How Ethos and Pathos Appeals Resonate in Reader Interpretations of Social Media Messages

Ewelina Gajewska[*], Katarzyna Budzynska[*], Jarosław A. Chudziak[*], Liesbeth Allein[†‡]

[*]Warsaw University of Technology, Poland
[†]Department of Computer Science, KU Leuven, Belgium
[‡] Department of Electronics and Information Systems, Ghent University, Belgium

Correspondence: ewelina.gajewska.dokt@pw.edu.pl

Abstract

Rhetorical strategies and their influence on audiences are often studied through social media posts and comments. However, this focus overlooks the universal audience, which is the majority of readers who remain silent and do not explicitly express how a message affects them. This study investigates how two classical modes of persuasion, ethos and pathos, resonate in the silent audience's interpretations of meaning. Using a dataset of social media sentences paired with human-written interpretations, we label both sources for ethos and pathos and assess whether these rhetorical appeals are preserved. Our analyses show that interpretations diverge from the original sentences in 30% of cases, with rhetorically charged content eliciting greater variability than neutral content. We further find that ethos and pathos in original sentences can predict audience attitudes toward the author, underscoring the subtle ways rhetoric shapes perception beyond visible engagement.

1 Introduction

Rhetoric, meaning, and audience perception are deeply intertwined in online communication. Yet analyzing their relationship is particularly challenging in ambiguous contexts such as social media. Ambiguity creates an *interpretive space* (Sandri et al., 2023), allowing the same message to be understood in multiple ways depending on the cultural background, knowledge, and cognitive frames of different audiences. Rhetorical devices are central to how this ambiguity is negotiated. This is clear in current digital ecosystems, where prominent figures exercise considerable rhetorical power. Through appeals to credibility (*ethos*) and emotion (*pathos*) (Aristotle, 1991), they shape how people perceive themselves, others, and society as a whole. Unlike devices like metaphors or narratives, these appeals directly target interpersonal trust and affective orientation. Investigating how such foundational modes of persuasion resonate with audiences is paramount to studying influence, trust, and polarization online (Del Vicario et al., 2016; Garimella et al., 2018; Buder et al., 2021).

So far, computational works on rhetoric mainly focus on identifying rhetorical strategies and fallacies in social media posts and comments (Habernal et al., 2017; Da San Martino et al., 2019; Sheng et al., 2021; Jin et al., 2022; Wiegand et al., 2023; Chen et al., 2024; Bagdon et al., 2025; Ramponi et al., 2025a). However, inferring audience perceptions only from explicit content presents strong limitations as only a small share of the audience is captured. The dynamics of interpretation and influence extend beyond vocal participants to the much larger *universal audience*. In classical and contemporary rhetorical theory (Aristotle, 1991; Perelman and Olbrechts-Tyteca, 1971), they are those who do not speak but remain rhetorically present, i.e., they observe, internalize, and shape discourse. This concept aligns with social media dynamics, where a majority of users consume content without commenting. In that respect, this work studies the rhetorical situation of interpretation rather than demographic segmentation.

This study investigates *if* and *how* rhetorical devices transform as they move from the original sentence to audience interpretations. Specifically, we examine whether ethos and pathos appeals in the sentence are retained, transformed, or omitted in internal representations of sentence meaning, i.e., reader interpretations. Interpretations are strong predictors of persuasive influence (Eagly, 2014). While influence can occur without full reproduction of devices, what is retained or omitted reveals which cues are salient and memorable. We further explore whether these appeals are predictive for variation in reader attitudes towards the sentence's author and describe additional factors that guide interpretation alongside ethos and pathos labels.

Proceedings of the 27th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 327–341
August 2–5, 2026.

Contributions Our work introduces *a novel computational perspective on rhetorical influence* by examining how rhetorical devices resonate in audience interpretations of social media content. We provide *large-scale empirical evidence* of ethos and pathos transfer from text to audience interpretations. This type of perspectivism has been rarely explored in computational research. Through quantitative and qualitative analyses, we identify *predictive factors driving ethos and pathos divergence* between source sentences and audience interpretations, and within interpretations. We further *enrich an existing interpretation modeling dataset with new annotation layers of ethos and pathos*, which are made publicly available[1].

2 Related Work

Variation in interpretation. This work is inspired by a growing literature advocating perspectivist approaches to language understanding (Plank, 2022; Cabitza et al., 2023; Fleisig et al., 2024; Frenda et al., 2025). Perspectivism moves away from single ground-truth interpretations to capture human variation. It is motivated by strong subjectivity and frequent annotation disagreement in tasks such as natural language inference (Pavlick and Kwiatkowski, 2019; Lee et al., 2023; Weber-Genzel et al., 2024), toxic and offensive language detection (Goyal et al., 2022; Sandri et al., 2023; Davani et al., 2024; Trusca and Allein, 2025), and values identification (Sorensen et al., 2024).

Unlike dominant approaches in perspectivism literature, this work does not assess the subjectivity of ethos and pathos through human label variation of the original sentence. Instead, we capture variation in language understanding through the set of audience interpretations that accompany the original sentence and how these differ in appeals to ethos and pathos. Our analysis of ethos and pathos thus takes place after interpretations are formulated. Consequently, the translation from original sentence to reader interpretations is not explicitly constrained by prior judgments of ethos and pathos, which is arguably closer to how interpretation occurs "in the wild".

Ethos and pathos. Many computational studies on rhetoric focus on rhetorical fallacies (Habernal et al., 2018; Goffredo et al., 2022; Jin et al., 2022; Musi et al., 2022; Glockner et al., 2025; Ramponi et al., 2025b), which may function as persuasive shortcuts (Walton, 2010). However, the influence of classic rhetorical strategies like pathos[2] and, especially, ethos on the audience remains underinvestigated. Evgrafova et al. (2024) analyzed pathos in user-generated argumentation and highlighted how emotional appeals can influence audience attitudes and intensify divisions. Duthie et al. (2016) built a computational model for extracting ethotic expressions from the UK parliamentary debates. Gajewska and Budzynska (2024) extended their approach to social media contexts, investigating how ethos attacks and supports contribute to the polarization of attitudes in climate change discussions.

This study investigates if and how ethos and pathos change when a sentence is translated to different interpretations among its audience.

3 Methodology

3.1 Data

Our analyses rely on the OrigamIM dataset (Allein and Moens, 2024), which was developed to support interpretation modeling (Allein et al., 2025). It contains 2,018 English sentences sampled from subreddit r/ChangeMyView, which represent views on controversial and polarizing topics, such as abortion and racism. In this subreddit, users articulate their stance on a topic and invite others to persuade them to reconsider. Each sentence is paired with approximately five human-written reformulations (9,851 interpretations in total) that reflect readers' interpretations of its implicit meanings, guided by diverse annotations of hidden moral judgments. Each interpretation is also paired with an attitude score on a five-point Likert scale, ranging from *very negative* (1) to *very positive* (5), indicating the reader's attitude toward the author of the original sentence.

