NYCU-NLP at EXIST 2025: An Empirical Study of Annotator-Aware Two-Stage Pipeline for Sexism Detection in Tweets

Notebook for the EXIST Lab at CLEF 2025

Joy Chrissetyo Prajogo[1,*], Lung-Hao Lee[2,*] and Hsien-I Lin[3]

[1]Department of Electrical Engineering and Computer Science, National Yang Ming Chiao Tung University, Taiwan
[2]Institute of Artificial Intelligence Innovation, National Yang Ming Chiao Tung University, Taiwan
[3]Institute of Electrical and Control Engineering, National Yang Ming Chiao Tung University, Taiwan

Abstract

This paper presents a comprehensive approach for automated sexism detection in bilingual tweets, evaluated within the context of the EXIST 2025 shared task. Our proposed framework explicitly integrates annotator demographics and leverages bilingual fusion, combining original and cross-translated tweets. We implement a novel two-stage hierarchical pipeline consisting of sexism identification (binary classification) followed by misogynistic intent classification (multi-class) and sexism type categorization (multi-label). Our experiments systematically compare three distinct modeling strategies within this pipeline: a fine-tuned transformer-based dual-encoder architecture with early and late fusion, a zero-shot auto-regressive (AR) large language model (LLM), and a zero-shot diffusion-based LLM. Evaluation results demonstrate that our transformer-based approach consistently achieves the highest performance across most metrics, emphasizing the effectiveness of explicitly modeling annotator disagreement and demographic context. Notably, the diffusion-based LLM demonstrates performance on par with, and in some cases superior to, the AR LLM, highlighting the potential of diffusion models as an alternative paradigm for complex text classification tasks.

Keywords

Transformers, Auto-Regressive LLM, Diffusion LLM, Sexism Identification

1. Introduction

Social media platforms, such as Facebook, X (formerly Twitter), LinkedIn, and TikTok, have dramatically changed how people connect, communicate, and share information. However, interactions on these platforms can be constructive, neutral or destructive. One prevalent form of destructive interaction is sexism, defined as prejudice or discrimination based on sex or gender, primarily targeted at women. On platforms like X, sexism often manifests in harmful messages or tweets, creating negative emotional and psychological impacts for other users. The widespread and damaging nature of online sexism highlights an urgent need for automatic identification and mitigation.

The sEXism Identification in Social neTworks (EXIST) shared task, initiated at CLEF in 2021, addresses this issue by promoting research into automated sexism detection on social media. EXIST has evolved significantly across its annual editions:

- EXIST 2021 initiated tasks with binary sexism identification and categorization [1].
- EXIST 2022 further emphasized the automation and scalability of sexism detection [2].
- EXIST 2023 introduced the detection of sexist intent (direct, reported, judgmental) and adopted the Learning with Disagreements (LeWiDi) paradigm, recognizing that different annotators might disagree on labels due to inconsistent subjective perceptions of sexism [3, 4].

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain
*Corresponding author.
jychpr.ee12@nycu.edu.tw (J. C. Prajogo); lhlee@nycu.edu.tw (L. Lee); sofin@nycu.edu.tw (H. Lin)
https://lunghao.weebly.com/ (L. Lee); https://www.arlabtw.com/ (H. Lin)

- EXIST 2024 expanded from text-based tweets to visual memes, still using the LeWiDi paradigm [5, 6].
- EXIST 2025 further broadens the task scope to multimedia data, including TikTok videos, thus covering text, images, and videos under the LeWiDi framework [7, 8].

In this edition, our team specifically addresses sexism detection in tweets, tackling the three subtasks designed for textual data:

- Task 1: Binary sexism detection (sexist vs. non-sexist tweets).
- Task 2: Misogynistic intent classification (direct, reported, judgmental).
- Task 3: Fine-grained sexism categorization, including ideological inequality, stereotyping and dominance, objectification, sexual violence, and misogyny (non-sexual violence).

Due to the continued use of Learning with Disagreements (LeWiDi) in EXIST 2025, each tweet is annotated by six annotators whose judgments may vary. Capturing and modeling this disagreement is crucial for developing robust detection systems.

To address these challenges, we propose a novel annotator-aware, two-stage bilingual sexism detection pipeline. Our approach integrates annotator demographics directly into the modeling process and exploits bilingual fusion through original tweets and their translations. We test this pipeline using three modeling strategies: a transformer-based dual-encoder architecture, an auto-regressive (AR) large language model (LLM), and a diffusion-based LLM. Empirical results indicate that diffusion-based LLM achieves performance comparable to that of the AR LLM, highlighting promising directions in leveraging diffusion models for text classification tasks.

The remainder of this paper is structured as follows: Section 2 reviews related work, Section 3 presents the methodologies and system architecture, Section 4 describes our experimental setup, including the dataset, parameters, and results, and Section 5 concludes with our findings and future directions.

2. Related Work

Sexism involves unjustified distinctions or discriminatory actions based on gender, primarily targeted against women, varying across cultural contexts. Various studies highlight the pervasive nature and diverse manifestations of online sexism. Early prominent sexism detection datasets include AMI [9] and HatEval [10], initially established for English-language detection. EXIST, launched at CLEF in 2021, uniquely combines binary detection, intent classification, and fine-grained categorization tasks [1].

Learning with Disagreement (LeWiDi) addresses scenarios where multiple annotators provide conflicting labels or interpretations due to subjective or ambiguous data. Instead of forcing consensus into a single "correct" label, LeWiDi explicitly incorporates annotator disagreement into model training and evaluation as informative signals. This approach enhances robustness and generalization by systematically leveraging diverse annotator perspectives, particularly useful in sentiment analysis, hate speech detection, and other subjective classification tasks [11].

EXIST adopted the LeWiDi paradigm starting from EXIST-2023 edition [3, 4], recognizing that annotator disagreement is inherent due to differing subjective perceptions of sexism [12]. Soft labeling, which quantifies annotator disagreements, has been shown to improve robustness compared to traditional single-label training [13, 11]. Prior solutions [14, 15] explicitly incorporated soft labeling, significantly enhancing system performance by modeling the inherent annotator disagreements.

