Says Who? Argument Convincingness and Reader Stance Are Correlated with Perceived Author Personality

Sabine Weber, Lynn Greschner, and **Roman Klinger**
Fundamentals of Natural Language Processing, University of Bamberg, Germany
{sabine.weber,lynn.greschner,roman.klinger}@uni-bamberg.de

Abstract

Alongside its literal meaning, text also carries implicit social signals: information that is used by the reader to assign the author of the text a specific identity or make assumptions about the author's character. The reader creates a mental image of the author which influences the interpretation of the presented information. This is especially relevant for argumentative text, where the credibility of the information might depend on who provides it. We therefore focus on the question: How do readers of an argument imagine its author? Using the ContArgA corpus, we study arguments annotated for convincingness and perceived author properties (level of education and Big Five personality traits). We find that annotators perceive an author to be similar to themselves when they agree with the stance of the argument. We also find that the envisioned personality traits and education level of the author are statistically significantly correlated with the argument's convincingness. We conduct experiments with four generative LLMs and a RoBERTa-based regression model showing that LLMs do not replicate the annotators judgments. Argument convincingness can however provide a useful signal for modeling perceived author personality when it is explicitly used during training.

1 Introduction

When interpreting a text, social clues about the author (also called *social meaning*, Nguyen et al., 2021) and referential meaning are often intrinsically linked. Reading a social media post by a disliked politician might lead to a different interpretation than reading the same text but assuming it was written by a friend. Similarly, properties of the text like word choice might give clues to an author's educational level or personality traits. This interplay is especially relevant with regards to persuasive text, where a judgment of the argument's convincingness is derived from both the argument itself and the source of the argument (Petty and Cacioppo, 1986). It is our hypothesis that, when only minimal contextual grounding is available, readers develop their mental image of the author mostly from the argumentative text, allowing us a glimpse into the construction of social meaning.

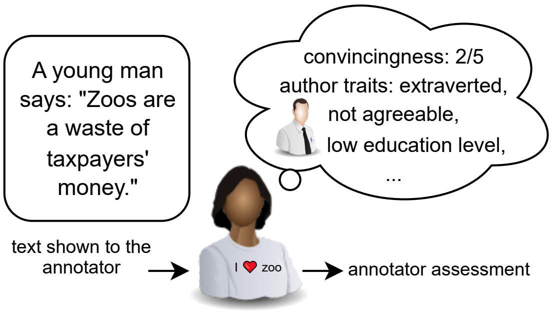
A young man says: "Zoos are a waste of taxpayers' money."
convincingness: 2/5
author traits: extraverted,
not agreeable,
low education level,
...
text shown to the annotator
I ❤ zoo
annotator assessment

Figure 1: Work flow example. The annotator's stance (pro zoos, as shown on the t-shirt) clashes with the stance of the argument. The argument is seen as unconvincing and leads to an unfavorable assessment of the author.

Previous work has found the assessment of argument convincingness to be a subjective task with low inter-annotator agreement (Quensel et al., 2025). While convincingness is derived from the argument text itself, it is also influenced by factors that are dependent on the reader of the argument: their familiarity with the topic and their stance towards it (Greschner et al., 2025). In this work, we examine both the argument-focused variable of convincingness as well as the annotator-focused variable of topic stance with regards to their role in the creation of a mental image of an argument's author. This way we move from an abstract view of arguments towards a perspectivist approach that integrates human differences.

To examine the phenomenon of author perception we investigate the connections between the convincingness of an argument and traits of its envisioned author (Big Five personality traits (Costa and McCrae, 1999) and education level). We find that convincingness is positively correlated with perceived education and the personality traits Agreeableness and Conscientiousness. We also investigate how properties of the annotator shape the envisioned author, specifically the annotators' agreement with the stance of the argument. We find that when the stance of the argument aligns with the stance of the annotator, they perceive the envisioned author to be more similar to themselves than when stances do not align.

We study the role of convincingness in modeling author personality by testing four generative LLMs, either with no convincingness signal, the annotator-assigned score, or a random value. Results show that the tested LLMs do not benefit from the convincingness signal. In contrast, a RoBERTa-based regression model trained with the convincingness signal better aligns with human annotations than the same model without it. This highlights convincingness as a valuable cue for modeling perceived author personality.

Our main research questions are:

RQ 1: Do annotators build an internal representation of an author when presented with an argument? *(Yes)*

RQ 2: What perceived author properties are associated with the individual assessment of convincingness in arguments? *(All personality traits show correlation with convincingness, with Agreeableness having the strongest correlation)*

RQ 3: How does annotator stance influence the similarity between annotator and envisioned author? *(When stances agree, there is a statistically significant correlation of all personality traits of annotator and envisioned author*)

RQ 4: Does the correlation in of convincingness and envisioned author personality traits (established in RQ2) carry over to computational modeling, helping to predict perceived author personality? *(Yes)*

Understanding how readers envision an argument's author is important because social inferences play a role in credibility judgments, yet they remain underexplored in computational argumentation. By examining how readers envision authors within a controlled setting, our work offers an empirical basis for understanding these social inferences. Our findings highlight that subjective factors like assessment of convincingness and reader stance are a non-negligible part of how people process argumentative text. We make all data and code publically available.[1]

2 Related Work

2.1 Social Meaning

Understanding and modeling the information that a text conveys about its author has been the focus of computational sociolinguistic research, looking especially at vernacular and dialect (Nguyen et al., 2016). Recent work has pointed out that the dimension of social meaning remains underexplored in the context of NLP, especially because modern NLP systems train on large data sets, where text is removed from the situational context of its creation, capturing the abstract patterns of language rather than its situation dependent use (Yang et al., 2025).