The annotators in OrigamIM processed the text presented to them and formed internal interpretations, akin to a universal audience. Since they were unconstrained in phrasing, interpretations may rephrase, summarize, or even expand the sentence's content. Summarization often highlights salient cues, whereas expansion reveals interpretive ambiguity. Note that recording these interpretations served as a methodological step to capture

[1] https://doi.org/10.17605/OSF.IO/DAC2T

[2] Related tasks such as emotion detection and sentiment analysis have a longstanding tradition in natural language processing. Yet, while these tasks center on the identification of emotional content expressed in text, pathos mining investigates the strategic use of emotions for persuasion.

cognitive responses, not evidence of active engagement.

Our analyses require multiple human-written interpretations and a context where rhetorical strategies are actively employed. OrigamIM satisfies these conditions, making it an ideal setting for isolating rhetorical effects. Platforms like X or Facebook differ in interaction style (e.g., more reactive) and may exhibit persuasion less prominently, which could reduce precision for the type of analysis conducted here. While five interpretations cannot fully capture the diversity of the universal audience (Allein et al., 2025), this indicative sample of human-written audience interpretation provides a solid starting point for systematic study of rhetorical resonance.

3.2 Ethos and Pathos Classification

For this preliminary study, we opt for a silver standard annotation pipeline to classify the rhetorical dimensions. Manually annotating a dataset of this scale, comprising nearly 10,000 text spans, for complex rhetorical dimensions like *ethos* and *pathos* presents significant resource and time constraints that fall outside the scope of this initial work. Furthermore, there exist datasets for training automated classifiers on these specific rhetorical appeals, therefore leveraging these models provides a scalable, computationally viable baseline for our exploratory analysis.

This section is structured as follows. We first define ethos and pathos, and illustrate these rhetorical devices with manually annotated examples from OrigamIM (§3.2.1). We then formalize the classification task (§3.2.2) and outline the experimental setup (§3.2.3). We report classification results (§3.2.4) and assess label reliability through automatic and manual evaluation (§3.2.5).

3.2.1 Ethos and Pathos Definition

Ethos is the speaker's credibility and moral character, which are pivotal for persuasive communication (Aristotle, 1991). It is constructed through the speaker's demonstration of practical wisdom (*phronesis*), virtue or moral excellence (*arete*), and goodwill towards the audience (*eunoia*). Ethos is not inherent but is established and perceived within the dynamic interaction between speaker and audience as the audience evaluates the speaker's trustworthiness and authority. Consequently, ethos functions as a rhetorical appeal that fundamentally relies on the audience's perception of the speaker's ethical character and expertise (Gajewska et al., 2024; Gajewska and Budzynska, 2024). The goal of ethos annotation is to identify sentences where a speaker references the credibility or character of an entity, either in a supportive or attacking way.

(1) [Ethos: -] *Reading / watching CNN is like eating junk food out of the trash bin.*

(2) [Ethos: +] *Man, she is actually a really good politician if she can handle all that.*

Pathos refers to the second mode of persuasion that appeals to affective states of the audience (Aristotle, 1991). By strategically evoking specific emotions in the audience, the speaker influences their judgment and decision-making. The goal of pathos annotation is to capture rhetorical appeals that aim to induce specific affective states in the hearer rather than simply express emotions.

(3) [Pathos: -] *I have a feeling there is going to be a HUGE genocide in the next 5-10 years.*

(4) [Pathos: +] *Look forward to see all the great things you will lead!*

Together, these rhetorical modes illustrate persuasion as contingent on shifting interpersonal and emotional perspectives rather than objective, static truths.

3.2.2 Task Formalization

We define ethos and pathos classifiers as functions mapping textual inputs to categorical label sets:

$$\phi_{\text{ethos}} : \mathcal{S} \rightarrow \mathcal{C}_{\text{ethos}}, \qquad \phi_{\text{pathos}} : \mathcal{S} \rightarrow \mathcal{C}_{\text{pathos}}$$

with $\mathcal{S}$ the space of all possible textual inputs, $\mathcal{C}_{\text{ethos}} = \{\textit{support}, \textit{attack}, \textit{neutral}\}$, and $\mathcal{C}_{\text{pathos}} = \{\textit{positive}, \textit{negative}, \textit{neutral}\}$. Since these label spaces are closely aligned, we unify them into a common scheme $\{+, -, 0\}$, respectively. Ternary labeling reflects a principled trade-off between label granularity and reliability. While binary labeling would not cover the differentiation in non-neutral cases, increasing the number of labels is well known to reduce reliability.

For each sentence s and each interpretation i in the interpretation-attitude pairs associated with s, denoted as $I_s = \{(i_1, a_1), ..., (i_N, a_N)\}$, with N the number of pairs, we assign:

$$et_s = \phi_{\text{ethos}}(s), \qquad pa_s = \phi_{\text{pathos}}(s),$$
$$et_{i_n} = \phi_{\text{ethos}}(i_n), \qquad pa_{i_n} = \phi_{\text{pathos}}(i_n)$$

as the ethos and pathos labels for the sentence s and each interpretation i_n, respectively.

3.2.3 Experimental Setup

We test with models finetuned on the respective sentence classification tasks and a pretrained large language model (LLM) for parameterizing functions ϕ_{ethos} and ϕ_{pathos}.

Finetuned classifiers Following Gajewska and Budzynska (2024), we establish baseline finetuned models for classifying ethos and pathos. To that end, we use the PolarIs dataset (Gajewska et al., 2024). It contains 15,588 sentences sourced from Reddit and X on topics such as COVID-19 vaccines, climate change, and U.S. presidential elections. Each sentence had been manually annotated by experts in rhetoric with ternary labels for ethos and pathos appeals. The label distribution for ethos is $\{+ : 854; - : 2,610; 0 : 12,124\}$ and for pathos $\{+ : 895; - : 4,329; 0 : 10,365\}$. Inter-annotator agreement, measured with Cohen's κ between an expert and trained annotators, indicates substantial agreement for ethos ($\kappa = 0.78$) and moderate agreement for pathos ($\kappa = 0.53$).

Fine-tuning is performed on 60% of the dataset (training set). We conduct a random search across model architectures (BERT and RoBERTa) and key hyperparameters, including learning rate (1e-5–3e-4), batch size ([8, 16, 32, 64]), and number of training epochs (2-6). The best model for each classification task is selected based on F1 scores on the validation set (20% of PolarIs data), i.e., RoBERTa-large (ethos) and RoBERTa-base (pathos). Hyperparameters configurations for reproducing the classification results are included in Appendix A.

LLM prompting Gemini 2.5 (Comanici et al., 2025) independently performs ethos and pathos classification through in-context learning. For each classification task, we formulate three prompts that establish a zero-shot, few-shot, or chain-of-though (CoT) prompting setting. The prompting templates can be found in Appendix B.

3.2.4 Classification Results for OrigamIM

Neutral labels dominate for both ethos and pathos, followed by negative and positive. This trend is consistent across both classification methods (see Figure 1).

3.2.5 Evaluation of Label Reliability

We assess the reliability of the silver-standard ethos and pathos labels for the OrigamIM dataset *automatically* and *manually*.