Annotators' demographic backgrounds strongly influence their perception of what constitutes sexism, leading to inconsistent results. Tredici et al. [16] demonstrated that incorporating annotator demographics via demographic embeddings can improve the accuracy of hate speech detection. Multi-annotator attention mechanisms have also been explored, notably using personalized toxicity classifiers [17]. Similarly, EXIST-2024 participants, such as Fang et al. [14] and Quan and Thin [15] effectively used annotator demographics and metadata, showcasing that explicitly modeling annotator features significantly enhances sexism detection performance.

Transformer-based architectures, particularly multilingual models like XLM-R [18] have proven highly effective for multilingual hate speech detection [19]. Data augmentation and multilingual fusion strategies, including cross-translation and back-translation, have been shown to enhance generalization in low-resource scenarios[20, 21]. Our pipeline uniquely uses bilingual fusion by simultaneously modeling original and cross-translated tweets, thereby enhancing robustness against linguistic variability and annotator disagreement.

Recently, zero-shot and few-shot large language models (LLMs) have shown impressive performance in toxicity detection tasks without task-specific fine-tuning. GPT-3 demonstrated that prompt-based LLMs can effectively identify toxic content through careful prompt engineering [22]. The LLaMA family[23], particularly the LLaMA 3.1 series, have been optimized for multilingual dialogue use cases and outperform many existing open-source and closed-source chat models on common industry benchmarks. Two teams [24, 10] conducted comprehensive evaluations demonstrating LLM competitiveness against fine-tuned models for detecting hate and toxic speech.

Diffusion-based language models represent a novel and promising approach to language generation tasks. Introduced by [25, 26], diffusion models generate text through iterative denoising steps, yielding more controlled and coherent outputs compared to autoregressive models. The recent Dream-v0 Instruct model, developed by the HKU NLP Group [27], demonstrated encouraging results for instruction-following tasks, showcasing the potential of diffusion models in text classification . Similarly, LLaDA, a diffusion model trained from scratch under the pre-training and supervised fine-tuning paradigm[28], has shown strong scalability and performance, rivaling autoregressive models like LLaMA 3 in various benchmarks . These developments highlight the viability of diffusion models as an alternative to traditional autoregressive approaches in language modeling.

Our proposed bilingual fusion approach, combining original tweets with their translated counterparts, is inspired by robust back-translation methods. Multi-view ensemble strategies further enhance robustness and generalization, as demonstrated by previous EXIST participants. Additionally, our two-stage pipeline (binary gate followed by multi-label classification) effectively reduces the complexity of sexism categorization, allowing models to specialize sequentially and improve overall detection accuracy.

In summary, while EXIST has significantly advanced the state-of-the-art in sexism detection through the introduction of multimedia tasks and the LeWiDi paradigm, no prior EXIST entry has systematically compared supervised dual-encoder transformers, zero-shot auto-regressive LLMs, or diffusion-based LLMs within a unified annotator-aware bilingual pipeline. Our study fills this important research gap.

3. Methodologies and Architecture

In this section, we detail our proposed framework for bilingual sexism detection, comprising multiple interconnected stages that form a comprehensive pipeline. The pipeline starts with an extensive data pre-processing stage that transforms raw tweet data into structured, annotated, bilingual data suitable for modeling. We then introduce our annotator-aware modeling pipeline, implemented as a two-stage hierarchical system. Within this pipeline, we explore and compare three distinct modeling strategies: a transformer-based dual-encoder, an auto-regressive large language model (LLM), and a diffusion-based LLM. The pipeline concludes with a data post-processing step that formats the final predictions appropriately for evaluation.

Our approach explicitly leverages annotator-provided metadata and directly addresses annotator disagreements. Furthermore, it exploits bilingual data by integrating original and cross-translated tweets to enhance the prediction accuracy.

3.1. Data Pre-Processing

Figure 1 illustrates our data pre-processing pipeline, which converts raw JSON annotation files into a structured, machine-readable CSV format. Below we describe each step in detail.

First, raw JSON annotation files are transformed into a structured CSV table, facilitating easier manipulation and subsequent processing. Following this, we generate bilingual tweet pairs by using

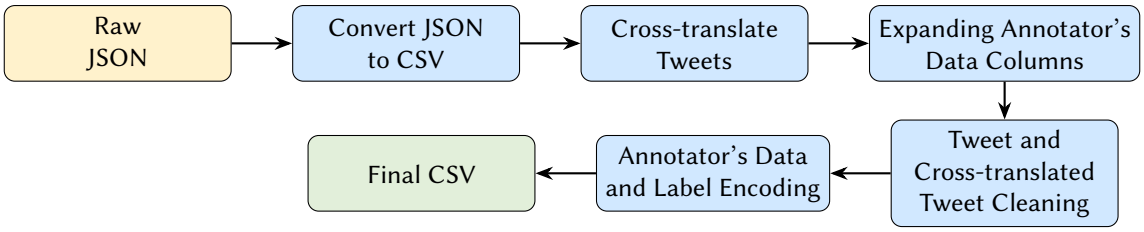

Figure 1: Data Preprocessing Flow

Table 1
Tweet and Cross-Translated Tweet Results Sample

ID	Original Tweet	New Cross-Translated Tweet
100008	[ES] @BestKabest Esta gringa sigue llorando por el gamergate, que "coincidencia" que tenga pronombres en su perfil	[EN] @Bestkabest This gringa continues to cry for the gamergate, that "coincidence" that has pronouns in its profile
202070	[EN] @BarbieReports @Londonist @TimeOutLondon @visitlondon the curse of phallogocentrism ! god forbid a woman should be upright :)	[ES] @Barbiereports @londonist @timeoutlondon @visitlondon ¡La maldición del fallogocentrismo! Dios no permita que una mujer debería estar en posición vertical :)

Table 2
Tweet and Cross-translated Tweet Cleaning Result Samples

ID	Original Text	Cleaned Text
101170	[ES] No queremos hijos valientes. Queremos hijos felices. Protejan a los menores con diversidad funcional para que dejen de ser víctimas del acoso escolar. #detidepende #laEspañaqueQueremos #derechoshumanos@sanchezcastejon @IreneMontero https://t.co/cZtiZfX5O6	[ES] no queremos hijos valientes. queremos hijos felices. protejan a los menores con diversidad funcional para que dejen de ser víctimas del acoso escolar.
	[EN] We do not want brave children. We want happy children. They protect minors with functional diversity to cease to be victims of bullying. #Detidepende #laspaña we will see #Haries Humanos @SanchezCastejon @irenemontero https://t.co/cztizfx5O6	[EN] we do not want brave children. we want happy children. they protect minors with functional diversity to cease to be victims of bullying. we will see humanos

Google Translate via Python's `deep-translator` library, translating Spanish tweets into English and vice versa. Table 1 shows samples of the original and cross-translated tweets.

