# Automated Definition Generation for Online Jargon Analysis

Helena Björnesjö, Axel Alness Borg, Katie Cohen, Björn Pelzer, and Erik Wachtmeister

FOI Swedish Defence Research Agency, Stockholm, Sweden
{helena.bjornesjo, axel.borg, katie.cohen, erik.wachtmeister}@foi.se

**Abstract.** Antagonistic online communities, such as the misogynist incel subculture, often develop a distinct linguistic style that is sometimes difficult for outsiders to interpret. A central challenge for law enforcement is identifying when common words are repurposed with new meanings. Such *polysemous* jargon complicates the detection of threats and radicalization cues, as unfamiliar word senses can be hard to recognize. We propose a prompt-based, corpus-level framework that leverages Large Language Models (LLMs) to generate contextualized definitions for polysemous terms and classify individual word usages accordingly. Using data from an online incel platform, we introduce an LLM-as-a-judge scoring method to evaluate the framework. Our results show that this corpus-level, prompt-based approach outperforms existing fine-tuned, sentence-level methods.

**Keywords:** Word sense disambiguation, Definition modeling, LLM-as-a-judge, Incels, Social media analysis

## 1 Introduction

Analysis of online communication is often complicated by the prevalence of *jargon*, i.e., specialized vocabularies and grammatical structures that demarcate social boundaries [4]. While jargon use can foster efficiency and cohesion among insiders [20], it also serves exclusionary functions, often deliberately obscuring meaning from outsiders. Given that violent extremist communities frequently employ jargon to obscure harmful content and evade detection [42], accounting for such language is essential in the analysis of online extremist discourse. Understanding how jargon is used in communication, is often crucial for interpreting threats or indications of violent intent. The ability to interpret these expressions also plays a key role in uncovering motivational dynamics and interrupting processes of isolation and radicalization. Since the volume of online data makes manual analysis difficult, automated methods for interpreting jargon offer clear value for forensic linguistics, law enforcement, and sociolinguistic research.

The incel subculture, a loosely connected online-based network of young heterosexual men who perceive themselves as romantically and sexually unsuccessful [30], is an example of a community with distinctive jargon. Incel forums often

promote conspiracy theories and violent misogyny, posing a latent threat to civil security through the normalization of radicalization and, in some cases, explicit calls to violence [37].

Incel jargon includes neologisms, such as 'foid' ('female android', a pejorative for women) and 'looksmaxxing' (the practice of improving one's physical appearance), as well as polysemous reappropriations of standard English terms. While neologisms are often easy to detect, polysemous words introduce multiple context-dependent meanings that are harder to resolve [40]. For example, 'toilet' is used as a pejorative to dehumanize women, while 'rope' functions as a verb referring to suicide.

Recent advances in Natural Language Processing (NLP) have highlighted the automatic generation of human-readable word definitions as a promising approach to both sense discrimination and sense-making for polysemous words. In the context of Word Sense Disambiguation (WSD), the set of possible senses is assumed to be known, and the task involves selecting the correct sense of a word based on its usage in a given context [11]. In contrast, Word Sense Induction (WSI) does not assume that the senses of a word are known. Instead, the senses must be induced from contextual use patterns [9]. Addressing the WSI challenge for words presumed to be polysemous due to jargon-specific reinterpretation via definition generation also enables compressing rich contextual information in a large dataset into more interpretable representations [32].

Building on recent work on definition generation [19, 31], we propose a definition generation framework for recognizing the jargon senses of polysemous words as they appear in online incel forums.

A notable shortcoming of previous approaches is that they generate separate definitions for each sentence where the target word appears, using each sentence as independent contextual input to the Large Language Model (LLM). This can result in definition inflation, the production of numerous distinct yet closely related definitions that reflect only minor contextual nuances. This highlights a central challenge in the task: generating all, but only the truly distinct, senses of a word. To account for this issue, we propose broadening the context used for generation to include multiple sentences, ranging from a few individual sentences to a modestly sized corpus. By doing so, we aim to reduce noise caused by sentence-level variation in word usage. Our method therefore produces contextually enriched definitions that reflect the meaning of a word sense across multiple word sense occurrences.

In contrast to prior work [19, 31], where the proposed methods rely on fine-tuning language models for definition generation, we adopt a simpler and more cost-effective prompt-based method, using LLMs to automatically recognize and generate word sense definitions in context, without the need for fine-tuning.

Finally, we leverage an LLM to assess the quality of the generated definitions. To our knowledge, this is the first application of the LLM-as-a-judge paradigm to the definition generation task [44]. Against this background, we explore the following research question:

*To what extent can LLMs be used to generate jargon specific word definitions from a corpus?*

We address this question by applying our prompt-based, corpus-level framework to incel forum data, and evaluate the approach against those of [19] and [31]. We explore the relative impact from different prompting techniques as well as the size of the corpus based on which the LLM generates a definition.

## 2 Related Work

Socialized forms of language use, such as jargon, have long attracted interest from sociolinguists investigating how language both shapes and is shaped by social phenomena [3, 23]. Such linguistic analyses can also support practical applications, such as online social community detection [34], analysis of extremist discourse [38] and hate speech [15, 29, 16, 7], and the monitoring of radicalization in digital environments [5].

The incel subculture has recently attracted growing attention in linguistic research. Incel-specific dictionaries have been compiled, and misogynistic expressions have been identified as a central component of incel jargon [21]. There is evidence that expressions of extremism within these communities have steadily increased over time [6]. Previous studies on incel jargon have also explored the relationship between semantic and behavioural features within incel forums [14, 13], the spread of incel jargon over time and across digital platforms [6], and jargon use in the broader so called manosphere [15].

