# StereoBusters at GSI:detect: LLM-Based Detection and Human Qualitative Analysis of Gender Stereotypes in Italian Short Texts

Salvatore Greco[1], Moreno La Quatra[2], Marta Marchiori Manerba[3], Ricardo Muñoz Sánchez[4] and Alessandra Teresa Cignarella[5]

[1]*Centre for Data Futures, King's College London, United Kingdom*
[2]*Speech Technologies and Machine Learning Lab, Kore University of Enna, Italy*
[3]*Computer Science Department, University of Turin, Italy*
[4]*Språkbanken Text, University of Gothenburg, Sweden*
[5]*Language and Translation Technology Team (LT3), Ghent University, Belgium*

**Abstract**

This paper presents the participation of the StereoBusters team in GSI:detect, a novel shared task at EVALITA 2026 focused on detecting and classifying gender stereotypes in short Italian texts. The competition consists of a main regression task that estimates the degree of gender-stereotypical content in a text, and a subtask that performs fine-grained classification into six gender stereotype categories. We describe our submitted systems in two settings: zero-shot and few-shot. We propose two LLM-based approaches: (1) using a single LLM to directly predict the degree of gender-stereotypical content and its category, and (2) using a panel of four LLMs, each predicting a binary gender-stereotypical label, and aggregating these predictions to obtain a continuous degree of stereotypical content, while using the best model in the panel to predict the category. We evaluate five prompting strategies: direct, guideline-based, Chain-of-Thought, Italian-language, and annotator-perspective prompting.

Our results show that: (1) while the approach aggregating a panel of LLMs performs better on the development data than a single continuous LLM approach, it generalizes less effectively on the test data; (2) prompts tend to benefit from the inclusion of annotation guidelines; and (3) even open-weight LLMs with fewer parameters can outperform larger models across both tasks and proposed approaches. In the official leaderboard, our best system ranked 2nd in the zero-shot setting of the main task (nMSE $= 0.626$), and 2nd in the few-shot setting of the main task (nMSE $= 0.645$). In the subtask, our best system ranked 1st in the zero-shot setting (F1 micro $= 0.646$), and 3rd in the few-shot setting (F1 micro $= 0.669$). Additionally, following the annotation guidelines, we qualitatively analyse the 200 development texts and provide additional annotations.

**Keywords**

Stereotype Detection, Gender Stereotypes, Italian, Prompting, Guidelines, LLMs, NLP

***Trigger warning:*** *This paper contains examples of stereotypical and potentially triggering content.*

## 1. Introduction

The exponential growth of social media has transformed online communication, enabling users to share opinions on virtually any topic at any time. However, this surge has also amplified the spread of harmful content, including hateful comments and prejudiced discourse. Such phenomena often stem from deeply ingrained stereotypes, which perpetuate discrimination and reinforce structural inequalities.

Natural Language Processing (NLP) can play a crucial role in addressing these challenges by providing tools to analyse linguistic manifestations of bias and automate their detection at scale. Within the Italian NLP community, several initiatives such as HaSpeeDe [1, 2], AMI [3, 4], HODI [5], and CALAMITA [6] have focused on hate speech detection and related phenomena, laying the groundwork for more nuanced investigations. Some works specifically focused on racial and ethnic stereotypes in Italian [7, 8, 9, 10]. Finally, other studies have explored NLP approaches for promoting gender-inclusive language, focusing

*EVALITA 2026: 9th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Feb 26 – 27, Bari, IT*
salvatore.greco@kcl.ac.uk (S. Greco); moreno.laquatra@unikore.it (M. La Quatra); marta.marchiorimanerba@unito.it (M. Marchiori Manerba); ricardo.munoz.sanchez@gu.se (R. Muñoz Sánchez); alessandrateresa.cignarella@ugent.be (A. T. Cignarella)
0000-0001-7239-9602 (S. Greco); 0000-0001-8838-064X (M. La Quatra); 0000-0003-2251-1824 (M. Marchiori Manerba); 0000-0002-9902-2925 (R. Muñoz Sánchez); 0000-0002-4409-6679 (A. T. Cignarella)

on rewriting or translating texts in a more gender-neutral manner [11, 12, 13, 14, 15, 16]. These efforts contribute to mitigating harmful content and fostering safer, more inclusive online spaces.

Despite these advances, research specifically targeting gender stereotypes in Italian texts remains scarce, compared to prior work on English and other high-resource languages. Furthermore, existing approaches often assume a binary conception of gender and lack robust evaluation baselines [17, 18]. The GSI:DETECT shared task [19], introduced at EVALITA 2026 [20], represents the first initiative in this domain for Italian. Its goal is to promote research on identifying stereotypical representations of gender across diverse short-text genres. To this end, the organizers released a manually curated dataset comprising 1,010 texts from social media and websites, annotated by four expert annotators to capture both aggregated and individual perspectives. The task is articulated into two parts:

- **Main task:** a regression task to determine the degree of gender-stereotypical content in a text.
- **Subtask:** a fine-grained classification task requiring systems to predict one of six stereotype categories.

In this paper, we present the systems developed by the **StereoBusters team** for the two tasks of GSI:DETECT. We explore the first two of the three experimental settings: *zero-shot*, *few-shot* and *fine-tuning*. We propose and evaluate two approaches leveraging large language models (LLMs) and diverse prompting strategies, including guideline integration and simulated annotator panels. We also performed an additional annotation experiment and a qualitative analysis (discussed in Appendix C).[1]

## 2. Task Description and Dataset

**Task Description** The GSI:DETECT task comprises (1) a main task to quantify the degree of Gender Stereotype (*GS value*) in a text; and (2) a subtask to classify the Gender Stereotype category (*GS category*).

Detecting stereotypes, however, is an inherently complex and subjective task, and the perception of stereotypes varies significantly across individuals [21]. For this reason, the GSI:DETECT task follows the perspectivist approach [22, 23]. Rather than relying on a single label obtained through majority voting, the dataset preserves non-aggregated annotations and provides a score designed to capture this variation. As a result, the *GS value* is defined as a single numerical value between 0 and 1 that reflects the fraction of annotators who labeled a text as containing stereotypes.

The *GS category* identifies the stereotype type in each text. Annotators assigned one category per text; if multiple applied, they chose the first occurring one. Categories are the following: **(1) Role stereotypes** encompass sociocultural expectations of the roles that people are expected to play regarding family or society; **(2) Personality stereotypes** attribute emotional, behavioral, or character traits to people based on their (perceived) gender; **(3) Competence stereotypes** are generalized judgments about professional, intellectual, or physical skills and abilities regardless of gender; **(4) Physical stereotypes** reflect expectations about the physical aspect of someone based on their (perceived) gender; **(5) Sexual stereotypes** are related to desire, sexuality, and sexual behavior. Insults with sexual connotations are also included in this category, regardless of whether they are direct or indirect; **(6) Relational stereotypes** refer to expectations of how people are expected to connect in interpersonal relationships, depending on their gender, including affective, romantic, and/or sexual relationships; **(7) No stereotype** means that none of the annotators considered that the text contains stereotypes.

**Overview of the GSI:detect Dataset** The dataset contains 1,010 short Italian texts: 200 in the development set and 810 in the test set. Texts were collected from social media and information websites to cover both informal and formal registers (a more detailed data analysis is discussed in Appendix A).

A total of four annotators sensitive to gender-related issues were involved in the annotation process. Each annotator labeled all texts by assigning (1) a binary yes/no gender stereotype label and (2) at most one stereotype category. Annotators were instructed to select the appropriate stereotype category when a text was judged to contain stereotypes, or to choose the "no" category when no stereotypes were present. If multiple stereotype categories appeared within the same text, annotators selected the category that occurred first in the text. The final GS value is computed by converting the four binary gender stereotype labels into numerical values (1 for "yes" 0 for "no") and taking their mean. The final GS category is determined via majority voting among annotators' category labels.

[1]The code repository is available at https://github.com/grecosalvatore/StereoBusters-GSI-Detect-Evalita2026.

# 3. Methodology

## 3.1. Gender Stereotypes Detection Approaches

Given the relatively limited size of the development data, our approach relies on zero-shot evaluation and in-context learning, rather than direct parameter tuning. However, since the dataset is annotated not only with an aggregate value but also with binary stereotype labels provided by individual annotators, we propose two alternative prediction targets for the *GS value* (main task):

**Approach 1** (CONTINUOUS): A single LLM is prompted to directly predict the aggregated GS value—the continuous gender stereotype score in [0, 1]. The LLM is expected to capture the degree of stereotypicity by implicitly modeling potential disagreement and subjectivity across multiple diverse viewpoints.

**Approach 2** (BINARY): Multiple LLMs are prompted to predict a binary gender stereotype label—whether the text contains a gender stereotype (0 = no, 1 = yes). The GS value is obtained by aggregating the binary predictions of multiple models. Specifically, we construct a panel of four models to mimic this process. Each model in the panel corresponds to a specific model-prompt combination selected to best replicate the labeling behavior of an individual annotator.

In both approaches, the *GS category* is predicted jointly with the main target when a stereotype is present. Under the CONTINUOUS approach, a single LLM is instructed to directly predict the GS value and, when the GS value is $\geq 0.25$, to also assign a stereotype category. In contrast, under the BINARY approach, each LLM predicts a binary gender-stereotype label and, if a stereotype is detected, additionally assigns its category. For the BINARY setting, the final GS value is obtained by aggregating the four binary predictions from the LLM panel, while the final category is determined either via majority vote or by designating one LLM as an oracle (i.e., super judge).

## 3.2. Prompting Strategies

For both approaches, we design and evaluate the following prompting strategies, reported in Figure 2 in Appendix B for the CONTINUOUS approach and in Figure 3 in Appendix B for the BINARY approach:

1. **Direct Prompt**: The LLM is directly instructed to assess whether the input text contains gender stereotypes and to output the corresponding prediction, with minimal additional guidance. Depending on the approach, the output is either a continuous GS value (Figure 2a) or a binary gender-stereotype label (Figure 3a).
2. **Chain-of-Thought (CoT) Prompt**: The LLM is instructed to reason step by step using a Chain-of-Thought strategy [24] before producing the final prediction (Figure 2b and Figure 3b). The underlying idea is that explicitly modeling intermediate reasoning steps may help the model better identify implicit gender stereotypes and arrive at more accurate predictions.
3. **Guidelines-Informed Prompt**: The LLM is provided with additional task-specific information, including detailed category definitions and short illustrative examples drawn directly from the annotation guidelines (Figure 2c and Figure 3c). The underlying idea is that explicitly incorporating the original annotation criteria may better align the model's predictions with human judgments.
4. **Italian Prompt**: The LLM is instructed to assess whether the input text contains gender stereotypes and to output the corresponding prediction, with minimal additional guidance. However, in this case, the prompt is written in Italian. The underlying idea is that, since the task involves detecting gender stereotypes in Italian texts, using the target language may better align the model's reasoning with linguistic and cultural cues specific to Italian (Figure 2d and Figure 3d).
5. **Perspectivist Prompt**: The LLM is instructed to emulate a panel of four annotators with differing sensitivities and perspectives. One annotator is highly sensitive to subtle stereotypes, two are moderately sensitive, and one identifies only explicit stereotypes. Note that this prompt is implemented only in the continuous setting (Figure 2e), as this approach mirrors the aggregation of the four binary models used to construct the panel for the binary approach.

All prompts predict the GS value or the binary gender stereotype label together with the GS category. Models are instructed to assign the "no" category when the GS value or binary stereotype label is 0.

### 3.3. Zero-shot and Few-shot Settings

All prompts are evaluated in both *zero-shot* and *few-shot* learning. The prompt structure is identical across the two settings; however, in few-shot, five illustrative examples are introduced in each prompt to guide the model's prediction, as shown in the red text in Figure 2f in Appendix B.

Given an input text, the five examples for the few-shot setting are identified by following these steps: (1) We split the annotated samples in the development set into two subsets: samples *with context* and samples *without context*; (2) We compute semantic embedding of the input text using the PARAPHRASE-MULTILINGUAL-MPNET-BASE-V2 model; (3) We retrieve the 5 most similar examples from the same subset (with or without context) based on cosine similarity between embeddings; and (4) The selected examples are formatted with their *text*, *GS value* for the CONTINUOUS approach or *binary gender stereotype label* for the BINARY approach (specific to one annotator), and *category*, and included in the system prompt.

The main differences between the few-shot implementations of the CONTINUOUS and BINARY approaches are as follows: (1) the ground-truth label provided in the examples differs, as the continuous version predicts the GS value, whereas the binary version predicts a binary gender stereotype label; and (2) the continuous approach needs to be executed only once per input text in the few-shot setting, since it produces a single aggregated score, while the binary approach must be run separately to emulate each specific annotator, as each sample has a binary stereotype label assigned by each annotator.

## 4. Results on the Development Data

In this section, we present the experimental results obtained on the development data. We first discuss the preliminary experiments conducted to identify the panels for the BINARY approach (Section 4.1), followed by the results for GS value prediction (Section 4.2) and GS category classification (Section 4.3).