Annotator metadata, originally aggregated in list form, is then expanded into separate, structured columns, allowing for efficient integration into downstream modeling. Next, rigorous data cleaning is applied to both the original and cross-translated tweets. This process involves removing usernames, URLs, emails, percentages, timestamps, phone numbers, hashtags, emojis, and other symbols. Additionally, tweets are converted to lowercase, and multiple spaces between words are removed. Table 2 illustrates tweet examples before and after cleaning.

Finally, annotator metadata and labels are encoded. Metadata and labels are transformed into numerical representations through one-hot encoding, ordinal encoding, and symbolic letter combinations, thereby reducing complexity and optimizing resource utilization during modeling. This culminates in a

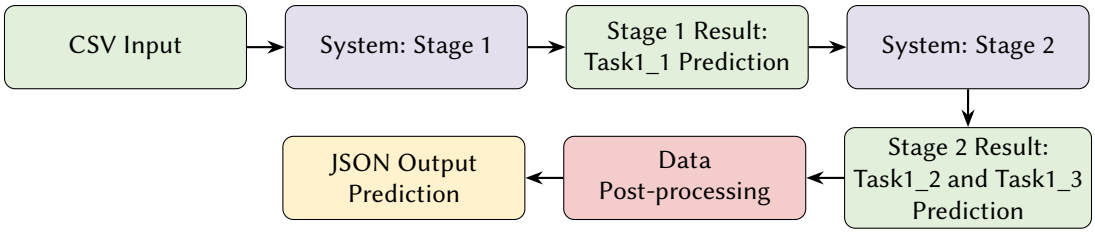

Figure 2: Annotator-Aware Two-Stage Pipeline System with Data Postprocessing

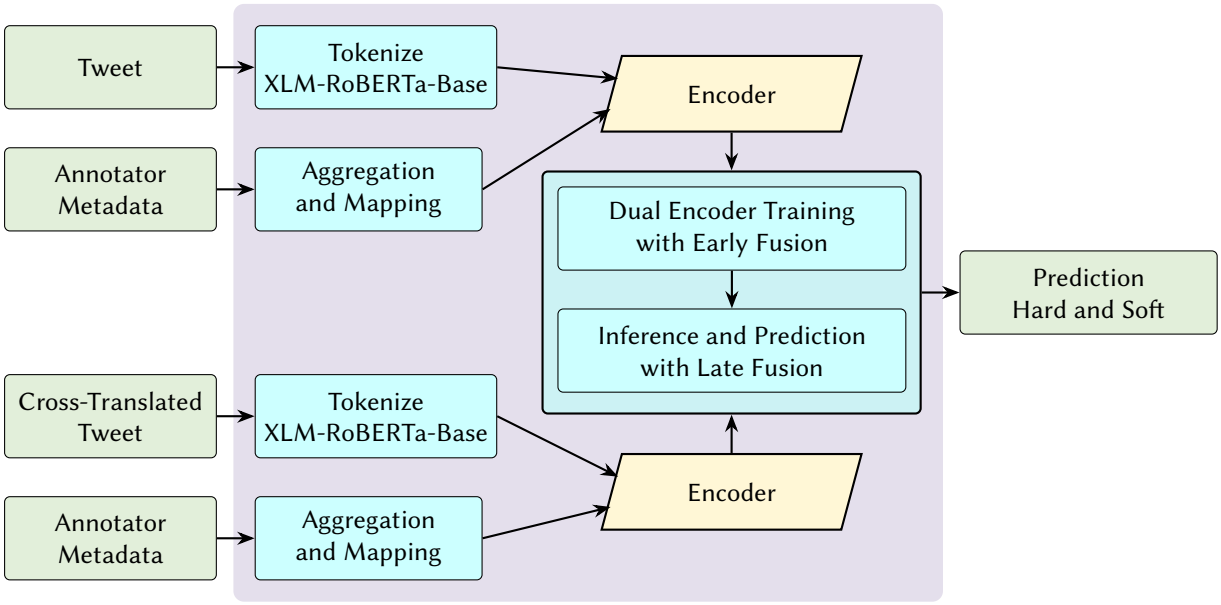

Figure 3: Simplified High-Level Diagram of Transformer-Based Approach

fully structured CSV dataset ready for use by the modeling pipeline.

3.2. Annotator-Aware Two-Stage Pipeline

Figure 2 illustrates our annotator-aware two-stage modeling pipeline, explicitly designed to address annotator disagreement. The pipeline operates hierarchically with two clearly defined stages:

Stage 1 addresses Task 1_1: Sexism Identification in Tweets, a binary classification task determining whether a tweet contains sexist content (YES or NO).

Stage 2 encompasses Task 1_2: Source Intention Classification and Task 1_3: Sexism Categorization. Task 1_2 is a multi-class classification of the author's intention behind sexist tweets, categorizing them as DIRECT, REPORTED, or JUDGEMENTAL. Task 1_3 performs a multi-label classification to determine the specific type of sexism, assigning labels such as IDEOLOGICAL-INEQUALITY, STEREOTYPING-DOMINANCE, OBJECTIFICATION, SEXUAL-VIOLENCE, and MISOGYNY-NON-SEXUAL-VIOLENCE.

To comprehensively evaluate framework performance, we implement three distinct modeling approaches: (1) a fine-tuned transformer-based dual-encoder using both early and late fusion, (2) a zero-shot auto-regressive large language model (LLM) with late fusion, and (3) a zero-shot diffusion-based LLM with late fusion. Each method explicitly incorporates bilingual information and annotator metadata, using an identical pipeline architecture (as depicted in Fig. 2) but differing in terms of modeling paradigms.

