SaFe Tweets at MultiPRIDE: From Multi-Model Ensembles to Expert LLMs for Reclamation Detection

Sara Visconti[†], Federica Manzi[1,†]

[1]*Independent Researcher*
[1]*Università degli Studi di Torino - Dipartimento di Informatica, Corso Svizzera 185, 10149 Torino, Italy*

Abstract

In this work, we describe our participation in the EVALITA 2026 MultiPRIDE task, which focuses on classifying Italian and English tweets by distinguishing between reclaiming and non-reclaiming uses of potentially offensive terms against the LGBTQIA+ community. In particular, we participated in subtasks A1 (Italian), A3 (English), and B1 (Italian with author context). For Italian in both subtasks, we employed an ensemble of lightweight models using majority voting to combine their predictions. The lightweight models included a logistic regression classifier, SetFit [1], and XLM-RoBERTa [2]. For English, due to the lack of such linguistic cues and the more severe class imbalance, we adopted a more complex two-step pipeline: a DeBERTaV3 [3] model with LoRA [4] fine-tuning for initial filtering, followed by LLaMa 4 Scout [5] to refine low-confidence predictions.
Despite leveraging powerful models, classification remained challenging due to annotation subjectivity, ambiguous tweets, and limited contextual information. Our results highlight the importance of both model choice and contextual understanding in distinguishing reclaiming from non-reclaiming language.

Keywords

reclaimed language, hate-speech detection, LGBTQIA+ tweets, social media analysis, context-aware classification, EVALITA 2026 MultiPRIDE, ensemble models, transformer-based models, LLMs pipeline, LoRA fine-tuning

Warning: this paper contains examples of content that, even though obfuscated, some readers may find offensive.

1. Introduction

Contextual information is crucial in the automatic detection of hate speech and offensive language, especially for reclaimed language, where terms historically used as slurs are reclaimed and used in positive or neutral ways by members of the targeted community. This distinction is important because, without it, social media platforms relying on automated detection systems might incorrectly flag reclaimed content as toxic, effectively censoring members of the LGBTQIA+ community [6]. For this reason, it is essential to develop systems that allow for a more nuanced and context-dependent distinction between such cases to avoid mislabeling reclaimed language, while still identifying genuinely offensive content [7].

In this paper, we describe our participation in the MultiPRIDE shared task at EVALITA 2026 [8]. The task is formulated as a binary classification problem, where systems are required to determine whether words and expressions in tweets that are potentially offensive toward the LGBTQIA+ community are used in a reclaiming or non-reclaiming way by the author of the tweet. The task is divided into two main subtasks:

- **Subtask A** - Textual Content: tweets must be classified as reclaimed or not-reclaimed using the textual content of the tweet exclusively. Participants can decide to work on one or more of the following languages: Italian (A1), Spanish (A2), and English (A3).

EVALITA 2026: 9th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Feb 26 – 27, Bari, IT
[†]These authors contributed equally.
sravisconti@gmail.com (S. Visconti); federica.manzi@unito.it (F. Manzi)
https://github.com/Sara-Vis (S. Visconti); https://github.com/fede-m (F. Manzi)
0009-0008-0618-5356 (F. Manzi)

- **Subtask B** - Contextual Content: in addition to the textual content of the tweet, participants can also use the biography of the author of the tweet (when available) as extra contextual content to make the prediction. For this subtask, only tweets in Italian (B1) and Spanish (B2) were available.

Our team participated in both subtasks: for subtask A, we focused on Italian (A1) and English (A3), whereas for subtask B only on Italian (B1). For each subtask and language, we adopted different models and approaches.

After a brief description of related work in Section 2, we describe these different approaches in Section 3. Section 4 presents the results divided by subtask and language. Finally, Section 5 discusses results and presents an error analysis, while Section 7 concludes the paper.

2. Background

In this section, we provide background on both the theoretical and computational aspects of our work. The first part introduces the theoretical framework of reclamation as a linguistic phenomenon and reviews previous attempts to identify reclaimed language in NLP. The second part focuses on technical background, including the main techniques, models, and architectures used in the MultiPRIDE task.

2.1. Theoretical Background

Reclamation of slurs refers to the phenomenon in which members of a targeted, typically minority group intentionally use slurs that out-group members have historically employed to harm them, but for non-derogatory purposes. In such cases, the same word or utterance can serve as either an insult or a marker of solidarity, empowerment, or protest, depending on the specific context in which it is produced. A prominent example is the reclamation of the term "queer" by the LGBTQIA+ community at the end of the 20th century. As explained in detail in [9], activist groups adopted the slur "queer", which had acquired a strongly negative and derogatory connotation in the course of the century, and overturned its discriminatory power into an aggressive and powerful tool of protest and political identity.

The mechanisms underlying such reclaiming uses have been analyzed from different theoretical perspectives. Following the *echoic* perspective proposed in [10], in-group members of a community may echo expressions that are derogatory against their own group, in contexts that signal dissociation from their original offensive meaning, often producing an ironic effect. Furthermore, [11] discusses the practice of *reading*, common in drag queen communities, in which slurs are used against in-group members as a way of forming a "thick skin" and learning how to promptly reply when those same slurs are used with harmful intent by out-group members.

All these accounts highlight how context and pragmatic cues are crucial for determining whether a slur is used in an offensive or reclaiming way. This is particularly relevant in the context of content moderation on social media. For example, [6] evaluates the levels of toxicity attributed to Twitter posts from drag queens compared to those from white nationalist accounts using the Perspective API[1], a system developed by Jigsaw for automatic content moderation. Their results show that tweets from drag queens were assigned higher toxicity scores, due to the higher frequency of slurs like "f*g" and "b*tch", as well as neutral terms like "gay" or "lesbian", highlighting the inability of such systems to account for context and accurately distinguish reclaiming usages. Similar results were also reported in [12], where the authors compare different commercial content moderation systems and note how they all show the tendency to over-moderate marginalized groups, including the LGBTQIA+ community.

Although the task remains largely underexplored, recent NLP work has started to investigate reclaimed language and, more broadly, the distinction between abusive and non-abusive slur usage depending on context. A notable example in the Italian setting is the ReCLAIM project [7]. The authors leveraged the HODI dataset [13], which contains tweets annotated for hate speech, filtered the dataset for tweets including slurs commonly directed at the LGBTQIA+ community, and re-annotated them as

[1] https://perspectiveapi.com/

reclaimed or not-reclaimed. The resulting dataset was then used to evaluate the zero-shot capabilities of an LLM to classify tweets as reclaimed or not. Furthermore, other relevant works on this topic are [14], [15], [16], and [17].