**Models** We evaluated non-commercial, open-weight LLMs from different families and sizes: Llama 3 [25, 26] (8B, 70B), Qwen [27] (8B, 32B), LLaMAntino [28, 29] (8B), and Gemma 3 [30] (4B, 12B, 27B).

### 4.1. Preliminary Experiments for Panels Compositions

This experiment aims to identify the most effective combinations of models and prompts to mimic each annotator in the binary gender stereotype labeling task, following the proposed BINARY approach.

**Evaluation metrics** We compute macro-averaged F1, Precision, and Recall for the prediction of the binary gender stereotype label, evaluated separately for each annotator using their corresponding binary ground-truth labels (0/1 for absence/presence of stereotype). In addition, we assess the agreement between the LLM predictions and each annotator by computing the Cohen's Kappa ($\kappa$) coefficient [31].

**Results** Table 1 reports the macro F1, Precision (P), Recall (R), and Cohen's $\kappa$ scores for the binary gender stereotype prediction task. Results are presented separately for each annotator (A1—A4). In the *zero-shot* setting, all model–prompt combinations are evaluated. In contrast, in *few-shot*, only the best-performing model–prompt combination for each model within the selected panels is tested.[2]

**Zero-shot Results** In the *zero-shot* setting (top section of Table 1), all model–prompt combinations achieve high F1 scores and annotator agreement, except for the CoT prompt, which consistently yields lower performance. The best model–prompt combinations based on Cohen's $\kappa$ are LLAMA-3.3-70B-INSTRUCT with the Italian prompt for A1, LLAMA-3.1-8B-INSTRUCT with the Direct prompt for A2, LLAMANTINO-3-ANITA-8B with the Guidelines prompt for A3, and GEMMA-3-27B-IT with the Direct prompt for A4. When optimizing for F1, the best combinations are LLAMA-3.3-70B-INSTRUCT with the Direct prompt for A1, LLAMA-3.3-70B-INSTRUCT with the Italian prompt for A2, GEMMA-3-12B-IT with the Direct prompt for A3, and GEMMA-3-27B-IT with the Direct prompt for A4. Interestingly, even smaller models with 8B or 12B parameters showed higher agreement with the annotators.

[2] As discussed above, in the binary approach, the ground-truth labels are annotator-specific. Consequently, each model–prompt combination must be evaluated separately for each annotator in *few-shot*, resulting in a very large experimental space.

**Table 1**
**Results on the Development Data for Binary Gender Stereotype Detection (Binary Approach).** Performance comparison of models in modeling individual annotators, used to identify the panels for the Binary approach. Results are reported per annotator (A1–A4) and include F1 score, Precision (P), Recall (R), and Cohen's Kappa ($\kappa$), evaluated separately under *zero-shot* and *few-shot* settings. In *few-shot*, only the best-performing models from the *zero-shot* experiments were evaluated. Best-performing *zero-shot* results for each metric are in bold and underscored. Model-prompt combinations selected for **Panel 1** (maximizing Cohen's $\kappa$) and **Panel 2** (maximizing F1) are indicated with colored labels **[P1Ax]** and **[P2Ax]**, where x denotes the annotator being modeled, with corresponding scores highlighted in gray.

| | | Per-Annotator Prediction | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | A1 | | | | A2 | | | | A3 | | | | A4 | | | |
| **Model** | **Prompt** | **F1** | **P** | **R** | **$\kappa$** | **F1** | **P** | **R** | **$\kappa$** | **F1** | **P** | **R** | **$\kappa$** | **F1** | **P** | **R** | **$\kappa$** |
| *Zero-Shot* | | | | | | | | | | | | | | | | | |
| Llama-3.1-8B-Instruct | Direct **[P1A2]** | .73 | .74 | .74 | .47 | .74 | **.76** | **.76** | **.50** | .75 | .76 | **.76** | .49 | .76 | .75 | .76 | .51 |
| | CoT | .55 | .71 | .61 | .21 | .49 | .67 | .59 | .15 | .60 | .74 | .63 | .27 | .62 | .76 | .64 | .31 |
| | Guidelines | .69 | .72 | .70 | .40 | .70 | .75 | .74 | .44 | .73 | .74 | .72 | .46 | .76 | .77 | .75 | .52 |
| | Italian | .69 | .70 | .70 | .40 | .69 | .72 | .72 | .41 | .74 | .75 | .74 | .49 | .74 | .75 | .74 | .48 |
| Llama-3.3-70B-Instruct | Direct **[P2A1]** | **.75** | .75 | **.75** | .50 | .74 | .75 | .76 | .49 | .72 | .72 | .72 | .44 | .74 | .74 | .74 | .48 |
| | CoT | .72 | **.77** | .73 | .46 | .66 | .73 | .71 | .37 | .72 | .75 | .72 | .45 | .72 | .75 | .72 | .45 |
| | Guidelines | .70 | .74 | .72 | .43 | .69 | .74 | .72 | .41 | .72 | .74 | .72 | .45 | .75 | .77 | .75 | .51 |
| | Italian **[P1A1]** - **[P2A2]** | **.75** | .75 | **.75** | **.51** | **.75** | .75 | **.76** | **.50** | .70 | .71 | .71 | .41 | .73 | .74 | .74 | .47 |
| Gemma-3-4B-it | Direct | .65 | .69 | .66 | .32 | .62 | .68 | .66 | .30 | .67 | .69 | .67 | .35 | .69 | .71 | .69 | .39 |
| | CoT | .34 | .57 | .50 | .01 | .31 | .54 | .50 | .00 | .38 | .62 | .51 | .02 | .37 | .45 | .50 | -.01 |
| | Guidelines | .67 | .69 | .68 | .36 | .65 | .69 | .68 | .34 | .69 | .70 | .69 | .38 | .73 | .74 | .73 | .46 |
| | Italian | .58 | .72 | .63 | .26 | .55 | .72 | .63 | .23 | .65 | **.78** | .67 | .36 | .64 | .75 | .65 | .33 |
| Gemma-3-12B-it | Direct **[P2A3]** | .72 | .73 | .73 | .45 | .73 | .74 | .74 | .46 | **.76** | .76 | **.76** | .51 | .74 | .73 | .74 | .47 |
| | CoT | .67 | .72 | .69 | .37 | .62 | .68 | .66 | .29 | .66 | .69 | .66 | .34 | .67 | .69 | .67 | .35 |
| | Guidelines | .70 | .73 | .71 | .42 | .68 | .74 | .72 | .40 | .75 | .77 | .74 | .50 | .77 | .78 | .76 | .54 |
| | Italian | .71 | .74 | .72 | .44 | .68 | .73 | .72 | .40 | .74 | .75 | .74 | .48 | .76 | .77 | .75 | .52 |
| Gemma-3-27B-it | Direct **[P1A4]** - **[P2A4]** | .69 | .73 | .71 | .41 | .68 | .73 | .71 | .39 | .73 | .75 | .73 | .47 | **.78** | **.80** | **.78** | **.57** |
| | CoT | .63 | .71 | .66 | .32 | .60 | .69 | .65 | .27 | .70 | .76 | .70 | .42 | .70 | .75 | .70 | .42 |
| | Guidelines | .67 | .72 | .69 | .37 | .66 | .73 | .70 | .36 | .72 | .75 | .71 | .44 | .77 | **.80** | .76 | .54 |
| | Italian | .67 | .71 | .68 | .36 | .65 | .71 | .69 | .35 | .73 | .76 | .73 | .48 | .77 | .79 | .76 | .54 |
| Qwen3-8B | Direct | .65 | .71 | .67 | .34 | .69 | .73 | .69 | .40 | .62 | .70 | .66 | .30 | .62 | .71 | .67 | .31 |
| | CoT | .69 | .70 | .69 | .38 | .69 | .69 | .69 | .38 | .62 | .65 | .64 | .27 | .61 | .64 | .63 | .25 |
| | Guidelines | .51 | .57 | .54 | .09 | .58 | .64 | .59 | .21 | .54 | .65 | .60 | .18 | .53 | .64 | .60 | .17 |
| | Italian | .67 | .70 | .67 | .35 | .72 | .74 | .71 | .44 | .64 | .69 | .67 | .32 | .66 | .72 | .70 | .36 |
| Qwen3-32B | Direct | .71 | .72 | .71 | .43 | .74 | .74 | .74 | .47 | .73 | .75 | .75 | .47 | .72 | .74 | .74 | .45 |
| | CoT | .72 | .74 | .73 | .46 | .69 | .73 | .72 | .41 | .70 | .71 | .70 | .40 | .74 | .75 | .74 | .48 |
| | Guidelines | .71 | .72 | .71 | .43 | .73 | .73 | .73 | .46 | .70 | .71 | .71 | .41 | .72 | .74 | .74 | .45 |
| | Italian | .69 | .70 | .69 | .38 | .71 | .72 | .71 | .42 | .70 | .73 | .72 | .42 | .67 | .70 | .69 | .36 |
| LLaMAntino-3-ANITA-8B | Direct | .71 | .71 | .71 | .42 | .71 | .72 | .72 | .43 | .75 | .75 | **.76** | .51 | .74 | .74 | .75 | .49 |
| | CoT | .68 | .69 | .68 | .37 | .68 | .71 | .71 | .38 | .70 | .70 | .69 | .39 | .73 | .73 | .72 | .45 |
| | Guidelines **[P1A3]** | .71 | .71 | .71 | .42 | .71 | .72 | .73 | .43 | **.76** | .76 | **.76** | **.52** | .75 | .75 | .75 | .50 |
| | Italian | .68 | .70 | .69 | .38 | .67 | .71 | .70 | .38 | .74 | .75 | .74 | .48 | .77 | .78 | .77 | .54 |
| *Few-Shot* | | | | | | | | | | | | | | | | | |
| Llama-3.1-8B-Instruct | Direct **[P1A2]** | - | - | - | - | .61 | .73 | .68 | .31 | - | - | - | - | - | - | - | - |
| Llama-3.3-70B-Instruct | Direct **[P2A1]** | .73 | .75 | .74 | .47 | - | - | - | - | - | - | - | - | - | - | - | - |
| Llama-3.3-70B-Instruct | Italian **[P1A1]** - **[P2A2]** | .74 | .74 | .74 | .49 | .77 | .78 | .77 | .54 | - | - | - | - | - | - | - | - |
| Gemma-3-12B-it | Direct **[P2A3]** | - | - | - | - | - | - | - | - | .76 | .76 | .75 | .51 | - | - | - | - |
| Gemma-3-27B-it | Direct **[P1A4]** - **[P2A4]** | - | - | - | - | - | - | - | - | - | - | - | - | .73 | .74 | .73 | .47 |
| LLaMAntino-3-ANITA-8B | Guidelines **[P1A3]** | - | - | - | - | - | - | - | - | .73 | .75 | .73 | .47 | - | - | - | - |

**Few-shot Results** In the *few-shot* setting (bottom section of Table 1), surprisingly, most models exhibit a decrease in the F1 score and lower Cohen's $\kappa$ agreement with the corresponding annotators. The only exception is Llama-3.3-70B-Instruct (Italian), which shows improved performance for Annotator 2. The decrease in performance in the few-shot setting may be attributed to the binary nature of the gender stereotype detection task in the Binary approach, where few examples can overly bias the model's decision boundary, particularly for smaller models.

**Summary of Results** The four best-performing model–prompt combinations to mimic individual annotators in the development data, were grouped into two panels (Table 2b). These panels were then used to compute the GS values in our development experiments and for the final submissions.

## 4.2. GS Value Prediction Results on Development Data (Main task)

This experiment aims to evaluate *GS value* prediction (main task) on development data across all models using the Continuous approach, and the two panels identified in Section 4.1 for the Binary approach.

**Evaluation metrics** We evaluate model performance using the complements of the Mean Squared Error (MSE) and the Mean Absolute Error (MAE), which quantify the discrepancy between the predicted and ground-truth GS values. Since GS values are continuous and lie in the interval $[0, 1]$, we report the metrics as $(1 - \text{MSE})$ and $(1 - \text{MAE})$, with higher values indicating better predictive performance.

**Results** Table 2a reports the complements of the MSE and MAE in *zero-shot* and *few-shot* settings for

all model–prompt combinations for the CONTINUOUS approach and for the two BINARY panels.

**Zero-shot Results** In the *zero-shot* setting, all model–prompt combinations perform well (1 - MSE ranging from 0.727 to 0.898). The best-performing systems are, for the CONTINUOUS approach: (1) LLAMA-3.3-70B-INSTRUCT with guidelines-informed prompt and (2) GEMMA-3-12B-IT with Italian prompt; and, for the BINARY approach, (3) Panel 1 based on Cohen's $\kappa$. All three systems achieve the highest score of 0.898. Other model–prompt combinations in the CONTINUOUS setting achieve very similar performance, such as LLAMA-3.3-70B-INSTRUCT with the Italian prompt and GEMMA-3-12B-IT with the direct prompt, with scores of 0.897 and 0.895. When considering the complement of the MAE, both panels in the BINARY approach outperform all CONTINUOUS systems, achieving scores of 0.800 and 0.795. This likely reflects the noise-reducing effect of ensembling four LLMs, which lowers average error.

