Stance-aware Definition Generation for Argumentative Texts

Natalia Evgrafova and **Loic De Langhe** and **Véronique Hoste** and **Els Lefever**
LT3, Ghent University, Belgium
{natalia.evgrafova,loic.delanghe,veronique.hoste,els.lefever}@ugent.be

Abstract

Definition generation models trained on dictionary data are generally expected to produce neutral and unbiased output while capturing the contextual nuances. However, previous studies have shown that generated definitions can inherit biases from both the underlying models and the input context. This paper examines the extent to which stance-related bias in argumentative data influences the generated definitions. In particular, we train a model on a slang-based dictionary to explore the feasibility of generating persuasive definitions that concisely reflect opposing parties' understandings of contested terms. Through this study, we provide new insights into bias propagation in definition generation and its implications for definition generation applications and argument mining.

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 abortion as murder, while another may describe it as a right of a woman to choose to terminate her pregnancy. While both groups agree that murder is immoral, they differ in how they interpret and categorize abortion. As a result, their definitions carry distinct emotive connotations aligned with their stance. Reflecting such subjectivity could be leveraged in argument mining to generate persuasive definitions that capture differing perspectives and understandings of the contested terms. This can aid in clustering arguments by perspective, summarizing key points of contention, and enhancing the understanding of diverse viewpoints within debates.

This paper examines how biased training data and biased input sequences influence the presence of bias in the generated definitions. 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 *"abortion is the act of deliberately killing a fetus"* produced by Llama-3-8b-Instruct[1] trained on three standard English

[1] https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md

Proceedings of the 12th Argument Mining Workshop, pages 168–180

Usage Example	Target Word	Definition
As long as death penalty is kept, this confirms that our society is founded on violence.	**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 an instance of the IBM argument corpus (Friedman et al., 2021).

dictionaries.

- Our findings confirm that a model fine-tuned on more expressive and loaded language, such as Llama-3-8b-Instruct fine-tuned on the Urban Dictionary (Ni and Wang, 2017), which is usually avoided (Periti et al., 2024), is more likely to capture and reproduce stance-related bias with examples as follows: "a punishment for someone who has committed a crime that is so bad that it should result in death", "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 can influence the degree of propagated stance-related bias in generated definitions.

- We provide (1) a manually annotated dataset evaluating the stance and plausibility of generated definitions, which can be used for neutral plausible definition detection or persuasive definition detection tasks; (2) a series of Llama-3-8b-Instruct definition generation models trained on dictionaries and combinations of dictionaries (including and excluding the Urban Dictionary) that have comparable performance to the state of the art[2].

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 the Urban Dictionary is generally avoided unless non-standard English is specifically targeted, as in the work of Ni and Wang (2017).

Recent 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, hu-

[2]The models, their training parameters, and data are available at huggingface.co/collections/LT3/stance-aware-definition-generation-for-argumentative-texts-6841456cadeec0116d0bad24.

man evaluations are conducted to assess the quality of the generated output, addressing, for example, '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 plausibility, or soundness, 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 stem from or involve a debate about how to define particular terms (Walton, 2005). The notions of persuasive definitions and quasi-definitions were introduced by Stevenson (1938, 1944). They 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 to have a child or not".

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, and implying a conclusion (Walton, 2005, p. 224).

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.

Similarly, 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 provide a stance-specific interpretation of a concept, highlighting 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., 2019a).

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 (and/or bias) in the generated definitions.

To explore these hypotheses, we instruction-tune Llama-3-8b-Instruct on the same dictionary data as in (Giulianelli et al., 2023; Periti et al., 2024): WordNet (Ishiwatari et al., 2019), Oxford (Gadetsky et al., 2018), as well as Wiktionary (Mickus et al., 2022), the standard English dictionaries. In addition to that, we include the online Urban Dictionary (Ni and Wang, 2017) in our training data. This crowd-sourced dictionary defines slang words, phrases, and cultural expressions. Previously, researchers abstained from using the Urban Dictionary as training data for non-slang applications to avoid unnecessary bias or possible errors (Periti et al., 2024). The train, validation, and test splits are used as in Ishiwatari et al. (2019)[3].

