared to Table 18, this is in particular prevalent for R and $F1$, both of which show performance below 0.01. Overall, we take this as evidence of a significantly harder task than LLM-generated vs. human classification.

LLM-generated comments vs. all human comments. To ensure the generalization of our classification results in §5.3.1 in settings where the number of human comments is much higher than the number of LLM-generated comments, and the choice of human comments is less well-informed,

Dimension	GPT-LLaMA	GPT-Mistral	LLaMA-Mistral
reference	**0.16**	**0.24**	0.08
emotion^{-}	0.07	0.06	0.01
story	0.06	0.08	0.02
interactivity^{-}	0.06	0.07	0.01
proposal	0.06	0.05	0.01
interactivity^{+}	0.05	0.08	0.04
justification^{+++}	0.05	0.05	0.00
emotion^{+}	0.05	0.04	0.02
argumentative	0.03	0.08	0.05
respect$^{\text{explicit}}$	0.03	0.06	0.03
justification^{+}	0.03	0.04	0.01
quality	0.03	0.03	0.01
narration	0.02	0.05	0.03
cogency	0.02	0.03	0.01
overall	0.02	0.03	0.01
impact	0.02	0.03	0.01
justification^{++}	0.02	0.02	0.01
clarity	0.02	0.02	0.00
respect$^{\text{implicit}}$	0.01	0.05	0.04
reasonableness	0.01	0.02	0.01
QforJustification	0.01	0.02	0.00
effectiveness	0.01	0.01	0.01
cgood$^{\text{universal}}$	0.01	0.01	0.00
cgood$^{\text{national}}$	0.00	0.01	0.00
empathy	0.00	0.00	0.00

Table 11: Wasserstein distances (WS) between model-generated comments per LLM pair in descending order for all argument quality features.

Feature	GPT-LLaMA	GPT-Mistral	LLaMA-Mistral
Type token ratio	0.13	0.14	0.01
Lemma token ratio	0.13	0.14	0.01
Avg. intensity anticipation	0.01	0.01	0.01
POS variability	0.03	0.03	0.01
Herdan's c	0.08	0.08	0.01
Avg. intensity trust	0.01	0.01	0.01

Table 12: Wasserstein distances (WS) between model-generated comments per LLM pair in aggregate for a distance threshold ≥ 0.1.

Feature	Com. - GPT	Com. - LLaMA	Com. - Mistral
Type token ratio	0.20	0.20	0.20
Lemma token ratio	0.19	0.19	0.19
Avg. intensity anticipation	0.13	0.13	0.13
POS variability	0.12	0.12	0.12
Herdan's C	0.11	0.11	0.11
Avg intensity trust	0.11	0.11	0.11

Table 13: Wasserstein distances (WS) between C^H and C^L in aggregate for a distance threshold ≥ 0.1. Com. is referring to human-written comments (C^L).

features	WS	D
ttr	0.20	0.20
lemma_token_ratio	0.19	0.19
avg_intensity_anticipation	0.13	-0.12
pos_variability	0.12	0.12
herdan_c	0.11	0.11
avg_intensity_trust	0.11	-0.10
avg_intensity_surprise	0.10	-0.10
avg_intensity_joy	0.09	-0.05
avg_intensity_anger	0.09	-0.07
n_high_intensity_trust	0.09	-0.09
n_high_Head_sensorimotor	0.08	-0.08
sichel_s	0.08	0.08
avg_intensity_sadness	0.08	-0.07
n_controversial_Gustatory_sensorimotor	0.07	-0.07
avg_intensity_fear	0.07	-0.03
n_controversial_Olfactory_sensorimotor	0.07	-0.07
n_low_intensity_joy	0.07	-0.07
cttr	0.07	-0.04
giroud_index	0.07	-0.04
rttr	0.07	-0.04
avg_aoa	0.07	-0.07
n_low_iconicity	0.06	-0.06
n_positive_sentiment	0.06	-0.06
n_controversial_Interoceptive_sensorimotor	0.06	-0.06
n_high_valence	0.06	-0.06
n_controversial_Auditory_sensorimotor	0.06	-0.06
lexical_density	0.06	-0.06
n_low_concreteness	0.06	-0.05
n_controversial_iconicity	0.06	-0.04
n_low_aoa	0.05	-0.04
avg_intensity_disgust	0.05	-0.04
avg_dominance	0.05	-0.05
n_high_dominance	0.05	-0.05
avg_sd_aoa	0.05	-0.05

Table 14: Wasserstein distances (WS) between between C^H and C^L in aggregate for a distance threshold ≥ 0.05.