In parallel, the field of computational linguistics has made significant progress, with machine learning models showing strong performance on classical linguistic tasks such as diachronic analysis and semantic change detection [8, 28, 32], as well as on the Word-in-Context (WiC) task [11], on WSD [11, 39, 45] and WSI [32, 36, 45], while definition modeling has emerged as a research area of its own [23, 18]. In particular, contextually informed generation of word definitions has shown promising results [24, 17].

Recently, Giulianelli et al. [19] proposed a method for generating word definitions for a given sentence containing the target word. Their approach involves fine-tuning a Flan-T5 XL (3B) [12] model on word usage data with annotated gold-standard definitions. Fine-tuning was done using word-sentence-definition triplets extracted from said datasets. To evaluate the quality of the generated definitions they compared the generated definitions to the gold standard definitions in the dataset using three different text similarity metrics: SacreBLEU [33] and ROUGE-L [26] to measure the overlap between the definitions, and BERTScore [43] for the semantic similarity.

Giulianelli et al. also propose a method of using the generated definitions for semantic change analysis by embedding them using SBERT [35] and clustering them with k-means using cosine similarity as a measure on distance between embeddings. They assign each cluster with a sense label by choosing the most prototypical definition in the cluster as the one with the smallest average distance to all other cluster embeddings.

Periti et al. [31] build upon the work of Giulianelli et al. by leveraging more recent LLMs. They fine-tune an instruction tuned Llama 3 (8B) [2] model for the definition generation task on the same dataset. To reduce the computational cost of fine-tuning the models, they use Low-Rank Adaptation (LoRA) [22]. They evaluate the generated definitions using the same text similarity metrics as [19]. They further evaluate the generated definitions by using them for WSI, WiC and Lexical Semantic Change (LSC) tasks. This is achieved by embedding the generated definitions similarly to [19]. Their results demonstrate that the proposed models match or outperform existing state-of-the-art methods across all evaluated tasks.

## 3 Method

We introduce and evaluate a prompt-based corpus-level approach to definition generation of polysemous jargon words and compare it to previous, definition generation approaches based on individual sentences and relying on model fine-tuning. We evaluate the approach on a dataset compiled from posts retrieved from an established online incel forum. We implement a simple baseline version of the new method, and also test various additional configurations of it for relative improvement. We also vary the number of context sentences provided to the LLM, which are used for generation of word sense definitions based on contextual use. Batches of sentences range in size from twenty-five sentences, up to the length of the entire corpus comprised of 100 retrieved sentences.

### 3.1 Datasets

To apply and evaluate our methods, we compiled a list of polysemous incel jargon terms, that is, words or abbreviations that have at least one widely understood meaning in standard English, and one or more additional senses specific to incel discourse. As a starting point, we collected incel jargon terminology from existing literature [21, 6, 25] and from an online glossary compiled by users of incels.is [1]. From these, we identified 17 recognizably polysemous words with distinct incel and non-incel senses, respectively. We manually crafted gold-standard definitions for these words, each receiving at least one definition representing common language use and at least one definition representing incel specific use. These definitions represent a ground truth against which the LLM-generated definitions can be compared. We then compiled two datasets of sentences containing occurrences of these words. The sentences were retrieved from two internet forums, each dataset representing one forum:[1]

[1] All retrieval, storage and processing of data was restricted to the secured infrastructure inside the premises of our institute. Only locally installed machine learning models were utilized. No identifiable samples of the data will be published or shared. The Swedish Ethical Review Authority has reviewed and approved this research (decision 2024-00576-01), including the retrieval, storage and processing of this data.

**Table 1.** The final list of polysemous incel jargon terms used for the experiments. The words are split into a validation and test used in different parts of the experimentation

| | |
|---|---|
| Validation | ER, Stacy, slayer, soy |
| Test | Becky, BF, bid, Brad, Chad, cope, rope, toilet, yellow fever, wizard |

- *incels.is*: With over 31,000 registered members, incels.is is the largest digital environment for male users identifying as incels. Founded in November 2017, it is also the oldest incel forum still in operation. To date, it has accumulated over 16 million posts, according to its own statistics. For our purposes this forum represents an incel environment where incel jargon has become commonplace. For each of the 17 polysemous words, we retrieved 100 randomly determined sentences, published on any of the public areas of the forum between November 2017 and April 2024.
- *AskReddit*: r/AskReddit with its approximately 100 million comments per year is a high-traffic subreddit on the Reddit platform, a large-scale, user-driven discussion forum. It invites open-ended, often informal questions across a broad spectrum of topics, making it a representative site for studying general-purpose online discourse. In this study, the dataset based on AskReddit serves as a control. We retrieved 100 sentences for each of the 17 words, randomly taken from the interval between January 2018 and January 2024.

The two datasets initially consisted of 1,700 sentences each, with 100 sentences per word. During manual annotation of the datasets, some previously unseen uses were detected, necessitating some adjustments of the definitions. One word was removed altogether, since it never occured in its incel jargon sense in the dataset. Two words (BBC and Albino) were singled out into a training set and used to test and explore prompts to be employed for our framework. Finally, the remaining 14 words were split into a validation and test set as shown in Table 1. The two sentence datasets used for the subsequent experiments thus consist of 1,400 sentences each, with 100 sentences per word. For the evaluation step, this set was further reduced, as it was not possible to assign a definition to every sentence during annotation.