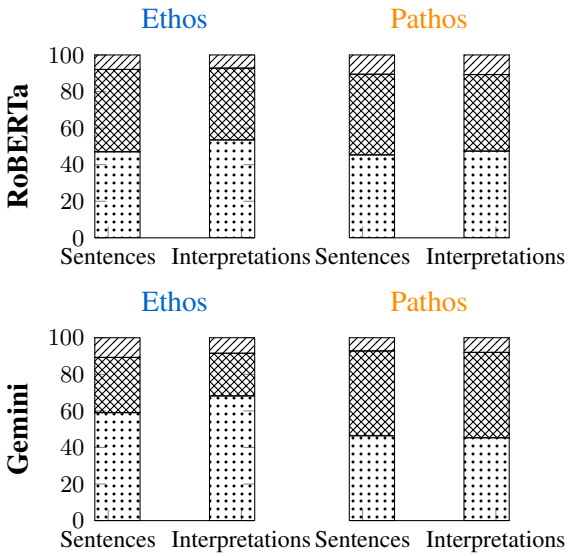


Figure 1: Ethos (left) and Pathos (right) label distributions in OrigamIM (percentages), with neutral (0, dots), negative (−, crosshatches), and positive labels (+, lines). The panel on top show the predicted distributions for RoBERTa, the panel at the bottom those for Gemini.

Automatic evaluation on gold standard external data We evaluate classification performance on two existing ethos/pathos datasets: PolarIs (held-out test set; three labels; 20% of the dataset) (Gajewska et al., 2024) and EU DisinfoTest (Sosnowski et al., 2024) (two labels). The latter dataset contains 1,344 narratives, which span a single or multiple sentences, on a variety of disinformation-sensitive topics such as migration and health. It provides ethos and pathos-heavy reformulations of the narratives produced by GPT-4. Ethos and pathos labeling is therefore binary, i.e., ethos/non-ethos and pathos/non-pathos, with 1,344 samples in each class. Likewise, we assess performance on PolarIs in a binary labeling setting, for which we aggregate the positive and negative labels into a joint class. This allows us to analyze the domain transferability of the classifiers. It is important to note that annotation in the PolarIs dataset regards others-directed ethos, while in EU DisinfoTest it regards self-referential ethos, which might lead to systematic performance differences due to the distinct pragmatic realizations of credibility cues across these datasets.

Table 1 reports the precision, recall, and F1-scores (mean and standard deviation) over three runs. Performance of Gemini closely aligns with

Model	Precision	Recall	F1
	PolarIs - 3 labels		
RoBERTa	$0.91_{\pm 0.01}$	$0.91_{\pm 0.01}$	$0.91_{\pm 0.01}$
Gemini 2.5	$0.79_{\pm 0.01}$	$0.77_{\pm 0.01}$	$0.78_{\pm 0.01}$
	PolarIs - 2 labels		
RoBERTa	$0.93_{\pm 0.01}$	$0.93_{\pm 0.01}$	$0.93_{\pm 0.01}$
Gemini 2.5	$0.80_{\pm 0.01}$	$0.80_{\pm 0.01}$	$0.80_{\pm 0.01}$
	EU DisinfoTest - 2 labels		
RoBERTa	$0.79_{\pm 0.01}$	$0.56_{\pm 0.01}$	$0.64_{\pm 0.01}$
Gemini 2.5	$0.87_{\pm 0.05}$	$0.49_{\pm 0.05}$	$0.57_{\pm 0.05}$

(a) Ethos classification.

Model	Precision	Recall	F1
	PolarIs - 3 labels		
RoBERTa	$0.88_{\pm 0.01}$	$0.88_{\pm 0.01}$	$0.88_{\pm 0.01}$
Gemini 2.5	$0.67_{\pm 0.00}$	$0.64_{\pm 0.01}$	$0.64_{\pm 0.01}$
	PolarIs - 2 labels		
RoBERTa	$0.90_{\pm 0.01}$	$0.90_{\pm 0.01}$	$0.90_{\pm 0.01}$
Gemini 2.5	$0.69_{\pm 0.00}$	$0.68_{\pm 0.01}$	$0.67_{\pm 0.01}$
	EU DisinfoTest - 2 labels		
RoBERTa	$0.85_{\pm 0.01}$	$0.50_{\pm 0.01}$	$0.54_{\pm 0.01}$
Gemini 2.5	$0.86_{\pm 0.00}$	$0.47_{\pm 0.01}$	$0.51_{\pm 0.01}$

(b) Pathos classification.

Table 1: **Automatic** validation of ethos/pathos label reliability: Performance on ethos (top) and pathos (bottom) classification across datasets.

high inter-annotator agreement for ethos ($\kappa = 0.78$) and a moderate agreement for pathos ($\kappa = 0.53$) reported for PolarIs, reflecting the greater interpretive ambiguity of emotional appeals. Fine-tuned RoBERTa shows a comparable performance for both classification tasks (F1=0.88-0.91), suggesting that in-domain training allows the model to approximate expert-level consistency on PolarIs. The performance drop in a domain-transfer condition indicates that these results partly reflect dataset-specific regularities. However, this result is expected since different annotation guidelines have been used for the two datasets, so the models are likely challenged by differences in perspective (self- vs. other-directed ethos), rhetorical framing, and topic distribution.

Analysis of inconsistent classification by the two approaches in EU DisInfo indicates that differences in perspective are primarily responsible for this disparate performance. Let's consider the following examples:

(5) Nobel Laureate Joseph Stiglitz has expressed concerns that the euro restricts economic policy options for member states, suggesting that countries lose economic autonomy upon adopting the euro. Moreover, the International Monetary Fund has documented persistent economic struggles in countries post-euro adoption, indicating a trend towards economic decline.

(6) The euro and Eurozone policies have been a failure, leading to increased economic disparities and hardships for member countries.

The divergence in classification stems from a fundamental mismatch in how the two models operationalize the ethos dimension based on their respective training sets. The model trained on the PolarIs dataset is optimized to recognize other-directed ethos appeals, such as the explicit reputation attacks on institutional policies seen in example (6)[3]. Conversely, the model exposed to the EU DisInfo guidelines is calibrated to capture instances where authority is outsourced to external entities, specifically flagging arguments from expert opinion like the citations of a Nobel Laureate in example (5)[4]. Consequently, when deployed across domains, the models fail to generalize because they are searching for fundamentally different rhetorical mechanisms: one tracks direct interpersonal or institutional confrontation, while the other tracks the structural integration of external credibility. This result confirms that the domain-transfer degradation is a structural misalignment in perspective modeling rather than mere lexical noise.

Manual evaluation on gold standard OrigamIM data We further validate label reliability *manually* on a randomly selected subset of 100 sentences and interpretations from OrigamIM. Two in-house expert annotators independently establish gold standard ethos and pathos labels for these samples, for which they follow the annotation guidelines used for constructing PolarIs (annotation details in Appendix C). Agreement between the two raters has been achieved in 71% of cases for ethos ($\kappa = 0.5$) and in 69% of cases for pathos ($\kappa = 0.4$). Note that these κ values likely underestimate true inter-annotator reliability, as several label categories are sparsely represented in this small sample (some occurring only 5-10 times), which is known to lower Cohen's κ. Due to the high cost and time requirements of expert annotation, we opted for a focused

[3] "Individuals, groups or organizations that participate in a discussion or are just a third party that is mentioned in the exchanges could be targets of such appeals." (Gajewska et al., 2024, p. 2)

[4] "Ethos is used to convey the writer's credibility and authority." (Sosnowski et al., 2024, p. 14718)

Model	Precision	Recall	F1
OrigamIM - 3 labels			
RoBERTa	0.75	0.63	0.64
Gemini 2.5	0.76	0.71	0.73

(a) Ethos classification.