We adhere to the Alpaca template (Taori et al., 2023) for instruction-tuning our dictionary models. This involves providing the model with a prompt consisting of an instruction and an input context sequence. The model is instructed to answer the following prompt: `What is the definition of {keyword} in the following text: {usage example}?`

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 the death penalty from the Webis args.me corpus (Ajjour et al., 2019b), as well as arguments on assisted suicide and capital punishment from the IBM Keypoint Dataset (Friedman et al., 2021), sourced from a debate platform described by Bar-Haim et al. (2020). The topics were selected based on the high number of available arguments and their contested nature, as agreed upon in discussions among the authors. We preprocess the datasets by retaining only the sentences that contain the target word: arguments are first split into sentences, and only those in which the target word appears are kept. Corpus statistics are shown in Table 2, with abortion being the most represented topic.

[3] https://github.com/shonosuke/ishiwatari-naacl2019

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

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 the stance (pro, con, neutral) and plausibility — a clear and accurate explanation of the term — in the generated definitions.

4 Results

4.1 Language Model Evaluation

We train unsloth/llama-3-8b-Instruct[4] on Oxford, Wordnet, and Urban dictionaries separately, in combination "All" — all dictionaries including Urban, and "NoSlang" — all dictionaries excluding the Urban Dictionary.

We evaluate the fine-tuned models' performance on dictionary test sets as in Ishiwatari et al. (2019), reporting the above-mentioned standard NLG metrics for comparison with previous work, including BERTScore (BERT-F1), ROUGE-L, BLEU, 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[5] (Giulianelli et al., 2023) and Llama[6] (Periti et al., 2024) models.

The performance of our models trained with the Unsloth framework[7] is comparable to state-of-the-art results but does not significantly exceed established benchmarks due to lightweight training and reduced training parameters, but it demonstrates the highest semantic similarity score for the Oxford-trained model and higher overlap rates for "NoSlang" combination.

[4] Llama-3-70b was also fine-tuned but showed only marginal improvement with the average BERT-F1 of 88.19 on test splits

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

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

[7] https://huggingface.co/unsloth

Model	BERTScore-F1	ROUGE-L	BLEU	NIST	SACREBLEU	METEOR	EX. MATCH
Oxford	**0.882**	0.293	0.091	0.498	9.200	0.259	13.650
WordNet	0.870	0.225	0.058	0.411	5.900	0.185	10.350
All	0.865	0.312	0.101	0.325	10.100	0.269	**49.800**
Slang	0.868	0.155	0.028	0.365	2.800	0.112	4.367
NoSlang	0.860	**0.426**	0.132	0.327	13.200	**0.381**	49.700
Flan-T5 XL	0.867	0.268	0.180	0.583	12.010	0.249	0.110*
Llama3 Dict	0.869	0.292	**0.191**	**0.680**	**13.729**	0.305	**50.093***

Table 3: Comparison of definition generation models across training data sources. The table presents average scores across all test sets (Oxford, WordNet, Wiki, Urban). Notably, the "Oxford" model achieves the highest BERTScore-F1, indicating superior semantic similarity, while the "NoSlang" model excels in ROUGE-L and METEOR scores, reflecting its effectiveness in capturing content overlap. Averages for the Flan-T5 XL (Giulianelli et al., 2023) and Llama3 Dict (Periti et al., 2024) baselines are based on results reported in Periti et al. (2024). *On seen data.

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. Initially, we expected Llama-Slang to produce a consistently more negative output, however, that was not confirmed: our results did not show any particular pattern for the models; what we observed was a general negative sentiment (-0.4) associated with the topics. We attribute this mostly to the negatively associated vocabulary in the chosen topics.

Next, we automatically annotated the stance of the generated definitions for the three topics of the preprocessed Webis args.me dataset. To do so, we fine-tuned microsoft/deberta-v3-base[9] models. The training utilized argumentative sentences containing target words from the Webis args.me corpus, with train, validation, and test splits as detailed in Table 4. The model's performance was evaluated on the test split, yielding the following results: for *gay marriage*, Macro-F1: 0.747 and Accuracy: 0.755; for *death penalty*, Macro-F1: 0.754 and Accuracy: 0.755; and for *abortion*, Macro-F1: 0.707 and Accuracy: 0.707.