### 3.2 Definition Generation

The method described here is intended for the overall task of identifying jargon-specific senses of polysemous words and for generating definitions of these senses. In our implementation, definitions were generated by prompting Llama 3.1 8B Instruct and Llama 3.1 70B Instruct, respectively. We evaluated Llama 3.1 8B Instruct to be able to more fairly compare results against previous methods. Following the assumption that a bigger model should be expected to perform better on any given task, we also evaluated Llama 3.1 70B Instruct. For all prompts we set the temperature to zero.

First, a base prompt was formulated, and in a zero-shot setting the LLM was instructed to define all word senses occurring in a provided sentence corpus. To

facilitate evaluation and further processing, the LLM output was post-processed by prompting the LLM to format the definitions into a list format. We then used regular expressions to extract the resulting definitions.

To systematically find an effective prompt for the task, the base approach was subsequently complemented with all possible combinations of three established prompt engineering techniques; few-shot prompting [10], chain-of-thought (CoT) prompting [41], and Self-refine (SR) prompting [27]. For example, one condition corresponded to the base prompt combined with a few-shot prompt. Due to the amount of prompts being tested, to avoid over-fitting on the test set, the prompt-improvement attempts were evaluated on a validation set, after which the best performing approach, in addition to the base prompt, was selected for evaluation on the test set. This was done for both the 8B and 70B Llama 3.1 Instruct models.

The prompt engineering techniques were modularly combined with the base prompt to cover all possible combinations. In the few-shot approach, two examples were prepended to the prompt, illustrating how to define the term 'BBC' using sentences sourced from Reddit and incel contexts, respectively. These sentences had a known word sense for 'BBC' following the manual annotation. The example answer was a listing of the gold standard definitions of 'BBC' occurring in the provided sentences. For the chain-of-thought configuration, an instruction to think step by step was appended. Self-refine prompting was implemented as an additional step in which the LLM was first asked to provide feedback on how to improve its initial answer based on the criterion that the provided definitions should be accurate and capture all word senses present in the sentences. The sentences were also provided as reference to the feedback generation step. Then, the LLM was instructed to improve the initial answer based on the generated feedback. The Self-refine approach is conceptually the same as described in [27]. However, for the sake of simplicity, we set the stop criteria to be one iteration and the attempt at improving the answer was therefore only done once.

In addition to the user prompt, a system prompt was also formulated. The system prompt instructed the LLM to take the role of a lexicographer persona; an expert on semantics and writing definitions. One goal of using the system message was to provide a formatting nudge for the resulting definitions. The system message was sourced initially from [31], but augmented to specify that the model should always complete the user's request, even when the content might be considered morally objectionable. This modification was motivated by early observations that the model sometimes refused to process incel jargon, possibly as a consequence of the offensive nature of content on the incel forum.

We also tested how the size of the input context affects definition generation by varying the number of input sentences to the LLM. All prompt-based experimental conditions were therefore tested both with and without batching. In the batching setup, instead of providing all sentences in the same prompt, the sentences were randomly sampled into groups of 25. Based on observations of the training set suggesting that the sentence order could influence which definitions were found, we employed oversampling, where the initial 100 sentences

were randomly sampled with replacement into 10 batches of 25 sentences each. The individual batches were then processed as before, the only difference being that the number of input sentences were fewer. After processing the batches, a summarization step followed in which the LLM was prompted to compile the generated definitions from the batches in a way that avoided duplicate definitions. Another motivation for batching was to reduce computational complexity by limiting the number of sentences analyzed simultaneously.

Finally, during preliminary testing on the training set we noticed that the set of definitions generated by the LLM sometimes contained hallucinated definitions, i.e. definitions that were not grounded in any word senses present in the corpus. To mitigate this, we explored post-processing the generated definitions through a self-consistency filter. For each of the input sentences, we subsequently asked the LLM to classify which, if any, of the generated definitions best describes the word sense in that sentence. The idea was to discard generated word senses that the LLM did not choose for classification, resulting in a reduced set of definitions. In the implementation we discarded word senses that the LLM did not select or only selected once in the list of 100 sentences. While this approach may inadvertently discard very rare word senses, it provides robustness against hallucinated definitions, as the LLM-based method for classifying what word sense definition, if any, fits a given sentence could itself misclassify.

### 3.3 Evaluation

To enable comparison with prior work, we applied the methods proposed in [19], using the Flan-T5-Definition-EN-XL model, and [31], using Llama3Dictionary, on our curated dataset. We then compared their results to the results of applying our own methods on the test set.

To evaluate our definition generation methods and compare them to previous methods which focus on sentence-based approaches [19, 31], we used the LLM-as-a-judge evaluation approach. To substantiate our evaluation results and enable comparison, we also used established metrics used to evaluate the sentence-based approaches. Additionally, the LLM-as-a-judge method was, in turn, evaluated in order to verify that it works on our task. Furthermore, we introduce and evaluate an LLM-based method for classifying which definition was used in a specific sentence given a set of corpus level definitions.

Comparing generated definitions against manually curated definitions at sentence level is difficult to automate. While there are metrics that have been used previously, none of them have been specifically developed for comparing definitions. To address this, we propose a new method: Instead of manually comparing and scoring the generated definitions against a gold standard definition, we employ an LLM judge [44] to automate this task. The LLM judge was instructed to make a semantic comparison between the generated definition and the gold standard definition at sentence level, and to assign a similarity score between 1-4.[2] For the experiments, Llama 3.1 70B Instruct was used as the judge model.