Model	Precision	Recall	F1
OrigamIM - 3 labels			
RoBERTa	0.73	0.66	0.67
Gemini 2.5	0.75	0.68	0.70

(b) Pathos classification.

Table 2: **Manual** validation of ethos/pathos label reliability: Performance on ethos (top) and pathos (bottom) classification on a subset of the OrigamIM dataset.

manual evaluation between the two individuals. To obtain the final labels, disagreements were resolved using raters' confidence levels for individual cases.

Table 2 reports model performance on the gold-standard subset of OrigamIM. For ethos, Gemini 2.5 outperforms RoBERTa (F1 = 0.73 vs. 0.64), mainly due to higher recall, indicating better coverage of credibility-related cues in this setting. For pathos, both models achieve comparable performance, with a modest advantage of Gemini 2.5 (F1 = 0.70 vs. 0.67), suggesting a slightly more robust alignment with human judgments on emotional appeals. Eventually, we combine RoBERTa and Gemini decisions through a majority voting scheme for all downstream data analyses.

4 Analysis

This section analyzes the OrigamIM data to examine how ethos and pathos expressed in an original sentence are reflected in audience interpretations.

RQ1: Are ethos and pathos in the original sentence reflected differently in audience interpretations?

When comparing the labels between sentence and interpretation, we observe that the interpretations do not consistently align with the original sentence in terms of ethos (i.e., $et_s \neq et_{i_n}$) and pathos (i.e., $pa_s \neq pa_{i_n}$). Only 74.2% and 70.4% of interpretations match the original sentence's ethos and pathos labels, respectively. Full alignment, where all interpretations of a given sentence reflect both the same ethos and pathos labels as the original, occurs in just 15.9% of the sentences. By contrast, complete divergence, where all interpretations differ from the original sentence in both dimensions, is rare, appearing in only one case. Figure 2 illustrates examples of full alignment and complete divergence. These findings show that rhetorical signals in a sentence are often perceived differently, highlighting the interpretive variability of ethos and pathos among the audience.

RQ2: Which ethos and pathos labels trigger greater variability in audience interpretations?

Original sentences with neutral ethos ($et_s = 0$) and pathos ($pa_s = 0$) are associated with significantly higher alignment in audience interpretations compared to non-neutral cases, with statistical significance at $p < 0.001$ for ethos and $p < 0.001$ for pathos (tested using the χ^2 statistic); see Figure 3a. This suggests that rhetorical neutrality facilitates consensus among readers. Sentences expressing ethos support ($et_s = +$) and positive pathos ($pa_s = +$) show the highest variability (see Figure 3b). We further observe that the same ethos and pathos label is expressed in all interpretations for only 36% and 28% of sentences, respectively. These findings indicate that rhetorically charged content tends to elicit a broader range of audience interpretations, underscoring the interpretive flexibility and complexity inherent in emotionally loaded discourse.

Predictor	β	SE	t	p
Intercept	1.54	0.02	87.37	***
[Pathos: +]	0.30	0.05	6.02	***
[Pathos: -]	0.11	0.03	4.50	***
Intercept	1.35	0.02	92.55	***
[Ethos: +]	0.46	0.05	10.26	***
[Ethos: -]	0.32	0.03	12.41	***

Table 3: OLS regression predicting **variability in interpretation labels from sentence labels**. Neutral labels are set as baselines. Statistical significance: *** ($p < .001$).

Table 3 presents the results of two OLS regression models predicting variability in interpretation labels from sentence-level rhetorical appeals. The intercepts represent the expected level of interpretive variability for sentences without marked pathos or ethos appeals. In the first model, pathos appeals are significantly associated with increased variability in sentence interpretations. Sentences containing positive pathos exhibit higher interpretive vari-

[*Complete Divergence*]

Sentence:
[Ethos: +] [Pathos: +]
It's a celebration of the white spirit and the lack of racism in so many white people.

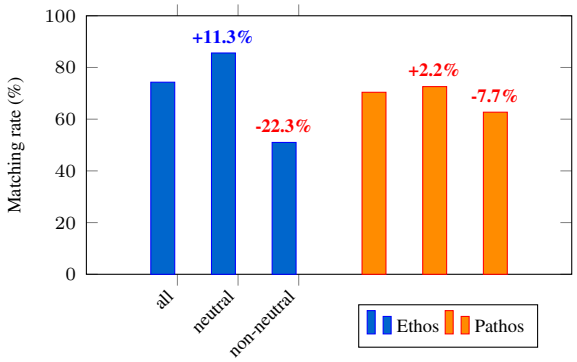
Interpretations:
[Ethos: -] [Pathos: -] The author is making the frankly bizarre claim that Juneteenth is a holiday about how some white people aren't racist.
[Ethos: 0] [Pathos: 0] People shouldn't be angry at Juneteenth.
[Ethos: 0] [Pathos: -] Many whites celebrate the enthusiasm of white spirit with no regards to racism.
[Ethos: 0] [Pathos: 0] Juneteenth celebrates the role white people played in the emancipation of American slaves.
[Ethos: 0] [Pathos: -] It's a celebration of the lack of racism white people.

[*Full Alignment*]

Sentence:
[Ethos: 0] [Pathos: 0]
They are essentially starting at $0.

Interpretations:
[Ethos: 0] [Pathos: 0] The writer states that people who do not have student debt or a degree are starting at a worse point than people with student debt and a degree.
[Ethos: 0] [Pathos: 0] Compensation for payments towards debt would be neutral.
[Ethos: 0] [Pathos: 0] People who clear their student loans have no money left with them at that point.
[Ethos: 0] [Pathos: 0] Their loan has been cancelled.
[Ethos: 0] [Pathos: 0] They are starting at $0.

Figure 2: Examples of complete divergence (*left*) and full alignment (*right*) in ethos and pathos labels between the original sentence (*top*, in italics) and corresponding audience interpretations (*bottom*).



(a) Matching rates between sentence and interpretation based on the sentence label for ethos (left) and pathos (right).

Label for s	# Labels in I_s	% [Ethos]	% [Pathos]
0	1 / 2 / 3	**67.2** / 31.0 / 1.7	**50.4** / 45.7 / 3.9
−	1 / 2 / 3	30.3 / **59.2** / 10.5	27.0 / **62.0** / 10.9
+	1 / 2 / 3	35.5 / **62.3** / 2.2	37.4 / **60.1** / 2.5

(b) Share of interpretation sets with 1, 2, or 3 distinct ethos/-pathos labels, conditioned on the original sentence label

Figure 3: Overview: (a) Matching rates by neutrality; (b) Distribution of number of labels in interpretations.

ability compared to neutral sentences ($\beta = 0.30$, $p < .001$), indicating that emotionally positive content elicits more divergent audience interpretations.