Mapping	Comparison	ρ
M, M - M, M	GPT, LLaMA - LLaMA, Mistral	0.98
	GPT, LLaMA - GPT, Mistral	0.98
	GPT, Mistral - LLaMA, Mistral	0.98
M, H - M, H	LLaMA, Com.$^{\downarrow}$ - Mistral, Com.$^{\downarrow}$	0.90
	GPT, Com.$^{\downarrow}$ - LLaMA, Com.$^{\downarrow}$	0.86
	GPT, Com.$^{\downarrow}$ - Mistral, Com.$^{\downarrow}$	0.84
	LLaMA, Com.$^{\uparrow}$ - Mistral, Com.$^{\uparrow}$	0.93
	GPT, Com.$^{\uparrow}$ - Mistral, Com.$^{\uparrow}$	0.90
	GPT, Com.$^{\uparrow}$ - LLaMA, Com.$^{\uparrow}$	0.88

Table 15: Feature Spearman rank-correlations between comparisons. M, M - M, M refers to Model, Model - Model, Model comparisons, M, H - M, H to Model, Human - Model, Human comparisons. Com.$^{\uparrow}$ is the most upvoted comment, Com.$^{\downarrow}$ the least upvoted comment.

Mapping	Comparison	ρ
OP, H - OP, H	OP, Com.$^{\uparrow}$ - OP, Com.$^{\downarrow}$	0.95
OP, M - OP, M	OP, GPT - OP, Mistral	0.91
	OP, LLaMA - OP, Mistral	0.84
	OP, GPT - OP, LLaMA	0.76

Table 16: Feature Spearman rank-correlations between comparisons. OP, M - OP, M refers to OP, Model - OP, Model comparisons, OP, H - OP, H to OP, Human - OP, Human comparisons. Com.$^{\uparrow}$ is the most upvoted comment, Com.$^{\downarrow}$ the least upvoted comment.

n-fold	A	P	R	$F1$
1	.9804	.9855	.9818	.9837
2	.9814	.9844	.9846	.9845
3	.9800	.9844	.9822	.9833
4	.9800	.9831	.9836	.9833
5	.9820	.9854	.9846	.9850
Avg.	.9808	.9846	.9834	.984

Table 17: 5-fold evaluation of the logistic regression models' performance on prediction of human-written and LLM-generated arguments based on argument quality scores.

n-fold	A	P	R	$F1$
1	.9881	.9927	.9875	.9901
2	.9879	.9915	.9883	.9899
3	.9901	.9948	.9886	.9917
4	.9886	.9921	.9889	.9905
5	.9876	.9924	.9869	.9897
Avg.	.9885	.9927	.9881	.9904

Table 18: 5-fold cross-validation performance for the logistic regression model (LLM vs. Human) using only linguistic features.

n-fold	A	P	R	$F1$
1	.9928	.9953	.9927	.9940
2	.9923	.9942	.9930	.9936
3	.9935	.9964	.9927	.9946
4	.9921	.9943	.9925	.9934
5	.9929	.9957	.9925	.9941
Avg.	.9927	.9952	.9927	.9939

Table 19: 5-fold evaluation of the logistic regression models' performance on prediction of human-written and LLM-generated arguments based on argument quality scores and linguistic features.

we run the same experiment on LLM-generated vs. all human comments.

As Table 24 shows, the performance not only does not drop compared to the results in Table 18, but the LLM-specific signal gets even clearer when contrasted with more human comments.

C.4 MAGE

This subsection provides details on the MAGE benchmark and presents our results in comparison to other methods and in detail per testbench.

We train a logistic regression classifier on features extracted from MAGE data and evaluate using their metrics – Area Under the Receiver Operating Characteristic curve (*AUROC*) (higher is better), recall for human (R_{C^H}), LLM (R_{C^L}), and average (R_{avg}) – as well as P, R, F_1, and A.

C.4.1 Datasets

Table 25 gives an overview of the datasets in MAGE, including details about the size, source, and domain. Table 26 gives an overview of the eight testbeds in MAGE.