While social meaning might be embedded along representational meaning in language models, these models do not actively draw on this knowledge (Lauscher et al., 2022). Nguyen et al. (2021) argue that linguistic forms with different social meaning should not receive the same representation if social meaning is relevant for the task at hand. This would require disentangling the two types of meaning with regards to social context and properties of the interaction participants.

Recent work in perspectivism shares in this criticism of socially unaware models (Frenda et al., 2024). Perspectivist authors point out that variations between annotators should not be seen as noise or a product of insufficient annotator training, but as a source of information about the task at hand (Kanclerz et al., 2022; Casola et al., 2025; Weber-Genzel et al., 2024).

Our work follows this advice by considering not only argument text, but also individual properties of the annotator like their personality traits, education level, stance towards the argument topic, assessment of argument convincingness and the mental image that annotators create of the author.

2.2 Personality Traits

The task of deducing an authors personality traits from text is examined in the domain of author profiling (Verhoeven et al., 2016; Kreuter et al., 2022). In contrast to this work, we do not aim to deduce the ground truth personality traits of the author of an argument, but rather what a reader of the argument thinks they are – assessing an imagined au-

[1] https://www.uni-bamberg.de/en/nlproc/projects/emcona

thor, that might be different for each reader, rather than an actually existing person.

Another line of work examines the way readers construct mental representations of fictional characters from text. Pizzolli and Strapparava (2019) use the Big Five personality trait model to create character profiles from dialogues in theater plays and Tiuleneva et al. (2024) release a data set of character utterances annotated with Big Five personality traits. While these works are more similar to our in that they are also concerned with the mental representation of personality from text alone, they do not integrate subjective assessments of the annotators into modeling, thus ignoring differences between readers.

2.3 Perceived Author Identity in Arguments

While there is some research on perceived author properties in other domains, there is only little research about implicit information conveyed about the author in textual arguments, one instance being Bender et al. (2011). The paper studies claims to authority and agreement in Wikipedia forum discussions. While Bender et al. (2011) do not study the assumptions that conversation participants make about one another, they examine how speaker identity and authority is constructed in text, calling this *identity work* in reference to sociolinguistic research (Bucholtz and Hall, 2010).

Another notable work is the ContArgA corpus (Greschner et al., 2025), which allows us to examine this identity work at play in one concentrated snapshot. Unlike a lengthy forum discussion that allows for opinions to develop slowly, the corpus offers an opportunity to see the mental model that annotators develop of the author, based only on a text, minimal demographic information and their own prior belief.

3 Experimental Settings

To answer our research questions, we take a two-step approach. We first conduct a detailed statistical analysis of the ContArgA courpus to answer RQs 1, 2, and 3 and then use the gained insights do design our modeling experiments, thereby answering RQ 4.

3.1 The ContArgA Corpus

To conduct our research, we require a corpus that combines short textual arguments with annotator assessments of the text's convincingness and properties of the envisioned author of the text, e.g., the author's education level and Big Five personality traits. We also need the corpus to contain the same information (education level and Big Five Personality traits) about the annotator. The ContArgA corpus (Greschner et al., 2025) satisfies these requirements which is why we select it for our study.

The ContArgA corpus contains 800 arguments that were sampled from two existing argument corpora and re-annotated for a variety of different variables. Each argument was annotated by 5 annotators resulting in a total of 4000 annotations. Each annotator annotated at least 2 arguments, but multiple participation was allowed. Annotators were recruited via the annotation platform Prolific and they represent an even distribution across ages and genders. For further details, refer to Greschner et al. (2025).

During the annotation process, annotators were instructed to imagine themselves as participants in a town hall discussion on a contentious issue, watching a speaker approaching the podium and presenting an argument in favor or against the issue (an example of this can be seen in Figure 1). They then provided judgments of the convincingness of the argument, emotions they experienced and an assessment of the person saying the argument. While the corpus contains a variety of annotations, we are specifically interested in variables pertaining to the annotator and the envisioned author of the argument.

Input. After being introduced to the scenario annotators see the textual argument along a minimal description of the person uttering the argument. The description presents the annotator either with an old or a young person and a man or a woman, e.g.: "An old woman approaches the microphone and makes a statement: . . . ". The input variables are age and gender of the author and the argument they present. Textual arguments that explicitly refer to the speakers age or gender were removed during corpus creation to avoid conflicting inputs.