Notably, negative pathos is also associated with a significant increase in variability ($\beta = 0.11$, $p < .001$), although the magnitude of this effect is smaller. The second model focuses on ethos appeals and reveals even stronger effects. Both positive and negative ethos are associated with statistically significant increases in interpretive variability relative to neutral sentences: positive ($\beta = 0.46$, $p < .001$) and negative ($\beta = 0.32$, $p < .001$). This pattern indicates that credibility-related cues, whether reinforcing or undermining someone's authority, markedly amplify disagreement or divergence in sentence interpretations.

RQ3: Can ethos and pathos in the original sentence predict attitudes toward speakers?

We examine whether ethos and pathos in the sentence, i.e., et_s and pa_s, can predict a reader's attitude towards the author, a_i. To that end, linear regression tested the effects of ethos and pathos on audience attitude ratings, treating attitudes as continuous numerical values. The model is significant for both pathos ($F(2, 9848) = 72.34, p < .001$) and ethos ($F(2, 9846) = 25.29, p < .001$). Table 4 indicate that sentences with positive pathos/ethos elicit more positive attitudes than those with neutral pathos/ethos. On the other hand, sentences with negative pathos/ethos trigger more negative attitudes than those with neutral pathos/ethos. The linear regression models for predicting *variability* in reader attitudes within the interpretation set accompanying a sentence, I_s, using sentence-level ethos and pathos are not statistically significant. Note that the limited size of I_s potentially biases attitude variability and the statistical findings regarding attitude variability.

Predictor	β	SE	t	p
Intercept	3.02	0.01	208.55	***
[Pathos: +]	0.37	0.04	8.89	***
[Pathos: -]	-0.12	0.02	-5.78	***
Intercept	3.00	0.01	234.26	***
[Ethos: +]	0.18	0.04	4.62	***
[Ethos: -]	-0.10	0.02	-4.46	***

Table 4: OLS regression predicting **audience attitudes from sentence-level pathos and ethos appeals**. Neutral labels are set as baselines. Statistical significance: *** ($p < .001$).

RQ4: Which factors beyond ethos and pathos in the sentence influence interpretation variability?

We manually analyze 100 randomly sampled cases, where at least one label differs between the interpretation and the sentence plus all cases, where all three ethos and pathos labels ({+, -, 0}) are present in a set of interpretations: there are 51 such sentences for ethos and 76 for pathos. Our analysis reveals eight factors that may explain why audience interpretations of the same sentence differ, alongside ethos and pathos (Table 5). These factors could be seen as interpreter-related, content- and form-based, and contextual factors.

Interpreter-related factors: These arise from the background, beliefs, and perspectives of the interpreter rather than from the sentence itself. This group includes *value system differences*, *political partisanship*, *social identity/position*, and *personal experience*, all of which reflect how readers project their own moral frameworks, identities, and lived experiences onto the content.

Content- and form-based factors: These are driven by linguistic or rhetorical properties of the sentence that invite multiple readings. This category covers *author opinion* (e.g., first- vs. third-person framing), *rhetorical questions*, and *sarcasm*, where stylistic or pragmatic cues alter perceived commitment, intent, or emotional tone.

Contextual factors: They stem from missing or underspecified contextual information. Particularly, *out-of-context* effects lead interpreters to infer referents, stance, or targets differently, resulting in divergent ethos and pathos assignments.

Extended definitions of these factors can be found in Appendix E.

5 Discussion

Our results align with theoretical accounts that emphasize mechanisms sustaining interpretive divides in digital communication. The notion of *echo chambers*, for instance, explains how selective exposure to ideologically homogeneous content reinforces prior beliefs and limits engagement with opposing views (Garimella et al., 2018; Del Vicario et al., 2016). Within such environments, ethos cues may be readily accepted when they align with in-group norms but dismissed or inverted when attributed to out-group speakers.

Furthermore, the observed relation between rhetorical charge and interpretation variance suggests a testable hypothesis: content with high rhetorical load may be more prone to divergent interpretations and conflict. This distinction raises the possibility of investigating "interpretive instability" as a separate dimension from traditional toxicity that triggers hostility. Such an approach could deepen our understanding of how persuasion interacts with audience cognition and inform research on the dynamics of conflict and consensus in online discourse.

These findings present critical implications for the computational modeling of discourse and pragmatic interpretation in multi-party interaction. By demonstrating a 30% divergence between original text and reader interpretations, this work challenges the widespread assumption that textual content can serve as a direct proxy for audience reception. Instead, the heightened variability observed in rhetorically charged content indicates that ground truth in communication data is inherently unstable, particularly when dealing with persuasive appeals like ethos and pathos. For the NLP community, this means that optimizing language models purely on semantic similarity metrics ignores a massive chasm in actual human communication. To model social phenomena like polarization, we must pivot from modeling what the text says to modeling the distribution of what the text evokes. This challenges the field to develop a new class of receiver-centric pragmatic models.

6 Conclusion

The paper presented a computational study on the transformation of rhetorical appeals (specifically ethos and pathos) within the interpretive space of social media discourse. Rather than viewing rhetoric as a static property of the text, we quanti-

Example sentence	Category	# Cases	% Cases
Sure, I do disagree with her on a lot but I think she's reasonable enough to work with people and not accuse them of being a Russian asset.	**Author opinion**	64	28.6
I've heard Christians talk about how it's actually good when someone's uncomfortable: it means that they're getting closer to actually having a breakthrough of faith or realizing the error in their non-christian ways.	**Value system**	44	19.6
Hell, Trump's done a lot of damage to the USA's reputation just because he tweets garbage non-stop.	**Political partisanship**	36	16.1
Black Lives Matter (as practiced by white people) DOES imply that white lives DON'T matter as much.	**Social identity/position**	30	13.4
She gets paid every month for something, and it's never going to end.	**Out of context**	20	8.9
Where are the federal agents going door to door telling people to not smoke tobacco and avoid high sugar foods?	**Rhetorical question**	16	7.1
I had discussions with some black people and (for lack of a better term) "woke" white people I know.	**Personal experience**	10	4.5
but I promise this isn't a 'white people are the real victims' rant and there will be new questions in here.	**Sarcasm**	4	1.8

Table 5: Factors influencing variability of ethos and pathos labels between sentences and interpretations.

fied the extent to which these modes of persuasion are preserved or distorted during the interpretation process. Our methodology relied on a statistical examination of the alignment between source and target rhetorical labels, employing regression analyses to isolate the specific effects of credibility and emotional appeals on interpretive variability and the formation of audience attitudes.

Our findings show that rhetorical appeals are significant predictors of interpretive divergence: while rhetorical neutrality fosters consensus, marked ethos and pathos appeals actively catalyze disagreement in meaning construction. Furthermore, we demonstrated that these appeals can predict reader attitudes toward the author, highlighting the direct link between rhetorical strategy and social perception. By disentangling the linguistic and cognitive factors that mediate this variation, such as value systems and political affiliation, this work advances a perspectivist approach to analyzing rhetorical influence.