[2] Our LLM Judge prompt was based on https://huggingface.co/learn/cookbook/llm_judge. (retrieved 2025-06-11). The full prompt is available upon request.

To verify that the LLM judge rated the generated definitions similarly to a human annotator, we manually annotated a subset of the inputs to the judge. Due to resource limitations, we selected three sentences per word and forum for the test set words. For all these sentences, we collected the generated definitions for each of the experimental configurations evaluated on the test set. In total, this corresponds to 600 rated generated definitions; 480 for our evaluated prompt-based methods and 120 for the fine-tuning based methods. The human annotator rated all the generated definitions against the gold standard definition for each sentence. The LLM judge performed the same rating. We compared the ratings assigned by the LLM judge and human annotator using Pearson Correlation Coefficient (PCC).

While the LLM judge evaluated the definitions at the sentence level, we drew inspiration from prior work to evaluate the definitions at the corpus level. We used two metrics for this evaluation, BERTScore [43], and embedding the definitions with an S-BERT model[3] followed by the use of cosine similarity to semantically compare the generated definitions and the gold standard definitions at the corpus level. BERTScore and S-BERT similarity were calculated per word by comparing each generated definition to all the manually curated definitions and assigning the score of the manually curated definition with highest similarity score. Finally, we averaged the scores of all generated definitions to achieve a, per method, similarity score between the generated and manually curated definitions.

To enable comparison between our prompt-based corpus-level method and prior sentence-based approaches, each sentence-level word sense occurrence in the test set was classified into one of the generated definitions. To this end, a sentence-level classifier was implemented by prompting Llama 3.1 70B Instruct to classify which of a given list of generated definitions match the word usage in a sentence the best, if any at all.

To evaluate the classifier, we randomly sampled three sentences per word and forum from the test set. For each of the methods evaluated on the test set, we then iterated over the sampled sentences and provided both the LLM-based classifier and a human annotator with the corresponding generated list of word sense definitions. Both the LLM-based classifier and human annotator were then tasked with classifying which word sense, if any, describes the usage of the word in the given sentence. Then the human annotation is seen as ground truth for this task, and we calculated the accuracy of the classifier by looking at how many word usages that were classified with the same definition.

Prompt templates are available upon request.

## 4 Results

In this section we present the evaluation results of the experiments. To provide context, we start by evaluating the judge and classifier.

[3] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

### 4.1 Judge and Classifier Evaluation

Table 2 displays the results of comparing the LLM judge ratings of generated definitions in our curated definitions with how a human annotator would rate the generated definition. The results show a moderate to high correlation between the LLM rating and the human rating. The results also show that the LLM judge consistently gives lower ratings than the human annotator.

**Table 2.** Pearson Correlation Coefficient between LLM judge and human annotator and average rating per model. Data is a random subset of the test set data.

| Model | PCC | LLM Rating | Human Rating |
|---|---|---|---|
| Flan-T5 XL | 0.699 | 1.70 | 2.02 |
| Llama Dict | 0.844 | 1.74 | 2.08 |
| Llama 8B | 0.698 | 2.43 | 3.21 |
| Llama 70B | 0.678 | 2.78 | 3.29 |

**Table 3.** Classification accuracy for generated definitions by different models and the average number of definitions to classify including "Unknown".

| Model | Accuracy | Def Count |
|---|---|---|
| Llama 8B | 0.560 | 7.25 |
| Llama 70B | 0.604 | 5.29 |

Classifier performance results are presented in Table 3. The accuracy achieved by the classifier is moderate. Often an expression has several generated semantically similar definitions, which introduces uncertainty into a comparison of definition choices: The classifier may end up choosing a different, yet equally reasonable, best match than the human annotator, resulting in a formal mismatch. The number of generated definitions also impacts the accuracy, which can be observed when comparing the different models that generate varying numbers of definitions.

### 4.2 Selection of prompting techniques

Results from evaluating the different prompting techniques on the validation set are presented in Table 4. No single prompting method consistently outperformed the others across the different models and batching configurations. For each of the four setups, the best performing prompt methods together with the base setting were selected for evaluation on the test set. For the Llama 3.1 8B batching setup, the base method performed the best, therefore the second-best method was selected together with the base method for evaluation on the test set.

### 4.3 Multi-sentence methods evaluated on test set

The evaluation of the selected methods on the test are shown in Table 5. All our prompt methods outperform the baselines on both forums. The methods consistently get higher ratings on the Reddit forum than on the incel forum. When comparing model sizes, the larger Llama model consistently receives a higher rating. For the smaller model, the prompt techniques can improve the

**Table 4.** Average LLM judge scores by Llama 3.1 70B Instruct on the validation set across both the incel and Reddit subsets. Best performing method split on model and batching is shown in bold.

