# ParsCN: A Persian Dataset for Counter-Narrative Generation to Combat Online Hate Speech

**Zahra Safdari Fesaghandis**[1], **Suman Kalyan Maity**[2]
[1]Bilkent University, Ankara, Turkey
[2]Missouri University of Science and Technology, Rolla, MO, USA
[1]zahra.safdari@bilkent.edu.tr, [2]smaity@mst.edu

## Abstract

Online hate speech threatens online civility, particularly in low-resource and multilingual environments. Counter-narratives offer a promising solution by promoting constructive responses to hate speech. However, automatic counter-narrative generation is hindered by the lack of high-quality data for low-resource languages like Persian. To bridge this gap, we introduce ParsCN, the first and most comprehensive Persian counter-narrative dataset. Consisting of 1,100 hate speech and counter-narrative pairs, it provides fine-grained annotations across six target groups and six countering strategies, tailored to the socio-cultural context of Persian online discourse. We propose a novel, scalable multi-stage framework that integrates culturally-informed human annotation with few-shot LLM-augmented generation, guided by semantic retrieval and rigorous manual curation. This approach enables the creation of diverse, high-quality counter-narratives while significantly reducing annotation costs—establishing a replicable paradigm for other low-resource settings. Comprehensive human and automatic evaluations confirm the quality of the dataset and the effectiveness of the generated responses. Human-written counter-narratives achieved the highest scores for relevance (4.23), Effectiveness (4.21), fluency (4.92), and tone appropriateness (4.79), with GPT-4o and Claude closely following. Automatic evaluations show strong semantic alignment (BERTScore F1 up to 0.709), high lexical diversity, and low toxicity across all sources. Finally, we conduct benchmark evaluations using mBART and PersianMind on a held-out test set. Results reveal that existing models struggle with fluency, cultural nuance, and safety—highlighting the need for Persian-specific resources like ParsCN. Our dataset serves as a foundational benchmark to advance research on Persian counter-narrative generation and foster safer, more inclusive digital spaces.

**Code** — https://github.com/zahrasafdari/ParsCN
**Datasets** — https://doi.org/10.5281/zenodo.18266930

## Introduction

Social media has profoundly reshaped human communication, facilitating global exchange while simultaneously magnifying challenges such as the spread of hate speech. Defined as language discriminating against individuals or groups based on ethnicity, religion, or gender, hate speech poses significant social and technical challenges, given the diversity of cultural norms, legal frameworks, and linguistic structures (Warner and Hirschberg 2012; Watanabe, Bouazizi, and Ohtsuki 2018; Fortuna and Nunes 2018). Traditional moderation methods like automated detection and removal face inherent limitations, including scale, censorship concerns, and difficulty capturing nuanced, evolving hateful language (Hee et al. 2024).

To address these limitations, counter-narratives have emerged as a proactive and non-confrontational alternative. They involve non-aggressive responses that challenge hateful narratives and promote civil discourse (Chung et al. 2019; Mathew et al. 2019), recognized as a scalable means to cultivate inclusive digital spaces. However, a major barrier to automating counter-narrative generation—especially for languages beyond English—is the scarcity of high-quality, task-specific datasets. This gap is particularly critical for low-resource languages like Persian, where distinct linguistic and socio-cultural factors shape both hate speech and effective counter responses.

We present ParsCN, a novel dataset for generating counter-narratives to mitigate hate speech in Persian.

Our key contributions are as follows:

- **ParsCN Dataset:** We present and publicly release ParsCN, the first dedicated dataset for counter-narrative generation in Persian. The dataset comprises 1,100 hate speech–counter-narrative pairs, annotated across six target groups and six distinct counter-narrative strategies, offering a nuanced and structured representation of online hate and its potential responses.
- **Addressing the Low-Resource Gap:** It addresses the vast resource gap for counter-narrative generation in low-resource languages, providing a high-quality, foundational dataset that is particularly well adapted to the unique linguistic and socio-cultural context of Persian online discourse.
- **Novel Multi-Stage Framework:** We present and utilize a novel multi-stage framework for dataset creation that effectively leverages expert human annotation with the assistance of state-of-the-art large language models (LLMs) and automated methods, thereby proposing a scalable framework for the creation of comparable re-

sources.

By releasing ParsCN, we aim to catalyse further research in automated counter-narrative generation for low-resource settings, contributing to the broader effort of fostering safer and more inclusive digital environments.

## Related Work

Counter-narrative generation is an active research area focused on combating online hate speech. This section reviews key datasets and generation methodologies, highlighting the novelty of our multi-stage framework for ParsCN in the context of existing approaches.

### Counter-Narrative Dataset Creation

Creating high-quality datasets is fundamental. Early efforts like CONAN (Chung et al. 2019) pioneered multilingual (English, French, Italian) expert-based datasets using nichesourcing. MultiTarget-CONAN (Fanton et al. 2021) introduced a human-in-the-loop approach for multi-target hate speech. More recently, datasets have expanded linguistically and more comprehensively, including IndicCONAN (Sahoo, Beria, and Bhattacharyya 2024) for Indian languages, PANDA (Bennie et al. 2025b) as the first Chinese dataset leveraging LLM-as-a-Judge, and CONAN-EUS (Bengoetxea et al. 2024) for Basque and Spanish via translation and post-editing, demonstrating models for low-resource contexts (Sahoo, Beria, and Bhattacharyya 2024; Bengoetxea et al. 2024). Diverse sourcing methods like crowdsourcing (Saha et al. 2024) and utilizing in-the-wild data from platforms like YouTube (Fialho et al. 2024) have provided varied resources. Datasets focusing on specific aspects like counter-narrative types (Saha et al. 2024; Sahoo, Beria, and Bhattacharyya 2024) and intent (Hengle et al. 2024) have also been developed. Despite these advances, a significant gap remains for many languages, particularly low-resource datasets like Persian.

ParsCN's novelty extends beyond being the first Persian CN dataset. It introduces: (1) a culturally tailored annotation framework addressing Persian-specific hate speech (e.g., targeting Afghan immigrants or Iran's regime); (2) fine-grained annotations across six target groups and six CN strategies; (3) a hybrid data sourcing approach combining Persian PHATE data with translated MultiTarget-CONAN and HateXplain instances, rigorously post-edited for cultural relevance; and (4) a scalable multi-stage framework integrating culturally informed human annotation, semantic retrieval using a SentenceTransformer model for few-shot LLM prompting, and rigorous manual curation of LLM outputs. Unlike simpler pipelines in IndicCONAN (direct human annotation) or CONAN-EUS (translation-based), ParsCN's framework prioritizes cultural relevance and scalability for low-resource settings. A detailed comparison with other low-resource CN datasets is provided in Appendix A.3, Table 8.

### Counter-Narrative Generation Methodologies

Various techniques have been explored for automated generation. Approaches leveraging external knowledge grounding (Chung, Tekiroglu, and Guerini 2021; Wilk et al. 2025) aim for more informative responses. Methods incorporating structural aspects like argumentative (Furman et al. 2023) or discourse structure (Hassan and Alikhani 2023) seek better control and quality. Constraint-based generation, such as optimizing for desired conversation outcomes using LLMs (Mun et al. 2023), addresses real-world impact. LLMs are central to many systems (Podolak et al. 2024), including retrieval-augmented approaches like ReZG (Jiang et al. 2025) for zero-shot generation. Research also analyzes the use and effectiveness of specific strategies for countering implied biases (Mun et al. 2023). Furthermore, developing multilingual models and context-aware approaches (Bennie et al. 2025a) is crucial for extending capabilities, especially in low-resource settings.

## Dataset

In this section, we present our **ParsCN** dataset designed to facilitate counter-narrative research in the Persian language, addressing the pressing issue of online hate speech in Persian-speaking communities. This dataset comprises pairs of hate speech and corresponding counter-narratives, annotated with hate speech target categories and counter-narrative types. **ParsCN** is the first dataset of its kind in Persian, filling a critical gap in resources for non-English counter-narrative research. Below, we describe the dataset's characteristics, the two-stage process of data collection involving manual and automated annotation, and the annotation strategies employed to ensure quality and diversity. Figure 1 illustrates the two-stage process followed for the creation of the **ParsCN** dataset, combining data sourcing, human annotation, and automated generation techniques.

### Hate Speech Datasets

To construct **ParsCN**, we sourced hate speech instances from three publicly available datasets, adapting them to Persian linguistic and cultural contexts:

**PHATE:** A Persian dataset of 7,000 manually annotated tweets labeled as normal or hate (subcategorized as violence, hate, or vulgar) (Delbari, Moosavi, and Pilehvar 2024). Annotations include target groups and justification spans, making it directly applicable to ParsCN.

**MultiTarget-CONAN:** An English dataset of $\sim$5,000 hate speech–counter-narrative pairs covering race, religion, and gender (Fanton et al. 2021). Created using a human-in-the-loop process with GPT-2 and NGO experts, it includes Persian-relevant targets (e.g., Muslims). We translated and manually validated culturally appropriate instances.

**HateXplain:** A 20,148-post English dataset from Twitter and Gab, annotated as hate, offensive, or normal (Mathew et al. 2021). We selected posts targeting Persian-relevant groups (e.g., ethnic and occupational), translating and validating them for cultural fit.

Together, these datasets provided a diverse pool of hate speech instances, which we filtered, translated, and annotated to construct ParsCN's hate–counter-narrative pairs.