Overall, Panel 1 emerges as the best-performing system on the development set for the main task. This panel achieves (1 − MSE) = 0.898 and (1 − MAE) = 0.800, indicating that model predictions differ from the gold standard by approximately 0.20 units on average. The MSE of 0.10 further suggests that large errors are infrequent, as squared deviations remain low on average.

**Few-shot Results** In the *few-shot* setting, most models and all panels achieve similar or slightly lower performance in predicting the GS values compared to the *zero-shot* setting. In contrast, some models benefit from the inclusion of five examples. In particular, GEMMA-3-12B-IT and GEMMA-3-27B-IT for the CONTINUOUS approach show consistent improvements across prompts when evaluated in *few-shot*. Overall, the best-performing model–prompt combination in the *few-shot* setting is the CONTINUOUS GEMMA-3-27B-IT with the guidelines-informed prompt, achieving a complement of the MSE of 0.899. This configuration exhibits the best performance across *zero-shot* and *few-shot* according to the MSE.

**Summary of Results** These experiments on the development data show that both the CONTINUOUS and BINARY (panel-based) approaches can effectively predict the GS value. The three best-performing models for the CONTINUOUS approach: LLAMA-3.3-70B-INSTRUCT with guidelines-informed and Italian prompts, and GEMMA-3-12B-IT with the Italian prompt, together with the two panels for the BINARY approach (Table 2b), were selected for the five *zero-shot* runs in the main task on the official test data (see Section 5). Similarly, for the five *few-shot* runs, we selected the same two panels for the BINARY approach, along with the three best-performing models for the few-shot CONTINUOUS setting: GEMMA-3-27B-IT with a guidelines-informed prompt, and GEMMA-3-12B-IT with both the direct and Italian prompts.

### 4.3. GS Category Classification Results on Development Data (Subtask)

This experiment aims to evaluate *GS category* classification (subtask) on development data across all model-prompt combinations using the CONTINUOUS and the BINARY approaches.

**Evaluation metrics** We compute the macro precision, recall, and F1 score between the predicted and gold GS categories. The metrics are averaged across seven categories: the six stereotype categories: *role*, *personality*, *competence*, *physical*, *sexual*, and *relation*; and additionally include the *no* category.

**Results** Table 2c reports the macro F1, precision (P), and recall (R) for all model–prompt combinations for the CONTINUOUS and BINARY approaches, in *zero-shot* and *few-shot*. For the BINARY approach, unlike the computation of the GS value (which requires aggregating four binary predictions), each LLM directly predicts a stereotype category. Consequently, in this experiment, we evaluate the ability of each model–prompt pair to predict the category independently, including for the BINARY approach.[3]

**Zero-shot Results** In the *zero-shot* setting, models prompted for the BINARY approach tend to outperform those for the CONTINUOUS ones in predicting the GS category. This is likely because LLMs are conditioned to predict the category only when a stereotype is detected, and the continuous nature of the CONTINUOUS approach may introduce more noise, which can propagate to the category classification. We also observed results that contrasted with those for GS value (or binary label) prediction. In particular, the CoT prompt often outperforms other prompts in category classification, whereas it was the worst-performing prompt for the main task. This suggests that LLMs benefit more from reasoning through

[3]We also evaluated majority voting within panels for category prediction; however, this strategy resulted in lower performance.

chains of thought when performing category classification than when predicting binary stereotype labels or GS values. Surprisingly, the guidelines-informed prompt is among the best prompts for the CONTINUOUS approach but the worst for the BINARY approach in GS category classification.

The best-performing models are as follows: for the BINARY approach, GEMMA-3-27B with direct prompt (F1 = 0.56), GEMMA-3-12B with CoT prompt (F1 = 0.56), and LLAMA-3.3-70B-INSTRUCT with CoT prompt (F1 = 0.53); for the CONTINUOUS approach, QWEN3-8B with CoT prompt (F1 = 0.52), GEMMA-3-27B with guidelines-informed prompt (F1 = 0.51), and GEMMA-3-27B with Italian prompt (F1 = 0.50).

**Few-shot Results** Most model–prompt configurations within the CONTINUOUS approach show improved GS category classification performance in the *few-shot* setting, with the exception of LLAMA-3.3-70B-INSTRUCT, which exhibits a performance decrease in few-shot not only for category classification but also for the main task. We do not report quantitative results for the BINARY approach, as each model–prompt configuration yields different outcomes depending on the annotator. Nevertheless, qualitative inspection suggests that few-shot prompting does not lead to substantial improvements.

**Summary of Results** Although some model–prompt combinations in the CONTINUOUS approach performed better on GS *category* prediction than those selected for the main task (GS value prediction), we prioritized the main task. Therefore, we submitted GS category predictions from the same model–prompt combinations used for GS value prediction (Section 2a). Specifically, for the *zero-shot* setting, we used LLAMA-3.3-70B-INSTRUCT with guidelines-informed and Italian prompts (F1 = 0.49), and GEMMA-3-12B-IT with the Italian prompt (F1 = 0.45). For the *few-shot* setting, we selected GEMMA-3-27B-IT with a guidelines-informed prompt (F1 = 0.57), and GEMMA-3-12B-IT with both the direct (F1 = 0.46) and Italian prompts (F1 = 0.46). In contrast, for the BINARY approach, we selected the GS category predicted by GEMMA-3-27B with the direct prompt (F1 = 0.56), as this model–prompt combination achieved the highest overall performance in GS category prediction in the zero-shot setting and is present in both experimental panels. As a result, this model–prompt combination serves as the oracle for GS category prediction in both panels. Our final submitted systems are summarized in Table 3.

# 5. Results on the Test Data and Official Rankings

## 5.1. GS Value Prediction Results on Test Data (Main task)

**Evaluation metrics** We evaluate the performance in predicting the GS value using the normalized Mean Squared Error (nMSE), computed as the MSE normalized by the variance of the gold GS values: $\text{nMSE} = \text{MSE}/\operatorname{Var}(y)$, where $y$ are the ground-truth GS values. Systems are ranked using the transformed score $1/(1 + \text{nMSE})$ so that higher values correspond to better performance.

**Results** Table 4 reports the official rankings for the main task for *zero-shot* and *few-shot* settings.

**Zero-shot Results** In the zero-shot setting, our best submission ranked 2nd overall with a score of 0.626. This system used GEMMA-3-12B with the guidelines-informed prompt under the CONTINUOUS approach. Our second-best system, LLAMA-3.3-70B-INSTRUCT with the Italian prompt, ranked 3rd with a score of 0.622. The top-performing system from team DIAG-Sapienza achieved a score of 0.700 using GPT-5. Notably, all three of our CONTINUOUS submissions (ranks 2, 3, and 6) outperformed the two BINARY panel submissions (ranks 11 and 12). This contrasts with our development set findings, where the panels achieved competitive performance. The gap between the CONTINUOUS and BINARY approaches on the test set may indicate that the panel compositions, optimized on the development data, did not generalize well to the test data distribution.

**Few-shot Results** In the few-shot setting, our best submission again ranked 2nd (score of 0.645). This system used GEMMA-3-27B with the guidelines-informed prompt. The GEMMA-3-12B model with the direct prompt ranked 3rd (0.629), and with the Italian prompt ranked 4th (0.612). All of these systems follow the CONTINUOUS approach. Compared to the zero-shot setting, our few-shot submissions showed some but modest improvements, with the best score increasing from 0.626 to 0.645. The BINARY Panel 2 (F1-based) achieved a score of 0.586 (rank 6), while Panel 1 (Cohen's $\kappa$-based) dropped to 0.549 (rank 13). This suggests that the few-shot examples may have introduced noise for certain panel configurations.

**Table 2**

**Results on the development data.** Results are reported for the GS value prediction (a) and GS category classification (c) tasks. For GS value prediction (a), the complement of nMSE and nMAE is reported for all models under the CONTINUOUS approach and for the two panels of the BINARY approach, summarized in (c). For GS category classification (c), the F1 score, Precision (P), and Recall (R) are reported for all model–prompt combinations under both approaches. Model–prompt combinations are grouped by approach, with *Zero-Shot* and *Few-Shot* results presented side by side. Bold values indicate the top-performing configurations. The scores of the five submitted configurations, separately for zero-shot and few-shot settings, are underlined and highlighted.

(a) GS Value Prediction (Main Task)

| Model | Prompt | *Zero-Shot* | | *Few-Shot* | |
|---|---|---|---|---|---|
| | | (1-MSE) ↑ | (1-MAE) ↑ | (1-MSE) ↑ | (1-MAE) ↑ |
| *Continuous Approach* | | | | | |
| Llama-3.1-8B-Instruct | Direct | .849 | .713 | .856 | .725 |
| | CoT | .737 | .618 | .778 | .651 |
| | Guidelines | .832 | .658 | .819 | .660 |
| | Italian | .830 | .687 | .791 | .655 |
| | Perspectivist | .826 | .671 | .826 | .690 |
| Llama-3.3-70B-Instruct | Direct | .888 | .743 | .573 | .493 |
| | CoT | .879 | .739 | .573 | .493 |
| | Guidelines | **.898** | .752 | .573 | .493 |
| | Italian | .897 | .764 | .573 | .493 |
| | Perspectivist | .867 | .741 | .573 | .493 |
| gemma-3-4b-it | Direct | .834 | .650 | .861 | .682 |
| | CoT | .769 | .618 | .824 | .648 |
| | Guidelines | .805 | .632 | .843 | .667 |
| | Italian | .852 | .676 | .815 | .645 |
| | Perspectivist | .727 | .604 | .788 | .649 |
| gemma-3-12b-it | Direct | .895 | .745 | .895 | .751 |
| | CoT | .841 | .686 | .881 | .730 |
| | Guidelines | .854 | .687 | .881 | .730 |
| | Italian | **.898** | .757 | .894 | .760 |
| | Perspectivist | .860 | .719 | .870 | .728 |
| gemma-3-27b-it | Direct | .891 | .745 | .891 | .757 |
| | CoT | .885 | .739 | .887 | .768 |
| | Guidelines | .869 | .715 | **.899** | .763 |
| | Italian | .885 | .731 | .886 | .751 |
| | Perspectivist | .831 | .688 | .867 | .714 |
| Qwen3-8B | Direct | .862 | .716 | .850 | .726 |
| | CoT | .797 | .696 | .802 | .701 |
| | Guidelines | .852 | .739 | .859 | .744 |
| | Italian | .848 | .721 | .816 | .690 |
| | Perspectivist | .796 | .675 | .821 | .690 |
| Qwen3-32B | Direct | .875 | .734 | .875 | .755 |
| | CoT | .810 | .700 | .830 | .729 |
| | Guidelines | .879 | .744 | .862 | .747 |
| | Italian | .886 | .748 | .853 | .727 |
| | Perspectivist | .869 | .738 | .869 | .736 |
| LLaMAntino-3 -ANITA-8B-Inst-DPO-ITA | Direct | .876 | .729 | .861 | .720 |
| | CoT | .820 | .692 | .829 | .703 |
| | Guidelines | .813 | .664 | .797 | .653 |
| | Italian | .871 | .720 | .723 | .603 |
| | Perspectivist | .855 | .743 | .835 | .705 |
| *Binary Approach* | | | | | |
| Panel 1 (Cohen's $\kappa$) | — | **.898** | **.800** | .882 | **.794** |
| Panel 2 (F1) | — | .883 | .795 | .868 | .760 |

(b) Panels composition

P1 - **Panel (Cohen's $\kappa$)**

| Annotator | Model | Prompt | k |
|---|---|---|---|
| A1 | Llama-3.3-70B-Instruct | Italian | 0.51 |
| A2 | Llama-3.1-8B-Instruct | Direct | 0.50 |
| A3 | LLaMAntino-3-ANITA-8B | Guidelines | 0.52 |
| A4 | gemma-3-27b-it | Direct | 0.57 |

P2 - **Panel (F1)**

| Annotator | Model | Prompt | F1 |
|---|---|---|---|
| A1 | Llama-3.3-70B-Instruct | Direct | 0.75 |
| A2 | Llama-3.3-70B-Instruct | Italian | 0.75 |
| A3 | gemma-3-12b-it | Direct | 0.76 |
| A4 | gemma-3-27b-it | Direct | 0.78 |

