Stance-aware Definition Generation for Argumentative Texts

Anonymous ACL submission

Abstract

Definition generation models trained on dictionary data are generally expected to produce neutral and unbiased output. However, previous studies have shown that generated definitions can inherit biases from both the underlying models and the input context. This paper examines how stance-related bias in argumentative data influences generated definitions, demonstrating that even dictionary-trained models can produce outputs that reflect subjective or emotive framing. Additionally, we explore the intentional generation of persuasive definitions, which express an opinion about the target word based on argumentative usage examples. Through this study, we provide new insights into bias propagation in definition generation and its implications for argument mining and other Natural Language Processing applications.

1 Introduction

The task of definition generation has been explored in the context of lexical semantic change analysis (Giulianelli et al., 2023), automated generation of definitions for unfamiliar terms in scientific contexts (August et al., 2022), and assisted language learning and reading (Huang et al., 2022).

Definition generation can be framed as a sequence-to-sequence problem: "Given an input sequence *C* containing a term *T*, generate a contextually appropriate, neutral definition *D* for *T*" (Giulianelli et al., 2023). As illustrated in Table 1, the model receives an input sequence — in this case, an argumentative usage example — and is prompted to define the term *death penalty* as used in context. The generated output is the corresponding definition.

Models fine-tuned on dictionary data are generally expected to produce neutral and unbiased output. However, previous research on definition generation has shown that generated definitions can exhibit bias or reflect stereotypes inherited from the underlying models (Giulianelli et al., 2023). Since definition generation relies on contextual embeddings of input sequences, we hypothesize that stance-related bias in the argumentative input sequence can also propagate into the generated definitions.

Not all bias in natural language is inherently negative (Shah et al., 2020). Some forms of bias reflect diverse cultural perspectives, values, and stances on a given topic. In argumentation, for instance, one group may define assisted suicide as murder, while another may describe it as a humane act. While both groups agree that murder is immoral, they differ in how they interpret and categorize assisted suicide. As a result, their definitions carry distinct emotive connotations aligned with their stance. Reflecting such subjectivity is undesirable in tasks like diachronic semantic change analysis, but it could be leveraged in argument mining to generate persuasive definitions that capture differing perspectives.

This paper examines how biased training data and biased input sequences influence the presence of bias in the generated output. It also explores the intentional generation of contextually biased, or persuasive, definitions that express an opinion about the target word based on usage examples from argumentative texts.

This study **contributes** the following:

- We demonstrate that stance-related bias from argumentative data can propagate to varying degrees into definitions generated by dictionary-trained models, resulting in outputs such as those produced by LLama-3-8b-Instruct (AI@Meta, 2024) trained on three neutral dictionaries: "abortion is the act of deliberately killing a fetus".

- Our findings confirm that models fine-tuned on more expressive and loaded language, such

Usage Example	**Target Word**	**Definition**
As long as death penalty is kept, this confirms that our society is founded on violence, and that violence and brute force solve.	**death penalty**	The punishment of death by a state or other legal system for a crime or offence

Table 1: An example of a definition generated by Flan-T5 Base (Giulianelli et al., 2023) on IBM argument corpus (Friedman et al., 2021).

as Llama-3-8b-Instruct fine-tuned on the Urban Dictionary (Ni and Wang, 2017), are more likely to capture and reproduce stance-related bias with examples as follows: "death penalty is the most effective deterrent against murder", "assisted suicide is a euphemism for murder". This model exhibits the biggest overlap between stances of the generated definitions and those of the corresponding argument.

- We show that inference-time prompts have a limited impact on controlling the degree of context-related bias in the output.
- We provide a manually annotated dataset[1] evaluating the stance and plausibility of generated definitions, which can be used for neutral plausible definition detection or persuasive definition detection tasks.
- We provide a series of Llama-3-8b-Instruct definition generation models[2] fine-tuned on dictionaries and a combination of dictionaries (including and excluding the Urban Dictionary) that have comparable performance to the state of the art.

2 Related work

2.1 Definition generation

In recent years, a number of studies have focused on generating contextual definitions, based on an input sequence and a target word (Giulianelli et al., 2023; Periti et al., 2024; Mickus et al., 2022). The generation of definitions has been successfully applied to a variety of tasks, such as interpretability of static embeddings (Gadetsky et al., 2018), learning and reading assistance (Ni and Wang, 2017; Zhang et al., 2022), and semantic change analysis (Giulianelli et al., 2023; Fedorova et al., 2024). Notably, Giulianelli et al. (2023) show that generated definitions, derived from word usage examples, enhance the interpretability of semantic change analysis, making it easier for lexicographers and other researchers to track diachronic shifts in meaning.