2.2. Technical Background

From a more technical perspective, the systems presented in the current work build upon several recent advances in transformer architectures and parameter-efficient fine-tuning. In this sub-section, we provide background on the key techniques and models we employ, including SetFit [1], DeBERTa [18], LoRA [4], and Mixture-of-Experts-based large language models such as the LLaMA 4 family.

Setfit (Sentence Transformer Fine Tuning) is a few-shot learning framework [19] that fine-tunes pre-trained sentence transformer models [20] to create a task-specific embedding space using contrastive learning. During training, triplets are formed, each consisting of a pair of training instances and a label indicating class similarity (1 if the examples belong to the same class, 0 otherwise). The model learns embeddings that bring pairs belonging to the same class closer while maximizing the distance between instances of different classes. This embedding space can then be used with a lightweight classifier, such as logistic regression, to perform classification. Being a framework for few-shot learning, the main advantage of using SetFit is that it is suitable for small or imbalanced datasets, as in our case, where the minority class contains relatively few examples.

The DeBERTa architecture, introduced in [18], builds on BERT and RoBERTa by introducing two new elements. The first is disentangled attention, which consists of representing tokens using two separate vectors: one encoding their content and the other their position. Attention weights among words are also disentangled: instead of having a single score per token-pair (encompassing both content and position information), the model computes three separate scores, namely content-to-content, content-to-position, and position-to-content. This enables the model to capture word order and relationships between tokens better. The second major novelty is an enhanced mask decoder, which improves the model's ability to learn representations for masked tokens by leveraging positional information to make the prediction.

LoRA (Low-Rank Adaptation) is a parameter-efficient technique for fine-tuning models on downstream tasks by learning a small number of additional trainable parameters while keeping the original model weights frozen. Instead of updating the full transformer parameters, LoRA learns low-rank matrices that approximate the required weight updates and are added to the original weights. This allows for fine-tuning models at reduced computational costs, making it efficient, especially as the number of model parameters scales up.

Finally, the core idea of the Mixture of Experts (MoE) paradigm, introduced in [21], consists of decomposing a model into multiple specialized experts. For a given input, a routing mechanism selects the most relevant expert or group of experts, and only their corresponding parameters are activated to produce the final prediction. This allows for reducing computational and memory requirements and scaling to a large number of parameters while keeping inference and training costs manageable. We refer to [22] for a more in-depth overview.

More recently, MoE architectures have been successfully adopted in transformer-based LLMs, such as Switch Transformers [23], Mixtral [24], DeepSeek-MoE [25], and the LLaMA 4 family [5], which we use in our experiments. Specifically, LLaMA 4 Scout[2] has a total of 109B parameters distributed across 16 experts, which allows to activate only 17B parameters for each input. Together with LLaMA 4 Maverik[3], which has 400B total parameters, 128 experts, and 17 B active parameters, LLaMA Scout was obtained by distilling from LLaMA 4 Behemoth, a much larger teacher model with approximately 2T parameters, 16 experts, and 288 B active parameters.

[2] https://huggingface.co/meta-llama/Llama-4-Scout-17B-16E

[3] https://huggingface.co/meta-llama/Llama-4-Maverick-17B-128E

3. Description of the System

In this section, we first describe the data and preprocessing steps and then the systems developed for the subtasks we participated in: subtask A in Italian (A1) and English (A3), and subtask B in Italian (B1). For each task and language, we adopted different architectures and modeling approaches, including an ensemble-based classifier and a two-step pipeline.

3.1. Data and Preprocessing

Table 1
Distribution of training data across subtasks, languages, and classes.

Subtasks	Language	Class 0	Class 1	Total
A, B	Italian	879	207	1086
A	English	938	88	1026

For the MultiPRIDE shared task, the organizers initially released a labeled training set for each language. The datasets for both subtasks were the same, except that the author's biography could not be used for subtask A. Each instance consisted of a tweet, the author's biography (available only for Italian in subtask B), the language, and the annotated binary label. The label indicated whether potentially offensive or neutral words and expressions targeting the LGBTQIA+ community in the tweet were used in a reclaimng (label 1) or non-reclaiming way (label 0).
It is crucial to stress that tweets belonging to label 0 included not only explicitly offensive tweets, but also instances that, even if not offensive or even positive, could not be classified as reclaimed.

Table 1 reports the distribution of each class per language. For both languages, the majority of instances belong to class 0, i.e., not-reclaimed, making both datasets imbalanced. This imbalance is even more pronounced in the English dataset, where class 1 accounts for only 8% of the dataset. To provide a more detailed overview of the linguistic properties of the data, Figure 1 shows the top 10 most frequent unigrams extracted from tweets in each class for English and Italian. The English dataset reveals a considerable lexical overlap between the two classes: the six most frequent unigrams, including the slur "f*ggot" and terms like "gay" and "queer", appear in both reclaimed and not-reclaimed tweets. This indicates that, in English, class distinction cannot rely on simple differences in the vocabulary. Conversely, Italian shows a much clearer separation of words across classes: class 1 is strongly associated with community-specific self-identifiers and markers of pride (e.g., "forci", "fr*cia", "l*lla", "pride"), whereas class 0 clusters around terms related to general discourse on transphobia and gender (e.g., "trans", "donne", "transfobia").
Furthermore, the log-odds ratios in Figure 2 highlight the most informative words for each class and confirm this trend. In Italian, the words most strongly correlated with each class are relevant to the reclamation topic, reflecting the same thematic difference (community self-identifications vs transphobia and gender) observed in unigrams. In contrast, English words for both classes are very generic and not relevant to the topic (e.g., "disease", "server", "everyone"). Based on these observations, we hypothesize that classifying tweets in Italian is likely to be more straightforward than in English, as clearer linguistic patterns can be identified.