(c) GS Category Classification (Subtask)

| Model | Prompt | *Zero-Shot* | | | *Few-Shot* | | |
|---|---|---|---|---|---|---|---|
| | | F1 | P | R | F1 | P | R |
| *Continuous Approach* | | | | | | | |
| Llama-3.1-8B-Instruct | Direct | .25 | .44 | .28 | .33 | .55 | .34 |
| | CoT | .42 | .48 | .47 | .41 | .46 | .45 |
| | Guidelines | .37 | .44 | .40 | .40 | .42 | .46 |
| | Italian | .29 | .39 | .32 | .30 | .32 | .33 |
| | Perspectivist | .28 | .34 | .31 | .31 | .30 | .36 |
| Llama-3.3-70B-Instruct | Direct | .44 | .50 | .47 | .04 | .02 | .14 |
| | CoT | .48 | .50 | .54 | .04 | .02 | .14 |
| | Guidelines | .49 | .50 | .53 | .04 | .02 | .14 |
| | Italian | .49 | .51 | .52 | .04 | .02 | .14 |
| | Perspectivist | .39 | .48 | .45 | .04 | .02 | .14 |
| gemma-3-4B | Direct | .31 | .54 | .37 | .33 | .45 | .40 |
| | CoT | .33 | .43 | .41 | .33 | .40 | .40 |
| | Guidelines | .38 | .40 | .46 | .40 | .49 | .47 |
| | Italian | .33 | .44 | .41 | .29 | .38 | .38 |
| | Perspectivist | .18 | .41 | .23 | .34 | .35 | .37 |
| gemma-3-12B | Direct | .40 | .50 | .42 | .46 | .53 | .48 |
| | CoT | .48 | .53 | .54 | .52 | .54 | .57 |
| | Guidelines | .45 | .49 | .52 | .51 | .53 | .56 |
| | Italian | .45 | .55 | .47 | .43 | .49 | .46 |
| | Perspectivist | .39 | .48 | .44 | .44 | .45 | .45 |
| gemma-3-27B | Direct | .44 | .49 | .49 | .51 | .52 | .54 |
| | CoT | .47 | .50 | .53 | .57 | .59 | .60 |
| | Guidelines | .51 | .55 | **.60** | **.57** | **.57** | **.63** |
| | Italian | .50 | .52 | .55 | .53 | .56 | .56 |
| | Perspectivist | .41 | .53 | .51 | .46 | .46 | .48 |
| Qwen3-8B | Direct | .40 | .47 | .41 | .49 | .59 | .50 |
| | CoT | .52 | .54 | .54 | .55 | .58 | .57 |
| | Guidelines | .49 | .49 | .51 | .54 | .55 | .55 |
| | Italian | .46 | .47 | .47 | .46 | .51 | .47 |
| | Perspectivist | .39 | .45 | .41 | .39 | .41 | .43 |
| Qwen3-32B | Direct | .44 | .49 | .47 | .53 | .57 | .55 |
| | CoT | .47 | .50 | .51 | .54 | .55 | .57 |
| | Guidelines | .54 | .54 | .56 | .55 | .55 | .57 |
| | Italian | .49 | .53 | .52 | .48 | .50 | .51 |
| | Perspectivist | .42 | .49 | .46 | .48 | .53 | .51 |
| LLaMAntino-3 -ANITA-8B-Inst-DPO-ITA | Direct | .31 | .48 | .32 | .35 | .51 | .37 |
| | CoT | .39 | .44 | .40 | .40 | .44 | .42 |
| | Guidelines | .34 | .38 | .38 | .34 | .34 | .39 |
| | Italian | .31 | .49 | .33 | .29 | .32 | .34 |
| | Perspectivist | .27 | .41 | .30 | .36 | .42 | .38 |
| *Binary Approach* | | | | | | | |
| Llama-3.1-8B-Instruct | Direct | .35 | .55 | .35 | — | — | — |
| | CoT | .43 | .49 | .45 | — | — | — |
| | Guidelines | .13 | .11 | .22 | — | — | — |
| | Italian | .37 | .47 | .37 | — | — | — |
| Llama-3.3-70B-Instruct | Direct | .47 | .56 | .46 | — | — | — |
| | CoT | .53 | .55 | .55 | — | — | — |
| | Guidelines | .15 | .12 | .24 | — | — | — |
| | Italian | .45 | .62 | .43 | — | — | — |
| gemma-3-4B | Direct | .33 | .56 | .36 | — | — | — |
| | CoT | .42 | .50 | .51 | — | — | — |
| | Guidelines | .14 | .11 | .23 | — | — | — |
| | Italian | .35 | .61 | .40 | — | — | — |
| gemma-3-12B | Direct | .33 | .56 | .36 | — | — | — |
| | CoT | **.56** | .57 | .57 | — | — | — |
| | Guidelines | .14 | .12 | .23 | — | — | — |
| | Italian | .47 | **.63** | .49 | — | — | — |
| gemma-3-27B | Direct (**Oracle** [P1] - [P2]) | **.56** | **.59** | **.56** | — | — | — |
| | CoT | .55 | .56 | .56 | — | — | — |
| | Guidelines | .14 | .12 | .23 | — | — | — |
| | Italian | .48 | .52 | .50 | — | — | — |
| Qwen3-8B | Direct | .33 | .54 | .32 | — | — | — |
| | CoT | .41 | **.63** | .37 | — | — | — |
| | Guidelines | .11 | .22 | .17 | — | — | — |
| | Italian | .30 | .55 | .30 | — | — | — |
| Qwen3-32B | Direct | .37 | .50 | .36 | — | — | — |
| | CoT | .52 | .58 | .52 | — | — | — |
| | Guidelines | .20 | .34 | .25 | — | — | — |
| | Italian | .39 | .61 | .38 | — | — | — |
| LLaMAntino-3 -ANITA-8B-Inst-DPO-ITA | Direct | .32 | .52 | .34 | — | — | — |
| | CoT | .39 | .53 | .39 | — | — | — |
| | Guidelines | .15 | .11 | .24 | — | — | — |
| | Italian | .31 | .55 | .33 | — | — | — |

**Table 3**
**Summary of Submitted Systems.**

| Run | Model ID | Config ID | Setting | Approach | GS Value Prediction | GS Category Classification |
|---|---|---|---|---|---|---|
| *Zero-Shot Submissions* | | | | | | |
| 1 | Panel_4LLMS | K_Cohen | Zero-shot | Binary | Panel 1: Aggregation of 4 binary labels | gemma-3-27b-it (Direct) |
| 2 | Panel_4LLMS | F1_score | Zero-shot | Binary | Panel 2: Aggregation of 4 binary labels | gemma-3-27b-it (Direct) |
| 3 | Llama_3.3_70B | P3GUIDE | Zero-shot | Continuous | Llama-3.3-70B-Instruct (Guidelines) | Llama-3.3-70B-Instruct (Guidelines) |
| 4 | Llama_3.3_70B | P5ITA | Zero-shot | Continuous | Llama-3.3-70B-Instruct (Italian) | Llama-3.3-70B-Instruct (Italian) |
| 5 | Gemma_3_12B | P5ITA | Zero-shot | Continuous | gemma-3-12b-it (Italian) | gemma-3-12b-it (Italian) |
| *Few-Shot Submissions* | | | | | | |
| 1 | Panel_4LLMS | K_Cohen | Few-shot | Binary | Panel 1: Aggregation of 4 binary labels | gemma-3-27b-it (Direct) |
| 2 | Panel_4LLMS | F1_score | Few-shot | Binary | Panel 2: Aggregation of 4 binary labels | gemma-3-27b-it (Direct) |
| 3 | Gemma_3_27B | P3GUIDE | Few-shot | Continuous | gemma-3-27b-it (Guidelines) | gemma-3-27b-it (Guidelines) |
| 4 | Gemma_3_12B | P1DIR | Few-shot | Continuous | gemma-3-12b-it (Direct) | gemma-3-12b-it (Direct) |
| 5 | Gemma_3_12B | P5ITA | Few-shot | Continuous | gemma-3-12b-it (Italian) | gemma-3-12b-it (Italian) |

**Summary of Results** Our systems achieved strong performance in the main task, ranking best among open-weight models, and overall, 2nd in both zero-shot and few-shot settings. The CONTINUOUS approach outperformed the BINARY approach on the test data. The guidelines-informed prompt proved effective across both settings and model sizes. These results suggest directly predicting the aggregated GS value is more robust than modeling individual annotators when generalizing to unseen data.

### 5.2. GS Category Classification Results on Test Data (Subtask)

**Evaluation metrics** We evaluate performance on GS category classification using the micro F1 score as the primary metric, following the official evaluation guidelines. We also report the macro F1 score.[4]

**Results** Table 5 reports the official rankings for the subtask, separately for *zero-shot* and *few-shot*.

**Zero-shot Results** In zero-shot, our submissions achieved the top four positions in the leaderboard. The best-performing system was Panel 2 (F1-based) with GEMMA-3-27B as the category oracle, achieving a micro F1 of 0.646. Panel 1 (Cohen's $\kappa$-based) ranked 3rd with a micro F1 of 0.643. The BINARY panels outperformed the CONTINUOUS systems for category classification, despite showing the opposite trend for GS value prediction. This indicates that the oracle model (GEMMA-3-27B with the direct prompt) used for category prediction in the panels was highly effective. The CONTINUOUS submissions using GEMMA-3-12B and LLAMA-3.3-70B ranked 2nd and 4th, with F1 micro of 0.644 and 0.634.

**Few-shot Results** In the few-shot setting, our best submission ranked 3rd with a micro F1 of 0.669. This system used GEMMA-3-27B with the guidelines-informed prompt under the CONTINUOUS approach. The top two positions were held by teams Prisma (0.706) and FestaDonato (0.671) using Claude and Gemini models, respectively. Both BINARY panels ranked 5th and 6th with identical micro F1 scores of 0.662. The remaining CONTINUOUS submissions (GEMMA-3-12B with direct and Italian prompts) ranked 11th and 12th. This drop in performance for these configurations suggests that the few-shot examples were less beneficial for smaller models in the category classification task. Similarly to the main task, our models ranked first in both settings, considering only non-commercial, open-weight models.

**Summary of Results** Our systems achieved 1st place in the zero-shot and 3rd place in the few-shot settings for category classification. In the zero-shot setting, the BINARY panels performed best, likely due to the strong category prediction ability of GEMMA-3-27B (direct prompt), which was used as the sole category predictor for both panels. In the few-shot setting, larger models with the guidelines-informed prompt performed best. Overall, category classification benefited from explicit reasoning about stereotype types, as reflected in the strong performance of the guidelines-informed prompt.

## 6. Conclusion

In this paper, we described our contribution to the GSI:DETECT task on stereotype detection at EVALITA 2026. We have presented two different LLM-based approaches: (1) directly predicting the degree of stereotype (CONTINUOUS), and (2) predicting the binary presence of gender stereotypes and aggregating multiple individual predictions to compute the degree of stereotype (BINARY). We also have proposed

[4]Official rankings F1 scores only over the six GS categories (*role*, *personality*, *competence*, *physical*, *sexual*, *relational*), excluding instances with no stereotype. This differs from our development analysis, which included the *no stereotype* category.

**Table 4**
Official GSI:detect leaderboard for **GS value** prediction **(Main Task)**, separately for *zero-shot* (a) and *few-shot* (b).

| | 1/1+nMSE | team_name | model | configuration |
|---|---|---|---|---|
| 1 | 0.700 | DIAG-Sapienza | GPT-5 | gen_with_explanation |
| 2 | 0.626 | StereoBusters | Llama_3.3_70B | P3GUIDE |
| 3 | 0.622 | StereoBusters | Llama_3.3_70B | P5ITA |
| 4 | 0.621 | FestaDonato | Gemini_2.5_Flash | engNegative_V2 |
| 5 | 0.618 | FestaDonato | Gemini_2.5_Flash | ensemble |
| 6 | 0.610 | StereoBusters | Gemma_3_12B | P5ITA |
| 7 | 0.610 | FestaDonato | Gemini_2.5_Flash | ITA_DIR-V3 |
| 8 | 0,609 | BASELINE | GPT-5 nano | split_prompt |
| 9 | 0.591 | Tiz | Gemma_3_12B | average |
| 10 | 0,590 | BASELINE | GPT-5 nano | unified_prompt |
| 11 | 0.570 | StereoBusters | Panel_4LLMS | F1_score |
| 12 | 0.565 | StereoBusters | Panel_4LLMs | K_Cohen |
| 13 | 0.557 | FestaDonato | Gemini_2.5_Flash | CoT-eng_V5 |
| 14 | 0.555 | FestaDonato | Gemini_2.5_Flash | CoT-ita_V4 |
| 15 | 0,543 | BASELINE | Qwen-3_14B | unified_prompt |
| 16 | 0,498 | BASELINE | N/A | GS_value:0.5 |
| 17 | 0.484 | Tiz | Gemma_3_12B | max3 |
| 18 | 0.483 | Tiz | Gemma_3_12B | max2 |
| 19 | 0.472 | Tiz | Gemma_3_12B | max1 |
| 20 | 0.430 | Tiz | Gemma_3_12B | all_zeros |
| 21 | 0.329 | Prisma | Claude_3.5_Sonnet | persona_A |
| 22 | 0.321 | Prisma | Claude_3.5_Sonnet | persona_C |
| 23 | 0.321 | Prisma | Claude_3.5_Sonnet | 4_annotators |
| 24 | 0.309 | Prisma | Claude_3.5_Sonnet | persona_D |
| 25 | 0.309 | Prisma | Claude_3.5_Sonnet | persona_B |