Annotations. The annotators provide three kinds of information: First, they provide information about themselves, specifically by disclosing their level of education and their stance towards the discussion topic and by filling a Big Five personality test (Gosling et al., 2003). Second, they provide their assessment of the argument, specifically by annotating its convincingness on a scale of 1 to 5, with 1 being least and 5 being most convincing. Third, they provide information about the en-

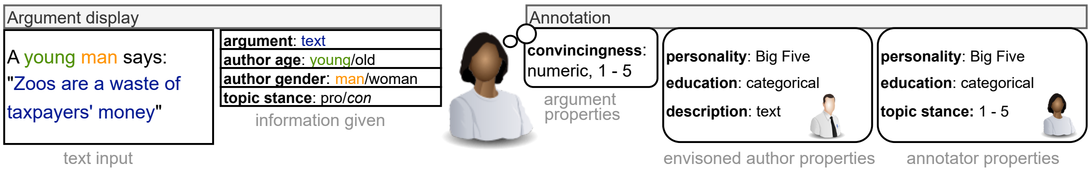

Figure 2: Overview of relevant variables collected in the ContArgA corpus.

visioned author, assigning them an education level, filling the same Big Five personality test they filled earlier for themselves for the author and lastly filling an optional free text field with further details about the envisioned author. An overview of the examined variables can be found in Figure 2.

3.2 Modeling

To examine the role of convincingness in modeling perceived author personality traits, we use the full ContArgA corpus as a test set. All LLMs receive the ContArgA corpus as input, while for the RoBERTa-based model we perform 5-fold cross-validation. This way we can report results for the RoBERTa model for all instances of the ContArgA corpus. In all test cases we compute the root mean squared error in comparison with the gold labels provided by the annotators for each of the Big Five personality traits. We release all models, training code and LLM outputs.[2]

LLM. To examine the capabilities of lightweight LLMs in modeling implicit assumptions about the author of a text we choose four recent open-weight LLMs (Gemma 8.5B (Mesnard et al., 2024), Mistral 7.2B (Jiang et al., 2023), Mixtral 46.7B (Jiang et al., 2024), 3.2B (Grattafiori et al., 2024)). We use a zero-shot prompting setup in which we put the full text that the annotators see as the user prompt. As system prompt we set the task of filling the Big Five personality test for the author of the argument. We also provide the not yet filled-out questionnaire itself, the same way as the human annotators were shown during the creation of the ContArgA corpus (the full prompt can be found in Appendix A). We set three experimental conditions: One providing the LLM with the convincingness value annotated by the human annotators (as an additional textual line in the prompt), one providing a random value, and one without this information. Building on the results of our earlier statistical analysis, we hypothesize that having the annotator-assigned convincingness values (in comparison to a random value or no value at all) should help the LLM to make author assessments that are more similar to human judgments.

All models are hosted on our own infrastructure via Ollama[3]. We use Ollama default parameters for all models, namely a temperature of 0.7, token limit of 1024 and Top P sampling set to 0.95.

RoBERTa-based Model. In addition to LLM prompting, we compare against a RoBERTa-based (Liu et al., 2019) regression model which is inspired by the FiLM model used for learning from input signals with different sized vector encodings (Perez et al., 2018). Each argument text is combined with the demographic data of the speaker (as in the example in Figure 2), tokenized and encoded by RoBERTa[4], and the representation of the [CLS] token is extracted as a fixed-length vector. To account for the role of perceived convincingness, we implement two model variants:

In the Baseline Model, the [CLS] embedding is passed through a two-layer feed-forward regression head (512 → 256 hidden units, ReLU activations, dropout rate 0.2), producing continuous predictions for the Big Five personality dimensions. In the Convincingness-Augmented Model, in addition to the argument text, the annotator-provided convincingness score is supplied as an auxiliary scalar input. The score is projected into the same dimensionality as the [CLS] embedding via a two-layer projection network, and used to generate feature-wise scaling and shifting parameters (FiLM modulation). These parameters are applied to the [CLS] embedding to yield a convincingness-conditioned representation, which is then passed through the same feed-forward regression head as the baseline. This architecture ensures that the augmented model has equivalent capacity to the baseline, with the only difference being the inclusion of convincingness information.

RoBERTa-based models were trained with early stopping, learning rate of 2×10^{-5} and a weight

[2] https://www.uni-bamberg.de/en/nlproc/projects/emcona

[3] https://github.com/ollama/ollama

[4] https://huggingface.co/FacebookAI/RoBERTa-base

decay of 0.01. The best performing model was trained for 9 epochs. All models were trained on a single Nvidia L40 GPU with one training run taking on average 1.5 minutes.

4 Results

In this following section we will outline the results of the analysis of the ContArgA data set and the modeling experiments.

4.1 RQ 1: Do annotators imagine an author when presented with arguments?

To answer this question we look at several statistical properties of the data presented in the ContArgA corpus.

Free Text Input. We investigate whether annotators use the free-text description field offered to give additional details about the author to attribute fully developed characteristics to them, and whether these descriptions differ depending on the argument or between annotators.

To address these questions, we analyze the free-text responses provided by annotators. We first examine the frequency and content of the entries, and then quantify variation using the Jaccard index, a measure of word overlap. Variation is assessed both within annotators (comparing their descriptions across different arguments) and across annotators (comparing descriptions for the same argument).

Overall, 69% of annotators completed the free-text field. Among the 100 most frequently used adjectives are “passionate” (22 mentions), “strong” (20 mentions), and “religious” (14 mentions), while common nouns include identity terms such as “father” (18 mentions), “student” (15 mentions), and “parent” (7 mentions). These patterns suggest that annotators ascribe specific roles and characteristics to the imagined authors.

To determine whether different arguments elicit different author descriptions, we calculated the Jaccard index for all entries from the same annotator. The resulting low average of 0.074 indicates that individual annotators provide distinct descriptions for different arguments. Exceptions exist, such as an annotator who consistently questioned whether each envisioned author was “of foreign descent.” Some annotators also maintain a repeated sentence structure (e.g., “She seemed ... ” or “The person is ...”) while varying the descriptors.