Regarding data preprocessing, different strategies were adopted for each subtask and language. For Italian in subtask A, the data were kept unchanged. For English, we experimented with several approaches to try to mitigate the severe class imbalance. In particular, we attempted to undersample the majority class (class 0) or to upsample the minority class (class 1). Undersampling was applied by randomly selecting a subset of class 0 instances equal in size to the number of class 1 instances. We also experimented with larger subsets corresponding to 2:1 and 3:1 ratios of class 0 to class 1.
Upsampling was implemented using two strategies. The first consisted of simply duplicating each class 1 tweet in the training set. The second and more elaborate approach relied on data augmentation using the *llama-4-scout-17b-16e-instruct* model. For each tweet classified as reclaimed, we asked the LLM

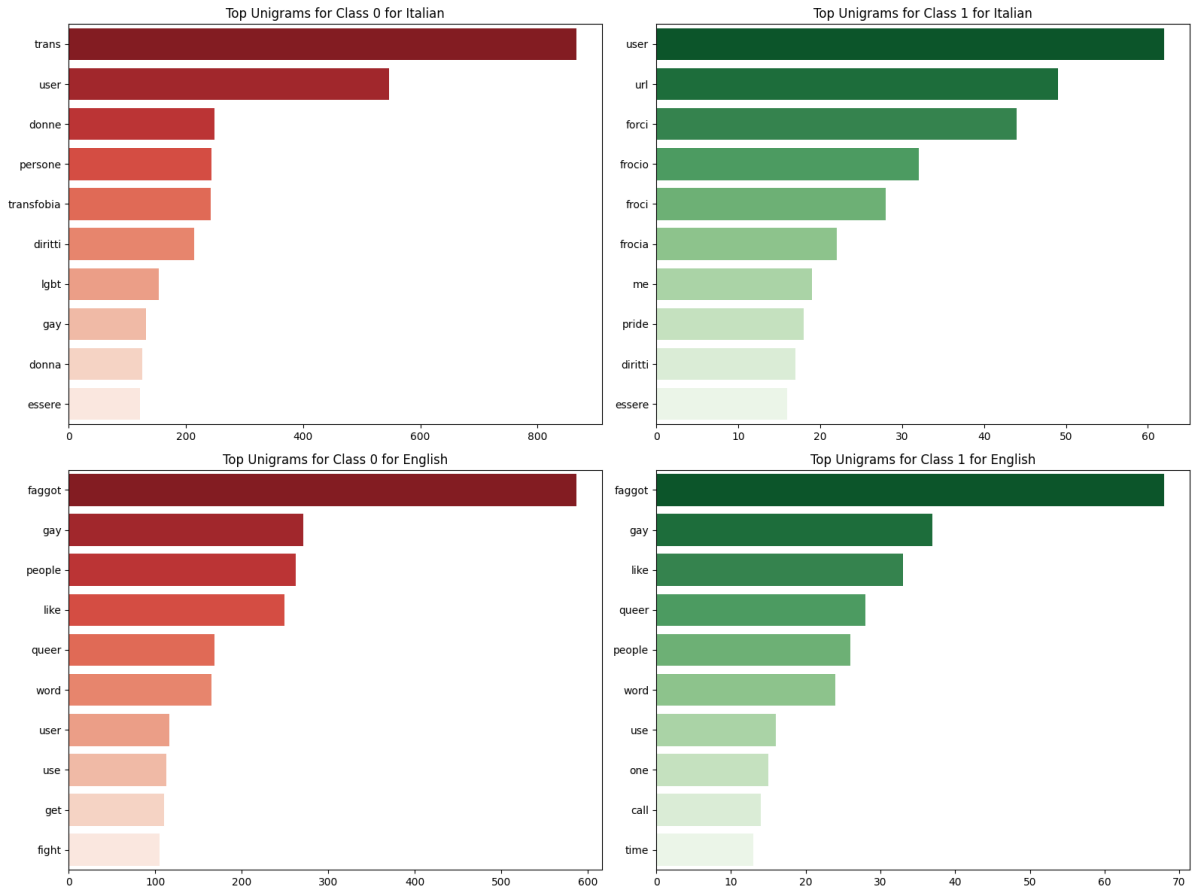

Figure 1: Comparison of the top 10 most frequent unigrams by class between Italian (top) and English (bottom). The red bars represent the not-reclaimed class (class 0), while the green bars represent the reclaimed class (class 1). Note the significant lexical overlap in English, where the top 6 unigrams are the same for both classes, compared to the distinct separation in Italian.

to generate two new paraphrased versions of the tweet. Since we did not want to give the model the responsibility to regenerate correct reclaimed versions (potentially changing offensive terms that could have changed the label of the tweet), we crafted a fixed list of words that had to stay the same in the paraphrased tweets. This ensured that the model only changed the non-crucial parts with no influence on the label of the tweet.

Finally, for subtask B, the author's biography was concatenated to the tweet text. To distinguish the biography from the tweet, it was introduced with the string "BIO" at the beginning, allowing the model to incorporate additional contextual information when making predictions.

3.2. Models and Approaches

As discussed above, we used different modeling approaches depending on the language and subtask. In the following subsections, we describe each system architecture and training strategy in detail.

3.2.1. Subtask A1 - Textual Content in Italian

For subtask A1, we proposed an ensemble system (Figure 3) composed of three models employed as separate judges: a logistic regression classifier, a SetFit model [1], and an encoder-only transformer-based model (XML-RoBERTa [2]). Each model was trained separately before being combined into the ensemble.

The logistic regression model was trained on TF-IDF embeddings of the tweets. Each tweet was