(a) Zero-shot setting

| | 1/1+nMSE | team_name | model | configuration |
|---|---|---|---|---|
| 1 | 0.685 | DIAG-Sapienza | GPT-5 | gen_with_explanation |
| 2 | 0.645 | StereoBusters | Gemma_3_27B | P3GUIDE |
| 3 | 0.629 | StereoBusters | Gemma_3_12B | P1DIR |
| 4 | 0.612 | StereoBusters | Gemma_3_12B | P5ITA |
| 5 | 0.587 | Prisma | Claude_3.5_Sonnet | stratified_sampling |
| 6 | 0.586 | StereoBusters | Panel_4LLMS | F1_score |
| 7 | 0.581 | MINDS | Qwen2.5_14B | KNN |
| 8 | 0.572 | FestaDonato | Gemini_2.5 Flash | Ensemble |
| 9 | 0.565 | FestaDonato | Gemini_2.5 Flash | Semantic_Clustering |
| 10 | 0.564 | MINDS | Qwen2.5_14B | LR |
| 11 | 0.561 | FestaDonato | Gemini_2.5 Flash | Semantic_Clustering |
| 12 | 0.550 | FestaDonato | Gemini_2.5 Flash | Hybrid |
| 13 | 0.549 | StereoBusters | Panel_4LLMs | K_Cohen |
| 14 | 0.500 | FestaDonato | Gemini_2.5 Flash | Manual |

(b) Few-shot setting

**Table 5**
Official GSI:detect leaderboard for **GS category** classification **(SubTask)**, for *zero-shot* (a) and *few-shot* (b).

| | F1 Micro | F1 Macro | team_name | model | configuration |
|---|---|---|---|---|---|
| 1 | 0.646 | 0.637 | StereoBusters | Panel_4LLMS | F1_score |
| 2 | 0.644 | 0.637 | StereoBusters | Llama_3.3_70B | P3GUIDE |
| 3 | 0.643 | 0.633 | StereoBusters | Panel_4LLMs | K_Cohen |
| 4 | 0.634 | 0.630 | StereoBusters | Llama_3.3_70B | P5ITA |
| 5 | 0.600 | 0.561 | FestaDonato | Gemini_2.5_Flash | CoT-eng_V5 |
| 6 | 0.597 | 0.552 | FestaDonato | Gemini_2.5_Flash | ensemble |
| 7 | 0.590 | 0.552 | FestaDonato | Gemini_2.5_Flash | engNegative_V2 |
| 8 | 0.586 | 0.543 | FestaDonato | Gemini_2.5_Flash | ITA_DIR-V3 |
| 9 | 0.581 | 0.519 | DIAG-Sapienza | GPT-5 | gen_with_explanation |
| 10 | 0.569 | 0.523 | FestaDonato | Gemini_2.5_Flash | CoT-ita_V4 |
| 11 | 0.539 | 0.534 | StereoBusters | Gemma_3_12B | P5ITA |
| 12 | 0.532 | 0.519 | BASELINE | GPT-5 nano | split_prompt |
| 13 | 0.516 | 0.504 | BASELINE | GPT-5 nano | unified_prompt |
| 14 | 0.394 | 0.386 | BASELINE | Qwen-3_14B | unified_prompt |
| 15 | 0.229 | 0.229 | Tiz | Gemma_3_12B | max2 |
| 16 | 0.229 | 0.229 | Tiz | Gemma_3_12B | max3 |
| 17 | 0.224 | 0.223 | Tiz | Gemma_3_12B | max1 |
| 18 | 0.224 | 0.290 | Prisma | Claude_3.5_Sonnet | 4_annotators |
| 19 | 0.222 | 0.222 | Tiz | Gemma_3_12B | average |
| 20 | 0.208 | 0.208 | Tiz | Gemma_3_12B | all_zeros |
| 21 | 0.181 | 0.179 | BASELINE | N/A | GS_category: random |
| 22 | 0.180 | 0.229 | Prisma | Claude_3.5_Sonnet | persona_C |
| 23 | 0.174 | 0.236 | Prisma | Claude_3.5_Sonnet | persona_A |
| 24 | 0.129 | 0.178 | Prisma | Claude_3.5_Sonnet | persona_B |
| 25 | 0.122 | 0.171 | Prisma | Claude_3.5_Sonnet | persona_D |

(a) Zero-shot setting

| | F1 Micro | F1 Macro | team_name | model | configuration |
|---|---|---|---|---|---|
| 1 | 0.706 | 0.612 | Prisma | Claude_3.5_Sonnet | stratified_sampling |
| 2 | 0.671 | 0.669 | FestaDonato | Gemini_2.5 Flash | Hybrid |
| 3 | 0.669 | 0.665 | StereoBusters | Gemma_3_27B | P3GUIDE |
| 4 | 0.666 | 0.656 | FestaDonato | Gemini_2.5 Flash | Ensemble |
| 5 | 0.662 | 0.653 | StereoBusters | Panel_4LLMs | K_Cohen |
| 6 | 0.662 | 0.654 | StereoBusters | Panel_4LLMS | F1_score |
| 7 | 0.657 | 0.653 | FestaDonato | Gemini_2.5 Flash | Semantic_Clustering |
| 8 | 0.650 | 0.635 | FestaDonato | Gemini_2.5 Flash | Semantic_Clustering |
| 9 | 0.636 | 0.627 | FestaDonato | Gemini_2.5 Flash | Manual |
| 10 | 0.606 | 0.549 | DIAG-Sapienza | GPT-5 | gen_with_explanation |
| 11 | 0.574 | 0.572 | StereoBusters | Gemma_3_12B | P1DIR |
| 12 | 0.565 | 0.561 | StereoBusters | Gemma_3_12B | P5ITA |

(b) Few-shot setting

different prompting strategies and evaluated how different LLMs perform for each of these. Finally, we conducted an additional annotation experiment using the annotation guidelines (see Appendix C).

Our results show that open-weight models, even with a relatively small number of parameters (e.g., Gemma 12B), can be highly effective, when properly prompted, at detecting both the presence and type of stereotypes, achieving performance close to that of more powerful commercial models. We found that both proposed approaches are promising for identifying and classifying gender stereotypes. However, while the BINARY approach outperformed the CONTINUOUS approach on the development data, it generalized less effectively to the test data. We also observed that few-shot prompting tends to yield limited performance improvements, except for the CONTINUOUS approach on the main task.

In general, all model-prompt combinations showed good performance, the main exception being the CoT prompt, which tends to exhibit lower performance for the main task. This pattern was partially flipped in the case of the subtask, where the CoT model showed competitive performance when compared to the other prompting strategies, if not straight outperforming them.

## Acknowledgments

The work of M. Marchiori Manerba is partially funded by the project CSP TRAPEZIO 2025-2025.0998 “Neikea-HS: NLP and Ethical AI for Knowledge-based Education Against Hate Speech”. The work of A.T. Cignarella is supported by the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Actions, Grant Agreement No. 101146287. Computational resources were provided by King’s Computational Research, Engineering and Technology Environment (CREATE) at King’s College London [32].

## Declaration on Generative AI

During the preparation of this work, the authors used **ChatGPT** and **Grammarly** to: **Grammar and spelling check, and Paraphrase and reword**. The authors reviewed the content and take full responsibility for it.

## References


[1] C. Bosco, F. Dell'Orletta, F. Poletto, M. Sanguinetti, M. Tesconi, Overview of the EVALITA 2018 Hate Speech Detection Task, in: Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Turin, Italy, December 12-13, 2018, volume 2263 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2018.

[2] M. Sanguinetti, G. Comandini, E. D. Nuovo, S. Frenda, M. Stranisci, C. Bosco, T. Caselli, V. Patti, I. Russo, HaSpeeDe 2 @ EVALITA2020: Overview of the EVALITA 2020 Hate Speech Detection Task, in: V. Basile, D. Croce, M. D. Maro, L. C. Passaro (Eds.), Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020), volume 2765 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020.

[3] 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. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), CEUR Workshop Proceedings, 2018.

[4] E. Fersini, D. Nozza, P. Rosso, AMI @ EVALITA2020: automatic misogyny identification, in: Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020), 2020.

[5] 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: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), volume 3473 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023.

[6] M. Nissim, D. Croce, V. Patti, P. Basile, G. Attanasio, et al., Challenging the abilities of large language models in italian: a community initiative, 2025. URL: https://arxiv.org/abs/2512.04759.

[7] C. Bosco, V. Patti, S. Frenda, A. T. Cignarella, M. Paciello, F. D'Errico, Detecting racial stereotypes: An italian social media corpus where psychology meets nlp, Information Processing Management 60 (2023) 103118. doi:https://doi.org/10.1016/j.ipm.2022.103118.

[8] W. Wolfgang Schmeisser-Nieto, G. Ricci, S. Frenda, M. Taule, C. Bosco, Implicit stereotypes: A corpus-based study for Italian, in: Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024), CEUR Workshop Proceedings, Pisa, Italy, 2024, pp. 997–1004. URL: https://aclanthology.org/2024.clicit-1.108/.

[9] B. Cristina, P. Marinella, F. Benamara, C. P. Giovanni, P. Viviana, M. Véronique, T. Mariona, et al., Sterheotypes project. detecting and countering ethnic stereotypes emerging from italian, spanish and french racial hoaxes, in: Proceedings of the Seminar of the Spanish Society for Natural Language Processing: Projects and System Demonstrations (SEPLN-CEDI-PD 2024), 2024.

[10] M. Berta, S. Greco, G. Tipaldo, T. Cerquitelli, Decoding narratives: Towards a classification analysis for stereotypical patterns in italian news headlines, in: 2024 IEEE International Conference on Big Data (BigData), 2024, pp. 5253–5262. doi:10.1109/BigData62323.2024.10825258.

[11] S. Greco, M. La Quatra, L. Cagliero, T. Cerquitelli, Towards ai-assisted inclusive language writing in italian formal communications, ACM Trans. Intell. Syst. Technol. 16 (2025). URL: https://doi.org/10.1145/3729237. doi:10.1145/3729237.

[12] B. Savoldi, G. Attanasio, E. Cupin, E. Gkovedarou, J. Hackenbuchner, A. Lauscher, M. Negri, A. Piergentili, M. Thind, L. Bentivogli, Mind the inclusivity gap: Multilingual gender-neutral translation evaluation with mGeNTE, in: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Suzhou, China, 2025, pp. 13709–13731. doi:10.18653/v1/2025.emnlp-main.692.

[13] M. La Quatra, S. Greco, L. Cagliero, T. Cerquitelli, Inclusively: An ai-based assistant for inclusive writing, in: Machine Learning and Knowledge Discovery in Databases: Applied Data Science

and Demo Track: European Conference, ECML PKDD 2023, Turin, Italy, September 18–22, 2023, Proceedings, Part VII, 2023, p. 361–365. doi:10.1007/978-3-031-43430-3_31.

[14] M. L. Quatra, S. Greco, L. Cagliero, M. Tonti, F. Dragotto, R. Raus, S. Cavagnoli, T. Cerquitelli, Building foundations for inclusiveness through expert-annotated data, in: EDBT/ICDT Workshops, 2024. URL: https://ceur-ws.org/Vol-3651/DARLI-AP-3.pdf.

[15] G. Attanasio, S. Greco, M. La Quatra, L. Cagliero, M. Tonti, T. Cerquitelli, R. Raus, E-mimic: Empowering multilingual inclusive communication, in: 2021 IEEE International Conference on Big Data (Big Data), 2021, pp. 4227–4234. doi:10.1109/BigData52589.2021.9671868.

[16] A. Piergentili, B. Savoldi, M. Negri, L. Bentivogli, Gender-neutral rewriting in Italian: Models, approaches, and trade-offs, in: Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025), CEUR Workshop Proceedings, Cagliari, Italy, 2025, pp. 899–910. URL: https://aclanthology.org/2025.clicit-1.84/.

[17] K. Stanczak, I. Augenstein, A Survey on Gender Bias in Natural Language Processing, arXiv preprint arXiv:2112.14168 (2021).

[18] A. T. Cignarella, A. Giachanou, E. Lefever, A Survey on Stereotype Detection in Natural Language Processing, ACM Computing Surveys 58 (2025) 1–33. doi:10.1145/3770754.

[19] G. Comandini, M. Speranza, S. Brenna, D. Testa, S. Cavagnoli, B. Magnini, Gsi:detect at evalita 2026: Overview of the task on detecting gender stereotypes in italian, 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.

[20] F. Cutugno, A. Miaschi, A. P. Aprosio, G. Rambelli, L. Siciliani, M. A. Stranisci, Evalita 2026: Overview of the 9th evaluation campaign of natural language processing and speech tools for italian, 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.