While this study establishes a baseline for modeling rhetorical interpretation within the silent audience, it also opens several avenues for future empirical investigation. First, while we isolate ethos and pathos as primary drivers of interpretative variance, future work expands the feature space to account for potentially correlated alternative dimensions. Specifically, we plan to disentangle rhetorical appeals from confounding variables such as ideological alignment, text complexity, and implicit topic bias, which may co-vary with ethos and pathos and contribute to the observed semantic divergence. Second, a vital next step is to bridge the gap between the latent cognitive states of the silent audience and the overt behaviors of active users. We intend to collect larger datasets that pair human written interpretations with platform-level engagement signals, such as upvotes, shares, and comment counts. Cross-referencing these behavioral metrics with our interpretation-level rhetorical patterns will allow us to determine whether the subtle shifts in audience perception we uncovered correlate with or diverge from high engagement digital discourse.

Limitations

The design of the study constrains the scope and robustness of its findings.

The analyses rely exclusively on the OrigamIM dataset, which contains English sentences collected from Reddit posts from the r/ChangeMyView subreddit and pairs each sentence with a set of five human-written interpretations. The single-language property of the sentences limits the generalizability of results to other languages. Moreover, the size of the interpretation set is too small to fully capture all possible audience interpretations of a given sentence. For this, a much larger set coming from a diverse group of readers is necessary to approximate true variation in sentence interpretations, which is, to our knowledge, non-existing.

Nonetheless, we contend that our findings possess a high degree of generalizability despite the data originating from a specific subreddit. That is because the underlying architecture of most social media platforms is fundamentally the same: they serve as digital arenas for the expression and exchange of personal opinions. By using the core feature of source opinions and their subsequent interpretations from Reddit, we are analyzing a basic

communicative unit, i.e., the message and its reception, that is a universal characteristic of social media interaction. While platform-specific norms vary, the cognitive mechanisms involved in processing rhetorically charged content, such as the projection of value systems or partisan filters, are structural features of human communication that transcend any single platform.

Ethos and pathos classification depends on fine-tuned RoBERTa-based models, which, despite reasonable performance, are subject to noise and moderate error rates inherent in subjective labeling tasks. The downstream analyses are based on statistically significant patterns rather than individual predictions such that observed trends reflect meaningful rhetorical phenomena rather than random noise. Nonetheless, the predicted labels should be interpreted as model-predicted rhetorical signals rather than absolute truths.

Ethical Considerations

This research involves analysis of social media text, raising several ethical considerations. First, the dataset used (OrigamIM) is publicly available and licensed for research (CC BY-SA 4.0), used solely according to intended purposes with consent from the dataset creators, thereby respecting data privacy and licensing norms. The subjective nature of rhetorical interpretation, influenced by cultural background and prior beliefs, may lead to overgeneralization or exclusion of minority viewpoints if models are deployed without careful contextualization. Furthermore, findings concerning polarization effects could be misapplied for manipulative purposes such as targeted disinformation. To mitigate harm, the work emphasizes interpretive variability and perspectivism, cautioning against simplistic or reductionist deployment of classification results.

Acknowledgments

This work has been funded in part by the Research Foundation - Flanders (FWO) under grant G0L0822N through the CHIST-ERA project "iTRUST Interventions against Polarisation in Society for Trustworthy Social Media. From Diagnosis to Therapy", in part by the National Science Centre, Poland (Chist-Era IV) under grant 2022/04/Y/ST6/00001, and in part by the National Science Centre, Poland under grant 2025/57/N/HS1/00480. Liesbeth Allein is further supported by a junior postdoctoral fellowship from the FWO under grant 12AGW26N.

References

Liesbeth Allein and Marie-Francine Moens. 2024. OrigamIM: A dataset of ambiguous sentence interpretations for social grounding and implicit language understanding. In *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*, pages 116–122, Torino, Italia. ELRA and ICCL.

Liesbeth Allein, Maria Mihaela Truşcă, and Marie-Francine Moens. 2025. Interpretation modeling: Social grounding of sentences by reasoning over their implicit moral judgments. *Artificial Intelligence*, 338:104234.

Aristotle. 1991. *The Art of Rhetoric*. Penguin Books, London, U.K.

Christopher Bagdon, Aidan Combs, Carina Silberer, and Roman Klinger. 2025. Donate or create? comparing data collection strategies for emotion-labeled multimodal social media posts. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17307–17330, Vienna, Austria. Association for Computational Linguistics.

Jürgen Buder, Lisa Rabl, Markus Feiks, Mandy Badermann, and Guido Zurstiege. 2021. Does negatively toned language use on social media lead to attitude polarization? *Computers in Human Behavior*, 116:106663.

Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a perspectivist turn in ground truthing for predictive computing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 6860–6868.

Guizhen Chen, Liying Cheng, Anh Tuan Luu, and Lidong Bing. 2024. Exploring the potential of large language models in computational argumentation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2309–2330, Bangkok, Thailand. Association for Computational Linguistics.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2019. Fine-grained analysis of propaganda in news articles. 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

5636–5646, Hong Kong, China. Association for Computational Linguistics.

Aida Davani, Mark Díaz, Dylan Baker, and Vinodkumar Prabhakaran. 2024. Disentangling perceptions of offensiveness: Cultural and moral correlates. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 2007–2021.

Michela Del Vicario, Gianna Vivaldo, Alessandro Bessi, Fabiana Zollo, Antonio Scala, Guido Caldarelli, and Walter Quattrociocchi. 2016. Echo chambers: Emotional contagion and group polarization on facebook. *Scientific reports*, 6(1):37825.

Rory Duthie, Katarzyna Budzynska, and Chris Reed. 2016. Mining ethos in political debate. *Computational Models of Argument*, pages 299–310.

Alice H Eagly. 2014. Recipient characteristics as determinants of responses to persuasion. In *Cognitive responses in persuasion*, pages 173–195. Psychology Press.

Natalia Evgrafova, Veronique Hoste, and Els Lefever. 2024. Analysing pathos in user-generated argumentative text. In *Proceedings of the Second Workshop on Natural Language Processing for Political Sciences @ LREC-COLING 2024*, pages 39–44, Torino, Italia. ELRA and ICCL.

Eve Fleisig, Su Lin Blodgett, Dan Klein, and Zeerak Talat. 2024. The perspectivist paradigm shift: Assumptions and challenges of capturing human labels. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2279–2292.

Simona Frenda, Gavin Abercrombie, Valerio Basile, Alessandro Pedrani, Raffaella Panizzon, Alessandra Teresa Cignarella, Cristina Marco, and Davide Bernardi. 2025. Perspectivist approaches to natural language processing: a survey. *Language Resources and Evaluation*, 59(2):1719–1746.

Ewelina Gajewska and Katarzyna Budzynska. 2024. Digital polarisation: Analysis of 'us' versus 'them' rhetoric in public discussions on social media. In *Progress in Polish Artificial Intelligence Research 5. Proceedings of the 5th Polish Conference on Artificial Intelligence (PP-RAI'2024)*, Warsaw, Poland. Warsaw University of Technology. Conference held April 18–20, 2024.

Ewelina Gajewska, Katarzyna Budzynska, Barbara Konat, Marcin Koszowy, Konrad Kiljan, Maciej Uberna, and He Zhang. 2024. Ethos and pathos in online group discussions: Corpora for polarisation issues in social media. *arXiv preprint arXiv:2404.04889*.

Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2018. Political discourse on social media: Echo chambers, gatekeepers, and the price of bipartisanship. In *Proceedings of the 2018 world wide web conference*, pages 913–922.

Max Glockner, Yufang Hou, Preslav Nakov, and Iryna Gurevych. 2025. Grounding fallacies misrepresenting scientific publications in evidence. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 9732–9767, Albuquerque, New Mexico. Association for Computational Linguistics.

Pierpaolo Goffredo, Shohreh Haddadan, Vorakit Vorakitphan, Elena Cabrio, and Serena Villata. 2022. Fallacious argument classification in political debates. In *Thirty-First International Joint Conference on Artificial Intelligence (IJCAI-22)*, pages 4143–4149. International Joint Conferences on Artificial Intelligence Organization.

Nitesh Goyal, Ian D Kivlichan, Rachel Rosen, and Lucy Vasserman. 2022. Is your toxicity my toxicity? exploring the impact of rater identity on toxicity annotation. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–28.

Ivan Habernal, Raffael Hannemann, Christian Pollak, Christopher Klamm, Patrick Pauli, and Iryna Gurevych. 2017. Argotario: Computational argumentation meets serious games. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 7–12, Copenhagen, Denmark. Association for Computational Linguistics.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018. Before name-calling: Dynamics and triggers of ad hominem fallacies in web argumentation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 386–396, New Orleans, Louisiana. Association for Computational Linguistics.

Zhijing Jin, Abhinav Lalwani, Tejas Vaidhya, Xiaoyu Shen, Yiwen Ding, Zhiheng Lyu, Mrinmaya Sachan, Rada Mihalcea, and Bernhard Schoelkopf. 2022. Logical fallacy detection. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7180–7198, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Noah Lee, Na Min An, and James Thorne. 2023. Can large language models capture dissenting human voices? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4569–4585, Singapore. Association for Computational Linguistics.

Elena Musi, Myrto Aloumpi, Elinor Carmi, Simeon Yates, and Kay O'Halloran. 2022. Developing fake news immunity: fallacies as misinformation triggers

during the pandemic. *Online Journal of Communication and Media Technologies*, 12(3).

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Chaïm Perelman and Lucie Olbrechts-Tyteca. 1971. Traité de l'argumentation. la nouvelle rhétorique. In *Collection de sociologie générale et de philosophie sociale. Éditions de l'Institut de Sociologie de Bruxelles, 2e éd*, volume 76. Revue de Métaphysique et de Morale.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682.

Alan Ramponi, Agnese Daffara, and Sara Tonelli. 2025a. Fine-grained fallacy detection with human label variation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 762–784, Albuquerque, New Mexico. Association for Computational Linguistics.

Alan Ramponi, Agnese Daffara, and Sara Tonelli. 2025b. Fine-grained fallacy detection with human label variation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 762–784.

Marta Sandri, Elisa Leonardelli, Sara Tonelli, and Elisabetta Jezek. 2023. Why don't you do it right? analysing annotators' disagreement in subjective tasks. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2428–2441, Dubrovnik, Croatia. Association for Computational Linguistics.

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2021. "nice try, kiddo": Investigating ad hominems in dialogue responses. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 750–767, Online. Association for Computational Linguistics.

Taylor Sorensen, Liwei Jiang, Jena D Hwang, Sydney Levine, Valentina Pyatkin, Peter West, Nouha Dziri, Ximing Lu, Kavel Rao, Chandra Bhagavatula, and 1 others. 2024. Value kaleidoscope: Engaging ai with pluralistic human values, rights, and duties. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19937–19947.

Witold Sosnowski, Arkadiusz Modzelewski, Kinga Skorupska, Jahna Otterbacher, and Adam Wierzbicki. 2024. EU DisinfoTest: a benchmark for evaluating language models' ability to detect disinformation narratives. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14702–14723, Miami, Florida, USA. Association for Computational Linguistics.

Maria Mihaela Trusca and Liesbeth Allein. 2025. Mimicking how humans interpret out-of-context sentences through controlled toxicity decoding. In *Proceedings of the 5th Workshop on Trustworthy NLP (TrustNLP 2025)*, pages 291–297, Albuquerque, New Mexico. Association for Computational Linguistics.

Douglas Walton. 2010. Why fallacies appear to be better arguments than they are. *Informal logic*, 30(2):159–184.

Leon Weber-Genzel, Siyao Peng, Marie-Catherine De Marneffe, and Barbara Plank. 2024. VariErr NLI: Separating annotation error from human label variation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2256–2269, Bangkok, Thailand. Association for Computational Linguistics.

Michael Wiegand, Jana Kampfmeier, Elisabeth Eder, and Josef Ruppenhofer. 2023. Euphemistic abuse – a new dataset and classification experiments for implicitly abusive language. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16280–16297, Singapore. Association for Computational Linguistics.

A Experimental Setup & Reproducibility

Ethos and pathos classification Table 6 includes all model and finetuning hyperparameters for reproducing the ethos and pathos classification. PolarIs dataset[5] is a collection of Reddit and Twitter discussions, centered around three topics: COVID-19 vaccines, climate change, and the 2016 U.S. elections. These social media posts were then split into individual sentences using the spaCy library. In total, it comprises 15,588 sentences annotated with appeals to ethos and pathos. Pre-processing involved the transformation of Twitter usernames into a common "@user" tag. Since the data is highly imbalanced (80% of sentences are $et_s = 0$, and 69% are $pa_s = 0$), we undersample the neutral category to constitute 50% of labels for both ethos and pathos. For the purpose of this study, we sample 60% of the data for training, 20% for testing, and 20% for validation in a label-stratified manner. All experiments, including finetuning of the RoBERTa models for ethos and pathos classification, were run on a NVIDIA Tesla T4 GPU available in Google Colab.

[5] https://osf.io/xjbaw/

Hyperparameter	Value
Ethos classification	
Model name	RoBERTa Large
Model URL	https://huggingface.co/FacebookAI/roberta-large
# Layers	24
Hidden size	1024
# Attention heads	16
# Model parameters	355M
Loss function	Cross-entropy loss
Optimizer	Adam
Learning rate	2e-5
# Epochs	6
Batch size	32
Pathos classification	
Model name	RoBERTa Base
Model URL	https://huggingface.co/FacebookAI/roberta-base
# Layers	12
Hidden size	768
# Attention heads	12
# Model parameters	125M
Loss function	Cross-entropy loss
Optimizer	Adam
Learning rate	4e-5
# Epochs	5
Batch size	32

Table 6: Hyperparameter settings for finetuning the ethos and pathos classifiers.

Interpretations All analyses were conducted over the entire OrigamIM dataset, for which we combined the train, validation, and test set provided by its creators.

Evaluation and analysis All analyses were conducted in the Python programming language, and specifically the Sklearn library for calculating classification performance metrics[6], Statsmodels for executing the linear regression model[7], Scipy for statistical testing[8], and Transformers for model finetuning[9].

Regression analyses with Gemini classification Below we present the results of two regression analyses with an alternative classification approach to ethos and pathos appeals, i.e. with the Gemini model (Tables 7 and 8).