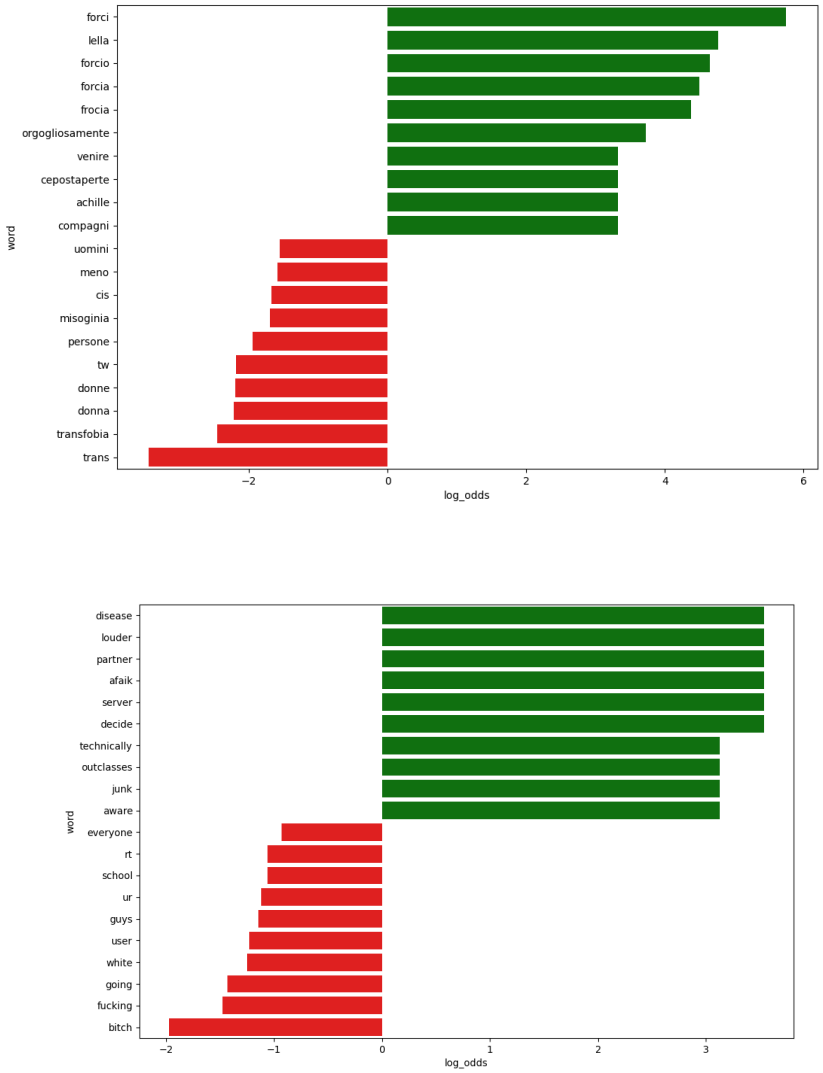

Figure 2: Comparison of the log-odds ratios of the top informative words for Italian (top) and English (bottom). Positive values (green bars) indicate words most strongly correlated to the reclaimed class (Class 1), while negative values (red bars) indicate words associated with the not-reclaimed class (Class 0). The Italian words show a strong semantic connection to the LGBTQIA+ community topic (e.g., *forcio*, *lella*), whereas the English words are very generic and do not have any semantic relevance to the topic.

encoded as a 6000-dimensional vector, corresponding to the 6000 most frequent unigrams and bigrams in the training set. Each vector entry represented the TF-IDF weight of the corresponding unigram or bigram in that tweet, reflecting its importance in the tweet relative to the entire training corpus.
For logistic regression, we used the *scikit-learn* implementation [26] and set the regularization coefficient C to 1, providing a moderate level of L2 regularization that balances underfitting and overfitting. To address the class imbalance, we set the *class weight* parameter to *balanced*, which automatically scales the weights inversely proportional to class frequencies, ensuring that the minority class has higher influence during training. The maximum number of iterations was increased to 1000 to ensure convergence, given the high dimensionality of the input vectors. The number of features and the n-gram range of TF-IDF, together with the regularization coefficient C were selected using grid search. All other logistic regression parameters were kept at their default values.

As mentioned in Section 2, the SetFit framework is composed of two steps: the first involving Sentence Transformers to embed the input text into a task-specific learned vector space, and the second relying

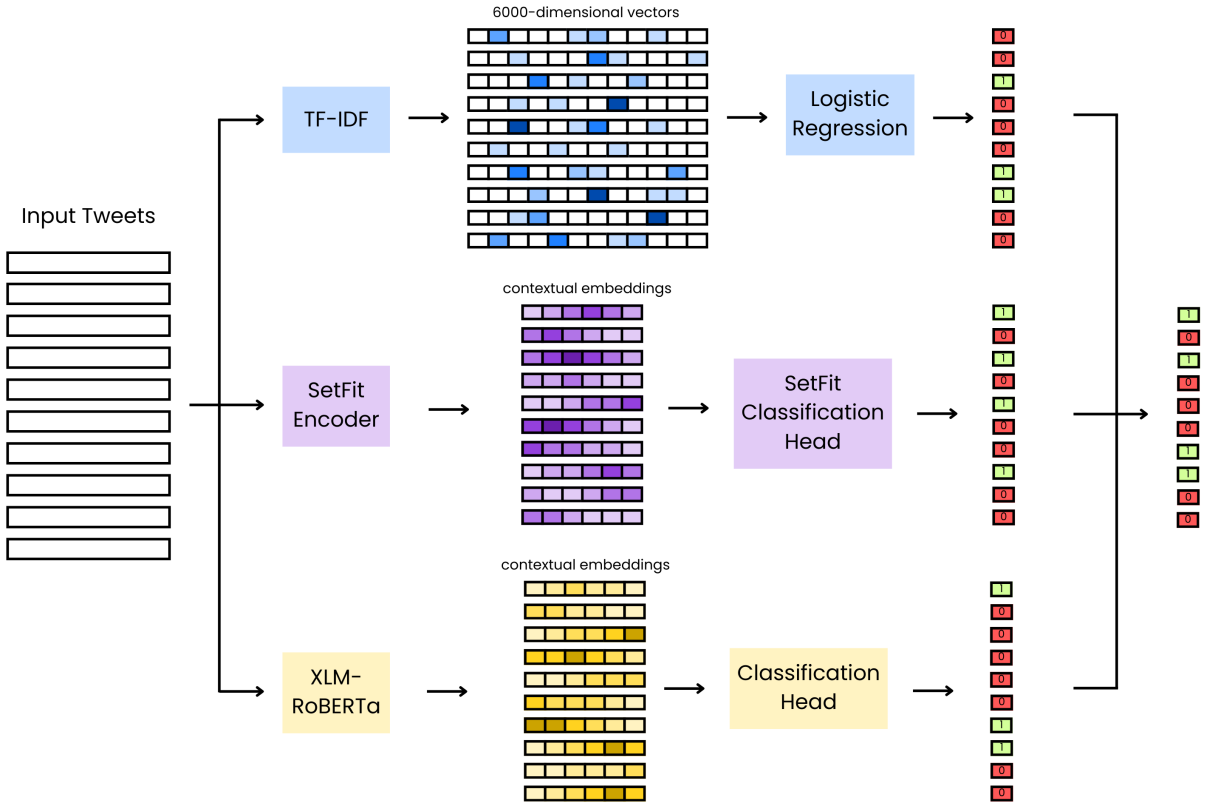

Figure 3: Ensemble system composed of three models employed as separate judges: logistic regression classifier, SetFit model, and encoder-only transformer-based model, proposed for subtask A1.

on a lightweight classification head to output the final prediction. In our experiments, embeddings were generated using *paraphrase-multilingual-MiniLM-L12-v2*[4]. Training was carried out for 10 epochs with a batch size of 8, setting the sampling strategy to undersampling to deal with class imbalance and reduce the number of instances of the majority class. As the classifier, the default logistic regression head was retained.
It is important to note that this approach differs from simply feeding a logistic regression classifier with sentence transformer embeddings, which we also experimented with, because SetFit focuses on learning task-specific embeddings rather than simple semantic representations.