Most English training data are sourced from traditional lexical resources such as the Oxford English Dictionary (Gadetsky et al., 2018), WordNet (Noraset et al., 2017), Wikipedia (Ishiwatari et al., 2019), and Wiktionary (Mickus et al., 2022), while Urban Dictionary is generally avoided unless non-standard English is specifically targeted, as in the work of Ni and Wang (2017).

Methods approach the task as a language modeling problem, where transformer-based Large Language Models are instruction-tuned (Zhang et al., 2023) to generate contextually appropriate definitions, as illustrated in Table 1. Several models have been explored in this setup, including sequence-to-sequence transformers like Flan-T5 (Giulianelli et al., 2023) and decoder-only architectures such as LLaMA2-Chat and LLaMA3-Instruct (Periti et al., 2024). These models are typically fine-tuned and evaluated on a combination of different dictionaries to assess their generalization ability.

In addition to instruction tuning, methods have been developed to enhance the quality of generated definitions such as adjusting their specificity (Huang et al., 2021) and complexity (August et al., 2022). These adjustments help tailor definitions to different contexts, making them more informative and interpretable across various applications.

The quality of generated definitions is typically assessed using standard natural language generation (NLG) metrics that measure overlaps with reference texts, such as BLEU (Papineni et al., 2002), SACREBLEU (Post, 2018) NIST (Doddington, 2002), ROUGE-L (Lin, 2004), and METEOR (Banerjee and Lavie, 2005), alongside semantic similarity measures such as BertScore (Zhang et al., 2020). Additionally, human evaluations are conducted to assess aspects such as ‘truthfulness’ and ‘fluency’, with inter-annotator agreement between 0.35 and 0.45 Krippendorff’s alpha (Giulianelli et al., 2023). Human annotations play a crucial role in evaluating the

[1] anonymised Github link

[2] anonymised Github link

plausibility of generated definitions, offering insights into how well they align with intended meanings based on specific evaluation criteria. Combining NLG metrics with human judgments ensures a more comprehensive and balanced evaluation, leveraging both quantitative and qualitative perspectives.

Generally, definition generation models have demonstrated the ability to capture fine-grained semantic nuances of target words depending on the context, highlighting their potential for broader applications in Natural Language Processing.

2.2 Definitions in argumentation

Work on argumentation theory has stated that many argumentative discussions involve a debate about how to define particular terms (Walton, 2005). So called *persuasive definitions*[3] often include loaded terms and rely on pathos, or emotive meaning, to make an argument about a topic: "Abortion is a murder of a human being". A pro-choice definition of abortion could then be "Abortion is the right of every woman to decide on her own body".

Formally, these statements function as definitions. However, they also serve as implicit arguments because they convey a stance. This contrasts with standard dictionary definitions, which aim to be objective and do not typically reflect an opinion. Dictionary definitions rely on common knowledge—accepted propositions that are not subject to dispute (Macagno and Walton, 2008)—whereas persuasive definitions act as implicit arguments, often reflecting the values and priorities of a particular group advocating for or against a topic.

Macagno and Walton (2008) describe persuasive definitions as those that align with two key argumentative schemes: argument from classification and argument from values. Stevenson (1938, 1944) identified two main strategies: altering the denotative meaning of a term by including or excluding specific objects (e.g., "Graffiti is art," redefining art to include graffiti), or modifying its emotive connotation without changing its meaning (e.g., "The death penalty is murder," framing the death penalty in morally charged terms). According to Macagno and Walton (2008), argument from classification involves redefining a term's denotation, while argument from values shifts its emotional connotation.

While exploring the shifting boundaries of such terms as art, justice, democracy etc. using NLP techniques presents an intriguing area for exploration, this paper focuses on analyzing definitions as potential arguments from values. Specifically, we aim to examine whether models trained on biased or unbiased data capture stance-related emotive connotations in the generated definitions.

As stated by Walton (2005), defining a term using loaded language constitutes an argument. While such definitions may not always be considered high-quality arguments, they nonetheless express a stance and provide a stance-specific interpretation of a concept. In addition, they highlight the value-based aspects that are most relevant to a given perspective — an approach also referred to as framing (Eemeren and Houtlosser, 1999; Ajjour et al., 2019).

In argument mining, the subjective and values-related nature of arguments has recently gained increased attention, leading to the adaptation of value taxonomies and the annotation of argumentative data for values (Kiesel et al., 2022), as well as the generation of arguments tailored to specific sets of morals (Alshomary et al., 2022). In this context, we investigate whether value-based information about opposing groups can be retrieved by generating context-dependent definitions that capture differing moral perspectives on a given topic.