We then assessed whether different annotators envision the same author differently by computing the Jaccard index across annotators for the same argument. The very low average of 0.035 confirms substantial variation between annotators. Although occasional overlap occurs (e.g., two annotators describing an author as “right-wing”), more often, annotators use divergent descriptors, such as one calling the author “well-educated” while another describes them as “naive”.

Demographic Variables of the Author. We hypothesize that if an annotator envisages an author, they might apply the same demographic biases to them as they would to an existing person. Specifically, we want to examine whether perceptions of an author’s age or gender influence how annotators envisage them, e.g. whether older authors are assumed to be wiser than younger authors, or whether female authors are envisioned as more emotional than male authors. In the ConArgA annotation process demographic information about the author was presented alongside the argument text (see Figure 1), which allows us to answer these questions.

To answer them, we assign numerical values to the successive categorical education levels and perform a T-Test between the respective groups. Because the personality traits are valued with numeric values, e.g., Extraversion = 2, we perform a T-Test between respective groups here, too. We do not find any statistically significant differences in envisioned education level between female and male authors, and only slight differences in assumed personality traits: Women are rated slightly (but statistically significantly) higher in Agreeableness and lower in Emotional Stability, which aligns with gender stereotypes in the region where the corpus was collected (Plant et al., 2000).

Demographic bias is more pronounced along the age axis than along the gender axis: Old authors are assumed to have a lower education level than young ones and are perceived as statistically significantly different in all personality traits, being assumed to be less extraverted and open and more agreeable, conscientious and emotionally stable than young authors. These findings point towards the annotators forming a complex mental image of the author when confronted with the arguments, rather than envisioning an average or random person.

Trait	
Extraversion	-0.08^{***}
Agreeableness	0.39^{***}
Conscientiousness	0.22^{***}
Emotional Stability	0.18^{***}
Openness	0.06^{**}

Table 1: Pearson correlation (r) between assumed author traits and argument convincingness. Significance levels $^{**}p < .01$, $^{***}p < .001$

Trait	Agree	Disagree
Extrav.	-0.06^{*}	-0.13^{***}
Agreeabl.	0.20^{***}	0.07^{*}
Conscient.	0.13^{***}	0.05
E. Stab.	0.10^{***}	0.02
Open.	0.06^{*}	0.02

Table 2: Pearson correlation (r) between participant traits and assumed author traits, by stance agreement. Significance levels $^{**}p < .01$, $^{***}p < .001$

4.2 RQ 2: What perceived author properties are associated with the individual assessment of convincingness in arguments?

We aim to investigate whether the author of a convincing argument is imagined as having specific personality traits. To address this question, we compute Pearson correlations between argument convincingness and the envisioned education level and personality traits of the author.

We observe a statistically significant positive correlation between convincingness and the author's perceived education level (Pearson's r = 0.20). Additionally, all personality traits show statistically significant correlations with convincingness, most of them positive, except for Extraversion. The correlation results can be found in Table 1.

This leads us to the following conclusion: The more convincing the argument a speaker is presenting, the more they are perceived to have a high education, to be agreeable, conscientious and emotionally stable. Openness and Extraversion play a less important role, with less convincing arguments associated with higher Extraversion.

4.3 RQ 3: How does annotator stance influence the similarity of annotator and envisioned author?

Overall similarity. To answer the question if the annotators imagine the author to be similar to themselves we compare the education level and personality traits of the annotators with the values they assigned to the author, by looking at average values and Pearson correlation. We find that the mean difference between the education levels of annotator and speaker is -0.86, corresponding to roughly one education level, and a Pearson correlation of 0.23.

We find small but statistically significant correlations between all personality traits of the annotator and the author, ranging from weakest -0.09 for Extraversion to strongest 0.13 for Agreeableness. This shows that while there are big differences between single annotators, the properties of the imagined author are not entirely determined by the argument text but also to a significant degree by the annotators themselves.

Stance Alignment. The ContArgA corpus provides both the stance of each argument (pro/con) and the annotator's own stance, enabling a comparison of aligned vs. opposing stance conditions. We partition annotations accordingly and compute correlations between annotators' own education level and personality traits and those attributed to imagined authors.

We find that the assumed author education correlates statistically significantly with the annotator education in both cases, but is stronger when annotator and argument agree (r = .28) then when they disagree (r = .21). When looking at personality traits (see Table 2) we see statistically significant correlations for all traits when stances agree, with the strongest correlation for Agreeableness. For non-agreeing stances only Extraversion and Agreeableness show statistically significant correlations. This shows that when a speaker agrees with the presented argument they assume the author to be similar to themselves, except for the trait of Extraversion, where there is a negative correlation.

4.4 RQ 4: Does the variable of convincingness help in modeling perceived author personality?

The data set analysis shows that argument convincingness has a strong statistically significant influence on perceived author personality traits, with higher correlations than stance alignment. For this reason we examine the role of argument convincingness in modeling perceived author personality.