| **Llama 3.1 70B** | | **Llama 3.1 70B batching** | |
|---|---|---|---|
| **Method** | **AVG** | **Method** | **AVG** |
| Base | 2.18 | Batching | 2.13 |
| Base + CoT | 2.07 | Batching + CoT | 2.27 |
| Base + SR | 2.21 | Batching + SR | 2.21 |
| Base + few shot | **2.58** | Batching + few shot | 2.45 |
| Base + CoT + SR | 2.08 | Batching + few shot + CoT | **2.52** |
| Base + few shot + CoT | 2.38 | Batching + few shot + SR | 2.49 |
| Base + few shot + SR | 2.43 | Batching + CoT + SR | 2.07 |
| Base + few shot + CoT + SR | 2.56 | Batching + few shot + CoT + SR | 2.48 |
| **Llama 3.1 8B** | | **Llama 3.1 8B batching** | |
| **Method** | **AVG** | **Method** | **AVG** |
| Base | 2.14 | Batching | **2.21** |
| Base + CoT | **2.29** | Batching + CoT | 1.99 |
| Base + SR | 1.87 | Batching + SR | 1.96 |
| Base + few shot | 1.86 | Batching + few shot | 2.08 |
| Base + CoT + SR | 2.27 | Batching + few shot + CoT | 2.05 |
| Base + few shot + CoT | 1.90 | Batching + few shot + SR | 2.00 |
| Base + few shot + SR | 1.93 | Batching + CoT + SR | 1.78 |
| Base + few shot + CoT + SR | 1.87 | Batching + few shot + CoT + SR | 1.84 |

**Table 5.** Mean LLM judge scores by Llama 3.1 70B Instruct on the test set. AVG is the mean score over all ratings from the Reddit and incel subsets combined. The standard deviations of the ratings are reported in parenthesis.

| **Method** | **AVG** | **Reddit** | **Incels** |
|---|---|---|---|
| **Baselines** | | | |
| Flan-T5 XL | 1.84 (0.93) | 2.14 (0.96) | 1.56 (0.80) |
| Llama Dict | 1.83 (0.99) | 2.12 (1.08) | 1.56 (0.80) |
| **Llama 3.1 8B** | | | |
| Base | 2.27 (1.20) | 2.28 (1.30) | 2.26 (1.10) |
| Base + CoT | 2.52 (1.19) | 2.64 (1.26) | 2.41 (1.11) |
| Batching | 2.42 (1.18) | 2.51 (1.20) | 2.33 (1.15) |
| Batching + few shot | 2.62 (1.17) | 2.83 (1.13) | 2.43 (1.17) |
| **Llama 3.1 70B** | | | |
| Base | 2.74 (1.14) | 2.84 (1.12) | 2.65 (1.15) |
| Base + few shot | 2.48 (1.14) | 2.43 (1.19) | 2.54 (1.08) |
| Batching | 2.58 (1.21) | 2.82 (1.23) | 2.36 (1.15) |
| Batching + few shot + CoT | 2.59 (1.05) | 2.54 (1.09) | 2.64 (1.00) |

performance over the base prompt, while for the larger model the difference is not as clear. Using the batching method did not improve the performance.

**Table 6.** BERTScore and S-BERT similarity for generated definitions using selected prompt methods on the test set. Best performing method per model and forum for each metric is shown in bold.

| | Reddit | | Incels | |
|---|---|---|---|---|
| Method | BERTScore | S-BERT | BERTScore | S-BERT |
| **Llama 3.1 8B** | | | | |
| Base | 0.850 | 0.493 | 0.855 | 0.507 |
| Base + CoT | 0.847 | **0.511** | 0.849 | **0.560** |
| Batching | 0.855 | 0.484 | 0.856 | 0.518 |
| Batching + Few Shot | **0.857** | 0.494 | **0.858** | 0.509 |
| **Llama 3.1 70B** | | | | |
| Base | 0.864 | 0.546 | 0.862 | 0.539 |
| Base + Few Shot | 0.856 | **0.560** | 0.868 | **0.601** |
| Batching | 0.857 | 0.500 | 0.864 | 0.530 |
| Batching + Few Shot + CoT | **0.866** | 0.551 | **0.874** | 0.579 |

In Table 6, we evaluate the average similarity between the generated definitions and their closest matching human definition on using BERTScore an S-BERT cosine similarity as similarity scores. The results are relatively similar to the LLM judge ratings, albeit with some small variations. In most cases there is also agreement between the BERTScore and S-BERT when it comes to the ranking of the methods for generating definitions. When comparing the models, the larger Llama model consistently gets higher similarity scores independently of the prompt method used.

### 4.4 Effect of self-consistency filter

A comparison between Table 7 and Table 5 shows that the self-consistency filter method does not significantly impact the score from the judge. Table 7 shows that the baseline methods generates many more definitions than our corpus level methods. Further, the self-consistency approach significantly reduces the amount of definitions generated by the different methods. Also, Llama 3.1 70B model generates fewer definitions than Llama 3.1 8B.

## 5 Limitations

Our study is subject to several limitations that should be considered when interpreting the results:

Our definition generation approach was tested on a small number of jargon terms drawn from incel discourse. In particular, optimization on the validation set was negatively affected by the small size of the validation set.

The context used for sense induction for incel jargon is semantically opaque and context-dependent, thus posing a challenge for human annotators as well as

**Table 7.** The effects of the self-consistency filter on the number of resulting definitions and the quality of the remaining definitions as measured by the Llama 3.1 70B Instruct judge. Results are calculated on the test set. The average number of definitions represent the mean over both the incel and Reddit subsets. The reported 'AVG' LLM judge score is the mean score over all ratings from the incel and Reddit subsets combined.