3 Methodology

As we have demonstrated above, definition generation has the potential to move beyond neutrality, offering a means to explore and represent stance-based perspectives in argumentative contexts.

Based on these considerations, this paper investigates the following hypotheses:

1. H1: The stance-related bias in argumentative data will seep into definitions generated by dictionary-trained models that are expected to produce neutral definitions.

2. H2: A model fine-tuned on more expressive and loaded language will capture stance-related bias more accurately.

3. H3: In instruction fine-tuned models, prompts for zero-shot inference can be used to control the degree of persuasiveness in the generated definitions.

To explore these hypotheses, we instruction-tune Llama-3-8b-Instruct (AI@Meta, 2024) on the same dictionary data as in (Giulianelli et al., 2023; Periti

[3] The notions of persuasive and quasi-definitions were introduced by Stevenson (1938; 1944).

et al., 2024): WordNet (Ishiwatari et al., 2019), Oxford (Gadetsky et al., 2018), as well as Wiktionary (Mickus et al., 2022). In addition to the standard dictionary data mentioned above, we train a model using definitions from the online Urban dictionary (Ni and Wang, 2017). This crowd-sourced dictionary defines slang words, phrases, and cultural expressions. The train, validation and test splits are used as in Ishiwatari et al. (2019)[4].

We adhere to the standard template for instruction-tuning dictionary models, which involves providing the model with a prompt containing an input context sequence. The model is then prompted to: *Provide an accurate and appropriate definition of TRG* where *TRG* refers to the target word in the context sentence. The fine-tuned dictionary models are then used to generate definitions for a target word in an argumentative input sequence. The target word is the topic of the argument, the input sequence is the argumentative sentence containing the target word. Each input sequence thus expresses a stance towards the target word – pro or contra; see Table 1 for an example.

The argumentative dataset comprises stance-annotated arguments on abortion, gay marriage, and death penalty from the Webis Argumentative Corpus (Friedman et al., 2021), sourced from a debate platform by Bar-Haim et al. (2020), as well as arguments on assisted suicide and capital punishment from the IBM ARG KeyPoint Corpus of arguments (Ajjour et al., 2020). We preprocess the datasets to keep only the sentences containing the target word. The resulting corpus statistics are illustrated in Table 2 with the predominant number of sentences for the topic of abortion.

Dataset	Topic	PRO	CON
Webis	Abortion	3773	3560
	Gay marriage	960	871
	Death penalty	947	1144
IBM	Assisted suicide	121	125
	Capital punishment	110	126
	Marriage	111	125

Table 2: Number of argumentative sentences per stance and topic

The generated definitions are evaluated using standard NLG metrics mentioned above, followed by a qualitative analysis assessing stance and plausibility of the generated definitions.

4 Results

4.1 Language Model Evaluation

We train unsloth/llama-3-8b-Instruct[5] on Oxford, Wordnet, and Urban separately, in combination "All" – all dictionaries including Urban, and "NoSlang" – all dictionaries excluding Urban.

We evaluate the fine-tuned models' performance on dictionary test sets, reporting the above-mentioned standard NLG metrics for comparison, including ROUGE-L, BLEU, BERT-F1, NIST, SacreBLEU, METEOR, and EXACT MATCH: these metrics demonstrate both exact lexical overlap between the generated output and the reference as well as semantic similarity (BERT-F1).

Table 3 presents the evaluation results of our trained Llama models compared to the recent state-of-the-art Flan-T5[6] (Giulianelli et al., 2023) and LLama[7] (Periti et al., 2024) models. The values represent the average scores across all test sets (Oxford, Wordnet, Wiki, Urban). The averages for the benchmarks are based on the observed results in Periti et al. (2024).

The performance of our models trained with Unsloth is comparable to state-of-the-art results but does not significantly exceed established benchmarks due to lightweight training and reduced training parameters.

4.2 Bias evaluation

As part of our preliminary analysis, we apply a sentiment classification model[8] to pre-annotate the sentiment of definitions on the three largest topics of our argumentative data – Abortion, Death Penalty and Gay Marriage. Each definition is scored on a continuous scale of -1 to +1, with higher scores signifying increasingly positive sentiment. This allows us to gain a high-level view on how the models differ in terms of the average sentiment of their generated definitions, as illustrated by Figure 1. Initially, we expected Llama-Slang to produce a consistently more negative output, however, that was not confirmed: Figure 1 does not show any particular pattern for the models, what we observe

[4] github.com/shonosuke/ishiwatari-naacl2019#download-dataset

[5] Llama-3-70b was also fine-tuned but showed only marginal improvement with the average BERT-F1 of 88.19 on test splits; all the trained models and code are made available on HuggingFace and GitHub

[6] https://huggingface.co/ltg/flan-t5-definition-en-xl

[7] https://huggingface.co/FrancescoPeriti/Llama3Dictionary

[7a] On seen data.