Trait	Gemma		LLaMA 3.2		Mistral		Mixtral		RoBERTa Reg.
	Rand	Gold	Rand	Gold	Rand	Gold	Rand	Gold	Gold
Extr.	+0.13	+0.13	+0.33	+0.36	-0.27	-0.31	+0.36	+0.22	+0.008
Agree.	+0.50	+0.50	+0.20	+0.36	-0.27	-0.29	+0.08	-0.55	+0.10
Consc.	+0.02	+0.05	+0.34	+0.20	-0.08	-0.10	+0.13	-0.24	+0.04
E. Stab.	+0.04	+0.04	-0.11	+0.04	+0.14	+0.14	+0.18	-0.28	+0.03
Open.	+0.48	+0.45	-0.09	-0.11	+0.18	+0.17	-0.02	-0.32	+0.002

Table 3: **Difference in Personality Trait Prediction Performance when Convincingness Signal is Added (avg RMSE)** Positive values (green) indicate improvement with the convincingness signal; negative values (red) indicate worse performance. For LLMs we compare improvements with a random convincingness signal and the gold convincingness signal. The RoBERTa model was trained and evaluated with the gold convincingness signal using 5-fold cross-validation. Absolute RMSE values are reported in Appendix C.

LLMs. We hypothesize that if LLMs mimic human assessments of an argument's author, then their predictions of the author's Big Five personality traits should improve when the annotator-assigned convincingness value is provided. To test this hypothesis, we use four different LLMs to annotate the arguments, using three prompting setups: One providing the human-annotated convincingness value with the prompt, one providing a random number as convincingness score and one only displaying the argument text without any convincingness information. We compute the root mean square error (RMSE) of the predicted personality trait values with the gold label provided by the annotator. Results can be seen in Table 3.

We can see that adding the annotator-assigned convincingness value to the prompt does not lead to a consistent improvement in prediction quality for Llama3.2, Mistral and Mixtral, worsening prediction quality for some traits. While prediction quality in Gemma is improved when the annotator-assigned convincingness value is provided, the same is also the case with a random value, suggesting that maybe the mention of convincingness as a keyword in the prompt leads the model to a different performance rather than the value itself.

To determine whether the correlation between convincingness values and specific personality trait is present in the LLM predictions, we run the same analyses as used for RQ3, computing the Peason correlation between the convincingness value assigned to the argument by a human annotator and the different LLM-predicted personality traits of the author. We find that the LLMs do not replicate the connection between higher convincingness and higher Agreeableness, Conscientiousness and Emotional Stability, showing no statistically significant correlation between convincingness and personality traits.

This can be seen as evidence that the tested LLMs are not intrinsically capable of reproducing this specific aspect of social meaning via zero-shot prompting.

RoBERTa-based Model. In the RoBERTa-based experiments, we evaluate whether incorporating annotator-assigned convincingness as an input signal improves performance relative to a model that omits this information, testing our hypothesis that convincingness aids in modeling perceived author personality. To do so we train both models and perform 5-fold cross-validation. We compute the root mean square error (RMSE) of the predicted personality trait values with the gold label provided by the annotator.

To test whether the differences between models' performance are statistically significant we use paired, instance-level resampling (Dror et al., 2018). For each predicted personality trait value, we compute RMSE for both models and form paired differences. We then perform a non-parametric paired bootstrap with 10,000 resamples to estimate the sampling distribution of the mean difference. The observed mean difference is -0.0221, with a 95 percent confidence interval of $[-0.0301, -0.0141]$, indicating a reliable overall advantage for the model that uses the subjective convincingness information. On the level of specific personality traits, the convicingness-informed model show statistically significant improvement in the prediction of Extraversion, Agreeableness and Openness. All results can be seen in Appendix B.

We find that unlike in the LLM prompting experiment, the RoBERTa model using the convincingness signal performs better on the prediction of all personality traits. While LLMs struggle to integrate the convincingness signal when given in a prompt, explicit integration of the signal in the RoBERTa-based model architecture helps the model to make use of the signal.

When computing the Peason correlation between the convincingness value assigned to the argument by a human annotator and the personality traits predicted by the RoBERTa-based model, we find that the correlation is much higher than the correlation in the human-annotated data, indicating that while the model uses the signal, it overly relies on this signal rather than learning other clues from text.

5 Discussion

Our results show that when annotators encounter an argument with little context, they still form an impression of the author of the argument. While these impressions are not universal across all annotators, there are nevertheless consistent trends: The argument's convincingness and the perceived author personality traits are correlated, with more convincing arguments being associated with a higher score in personality traits like Agreeableness, Conscientiousness and Emotional Stability.

This could be seen as a textual expression of the *Halo Effect* known to social psychology (Thorndike, 1920). The Halo Effect describes how a single salient positive trait such as physical attractiveness can create a favorable impression of a person, which then shapes the overall perception of their other qualities, such as assumed intelligence or trustworthiness. In our case, instead of attractiveness, the convincingness of the argument could influence the perception of the personality traits. Previous work shows this effect in multi-modal LLMs used for making hiring decisions (Kim et al., 2025) and in the reproduction of body image stereotypes by LLMs (Asad et al., 2025). While this phenomenon is well known in psychology and taken into account when designing studies, this is to our knowledge the first study to show this phenomenon in textual arguments and to use it in the design of a computational model for author perception.