For the third judge of the ensemble system, we fine-tuned XLM-RoBERTa [2], the multilingual version of the RoBERTa transformer model [27], on the full training set for 4 epochs and with a batch size of 8, with a learning rate of 1.03e-5, and a weight decay of 0.06. We performed hyperparameter optimization using Optuna [28]. We used the default cross-entropy loss function and the AdamW optimizer. The maximum sequence length was set to 128 tokens, as tweets are generally short and do not require the full default length of 512 tokens.

At inference time, each tweet was first classified by the three models (logistic regression, Setfit, and XLM-RoBERTa) independently. The final prediction was then obtained via majority voting, by looking at the class predicted by at least two of the three models.

3.2.2. Subtask A3 - Textual Content in English

For Subtask A3 (English), the severe class imbalance in the English dataset posed a major challenge. As discussed in 3.1, there was no clear semantic and lexical distinction between words from different classes, making it difficult for models to learn meaningful patterns. To mitigate this effect, we tried

[4]https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2

different approaches for both data preprocessing and problem modeling. First, we undersampled the majority class (class 0) to contain only twice as many instances as class 1. The resulting dataset consisted of all 88 examples of class 1 and only 176 examples of class 0.

The final system we proposed is a two-step pipeline showed in Figure 4. In the first step, we used the DeBERTaV3 pre-trained transformer model [3], based on the DeBERTa architecture presented in 2, which we fine-tuned on the undersampled dataset using Low-Rank Adaptation (LoRA). This approach not only reduces computational and memory costs, but is also useful in few-shot scenarios or imbalanced datasets, such as the English subtask considered here. Training was carried out for 10 epochs with a batch size of 16, a learning rate of 1e-3, and a weight decay of 0.01. No class weights were needed due to the undersampling strategy.
For LoRA, we set the rank to 8: we deliberately chose a small value to limit how much model weights could change and reduce the risk of overfitting given the size of the dataset. This choice was motivated by early experiments, in which full fine-tuning of DeBERTa without LoRA led to overfitting. Moreover, the scaling factor alpha was set to 16 and dropout to 0.1 for regularization. At inference, the logits produced by DeBERTaV3 in the output layer are converted to probabilities using softmax. The confidence of the prediction is defined as the maximum probability across classes. We defined a threshold of 0.9, so that if the confidence falls below this threshold, the prediction is considered uncertain and the tweet undergoes the second step of the pipeline for further analysis.

The second step leveraged LLaMa 4 Scout, a powerful large language model based on a Mixture-of-Experts (MoE) architecture, to reclassify tweets for which the deBERTaV3 prediction was uncertain, i.e., whose confidence was falling below the 0.9 threshold. In the prompt fed to LLaMa 4, the task was framed as a three-class classification problem. Class 1 corresponded to reclaimed tweets, while the non-reclaimed class was subdivided into class 0, which included openly offensive tweets, and class 2, which covered neutral or even positive tweets towards the LGBTQIA+ community that are not inherently reclamatory.

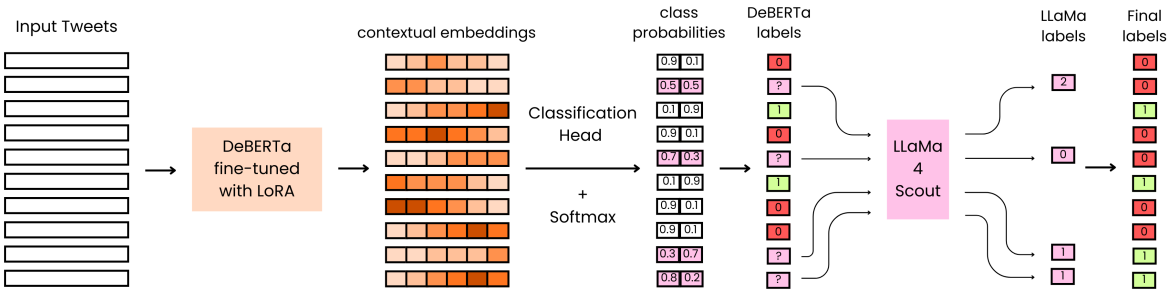

Figure 4: Two-step pipeline: pre-trained transformer model fine-tuned with LoRA, followed by an LLM to reclassify uncertain predictions, proposed for subtask A3.

The model received as input the original tweet together with the DeBERTaV3 prediction and was asked to judge whether the prediction of the smaller model was correct, and to provide an appropriate final label (0, 1, or 2). In a post-processing step, predictions assigned to class 2 were mapped to class 0 to match the binary label scheme of the task. The model was used in a zero-shot setting, meaning that no labeled tweets or in-context examples were provided in the prompt. Instead, the LLM was guided in its prediction by a set of rules, such as identifying first-person pronouns paired with slurs as markers of in-group membership and thus of reclaimed usage, as well as criteria to distinguish offensive from neutral content. To serve the LLM (specifically *llama-4-scout-17b-16e-instruct*), we used Groq[5], a specialized hardware and software platform designed for low-latency inference. Additionally, we employed the Python *instructor* [29] library to enforce a structured JSON output, so that we did not have to deal with free text.

During the experimentation phase, we conducted several ablation studies for both steps of the

[5]https://groq.com/

pipeline. For step 1, we evaluated alternative classifiers, such as SetFit and logistic regression. Both approaches resulted in a recall stuck at 0.5, indicating that the models were predicting the majority class for all instances. Moreover, the output probabilities for both models were very similar to each other and therefore not very informative: logistic regression produced probabilities concentrated around 0.5, while SetFit around 1.0. As a result, it was not possible to establish a meaningful confidence threshold for further refinement with an LLM.
For step 2, we experimented with using LLaMa 4 Scout to directly predict all tweets, without the DeBERTa pre-filtering step, but this yielded worse results. Even with the pre-filtering step, we found that framing the task as an open classification task instead of as a correction of the output of DeBERTa also produced worse results. Other open-source LLMs were tested in Step 2, but LLaMa 4 Scout consistently outperformed alternatives, including LLaMa 4 Maverik and Qwen3 [30].

3.2.3. Subtask B1 - Contextual Content in Italian

For this subtask, we adopted the same ensemble architecture used for subtask A1, consisting of logistic regression, SetFit, and XLM-RoBERTa.
As described in Section 3.1, the only change concerned the input, where the tweet text was concatenated with the author's biography. The rationale behind this choice was to allow the model to leverage potentially informative cues present in user bios, such as self-identifiers and markers of pride that may signal membership in the LGBTQIA+ community.