[8] https://huggingface.co/tabularisai/multilingual-sentiment-analysis

	Oxford	Wordnet	All	Slang	NoSlang	Flan-T5 XL	LLaMA3 Dict
ROUGE-L	0.293	0.225	0.312	0.155	0.426	0.268	0.292
BLEU	0.091	0.058	0.101	0.028	0.132	0.180	0.191
BERT-F1	0.882	0.870	0.865	0.868	0.860	0.867	0.869
NIST	0.498	0.411	0.325	0.3648	0.327	0.583	0.680
SACREBLEU	9.200	5.900	10.100	2.800	13.200	12.01	13.729
METEOR	0.259	0.185	0.269	0.112	0.381	0.249	0.305
EX. MATCH	13.650	10.350	49.800	4.367	49.700	0.110^{α}	50.093^{α}

Table 3: Comparison of Definition Generation Models Across Different Training Data Sources

is a general negative sentiment associated with the topics.

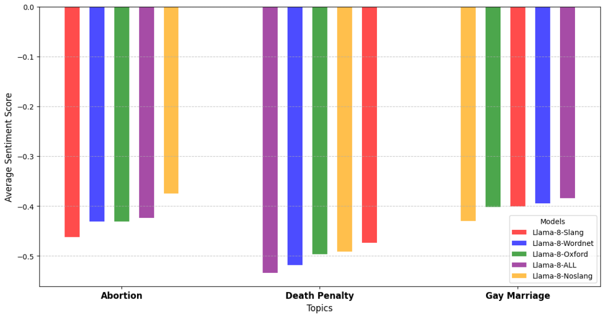

Figure 1: Average sentiment per model

Regarding Flan-T5 models, the sentiment scores ranged from -0.36 for Flan-T5-Base to -0.51 for Flan-T5-Large[9]. All models exhibited clearly negative sentiments within the definitions they generated. We attribute this mostly to the negatively associated vocabulary in the chosen topics.

Next, we automatically annotated stance of the generated definitions for the three topics of the Webis dataset. To do so, we fine-tuned microsoft/deberta-v3-base[10] models with Macro-F1: 0.747, Accuracy: 0.755 for the topic of *gay marriage*, Macro-F1: 0.754, Accuracy 0.755 for *death penalty*, and Macro-F1: 0.707, Accuracy: 0.707 for *abortion*. To train the models, we extracted the argumentative sentences containing target words from the Webis corpus with train, validation and test splits shown in Table 4.

This allowed us to compare each definition's detected stance with a corresponding argument's stance. The results (see Figure 2) indicate the percentage of the overlap per Llama model and a prompt (see Table 5) that was used to generate the definition. The largest proportion of the definitions reflecting the stance of the argument was observed for LLama trained on the Urban dictionary (Slang) followed by Llama trained on all dictionaries (All). Llama trained on WordNet, which provides a considerable number of word senses and examples, also showed a larger proportion of the stances overlap for *death penalty*. Using different prompts at inference for a more context-aware definition did not change much the stance presence in the generated output; however, prompts did consistently influence the length of the output with definitions generated with prompts 3 and 4 being 5-10 tokens longer on average.

Topic	**Train**	**Dev**	**Test**
Abortion	3480	1160	1160
Gay Marriage	1005	335	336
Death Penalty	1397	466	466

Table 4: Dataset splits for stance-detection training.

#	**Prompt Text**
0	What is the definition of {keyword} in the following text?
1	What is the contextual definition of {keyword} in this text?
2	In what sense is the {keyword} used in the following text?
3	What is the persuasive definition of {keyword} in the following text?
4	What is the emotionally charged definition of {keyword} in the following text?

Table 5: Prompts used for definition generation.

4.3 Definitions Topic Modeling

Previous research has explored clustering methods for retrieving various word senses (Giulianelli et al., 2023). In this study, we investigate whether soft clusters obtained through unsupervised topic models exhibit stance-related bias. To this end, we apply a BERTtopic model (Grootendorst, 2022) on definitions of the term "abortion" generated by both the Llama-Slang (which is expected to pro-

[9] The sentiment scores for 'death penalty' and 'gay marriage' did not vary significantly across Flan-T5 models. However, the 'Large' and 'XL' variants were slightly more negative (by less than 0.05) compared to other topics.