We also show that people perceive an author to be more similar to themselves when they agree with the stance of the argument. We find this to be in line with social projection theory (Machunsky et al., 2014). When faced with an argument and very little other information about the author, the annotator might place the imagined author as an in-group or an out-group member based on stance alignment, and therefore assign them more similarity if stances match. This in turn can lead to a different assessment of the information conveyed in the text or in future interactions with the author, e.g. finding statements more trustworthy because they are perceived to originate from ones in-group. While there is some previous work investigating social projection in LLMs outputs (Sumita et al., 2025), this topic seems less explored in NLP and can offer an avenue for future work.

Lastly, we find that LLMs, at least in a zero-shot prompting approach, do not necessarily mimic these human behaviors. This points to the possibility that the tested light-weight models lack the implicit social reasoning or contextual inference abilities required to reconstruct perceived author characteristics from argumentative text alone, which is in line with previous works about the shortcomings of LLMs with regards to social reasoning (Lauscher et al., 2022). We also show that this social knowledge can be learned, using a lightweight approach that does not need adaptation of language models but relies on a regression head on top of fixed RoBERTa embeddings.

Our work may raise the question of desired model behavior. Should LLMs or other models that humans interact with represent envisioned author personality in line with their users? Previous work shows that humans integrate author-specific information when judging the convincingness of arguments (Petty and Cacioppo, 1986), which suggests that models approximating such judgments may require mechanisms for representing this information.

In this work, the LLM experiments are not intended to encourage anthropomorphic interpretations or to propose that models ought to construct a stereotypical author profile. Instead, they serve to probe whether current systems can use author-related cues in a way that is informative for modeling human assessments. This perspective frames LLMs not as stand-ins for human annotators, but as instruments for exploring how particular contextual factors may or may not be captured computationally. Understanding the boundaries of these capabilities is essential for designing methods that reflect human argument evaluation.

Our findings have implications with regards to the study of persuasion and misinformation spread on social media, where readers encounter arguments on contentious topics without the grounding of a personal relationship with an author or a longer discourse to contextualize statements. Readers are likely to construct their own mental models of authors based solely on textual cues, which can influence how persuasive they find an argument. This may reinforce existing biases or in-group preferences – a dynamic that plays a critical role in the amplification of polarizing content online. We therefore encourage the explicit modeling of social phenomena like the Halo Effect and social projection theory in future work.

6 Conclusion and Future Work

In this paper we examine argumentative texts with regards to the creation of social meaning, asking wether the readers of arguments imagine an author of the argument and what factors influence the properties of this envisioned author. Using statistical analysis we find that readers do imagine an author, showing that readers assign social roles and qualities based on the text and that their judgments in part reflect predominant demographic stereotypes. We also show that the convincingness of an argument is correlated with the envisioned author's education level and their personality traits, linking more convincing arguments with more educated and more agreeable and conscientious envisioned authors. We also find that annotators envision authors to be more similar to themselves if the stance of the presented argument (for or against a certain topic) aligns with the stance of the annotator.

Our work builds a connection to work in social psychology, where the Halo Effect is a well described phenomenon. The correlation between the convincingness of an argument and personality traits assigned to the envisioned author of the argument can be seen as one expression of the Halo Effect, where one perceived positive trait of a person (in our case high convincingness) influences the assessment of other unrelated traits. As to our knowledge this is the first work to show this effect in connection with argumentative text.

Future work should build on these findings by integrating them into models of argument and social interaction, addressing the need for representations of social meaning that are disentangled from denotational meaning. Ultimately, modeling social meaning can enable language representations that move beyond surface forms of text to capture the nuances of different usage contexts.

7 Limitations

Our work is limited to the data presented in the ContArgA corpus, which was created by annotators from a relatively constrained geographic area (the UK and Ireland). This limits the strength of the deductions based on it. It also contains only English language arguments. Despite these limitations the ContArgA corpus is to our knowledge the only corpus that examines envisioned author properties in the context of arguments, which makes it the best option to answer our research questions.

Work in perspectivism calls for modeling of single annotators, or a distribution of annotator judgments rather than one gold label. When using convincingness as a signal during modeling, we do so in an unaggregated manner, using the convincingness judgment of a single annotator as input and evaluating model performance against that same annotator's gold labels. Nevertheless, we report model performance averaged over all data points, which could be seen as a break from perspectivist modeling principles. Future work could be dedicated to exmaining model performance for different groups of annotators or as a distribution over annotators.

8 Ethical Considerations

Central to this work is the ContArgA corpus, which was collected prior to this work and is publicly available data. The collection of the ContArgA corpus was approved by the ethics boards, and conducted via online crowdsourcing for which the annotators were payed and provided consent for the usage of the data. The corpus does not contain any data that would allow for personal identification.

A guiding question for ethical consideration is who profits from our work and who is likely to get harmed by intended or unintended uses of it. We aim for this work to help in the modeling of social interactions online, specifically when arguments are encountered with little contextual information about the author, e.g., in social media contexts. Our findings can help to shine a light on the spread of misinformation or the mechanics of radicalization in online spaces. This can help to make these spaces safer for all participants.

We do not attempt to predict ground truth personality traits of real people, which would be a violation of privacy if used without consent. We are modeling what author readers imagine when they read an argument. This knowledge can be used not only to study, but also to manipulate author perceptions. When used with malicious intent this knowledge could be used in the creation of convincing chatbots that exploit the connection between stance alignment and perception of personality similarity to manipulate people.

Lastly, AI assistance was used during the creation of this paper. We used GitHub Copilot 0.36.2[5] for coding and ChatGPT-4[6] for assistance with LaTeX formatting of tables and rephrasing of text for clarity and grammatical correctness.