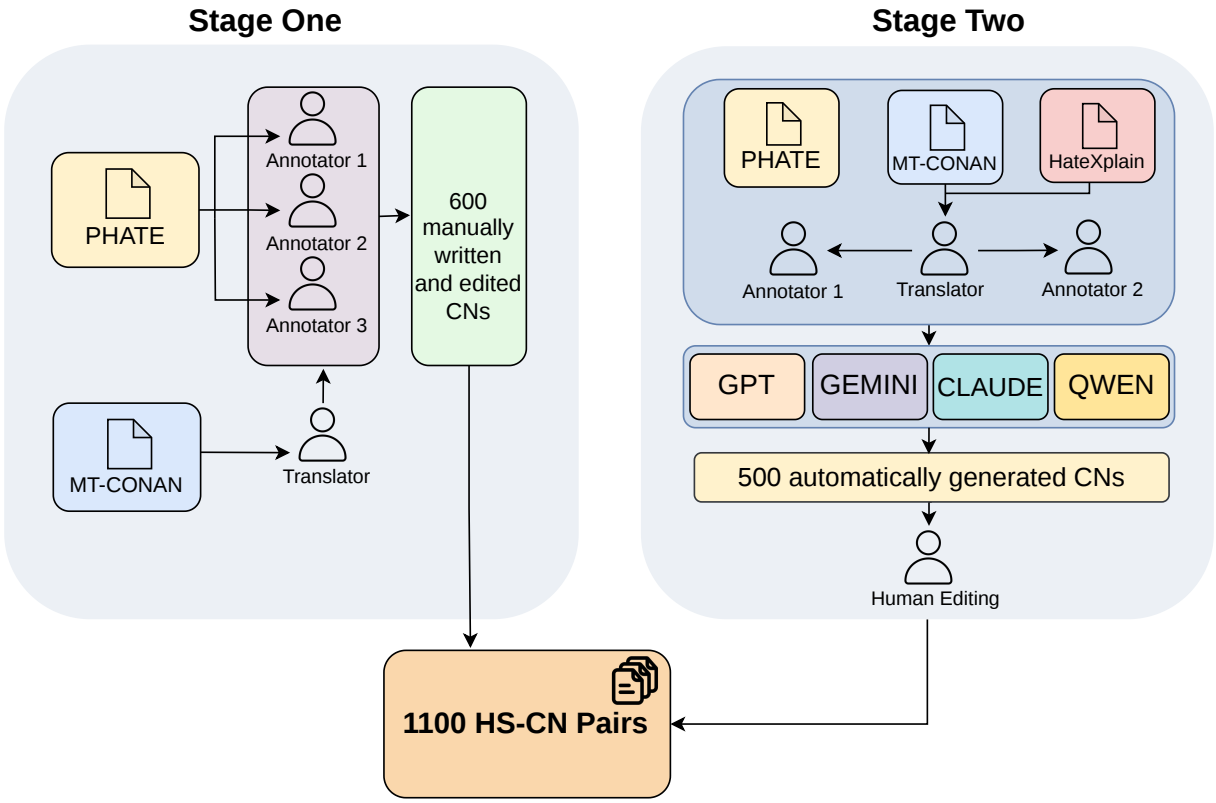


Figure 1: ParsCN Dataset Creation Pipeline

## Hate Speech Categories

To enable fine-grained analysis and generation of counter-narratives, we categorize hate speech in **ParsCN** into six types, directly adopted from the PHATE dataset Delbari, Moosavi, and Pilehvar (2024) and tailored to the socio-cultural context of Persian-speaking regions (Delbari, Moosavi, and Pilehvar 2024). These categories are as follows - **i) Religious**: Statements targeting religious groups, such as Islam or the Jewish religion; **ii) Racial**: Speech discriminating against ethnic/racial groups like Azeris, Kurds, or Black Africans; **iii) Gender**: Statements derogatory to men, women, and including activists; **iv) Political**: Expressions targeting the government, politicians, or country laws; **v) Occupational**: Statements targeting professions, such as police or teachers; **vi) National**: Speech targeting national or immigrant groups like Afghans, Chinese, or Arabs. Table 1 provides one example per category, showcasing the dataset's diversity.

## Counter-Narrative Approaches

To produce effective counter-narratives against hate speech in Persian online platforms, **ParsCN** incorporates a diverse set of six counter-narrative strategies, adapted from (Sahoo, Beria, and Bhattacharyya 2024) (see Table 2 and Appendix A.2). These strategies are tailored to address hate speech in six target categories: Religious, Racial, Gender, Political, Occupational, and National. Each type, represented in **ParsCN**'s dataset (see Table 1 for a few examples from our dataset), ensures contextually appropriate counter-narratives for Persian hate speech.

## Stage One: Data Collection and Annotation

In the first stage of **ParsCN** development, we constructed a dataset of 600 hate speech–counter-narrative pairs to facilitate automated counter-narrative generation in Persian across the target categories. This phase involved curating a diverse set of hate speech instances, generating contextually appropriate counter-narratives based on the PHATE dataset, and incorporating translated examples from MT-CONAN. All entries were annotated according to the six counter-narrative strategies outlined in the Counter-Narrative Approaches section. The resulting dataset–balanced with 100 instances per category–provides a robust foundation for model training in Stage Two. A summary of the dataset composition is provided in Table 3, with further methodological details described below.

**Data Sourcing** To address the scarcity of Persian hate speech data, we sourced 450 hate speech instances from the PHATE dataset (Delbari, Moosavi, and Pilehvar 2024), selected to represent diverse expressions of hate across six categories. However, PHATE provided only 20 and 30 unique and acceptable instances for Gender and Racial categories, respectively, necessitating supplementation to achieve 100 instances per category. To ensure cultural relevance for Persian contexts, we carefully selected 150 instances from

| Target Group | Hate Speech (Persian/English) | Counter Narrative (Persian/English) | Counter Type |
|---|---|---|---|
| Religious | اسلام اگه خوب بود، خدا برای خودش زن. مسیحی نمی گرفت<br>If Islam were good, God wouldn't have taken a Christian wife for himself. | اگه منظورت حضرت مریمه که اونموقع هنور حرفی از اسلام زده نشده بود و حدود هزار سال بعدش اسلام و قرآن توسط حضرت محمد توسعه پیدا کرد<br>If you mean Mary, there was no mention of Islam at that time, and about a thousand years later, Islam and the Quran were developed by Muhammad. | Fact-based |
| Racial | من از سیاه‌پوستان متنفرم: آن‌ها یک پلیس را! کشتند<br>I hate black people: they killed a policeman! | را مقصر می‌دانی؟ فرض کن در گروهی ۱۰ نفره هستی و یکی قاتل است، آیا این باعث می‌شود همه قاتل باشند؟ جدی می‌گویی؟<br>Do you blame them all? Suppose you are in a group of 10 people and one is a murderer, does that make them all murderers? Are you serious? | Counter Questions |
| Gender | زنی که به غریبه رواست، به شوهر حرام است<br>A woman who is permissible to a stranger is forbidden to her husband. | بر چه اساسی این حرف را میزنید؟ قضاوت درباره دیگران بدون شناخت شرایط و انتخاب‌هایشان، تنها باعث ایجاد فاصله و سوءتفاهم می‌شود<br>On what basis do you say this? Judging others without understanding their circumstances and choices only creates distance and misunderstanding. | Warning of Conse-quences, Counter Question |
| Political | بخاطر یه حجاب همه رو تیکه تیکه کردن. وحشیا<br>They tore everyone to pieces because of a hijab. savages! | بیا فحش دادن را بگذاریم کنار و وقتمان را برای رسیدن به هدفمان بگذاریم<br>Let's stop cursing and take our time to achieve our goal. | Positive Response |
| Occupa-tional | متاسفانه اگه مشکلی برات پیش بیاد هرچقدر زنگ بزنی به نیروی انتظامی بهونه نیارن ک نیان بعدم ک میان میگن نمیخاد بری شکایت کنی دستت ب جایی بند نیس.فقط کافیه یجا پارتی گرفته باشن یا ی تجمع باشه مثل مور و ملخ میریزن اونجا<br>Unfortunately, if something happens to you, no matter how many times you call the police, they'll make excuses not to come. And when they do show up, they say, 'No need to file a complaint — it won't get you anywhere.' But if there's a party or a public gathering somewhere, they swarm the place like ants and locusts. | فکر کنم شما براساس یک تجربه شخصی اینطوری قضاوت میکنید. اگر شکایتی نسبت به عملکرد نیروی های انتظامی دارید بهتر هست با نهاد های مرتبط در میان بگذارید<br>I think you're judging based on a personal experience. If you have a complaint about the performance of the police forces, it's better to raise it to the relevant authorities. | Positive Response, Denouncing |
| National | ترودو، و وزیر مهاکت سومالیایی‌اش، احمد حسان، می‌خواهند ۱۰،۰۰۰ سومالیایی عقب‌مانده را به کانادا بیاورند تا این کشور را با احمق‌ترین مردم زمین پر کنند<br>The idiot son, Trudeau, and his Somali immigration minister, Ahmed Hussen, want to bring 10,000 backward Somalis to Canada to fill the country with the dumbest people on Earth. | سومالیایی‌ها نه تنها عقب‌مانده نیستند، بلکه در جوامع میزبان هم نقش‌آفرین و موفق‌اند؛ مهاجرت فرصتی برای رشد همگان است، نه تهدید<br>Somalis are not only not backward, but they are also active and successful members of host communities; migration is an opportunity for everyone's growth, not a threat. | Contradic-tion, Positive Response |