[10] https://huggingface.co/microsoft/deberta-v3-base

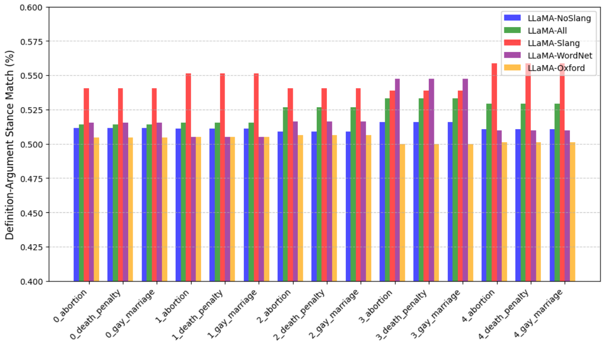

Figure 2: Prompt-Definition Stance Analysis

duce stance-related clusters) and Llama-NoSlang (which is expected to be neutral) models on the same dataset.

Our results (Figure 3, Figure 4) indicate that Llama-Slang, in addition to using more loaded and emotive language, tends to produce topics that reflect opposing perspectives on abortion. Interestingly, both sides of the argument are reflected in the output, with some clusters focusing on keywords "right to choose" while others contain negatively associated words such as "killing unborn baby" or "innocent/killing/murder". This is in contrast to Llama-NoSlang where topics tend to be fairly uniform and lacking the more charged language of the context sentences.

These findings lead us to believe that contextual bias from the test data seeps into the generated definitions, primarily when the model is trained on emotionally charged data. This model's awareness of bias, as we have seen, can better reflect varying perspectives. This also highlights the model's sensitivity to argumentative framing, making it a potential tool for analyzing value-based perspectives in the contested discourse. In contrast, we find Llama-NoSlang to be much more robust with respect to context variation, with most clusters corresponding to what one would intuitively consider a neutral and plausible definition for the term. Nevertheless, a thorough analysis of all the generated definitions shows that a "neutral" model might still generate biased output based on the input: "abortion is the act of deliberately killing a fetus", "death penalty is the judicial killing of a human", "assisted suicide is a deliberate act of self-destruction that is facilitated by another person" – these definitions are generated by one of our most robust models – Llama-NoSlang.

4.4 Annotated stance and plausibility across models

Generated definition evaluation is often supplemented by qualitative analysis and human annotations. In spite of a decent Bert-F1 score (0.87) across models as shown in Table 3, generated definitions might not be plausible because they are too general or subjective (Huang et al., 2021). Additionally, some models may fail to produce meaningful outputs at all, further necessitating human assessment.

In order to provide a thorough qualitative assessment of the generated definitions, we set up a two-dimensional annotation task where we aim to analyze the presence of stance in generated definitions and assess the general plausibility of the generated definitions. In this set-up plausibility is understood as clarity and accuracy of the definition. The annotations were performed by two human annotators, both graduate-level NLP researchers, authors of this paper. In the task, annotators were presented with a target word, its corresponding generated definition and were asked to evaluate:

- **Stance:** What stance is expressed in the definition towards the topic?
  (Options: Pro, Contra, Neutral)

- **Plausibility:** Does the generated text function as a proper definition by providing a clear and accurate explanation of the term?
  (Options: Yes, No)

In total, 500 definitions were annotated, selected as random samples of 100 definitions generated by each of the following models: Llama-Slang, Llama-NoSlang, Llama-All; Flan-T5-Base and Flan-T5-XL (Giulianelli et al., 2023). To ensure a fair comparison and avoid dependence on a single type of argumentative data, we excluded "abortion" from the Llama models' analysis, as it has been extensively studied and proved to be one of the most stance-dependent topics. Instead, we focused on "death penalty" and "assisted suicide" to provide consistent data across the three Llama models. Additionally, to assess whether the Flan-T5 models differ significantly from the Llama models, we annotated samples from the Flan-T5 models including diverse topics to gain general insights into these models' stance and plausibility results.

Both stance and plausibility judgments involve a degree of subjectivity, with agreement scores influenced not only by individual annotator interpretations but also by the diversity and distribution of annotated instances.