[21] M. Sanguinetti, F. Poletto, C. Bosco, V. Patti, M. Stranisci, An Italian Twitter corpus of hate speech against immigrants, in: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), European Language Resources Association (ELRA), Miyazaki, Japan, 2018. URL: https://aclanthology.org/L18-1443/.

[22] V. Basile, et al., It's the end of the gold standard as we know it. on the impact of pre-aggregation on the evaluation of highly subjective tasks, volume 2776, CEUR-WS, 2020, pp. 31–40.

[23] F. Cabitza, A. Campagner, V. Basile, Toward a perspectivist turn in ground truthing for predictive computing, Proceedings of the AAAI Conference on Artificial Intelligence 37 (2023) 6860–6868. doi:10.1609/aaai.v37i6.25840.

[24] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le, D. Zhou, Chain-of-thought prompting elicits reasoning in large language models, in: Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS '22, 2022.

[25] A. Grattafiori, A. Dubey, A. Jauhri, et al., The llama 3 herd of models, 2024. arXiv:2407.21783.

[26] AI@Meta, Llama 3 model card (2024). URL: https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.

[27] Q. Team, Qwen3 technical report, 2025. URL: https://arxiv.org/abs/2505.09388.

[28] M. Polignano, P. Basile, G. Semeraro, Advanced natural-based interaction for the italian language: Llamantino-3-anita, 2024. arXiv:2405.07101.

[29] P. Basile, E. Musacchio, M. Polignano, L. Siciliani, G. Fiameni, G. Semeraro, Llamantino: Llama 2 models for effective text generation in italian language, 2023. arXiv:2312.09993.

[30] G. Team, Gemma 3 (2025). URL: https://goo.gle/Gemma3Report.

[31] J. Cohen, A coefficient of agreement for nominal scales, Educational and psychological measurement 20 (1960) 37–46.

[32] King's College London, King's computational research, engineering and technology environment (create), https://doi.org/10.18742/rnvf-m076, 2025. Retrieved December, 2025.

[33] M. Sap, S. Gabriel, L. Qin, D. Jurafsky, N. A. Smith, Y. Choi, Social bias frames: Reasoning about social and power implications of language, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020. doi:10.18653/v1/2020.acl-main.486.

# Appendix

The appendix is organised as follows: Appendix A presents additional analysis of the GSI:DETECT dataset; Appendix B provides the LLMs' prompts under the CONTINUOUS and BINARY approaches; and Appendix C discusses an additional human annotation experiment and a qualitative analysis of the data.

## A. Analysis of the GSI:detect dataset

As introduced in Section 2, the GSI:DETECT dataset comprises 1,010 short Italian texts collected from both social media and information websites, covering both formal and informal registers. The dataset is split into 200 instances for development and 810 instances for testing. The information and annotations provided for each text are summarised in Table 6.

The annotation process involved four annotators with expertise and sensitivity to gender-related issues. Demographic characteristics of the annotators were released midway through the development phase of the task (see Table 7) to support the design of systems that better reflect annotator diversity.

The distribution of the *GS value* across the dataset is shown in Figure 1a. The development and test sets present similar distributions: approximately one third of the texts have a GS value of 0 (i.e., none of the annotators identified stereotypes), and approximately one third have a GS value of 1 (i.e., all annotators identified stereotypes). However, the test set contains slightly more texts with a GS value of 1, while the development set contains fewer texts with intermediate values (0.25 and 0.5).

Similarly, the distributions of the *GS category* follow a comparable pattern in the development and test sets, as shown in Figure 1b. The most common category other than "no stereotype" is the *competence* stereotype category, followed closely by *role* and *personality* stereotypes. The least represented categories are *sexual* and *relational* stereotypes. Although the overall distributions are similar across the two splits, *competence*, *role*, and *personality* stereotypes are more prevalent in the development set, whereas *physical*, *sexual*, and *relational* stereotypes occur more frequently in the test set.

We computed Cohen's $\kappa$ to quantify the agreement between annotators. Table 8 reports the pairwise agreement on the presence of stereotypes (i.e., yes/no label) for the development set, while Table 9 reports the agreement for the test set. The values for the development dataset lie between 0.45 and 0.63, which indicates moderate to substantial agreement. With the exception of two of the annotators (A1—A3), the values for the test set lie between 0.61 and 0.69, which indicates substantial agreement.

## B. LLM Prompts

As discussed in Section 3, we proposed and evaluated five prompting strategies: (1) direct instructions, (2) integration of detailed guidelines, (3) Chain-of-Thought prompting, (4) use of Italian prompts, and (5) modeling the perspectives of different annotators. LLMs' prompts are reported in Figure 2 for the CONTINUOUS approach and in Figure 3 for the BINARY approach.

For the *few-shot* setting, the prompt structure is identical across the two settings; however, in the few-shot setting, five illustrative examples are appended to the end of each prompt, as shown in the red text in Figure 2f. We report only this example, as all other prompts follow the same structure.

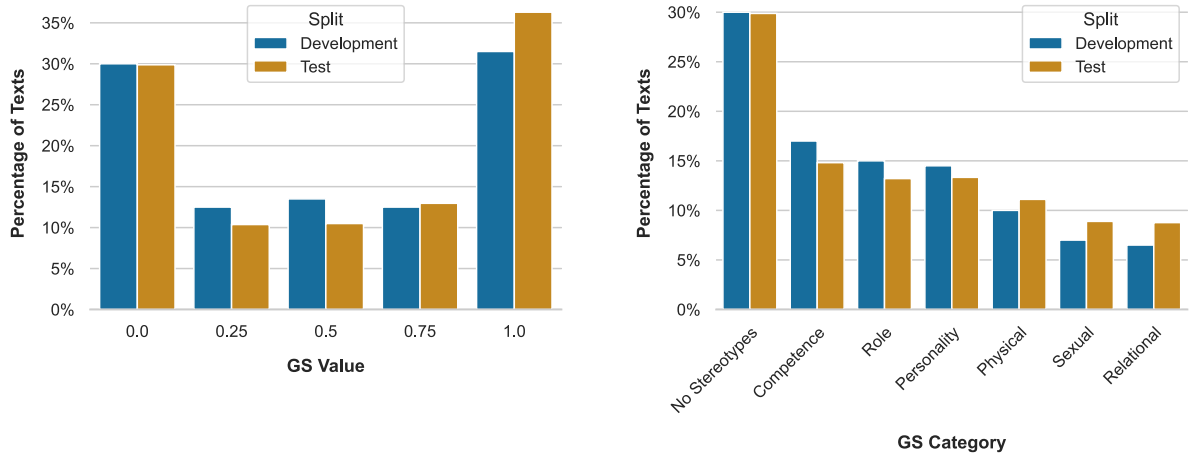
We designed all prompts to predict either the GS value or the binary gender stereotype label, together with the GS category. Models are instructed to assign the "no" category only when the GS value or label is 0 (i.e., when no stereotypes are detected).

**Table 6**
**List of variables and annotations provided for each text within the GSI:detect dataset.**

| Variable | Description |
|---|---|
| Text | The short text to be used for the task and any additional context. |
| Annotations | A list containing four binary annotations denoting the presence of stereotypes in the text, one entry for each annotator. |
| GS Value | The GS value for the text. As mentioned in Section ??, this is a numerical value between 1 and 0. |
| GS Category | The GS category for the text. |
| Context | A binary label denoting whether additional information regarding the context in which the text appears. |
| Annotations | A list containing four binary annotations (yes/no) about the presence of stereotypes in the text, one for each annotator. |

**Table 7**
**Demographic characteristics of the four annotators.**

| Characteristic | Value | Count |
|---|---|---|
| Gender | Cis woman | 3 |
| | Cis man | 1 |
| Age | 20-30 | 2 |
| | 30-40 | 1 |
| | 40+ | 1 |
| Educational Level | PhD | 1 |
| | Masters | 3 |
| First Language | Italian | 4 |




(a) GS values distribution

(b) GS categories distribution (sorted by frequency)

**Figure 1: Distribution of the GS values and categories across the development and test sets.**

**Table 8**
Pairwise agreement of the development set.

| Ann_A | Ann_B | Cohen's $\kappa$ |
|---|---|---|
| A1 | A2 | 0.6172 |
| A1 | A3 | **0.4528** |
| A1 | A4 | 0.5527 |
| A2 | A3 | 0.5603 |
| A2 | A4 | 0.6203 |
| A3 | A4 | **0.6328** |

**Table 9**
Pairwise agreement of the test set.

| Ann_A | Ann_B | Cohen's $\kappa$ |
|---|---|---|
| A1 | A2 | 0.6250 |
| A1 | A3 | **0.4972** |
| A1 | A4 | 0.6534 |
| A2 | A3 | 0.6107 |
| A2 | A4 | **0.6912** |
| A3 | A4 | 0.6772 |

```
system: |
  You are an expert in detecting gender stereotypes in Italian text.
  Your task is to analyze texts and provide two outputs:
  1. A GS value (0.0 to 1.0) indicating the degree of gender stereotype present
  2. The category of stereotype if present

  GS Value Scale:
  - 0.0 = No stereotype
  - 0.25 = Weak stereotype (1 of 4 annotators would identify it)
  - 0.50 = Moderate stereotype (2 of 4 annotators would identify it)
  - 0.75 = Strong stereotype (3 of 4 annotators would identify it)
  - 1.0 = Very strong stereotype (all annotators would identify it)

  Categories:
  - role: Social/cultural expectations about what men/women should do
  - personality: Emotional/behavioral traits assigned by gender
  - competence: Judgments about abilities based on gender
  - physical: Expectations about appearance and body
  - sexual: Attitudes about sexuality and sexual behavior
  - relational: How genders should behave in relationships
  - no: No stereotype present

  IMPORTANT: Respond ONLY with the JSON output. Do not include any explanations
      before or after.
  Format your response EXACTLY as: JSON_OUTPUT:{"gs_value": X.XX, "gs_category":
      "category"}
user: |
  Analyze this Italian text for gender stereotypes:

  Text: "{text}"

  Provide your analysis as JSON_OUTPUT:{{"gs_value":<number between 0.0 and 1.0>,
      "gs_category":"<role|personality|competence|physical|sexual|relational|no>"}}
user_with_context: |
  Analyze this Italian text for gender stereotypes. The text includes context
      that should inform your evaluation.

  Context: "{context}"

  Text to evaluate: "{text}"

  IMPORTANT: Evaluate the main text considering the provided context. The context
      helps understand the situation but your analysis should focus on the main
      text.

  Provide your analysis as JSON_OUTPUT:{{"gs_value":<number between 0.0 and 1.0>,
      "gs_category":"<role|personality|competence|physical|sexual|relational|no>"}}
```

(a) **Zero-Shot Prompt 1 (Continuous): Direct**

```
system: |
  You are an expert annotator trained in detecting gender stereotypes according
      to specific guidelines.

  DEFINITION: Gender stereotypes are rigid and generalized beliefs about roles,
      behaviors, abilities,
  or characteristics that men and women "should" have based on their sex or
      gender.

  DETAILED CATEGORIES:

  1. ROLE STEREOTYPES
     - Men: breadwinners, heads of household, financial support, protection
     - Women: mothers, housewives, child-rearing, domestic work
     - Note: Work aspects usually fall under "competence"

  2. PERSONALITY STEREOTYPES
     - Women: empathetic, sensitive, emotional, indecisive, caring
     - Men: strong, assertive, rational, confident, unemotional
     - Includes gendered insults (e.g., "femminuccia", "avere le palle")

  3. COMPETENCE STEREOTYPES
     - Women: suited for humanities, supportive roles
     - Men: skilled in STEM, logic, technology, leadership
     - Includes sports, games, professional abilities

  4. PHYSICAL STEREOTYPES
     - Women: slim, young, attractive, well-groomed (judged harshly)
     - Men: robust, can be scruffy (more accepted)
     - IMPORTANT: Comments on appearance are stereotypes when contextually
       irrelevant

  5. SEXUAL STEREOTYPES
     - Women: passive, pure, judged for sexual freedom
     - Men: active, "hunters", naturally sexual
     - Includes sexual insults like "puttana", "zoccola", "figlio/a di puttana"

  6. RELATIONAL STEREOTYPES
     - Women: caring, faithful, need protection, submissive
     - Men: dominant, autonomous, emotionally detached
     - Includes judgments about singlehood, marriage motivations, legal bias
       claims

  ANNOTATION RULE: If multiple stereotypes appear, choose the one that appears
      FIRST in the text.

  IMPORTANT: End your response with ONLY the JSON output.
  Format: JSON_OUTPUT: {"gs_value": X.XX, "gs_category": "category"}
user: |
  Text to analyze: "{text}"

  Following the guidelines above:
  1. Does this text contain a gender stereotype?
  2. How obvious/strong is it? (Consider: would most people recognize it?)
  3. Which category does it belong to?
  4. What GS value should it receive? (0.0 = none, 1.0 = maximum)

  Provide your analysis ending with: JSON_OUTPUT:{{"gs_value": <0.00-1.00>,
      "gs_category":"<role|personality|competence|physical|sexual|relational|no>"}}
user_with_context: |
  Context: "{context}"

  Text to analyze: "{text}"

  Following the guidelines above and considering the context:
  1. Does the main text contain a gender stereotype?
  2. How does the context affect the interpretation?
  3. How obvious/strong is the stereotype? (Consider: would most people recognize
      it?)
  4. Which category does it belong to?
  5. What GS value should it receive? (0.0 = none, 1.0 = maximum)

  Provide your analysis ending with: JSON_OUTPUT:{{"gs_value":<0.00-1.00>,
      "gs_category":"<role|personality|competence|physical|sexual|relational|no>"}}
```