| | Average number of definitions before and after filter | | Averged LLM judge scores after filter was applied | | |
|---|---|---|---|---|---|
| Method | AVG before | AVG after | AVG | Reddit | Incels |
| **Baselines** | | | | | |
| Flan-T5 XL | 65.7 | | | | |
| Llama Dict | 70.0 | | | | |
| **Llama 3.1 8B** | | | | | |
| Base | 7.0 | 5.6 | 2.28 | 2.22 | 2.35 |
| Base + CoT | 6.9 | 5.9 | 2.51 | 2.67 | 2.36 |
| Batching | 8.9 | 7.0 | 2.45 | 2.51 | 2.40 |
| Batching + few shot | 7.8 | 5.6 | 2.60 | 2.82 | 2.39 |
| **Llama 3.1 70B** | | | | | |
| Base | 4.8 | 4.7 | 2.74 | 2.89 | 2.60 |
| Base + few shot | 4.0 | 2.7 | 2.45 | 2.44 | 2.45 |
| Batching | 6.3 | 5.7 | 2.62 | 2.86 | 2.40 |
| Batching + few shot + CoT | 4.5 | 3.9 | 2.61 | 2.53 | 2.69 |

automated systems. However, our results indicate that LLMs can still generate high-quality definitions, relative to the human annotated gold-standard.

While our suggested method was tested on online incel jargon, its performance on other types of jargon, such as medical or legal terminology, remains unknown. Generalizability may be affected by differences in social context, language complexity, and semantic ambiguity. Furthermore, since incel jargon has to some extent spread from incel communities to a broader online discourse, LLMs may have encountered incel jargon during training. Thus it is unclear whether our method would perform equally well for newly emerging or highly niche jargons that may not be present in LLM training data.

Since we did not implement an error analysis of the self-consistency filter, it is unclear how many valid definitions were erroneously filtered out. However, the filtering did not seem to affect performance, suggesting that filtered out definitions were indeed not valid or redundant.

From a practical perspective, LLMs require significant hardware resources, especially if, as in our case, they must be run locally. This may limit the scalability of our approach in some applications.

## 6 Discussion

This study proposes a prompt-based, corpus-level framework for generating word sense definitions of polysemous jargon. Our results from using the framework on incel jargon suggest that this approach is well-suited for forensic and intelligence work, particularly in domains characterized by coded or opaque language. By providing interpretable and community-specific definitions, the method supports tasks such as threat assessment, radicalization detection, and discourse analysis.

A central advantage of our approach is the deployment of corpus-level definition generation. While prior methods risk definition inflation by treating each sentence in isolation, our method produces a compact, semantically distinct set of definitions. Combined with a classifier, the definitions remain applicable to individual sentences. Additionally, a corpus-based approach results in definitions that reflect the collective usage within a community rather than isolated or idiosyncratic word usage. This contributes to a more accurate understanding of group discourse and reduces the risk of misinterpretation.

Another contribution is our prompt-only strategy. In contrast to resource-intensive fine-tuning approaches, our method performs strongly with no additional model training.

To evaluate definition quality, we introduced an automated LLM-as-a-judge scoring procedure. This method approximated human judgments well and offers a scalable solution to the otherwise manual and time-consuming process of evaluating generated definitions. The LLM was conservative in scoring, which may be preferable in high-stakes applications requiring cautious interpretations.

These improvements also suggest applicability to related tasks such as WSI. As shown in [31], generated definitions can be used in the WSI task where it may match or outperform other state-of-the-art methods.

While our LLM-as-a-judge evaluation approach shows promising correlation with human judgment, it also introduces potential sources of bias. The model's judgments may reflect training data biases or content moderation heuristics, especially when dealing with extremist or toxic language. More validation is needed to establish its reliability. Similarly, the classifier evaluation may be subject to related biases, as well as ambiguity caused by redundant definitions, which can lead to multiple plausible interpretations for a single instance.

Regarding model size, we observed that smaller models benefited substantially from few-shot and chain-of-thought prompting, while larger models performed well even with a simple base prompt, suggesting that the additional complexity of engineered prompts yields diminishing returns as model capacity increases. Similarly, our batching strategy improved results for smaller models, while offering limited benefit for the larger model, likely due to its greater context-handling ability. Another possible explanation is the modest corpus size used in our study; batching may become more advantageous with larger datasets. This observation also suggests that the quality of generated definitions is likely to improve over time, as models continue to improve.

We also observed a consistent difference in model performance between the Reddit and incel forum datasets. Definitions generated from Reddit data tended

to receive higher evaluation scores, likely due to the more conventional language used on Reddit. The LLMs' internal safety mechanisms may also have been triggered by offensive language and hate speech on the incel forums, limiting their engagement with such content.

Together, these findings support the viability of corpus-level definition generation using LLMs and prompt-based methods, and anwers our research questions in the affirmative.

## 7 Conclusions and Future Work

Compared to earlier fine-tuned, sentence-level methods, our prompt-based, corpus-level approach produces more concise and semantically distinct definitions without additional training, resulting in stronger evaluation scores and greater practical flexibility. The results suggest that fine-tuning is not necessary for effective definition generation. The method offers potential utility in digital forensics, law enforcement, and sociolinguistics. By generating community-wide definitions, it supports the interpretation of coded or opaque language, enabling the detection of semantic shifts, radicalization cues, and the spread of extremist jargon.

The introduction of an LLM-as-a-judge evaluation method also contributes to the field by providing a scalable and effective alternative to manual evaluation. Future work should include model-family diversification with regard to the LLM judge, to improve the reliability of the judge.

Next steps include assessing generalizability by expanding the incel jargon vocabulary and corpus size, as well as applying the method to other domains, such as metaphorical or mainstream-adapted jargon. This includes contexts with sparse usage data or greater semantic variability.