C.4.2 LLMs

The LLM texts from MAGE are generated using the models listed in Table 27 (see Li et al. (2024a) for further details).

C.4.3 Results (Comparison)

We compare the results for different generated content detection methods reported by Li et al. (2024a) with the results of a simple logistic regression trained on linguistic features.

Table 28 and 29 reveal that our simple approach not only outperforms humans and ChatGPT but also shows performance on par with computationally much more expensive methods in fixed-domain fixed-model detection scenarios.

This trend holds for other in-distribution scenarios, as Table 30 shows. Using linguistic features with a simple logistic regression for machine-generated content detection is surprisingly robust, showing much less performance degradation than other methods.

Overall, given the simplicity and interoperability of our approach, it is a viable alternative to some other methods, in particular in scenarios with limited resources or where interpretability is vital.

C.4.4 Results (Ours, detailed)

In the following, we present the full results per testbed of our method using logistic regression with linguistic features.

Testbed 1: Fixed-Domain & Model-specific As shown in Table 31, our approach generally performs well (over 0.8 for most domains for most metrics) with some variation across domains. Given the particularly low R_{C^L} for ELI5 (0.625) and Yelp (0.557), they appear to be more challenging domains.

Testbed 2: Arbitrary-domains & Model-specific As Table 32 shows, there are marked differences between models. Text from open-weight/-source models appears to be significantly easier to detect with linguistic features than the closed OpenAI models, as evidenced by the systematically lower R_{C^L} in particular for `text-davinci-002` and `text-davinci-003`. There is no apparent parameter scale effect within model families.

Testbed 3: Fixed-domain & Arbitrary-models As Table 33 shows, the trends are generally the same as in testbed 1. While the performance is generally lower across metrics and domains, dropping by around 10%, Yelp stays the most challenging domain.

Testbed 5: Unseen Models We present the results for the detection of unseen models per model in Table 34. Given information during training about general markers of LLM-generated text, linguistic features are useful across models, with a R_{C^L} across models over 0.66. The most marked difference can be found between the closed OpenAI models and open-weight/-source models. While the latter show a R_{C^L} consistently over 0.8, the formers' is consistently lower. Again, we do not find systematic parameter scale effects within model families.

	1	2	3	4	5	Avg.
quality	14.83	14.53	13.91	14.27	14.68	14.45
clarity	-4.62	-4.40	-4.20	-4.57	-4.20	-4.40
cogency	1.54	1.44	1.79	1.26	1.42	1.49
effectiveness	16.12	15.62	16.15	16.19	15.42	15.90
reasonableness	-3.80	-3.73	-3.86	-3.92	-3.39	-3.74
overall	3.95	4.01	4.10	4.05	3.86	3.99
impact	1.38	1.66	1.54	1.46	1.53	1.51
argumentative	-8.78	-8.62	-8.54	-8.57	-8.42	-8.58
emotion^{+}	-0.89	-0.95	-0.96	-0.95	-0.91	-0.93
emotion^{-}	-3.12	-3.21	-3.09	-2.90	-3.19	-3.10
narration	-0.29	-0.17	-0.15	-0.67	-0.58	-0.37
QforJustification	-4.11	-4.02	-4.19	-3.85	-3.72	-3.98
empathy	-0.14	-0.50	-0.54	-0.25	-0.21	-0.33
proposal	-0.13	-0.42	-0.02	-0.13	-0.44	-0.23
reference	-2.20	-2.30	-2.21	-2.25	-2.19	-2.23
justification^{+}	-2.18	-2.16	-2.28	-2.23	-2.22	-2.22
justification^{++}	1.63	1.58	1.46	1.57	1.33	1.51
justification^{+++}	-0.23	-0.21	-0.37	-0.10	-0.09	-0.20
commonGood$^{\text{national}}$	0.69	0.81	0.33	0.66	0.33	0.56
commonGood$^{\text{universal}}$	1.83	1.78	1.78	1.77	1.88	1.81
respect$^{\text{implicit}}$	-0.33	-0.33	-0.36	-0.24	-0.32	-0.32
respect$^{\text{explicit}}$	1.97	1.86	1.89	2.07	1.97	1.95
story	0.75	0.61	0.66	0.77	0.78	0.71
interactivity^{-}	-1.95	-1.90	-1.55	-1.67	-1.81	-1.78
interactivity^{+}	4.32	4.41	4.71	4.58	4.34	4.47

Table 20: Logistic regression coefficients for each of the 5 folds and on average for the 15 most predictive argument quality dimensions, showing the relative saliency of features in predicting Human vs. LLM-authored counter-argument comments.