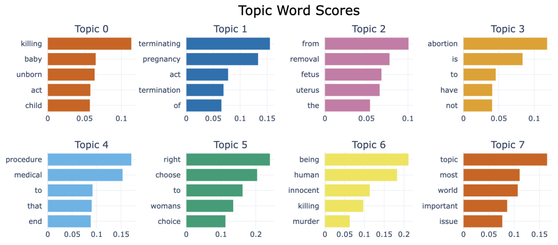

Figure 3: LLama-Slang abortion definitions

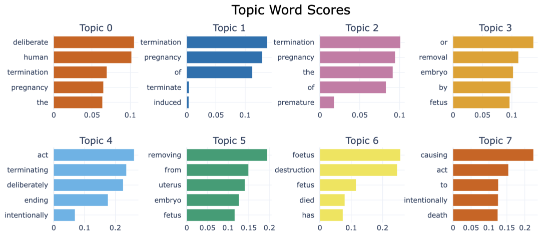

Figure 4: LLama-NoSlang abortion definitions

In Table 6, we report both the percentage of agreement between the two annotators and Cohen's Kappa (κ) to assess inter-annotator reliability for stance and plausibility annotations. Although the overall agreement is moderate, the highest inter-annotator agreement is observed for Llama-Slang in stance annotation (84%, $\kappa = 0.688$), which corresponds to the model with the largest proportion of biased definitions. This suggests that Llama-Slang produced more polarized definitions that facilitated stronger annotator agreement. The polarized definitions were not marked as plausible in most cases, as they were too subjective for a standard definition; however, they were good examples of persuasive definitions. For other models, annotators often detected slight biases that were insufficient to be annotated as pro or contra stance; thus, they were marked 'neutral'. The lower agreement rate corresponds to the amount of doubt annotators had when deciding on stance. The cases with higher percentages and low κ in Table 6 indicate cases where most stances were annotated as 'neutral'.

For example, Llama-NoSlang, which was expected to generate more neutral definitions, showed the highest percentage of agreement for stance (94%), but worse-than-chance Kappa score ($\kappa = -0.017$), which was the result of most generated definitions being neutral, suggesting that Llama-NoSlang is generally successful in generating neutral, dictionary-like definitions.

For plausibility judgments, agreement scores are generally lower than for stance, with Llama-Slang reaching $\kappa = 0.440$ and Llama-NoSlang showing the lowest reliability ($\kappa = 0.222$) with most examples being neutral and plausible; a larger-scale plausibility annotation might help evaluate the models better. In this setting, these metrics are an indicator of more homogeneous data with little variation in stance. This further underscores the subjective nature of plausibility and quality assessments, where annotators may have different interpretations of whether a definition is sufficiently informative and accurate.

Models	**Stance** (%)	**Plaus.** (%)	**Stance** (κ)	**Plaus.** (κ)
Llama-Slang	84.00	72.00	0.688	0.440
Llama-All	85.00	71.00	0.454	0.430
Llama-NoSlang	94.00	66.00	-0.017	0.222
Flan-T5-Base	95.00	82.00	0.519	0.572
Flan-T5-XL	97.00	76.00	0.652	0.465

Table 6: Inter-Annotator Agreement Llama output

Flan-T5 models presented generally more diverse comments, which resulted in more varied annotations. While for all the Llama models neutral stance would be associated with plausibility, annotators observed that Flan-T5 had most cases of neutral definitions that are not plausible, like: "Gay marriage is the practice of marrying people who are not your mate" (Flan-T5-XL). These models would also reproduce bias from the input sequence as in this definition of *death* penalty: "The infliction of the death penalty, in particular, the killing of an innocent person as a form of punishment".

In Figure 5, we present the Pearson correlation coefficients measuring the relationship between annotated definition stances and the stance of the corresponding argument. Additionally, we report the correlation between annotators, assessing their consistency in assigning stance labels.

These findings support the hypothesis that Llama-Slang captures more stance-related bias since the correlation between the annotated definition stance and argument stance is highest for this model. Notably, Annotator 2's stance annotations correlate slightly more strongly with the argument stance suggesting potential differences in how strictly each annotator perceived stance in definitions. The inter-annotator correlation (0.732) for Llama-Slang indicates a strong agreement between

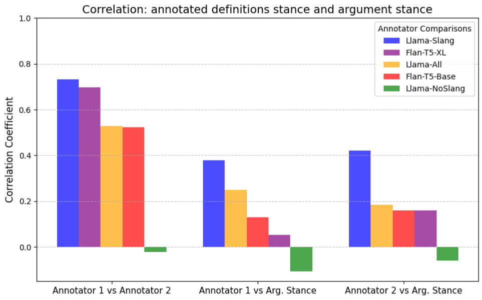

Figure 5: Definition stance correlation

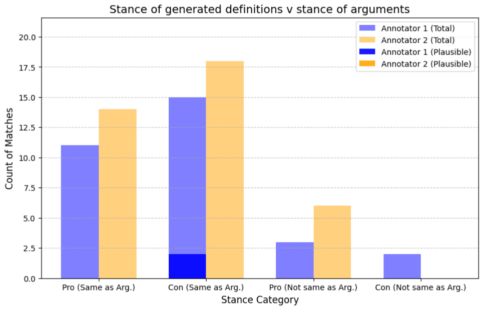

Figure 6: Llama-Slang: Stances of definitions and arguments

annotators, despite the subjective nature of stance detection. This suggests that while stance annotation involves some interpretative variation, annotators were largely consistent in their judgments. The decrease for Llama-All and even negative correlation for Llama-NoSlang can be explained, as mentioned before, by the high proportion of neutral definitions in this model's output.