9 Acknowledgment

This project has been conducted as part of the Emcona (The Interplay of Emotions and Convincingness in Arguments) project, which is funded by the German Research Foundation (DFG, project KL2869/12–1, project number 516512112).

References

Narjis Asad, Nihar Ranjan Sahoo, Rudra Murthy, Swaprava Nath, and Pushpak Bhattacharyya. 2025. "You are beautiful, body image stereotypes are ugly!" BIStereo: A benchmark to measure body image stereotypes in language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 24471–24496, Vienna, Austria. Association for Computational Linguistics.

Emily M. Bender, Jonathan T. Morgan, Meghan Oxley, Mark Zachry, Brian Hutchinson, Alex Marin, Bin Zhang, and Mari Ostendorf. 2011. Annotating social acts: Authority claims and alignment moves in Wikipedia talk pages. In *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, pages 48–57, Portland, Oregon. Association for Computational Linguistics.

Mary Bucholtz and Kira Hall. 2010. *Locating identity in language*, pages 18–28. Language and Identities, Edinburgh University Press.

Silvia Casola, Yang Janet Liu, Siyao Peng, Oliver Kraus, Albert Gatt, and Barbara Plank. 2025. References matter: Investigating the impact of reference set variation on summarization evaluation. In *Proceedings of the 18th International Natural Language Generation Conference*, pages 274–291, Hanoi, Vietnam. Association for Computational Linguistics.

Paul Costa and Robert McCrae. 1999. A five-factor theory of personality. *Handbook of personality: Theory and research*, 2(01):1999.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392, Melbourne, Australia. Association for Computational Linguistics.

Simona Frenda, Gavin Abercrombie, Valerio Basile, Alessandro Pedrani, Raffaella Panizzon, Alessandra Teresa Cignarella, Cristina Marco, and Davide Bernardi. 2024. Perspectivist approaches to natural language processing: a survey. *Lang. Resour. Eval.*, 59(2):1719–1746.

Samuel Gosling, Peter Rentfrow, and William Swann Jr. 2003. Ten-item personality inventory. *Journal of Research in Personality*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, and (many others). 2024. The llama 3 herd of models. *arXiv preprint*, arXiv:2407.21783.

Lynn Greschner, Sabine Weber, and Roman Klinger. 2025. Trust me, I can convince you: The contextualized argument appraisal framework. *arXiv preprint arXiv:2509.17844*. Accessed: 2025-10-02.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *arXiv preprint*, arXiv:2310.06825.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts. ArXiv:2401.04088 [cs].

Kamil Kanclerz, Marcin Gruza, Konrad Karanowski, Julita Bielaniewicz, Piotr Milkowski, Jan Kocon, and Przemyslaw Kazienko. 2022. What if ground truth is subjective? personalized deep neural hate speech detection. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, pages 37–45, Marseille, France. European Language Resources Association.

Kyusik Kim, Jeongwoo Ryu, Hyeonseok Jeon, and Bongwon Suh. 2025. Blinded by context: Unveiling the halo effect of MLLM in AI hiring. In *Findings of the Association for Computational Linguistics: ACL*

[5] https://github.com/microsoft/vscode-copilot-chat

[6] https://chatgpt.com

2025, pages 26067–26113, Vienna, Austria. Association for Computational Linguistics.

Anne Kreuter, Kai Sassenberg, and Roman Klinger. 2022. Items from psychometric tests as training data for personality profiling models of Twitter users. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 315–323, Dublin, Ireland. Association for Computational Linguistics.

Anne Lauscher, Federico Bianchi, Samuel R. Bowman, and Dirk Hovy. 2022. SocioProbe: What, when, and where language models learn about sociodemographics. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7901–7918, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized bert pretraining approach. *CoRR*, abs/1907.11692.

Maya Machunsky, Claudia Toma, Vincent Yzerbyt, and Olivier Corneille. 2014. Social projection increases for positive targets: Ascertaining the effect and exploring its antecedents. *Personality and Social Psychology Bulletin*, 40(10):1373–1388.

Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L. Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu-hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, Kathleen Kenealy, et al. 2024. Gemma: Open models based on Gemini research and technology. *arXiv preprint*, arXiv:2403.08295.

Dong Nguyen, A. Seza Doğruöz, Carolyn P. Rosé, and Franciska de Jong. 2016. Computational sociolinguistics: A Survey. *Computational Linguistics*, 42(3):537–593.

Dong Nguyen, Laura Rosseel, and Jack Grieve. 2021. On learning and representing social meaning in NLP: a sociolinguistic perspective. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 603–612, Online. Association for Computational Linguistics.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. 2018. Film: Visual reasoning with a general conditioning layer. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Richard E. Petty and John T. Cacioppo. 1986. The elaboration likelihood model of persuasion. volume 19 of *Advances in Experimental Social Psychology*, pages 123–205. Academic Press.

Daniele Pizzolli and Carlo Strapparava. 2019. Personality traits recognition in literary texts. In *Proceedings of the Second Workshop on Storytelling*, pages 107–111, Florence, Italy. Association for Computational Linguistics.

E Ashby Plant, Janet Shibley Hyde, Dacher Keltner, and Patricia G Devine. 2000. The gender stereotyping of emotions. *Psychology of women quarterly*, 24(1):81–92.