Table 1: Examples of Hate Speech and Counter Narratives, Along with Their Target Groups and Counter Types, from ParsCN

| Type | Purpose |
|---|---|
| Positive Response | Fosters empathy and inclusion |
| Counter questions | Prompts reflection on biases |
| Denouncing | Rejects harmful rhetoric |
| Fact-based | Corrects misinformation with evidence |
| Warning of consequences | Highlights negative outcomes |
| Contradiction | Exposes logical inconsistencies |

Table 2: Summary of Counter-Narrative Types in ParsCN

| Source | Instances | Annotation |
|---|---|---|
| PHATE | 450 | Wrote CNs and labeled their types |
| MT-CONAN | 150 | Translated CNs and labeled their types |

Table 3: ParsCN Dataset Composition

MultiTarget-CONAN (Fanton et al. 2021) and HateXplain (Mathew et al. 2021). From MultiTarget-CONAN, we chose 80 instances targeting Women for the Gender category, reflecting prevalent Persian socio-cultural issues like gender-based discrimination (e.g., claims that women are unfit for politics, countered with arguments for gender equality). For the Racial category, we selected 70 instances targeting People of Color, ensuring alignment with ethnic tensions in Persian discourse (e.g., derogatory remarks about Arabs, countered with calls for humanity). From HateXplain, we selected instances targeting groups like Chinese and Arabs, which frequently appear in Persian online discourse due to geopolitical and cultural interactions. Full examples, with Persian and English translations to illustrate cultural fit are shown in Table 4 (see Appendix A.5 for details). All instances from MultiTarget-CONAN and HateXplain were translated into Persian using Google Translate and manually post-edited by native Persian annotators to ensure linguistic precision, grammatical accuracy, and cultural congruence, aligning with Persian-specific target categories. This process yielded a balanced dataset of 600 instances, ensuring that translated samples reflect the socio-cultural nuances of Persian online discourse. To mitigate potential translation bias, our post-editing process functioned as content localization rather than literal translation. We replaced Western-centric hate tropes with concepts relevant to Persian socio-cultural dynamics (e.g., shifting general racial slurs to specific regional ethnic tensions). The success of this approach is evidenced by our human evaluation results (Section titled "Human Evaluation" and Appendix A.9), where translated-and-adapted samples achieved strong pair-level human evaluation scores that were broadly comparable to native Persian instances, especially in fluency, semantic clarity, and tone appropriateness.

**Counter-Narrative Annotation** Securing multiple annotations for a low-resource task like Persian counter-narrative generation is inherently challenging due to the limited availability of individuals who are both fluent in Persian and experienced in hate speech research. This constraint is also evident in prior work, such as the IndicCONAN dataset (Sahoo, Beria, and Bhattacharyya 2024), which relied on only two annotators. In our case, annotation was conducted by three native Persian speakers with prior research experience in hate speech and counter speech (See Appendix A.6 for detailed annotator information). For the 600 human-generated hate speech–counter-narrative pairs, the annotation process began with a shared batch of 80 samples, which all three annotators worked on independently. They then discussed their responses collaboratively to identify and correct mistakes, refining their annotation strategy and aligning their interpretations to ensure high-quality outputs. After this initial phase, the remaining samples were distributed among the annotators for individual annotation. For each hate speech instance, annotators generated a single counter-narrative, guided by comprehensive guidelines adapted from PEN America and Get The Trolls Out ( https://tinyurl.com/ycxct6wh). These guidelines outlined strategies for crafting respectful, empathetic, and effective counter-narratives, emphasizing the use of credible information and appropriate tone. Counter-narrative type selection (e.g., Positive Response, Denouncing, Fact-based, as defined in Appendix A.2) was informed by these guidelines and the context of the hate speech, with a focus on producing impactful and culturally resonant responses. While no rigid template was used—preserving cultural nuance—the use of guidelines and initial collaborative annotation helped minimize subjectivity. For example, a gender-based hate speech suggesting that women are untrustworthy was countered with: "Labeling all women this way is entirely unjustified." Throughout the annotation process, we aimed to balance diversity in counter-narrative types—avoiding over-reliance on strategies like Denouncing—while maintaining cultural sensitivity. Each counter-narrative was labeled with its type, in accordance with the framework outlined in the Counter-Narrative Approaches section.

**MT-CONAN Integration** To enhance dataset diversity, the 150 MT-CONAN instances were paired with their original counter-narratives, translated into Persian using Google Translate and post-edited by annotators for fluency and cultural relevance. Annotators labeled each translated counter-narrative with its corresponding type, ensuring consistency with the PHATE annotations. This process, focused on the Gender (80 women-targeted instances) and Racial (70 poc targeted instances) categories, introduced varied counter-narrative styles while maintaining dataset coherence. Quality control during post-editing addressed challenges like preserving the original counter-narratives' intent in Persian contexts. The combined dataset of 600 hate speech–counter-narrative pairs provides a robust foundation for Stage Two's model training.

| Hate Speech (Persian) | Counter-Narrative (Persian) | Hate Speech (English) | Counter-Narrative (English) | Source |
|---|---|---|---|---|
| | | **MultiTarget-CONAN** | | |
| به همین دلیل زنان جایی در سیاست ندارند. اگر مردی مثل آن‌ها شکایت می‌کرد، ترسو خطابش می‌کردند و می‌گفتند از کار بیرون برو. برابری یک دروغ است و فمینیسم حکم مرگ هر تمدنی است. | برابری جنسیتی کلید یک خانواده و اقتصاد مولد است. محدود کردن زنان از ورود به همه حوزه‌های جامعه، پتانسیل هر کشوری را کاهش می‌دهد. | This is why women have no place in politics. If a man complained like them, he'd be called a coward and told to quit his job. Equality is a lie, and feminism is the death knell of any civilization. | Gender equality is key to a productive family and economy. Restricting women from entering all societal domains reduces a country's potential. | Gender |
| از اینکه زنان در بحث‌های سیاسی شرکت می‌کنند خسته شده‌ام. فمینیسم یک نیروی تفرقه‌افکن در جامعه است. | فمینیسم دیدگاه جدیدی به مسائل اجتماعی و سیاسی سنتی اضافه می‌کند. | I'm tired of women participating in political discussions. Feminism is a divisive force in society. | Feminism adds a fresh perspective to traditional social and political issues. | Gender |
| زنان قادر به مدیریت مسئولیت‌های عمومی نیستند. آن‌ها باید در خانه بمانند و بچه‌ها را بزرگ کنند. | زنان بیش از حد قادر به تصدی موقعیت‌های قدرتمند در جامعه ما بوده‌اند و اغلب با مناصب عمومی پاداش می‌گیرند که کمتر کسی جرأت آرزوی آن را دارد. | Women are incapable of handling public responsibilities. They should stay home and raise children. | Women have proven more than capable of holding powerful positions in our society, often earning public roles that few dare to aspire to. | Gender |
| | | **HateXplain** | | |
| ظاهراً در یمن وبا شیوع پیدا کرده، چند عرب کمتر. | وبا یک بیماری است و ربطی به ملیت ندارد. آیا انسانیت ما حکم نمی‌کند که برای سلامتی همه نگران باشیم؟ | Apparently, cholera has spread in Yemen—fewer Arabs, then. | Cholera is a disease, not tied to nationality. Doesn't our humanity compel us to care about everyone's health? | National |
| اگر اینجا را دوست ندارند، باید گورشان را از سرزمین‌های ما گم کنند و به کشور "کثافت عربی" برگردند. جنگ صلیبی! مسلمانان، بیرون! | آیا بهتر نیست به جای نفرت‌پراکنی و جنگ‌طلبی، با احترام و تفاهم در کنار هم زندگی کنیم؟ همه ما انسان هستیم و سزاوار صلح و آرامشیم. | If they don't like it here, they should get lost from our lands and go back to their "filthy Arab" country. Crusade! Muslims, out! | Isn't it better to live together with respect and understanding instead of spreading hate and warmongering? We're all human and deserve peace. | National |
| من کمی گیج هستم زیرا مردم چین نمی‌توانند به توییتر دسترسی پیدا کنند. این "چینگ چانگ" چگونه از آن استفاده می‌کند؟ فکر کنم پاکستانی باشه. | ممکنه بهتر نباشه به جای تمسخر، از خودمون بپرسیم چرا دیدن موفقیت یک نفر از آسیای شرقی یا جنوبی انقدر ما رو به قضاوت می‌کشونه؟ مگر انسان‌ها را با دسترسی به شبکه‌های اجتماعی قضاوت می‌کنیم یا با شعور و شخصیت‌شون؟ | I'm a bit confused since Chinese people can't access Twitter. How does this "ching chong" use it? Must be Pakistani. | Wouldn't it be better, instead of mocking, to ask ourselves why seeing someone from East or South Asia succeed prompts us to judge? Do we judge people by their social media access or their character? | National |

Table 4: Examples of Culturally Relevant Hate Speech and Counter-Narratives from MT-CONAN and HateXplain with English Translations

## Stage Two: Automated Counter-Narrative Generation

The second stage focused on expanding the dataset by leveraging LLMs for automated counter-narrative generation. This stage aimed to produce an additional 500 hate speech–counter-narrative pairs across five target categories, utilizing the manually curated Stage One data as few-shot examples to guide the LLMs towards generating high-quality, contextually relevant responses in Persian. The process involved sourcing new hate speech instances, retrieving relevant examples from Stage One, prompting the LLMs for generation, and performing manual curation and quality control on the generated outputs.