This setup allowed us to compare the pro and contra stance detected in each generated definition with the gold-standard stance of the corresponding argument. The results (see Figure 1) show the percentage of stance overlap for each Llama model and prompt: a neutral Prompt 0, context-sensitive Prompts 1-2, persuasive Prompt 3, and emotionally charged Prompt 4, as shown in Table 5. The highest proportion of stance-aligned definitions — those that expressed the same stance as the argument in the original corpus — was observed for the model trained on Urban Dictionary data (Llama-Slang), followed by the model trained on all dictionaries (Llama-All). The WordNet-trained model also achieved relatively high stance alignment in the *abortion* topic, likely due to its broad lexical coverage and usage examples. Prompts 1 and 4 consistently resulted in slightly higher stance alignment rates, while Prompts 3 and 4 tended to produce longer definitions — by approximately 5–10 tokens on average.

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

Table 4: Dataset splits illustrating the number of sentences containing the target word used for training topic-based stance-detection deberta-v3-base models.

In addition, we computed how often Llama-Slang's higher proportion of stances overlap was statistically significant compared to other models using a two-proportion z-test (see Table 10 in the Appendix for the detailed scores). As summarized in Figure 6, Llama-Slang showed statistically significant results in seven comparisons against Llama-NoSlang, four against Llama-WordNet, and nine against Llama-Oxford. These effects were most prominent for Prompt 1 and Prompt 4, which

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

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

explicitly encouraged contextually or emotionally framed definitions (see Figure 7).

#	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 BERTopic model (Grootendorst, 2022) on definitions of the term "abortion" generated by both the Llama-Slang — which is expected to produce stance-related clusters — and Llama-Oxford — which showed best results for the definition generation similarity to dictionary gold standards — models on the same dataset.

Our results (see Appendix for visualizations) 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-Oxford where topics tend to be fairly uniform and lack 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. The model's awareness of bias can help better reflect varying perspectives, making it a potential tool for analyzing how different ideological groups use and frame a particular term. In contrast, we find Llama-Oxford to be much more robust with most clusters corresponding to what one would intuitively consider a neutral and plausible definition. A thorough analysis of the generated definitions shows, however, 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

Evaluation of generated definitions is often supplemented by qualitative analysis and human annotations. Despite a decent BERTScore-F1 (0.87) across models as shown in Table 3, generated definitions might not be plausible because they are too general, subjective (Huang et al., 2021), or not meaningful.

In order to provide a thorough assessment of the generated definitions, we set up an annotation task where we analyze the presence of stance in the definition — pro, contra, and neutral — and assess the general plausibility of the generated definitions. In this setup, 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, 1000 definitions were annotated. First, we selected 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 explore stance-related bias in both Llama and Flan-T5 model outputs. In addition, we took a closer look at all five Llama models trained on Oxford, Wordnet, Slang, All and NoSlang data — specifically for the topic of abortion — to assess how training data influences the stance-related bias in generated definitions.

Both stance and plausibility judgments involve a degree of subjectivity, with agreement scores in-

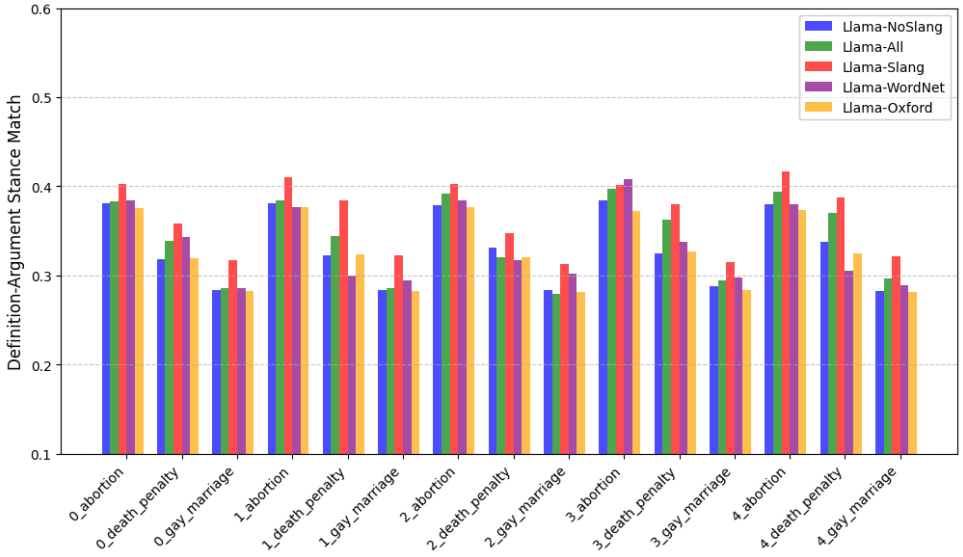

Figure 1: The overlap between the stance detected in the generated definitions and the gold-standard stance of the arguments used as context during generation.