While this approach does not introduce any substantial subtask-specific architectural change, we believe it provides a simple and straightforward baseline for evaluating the contribution of the bio information without introducing additional complexity. More sophisticated strategies, such as explicitly predicting the author's community membership and using this information to guide tweet classification, would require defining labels for LGBTQIA+ membership, which may be highly subjective and difficult to determine automatically. Although some bios contain explicit membership cues (e.g., pride flags) and tweets labeled as reclaimed are assumed to be authored by community members (following the assumption also used in [7]), explicitly modeling membership without relying on true labels could introduce additional bias and subjectivity. For this reason, we chose to rely solely on the provided labels and to supply the model with the biography as mere additional context to learn from.

4. Results

As mentioned earlier, our team participated in subtask A for Italian (A1) and English (A3), and subtask B for Italian only (B1). The test sets were released after the experimental phase without true labels. For each subtask, we applied our systems on the corresponding test data, collected the predicted labels and submitted them to the task organizers for evaluation.
Although up to two runs with different architectures and systems could be submitted for each subtask, we submitted a single run per task. Thus, only one set of results is available for each task.

The official ranking for the shared task was based on the macro F1 score. Accordingly, Table 2 reports macro-averaged precision, recall, and F1-score. Among the three subtasks, we can note that subtask A3 (English) exhibits the lowest performance across all metrics. This observation is consistent with the hypothesis we formulated during data analysis in Section 3.1, where we highlighted the lack of clear lexical and semantic patterns distinguishing the two classes in the English training set.
Furthermore, the results for subtask B1 are lower than those for A1, despite using the same system architecture. This suggests that simply including additional contextual information from the author's biography, without adapting the underlying system accordingly, did not improve performance and may have even introduced noise.

Table 2
Results per subtasks

Subtask	Language	Precision (macro)	Recall (macro)	F1-Score (macro)
A1	Italian	0.9263	0.8620	0.8895
A3	English	0.6260	0.6412	0.6329
B1	Italian	0.9251	0.7664	0.8164

5. Discussion

The results highlight differences in the ability of systems to distinguish reclaiming from non-reclaiming usages across both languages and subtasks.

The linguistic properties of the data can largely explain the difference in performance between English and Italian in subtask A. The lexical analysis performed in Section 3.1 reveals that Italian exhibits strong correlations between specific vocabulary (e.g., community self-identifiers) and the reclamation label, making it easier to distinguish between classes by simply examining the linguistic features of the tweets. This made it possible to employ an ensemble system composed of smaller models such as Logistic Regression, SetFit, and XLM-RoBERTa for subtask A1.

In contrast, the English dataset exhibits high vocabulary overlap between classes, with the same frequent terms appearing in both classes. These different characteristics of the two languages help explain why classification in Italian (A1) achieved better performance than in English (A3), where distinguishing reclamation could not rely on explicit linguistic cues.

A further challenge was represented by class imbalance, which, although present in both datasets, was particularly severe in the English setting, where reclaiming tweets (class 1) made up only 8% of the entire dataset. Furthermore, unlike Italian, where this imbalance could be partially mitigated by a clearer semantic and lexical separation between classes, English could not rely on such linguistic distinctions. For several of the systems and models we experimented with, including SetFit and logistic regression, this resulted in recall values stuck at 0.5, with the models predicting the majority class for every instance. Moreover, the confidence scores associated to these predictions were tightly clustered around 1 for SetFit and 0.5 for logistic regression, making them unsuitable for confidence-based refinement strategies.

For these reasons, a more complex solution was required for the English subtask, motivating the adoption of the proposed two-step pipeline. Among the tested models, DeBERTaV3 fine-tuned with LoRA was the only configuration that achieved a recall above 0.5, indicating that it was able to learn patterns rather than just collapsing to predicting the majority class. Additionally, unlike simpler models, its confidence scores were sufficiently diverse to allow for the introduction of a threshold and the consequent second LLM-based refinement step.

In this second step, tweets for which DeBERTaV3's prediction fell below the confidence threshold were re-evaluated using a more powerful model. Using an LLM allowed us to provide a more detailed task description in natural language and clear rules and patterns to guide the model's decision process. Moreover, introducing a "neutral" class (class 2) helped the model distinguish actually reclaiming tweets from the ones whose content was neutral or even positive toward the LGBTQIA+ community but not inherently reclaiming. This allowed to reduce the number of false positives for class 1 in ambiguous cases. Interestingly, we also noticed that framing the task as a correction of the output of a smaller model yielded better results than directly asking the LLM to perform open classification of the tweets.

When comparing results across subtasks, we can observe a decrease in performance for subtask B1 compared to subtask A1. This suggests that simply concatenating additional contextual information about the author to the tweet text may introduce noise and worsen performance, rather than improve it, when such information is not explicitly modeled. This result highlights the need for a more structured approach to incorporate such information, ensuring that it contributes meaningfully to the prediction instead of introducing noise. However, our experiments also highlighted that determining whether an author belongs to the LGBTQIA+ community, and therefore whether they can use slurs and offensive

terms in a reclaiming way, is a highly non-trivial task, even for human annotators.

6. Error Analysis

The following error analysis highlights both some of the limitations of our approach and inherently ambiguous cases that are difficult to resolve automatically. For subtask A1 (Italian), the ensemble system sometimes struggled to distinguish between reclaiming and non-reclaiming uses of pejorative words such as "fr*cio", even when they co-occurred with first-person pronouns or markers of pride.

Tweet: "Che poi non è tanto per **noi fr*ci**, è che non so davvero come si possa esultare per aver negato tutele e diritti a delle persone disabili"

Translation: "It's not even that much about **us f*ggots**, I just really don't understand how you can cheer for denying protections and rights from disabled people"

Comment: Here, the slur in the tweet is clearly used in a reclaiming way, since the pronoun "us" identifies the author as belonging to the LGBTQIA+ community. However, our model did not label this tweet as reclaimed.

In contrast, there were instances where relying solely on these linguistic markers could be misleading. For example, the author of a tweet specifies that they belong to the LGBTQIA+ community, but they are offended by slur words such as "fr*cio" or "r*cchione". In this case, the non-reclaiming nature of the tweet depends on its content rather than the presence of in-group markers, since the author explicitly rejects their reclaiming use, making classification challenging.