Figure 6 presents an overlap of definitions of pro and contra stances with the stance of the corresponding arguments and the number of times the annotated stance was different from that of the argument. Despite the dataset being small, we do see that for most cases where the stance of the definition was annotated as pro or contra, it corresponds to the argument stance with only few examples of the "wrongly annotated" stances for definitions. Note also the lack of plausible definitions among the ones that have an explicit stance, as these are often judged too subjective or emotional to be considered as an appropriate definition of the term.

5 Conclusions and Future Work

This study explored how stance-related bias in argumentative data influences definition generation, using models trained on dictionary data. Our findings confirm key hypotheses regarding the propagation of bias, the role of training data in stance capture, but do not support the feasibility of prompt-based control over persuasiveness.

H1: Stance-related bias in argumentative data seeps into definitions generated by dictionary-trained models. Our results demonstrate that Llama and Flan-T5 models trained on neutral dictionary data might be influenced by bias present in the input sequence to a different extent. The best results in terms of neutrality were demonstrated by Llama-NoSlang trained on a few standard dictionaries and Llama-Oxford that shows the least changes when prompted to generate more contextually sensitive definitions.

H2: Models fine-tuned on more expressive and loaded language capture stance-related bias more accurately. In this paper, we showed that Llama-Slang, fine-tuned on the Urban Dictionary, showed the highest degree of definition stance alignment with the corresponding argument sentence. LLama-All, trained on all the dictionaries including Urban, showed second best sensitivity to stance-related bias in the input sequence among Llama models.

H3: Instruction fine-tuned models allow for prompt-based control over persuasiveness. Our experiments with different inference-time prompts showed an increase in definition length when prompts explicitly requested more context-aware definitions. However, generating differently biased definitions solely through prompting proved to be unrealistic. The training data of the model and the input sequence data would usually have a stronger effect on the generated output than a prompt.

Overall, our study provides insights into how stance-related biases of the argumentative data manifest in automated definition generation of the words that represent a topic of an argument across Llama and Flan-T5 models. The results highlight opportunities for refining models to better balance neutrality and context awareness. Additionally, leveraging context-dependent bias can offer valuable insights into underlying opinions and perspectives in argumentative discourse. Future work can focus on developing robust methods for controlling the degree of contextual bias in generated output and fine-tuning models specifically tailored for persuasive definition generation. Additionally, new evaluation metrics could be introduced to provide deeper insights into the plausibility of generated definitions.

Limitations

The limitations of this study are the following. First, the study is limited to English-language data and perspective only: what is plausible may differ across languages and countries depending on, for example, whether the death penalty, abortion, gay marriage, etc. is a legal practice or not. Second, we only trained and evaluated a series of comparatively smaller generative Llama models (llama-8b), and only marginally touched upon other models, like Flan-T-5. It is possible that our observations of stance and bias do not fully generalize to other models. Here, we anticipate two key possibilities: different or larger models could potentially be more robust against contextual variation in the input prompt, or they might become more reliant on their original training data, potentially reinforcing certain biases and failing to capture context entirely. Third, we only annotated a limited number of the generated definitions for the stance dataset. As a result, the analysis presented in the paper only provides a snapshot of the broader picture. While our sample size is sufficient for initial insights, future work should aim to extend the annotation process and provide a more complete human evaluation of the generated data. Fourth, we limited ourselves to target words that corresponded to topics of arguments, however, the arguments might have other interesting target words that can be defined persuasively eg. fetus in a debate on abortion. Finally, there is a lot of room to explore not only arguments from values but also arguments from classification: understanding the boundaries of abstract concepts that are commonly used in arguments is an exciting area for further research that could provide insights into questions like "What is understood with terms like extremism, terrorism, justice, democracy across languages and cultures?", etc.

References

AI@Meta. 2024. Llama 3 model card.

Yamen Ajjour, Milad Alshomary, Henning Wachsmuth, and Benno Stein. 2019. Modeling frames in argumentation. 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 2922–2932, Hong Kong, China. Association for Computational Linguistics.

Yamen Ajjour, Henning Wachsmuth, Johannes Kiesel, Martin Potthast, Matthias Hagen, and Benno Stein. 2020. args.me corpus.