Feature	Mean Coeff.	Std.
maas_index	5.0708	0.5674
n_hapax_dislegomena	-3.8671	0.2416
n_dependency_nsubj	-3.5075	0.3399
n_types	-3.3919	0.3822
n_pron	-3.0315	0.2529
n_lemmas	-2.6867	0.3884
n_sentences	2.1243	0.1780
ttr	-2.1208	0.3868
n_PUNCT_PunctType_Comm	2.0851	0.0858
n_dependency_punct	-2.0215	0.1446
summer_index	1.9347	0.4260
n_monosyllables	-1.8784	0.1326
n_polysyllables	1.8221	0.1305
n_low_age of acquisition rating	1.8138	0.2083
n_dependency_dep	-1.6811	0.1625

Table 21: Logistic regression mean coefficient and standard deviation across 5 folds for the 15 most predictive linguistic features, showing the relative saliency of features in predicting Human vs. LLM-authored counter-argument comments.

n-fold	A	P	R	$F1$
1	0.9148	0.8462	0.0041	0.0081
2	0.9147	0.7273	0.0030	0.0059
3	0.9147	0.8889	0.0030	0.0059
4	0.9147	0.7500	0.0033	0.0066
5	0.9147	0.7500	0.0022	0.0044
Avg.	0.9147	0.7925	0.0031	0.0062

Table 22: 5-fold cross validation of the logistic regression models' performance on prediction of CMV↑ vs. other human comments using linguistic features.

n-fold	A	P	R	$F1$
1	0.9145	0.5294	0.0033	0.0066
2	0.9146	0.6429	0.0033	0.0066
3	0.9147	0.6923	0.0033	0.0066
4	0.9146	0.6000	0.0033	0.0066
5	0.9147	0.6429	0.0033	0.0066
Avg.	0.9146	0.6215	0.0033	0.0066

Table 23: 5-fold cross validation of the logistic regression models' performance on prediction of CMV↓ vs. other human comments using linguistic features.

n-fold	A	P	R	$F1$
1	0.9999	0.9996	1.0000	0.9998
2	0.9999	0.9998	1.0000	0.9999
3	0.9999	0.9996	0.9998	0.9997
4	0.9998	0.9996	0.9994	0.9995
5	0.9999	0.9998	0.9998	0.9998
Avg.	0.9999	0.9997	0.9998	0.9997

Table 24: 5-fold cross validation of the logistic regression models' performance on prediction of LLM-generated vs. all human comments using linguistic features.

Dataset	Size	Domain	Source
CMV	804	Opinion statement	Tan et al. (2016)
Yelp	1000	Opinion statement	Zhang et al. (2015)
XSum	1000	News articles	Narayan et al. (2018)
TLDR	777	News articles	TLDR_news[9]
ELI5	1000	QA Answers	Fan et al. (2019)
WP	1000	Stories	Fan et al. (2018)
ROCStories	1000	Stories	Mostafazadeh et al. (2016)
HellaSwag	1000	Commonsense	Zellers et al. (2019)
SQuAD	1000	Wikipedia	Rajpurkar et al. (2016)
SciXGen	1000	Scientific writing	Chen et al. (2021)

Table 25: Datasets in MAGE, among which *CMV* and *Yelp* have an inherently argumentative nature. Also, *CMV*, *ELI5*, and *WP* (WritingPrompts) are all sub-communities in Reddit.

Testbed	Description
1	Fixed domain & fixed model
2	Arbitrary domains & fixed model
3	Fixed domain & arbitrary models
4	Arbitrary domains & arbitrary models (known set)
5	Unseen models (out-of-distribution)
6	Unseen domains (out-of-distribution)
7	Unseen domains & models (out-of-distribution)
8	Robustness to paraphrase attacks

Table 26: Overview of the eight test-beds in MAGE used to evaluate detection performance across varying levels of domain and model familiarity from Li et al. (2024a).

Testbed 6: Unseen Domains We present the results for the detection of LLM-generated content in unseen domains in Table 35. While all domains except for Yelp and Sci Gen show robust performance for R_{C^L}, there are marked differences between them. CMV and WP retain particularly high performance, indicating useful information on LLM-specific writing styles in these domains being present in other domains, too.