Carlotta Quensel, Neele Falk, and Gabriella Lapesa. 2025. Investigating subjective factors of argument strength: Storytelling, emotions, and hedging. In *Proceedings of the 12th Argument mining Workshop*, pages 126–139, Vienna, Austria. Association for Computational Linguistics.

Yasuaki Sumita, Koh Takeuchi, and Hisashi Kashima. 2025. Cognitive biases in large language models: A survey and mitigation experiments. In *Proceedings of the 40th ACM/SIGAPP Symposium on Applied Computing*, SAC '25, page 1009–1011, New York, NY, USA. Association for Computing Machinery.

Edward Thorndike. 1920. A constant error in psychological ratings. *Journal of applied psychology*, 4(1):25–29.

Marina Tiuleneva, Vadim A. Porvatov, and Carlo Strapparava. 2024. Big-five backstage: A dramatic dataset for characters personality traits & gender analysis. In *Proceedings of the Workshop on Cognitive Aspects of the Lexicon @ LREC-COLING 2024*, pages 114–119, Torino, Italia. ELRA and ICCL.

Ben Verhoeven, Walter Daelemans, and Barbara Plank. 2016. TwiSty: A multilingual Twitter stylometry corpus for gender and personality profiling. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1632–1637, Portorož, Slovenia. European Language Resources Association (ELRA).

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.

Diyi Yang, Dirk Hovy, David Jurgens, and Barbara Plank. 2025. Socially aware language technologies: Perspectives and practices. *Computational Linguistics*, 51:689–703.

A LLM prompt

Figure 3 shows the full prompt used in our LLM experiments. It uses the same text as the text shown to the human annotators during the creation of the ContArgA corpus.

B Bootstrap Levels

Table 4 shows Bootstrap levels for all personality traits for the RoBERTa-based models.

Trait	95% CI Low.	95% CI Up.
Extraversion	-0.0141	-0.0072
Agreeableness	-0.1060	-0.0617
Conscient.	-0.0135	0.0169
Emot. Stability	-0.0190	0.0057
Openness	-0.0176	-0.0046

Table 4: Bootstrap 95% confidence intervals for the difference in RMSE ($\Delta = \text{RMSE}_A - \text{RMSE}_B$). Negative values indicate better performance by the convincingness-informed model. Intervals excluding zero denote statistically significant differences.

C Full RMSE values for all tested LLMs

Table 5 shows all RMSE values for all tested systems.

```
<|system|>
You are an expert human annotator. You must fill out the TIPI questionnaire.
Use ONLY numbers 1-7. If unsure, choose the closest valid number.
<|user|>
You are in a bustling town hall, a bright room with folding chairs and a low hum of conversation.
Posters line the walls, showcasing various perspectives on the issue at hand. A panel of speakers
sits at a long table in front, ready to listen as community members approach the podium. The
crowd is a mix of familiar faces, local officials, advocates, and reporters poised with cameras.
As the moderator calls for public comments, silence falls. Curious and skeptical eyes focus on
the podium as each speaker takes their turn.
{speaker_age} {speaker_gender} approaches the microphone and makes a statement: {argument}
You judge the convincingness of the statement on a scale from 1 (not convincing) to
5 (very convincing) as {convincingness}.

Here are a number of personality traits that may or may not apply to the person.
Please choose a number from 1 (Disagree strongly) to 7 (Agree strongly) to indicate the extent
to which you think the person would agree or disagree with that statement.
You should rate the extent to which the pair of traits applies to the person, even if one
characteristic applies more strongly than the other.

Answer in this format:
1. [rating]
2. [rating]
3. [rating]
4. [rating]
5. [rating]
6. [rating]
7. [rating]
8. [rating]
9. [rating]
10. [rating]

The person is extraverted, enthusiastic.
The person is critical, quarrelsome.
The person is dependable, self-disciplined.
The person is anxious, easily upset.
The person is open to new experiences, complex.
The person is reserved, quiet.
The person is sympathetic, warm.
The person is disorganized, careless.
The person is calm, emotionally stable.
The person is conventional, uncreative.

---
Important:
- Base your ratings only on the information given.
- Do not explain your ratings. Just output the numbers as shown above.
```

Figure 3: LLM prompt used for TIPI annotation.

Model	Setting	Extraversion	Agreeableness	Conscient.	Emotional Stab.	Opennness
Gemma	With Conv	1.73	1.75	1.22	1.40	2.64
Gemma	Random Conv	1.73	1.75	1.25	1.40	2.61
Gemma	No Conv	1.86	2.25	1.27	1.44	3.09
LLaMA 3.2	With Conv	1.65	1.98	1.81	1.93	2.42
LLaMA 3.2	Random Conv	1.68	1.87	1.67	2.08	2.40
LLaMA 3.2	No Conv	2.01	2.07	2.01	1.97	2.31
Mistral	No Conv	1.74	2.04	1.42	1.59	2.88
Mistral	Random Conv	2.01	2.31	1.50	1.45	2.70
Mistral	With Conv	2.05	2.33	1.52	1.45	2.71
Mixtral	With Conv	1.25	2.52	1.74	1.97	2.51
Mixtral	Random Conv	1.11	1.89	1.37	1.51	2.21
Mixtral	No Conv	1.47	1.97	1.50	1.69	2.19

Table 5: RMSE by personality trait across models and settings.