**Acknowledgments** This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 101021797.

## References

1. Incel Glossary — incels.wiki. https://incels.wiki/w/Incel_Glossary, [Accessed 26-02-2025]
2. AI@Meta: Llama 3 model card (2024), https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md
3. AlJadda, K., Korayem, M., Grainger, T., Russell, C.: Crowdsourced query augmentation through semantic discovery of domain-specific jargon. In: 2014 IEEE International Conference on Big Data (Big Data). pp. 808–815 (2014)
4. Allan, K.: Jargon. In: Brown, K. (ed.) Encyclopedia of Language & Linguistics (Second Edition), pp. 109–112. Elsevier, Oxford, second edition edn. (2006)
5. Araque, Ó., Sánchez-Rada, J.F., Carrera, Á., Iglesias, C.A., Tardío, J., García-Grao, G., Musolino, S., Antonelli, F.: Making sense of language signals for monitoring radicalization. Applied Sciences 12(17) (2022), https://www.mdpi.com/2076-3417/12/17/8413

6. Baele, S., Brace, L., Ging, D.: "a diachronic cross-platforms analysis of violent extremist language in the incel online ecosystem". Terrorism and Political Violence 36(3), 382–405 (2023)
7. Berglind, T., Pelzer, B., Kaati, L.: Levels of hate in online environments. In: Spezzano, F., Chen, W., Xiao, X. (eds.) ASONAM '19: International Conference on Advances in Social Networks Analysis and Mining, Vancouver, British Columbia, Canada, 27-30 August, 2019. pp. 842–847. ACM (2019)
8. Bevilacqua, M., Maru, M., Navigli, R.: Generationary or "how we went beyond word sense inventories and learned to gloss". In: Webber, B., Cohn, T., He, Y., Liu, Y. (eds.) Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 7207–7221. Association for Computational Linguistics, Online (Nov 2020)
9. Biemann, C.: Word Sense Induction and Disambiguation, pp. 145–155. Springer Berlin Heidelberg, Berlin, Heidelberg (2012)
10. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language models are few-shot learners (2020), https://arxiv.org/abs/2005.14165
11. Cassotti, P., Siciliani, L., DeGemmis, M., Semeraro, G., Basile, P.: XL-LEXEME: WiC pretrained model for cross-lingual LEXical sEMantic changE. In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 1577–1585. Association for Computational Linguistics, Toronto, Canada (Jul 2023)
12. Chung, H.W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, E., Wang, X., Dehghani, M., Brahma, S., Webson, A., Gu, S.S., Dai, Z., Suzgun, M., Chen, X., Chowdhery, A., Narang, S., Mishra, G., Yu, A., Zhao, V., Huang, Y., Dai, A., Yu, H., Petrov, S., Chi, E.H., Dean, J., Devlin, J., Roberts, A., Zhou, D., Le, Q.V., Wei, J.: Scaling instruction-finetuned language models (2022)
13. Danescu-Niculescu-Mizil, C., West, R., Jurafsky, D., Leskovec, J., Potts, C.: No country for old members: user lifecycle and linguistic change in online communities. In: Proceedings of the 22nd International Conference on World Wide Web. p. 307–318. WWW '13, Association for Computing Machinery, New York, NY, USA (2013)
14. Demirbas, A.Y., Hossain, J., Sarıyüce, A.E.: From words to actions: A comprehensive approach to identifying incel behavior on reddit. In: 2024 IEEE International Conference on Big Data (BigData). pp. 5745–5754 (2024)
15. Farrell, T., Araque, O., Fernandez, M., Alani, H.: On the use of jargon and word embeddings to explore subculture within the reddit's manosphere. In: Proceedings of the 12th ACM Conference on Web Science. p. 221–230. WebSci '20, Association for Computing Machinery, New York, NY, USA (2020)
16. Fontanella, L., Chulvi, B., Ignazzi, E., Sarra, A., Tontodimamma, A.: How do we study misogyny in the digital age? A systematic literature review using a computational linguistic approach. Palgrave Communications 11(1), 1–15 (December 2024), https://ideas.repec.org/a/pal/palcom/v11y2024i1d10.1057_s41599-024-02978-7.html
17. Gadetsky, A., Yakubovskiy, I., Vetrov, D.: Conditional generators of words definitions. In: Gurevych, I., Miyao, Y. (eds.) Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 266–