(c) **Zero-Shot Prompt 3 (Continuous): Guidelines**

```
system: |
  You are an expert linguist specializing in gender stereotype detection in
      Italian text.

  After your reasoning, you MUST end with ONLY the JSON output on a new line.
  Format: JSON_OUTPUT: {"gs_value": X.XX, "gs_category": "category"}
user: |
  Analyze the following Italian text for gender stereotypes using step-by-step
      reasoning:

  Text: "{text}"

  Follow these steps:
  1. Identify if any gender-based generalizations are present
  2. Determine if these generalizations are stereotypical
  3. Assess the strength/obviousness of the stereotype
  4. Classify the type of stereotype
  5. Assign a numerical score

  Categories:
  - role: What men/women should do/be (family roles, social expectations)
  - personality: Character traits by gender (emotional, rational, etc.)
  - competence: Ability judgments (STEM, leadership, sports)
  - physical: Appearance expectations (beauty, grooming, body)
  - sexual: Sexuality and sexual behavior norms
  - relational: Behavior in relationships and interactions
  - no: No stereotype present

  GS Value (0.0-1.0):
  - 1.0 = Explicit, clear stereotype
  - 0.75 = Strong but may have slight ambiguity
  - 0.50 = Moderate, some would debate
  - 0.25 = Weak, subtle, or context-dependent
  - 0.0 = No stereotype

  After your reasoning, provide ONLY: JSON_OUTPUT: {{"gs_value": <0.0-1.0>,
      "gs_category": "<category>"}}
user_with_context: |
  Analyze the following Italian text for gender stereotypes using step-by-step
      reasoning. Consider the context provided.

  Context: "{context}"

  Text to evaluate: "{text}"

  Follow these steps:
  1. Consider how the context frames the main text
  2. Identify if any gender-based generalizations are present in the main text
  3. Determine if these generalizations are stereotypical
  4. Assess the strength/obviousness of the stereotype
  5. Classify the type of stereotype
  6. Assign a numerical score

  Categories:
  - role: What men/women should do/be (family roles, social expectations)
  - personality: Character traits by gender (emotional, rational, etc.)
  - competence: Ability judgments (STEM, leadership, sports)
  - physical: Appearance expectations (beauty, grooming, body)
  - sexual: Sexuality and sexual behavior norms
  - relational: Behavior in relationships and interactions
  - no: No stereotype present

  GS Value (0.0-1.0):
  - 1.0 = Explicit, clear stereotype
  - 0.75 = Strong but may have slight ambiguity
  - 0.50 = Moderate, some would debate
  - 0.25 = Weak, subtle, or context-dependent
  - 0.0 = No stereotype

  After your reasoning, provide ONLY: JSON_OUTPUT: {{"gs_value": <0.0-1.0>,
      "gs_category": "<category>"}}
```

(b) **Zero-Shot Prompt 2 (Continuous): CoT**

```
system: |
  Sei un esperto di stereotipi di genere nella cultura e lingua italiana.

  Il tuo compito è analizzare testi brevi e identificare:
  1. Il livello di stereotipo di genere (valore GS da 0.0 a 1.0)
  2. La categoria dello stereotipo

  CATEGORIE:
  - role: Ruoli familiari e sociali (es. uomo capofamiglia, donna casalinga)
  - personality: Tratti emotivi/comportamentali (es. donna emotiva, uomo
      razionale)
  - competence: Capacit‘a basate sul genere (es. uomini bravi in matematica)
  - physical: Aspetto fisico e cura personale (es. donne devono essere belle)
  - sexual: Comportamento sessuale (es. donna pura, uomo cacciatore)
  - relational: Dinamiche relazionali (es. uomo dominante, donna sottomessa)
  - no: Nessuno stereotipo presente

  VALORI GS:
  - 1.0 = Stereotipo molto chiaro ed esplicito
  - 0.75 = Stereotipo forte
  - 0.50 = Stereotipo moderato, può essere dibattuto
  - 0.25 = Stereotipo debole o sottile
  - 0.0 = Nessuno stereotipo

  IMPORTANTE: Termina la risposta SOLO con l’output JSON.
  Formato: JSON_OUTPUT: {"gs_value": X.XX, "gs_category": "category"}
user: |
  Analizza questo testo italiano:

  "{text}"

  Termina con: JSON_OUTPUT: {{"gs_value": <0.0-1.0>, "gs_category":
      "<role|personality|competence|physical|sexual|relational|no>"}}
user_with_context: |
  Analizza questo testo italiano considerando il contesto fornito.

  Contesto: "{context}"

  Testo da valutare: "{text}"

  IMPORTANTE: Valuta il testo principale considerando il contesto fornito. Il
      contesto aiuta a comprendere la situazione ma l’analisi deve concentrarsi
      sul testo principale.

  Termina con: JSON_OUTPUT: {{"gs_value": <0.0-1.0>, "gs_category":
      "<role|personality|competence|physical|sexual|relational|no>"}}
```

(d) **Zero-Shot Prompt 4 (Continuous): Italian**

**Figure 2: Zero-Shot Continuous Classification Prompts.**

```
system: |
  You are simulating a panel of 4 annotators evaluating gender stereotypes in
      Italian text.

  Some texts are clearly stereotypical (all would agree), others are ambiguous
      (annotators may disagree).
  Your task is to estimate how many of 4 annotators would identify a gender
      stereotype.

  Categories (choose ONE):
  - role: Social roles men/women should have
  - personality: Character/emotional traits by gender
  - competence: Ability judgments based on gender
  - physical: Appearance/body expectations
  - sexual: Sexual behavior and attitudes
  - relational: Relationship dynamics and expectations
  - no: No stereotype present

  IMPORTANT: End your response with ONLY the JSON output.
  Format: JSON_OUTPUT: {"gs_value": X.XX, "gs_category": "category"}
user: |
  Text: "{text}"

  Think like 4 different annotators:
  - Annotator 1: Very sensitive to subtle stereotypes
  - Annotator 2: Moderately sensitive, needs clearer signals
  - Annotator 3: Moderately sensitive, context matters
  - Annotator 4: Only identifies explicit stereotypes

  For each annotator, would they say "yes, this is a stereotype" or "no"?

  Count the "yes" responses:
  - 4 yes  gs_value = 1.0
  - 3 yes  gs_value = 0.75
  - 2 yes  gs_value = 0.50
  - 1 yes  gs_value = 0.25
  - 0 yes  gs_value = 0.0

  Then classify the category (if stereotype present).

  End with: JSON_OUTPUT: {{"gs_value": <0.0-1.0>, "gs_category": "<category>"}}
user_with_context: |
  Context: "{context}"

  Text to evaluate: "{text}"

  Consider the context when evaluating. Think like 4 different annotators:
  - Annotator 1: Very sensitive to subtle stereotypes
  - Annotator 2: Moderately sensitive, needs clearer signals
  - Annotator 3: Moderately sensitive, context matters
  - Annotator 4: Only identifies explicit stereotypes

  For each annotator, considering the context, would they say "yes, this is a
      stereotype" or "no"?

  Count the "yes" responses:
  - 4 yes  gs_value = 1.0
  - 3 yes  gs_value = 0.75
  - 2 yes  gs_value = 0.50
  - 1 yes  gs_value = 0.25
  - 0 yes  gs_value = 0.0

  Then classify the category (if stereotype present).

  End with: JSON_OUTPUT: {{"gs_value": <0.0-1.0>, "gs_category": "<category>"}}
```

(e) **Zero-Shot Prompt 5 (CONTINUOUS): Perspectivist**

```
system: |
  You are an expert in detecting gender stereotypes in Italian text.
  Your task is to analyze texts and provide two outputs:
  1. A GS value (0.0 to 1.0) indicating the degree of gender stereotype present
  2. The category of stereotype if present

  GS Value Scale:
  - 0.0 = No stereotype
  - 0.25 = Weak stereotype (1 of 4 annotators would identify it)
  - 0.50 = Moderate stereotype (2 of 4 annotators would identify it)
  - 0.75 = Strong stereotype (3 of 4 annotators would identify it)
  - 1.0 = Very strong stereotype (all annotators would identify it)

  Categories:
  - role: Social/cultural expectations about what men/women should do
  - personality: Emotional/behavioral traits assigned by gender
  - competence: Judgments about abilities based on gender
  - physical: Expectations about appearance and body
  - sexual: Attitudes about sexuality and sexual behavior
  - relational: How genders should behave in relationships
  - no: No stereotype present

  Here are some examples:
  {examples}

  IMPORTANT: Respond ONLY with the JSON output. Do not include any explanations
      before or after.
  Format your response EXACTLY as: JSON_OUTPUT: {{"gs_value": X.XX,
      "gs_category": "category"}}
user: |
  Now analyze this Italian text for gender stereotypes:

  Text: "{text}"

  Provide your analysis as JSON_OUTPUT:{{"gs_value":<number between 0.0 and
      1.0>,"gs_category":
      "<role|personality|competence|physical|sexual|relational|no>"}}
```

(f) **Few-Shot Prompt 1 (CONTINUOUS): Direct**

**Figure 2: Additional zero-shot CONTINUOUS Classification Prompts, and an example of few-shot (f).**

```
system: |
  You are an expert in detecting gender stereotypes in Italian.

  TASK: Determine if the text contains a gender stereotype (0 = no, 1 = yes)

  STEREOTYPE = 1 when the text makes claims about "men" or "women" in general
      with stereotypical patterns.
  - Are like: "Gli uomini sono razionali"
  - Should do: "Le donne devono occuparsi della casa"
  - Cannot do: "Un uomo non può essere sensibile"
  STEREOTYPE = 0 when the text describes:
  - Specific individuals: "Una donna entra nel negozio"
  - Defined subgroups: "Gli uomini della mia famiglia"
  - Gender mention without generalization: "Il ragazzo serve al tavolo"

  CATEGORIES (select only the FIRST category that appears in the text):
  - role: Family/social roles and expectations
  - personality: Character traits (emotional, rational, etc.)
  - competence: Abilities (STEM, leadership, sports, etc.)
  - physical: Appearance expectations
  - sexual: Sexuality and behavior norms
  - relational: Relationship behaviors
  - no: Use only when stereotype = 0

  OUTPUT FORMAT:
  {"stereotype": 0, "category": "no"}

  Evaluate ONLY the given text. Do not infer missing context.

user: |
  Analyze this Italian text for gender stereotype:

  TEXT: "{text}"

  JSON_OUTPUT: {{"stereotype": <0 or 1>, "category": "<category>"}}

user_with_context: |
  Analyze this Italian text for gender stereotype within the provided context.

  CONTEXT: "{context}"

  TEXT: "{text}"

      Evaluate ONLY the main text. Context is for interpretation, not
      classification.

      JSON_OUTPUT: {{"stereotype": <0 or 1>, "category": "<category>"}}
```

(a) **Zero-Shot Prompt 1 (Binary): Direct**

```
system: |
  You are an expert in detecting gender stereotypes in Italian.

  After your reasoning, you MUST end with ONLY the JSON output on a new line.
  Format: JSON_OUTPUT: {"stereotype": 0, "category": "no"}

user: |
  Analyze the following Italian text for gender stereotypes using step-by-step
      reasoning:

  TEXT: "{text}"

  Follow these steps:
  1. Identify if any gender-based generalizations are present
  2. Determine if these generalizations are stereotypical
  3. Classify the type of stereotype if present
  4. Make a binary decision (stereotype present or not)

  Binary Classification:
  - stereotype = 1: Gender stereotype IS present
  - stereotype = 0: Gender stereotype is NOT present

  Categories:
  - role: What men/women should do/be (family roles, social expectations)
  - personality: Character traits by gender (emotional, rational, etc.)
  - competence: Ability judgments (STEM, leadership, sports)
  - physical: Appearance expectations (beauty, grooming, body)
  - sexual: Sexuality and sexual behavior norms
  - relational: Behavior in relationships and interactions
  - no: No stereotype present

  After your reasoning, provide ONLY: JSON_OUTPUT: {{"stereotype": <0 or 1>,
      "category": "<category>"}}

user_with_context: |
  Analyze the following Italian text for gender stereotypes using step-by-step
      reasoning. Consider the context provided.

  CONTEXT: "{context}"

  TEXT: "{text}"

  Follow these steps:
  1. Consider how the context frames the main text
  2. Identify if any gender-based generalizations are present in the main text
  3. Determine if these generalizations are stereotypical
  4. Classify the type of stereotype if present
  5. Make a binary decision (stereotype present or not)

  Binary Classification:
  - stereotype = 1: Gender stereotype IS present
  - stereotype = 0: Gender stereotype is NOT present

  Categories:
  - role: What men/women should do/be (family roles, social expectations)
  - personality: Character traits by gender (emotional, rational, etc.)
  - competence: Ability judgments (STEM, leadership, sports)
  - physical: Appearance expectations (beauty, grooming, body)
  - sexual: Sexuality and sexual behavior norms
  - relational: Behavior in relationships and interactions
  - no: No stereotype present

  After your reasoning, provide ONLY: JSON_OUTPUT: {{"stereotype": <0 or 1>,
      "category": "<category>"}}
```