**Expansion of Dataset Sourcing** To provide the LLMs with new hate speech instances for generating counter-narratives, we sourced 100 new hate speech instances for each of the five target categories: Religious, Political, Gender, Racial, and National. Instances for the Religious and Political categories were collected from the PHATE dataset (Delbari, Moosavi, and Pilehvar 2024). For the Gender (specifically targeting women) and Racial categories, we selected instances from the MT-CONAN dataset (Fanton et al. 2021). Instances for the National category were sourced from the HateXplain dataset (Mathew et al. 2021). We carefully selected these instances to ensure their relevance and appropriateness for generating counter-narrative responses in the Persian context, as we discussed in the Data Sourcing

section. The Occupational target group was excluded in this stage due to difficulty in identifying a sufficient number of high-quality instances suitable for counter-narrative generation that aligned with the dataset's goals. For the English-language sources (MT-CONAN and HateXplain), the selected hate speech instances were translated into Persian using Google Translate and subsequently reviewed and manually edited by annotators to ensure linguistic accuracy, fluency, and cultural coherence within the Persian online discourse.

**Retrieval of Relevant Examples** To implement the few-shot prompting strategy effectively, we developed a method to retrieve the most semantically similar hate speech–counter-narrative pairs from the Stage One dataset (600 pairs) for each new hate speech instance collected in Stage Two. We utilized a pre-trained SentenceTransformer model, specifically `'paraphrase-multilingual-mpnet-base-v2'`, known for its effectiveness in generating multilingual sentence embeddings. Each new hate speech instance from Stage Two was embedded using this model, as were all hate speech instances from the Stage One dataset. Cosine similarity was then calculated between the embedding of the new instance and the embeddings of all Stage One instances. For each new hate speech instance, the top 10 most similar pairs from Stage One, along with their similarity scores and metadata (original hate speech, counter-narrative, target group, counter type(s)), were retrieved. These retrieved pairs served as examples in the LLM prompts, providing contextual guidance on the desired style and content of effective Persian counter-narratives for similar hate speech. The retrieved pairs were manually reviewed before inclusion in the prompts to select the most relevant and high-quality examples.

**LLM-Based Counter-Narrative Generation** The automated generation of counter-narratives was performed using four LLMs: GPT-4o, Gemini 2.0 Flash, Claude 3.7 Sonnet, and Qwen3-235B-A22B, selected for their strong multilingual capabilities. A few-shot prompting approach was employed, where the input to the LLM included the new hate speech instance requiring a response, along with the corresponding hate speech-counter-narrative pairs retrieved from Stage One, and explicit instructions regarding the desired output format and content. The prompt instructed the models to generate a short counter-narrative in Persian, ideally between 1 and 2 lines and under 50 words. Furthermore, the prompt encouraged the LLMs to utilize a combination of two counter-narrative types from the predefined list (Positive Response, Counter Questions, Denouncing, Fact-based, Warning of Consequences, Contradiction and Hypocrisy), or to select the single most effective type based on the provided examples and type definitions. Across the five target categories, each of the four LLMs generated 25 counter-narratives per category. This process resulted in a total of 4 LLMs $\times$ 5 categories $\times$ 25 instances/category = 500 automatically generated counter-narratives.

**Persian Counter-Narrative Generation Prompt**

Generate a counter-narrative response in Persian to the hate speech: **{HATE_SPEECH}**. The response must:

- Be a short paragraph (1–2 lines, under 50 words).
- Be in Persian.
- Use a mix of **{TYPE 1}** and **{TYPE 2}** counter-narrative types or select the most effective type from the following, based on these sample counter-narrative responses:
  - **Positive Response**: Statements presenting an optimistic, constructive, or supportive viewpoint to promote understanding, empathy, or positive change.
  - **Counter Questions**: Thought-provoking questions that challenge underlying assumptions, biases, or implications to stimulate critical thinking.
  - **Denouncing**: Statements openly condemning harmful ideas or actions, highlighting negative effects or ethical concerns.
  - **Fact-based**: Statements providing factual evidence to correct misperceptions or misinformation.
  - **Warning of Consequences**: Statements informing of potential negative outcomes of a viewpoint or action.
  - **Contradiction and Hypocrisy**: Statements pointing out inconsistencies or hypocrisy in the original statement.

**Examples:**

- {Hate Speech1}, {Counter-Narrative1}, {Target Group}, {Counter Type}
- {Hate Speech2}, {Counter-Narrative2}, {Target Group}, {Counter Type}
- {Hate Speech n}, {Counter-Narrative n}, {Target Group}, {Counter Type}

**Do not include explanations, notes, or multiple responses—only the counter-narrative paragraph.**

**Manual Curation and Quality Control** Although the counter-narratives were generated using advanced LLMs, a crucial manual curation and quality control step was performed to ensure the high quality and usability of the expanded dataset. All 500 automatically generated counter-narratives were meticulously reviewed by human annotators. This review process focused on several key aspects: evaluating the fluency and grammatical correctness of the Persian text, assessing the cultural appropriateness, verifying adherence to the specified length constraints (under 50 words), and confirming that the generated response effectively employed one or a combination of the intended counter-narrative types. Annotators made necessary edits to correct errors, improve clarity, and ensure that each generated counter-narrative met the dataset's quality standards. This manual refinement step was essential for producing a reliable and high-quality dataset extension for training and evaluating counter-narrative models in Persian.

## FAIR Compliance

The data set has been developed according to the FAIR principles to ensure that it is Findable, Accessible, Interopera-

ble, and Reusable. The dataset is assigned a persistent DOI, making it easily findable. The dataset is hosted on a data-sharing platform that allows for authorized access to researchers, ensuring that it can be retrieved using standard protocols. The data is stored in standard CSV format, which is widely used and compatible with common data analysis tools. This promotes interoperability with various software applications. The dataset is shared under Creative Commons Attribution 4.0 International, which allows for reuse in academic research.

### Potential Research Applications and Usage

This dataset enables the development and evaluation of automated systems for Persian counter-narrative detection, a task that has been largely unexplored due to the lack of annotated resources. It can be used to train and benchmark machine learning models for counter-narrative classification, stance analysis, and response generation. It also supports sociolinguistic research on how Persian-speaking communities respond to online hate, facilitating cross-cultural comparisons and multilingual transfer learning. In practical settings, it can help to create moderation tools that recommend constructive responses, and promote healthier online interactions. Additionally, the resource provides a benchmark for assessing the cultural sensitivity and safety of large language models in Persian. Overall, it contributes to advancing computational methods and real-world interventions for mitigating online hate in Persian digital spaces.

## Data Statistics

This section presents key statistics describing the composition and characteristics of the **ParsCN** dataset.

### Dataset Composition

Table 5 presents a combined view of the dataset's composition by target group and the average word counts for both hate speech instances and their corresponding counter-narratives within each group. These statistics provide insight into the distribution of pairs across the six defined target groups and the typical verbosity of hate speech and counter-narratives. The dataset comprises a total of 1100 hate speech–counter-narrative pairs. We observe variations in average word counts across target groups. The Religious group has the longest average counter-narrative word count (32.04 words), while the Political group has the shortest (23.98 words). We limited counter-narratives to under 50 words (typically 1-3 sentences). This choice is empirically motivated by social media consumption patterns on platforms like Twitter/X, Instagram etc., where concise responses demonstrate significantly higher readability, mobile engagement, and likelihood of dissemination compared to long-form rebuttals.

### Distribution of Counter-Narrative Types

The dataset's counter-narratives are annotated with six primary types, often appearing in combination. Among these, Positive Response is the most frequently occurring type, appearing in 547 counter-narratives, reflecting a strong preference for constructive, empathetic, or inclusive strategies.

| Target Group | Number of Pairs | Average HS Word Count | Average CN Word Count |
|---|---|---|---|
| Gender | 200 | 18.96 | 26.18 |
| Political | 200 | 37.48 | 23.98 |
| National | 200 | 31.85 | 24.55 |
| Racial | 200 | 17.07 | 26.28 |
| Religious | 200 | 35.77 | 32.04 |
| Occupational | 100 | 36.16 | 25.01 |
| **Overall Avg** | **1100** | **29.55** | **26.34** |

Table 5: Dataset Composition and Average Word Counts (HS/CN) by Target Group

Denouncing follows with 400 occurrences, indicating that directly condemning hate speech is also a common approach. Counter Questions appear 323 times, suggesting moderate use of reflective or rhetorical challenges. Less frequent are Fact-based counter-narratives (221), Warning of Consequences (185), and Contradiction (168), which may point to the difficulty of countering emotionally charged hate speech with logic, or to the specific nature of hate speech in the dataset. Overall, the prevalence of Positive Response highlights the dataset's emphasis on fostering positive discourse as a central countering mechanism. This imbalance is partly intentional and partly data-driven. Our annotation guidelines encouraged constructive, respectful, and non-escalatory responses, which made Positive Response broadly applicable across many hate speech scenarios. By contrast, strategies such as Fact-based, Warning of Consequences, and Contradiction often require stronger contextual, factual, or rhetorical grounding and were therefore less universally suitable. We also acknowledge that some portion of the observed distribution may reflect annotator preference, despite our best effort to maintain diversity across response types. We view this distribution not as a flaw, but as an informative characteristic of how counter-narrative is naturally operationalized in Persian online contexts.

## Human Evaluation

To assess the quality of the generated counter-narratives from different sources (LLMs, MT-CONAN, and human writers), we conducted a human evaluation. Two expert annotators evaluated a subset of the generated counter-narratives based on several key metrics.