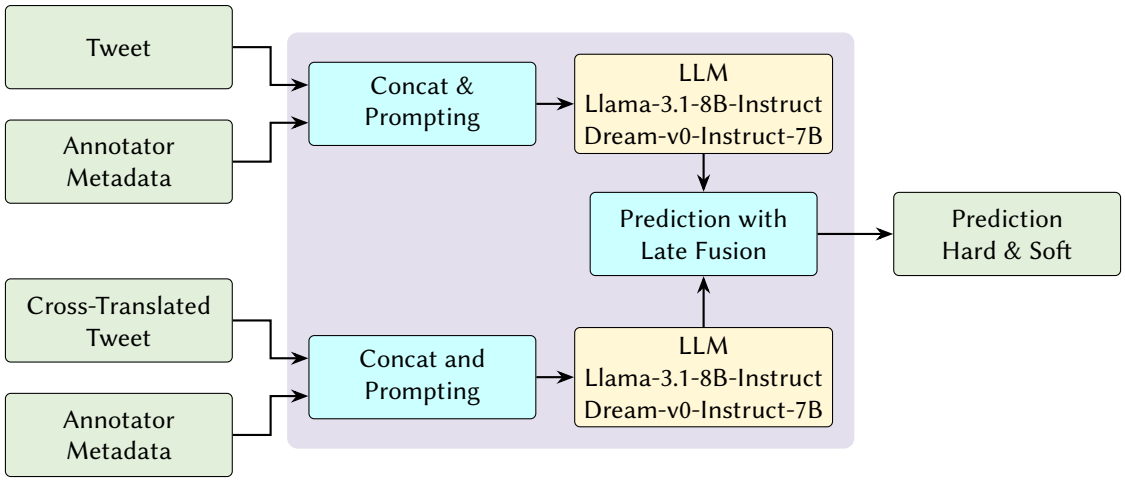

Figure 4: Simplified High-Level Diagram of LLM-Based System

3.2.1. First System: Transformer-Based Approach

Our first system, depicted in Fig. 3, uses a transformer-based dual-encoder architecture with XLM-RoBERTa-Base (XLM-R), explicitly integrating bilingual tweet data and annotator metadata. The detailed steps are as follows:

Tweets (original and cross-translated) undergo independent tokenization using the XLM-R tokenizer, producing token IDs and attention masks. Simultaneously, annotator metadata is mapped into numerical indices and then aggregated, forming structured tensors. The resulting tokenized tweets and metadata tensors are combined within a dual encoder that comprises the XLM-R backbone and an annotator-aware sub-encoder using Multi-Head Attention (MHA).

During training (Stages 1 and 2), embeddings from the original and cross-translated tweets are fused early using a weighted sum (alpha = 0.8). Subsequently, during inference, predictions from each encoder branch (original and cross-translated) are fused using the same weighted averaging (alpha = 0.8). This produces both hard (binary or categorical) and soft (probability) predictions. Importantly, the same transformer-based framework is reused for both stages, differing only in terms of the classification heads and labels specific to each task.

3.2.2. Second System: Auto-regressive LLM-Based Approach

The second system, shown in Fig. 4, adopts a zero-shot auto-regressive LLM approach using the LLaMA-3.1-8B-Instruct model without additional fine-tuning. The approach differs notably from the transformer-based method by relying on carefully engineered textual prompts instead of encoder embeddings.

In this system, each tweet and its associated annotator metadata are concatenated into structured textual prompts, provided separately for the original and cross-translated tweets. The auto-regressive LLM directly computes log probabilities for targeted tokens such as "Yes," "No," or category codes ("1," "2," "3"), depending on the specific task and classification type (binary or multi-class). Original and translated log probabilities are subsequently fused using late fusion (alpha = 0.55), enhancing prediction robustness without task-specific training.

Similar to the first system, an identical process is repeated for Stages 1 and Stage 2, differing only in terms of prompt content corresponding to task objectives.

3.2.3. Third System: Diffusion LLM-Based Approach

Our third modeling approach introduces a diffusion-based LLM, specifically the Dream-v0-Instruct-7B model. The pipeline architecture for this system (illustrated in Fig. 4) matches the second system exactly,

differing only in the underlying model. This approach demonstrates a novel text classification paradigm that leverages diffusion models' generative capabilities for classification tasks.

The diffusion-based model uses the same prompt-engineering strategy, combining bilingual tweet pairs and annotator metadata into structured prompts. It similarly extracts log probabilities for target tokens ("Yes," "No," or category labels), which are fused using late fusion (alpha = 0.65). Comparing the diffusion LLM against traditional transformer and auto-regressive LLM methods, we evaluate and highlight its effectiveness for complex tasks such as sexism detection.

3.3. Data Post-Processing

Following prediction generation, the results are consolidated into unified CSV files containing all fused predictions. We then use a lightweight post-processing script to convert these predictions into the six separate JSON files required by the evaluation toolkit. For each dataset (dev and test) and each system, the script iterates over each tweet prediction and extracts the following: a single probability and binary label for Task 1_1, a four-element probability vector with a corresponding integer label for Task 1_2, and a six-element probability vector paired with six binary labels for Task 1_3. The extracted predictions are formatted and written into distinct JSON files, thus conforming precisely to the format expected by the official evaluation scorers.

4. Evaluation

In this section, we present a detailed evaluation of our proposed systems. The evaluation is structured into subsections covering the dataset, parameters and settings, results and ablation studies, and the final outcomes in terms of performance metrics.