(b) **Zero-Shot Prompt 2 (Binary): CoT**

```
system: |
  You are an expert in detecting gender stereotypes in Italian.

  TASK: Determine if the text contains a gender stereotype (0 = no, 1 = yes)

  STEREOTYPE = 1 when the text makes claims about "men" or "women" in general.
  - Are like: "Gli uomini sono razionali"
  - Should do: "Le donne devono occuparsi della casa"
  - Cannot do: "Un uomo non può essere sensibile"
  STEREOTYPE = 0 when the text describes:
  - Specific individuals: "Una donna entra nel negozio"
  - Defined subgroups: "Gli uomini della mia famiglia"
  - Gender mention without generalization: "Il ragazzo serve al tavolo"

  CATEGORIES (select only the FIRST category that appears in the text):

  1. ROLE STEREOTYPES
     - Men: breadwinners, heads of household, financial support, protection
     - Women: mothers, housewives, child-rearing, domestic work
     - Note: Work aspects usually fall under "competence"

  2. PERSONALITY STEREOTYPES
     - Women: empathetic, sensitive, emotional, indecisive, caring
     - Men: strong, assertive, rational, confident, unemotional
     - Includes gendered insults (e.g., "femminuccia", "avere le palle")

  3. COMPETENCE STEREOTYPES
     - Women: suited for humanities, supportive roles
     - Men: skilled in STEM, logic, technology, leadership
     - Includes sports, games, professional abilities

  4. PHYSICAL STEREOTYPES
     - Women: slim, young, attractive, well-groomed (judged harshly)
     - Men: robust, can be scruffy (more accepted)
     - IMPORTANT: Comments on appearance are stereotypes when contextually
       irrelevant

  5. SEXUAL STEREOTYPES
     - Women: passive, pure, judged for sexual freedom
     - Men: active, "hunters", naturally sexual
     - Includes sexual insults like "puttana", "zoccola", "figlio/a di puttana"

  6. RELATIONAL STEREOTYPES
     - Women: caring, faithful, need protection, submissive
     - Men: dominant, autonomous, emotionally detached
     - Includes judgments about singlehood, marriage motivations, legal bias
       claims

  OUTPUT FORMAT:
  {"stereotype": 0, "category": "no"}

  Evaluate ONLY the given text. Do not infer missing context.

user: |
  Analyze this Italian text for gender stereotype:

  TEXT: "{text}"

  JSON_OUTPUT: {{"stereotype": <0 or 1>, "category": "<category>"}}

user_with_context: |
  Analyze this Italian text for gender stereotype within the provided context.

  CONTEXT: "{context}"

  TEXT: "{text}"

  Evaluate ONLY the main text. Context is for interpretation, not classification.

  JSON_OUTPUT: {{"stereotype": <0 or 1>, "category": "<category>"}}
```

(c) **Zero-Shot Prompt 3 (Binary): Guidelines**

```
system: |
  Sei un esperto nel rilevamento di stereotipi di genere in italiano.

  COMPITO: Determina se il testo contiene uno stereotipo di genere (0 = no, 1 =
      s'i)

  STEREOTIPO = 1 quando il testo fa affermazioni su "uomini" o "donne" in
      generale.
  - Sono come: "Gli uomini sono razionali"
  - Dovrebbero fare: "Le donne devono occuparsi della casa"
  - Non possono fare: "Un uomo non può essere sensibile"
  STEREOTIPO = 0 quando il testo descrive:
  - Individui specifici: "Una donna entra nel negozio"
  - Sottogruppi definiti: "Gli uomini della mia famiglia"
  - Menzione del genere senza generalizzazione: "Il ragazzo serve al tavolo"

  CATEGORIE (seleziona solo la PRIMA categoria che compare nel testo):
  - role: Ruoli e aspettative familiari/sociali
  - personality: Tratti caratteriali (emotivo, razionale, ecc.)
  - competence: Abilit'a (STEM, leadership, sport, ecc.)
  - physical: Aspettative sull'aspetto fisico
  - sexual: Sessualit'a e norme comportamentali
  - relational: Comportamenti nelle relazioni
  - no: Usa solo quando stereotype = 0

  FORMATO OUTPUT:
  {"stereotype": 0, "category": "no"}

  Valuta SOLO il testo fornito. Non inferire contesto mancante.

user: |
  Analizza questo testo italiano per rilevare stereotipi di genere:

  TESTO: "{text}"

  JSON_OUTPUT: {{"stereotype": <0 o 1>, "category": "<categoria>"}}

user_with_context: |
  Analizza questo testo italiano per rilevare stereotipi di genere nel contesto
      fornito.

  CONTESTO: "{context}"

  TESTO: "{text}"

  Valuta SOLO il testo principale. Il contesto serve per l'interpretazione, non
      per la classificazione.

  JSON_OUTPUT: {{"stereotype": <0 o 1>, "category": "<categoria>"}}
```

(d) **Zero-Shot Prompt 4 (Binary): Italian**

**Figure 3: Zero-Shot Binary Classification Prompts.**

**Table 10**
Pairwise agreement of the Main Task (our annotations).

| Ann_A | Ann_B | N | Cohen's $\kappa$ |
|---|---|---|---|
| *A1 | *A2 | 200 | 0.4505 |
| *A1 | *A3 | 200 | 0.5697 |
| *A1 | *A4 | 200 | **0.6139** |
| *A2 | *A3 | 200 | 0.5635 |
| *A2 | *A4 | 200 | **0.3891** |
| *A3 | *A4 | 200 | 0.4901 |

**Table 11**
Pairwise agreement of the Subtask (our annotations).

| Ann_A | Ann_B | N | Cohen's $\kappa$ |
|---|---|---|---|
| *A1 | *A2 | 80 | 0.555 |
| *A1 | *A3 | 96 | **0.594** |
| *A1 | *A4 | 114 | 0.495 |
| *A2 | *A3 | 79 | 0.479 |
| *A2 | *A4 | 84 | **0.390** |
| *A3 | *A4 | 100 | 0.564 |

## C. Additional Human Annotation Experiment

In this section, we describe an additional annotation experiment. By following the guidelines provided by the organizers, four Italian native speakers (co-authors of this paper) have annotated the 200 texts of the dev set. This work has been conducted to explore the subjective nature of the social phenomena the task is trying to model, i.e., identifying if a text contains a gender stereotype and of which kind. To quantify the agreement between annotators, we compute Cohen's $\kappa$. Table 10 reports the pairwise agreement for the main task, i.e., has_stereotype: YES/NO (binary), for the total of 200 instances, while Table 11 reports the agreement for the subtask, i.e, the stereotype categories (multiclass), for 68 texts.[5] Annotators are indicated with * to denote they are not the official annotators, but members of our team.

The agreement scores are consistently higher than 0.45, indicating a moderate agreement: the highest score is reached for the main task (0.61). We point out one notable exception: in both tasks, the same pair of annotators register the lowest agreement (0.39). In order to better understand what caused such misalignment, a collaborative approach was taken during a group meeting to share general insights, ensuring that any disagreements were addressed through discussions and ultimately resolved through consensus. In the following, we report considerations that emerged from this collective manual review.

We first highlight a clear tendency: one annotator assumed an over-identification position identifying an over-presence of stereotypes; while the other adopted the opposite, under-estimating their presence.

Example 1

**IT:** *Madonna che str0nza acida questa [USER] di questo profilo. Ma che cazz0 ci campi a fare?*

**EN:** *Holy cow, what a sour bitch this [USER] on this profile is. What the hell are you doing here?*

This frequent pattern and the **great subjectivity of the task** itself, prompted us to question what we are truly modeling, and, in relation to the real-world application, if we are interested in limiting false negatives or maximizing true positives.

We acknowledge the strong subjective and ideological nature of the task, which could lead to marking as stereotype texts that express a **dissonant belief or faulty stance** with respect to the annotator's perspective, or, similarly, to not properly recognize a stereotype if the annotator agreed with the position expressed. We recognize the crucial role of the context to support the annotation. Indeed, in several instances, the content was labeled as stereotypical strictly in **relation to the context** (which was sometimes provided and often times was not).

Example 2

**IT:** *[Commento a post dalla dicitura "Quale sarebbe la tua reazione se adesso scoprissi di essere incinta??"] Vorrei tantissimo una femminuccia xké ho tre bellissimi maschietti ma ho paura,paura del 4 maschio,ho paura del giudizio della famiglia...insomma non sarei felicissima*

**EN:** *[Comment on a post with the caption "What would be your reaction if you found out you were pregnant now?"] I would really like a girl because I have three beautiful boys but I'm scared, scared of the fourth boy, scared of the family's judgement...in short, I wouldn't be very happy*

[5] 68 were the common instances to compare where all annotators have labeled a category within the subtask.

In the absence of the preceding conversational or descriptive context, identifying a text as stereotypical becomes a very challenging (and arbitrary) task. In several cases, **utterances that were trying to counter a stereotype were expressing stereotypes themselves**: this particular but frequent scenario caused a different interpretation of the guidelines.

Example 3

**IT:** *Anche le figlie se si sposano e vanno in altra città, quando hanno i figli non le senti e non le vedi,e questione di carattere, donne ho uomini e lo stesso.*

**EN:** *Even daughters, if they get married and go to another city, when they have children you don't hear from them or see them, it's a question of character, women and men are the same.*

Moreover, the **use of irony and/or sarcasm** complicated stereotype identification, as stereotypical content may be expressed implicitly while simultaneously being challenged or mocked.

Example 4

**IT:** *[Commento ad articolo con titolo "Non esistono due cervelli, maschile e femminile. Il cervello è plastico e plasmato dall'ambiente".] Al Festivaletteratura di Mantova la neuroscienziata Martina Ardizzi, docente all'Università di Padova, ha spiegato che differenze considerate biologiche, come emozioni, abilità spaziali o matematiche, dipendono in realtà dalla socializzazione. ...però il cervello, quando è nella testa di una donna, diventa miracolosamente... migliore!*

**EN:** *[Comment on an article titled "There are not two brains, male and female. The brain is plastic and shaped by the environment."] At the LiteratureFestival in Mantua, neuroscientist Martina Ardizzi, a professor at the University of Padua, explained that differences considered biological, such as emotions, spatial abilities, or mathematical abilities, are actually the result of socialization. ...but the brain, when it's in a woman's head, miraculously becomes... better!*

Regarding the categories proposed by the guidelines, we noticed that (1) occasionally the available categories did not cover the angle of the stereotype expressed, and (2) the difficulty of assigning only one category when overlapping dimensions were present. These two factors could potentially lead to inconsistent annotations, inconsistent data, and ultimately inconsistent models. Finally, in multiple samples, the stereotype identification was rooted in the literal reading of the text, but rather it was implicit: recognizing it involved several logical inferences. On this aspect, it would be an added value for the data to ask annotators to (1) highlight the span of text expressing the stereotype (i.e., a rationale), and (2) to write down the **implicit belief underlying the text**, as performed by Sap et al. [33].

Example 5

**IT:** *Ma trovo vergognoso strumentalizzare questa storia per farne l' ennesimo dogma ideologico per giustificare la violenza con cui le neo femministe rivendicano determinati torti. Accusando TUTTI i maschi indistintamente...*

**EN:** *But I find it shameful to exploit this story to make it yet another ideological dogma to justify the violence with which neo-feminists claim certain wrongs. Accusing ALL men without distinction...*

Overall, these observations confirm that identifying and categorizing stereotypes is a highly challenging task even for humans. Consequently, expecting models to learn this phenomenon reliably from imperfect and subjective annotations represents a sensitive direction.

We release our additional annotations publicly to support future research. For example, these annotations could be used to estimate a more robust GS Value by aggregating binary gender-stereotype labels from eight annotators instead of four, thereby better capturing the diversity and subjectivity of the task. Additionally, agreement on the development set between the original annotators (Table 8) was slightly higher than the agreement among the new annotators (Table 10). It would be interesting to investigate whether this difference affects model behavior and performance, and if so, to what extent.[6]

[6]The additional annotations are available at: https://github.com/grecosalvatore/StereoBusters-GSI-Detect-Evalita2026.