### Evaluation Metrics

We used the following metrics to evaluate the counter-narratives:

**Relevance**: Measures how well the counter-narrative addresses the original hate speech content and context.

**Effectiveness**: Evaluates the persuasive power or potential impact of the counter-narrative in mitigating the hateful intent or sentiment of the original hate speech.

**Fluency**: Assesses the grammatical correctness and readability of the generated text in Persian.

**Tone Appropriateness**: Evaluates whether the counter-narrative uses a respectful, empathetic, and appropriate tone

without escalating or causing further harm.

### Inter-Annotator Agreement Scores

To ensure the reliability and consistency of the manual annotation process, particularly for counter-narrative quality assessment metrics defined above, we calculated Cohen's Kappa scores between the two independent expert annotators. As the original human evaluation scores were assigned on a continuous-like scale (1 to 5 with 0.5 increments), and Cohen's Kappa is typically used for categorical data, we categorized the scores into three distinct levels: Low' ([1, 1.5, 2]), Medium' ([2.5, 3, 3.5]), and 'High' ([4, 4.5, 5]). This categorization, consistent with prior NLP annotation studies (Landis and Koch 1977), avoids overlap and ensures clarity in distinguishing score ranges. Kappa was then computed for each annotation criterion on a subset of the annotated data used in the human evaluation. The resulting agreement scores demonstrate substantial inter-annotator agreement: 0.724 for Relevance, 0.692 for Effectiveness, 0.655 for Fluency, and 0.770 for Tone Appropriateness. These results indicate a high level of consistency and reliability.

### Evaluation Results

Table 6 presents the average human evaluation scores for each source across the four metrics, as rated by two independent annotators (A1 and A2), along with their combined average. Scores were rated on a scale from 1-5. We observe that the Human-written counter-narratives receive the highest average scores across all metrics, confirming the effectiveness of human intuition, context-awareness, and nuanced language generation in sensitive scenarios. Among the models, GPT-4o and Claude closely follow in performance, especially in relevance and tone appropriateness, showing strong generalization and emotional calibration. Gemini scores well in fluency, while Qwen leads in relevance but slightly lags in fluency. MT-CONAN, while competitive in fluency, exhibits lower effectiveness, indicating a potential trade-off between linguistic smoothness and persuasive impact.

## Automatic Evaluation

To complement human evaluation and provide quantitative insights into the quality and characteristics of the generated counter-narratives, we conducted an automatic evaluation using several widely adopted metrics. This evaluation assessed the fluency, semantic similarity, lexical diversity, and toxicity of counter-narratives generated by different LLMs, translated from MT-CONAN, and written by native speakers.

### Evaluation Metrics and Results

We employed several automatic evaluation metrics to assess different aspects of the generated counter-narratives. Table 7 presents a combined view of the results for Perplexity (Jelinek et al. 1977), Semantic Similarity (BERTScore) (Zhang et al. 2019), Lexical Diversity (Distinct-n) (Li et al. 2015), and Toxicity (Imani et al. 2023) across different data sources (LLMs, MT-CONAN, and Human).

**Perplexity:** LLM-generated CNs generally exhibit lower perplexity than MT-CONAN translations and human samples, suggesting higher fluency and consistency with typical language patterns learned by the models. Human-written CNs show the highest perplexity (158.45). While numerically higher perplexity might seem undesirable, in this context, it is a valuable indicator of native speaker written data's linguistic richness, creativity, and inclusion of more complex or nuanced phrasing that is less predictable by standard language models. This inherent variability is crucial for training models that can generate diverse and human-like responses.

**Semantic Similarity (BERTScore):** MT-CONAN translated CNs achieve the highest semantic similarity scores, indicating strong alignment with their hate speech text. LLM-generated CNs follow closely, demonstrating their capability to produce semantically relevant counter-narratives. Human-written CNs show slightly lower scores and reflect the more varied and interpretive ways humans respond, adding valuable layers of meaning and context.

**Lexical Diversity (Distinct-n):** The MT-CONAN responses (0.892 Distinct-2) and human-written CNs (0.384 Distinct-1, 0.879 Distinct-2) exhibit high lexical diversity, reflecting a broad range of vocabulary and phrasing. LLMs, particularly GPT-4o (0.410 Distinct-1), also demonstrate competitive lexical diversity, balancing fluency with varied language use.

**Toxicity:** We estimated toxicity using the Glot500 classifier (Imani et al. 2023). Lower scores indicate less toxicity. All sources produce predominantly non-toxic counter-narratives. LLM-generated responses, especially from Gemini and Claude, show extremely low toxicity. Human responses maintain a moderate level, balancing assertiveness with civility. MT-CONAN samples have higher toxicity, potentially influenced by translation nuances or the original source's tone.

The automatic evaluation metrics collectively highlight the strengths of the ParsCN dataset, which integrates data from multiple sources. LLMs contribute fluency and semantic relevance, while MT-CONAN data brings high semantic similarity and lexical diversity. Human-authored counter-narratives offer invaluable linguistic richness and realistic variation. The low toxicity scores across all sources confirm the dataset's focus on generating appropriate and non-harmful counter-narrative. This blend of characteristics makes ParsCN a robust and diverse resource for developing and evaluating counter-narrative generation models in Persian.

## Baseline Model Evaluation

To underscore the necessity of ParsCN, we evaluated two baseline models, mBART and PersianMind, on a representative subset of 125 Persian hate speech samples from the ParsCN dataset. We selected this subset to balance diversity across target groups with the practical cost of manual inspection and human evaluation. The experiment setup and results are presented in Appendix A.4 (Tables 9, 10, and 11). The models were assessed using automated metrics (BLEU, ROUGE-L, METEOR, BERTScore F1, per-

| Source | Relevance | | | Effectiveness | | | Fluency | | | Tone Appropriateness | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | A1 | A2 | Avg | A1 | A2 | Avg | A1 | A2 | Avg | A1 | A2 | Avg |
| Claude | 4.175 | 4.088 | 4.132 | **4.050** | 4.138 | 4.094 | 4.850 | 4.800 | 4.825 | 4.850 | 4.488 | 4.669 |
| GPT-4o | 4.162 | 4.125 | 4.144 | 4.037 | 4.188 | 4.113 | 4.750 | 4.688 | 4.719 | 4.825 | 4.562 | 4.693 |
| Gemini | 4.225 | 4.000 | 4.112 | 3.925 | 3.950 | 3.938 | 4.888 | 4.725 | 4.807 | 4.700 | 4.412 | 4.556 |
| Human | 4.150 | 4.312 | **4.231** | 4.025 | 4.388 | **4.206** | 4.950 | 4.888 | **4.919** | 4.962 | 4.612 | **4.787** |
| MT-CONAN | 4.162 | 3.888 | 4.025 | 3.775 | 3.738 | 3.756 | 4.712 | 4.800 | 4.756 | 4.588 | 4.425 | 4.507 |
| Qwen | **4.250** | 4.050 | 4.150 | 3.988 | 4.025 | 4.006 | 4.638 | 4.362 | 4.500 | 4.850 | 4.625 | 4.738 |

Table 6: Human Evaluation Scores by Source and Annotator Average

| | Perplexity | BERTScore | | | Distinct-n | | Toxicity |
|---|---|---|---|---|---|---|---|
| Source | | P | R | F1 | Distinct-1 | Distinct-2 | Score |
| Gemini | **39.15** | 0.660 | 0.666 | 0.662 | 0.303 | 0.754 | **0.037** |
| Claude | 49.37 | 0.659 | 0.667 | 0.663 | 0.317 | 0.782 | 0.053 |
| Qwen | 61.70 | 0.666 | 0.672 | 0.669 | 0.302 | 0.740 | 0.078 |
| GPT-4o | 99.60 | 0.664 | 0.669 | 0.666 | **0.410** | 0.873 | 0.088 |
| MT-CONAN | 100.42 | **0.692** | **0.728** | **0.709** | 0.405 | **0.892** | 0.106 |
| Human | **158.45** | 0.643 | 0.628 | 0.635 | **0.384** | **0.879** | 0.061 |

Table 7: Automatic Evaluation Metrics by Source. (P: Precision, R: Recall, F1: F1 Score)

plexity, toxicity) and human evaluations (relevance, effectiveness, tone appropriateness, fluency, cultural relevance). mBART, fine-tuned on English MT-CONAN data, showed slightly better lexical similarity (BLEU: 0.0096, ROUGE-L: 0.0855, METEOR: 0.0892) but had high toxicity (0.2640 vs. 0.0163 for gold references) and low human scores (e.g., relevance: 2.35, cultural relevance: 1.95 vs. 3.97 and 3.68 for gold references). PersianMind, a monolingual Persian model, scored near-zero on BLEU and ROUGE-L, with high perplexity (316.3498 vs. 116.1709 for gold references), indicating poor fluency. Its human evaluation scores (e.g., effectiveness: 2.02 vs. 4.04 for gold references) further highlight limitations in producing relevant and culturally aligned responses. Despite reasonable BERTScore F1 scores (0.8305–0.8464), both models generated outputs that were generic and lacked the specificity and cultural nuance present in ParsCN's human-annotated responses. These results reveal significant gaps in existing models' ability to generate precise, fluent, safe, and culturally appropriate Persian counter-narratives—primarily due to the lack of Persian-specific training data. mBART's reliance on English data fails to capture Persian linguistic and cultural complexities, while PersianMind lacks fine-tuning on domain-specific counter-narrative tasks. ParsCN addresses this critical gap by providing a culturally tailored, high-quality annotated dataset to enable the development of effective counter-narrative generation models for Persian-speaking communities.