4.1. Dataset

We use the EXIST 2025 dataset [7, 8], which includes tweets specifically curated for Tasks 1_1, 1_2, and 1_3. The dataset contains 6,920 tweets for training, 1,038 tweets for development, and 2,076 tweets for testing. The tweets are evenly balanced between English and Spanish. Each tweet in the dataset is provided as a separate JSON file with the following attributes: 1) `id_EXIST`: a unique identifier for the tweet; 2) `lang`: the language of the text ("en" or "es"); 3) `tweet`: the text content; 4) `number_annotators`: the number of annotators; 5) `annotators`: unique identifiers for each annotator; 6) `gender_annotators`: the gender of each annotator ("F" or "M"); 7) `age_annotators`: the age group of each annotator ("18–22", "23–45", or "46+"); 8) `ethnicity_annotators`: the self-reported ethnicity ("Black or African American", "Hispanic or Latino", "White or Caucasian", "Multiracial", "Asian", "Asian Indian", or "Middle Eastern"); 9) `study_level_annotators`: self-reported education level ("No high school diploma", "High school degree or equivalent", "Bachelor's degree", "Master's degree", or "Doctorate"); 10) `country_annotators`: the self-reported country of residence; 11) `labels_task1_1`: labels indicating whether the tweet contains sexist expressions or refers to sexist behaviors ("YES" or "NO"); 12) `labels_task1_2`: labels for the author's intention ("DIRECT", "REPORTED", "JUDGMENTAL", "-", or "UNKNOWN"); 13) `labels_task1_3`: arrays of labels indicating the type(s) of sexism in the tweet ("IDEOLOGICAL_INEQUALITY", "STEREOTYPING-DOMINANCE", "OBJECTIFICATION", "SEXUAL-VIOLENCE", "MISOGYNY-NON-SEXUAL-VIOLENCE", "-", or "UNKNOWN"); and 14) `split`: the subset ("TRAIN", "DEV", or "TEST") with language code ("EN"/"ES").

Each tweet is annotated by multiple annotators, who may differ in demographic characteristics, thus promoting diverse viewpoints aligned with the organizers' concept of learning with disagreement (LeWiDi).

Table 3
Task1_1 Result for All Systems

Evaluation Metrics	Transformer-Based	AR LLM-Based	Diffusion LLM-Based
ICM ↑	**0.55050**	0.13473	0.13561
ICM-Norm ↑	**0.77538**	0.56740	0.56784
F1 Score ↑	**0.80532**	0.67281	0.67222
ICM-Soft ↑	-1.82860	-1.78676	**-1.69095**
ICM-Soft-Norm ↑	0.20459	0.21135	**0.22683**
Cross Entropy ↓	4.69428	**4.55972**	4.56743

4.2. Parameters and Settings

We configure distinct parameters for each of our three systems. The first system fine-tunes an `XLM-RoBERTa-base` backbone with a maximum sequence length of 256 tokens, augmented by a small annotator metadata encoder (`AnnEncoder`). Within this encoder, annotator demographics (gender, country, ethnicity, and study level) are respectively embedded into vectors with 4, 8, 8, and 4 dimensions, while annotator age is projected to a 4-dimensional vector. These embeddings are aggregated via a 4-head multi-head attention layer to produce a unified 32-dimensional metadata summary. During Stage 1, we produce a 768-dimensional CLS embedding from both the original and cross-translated tweets, fuse them using an alpha weight ($\alpha_{\text{enc}} = 0.8$), concatenate the result with the metadata embedding (yielding an 800-dimensional vector), and train a binary classifier head (`Linear(800→1)`) using `BCEWithLogitsLoss` for 2 epochs (batch size = 16, learning rate = 2×10^{-5}, dropout = 0.1). Encoder weights are then saved and reloaded for Stage 2, which adds two additional classification heads: a 4-class head for Task 1_2 and a 6-class head for Task 1_3. The Stage 1 predicted probability is embedded (1→4 dimensions) and concatenated (resulting in 804 dimensions). Stage 2 training occurs over 3 epochs (batch size = 16, learning rate = 2×10^{-5}), using `CrossEntropyLoss` (Task 1_2) and `BCEWithLogitsLoss` (Task 1_3). Inference for both stages performs late fusion of logits (original vs. cross-translated) with $\alpha_{\text{inf}} = 0.8$. All models are run on GPUs with random seeds fixed to 42 for reproducibility.

Our second system uses a quantized, autoregressive LLaMA-3.1-8B-Instruct model with 4-bit NF4 quantization and double quantization (compute dtype = `torch.float16`). This model, along with its corresponding tokenizer (maximum prompt length = 256), is used in a zero-shot setting without fine-tuning. Tweets and annotator metadata are structured into textual prompts, and predictions are generated directly from model-derived token log probabilities. Original and cross-translated predictions are combined using late fusion ($\alpha = 0.55$), producing both soft (probabilities) and hard (binary/categorical) outputs. Inference occurs on a GPU with a fixed random seed of 42. Predictions are saved as CSV files and later converted to JSON format for evaluation.

The third system mirrors the second but replaces the autoregressive LLaMA model with the diffusion-based Dream-v0-Instruct-7B model, using identical quantization settings (4-bit NF4, compute dtype = `torch.float16`) and tokenizer parameters. The fusion weight for combining original and translated log probabilities is set slightly higher ($\alpha = 0.65$). As in the second system, inference runs without fine-tuning, and results are similarly post-processed from CSV to JSON format.

4.3. Results

Tables 3,4, and5 summarize the evaluation results on the development set across the three tasks (Task 1_1, Task 1_2, and Task 1_3). The evaluation metrics include ICM, ICM-Norm, F1 Score, ICM-Soft, ICM-Soft-Norm, and Cross Entropy [7, 8], with higher scores indicating better performance except for Cross Entropy, where lower values are preferable.

For Task 1_1 (Table 3), the Transformer-Based system achieves the best results in the hard label evaluation metrics (ICM, ICM-Norm, and F1 Score). Interestingly, in the soft probability metrics (ICM-

Table 4
Task1_2 Result for All Systems

Evaluation Metrics	Transformer-Based	AR LLM-Based	Diffusion LLM-Based
ICM ↑	**0.26836**	-0.46884	-0.46814
ICM-Norm ↑	**0.58392**	0.35339	0.35360
F1 Score ↑	**0.51215**	0.27064	0.26409
ICM-Soft ↑	**-2.04240**	-3.75278	-4.30655
ICM-Soft-Norm ↑	**0.33612**	0.19888	0.15444
Cross Entropy ↓	**2.31502**	4.09682	4.21338

Table 5
Task1_3 Result for All Systems

Evaluation Metrics	Transformer-Based	AR LLM-Based	Diffusion LLM-Based
ICM ↑	**-0.98916**	-1.20316	-1.40393
ICM-Norm ↑	**0.29310**	0.24833	0.20634
F1 Score ↑	**0.31354**	0.27177	0.24021
ICM-Soft ↑	**-5.75233**	-11.36775	-24.04366
ICM-Soft-Norm ↑	**0.20160**	0.00000	0.00000

Soft, ICM-Soft-Norm, and Cross Entropy), both LLM-based systems outperform the Transformer-Based model. Specifically, the Diffusion LLM slightly surpasses the AR LLM in ICM-Soft and ICM-Soft-Norm, whereas the AR LLM achieves the lowest Cross Entropy.