Tweet: "@USER Io sono orgoglioso di essere gay. Non mi offendo se mi chiami gay. **Mi offendo** se mi chiami fr*cio, f*nocchio, r*cchione, ch*cca e chi più ne ha più ne metta"

Translation: "@USER I'm proud of being gay. I won't be offended if you call me gay. **I'll get offended** if you call me f*ggot, q*eer, s*ssy, and so on and so forth."

Comment: Here, there are markers of pride that help identify the author as part of the LGBTQIA+ community. However, the author says that they get offended if someone uses slurs towards them. Thus, the slurs they wrote in the tweet were not written to reclaim them, making the tweet not-reclaimed. Our system labeled this instance as reclaimed.

This shows that, despite generally achieving better performance on Italian than English, simpler models alone may not suffice for more complex cases, where more reasoning capabilities are needed to interpret the context correctly. However, some ambiguous cases like the one in the example remain challenging, possibly even for humans.

For subtask A3 (English), errors reveal additional challenges. The annotation appears less consistent than in Italian, and it is often unclear why a particular tweet was labeled as reclaimed (class 1) rather than not-reclaimed (class 0). Some annotation rules, such as requiring the author to belong to the community for reclaiming usage, seemed less consistently applied, especially since the English dataset did not include author biographies. This increases subjectivity and makes it harder for models to learn clear patterns to distinguish between classes.

Tweet 1: "F*ck Louis CK and any straight d*uchebag who feels legitimate calling someone a "f*ggot" because of his bullshit stand-up. Comedians don't get to decide what is acceptable and what is not.

Tweet 2: "Sometimes I think that being gay is completely accepted in our society and then I'm quickly reminded by someone at another table that I'm a "f*ggot"'"

Comment: Here, we cannot be certain whether the tweets' authors belong to the LGBTQIA+ community. However, they are both reporting cases in which the same slur was used in a non-reclaiming way by someone who is not part of the community. Thus, one might want to classify these tweets as not-reclaimed (label 1). Nevertheless, one might assume the authors are indeed part of the community, since they seem knowledgeable enough on the topic to avoid using the slurs when not belonging to the community. In this scenario, they might have decided to use the exact slur that was used against them to reclaim it. Therefore, the reclaimed class (label 0) would also be appropriate. In the dataset, the first tweet had label 1 while the second had label 0.

Furthermore, the limited performance observed even when using more powerful LLMs suggests that simply increasing model capacity may not necessarily lead to better results. Even with prompting, the rules provided to distinguish between classes need to be aligned with the annotator's perceptions and annotation guidelines for them to be effective. This emphasizes the inherent difficulty of the task and its leaning towards subjectivity and indicates that providing clearer annotation guidelines could benefit automatic classification approaches.

Overall, this analysis highlights that the task is challenging even for humans, particularly when linguistic cues are subtle or the annotation schema is ambiguous, limiting the effectiveness of automatic systems, even for powerful LLMs. For LLMs to work effectively, clear and detailed instructions are essential. However, providing such specific instructions is challenging in a task that is inherently subjective and open to interpretation, as it requires an in-depth description of how to handle diverse and nuanced cases.

7. Conclusion and Future Work

In this work, we described our participation in the EVALITA 2026 MultiPRIDE task. This task focused on classifying Italian and English tweets containing words and slurs potentially offensive to the LGBTQIA+ community, distinguishing between reclaiming and non-reclaiming uses. The task proved difficult due to a series of challenges. On a technical level, the high class imbalance was problematic, especially for the English dataset. On a theoretical level, the task is inherently difficult, as determining what counts as reclaiming is subjective and not always consistent.

Our experiments showed that Italian tweets (subtask A1) were easier to classify due to clearer lexical and semantic distinctions between classes, allowing simpler models to perform well, without the need for powerful reasoning models. In this case, using an ensemble model composed of logistic regression, SetFit, and XLM-RoBERTa with majority vote was sufficient to achieve decent performance.
In contrast, English tweets (subtask A3) presented further challenges, due to the lack of clear linguistic clues to distinguish the reclaimed from the not-reclaimed class. This required the usage of a more complex two-step pipeline, which leverages DeBERTaV3 with LoRA fine-tuning as a first pre-filtering step, and a more powerful model, LLaMa 4 Scout, in the second step to refine low-confidence predictions. However, despite using powerful LLMs, classification remained difficult due to annotation subjectivity, ambiguous tweets, and the limited availability of contextual information in English.

Further work could explore more structured methods for incorporating author context (subtask B1), a more balanced dataset, especially for English, and clearer annotation guidelines that can be used to structure prompts and explain the task to LLMs more effectively. Overall, our study highlights the potential and limitations of using both traditional machine learning models and more powerful LLMs for nuanced and context-dependent detection of hate speech and offensive language in social media.

Declaration on Generative AI

During the preparation of this work, the authors used GPT-4o, Gemini 3 Pro, and Grammarly to correct grammar, spelling and improve the readability of the paper through sentence polishing and rephrasing. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] L. Tunstall, N. Reimers, U. E. S. Jo, L. Bates, D. Korat, M. Wasserblat, O. Pereg, Efficient Few-Shot Learning Without Prompts, arXiv (2022). doi:10.48550/arXiv.2209.11055.

[2] 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: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 8440–8451. doi:10.18653/v1/2020.acl-main.747.

[3] P. He, J. Gao, W. Chen, DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing, 2021. arXiv:2111.09543.

[4] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, LoRA: Low-Rank Adaptation of Large Language Models, in: International Conference on Learning Representations, 2022. URL: https://openreview.net/forum?id=nZeVKeeFYf9.

[5] Meta AI, The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation, 2025. URL: https://ai.meta.com/blog/llama-4-multimodal-intelligence/.

[6] T. Oliva, Fighting Hate Speech, Silencing Drag Queens? Artificial Intelligence in Content Moderation and Risks to LGBTQ Voices Online, Sexuality & Culture (2021). doi:10.1007/s12119-020-09790-w.

[7] L. Draetta, C. Ferrando, M. Cuccarini, L. James, V. Patti, ReCLAIM Project: Exploring Italian Slurs Reappropriation with Large Language Models, in: Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024), CEUR Workshop Proceedings, Pisa, Italy, 2024, pp. 335–342.

[8] C. Ferrando, L. Draetta, M. Madeddu, M. Sosto, V. Patti, P. Rosso, C. Bosco, J. Mata, E. Gualda, MultiPRIDE at EVALITA 2026: Overview of the Multilingual Automatic Detection of Slur Reclamation in the LGBTQ+ Context Task, in: Proceedings of the Ninth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2026), CEUR.org, Bari, Italy, 2026.