C.5 Results (Linguistic Features, Argument Quality, All)

We test the same three settings as in Section 5.3.1: Training a logistic regression on (a) linguistic features, (b) argument quality dimensions, and (c) both. Given that argument quality dimensions are only fully applicable to argumentative domains, we only run this experiment on CMV and Yelp on testbed 1 and 3 of MAGE.

The results for testbed 1 are presented in Table 36. While the classifier based on the argument quality dimensions is systematically lower than the one based on linguistic features for both domains and across metrics, the information of both appears to be complementary, as the best performance is obtained using both. While the tendencies are less

Model Family	Variants
OpenAI GPT (Brown et al., 2020)	text-davinci-002, text-davinci-003, gpt-3.5-turbo
LLaMA (Touvron et al., 2023)	6B, 13B, 30B, 65B
GLM (Zeng et al., 2023)	GLM-130B
FLAN-T5 (Chung et al., 2022)	small, base, large, xl, xxl
OPT (Zhang et al., 2022)	125M, 350M, 1.3B, 2.7B, 6.7B, 13B, 30B, iml-1.3B, iml-30B
BigScience (Sanh et al., 2022)	T0-3B, T0-11B, BLOOM-7B1
EleutherAI (Wang, 2021)	GPT-J-6B, GPT-NeoX-20B

Table 27: Overview of LLM families and their variants in MAGE.

Detector	R_{C^H}	R_{C^L}	R_{avg}
ChatGPT	96.98%	12.03%	54.51%
Human	61.02%	47.98%	54.50%
Ours	96.13%	79.70 %	87.92%

Table 28: Detection performance of ChatGPT and humans reported in Li et al. (2024a) compared to our approach.

Methods	Human/Machine	R_{avg}	AUROC
FastText	94.72% / 94.36%	94.54%	0.98
GLTR	90.96% / 83.94%	87.45%	0.94
Longformer	97.30% / 95.91%	96.60%	0.99
DetectGPT	91.68% / 81.06%	86.37%	0.92
Ours	96.13% / 79.70 %	87.92%	0.88

Table 29: (Testbed 1) White-box detection performance. "Human/Machine" denotes R_{C^H} and R_{C^L}, respectively, reported in Li et al. (2024a) compared to our approach.

clear, these trends are reflected in testbed 3 (Table 37).

D Implementation Details and Used Resources

D.1 Implementation Details & Reproducibility

We link our GitHub repository for the implementation details and reproducibility. Package versions can be found in *requirements.txt*.

D.2 Resources

Open-code LLM inference was performed using the HuggingFace text generation pipeline at `www.huggingface.co/docs/text-generation-inference` and the inference was run on in-house NVIDIA RTX A6000 GPUs. API-access GPT-3.5-turbo inference was run via OpenAI text completion API at `https://platform.openai.com/docs/guides/gpt` (accessed between Dec. 8 - 19. 2024).

E Usage of AI Assistants

In this work, GitHub Copilot[10] was used as a code completion/suggestion tool. Additionally, AI-assisted writing tools like Grammarly[11] have been used for spelling checks and grammar corrections.