In Task 1_2 (Table 4), the Transformer-Based system consistently outperforms both LLM-based systems across all metrics for both hard and soft evaluations, while the two LLM systems provide similar performance. The Diffusion LLM performs slightly better on ICM and ICM-Norm, while the AR LLM achieves a higher F1 Score in the hard metrics. In the soft metrics, the AR LLM system significantly outperforms the Diffusion-based model.

For Task 1_3 (Table 5), the Transformer-Based system achieves superior performance across all metrics. The AR LLM system ranks second, followed by the Diffusion LLM system, indicating a clear performance advantage for the Transformer-based architecture on this complex multi-label task.

These results reflect the best performance achieved following extensive ablation studies. These studies explored multiple configurations, including data pre-processing variants (original vs. fully translated vs. cross-translated tweets), different fusion strategies (equal vs. adjustable weighting), epoch counts, and various early and late fusion α values.

Overall, our evaluation demonstrates that the Transformer-Based dual-encoder model offers robust and superior performance on most metrics and tasks. Meanwhile, the novel Diffusion LLM-based system shows performance comparable to or slightly better than the more established AR LLM-based approach on selected metrics, indicating promising potential for diffusion-based LLMs in challenging classification scenarios.

4.4. Final Rankings

This section presents the final rankings and official results of our systems evaluated on the test set. Tables 6, 7, and 8 show the published rankings from the organizers across all three tasks. Our three submitted systems correspond to: System 1 (run 1) — the Transformer-based model; System 2 (run 2) — the AR LLM-based model; and System 3 (run 3) — the Diffusion LLM-based model. The number of participating submissions per task varies: for hard parameter evaluation, there were respectively 160, 140, and 132 submissions for Task1_1, Task1_2, and Task1_3, as opposed to 67, 56, and 53 for the soft parameter evaluation.

From the official test results, it is evident that System 1 (transformer-based) consistently outperformed the other two across all tasks, securing the highest rankings and results metrics among our submissions.

When comparing the two LLM-based systems, System 3 (Diffusion LLM-based) sometimes outperformed System 2 (AR LLM-based), depending on the task and language-specific setting.

Overall, these final test rankings align with our development set findings. The transformer-based system demonstrated superior performance across the board, while the diffusion-based LLM showed competitive results, occasionally matching or even surpassing the AR LLM-based system. This highlights the potential of diffusion models as viable alternatives for multilingual, multi-task text classification.

Table 6
Final Results Test Set on Task1_1

Language	System	Hard Rank	ICM-Hard ↑	ICM-Hard Norm ↑	Macro F1 ↑	Soft Rank	ICM-Soft ↑	ICM-Soft Norm ↑	Cross Entropy ↓
All	1	60	**0.4912**	**0.7469**	**0.7512**	17	**0.6569**	**0.6053**	**0.8302**
	2	135	0.0706	0.5355	0.6444	51	-0.7700	0.3765	0.9325
	3	136	0.0504	0.5253	0.6353	50	-0.7152	0.3853	0.9415
ES	1	40	**0.5085**	**0.7543**	**0.7740**	17	**0.7204**	**0.6155**	**0.7975**
	2	137	0.0337	0.5168	0.6701	51	-0.7979	0.3720	0.9281
	3	135	0.0640	0.5320	0.6615	50	-0.7196	0.3846	0.9319
EN	1	107	**0.4587**	**0.7341**	**0.7214**	27	**0.5383**	**0.5864**	**0.8670**
	2	145	0.0855	0.5437	0.6077	55	-0.7786	0.3750	0.9376
	3	147	0.0247	0.5126	0.6037	54	-0.7365	0.3818	0.9522

Table 7
Final Results Test Set on Task1_2

Language	System	Hard Rank	ICM-Hard ↑	ICM-Hard Norm ↑	Macro F1 ↑	Soft Rank	ICM-Soft ↑	ICM-Soft Norm ↑	Cross Entropy ↓
All	1	50	**0.1529**	**0.5497**	**0.4798**	14	**-1.9749**	**0.3409**	**2.2312**
	2	91	-0.4735	0.3461	0.2658	25	-3.9860	0.1788	4.1786
	3	94	-0.5610	0.3176	0.2526	31	-4.6004	0.1293	4.3754
ES	1	34	**0.2090**	**0.5653**	**0.5066**	15	**-1.6782**	**0.3656**	**2.2205**
	2	94	-0.5286	0.3349	0.2613	25	-4.1051	0.1712	4.2326
	3	93	-0.5244	0.3362	0.2608	26	-4.1967	0.1639	4.3064
EN	1	61	**0.0655**	**0.5227**	**0.4434**	17	**-2.5469**	**0.2918**	**2.2431**
	2	100	-0.4432	0.3466	0.2648	23	-3.9332	0.1785	4.1181
	3	105	-0.6250	0.2837	0.2423	35	-5.6403	0.0390	4.4528

Table 8
Final Results Test Set on Task1_3

Language	System	Hard Rank	ICM-Hard ↑	ICM-Hard Norm ↑	Macro F1 ↑	Soft Rank	ICM-Soft ↑	ICM-Soft Norm ↑
All	1	74	**-0.5419**	**0.3742**	**0.3210**	22	**-5.8946**	**0.1887**
	2	82	-0.9882	0.2705	0.2779	33	-12.2392	0.0000
	3	95	-1.1945	0.2226	0.2418	50	-28.4446	0.0000
ES	1	72	**-0.6233**	**0.3608**	**0.3206**	22	**-6.1782**	**0.1785**
	2	85	-1.1044	0.2534	0.2731	34	-12.9536	0.0000
	3	90	-1.2585	0.2190	0.2414	44	-24.8731	0.0000
EN	1	79	**-0.4696**	**0.3849**	**0.3197**	20	**-5.4105**	**0.2036**
	2	88	-0.8756	0.2854	0.2818	30	-11.4012	0.0000
	3	96	-1.1282	0.2235	0.2413	48	-33.3211	0.0000