[9] E. Nossem, Queer, Frocia, Femminiellə, Ricchione et al. – Localizing 'Queer' in the Italian Context., Gender/Sexuality/Italy 6 (2019) 1–27. doi:10.15781/31yc-ys20.

[10] C. Bianchi, Slurs and Appropriation: An Echoic Account, Journal of Pragmatics 66 (2014) 35–44. doi:10.1016/j.pragma.2014.02.009.

[11] S. McKinnon, "Building a thick skin for each other": The use of 'reading' as an interactional practice of mock impoliteness in drag queen backstage talk, Journal of Language and Sexuality 6 (2017) 90–127. doi:10.1075/jls.6.1.04mck.

[12] D. Hartmann, A. Oueslati, D. Staufer, L. Pohlmann, S. Munzert, H. Heuer, Lost in Moderation: How Commercial Content Moderation APIs Over- and Under-Moderate Group-Targeted Hate Speech and Linguistic Variations, in: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, CHI '25, ACM, 2025, p. 1–26. doi:10.1145/3706598.3713998.

[13] D. Nozza, A. T. Cignarella, G. Damo, T. Caselli, V. Patti, HODI at EVALITA 2023: Overview of the first Shared Task on Homotransphobia Detection in Italian, in: EVALITA, volume 3473 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023.

[14] E. Zsisku, A. Zubiaga, H. Dubossarsky, Hate Speech Detection and Reclaimed Language: Mitigating False Positives and Compounded Discrimination, Proceedings of the 16th ACM Web Science Conference (2024). URL: https://api.semanticscholar.org/CorpusID:269228809.

[15] E. W. Pamungkas, V. Basile, V. Patti, Do You Really Want to Hurt Me? Predicting Abusive Swearing in Social Media, in: International Conference on Language Resources and Evaluation, 2020. URL: https://api.semanticscholar.org/CorpusID:218974299.

[16] A. T. Cignarella, M. Sanguinetti, S. Frenda, A. Marra, C. Bosco, V. Basile, QUEEREOTYPES: A Multi-Source Italian Corpus of Stereotypes towards LGBTQIA+ Community Members, in: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, Torino, Italia, 2024, pp. 13429–13441. URL:

https://aclanthology.org/2024.lrec-main.1176/.
[17] L. Veloso, L. Hirlimann, P. Wicke, H. Schütze, SLAyiNG: Towards Queer Language Processing (2025). doi:10.48550/arXiv.2509.17449.
[18] P. He, X. Liu, J. Gao, W. Chen, DeBERTa: Decoding-enhanced BERT with Disentangled Attention, in: International Conference on Learning Representations, 2021. URL: https://openreview.net/forum?id=XPZIaotutsD.
[19] H. Liu, D. Tam, M. Muqeeth, J. Mohta, T. Huang, M. Bansal, C. Raffel, Few-Shot Parameter-Efficient Fine-Tuning is Better and Cheaper than In-Context Learning, arXiv (2022). doi:10.48550/arXiv.2205.05638.
[20] N. Reimers, I. Gurevych, Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Hong Kong, China, 2019, pp. 3982–3992. doi:10.18653/v1/D19-1410.
[21] J. Robert, J. Michael, N. Steven, H. Geoffrey, Adaptive Mixtures of Local Experts, Neural Computation 3 (1991) 79–87. doi:10.1162/neco.1991.3.1.79.
[22] S. Mu, S. Lin, A Comprehensive Survey of Mixture-of-Experts: Algorithms, Theory, and Applications, 2025. URL: https://arxiv.org/abs/2503.07137. arXiv:2503.07137.
[23] W. Fedus, B. Zoph, N. Shazeer, Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity, Journal of Machine Learning Research 23 (2022) 1–39. URL: http://jmlr.org/papers/v23/21-0998.html.
[24] A. Q. Jiang, A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bamford, D. S. Chaplot, D. de las Casas, E. B. Hanna, F. Bressand, G. Lengyel, G. Bour, G. Lample, L. R. Lavaud, L. Saulnier, M.-A. Lachaux, P. Stock, S. Subramanian, S. Yang, S. Antoniak, T. Le Scao, T. Gervet, T. Lavril, T. Wang, T. Lacroix, W. El Sayed, Mixtral of Experts, 2024. doi:10.48550/arXiv.2401.04088. arXiv:2401.04088.
[25] D. Dai, C. Deng, C. Zhao, R. X. Xu, H. Gao, D. Chen, J. Li, W. Zeng, X. Yu, Y. Wu, Z. Xie, Y. K. Li, P. Huang, F. Luo, C. Ruan, Z. Sui, W. Liang, DeepSeekMoE: Towards Ultimate Expert Specialization in Mixture-of-Experts Language Models, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 1280–1297. doi:10.18653/v1/2024.acl-long.70.
[26] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, A. Müller, J. Nothman, G. Louppe, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, É. Duchesnay, Scikit-learn: Machine Learning in Python, 2018. URL: https://arxiv.org/abs/1201.0490. arXiv:1201.0490.
[27] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, RoBERTa: A Robustly Optimized BERT Pretraining Approach, in: Proceedings of the 20th Chinese National Conference on Computational Linguistics, Chinese Information Processing Society of China, Huhhot, China, 2021, pp. 1218–1227. URL: https://aclanthology.org/2021.ccl-1.108/.
[28] T. Akiba, S. Sano, T. Yanase, T. Ohta, M. Koyama, Optuna: A Next-generation Hyperparameter Optimization Framework, 2019. doi:10.48550/arXiv.1907.10902. arXiv:1907.10902.
[29] J. Liu, Contributors, Instructor: A library for structured outputs from large language models, 2024. URL: https://github.com/instructor-ai/instructor.
[30] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, C. Zheng, D. Liu, F. Zhou, F. Huang, F. Hu, H. Ge, H. Wei, H. Lin, J. Tang, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Zhou, J. Lin, K. Dang, K. Bao, K. Yang, L. Yu, L. Deng, M. Li, M. Xue, M. Li, P. Zhang, P. Wang, Q. Zhu, R. Men, R. Gao, S. Liu, S. Luo, T. Li, T. Tang, W. Yin, X. Ren, X. Wang, X. Zhang, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Zhang, Y. Wan, Y. Liu, Z. Wang, Z. Cui, Z. Zhang, Z. Zhou, Z. Qiu, Qwen3 Technical Report, 2025. doi:10.48550/arXiv.2505.09388. arXiv:2505.09388.