[10] `https://github.com/features/copilot`
[11] `https://app.grammarly.com/`

Settings	Methods	R_{C^H}	R_{C^L}	R_{avg}	AUROC
Testbed 2: In-distribution Detection Arbitrary-domains & Model–specific	FastText (Joulin et al., 2017)	88.96%	77.08%	83.02%	0.89
	GLTR (Gehrmann et al., 2019)	75.61%	79.56%	77.58%	0.84
	Longformer (Beltagy et al., 2020)	95.25%	96.94%	96.10%	0.99
	DetectGPT⋆ (Mitchell et al., 2023)	48.67%	75.95%	62.31%	0.60
	Ours	98.92%	69.23%	84.08 %	0.84
Testbed 3: In-distribution Detection Fixed-domain & Arbitrary-models	FastText (Joulin et al., 2017)	89.43%	73.91%	81.67%	0.89
	GLTR (Gehrmann et al., 2019)	37.25%	88.90%	63.08%	0.80
	Longformer (Beltagy et al., 2020)	89.78%	97.24%	93.51%	0.99
	DetectGPT⋆ (Mitchell et al., 2023)	86.92%	34.05%	60.48%	0.57
	Ours	75.30%	90.04%	82.67%	0.83
Testbed 4: In-distribution Detection Arbitrary-domains & Arbitrary-models	FastText (Joulin et al., 2017)	86.34%	71.26%	78.80%	0.83
	GLTR (Gehrmann et al., 2019)	12.42%	98.42%	55.42%	0.74
	Longformer (Beltagy et al., 2020)	82.80%	98.27%	90.53%	0.99
	DetectGPT⋆ (Mitchell et al., 2023)	86.92%	34.05%	60.48%	0.57
	Ours	61.62%	89.73%	75.67%	0.76
Testbed 5: Out-of-distribution Detection – Unseen Models	FastText (Joulin et al., 2017)	83.12%	54.09%	68.61%	0.74
	GLTR (Gehrmann et al., 2019)	25.77%	89.21%	57.49%	0.65
	Longformer (Beltagy et al., 2020)	83.31%	89.90%	86.61%	0.95
	DetectGPT⋆ (Mitchell et al., 2023)	48.67%	75.95%	62.31%	0.60
	Ours	75.81%	88.51%	82.16%	0.82
Testbed 6: Out-of-distribution Detection – Unseen Domains	FastText (Joulin et al., 2017)	54.29%	72.79%	63.54%	0.72
	GLTR (Gehrmann et al., 2019)	15.84%	97.12%	56.48%	0.72
	Longformer (Beltagy et al., 2020)	38.05%	98.75%	68.40%	0.93
	DetectGPT⋆ (Mitchell et al., 2023)	86.92%	34.05%	60.48%	0.57
	Ours	94.79%	67.38%	81.09%	0.81

Table 30: (Testbeds 2–6) Detection performance of different methods. Out-of-distribution settings evaluate detection on texts from unseen domains or texts generated by previously unseen LLMs. ⋆ denotes unsupervised detection reported in Li et al. (2024a).

Domain	R_{C^H}	R_{C^L}	R_{avg}	AUROC
CMV	0.979	0.870	0.925	0.925
ELI5	0.993	0.625	0.809	0.809
HellaSWAG	0.987	0.887	0.937	0.937
ROCT	0.728	0.846	0.787	0.787
Sci Gen	0.984	0.845	0.915	0.915
SQuAD	0.995	0.728	0.861	0.861
TLDR	0.975	0.837	0.906	0.906
WP	0.993	0.936	0.965	0.965
Yelp	0.998	0.557	0.778	0.778
XSUM	0.980	0.839	0.909	0.909
Avg.	0.961	0.797	0.879	0.879

Table 31: MAGE results: Testbed 1

Model	R_{C^H}	R_{C^L}	R_{avg}	AUROC
gpt-3.5-turbo	0.981	0.569	0.775	0.775
text-davinci-002	0.980	0.198	0.589	0.589
text-davinci-003	0.985	0.269	0.627	0.627
GLM130B	0.994	0.709	0.852	0.852
t0 3b	0.983	0.561	0.772	0.772
t0 11b	0.984	0.608	0.796	0.796
flan t5 small	0.979	0.605	0.792	0.792
flan t5 base	0.978	0.425	0.702	0.702
flan t5 xl	0.981	0.453	0.717	0.717
flan t5 xxl	0.989	0.679	0.834	0.834
bloom 7b	0.998	0.939	0.969	0.969
flan t5 large	0.981	0.397	0.689	0.689
gpt j	0.998	0.962	0.980	0.980
gpt neox	0.998	0.941	0.970	0.97
opt 125m	0.994	0.790	0.892	0.892
opt 350m	0.995	0.826	0.911	0.911
opt 2.7b	0.997	0.896	0.946	0.946
opt 1.3b	0.995	0.850	0.923	0.923
opt iml max 1.3b	0.991	0.770	0.881	0.881
opt 6.7b	0.995	0.840	0.918	0.918
opt 13b	0.998	0.907	0.952	0.952
opt 30b	0.998	0.900	0.949	0.949
opt iml 30b	0.995	0.848	0.922	0.922
LLaMA 7B	0.989	0.743	0.866	0.866
LLaMA 13B	0.984	0.715	0.850	0.850
LLaMA 30B	0.985	0.642	0.813	0.813
LLaMA 65B	0.983	0.648	0.816	0.816
Avg.	0.989	0.692	0.841	0.841