## Conclusions and Future Work

We introduce **ParsCN**, the first large-scale, carefully curated corpus for Persian counter-narrative generation. The dataset comprises 1,100 hate speech–counter-narrative pairs spanning six target groups and six countering strategies. Built through a multi-stage pipeline that combines expert annotation, strategic translation, and few-shot prompting of LLMs—each step followed by rigorous human curation—**ParsCN** offers a rich, balanced, and culturally grounded resource for both training and evaluating counter-narrative systems in a low-resource language context.

**ParsCN**'s fine-grained annotations and balanced structure make it an immediate test bed for advancing counter-narrative generation and evaluation in Persian. Beyond its intrinsic value, the dataset illustrates how a human-in-the-loop, LLM-assisted approach can efficiently generate high-quality, culturally appropriate resources at a fraction of the usual cost—offering a scalable and cost-effective template for supporting counter-narrative research in other low-resource languages.

Looking ahead, several promising directions can extend and build upon **ParsCN**. The dataset can be enriched by expanding its coverage to include additional target groups (e.g., age, disability), adopting more nuanced or culturally specific counter-narrative strategies, and incorporating fine-grained annotations—such as the severity of hate speech or the perceived effectiveness of responses. Potential future work may include designing and experimenting with automatic counter-narrative generation models, e.g., fine-tuning LLMs, exploring retrieval-augmented approaches, and examining strategy-optimized architectures. Future work should also move beyond offline dataset construction and evaluation to study how counter-narratives perform in real online environments. In particular, an important next step is to examine whether generated responses can meaningfully influence user attitudes, reduce hostility, or support de-escalation in practice. In this sense, **ParsCN** should be viewed as a foundational benchmark for Persian counter-narrative generation and a basis for future user-centered and deployment-oriented evaluation. Moreover, the dataset provides a foundation for cross-lingual research, enabling the transfer of counter-narrative capabilities to other languages and supporting the creation of robust multilingual frameworks for combating online hate.

## References


Bengoetxea, J.; Chung, Y.-L.; Guerini, M.; and Agerri, R. 2024. Basque and Spanish counter narrative generation: Data creation and evaluation. *arXiv preprint arXiv:2403.09159*.

Bennie, M.; Xiao, B.; Liu, C. X.; Zhang, D.; Meng, J.; and Tripp, A. 2025a. CODEOFCONDUCT at Multilingual Counterspeech Generation: A Context-Aware Model for Robust Counterspeech Generation in Low-Resource Languages. *arXiv preprint arXiv:2501.00713*.

Bennie, M.; Zhang, D.; Xiao, B.; Cao, J.; Liu, C. X.; Meng, J.; and Tripp, A. 2025b. PANDA–Paired Anti-hate Narratives Dataset from Asia: Using an LLM-as-a-Judge to Create the First Chinese Counterspeech Dataset. *arXiv preprint arXiv:2501.00697*.

Chung, Y.-L.; Kuzmenko, E.; Tekiroglu, S. S.; and Guerini, M. 2019. CONAN–COunter NArratives through Nichesourcing: a multilingual dataset of responses to fight online hate speech. *arXiv preprint arXiv:1910.03270*.

Chung, Y.-L.; Tekiroglu, S. S.; and Guerini, M. 2021. Towards knowledge-grounded counter narrative generation for hate speech. *arXiv preprint arXiv:2106.11783*.

Delbari, Z.; Moosavi, N. S.; and Pilehvar, M. T. 2024. Spanning the spectrum of hatred detection: a persian multi-label hate speech dataset with annotator rationales. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 17889–17897.

Fanton, M.; Bonaldi, H.; Tekiroglu, S. S.; and Guerini, M. 2021. Human-in-the-loop for data collection: a multi-target counter narrative dataset to fight online hate speech. *arXiv preprint arXiv:2107.08720*.

Fialho, P.; Ribeiro, R.; Batista, F.; Ramos, G.; Fonseca, A.; Moro, S.; Guerra, R.; Carvalho, P.; Marques, C.; and Silva, C. 2024. Counter Hate Speech Detection in Youtube Conversations. In *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, 94–105. Springer.

FORCE11. 2020. The FAIR Data principles. https://force11.org/info/the-fair-data-principles/.

Fortuna, P.; and Nunes, S. 2018. A survey on automatic detection of hate speech in text. *Acm Computing Surveys (Csur)*, 51(4): 1–30.

Furman, D.; Torres, P.; Rodríguez, J.; Letzen, D.; Martinez, M.; and Alemany, L. 2023. High-quality argumentative information in low resources approaches improve counter-narrative generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2942–2956.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.

Hassan, S.; and Alikhani, M. 2023. Discgen: A framework for discourse-informed counterspeech generation. *arXiv preprint arXiv:2311.18147*.

Hee, M. S.; Sharma, S.; Cao, R.; Nandi, P.; Nakov, P.; Chakraborty, T.; and Lee, R. 2024. Recent Advances in Online Hate Speech Moderation: Multimodality and the Role of Large Models. *Findings of the Association for Computational Linguistics: EMNLP 2024*, 4407–4419.

Hengle, A.; Kumar, A.; Singh, S.; Bandhakavi, A.; Akhtar, M. S.; and Chakroborty, T. 2024. Intent-conditioned and non-toxic counterspeech generation using multi-task instruction tuning with rlaif. *arXiv preprint arXiv:2403.10088*.

Imani, A.; Lin, P.; Kargaran, A. H.; Severini, S.; Sabet, M. J.; Kassner, N.; Ma, C.; Schmid, H.; Martins, A. F.; Yvon, F.; et al. 2023. Glot500: Scaling multilingual corpora and language models to 500 languages. *arXiv preprint arXiv:2305.12182*.

Jelinek, F.; Mercer, R. L.; Bahl, L. R.; and Baker, J. K. 1977. Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, 62(S1): S63–S63.

Jiang, S.; Tang, W.; Chen, X.; Tang, R.; Wang, H.; and Wang, W. 2025. ReZG: Retrieval-augmented zero-shot counter narrative generation for hate speech. *Neurocomputing*, 620: 129140.

Landis, J. R.; and Koch, G. G. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33 1: 159–74.

Li, J.; Galley, M.; Brockett, C.; Gao, J.; and Dolan, B. 2015. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*.

Mathew, B.; Saha, P.; Tharad, H.; Rajgaria, S.; Singhania, P.; Maity, S. K.; Goyal, P.; and Mukherjee, A. 2019. Thou shalt not hate: Countering online hate speech. In *Proceedings of the international AAAI conference on web and social media*, volume 13, 369–380.

Mathew, B.; Saha, P.; Yimam, S. M.; Biemann, C.; Goyal, P.; and Mukherjee, A. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 14867–14875.

Mun, J.; Allaway, E.; Yerukola, A.; Vianna, L.; Leslie, S.-J.; and Sap, M. 2023. Beyond denouncing hate: Strategies for countering implied biases and stereotypes in language. *arXiv preprint arXiv:2311.00161*.

Podolak, J.; Łukasik, S.; Balawender, P.; Ossowski, J.; Piotrowski, J.; Bakowicz, K.; and Sankowski, P. 2024. LLM generated responses to mitigate the impact of hate speech. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Findings of the Association for Computational Linguistics: EMNLP 2024*, 15860–15876. Miami, Florida, USA: Association for Computational Linguistics.

Saha, P.; Datta, A.; Jana, A.; and Mukherjee, A. 2024. CrowdCounter: A benchmark type-specific multi-target counterspeech dataset. *arXiv preprint arXiv:2410.01400*.

Sahoo, N. R.; Beria, G. P.; and Bhattacharyya, P. 2024. Indicconan: A multilingual dataset for combating hate speech in indian context. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 22313–22321.

Warner, W.; and Hirschberg, J. 2012. Detecting hate speech on the world wide web. In *Proceedings of the second workshop on language in social media*, 19–26.

Watanabe, H.; Bouazizi, M.; and Ohtsuki, T. 2018. Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection. *IEEE access*, 6: 13825–13835.

Wilk, B.; Shomee, H. H.; Maity, S. K.; and Medya, S. 2025. Fact-based Counter Narrative Generation to Combat Hate Speech. In *Proceedings of the ACM on Web Conference 2025*, 3354–3365.

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

## Paper Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes
   (e) Did you describe the limitations of your work? Yes, see Appendix A.10
   (f) Did you discuss any potential negative societal impacts of your work? Yes, see Appendix A.7
   (g) Did you discuss any potential misuse of your work? NA
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? NA
   (b) Have you provided justifications for all theoretical results? NA
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? NA
   (e) Did you address potential biases or limitations in your theoretical framework? NA
   (f) Have you related your theoretical results to the existing literature in social science? NA
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? NA
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? NA
   (b) Did you include complete proofs of all theoretical results? NA
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes, see section Appendix A.1
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? NA
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? NA
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? No
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? NA
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
   (a) If your work uses existing assets, did you cite the creators? Yes
   (b) Did you mention the license of the assets? Yes
   (c) Did you include any new assets in the supplemental material or as a URL? Yes
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? NA
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes, the dataset does not contain PII, however, since it includes hate speeches, offensive content is included.
   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see FORCE11 (2020))? Yes, see section titled "FAIR Compliance"
   (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Gebru et al. (2021))? Yes
6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...
   (a) Did you include the full text of instructions given to participants and screenshots? Yes

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? The study is exempted by the IRB

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? The annotators voluntarily contributed to the work. No hourly wage was paid.