Baseline Model	↑ Accuracy	p < 0.05
Llama-NoSlang	15/15	7/15
Llama-WordNet	14/15	4/15
Llama-Oxford	15/15	9/15

Table 6: Llama-Slang vs. baselines across 15 prompts. Columns show how often Llama-Slang outperformed each model in stance accuracy and in how many cases the difference was statistically significant ($p < 0.05$).

Prompt 0	Prompt 1	Prompt 2	Prompt 3	Prompt 4
0.1334	**0.0250**	0.2172	0.1720	**0.0184**

Table 7: Average p-values across model comparisons per prompt for Llama-Slang. Prompts 1 and 4 showed significant differences ($p < 0.05$), indicating stronger stance alignment under subjectively framed instructions.

Models	Stance (%)	Plaus. (%)	Stance (κ)	Plaus. (κ)
Llama-Slang	84	72	**0.688**	0.440
Llama-All	85	71	0.454	0.430
Llama-NoSlang	94	66	-0.017	0.222
Flan-T5-Base	95	82	0.519	0.572
Flan-T5-XL	**97**	76	0.652	0.465
Llama-Wordnet: abortion	78	74	0.541	0.470
Llama-Oxford: abortion	91	81	0.469	0.313
Llama-NoSlang: abortion	94	**93**	0.603	0.682
Llama-All: abortion	82	84	0.621	**0.684**
Llama-Slang: abortion	76	85	0.574	0.676

Table 8: Inter-Annotator Agreement for Llama and Flan-T5 definition annotations for stance and plausibility.

fluenced not only by individual annotator interpretations but also by the diversity and distribution of annotated instances. In Table 8, 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. For other models, annotators often detected slight biases that were insufficient to be annotated as pro or contra stance; thus, they were marked *neutral*. Higher percentages and low κ in Table 8 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 ex-

Model	% Stance-Taking	Match Rate (if stance)	Overall Match Rate	Avg. Plausibility (%)
Llama-Slang	**38.75**	76.84	**29.75**	37.25
Llama-All	24.00	75.27	17.75	53.25
Llama-WordNet	35.00	49.66	17.50	43.00
Llama-Oxford	9.00	54.17	5.50	**84.50**
Llama-NoSlang	5.50	35.42	2.00	79.75
Flan-T5-Base	5.50	81.67	4.50	70.00
Flan-T5-XL	4.50	75.00	3.50	66.00

Table 9: Comparison of stance sensitivity and definition plausibility across models. Llama-Slang produced the highest number of stance-aligned definitions, while Llama-Oxford and Flan models received the highest plausibility ratings.

amples being *neutral* and plausible; a larger-scale plausibility annotation might help evaluate the models better.

While for all the Llama models neutral stance would be associated with plausibility, annotators observed that Flan-T5 had 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".

The moderate and substantial inter-annotator agreement suggests that while stance annotation involves some interpretative variation, annotators were largely consistent in their judgments when evaluating biased definitions.

Table 9 presents a comparison of models based on their stance sensitivity (both pro and con) and definition plausibility. Llama-Slang stands out with the highest percentage of stance-taking definitions (38.75%) and the highest overall stance match rate between definitions and original arguments (29.75%), indicating that training on informal or biased data (like slang) can steer models to produce more context-sensitive outputs. However, these benefits come at the cost of plausibility: LLama-Slang received the lowest average plausibility rating (37.25%), mostly because annotators would not perceive biased definitions as plausible. In contrast, models like LLama-Oxford and Flan-T5-Base produced significantly fewer stance-taking definitions but were rated as more plausible, with Oxford achieving the highest plausibility score (84.5%). The model choice should therefore be guided by the specific goals of the application, whether to prioritize contextual and/or stance- sensitivity or definitional neutrality.