[6]https://scikit-learn.org/stable/
[7]https://www.statsmodels.org/stable/index.html
[8]https://scipy.org/
[9]https://pypi.org/project/transformers/

Predictor	β	SE	t	p
Intercept	1.54	0.02	79.33	***
[Pathos: +]	0.21	0.05	4.17	***
[Pathos: -]	0.06	0.03	2.29	**
Intercept	1.33	0.02	86.91	***
[Ethos: +]	0.46	0.04	11.93	***
[Ethos: -]	0.37	0.03	13.98	***

Table 7: OLS regression predicting **variability in interpretation labels from sentence labels** with Gemini. Neutral labels are set as baselines. Statistical significance: *** ($p < .001$), ** ($p < .01$).

Predictor	β	SE	t	p
Intercept	3.03	0.02	193.94	***
[Pathos: +]	0.37	0.04	9.03	***
[Pathos: -]	-0.14	0.02	-6.57	***
Intercept	3.00	0.01	225.43	***
[Ethos: +]	0.14	0.03	4.13	***
[Ethos: -]	-0.10	0.02	-4.29	***

Table 8: OLS regression predicting **audience attitudes from sentence-level pathos and ethos appeals** with Gemini. Neutral labels are set as baselines. Statistical significance: *** ($p < .001$).

B Prompting Templates

Prompt templates are given on an example of ethos annotation; templates for pathos are formulated in an analogous manner.

Zero-shot. Your task is to identify ethos appeals in a list of sentences. Ethos appeals are references to a speaker's character or credibility. These references may be: • favorable (supporting someone's credibility or character), • unfavorable (attacking or undermining credibility or character), • non-ethos (no reference to character/credibility).

Few-shot. Your task is to identify ethos appeals in a list of sentences. Ethos appeals are references to a speaker's character or credibility. These references may be: • favorable (supporting someone's credibility or character), • unfavorable (attacking or undermining credibility or character), • non-ethos (no reference to character/credibility). Example of a favorable ethos appeal: Great initiative by MRIs. Example of a unfavorable ethos appeal: To me it's just a control/manipulation tactic by the government and elites. Example of a non-ethos appeal: A few months ago we had the worst flood our area has had in over 100 years.

Chain-of-thought. Your task is to identify ethos appeals in a list of sentences. Ethos appeals are references to a speaker's character or credibility. These references may be: • favorable (supporting someone's credibility or character) • unfavorable (attacking or undermining credibility or character) • non-ethos (no reference to character/credibility) Use step-by-step reasoning internally to identify ethos appeals and output the required labels. Below are three worked examples that illustrate the reasoning style.
#Example 1 (unfavorable ethos) Sentence: "To me it's just a control/manipulation tactic by the government and elites."
Demonstrated CoT: The sentence attacks the credibility and intentions of "the government and elites." → This is an unfavorable ethos appeal. Final output format: ID 1: unfavorable.
#Example 2 (favorable ethos) Sentence: "Great initiative by MRIs."
Demonstrated CoT: The sentence praises MRIs, implying they are doing something commendable. → This supports their credibility/character. Final output format: ID 2: favorable.
#Example 3 (non-ethos) Sentence: "A few months ago we had the worst flood our area has had in over 100 years."
Demonstrated CoT: The sentence reports an event, with no reference to character or credibility. → This is non-ethos. Final output format: ID 3: non-ethos.
Now process the following sentences using the same rules and output labels in the specified format.

C Manual Validation: Annotation Guidelines

The corpora are annotated at the sentence level for two rhetorical strategies from Aristotelian rhetoric: ethos (credibility appeals) and pathos (emotional appeals). The following guidelines come from the PolarIs dataset (Gajewska et al., 2024), which we use also for our manual verification of classification reliability in the target OrigamIM dataset. Here, the annotation is conducted on a randomly sampled 100 cases from the OrigamIM dataset and annotated independently by two trained raters. Both annotators are females with a previous experience in annotation studies and linguistic research, who voluntarily performed the task.

Ethos Annotation. The goal of ethos annotation is to identify sentences where a speaker references the credibility or character of an entity, either in a supportive or attacking way. The procedure for identifying such references is as follows. First, the rater should detect an ethos appeal if a sentence mentions an entity (a person/group/organization) and this entity is described in a favorable (support) or unfavorable (attack) manner. Second, the rater should mark the sentence as a support if the entity is mentioned in a favorable manner or as an attack if the entity is mentioned in an unfavorable manner.

Pathos Annotation. The goal of pathos annotation is to capture rhetorical appeals that aim to induce emotions in the hearer rather than simply express a sentiment. The procedure for identifying such appeals is as follows. First, the rater should determine whether the speaker's language intends rhetorical gain through emotional cues. Second, the rater should mark positive pathos when the speaker intended to induce positive emotional states in the hearers or negative pathos when the speaker intended to induce negative emotional states in the hearers. Linguistic signals for the identification fo pathos appeals include emotion-eliciting words, phrases, and rhetorical figures.

D Access to Existing and New Artifacts

Our analyses relied on the OrigamIM dataset (Allein and Moens, 2024), which is licensed under the CC BY-SA 4.0 license. We used OrigamIM for research purposes only, which is consistent with its intended use. We retrieved the entire dataset from a publicly accessible GitHub repository[10]. We will release the OrigamIM dataset with the labels for ethos and pathos inferred in this work with the same CC BY-SA 4.0 license in a public repository, for which we have oral consent from the creators of OrigamIM.

E Definitions of Variability Factors

Below we provide definitions of eight factors driving the variability between the sentence and interpretations beyond ethos and pathos labels.

> **Author opinion**: Sentences written in the first person are frequently rephrased into third-person statements, altering perceived speaker commitment and ethical stance reflected in the interpretations.

[10] https://github.com/laallein/origamIM

Value system differences: Interpreters impose their own moral or ideological priorities onto the sentence, leading the same content to be framed as either ethically responsible or ethically problematic.

Political partisanship: References to political actors (e.g., Donald Trump) are often interpreted through partisan lenses, resulting in polarity shifts such as praise being rephrased into criticism.

Social identity or social position: Mentions of race, gender or group membership activate different social meanings, affecting both perceived intent and emotional appeal.

Out-of-context effects: Sentences that are very short or lack the necessary contextual information to be understood on their own lead to interpreters having to infer the referent, stance or target of the sentence. This, in turn, may lead to divergent or incorrect assumptions about what or who the sentence is addressing.

Rhetorical questions: Rhetorical questions are frequently normalized into declarative claims, changing their persuasive force and associated ethos or pathos labels.

Personal experience: This is closely related to author opinion. However, personal experience specifically refers to cases where sentence authors invoke anecdotal or lived experiences to support a broader claim. While author opinion reflects a general personal stance, personal experience grounds that stance in a specific situation the author reports having encountered. Interpretations of such sentences vary depending on whether readers treat the anecdote as credible evidence, a subjective account, or a rhetorical device, which can lead to differences in how ethos and pathos are assigned.

Sarcasm: ironic or mocking intent is not consistently detected, resulting in marked disagreement in emotional and ethical labeling across interpretations.

F On the Use of AI Assistants in Research or Writing

AI assistants were used in this research to assist in writing (ChatGPT, Copilot). After using these tools/services, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.