271. Association for Computational Linguistics, Melbourne, Australia (Jul 2018), https://aclanthology.org/P18-2043/
18. Gardner, N., Khan, H., Hung, C.C.: Definition modeling: literature review and dataset analysis. Applied Computing and Intelligence 2(1), 83–98 (2022)
19. Giulianelli, M., Luden, I., Fernandez, R., Kutuzov, A.: Interpretable word sense representations via definition generation: The case of semantic change analysis (2023), https://arxiv.org/abs/2305.11993
20. Gonzales, A.L., Hancock, J.T., Pennebaker, J.W.: Language style matching as a predictor of social dynamics in small groups. Communication Research 37(1), 3–19 (2010)
21. Gothard, K., Dewhurst, D.R., Minot, J.R., Adams, J.L., Danforth, C.M., Dodds, P.S.: "the incel lexicon: Deciphering the emergent cryptolect of a global misogynistic community" (2021), https://arxiv.org/abs/2105.12006
22. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models (2021), https://arxiv.org/abs/2106.09685
23. Huang, J., Shao, H., Chang, K.C.C., Xiong, J., mei Hwu, W.: Understanding jargon: Combining extraction and generation for definition modeling (2022), https://arxiv.org/abs/2111.07267
24. Ishiwatari, S., Hayashi, H., Yoshinaga, N., Neubig, G., Sato, S., Toyoda, M., Kitsuregawa, M.: Learning to describe unknown phrases with local and global contexts. In: Burstein, J., Doran, C., Solorio, T. (eds.) 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). pp. 3467–3476. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019)
25. Klein, E., Golbeck, J.: A lexicon for studying radicalization in incel communities. In: Proceedings of the 16th ACM Web Science Conference. p. 262–267. WEBSCI '24, Association for Computing Machinery, New York, NY, USA (2024), https://doi.org/10.1145/3614419.3644005
26. Lin, C.Y.: ROUGE: A package for automatic evaluation of summaries. In: Text Summarization Branches Out. pp. 74–81. Association for Computational Linguistics, Barcelona, Spain (Jul 2004), https://aclanthology.org/W04-1013/
27. Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegreffe, S., Alon, U., Dziri, N., Prabhumoye, S., Yang, Y., Gupta, S., Majumder, B.P., Hermann, K., Welleck, S., Yazdanbakhsh, A., Clark, P.: Self-refine: Iterative refinement with self-feedback (2023), https://arxiv.org/abs/2303.17651
28. Madaan, N., Manjunatha, A., Nambiar, H., Goel, A.K., Kumar, H., Saha, D., Bedathur, S.: Detail : A tool to automatically detect and analyze drift in language (2022), https://arxiv.org/abs/2211.04250
29. Md Saroar Jahan, M.O.: "a systematic review of hate speech automatic detection using natural language processing". Neurocomputing 546 (2023), http://dx.doi.org/10.1145/3672393
30. Pelzer, B., Kaati, L., Cohen, K., Fernquist, J.: Toxic language in online incel communities. SN Social Sciences 1 (08 2021)
31. Periti, F., Alfter, D., Tahmasebi, N.: Automatically generated definitions and their utility for modeling word meaning. In: Al-Onaizan, Y., Bansal, M., Chen, Y.N. (eds.) Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. pp. 14008–14026. Association for Computational Linguistics, Miami, Florida, USA (Nov 2024)

32. Periti, F., Montanelli, S.: "lexical semantic change through large language models: a survey". ACM Computing Surveys 56(11), 1–38 (Jun 2024), http://dx.doi.org/10.1145/3672393
33. Post, M.: A call for clarity in reporting BLEU scores. In: Proceedings of the Third Conference on Machine Translation: Research Papers. pp. 186–191. Association for Computational Linguistics, Belgium, Brussels (Oct 2018), https://www.aclweb.org/anthology/W18-6319
34. Puertas, E., Moreno Sandoval, L.G., Redondo, Alvarado, J., Pomares Quimbaya, A.: Detection of sociolinguistic features in digital social networks for the detection of communities. Cognitive Computation 13, 20 (03 2021)
35. Reimers, N., Gurevych, I.: Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In: Inui, K., Jiang, J., Ng, V., Wan, X. (eds.) Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 3982–3992. Association for Computational Linguistics, Hong Kong, China (Nov 2019)
36. Stekel, M., Azaria, A., Gordin, S.: Word sense induction with attentive context clustering. In: Hämäläinen, M., Alnajjar, K., Partanen, N., Rueter, J. (eds.) Proceedings of the Workshop on Natural Language Processing for Digital Humanities. pp. 144–151. NLP Association of India (NLPAI), NIT Silchar, India (Dec 2021), https://aclanthology.org/2021.nlp4dh-1.17/
37. Stephane J. Baele, L.B., Coan, T.G.: From "incel" to "saint": Analyzing the violent worldview behind the 2018 toronto attack. Terrorism and Political Violence 33(8), 1667–1691 (2021)
38. Torregrosa López, F.J., Bello Orgaz, G., Martínez-Cámara, E., Del Ser, J., Camacho, D.: A survey on extremism analysis using natural language processing: definitions, literature review, trends and challenges. Journal of Ambient Intelligence and Humanized Computing 14 (01 2022)
39. Wang, M., Wang, Y.: Word sense disambiguation: Towards interactive context exploitation from both word and sense perspectives. In: Zong, C., Xia, F., Li, W., Navigli, R. (eds.) Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 5218–5229. Association for Computational Linguistics, Online (Aug 2021)
40. Waśniewska, M.: The red pill, unicorns and white knights: Cultural symbolism and conceptual metaphor in the slang of online incel communities. Cultural conceptualizations in language and communication pp. 65–82 (2020)
41. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., Zhou, D.: Chain-of-thought prompting elicits reasoning in large language models (2023), https://arxiv.org/abs/2201.11903
42. Weimann, G., Am, A.B.: Digital dog whistles: The new online language of extremism. International Journal of Security Studies 2(1), 4 (2020)
43. Zhang*, T., Kishore*, V., Wu*, F., Weinberger, K.Q., Artzi, Y.: Bertscore: Evaluating text generation with bert. In: International Conference on Learning Representations (2020), https://openreview.net/forum?id=SkeHuCVFDr
44. Zheng, L., Chiang, W.L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E.P., Zhang, H., Gonzalez, J.E., Stoica, I.: Judging llm-as-a-judge with mt-bench and chatbot arena (2023), https://arxiv.org/abs/2306.05685
45. Škvorc, T., Robnik-Šikonja, M.: Solving word-sense disambiguation and word-sense induction with dictionary examples (2025), https://arxiv.org/abs/2503.04328