(d) Did you discuss how data is stored, shared, and deidentified? Yes

# Appendix A

## A.1 Data and Code Availability

Both the ParsCN dataset and the full codebase used for automatic evaluation (BERTScore, BLEU, ROUGE, METEOR, Distinct-n, Perplexity, Toxicity) as well as for baseline model evaluation (mBART and PersianMind) are available on Github at: https://github.com/zahrasafdari/ParsCN

The repository includes:

- Evaluation scripts for all automatic metrics described in the paper.
- Baseline model training and inference code.
- Pre-processing routines for dataset tokenization and formatting.
- The proposed ParsCN dataset that contains hate speech-counter-narrative pairs along with target groups and counter-narrative types.

**License:** The materials are publicly released under a CC BY 4.0 license.

## A.2 Detailed Counter-Narrative Types' Definition

The counter-narrative types used in the dataset are described below. These definitions are derived from Sahoo, Beria, and Bhattacharyya (2024):

- *Positive Response:* These counter-narratives counter hostility with inclusive, supportive statements that encourage empathy and unity. For example, in response to gender-based hate, they might advocate mutual respect in relationships to promote harmony. This approach seeks to reframe divisive rhetoric into constructive dialogue.
- *Counter questions:* This type uses probing questions to challenge the biases or assumptions in hate speech, urging reflection. For instance, questioning claims about Sunni minorities' ambitions can prompt reconsideration of stereotypes, fostering critical thinking.
- *Denouncing:* These statements firmly reject harmful rhetoric, condemning its ethical or social damage. For example, denouncing slurs against religious groups as divisive aims to curb their spread by highlighting their harm.
- *Fact-based:* This approach refutes misinformation with credible evidence, enhancing discourse accuracy. For instance, citing statistics to debunk myths about women's driving in Iran corrects false narratives and builds trust in factual dialogue.
- *Warning of consequences:* These counter-narratives caution against the adverse effects of hate speech, such as social division from targeting Afghan immigrants. By emphasizing risks, they encourage more responsible perspectives.
- *Contradiction:* This type exposes inconsistencies in hate speech, such as criticizing women's autonomy while ignoring men's. By highlighting logical flaws, it undermines the credibility of harmful claims.

## A.3 Comparison of Low-Resource Counter-Narrative Datasets

This section provides a detailed comparison of ParsCN with other low-resource counter-narrative (CN) datasets to highlight its novel contributions in scope and methodology. Table 8 summarizes key differences in language, data sourcing, annotation approach, and scalability.

| Dataset | Lang. | Sourcing | Annotation | Scal. |
|---|---|---|---|---|
| CONAN | Eng., Fr., It. | Nichesourcing | Expert-based | Limited |
| IndicCONAN | Indian languages. | Crowdsourcing | Human-only | Moderate |
| PANDA | Chinese | LLM-as-a-Judge | Automated | High |
| CONAN-EUS | Basque, Spanish | Translation | Post-edited | Moderate |
| **ParsCN** | **Persian** | **Hybrid (PHATE, translated)** | **Human+LLM, culturally tailored** | **High** |

Table 8: Comparison of Low-Resource CN Datasets

## A.4 Baseline Model Evaluation Details

This section provides the detailed experiment setup and results for evaluating baseline models (mBART and PersianMind) on generating counter-narratives for 125 Persian hate speech samples from the ParsCN dataset.

### Experiment Setup

We assessed two models on generating counter-narrative responses for 125 diverse hate speech samples from the ParsCN dataset:

- **mBART**: A multilingual sequence-to-sequence model fine-tuned on the English MT-CONAN dataset. Persian hate speech samples were translated into English by human translators with post-editing, and mBART generated English counter-narratives, compared against gold-standard English-translated Persian counter-narratives.
- **PersianMind** (universitytehran/PersianMind-v1.0-8-bit quantized): A Persian language model prompted directly in Persian, with outputs compared to the same gold-standard Persian counter-narratives.

**Evaluation Metrics:**

- **Automated Metrics:**
  - **Standard Metrics**: BLEU, ROUGE-L, METEOR (measuring lexical and structural similarity), and BERTScore F1 (measuring semantic similarity).

  - **Quality and Safety Metrics**: Perplexity (fluency, lower is better; measured using a GPT-2 Persian model for PersianMind and a standard language model for mBART) and toxicity (lower is safer, measured with `texttox/glot500-toxicity-classifier` (Imani et al. 2023) for PersianMind).
- **Human Evaluation Metrics** (for mBART and PersianMind, scored on a 1–5 scale, higher is better):
  - **Relevance**: How well the response addresses the hate speech content.
  - **Effectiveness**: The response's ability to mitigate hate speech or promote constructive dialogue.
  - **Tone Appropriateness**: Suitability of the tone for counter-narrative responses (e.g., empathetic, non-confrontational).
  - **Fluency**: Linguistic coherence and naturalness of the response.
  - **Cultural Relevance**: Alignment with Persian cultural norms and context.

## Results

The performance of the evaluated models—mBART and PersianMind—is summarized in Tables 9, 10, and 11.

Lexical and Structural Similarity: As shown in Table 9, both models exhibit extremely low BLEU, ROUGE-L, and METEOR scores, highlighting minimal overlap with the gold-standard references in terms of lexical choices or syntactic structure. mBART performs slightly better (BLEU: 0.0096, ROUGE-L: 0.0855, METEOR: 0.0892), but these values remain insufficient for practical use. PersianMind fails to achieve any non-zero score on BLEU and ROUGE-L, with a low METEOR score (0.0626), indicating limited synonym or stem-level alignment.

Semantic Similarity: Despite weak lexical resemblance, both models achieve relatively high BERTScore F1 values (mBART: 0.8464, PersianMind: 0.8305), suggesting some semantic proximity to the gold references. However, this similarity is primarily due to the models generating generic counter-narrative with broadly positive sentiments (e.g., promoting empathy and peace). These outputs often lack the contextual specificity and cultural grounding necessary for persuasive, situation-sensitive counter-narrative.

Fluency and Safety (Automated Metrics): As shown in Table 10, mBART demonstrates the lowest perplexity (65.73), indicating high fluency, though this may come at the cost of generating templated or repetitive text. PersianMind, while more fluent than other unreported baselines, still exhibits a high perplexity score (316.35), far above the gold Persian references (116.17), suggesting limited coherence and linguistic naturalness. On the safety front, mBART's outputs are notably more toxic (0.2640) than human-authored gold responses (0.0163 in English, 0.0598 in Persian), raising red flags for deployment in sensitive domains. PersianMind performs better (toxicity: 0.1157), but still exceeds safe thresholds compared to human benchmarks, signaling a risk of inappropriate or counterproductive content generation.

Human Evaluation: Table 11 presents a detailed comparison of human ratings across five dimensions.

Relevance and Effectiveness: Both models received low scores, with mBART scoring 2.35 for relevance and 1.86 for effectiveness, and PersianMind at 2.24 and 2.02, respectively. These results reflect the models' inability to directly and constructively address hate speech content.

Tone Appropriateness: PersianMind outperformed mBART (3.41 vs. 2.49), showing a relatively more empathetic and non-confrontational tone, yet still lagging behind the gold standard (4.34).

Fluency: Both models demonstrated moderate fluency (mBART: 3.30, PersianMind: 3.58), aligning with automated perplexity trends. However, both remained below human-authored responses (4.61).

Cultural Relevance: Scores for cultural relevance were low across the board (mBART: 1.95, PersianMind: 2.08), underscoring a disconnect between generated outputs and Persian socio-cultural norms. This gap limits the models' utility in real-world applications, where cultural sensitivity is paramount.

Overall Assessment: These results collectively highlight serious limitations in current baseline systems for Persian counter-narrative generation. The outputs are often semantically generic, lexically distant from human responses, and culturally misaligned. Moreover, high toxicity levels and poor human evaluation scores raise concerns about both safety and efficacy. These deficiencies primarily stem from the lack of Persian-specific training data. While mBART relies on English training data and cross-lingual transfer, it fails to capture Persian nuance even with high-quality translations. PersianMind, despite being monolingual, lacks fine-tuning on domain-specific counter-narrative tasks, resulting in bland or misdirected outputs.

The findings emphasize the need for culturally anchored, Persian-language resources like ParsCN to train models that can generate precise, fluent, safe, and context-aware counter-narratives for online hate mitigation.

| Metric | mBART | PersianMind |
|---|---|---|
| BLEU | 0.0096 | 0.0000 |
| ROUGE-L | 0.0855 | 0.0000 |
| METEOR | 0.0892 | 0.0626 |
| BERTScore F1 | 0.8464 | 0.8305 |

Table 9: Standard Evaluation Metrics for Model Generations vs. Gold References

| Metric | mBART | PersianMind | Gold References |
|---|---|---|---|
| Perplexity | **65.73** | 316.35 | **102.23 (En), 116.17 (Fa)** |
| Toxicity | 0.2640 | 0.1157 | **0.0163 (En), 0.0598 (Fa)** |

Table 10: Quality and Safety Metrics for Model Generations vs. Gold References

| Metric | mBART | PersianMind | Gold References |
|---|---|---|---|
| Relevance | 2.35 | 2.24 | **3.97** |
| Effectiveness | 1.86 | 2.02 | **4.04** |
| Tone Appropriateness | 1.49 | 3.41 | **4.34** |
| Fluency | 3.30 | 3.58 | **4.61** |
| Cultural Relevance | 1.95 | 2.08 | **3.68** |

Table 11: Human Evaluation Metrics for Model Generations vs. Gold References (1–5 Scale)

## A.5 Examples of Culturally Relevant Samples

This section provides examples of hate speech and counter-narrative pairs from MultiTarget-CONAN and HateXplain, selected and translated to align with Persian socio-cultural contexts, as described in the Data Sourcing section. These examples, presented in both Persian and English, demonstrate the cultural relevance of the selected samples, addressing concerns about their fit within Persian online discourse. Table 4 presents three examples from each dataset, with hate speech and corresponding counter-narratives in Persian and their English translations.