5. Conclusions

We introduce an annotator-aware, bilingual sexism detection pipeline evaluated on the EXIST-2025 shared task. Our methodology explicitly addresses annotator disagreement by incorporating annotator demographics directly into the modeling process, enhancing prediction robustness through the bilingual fusion of original and cross-translated tweets. We implemented and evaluated three distinct modeling strategies: a transformer-based dual-encoder architecture with fine-tuning and fusion, a zero-shot auto-regressive LLM, and a zero-shot diffusion-based LLM.

Experimental results clearly indicate that our transformer-based approach delivers superior performance across all three EXIST subtasks —sexism detection, misogynistic intent classification, and sexism type categorization— and this advantage is especially evident in metrics such as ICM, ICM-Norm, and F1 scores. While the AR LLM performed effectively in zero-shot scenarios, the diffusion-based LLM notably demonstrated competitive performance, occasionally surpassing the AR LLM on selected metrics. This finding suggests that diffusion-based LLMs hold considerable potential for text classification tasks, warranting further exploration and optimization. Our findings were also confirmed by the final rankings published by the organizers.

Future research directions can include deeper integration of annotator disagreement into the diffusion-based models and hybrid architectures combining transformer encoders with diffusion mechanisms. Extending the proposed pipeline to multimedia contexts, such as image-based memes and video content, could help broaden the applicability and generalization capabilities of automated sexism detection systems.

Declaration on Generative AI

During the preparation of this work, the authors used Open AI ChatGPT model GPT-4.5 for: Abstract drafting, drafting content, generating the literature review, grammar and spelling checking, paraphrasing and rewording, and plagiarism detection. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content.

References

[1] F. Rodríguez-Sánchez, J. C. de Albornoz, L. Plaza, J. Gonzalo, P. Rosso, M. Comet, T. Donoso, Overview of exist 2021: sexism identification in social networks, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021) at SEPLN 2021, volume 2943 of *CEUR Workshop Proceedings*, CEUR-WS.org, Málaga, Spain, 2021, pp. 195–207. URL: http://ceur-ws.org/Vol-2943/.

[2] F. Rodríguez-Sánchez, J. C. de Albornoz, L. Plaza, A. Mendieta-Aragón, G. Marco-Remón, M. Makeienko, M. Plaza, J. Gonzalo, D. Spina, P. Rosso, Overview of exist 2022: sexism identification in social networks, Procesamiento del Lenguaje Natural 69 (2022) 229–240. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6443. doi:10.26342/2022-69-20.

[3] L. Plaza, J. C. de Albornoz, R. Morante, E. Amigó, J. Gonzalo, D. Spina, P. Rosso, Overview of exist 2023 – learning with disagreement for sexism identification and characterization, in: A. Arampatzis, et al. (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. CLEF 2023, volume 14163 of *Lecture Notes in Computer Science*, Springer, Cham, 2023, pp. 398–418. URL: https://doi.org/10.1007/978-3-031-42448-9_23. doi:10.1007/978-3-031-42448-9_23.

[4] L. Plaza, J. C. de Albornoz, R. Morante, E. Amigó, J. Gonzalo, D. Spina, P. Rosso, Overview of exist 2023 – learning with disagreement for sexism identification and characterization (extended overview), in: Working Notes of CLEF 2023 – Conference and Labs of the Evaluation Forum, volume 3497 of *CEUR-WS.org*, CEUR Workshop Proceedings, Thessaloniki, Greece, 2023. URL: http://ceur-ws.org/Vol-3497/, presented at CLEF 2023, September 18–21, 2023.

[5] L. Plaza, J. C. de Albornoz, V. Ruiz, A. Maeso, B. Chulvi, P. Rosso, E. Amigó, J. Gonzalo, R. Morante, D. Spina, Overview of exist 2024 - learning with disagreement for sexism identification and

characterization in tweets and memes, in: L. Goeuriot, et al. (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction, Lecture Notes in Computer Science, Springer Nature Switzerland, Cham, 2024, pp. 93–117. To appear.

[6] L. Plaza, J. C. de Albornoz, R. Morante, E. Amigó, J. Gonzalo, D. Spina, P. Rosso, Overview of exist 2024 – learning with disagreement for sexism identification and characterization in tweets and memes (extended overview), in: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum, volume 3740 of *CEUR-WS.org*, CEUR Workshop Proceedings, Grenoble, France, 2024. URL: https://ceur-ws.org/Vol-3740/, presented at CLEF 2024, September 9–12, 2024.

[7] L. Plaza, J. C. de Albornoz, I. Arcos, P. Rosso, D. Spina, E. Amigó, J. Gonzalo, R. Morante, Overview of exist 2025: Learning with disagreement for sexism identification and characterization in tweets, memes, and tiktok videos, in: J. C. de Albornoz, J. Gonzalo, L. Plaza, A. G. S. de Herrera, J. Mothe, F. Piroi, P. Rosso, D. Spina, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction, Proceedings of the Sixteenth International Conference of the CLEF Association (CLEF 2025), 2025.

[8] L. Plaza, J. C. de Albornoz, I. Arcos, P. Rosso, D. Spina, E. Amigó, J. Gonzalo, R. Morante, Overview of exist 2025: Learning with disagreement for sexism identification and characterization in tweets, memes, and tiktok videos (extended overview), in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), CLEF 2025 Working Notes, 2025.

[9] E. Fersini, D. Nozza, P. Rosso, Overview of the evalita 2018 task on automatic misogyny identification (ami), in: Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2018), volume 2263 of *CEUR Workshop Proceedings*, CEUR-WS.org, Turin, Italy, 2018. URL: http://ceur-ws.org/Vol-2263/.