5 Conclusions and Future Work

This study explored how stance-related bias in argumentative data is reflected in the definitions generated by models trained on dictionary data. Our findings confirm key hypotheses regarding bias propagation, demonstrating how both training data and prompts influence models to produce more context-sensitive and stance-aware definitions.

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. We demonstrated that Llama-Slang, fine-tuned on the Urban Dictionary, had 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. We observed a statistically significant improvement in stance match accuracy for the model trained on slang data when prompts encouraged contextually or emotionally framed definitions. These prompts also resulted in longer definitions, indicating that explicitly requesting more context led to more elaborate and persuasive outputs.

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 may 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.

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-T5. 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 e.g. 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?".

Acknowledgements

This work was supported by the Research Foundation — Flanders (FWO) under grant FWO.OPR.2023.0004.01 (G019823N).

References

Yamen Ajjour, Milad Alshomary, Henning Wachsmuth, and Benno Stein. 2019a. 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. 2019b. Data acquisition for argument search: The args.me corpus. In *KI 2019: Advances in Artificial Intelligence*, pages 48–59, Cham. Springer International Publishing.

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.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca:

An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

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)*.

Appendix

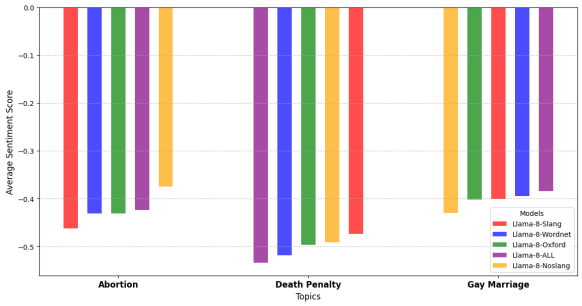

Figure 2: Average sentiment score per model across topics.

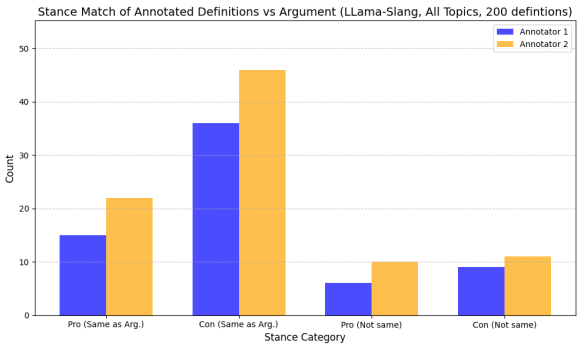

Figure 3: Llama-Slang: overlap between annotated stance of definitions and their corresponding argument stances.

Topic	Definition	Model B	Acc A (%)	Acc B (%)	Z	p-value
Death	definition_1_death_penalty	Llama-WordNet	38.41	29.98	3.673	0.0002
Death	definition_4_death_penalty	Llama-WordNet	38.76	30.56	3.559	0.0004
Abortion	definition_4_abortion	Llama-Oxford	41.62	37.33	3.096	0.0020
Death	definition_4_death_penalty	Llama-Oxford	38.76	32.44	2.729	0.0064
Death	definition_1_death_penalty	Llama-NoSlang	38.41	32.20	2.683	0.0073
Abortion	definition_4_abortion	Llama-WordNet	41.62	37.98	2.630	0.0085
Death	definition_1_death_penalty	Llama-Oxford	38.41	32.32	2.632	0.0085
Abortion	definition_4_abortion	Llama-NoSlang	41.62	38.02	2.601	0.0093
Abortion	definition_1_abortion	Llama-Oxford	41.06	37.62	2.490	0.0128
Abortion	definition_1_abortion	Llama-WordNet	41.06	37.62	2.490	0.0128
Death	definition_3_death_penalty	Llama-NoSlang	37.94	32.44	2.381	0.0172
Gay Marriage	definition_4_gay_marriage	Llama-Oxford	32.19	28.18	2.289	0.0221
Gay Marriage	definition_1_gay_marriage	Llama-Oxford	32.26	28.25	2.287	0.0222
Death	definition_3_death_penalty	Llama-Oxford	37.94	32.67	2.278	0.0227
Gay Marriage	definition_4_gay_marriage	Llama-NoSlang	32.19	28.25	2.247	0.0247
Gay Marriage	definition_1_gay_marriage	Llama-NoSlang	32.26	28.32	2.245	0.0248
Abortion	definition_1_abortion	Llama-NoSlang	41.06	38.06	2.170	0.0300
Death	definition_4_death_penalty	Llama-NoSlang	38.76	33.72	2.164	0.0304
Abortion	definition_3_abortion	Llama-Oxford	40.14	37.21	2.120	0.0340
Gay Marriage	definition_0_gay_marriage	Llama-Oxford	31.75	28.25	2.001	0.0454