Table 32: MAGE results: Testbed 2

Domain	R_{C^H}	R_{C^L}	R_{avg}	AUROC
CMV	0.807	0.961	0.884	0.884
ELI5	0.842	0.847	0.845	0.845
HellaSWAG	0.786	0.972	0.879	0.879
ROCT	0.280	0.933	0.606	0.606
Sci Gen	0.745	0.939	0.842	0.842
SQuAD	0.900	0.825	0.862	0.862
TLDR	0.631	0.949	0.790	0.790
WP	0.919	0.965	0.942	0.942
Yelp	0.909	0.667	0.788	0.788
XSUM	0.712	0.946	0.829	0.829
Avg.	0.753	0.900	0.827	0.827

Table 33: MAGE results: Testbed 3

Model	R_{C^H}	R_{C^L}	R_{avg}	AUROC
gpt-3.5-turbo	0.765	0.663	0.714	0.714
text-davinci-002	0.766	0.742	0.754	0.754
text-davinci-003	0.770	0.666	0.718	0.718
GLM130B	0.757	0.924	0.841	0.841
t0 3b	0.759	0.891	0.825	0.825
t0 11b	0.757	0.893	0.825	0.825
flan t5 base	0.762	0.827	0.795	0.795
flan t5 xl	0.761	0.800	0.781	0.781
flan t5 xxl	0.758	0.903	0.831	0.831
gpt j	0.753	0.988	0.871	0.871
gpt neox	0.753	0.992	0.873	0.873
bloom 7b	0.753	0.986	0.870	0.870
flan t5 small	0.760	0.846	0.803	0.803
flan t5 large	0.762	0.809	0.786	0.786
opt 125m	0.756	0.912	0.834	0.834
opt 350m	0.755	0.911	0.833	0.833
opt 1.3b	0.755	0.962	0.858	0.858
opt iml max 1.3b	0.756	0.933	0.845	0.845
opt 2.7b	0.754	0.973	0.864	0.864
opt 6.7b	0.754	0.955	0.855	0.855
opt 13b	0.754	0.973	0.864	0.864
opt 30b	0.753	0.979	0.866	0.866
opt iml 30b	0.754	0.962	0.858	0.858
LLaMA 7B	0.758	0.883	0.821	0.821
LLaMA 13B	0.759	0.878	0.819	0.819
LLaMA 30B	0.761	0.833	0.797	0.797
LLaMA 65B	0.761	0.810	0.785	0.785
Avg.	0.758	0.885	0.822	0.822

Table 34: MAGE results: Testbed 5

Domain	R_{C^H}	R_{C^L}	R_{avg}	AUROC
CMV	0.967	0.857	0.912	0.912
ELI5	0.981	0.653	0.817	0.817
HellaSWAG	0.982	0.723	0.852	0.852
ROCT	0.637	0.609	0.623	0.623
Sci Gen	0.995	0.536	0.766	0.766
SQuAD	0.989	0.692	0.841	0.841
TLDR	0.960	0.740	0.850	0.850
WP	0.991	0.883	0.937	0.937
Yelp	0.994	0.365	0.679	0.679
XSUM	0.984	0.679	0.831	0.831
Avg.	0.948	0.674	0.811	0.811

Table 35: MAGE results: Testbed 6

Domain	Features	R_{C^H}	R_{C^L}	R_{avg}	AUROC
CMV	Argument Quality	0.975	0.484	0.729	0.729
	Linguistic	0.979	0.870	0.925	0.925
	All	0.980	0.893	0.937	0.937
Yelp	Argument Quality	0.999	0.070	0.534	0.534
	Linguistic	0.998	0.557	0.778	0.778
	All Features	0.998	0.580	0.789	0.789

Table 36: Testbed 1: Comparison feature sets for cmv, yelp

Domain	Features	R_{C^H}	R_{C^L}	R_{avg}	AUROC
CMV	Argument Quality	0.376	0.958	0.667	0.667
	Linguistic	0.807	0.961	0.884	0.884
	All	0.855	0.966	0.911	0.911
Yelp	Argument Quality	0.833	0.429	0.631	0.631
	Linguistic	0.998	0.557	0.778	0.778
	All	0.912	0.707	0.810	0.810

Table 37: Testbed 3: Comparison feature sets for CMV, Yelp