[10] V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. Rangel Pardo, P. Rosso, M. Sanguinetti, SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter, in: J. May, E. Shutova, A. Herbelot, X. Zhu, M. Apidianaki, S. M. Mohammad (Eds.), Proceedings of the 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 54–63. URL: https://aclanthology.org/S19-2007/. doi:10.18653/v1/S19-2007.

[11] A. Uma, T. Fornaciari, A. Dumitrache, T. Miller, J. Chamberlain, B. Plank, E. Simpson, M. Poesio, SemEval-2021 task 12: Learning with disagreements, in: A. Palmer, N. Schneider, N. Schluter, G. Emerson, A. Herbelot, X. Zhu (Eds.), Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), Association for Computational Linguistics, Online, 2021, pp. 338–347. URL: https://aclanthology.org/2021.semeval-1.41/. doi:10.18653/v1/2021.semeval-1.41.

[12] E. Leonardelli, G. Abercrombie, D. Almanea, V. Basile, T. Fornaciari, B. Plank, V. Rieser, A. Uma, M. Poesio, SemEval-2023 task 11: Learning with disagreements (LeWiDi), in: A. K. Ojha, A. S. Doğruöz, G. Da San Martino, H. Tayyar Madabushi, R. Kumar, E. Sartori (Eds.), Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 2304–2318. URL: https://aclanthology.org/2023.semeval-1.314/. doi:10.18653/v1/2023.semeval-1.314.

[13] A. Uma, T. Fornaciari, D. Hovy, S. Paun, B. Plank, M. Poesio, Learning from disagreement: A survey, International Journal of Language, Translation and Intercultural Communication 10 (2021) 1392–1470. URL: https://doi.org/10.12681/ijltic.29238.

[14] Y.-Z. Fang, L.-H. Lee, J.-D. Huang, Nycu-nlp at exist 2024: Leveraging transformers with diverse annotations for sexism identification in social networks, in: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, CEUR-WS.org, Grenoble, France, 2024. URL: http://ceur-ws.org/, notebook for the EXIST Lab at CLEF 2024.

[15] L. M. Quan, D. V. Thin, Sexism identification in social networks with generation-based language models: Notebook for the exist lab at clef 2024, in: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, CEUR-WS.org, Grenoble, France, 2024. URL: http://ceur-ws.org/, to appear.

[16] M. Del Tredici, D. Marcheggiani, S. Schulte im Walde, R. Fernández, You shall know a user by the

company it keeps: Dynamic representations for social media users in NLP, in: K. Inui, J. Jiang, V. Ng, X. Wan (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), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 4707–4717. URL: https://aclanthology.org/D19-1477/. doi:10.18653/v1/D19-1477.

[17] B. Vidgen, A. Harris, D. Nguyen, R. Tromble, S. Hale, H. Margetts, Challenges and frontiers in abusive content detection, in: S. T. Roberts, J. Tetreault, V. Prabhakaran, Z. Waseem (Eds.), Proceedings of the Third Workshop on Abusive Language Online, Association for Computational Linguistics, Florence, Italy, 2019, pp. 80–93. URL: https://aclanthology.org/W19-3509/. doi:10.18653/v1/W19-3509.

[18] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 8440–8451. URL: https://aclanthology.org/2020.acl-main.747/. doi:10.18653/v1/2020.acl-main.747.

[19] D. Nozza, F. Bianchi, D. Hovy, Pipelines for social bias testing of large language models, in: A. Fan, S. Ilic, T. Wolf, M. Gallé (Eds.), Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models, Association for Computational Linguistics, virtual+Dublin, 2022, pp. 68–74. URL: https://aclanthology.org/2022.bigscience-1.6/. doi:10.18653/v1/2022.bigscience-1.6.

[20] R. Sennrich, B. Haddow, A. Birch, Improving neural machine translation models with monolingual data, in: K. Erk, N. A. Smith (Eds.), Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Berlin, Germany, 2016, pp. 86–96. URL: https://aclanthology.org/P16-1009/. doi:10.18653/v1/P16-1009.

[21] T. O. Ajayi, M. Arcan, P. Buitelaar, Cross-lingual transfer and multilingual learning for detecting harmful behaviour in African under-resourced language dialogue, in: T. Kawahara, V. Demberg, S. Ultes, K. Inoue, S. Mehri, D. Howcroft, K. Komatani (Eds.), Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue, Association for Computational Linguistics, Kyoto, Japan, 2024, pp. 579–589. URL: https://aclanthology.org/2024.sigdial-1.49/. doi:10.18653/v1/2024.sigdial-1.49.

[22] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 1877–1901. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[23] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample, Llama: Open and efficient foundation language models, 2023. URL: https://arxiv.org/abs/2302.13971. arXiv:2302.13971.

[24] T. Hartvigsen, S. Gabriel, H. Palangi, M. Sap, D. Ray, E. Kamar, ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection, in: S. Muresan, P. Nakov, A. Villavicencio (Eds.), Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 3309–3326. URL: https://aclanthology.org/2022.acl-long.234/. doi:10.18653/v1/2022.acl-long.234.

[25] S. Gong, M. Li, J. Feng, Z. Wu, L. Kong, DiffuSeq: Sequence to sequence text generation with diffusion models, in: International Conference on Learning Representations, ICLR, 2023.

[26] J. Austin, D. D. Johnson, J. Ho, D. Tarlow, R. van den Berg, Structured denoising diffusion

models in discrete state-spaces, in: M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, J. W. Vaughan (Eds.), Advances in Neural Information Processing Systems, volume 34, Curran Associates, Inc., 2021, pp. 17981–17993. URL: https://proceedings.neurips.cc/paper_files/paper/2021/file/958c530554f78bcd8e97125b70e6973d-Paper.pdf.

[27] J. Ye, Z. Xie, L. Zheng, J. Gao, Z. Wu, X. Jiang, Z. Li, L. Kong, Dream 7b, 2025. URL: https://hkunlp.github.io/blog/2025/dream.

[28] S. Nie, F. Zhu, Z. You, X. Zhang, J. Ou, J. Hu, J. Zhou, Y. Lin, J.-R. Wen, C. Li, Large language diffusion models, arXiv preprint arXiv:2502.09992 (2025).