Table 10: Results of z-tests comparing stance-match accuracy between Llama-Slang and baseline models. Only definition-topic pairs with statistically significant differences ($p < 0.05$) are shown.

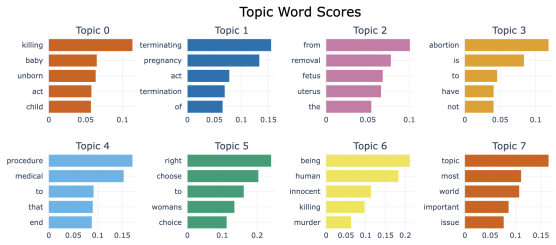

Figure 4: Llama-Slang abortion definitions

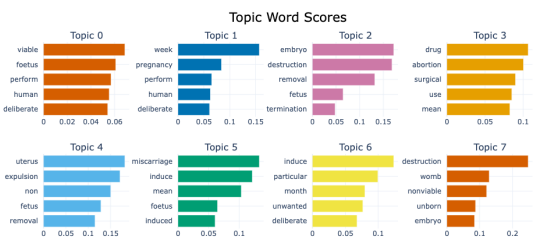

Figure 5: Llama-Oxford abortion definitions

Model	Topic	Stance	Generated Definition
Llama-Slang	Abortion	neutral	the act of a woman to remove an unwanted fetus from her uterus
Llama-Slang	Abortion	pro	a medical procedure that is both a right and a responsibility of women, and should not be illegal
Llama-Slang	Abortion	contra	the act of a woman killing a fetus with her body
Llama-All	assisted suicide	neutral	a method of suicide in which a person deliberately ends their own life with the assistance of another person or a device
Llama-All	assisted suicide	contra	the act of killing another human
Llama-All	assisted suicide	pro	the right to choose to end one's life
Llama-NoSlang	Abortion	neutral	the deliberate termination of a human pregnancy, most commonly performed as a surgical procedure by a qualified health professional
Llama-NoSlang	Abortion	pro	the deliberate termination of a human pregnancy, as a means of birth control, or because it poses health risks to the mother; the induced expulsion of a fetus or embryo from the uterus before the viability of birth.
Llama-NoSlang	Abortion	contra	the act of deliberately killing a human fetus
Llama-Oxford	Abortion	contra	a violent or illegal termination of pregnancy
Llama-Oxford	Abortion	neutral	the termination of a pregnancy by the removal or expulsion from the uterus of a non-viable fetus, or a fetus that does not stand a viable chance of survival after birth
Llama-Oxford	Abortion	neutral	the deliberate termination of a human pregnancy, most often performed before the fetus is viable, by various medical means, in order to remove a fetus that has serious abnormalities or is otherwise unsuitable for delivery or would otherwise produce a child that would suffer.
Flan-T5-Base	Death penalty	neutral	The punishment of death, in particular the execution of a condemned person by hanging
Flan-T5-Base	Gay marriage	contra	The practice of marrying people who are not your mate
Flan-T5-Base	Abortion	neutral	The act of terminating a pregnancy, either naturally or by artificial means
Flan-T5-Base	Death penalty	contra	The infliction of the death penalty, in particular the killing of an innocent person as a form of punishment
Flan-T5-Base	Death penalty	neutral	The punishment of death, especially as a legally mandated part of a state's criminal code
Flan-T5-XL	Gay marriage	neutral	The practice of marrying people who are not your mate
Flan-T5-XL	Abortion	neutral	The act of terminating a pregnancy, either naturally or by artificial means
Flan-T5-XL	Death penalty	contra	The infliction of the death penalty, in particular the killing of an innocent person as a form of punishment

Table 11: Generated definitions across models and stances for contested topics. Each definition reflects a perspective that aligns with a stance (pro, contra, neutral) as annotated by human annotators.