**Relevance to Persian Context** The examples in Table 4 were selected and post-edited to align with socio-cultural issues prevalent in Persian online discourse, ensuring their relevance to Persian-speaking communities. Below, we explain how each example reflects specific cultural and social dynamics in Persian contexts:

- **MultiTarget-CONAN Examples (Gender)**: The three examples targeting women address gender-based discrimination, a prominent issue in Persian online spaces. The hate speech samples reflect common misogynistic narratives in Persian discourse, such as dismissing women's political participation or confining them to domestic roles, which resonate with ongoing debates about gender equality in Iran and other Persian-speaking regions. For instance, the claim that "women have no place in politics" mirrors sentiments often expressed in Persian social media, where traditional gender roles are debated. The counter-narratives, emphasizing equality and women's capabilities, align with progressive movements advocating for women's rights in public and political spheres, making them highly relevant to Persian audiences.
- **HateXplain Examples (National)**: The three examples targeting Arabs and Chinese individuals reflect ethnic and geopolitical tensions prevalent in Persian online discourse. Anti-Arab sentiment, as seen in references to "fewer Arabs" or "filthy Arab country," is rooted in historical and political rivalries in the Middle East, often amplified in Persian social media due to regional conflicts (e.g., Iran's relations with Arab states). Similarly, the derogatory "ching chong" remark targeting Chinese individuals reflects stereotypes that emerge in Persian online spaces, influenced by global media and geopolitical perceptions of China. The counter-narratives, which promote humanity, peace, and character-based judgment, resonate with Persian cultural values of hospitality and respect, countering hate with calls for unity that are contextually appropriate for Persian-speaking communities.

These examples were carefully selected to address issues like gender discrimination and ethnic prejudice, which are prevalent in Persian online discourse. Native Persian annotators post-edited the translated samples to ensure linguistic accuracy and cultural congruence, as evidenced by high human evaluation scores for relevance (4.025–4.231) and tone appropriateness (4.507–4.787) (Table 6). This process ensures that the ParsCN dataset captures the "Persian soul" by reflecting authentic socio-cultural dynamics.

## A.6 Annotator Details

All annotators are native Persian speakers with graduate or doctoral backgrounds in Computational Linguistics or Social Computing, each with prior experience in hate speech or counter-narrative research. Their academic and linguistic backgrounds ensured a deep familiarity with the socio-cultural dynamics relevant to Persian online discourse. Annotators were based in Persian-speaking regions (including major urban centers like Tehran and Qazvin), with a balanced gender distribution and an age range between early 20s and early 30s. Before annotation, they completed a two-hour calibration workshop and a qualification quiz (20 samples) to ensure conceptual alignment. Inter-annotator agreement reached Cohen's $\kappa = 0.71$ (substantial), and semantic consistency on shared samples achieved a BERTScore $F1 = 0.82$, confirming both reliability and coherence. Ethical approval was obtained under institutional research guidelines, with full anonymization and voluntary participation.

## A.7 Ethical Considerations

The development of resources to combat hate speech must carefully address ethical considerations. Given that the dataset contains sensitive and potentially offensive content, strict protocols for storage, access, and distribution are essential to prevent misuse and ensure researcher safety. Bias is a significant concern, stemming from the source data, annotation decisions, and LLMs utilized; this would result in datasets or models that unintentionally under-represent specific types of hate speech or generate biased or ineffective counter-narratives to specific target groups. Finding the right balance between effective counter-narrative and avoiding censorship is a particularly sensitive challenge; it's vital that systems trained on resources like ParsCN are designed to promote healthy dialogue without suppressing legitimate, critical expression. Finally, the actual effectiveness and real impact of automatically generated counter-narratives in countering online hate speech is a complex problem that calls for further investigation. The question needs to go beyond simply generating syntactically correct answers to analyzing their social and practical implications in the specific Persian online culture, and also considering the responsibility for the generated outputs.

## A.8 Qualitative Error Analysis

A qualitative inspection of the automatically generated counter-narratives revealed several recurring failure modes

| Source | Cultural Naturalness | | | Semantic Clarity | | | Fluency | | | Tone Appropriateness | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | A1 | A2 | Avg | A1 | A2 | Avg | A1 | A2 | Avg | A1 | A2 | Avg |
| PHATE | 4.250 | 4.033 | 4.142 | 4.100 | 3.867 | 3.983 | 4.417 | 4.183 | 4.300 | 4.517 | 4.333 | 4.425 |
| Translated | 3.767 | 3.733 | 3.750 | 4.150 | 3.983 | 4.067 | 4.350 | 4.167 | 4.258 | 4.250 | 4.100 | 4.175 |

Table 12: Pair-level human evaluation comparing native Persian PHATE pairs and translated-and-adapted MT-CONAN pairs. Scores are on a 1–5 scale.

prior to manual curation. First, some responses were overly generic and failed to engage with the specific hateful claim, instead producing broad appeals to kindness or peace. Second, some fact-based responses introduced unsupported or weakly grounded claims, which risked sounding hallucinatory or unpersuasive. Third, a subset of outputs exhibited cultural misalignment, including phrasing that was grammatically acceptable in Persian but pragmatically unnatural or insufficiently adapted to Persian socio-political discourse. Fourth, some responses used an overly confrontational or sarcastic tone, which conflicted with our goal of promoting constructive and non-escalatory counter-narrative. Manual curation addressed these issues by revising factual content, improving specificity, adapting culturally incongruent wording, and softening tone where necessary. This analysis further supports the importance of human oversight in low-resource, culturally sensitive counter-narrative generation.

## A.9 Comparison of Native Persian and Translated-Source Pairs

To further examine whether translated source material introduced noticeable degradation in the final Persian hate speech–counter-narrative pairs, we conducted a focused comparative human evaluation between native Persian pairs derived from PHATE and translated-and-adapted pairs derived from MT-CONAN. Since MT-CONAN was incorporated only for the overlapping Gender and Racial target groups in Stage One, we restrict this analysis to those categories to ensure a controlled comparison between native and translated-source content.

We randomly sampled a balanced set of pairs from the two sources and asked two native Persian annotators to independently evaluate each hate speech–counter-narrative pair as a whole. Rather than evaluating the hate speech and counter-narrative separately, annotators scored the pair-level quality using four criteria: *Cultural Naturalness*, *Semantic Clarity*, *Fluency*, and *Tone Appropriateness*. Cultural Naturalness measures whether the pair feels plausible and natural in contemporary Persian discourse; Semantic Clarity assesses whether the meaning of the hate speech, the response, and their relationship are clear and coherent; Fluency evaluates the overall linguistic well-formedness of the pair; and Tone Appropriateness captures whether the counter-narrative responds in a respectful and contextually suitable manner. All scores were assigned on a 1–5 scale.

Table 12 reports the average scores for each annotator and the combined mean. The PHATE pairs achieve slightly higher scores in Cultural Naturalness (4.14 vs. 3.75) and Tone Appropriateness (4.43 vs. 4.18), while translated-source pairs remain highly competitive in Fluency (4.26 vs. 4.30) and slightly exceed PHATE in Semantic Clarity (4.07 vs. 3.98). Overall, the translated-and-adapted pairs receive strong scores across all four dimensions, indicating that the translation and manual post-editing pipeline preserved high quality in the final Persian dataset.

These results support our design choice of using translation followed by careful human adaptation for low-resource dataset construction. Although native Persian pairs remain somewhat more culturally natural on average, the gap is modest, and the translated-source pairs remain fluent, semantically clear, and tone-appropriate. This suggests that our localization process substantially mitigated translation artifacts and enabled translated examples to serve as reliable additions to the dataset, especially in categories where native Persian coverage was limited. Importantly, the translated-source pairs do not exhibit a broad degradation in quality: their scores remain above 4.0 in Semantic Clarity, Fluency, and Tone Appropriateness, and only Cultural Naturalness shows a noticeable but limited decrease relative to native Persian pairs. We therefore view translation followed by culturally informed human adaptation as a practical and effective strategy for expanding Persian counter-narrative resources when native data is scarce.

## A.10 Limitations

While introducing the first Persian counter-narrative dataset, this study faces some limitations. The 1100 pair dataset size, while substantial for a low-resource language baseline, is smaller than for high-resource languages and may possibly limit model generalizability. Relying on translated and adapted hate speech examples from English datasets, despite manual post-editing, cannot fully capture the entire scope and varieties of native Persian online hate speech and counter-narratives. Furthermore, the utilization of LLMs for generation in Stage Two, for all its scalability, means that generated counter-narratives can reflect biases present in the LLM training data or miss out on the spontaneity and contextual richness of actual human responses, even following manual curation. In addition, our evaluation is limited to offline human and automatic assessment; we do not study how generated counter-narratives perform in live online environments or whether they influence user beliefs, engagement, or de-escalation outcomes.

## A.11 AI assistance in Writing

AI Assistants were used solely to assist in improving the writing clarity and language of this paper. Specifically, AI-

assisted refinements were applied to enhance readability, coherence, and grammatical accuracy. No AI-generated content was used to replace critical thinking or fabricate results. Ideas, methodology, experimental design, analysis, and conclusions were entirely conceived, developed, and executed by the authors.