Milad Alshomary, Roxanne El Baff, Timon Gurcke, and Henning Wachsmuth. 2022. The moral debater: A study on the computational generation of morally framed arguments. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8782–8797, Dublin, Ireland. Association for Computational Linguistics.

Tal August, Katharina Reinecke, and Noah A. Smith. 2022. Generating scientific definitions with controllable complexity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8298–8317, Dublin, Ireland. Association for Computational Linguistics.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72.

Roy Bar-Haim, Lilach Eden, Roni Friedman, Yoav Kantor, Dan Lahav, and Noam Slonim. 2020. From arguments to key points: Towards automatic argument summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4029–4039, Online. Association for Computational Linguistics.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the Second International Conference on Human Language Technology Research (HLT 2002)*, pages 138–145.

Frans Van Eemeren and Peter Houtlosser. 1999. Strategic manoeuvring in argumentative discourse. *Discourse Studies*, 1(4):479–497.

Mariia Fedorova, Andrey Kutuzov, and Yves Scherrer. 2024. Definition generation for lexical semantic change detection. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5712–5724, Bangkok, Thailand. Association for Computational Linguistics.

Roni Friedman, Lena Dankin, Yufang Hou, Ranit Aharonov, Yoav Katz, and Noam Slonim. 2021. Overview of the 2021 key point analysis shared task. In *Proceedings of the 8th Workshop on Argument Mining*, pages 154–164, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Artyom Gadetsky, Ilya Yakubovskiy, and Dmitry Vetrov. 2018. Conditional generators of words definitions. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 266–271, Melbourne, Australia. Association for Computational Linguistics.

Mario Giulianelli, Iris Luden, Raquel Fernandez, and Andrey Kutuzov. 2023. Interpretable word sense representations via definition generation: The case of semantic change analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3130–3148, Toronto, Canada. Association for Computational Linguistics.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Han Huang, Tomoyuki Kajiwara, and Yuki Arase. 2021. Definition modelling for appropriate specificity. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2499–2509, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Han Huang, Tomoyuki Kajiwara, and Yuki Arase. 2022. JADE: Corpus for Japanese definition modelling. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6884–6888, Marseille, France. European Language Resources Association.

Shonosuke Ishiwatari, Hiroaki Hayashi, Naoki Yoshinaga, Graham Neubig, Shoetsu Sato, Masashi Toyoda, and Masaru Kitsuregawa. 2019. Learning to describe unknown phrases with local and global contexts. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3467–3476, Minneapolis, Minnesota. Association for Computational Linguistics.

Johannes Kiesel, Milad Alshomary, Nicolas Handke, Xiaoni Cai, Henning Wachsmuth, and Benno Stein. 2022. Identifying the human values behind arguments. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4459–4471, Dublin, Ireland. Association for Computational Linguistics.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81.

Fabrizio Macagno and Douglas Walton. 2008. Persuasive definitions: Values, meanings and implicit disagreements. *Informal Logic*, 28(3):203–228. 26 Pages, Posted: 23 Jan 2011.

Timothee Mickus, Kees Van Deemter, Mathieu Constant, and Denis Paperno. 2022. Semeval-2022 task 1: CODWOE – comparing dictionaries and word embeddings. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1–14, Seattle, United States. Association for Computational Linguistics.

Ke Ni and William Yang Wang. 2017. Learning to explain non-standard English words and phrases. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 413–417, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Thanapon Noraset, Chen Liang, Larry Birnbaum, and Doug Downey. 2017. Definition modeling: Learning to define word embeddings in natural language. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence (AAAI)*, pages 3259–3266.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318.

Francesco Periti, David Alfter, and Nina Tahmasebi. 2024. Automatically generated definitions and their utility for modeling word meaning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14008–14026, Miami, Florida, USA. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting bleu scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191.

Deven Santosh Shah, H. Andrew Schwartz, and Dirk Hovy. 2020. Predictive biases in natural language processing models: A conceptual framework and overview. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5248–5264, Online. Association for Computational Linguistics.

Charles L. Stevenson. 1938. Persuasive definitions. *Mind*, 47:331–350.

Charles L. Stevenson. 1944. *Ethics and Language*. Yale University Press, New Haven.

Douglas Walton. 2005. *Fundamentals of Critical Argumentation*. Cambridge University Press.

Hengyuan Zhang, Dawei Li, Shiping Yang, and Yanran Li. 2022. Fine-grained contrastive learning for definition generation. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1001–1012, Online only. Association for Computational Linguistics.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, et al. 2023. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *Proceedings of the*

8th International Conference on Learning Representations (ICLR).