GSI:detect at EVALITA 2026: Overview of the Task on Detecting Gender Stereotypes in Italian*

Gloria Comandini[1,*], Manuela Speranza[2], Sofia Brenna[2,3], Davide Testa[2,4], Stefania Cavagnoli[5] and Bernardo Magnini[2]

[1]*Italian Institute of Germanic Studies (IISG), Rome, Italy*
[2]*Fondazione Bruno Kessler (FBK), Trento, Italy*
[3]*Free University of Bozen-Bolzano, Bolzano, Italy*
[4]*University of Rome La Sapienza, Rome, Italy*
[5]*University of Rome Tor Vergata, Rome, Italy*

Abstract

GSI:detect is a new shared task for the recognition and classification of gender stereotypes (GSs) presented at EVALITA 2026. The task adopts a perspectivist approach in order to enhance the high subjectivity of GS recognition and analysis on a dataset of challenging short texts in Italian. GSI:detect is organized in: A) a Main Task (GS Detection) in which systems have to assign to a text the GS value, a numerical score that quantifies the extent to which a given text exhibits or refers to a GS; B) an optional Subtask (GS Classification) in which systems, given six pre-defined categories (e.g. role, relational, etc.) must assign one to each text. Seven teams from academic and non-academic environments took part in the challenge, with a total of 50 submitted runs for the Main Task and a total of 43 submitted runs for the optional Subtask. We present here first an overview of the GSI:detect task, the dataset and the evaluation criteria, then outline and discuss the participants' results. **Content warning**: Examples taken from the GSI:detect dataset may contain sensitive, offensive, or otherwise distressing content.

Keywords

gender stereotypes, perspectivism, linguistic resource, evaluation, LLMs,

1. Introduction and Motivation

The GSI:detect Task, organised within EVALITA 2026 [1], aims to take one step further in the state-of-the-art detection of gender stereotypes (GSs)[1]. Gender stereotypes have recently been the object of extensive research in the context of automatic recognition of GSs in misogynistic hate speech [3, 4, 5] and also as far as large language models' production of stereotyped and biased material is concerned in text generation [6, 7] and translation [8].

However, with GSI:detect we aimed to first expand our focus beyond the context of hate speech, because GSs can also appear in non-hateful communication (even as a compliment: "women's nurturing nature will save the world!"). In fact, GSs are sometimes produced even by their own targets (e.g., a man who says "You know, men only have one thing in mind", or a woman who says "I failed in the math exam. Oh well, girls don't do well in math anyway, ah-ah"), who have internalized these biased views on gender.

Secondly, we want to underline that the recognition and therefore the analysis of stereotypes can be deeply subjective, as seen for example in the low inter-annotator agreement (IAA) (0.41, Cohen's

EVALITA 2026: 9th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Feb 26 – 27, Bari, IT
*Corresponding author
comandini@studigermanici.it (G. Comandini); manspera@fbk.eu (M. Speranza); sbrenna@fbk.eu (S. Brenna); dtesta@fbk.eu (D. Testa); stefania.cavagnoli@uniroma2.it (S. Cavagnoli); magnini@fbk.eu (B. Magnini)
https://huggingface.co/GloriaComandini (G. Comandini); https://linktr.ee/davide.testa (D. Testa)
0000-0003-3406-2819 (G. Comandini); 0009-0001-3748-1448 (S. Brenna); 0009-0002-2489-5323 (D. Testa); 0000-0003-1677-6455 (S. Cavagnoli); 0000-0002-0740-5778 (B. Magnini)

[1]We intend GSs as socially constructed beliefs about the 'appropriate' roles, behaviours, and appearances of a person regarding their gender; although GSI:detect focuses on GSs about men and women for the sake of simplicity, we describe gender as "a nonbinary construct" [2] and we intend to include GSs regarding non-binary people in future research.

k) in the recognition of racist stereotypes by [9]. This situation is common to most highly subjective NLP tasks, such as hate speech recognition [10], irony detection [9] or sentiment analysis applied to complex texts (e.g. newspaper articles or latin poetry) [11, 12, 13], where annotators judgments can be influenced by several factors, such as agreement or disagreement with a statement, the fact that the annotator belongs to the group targeted by hate or stereotypes [10], personal opinions [14] and interiorized biases [15, 16].

For all these reasons, we decided to adopt a Perspectivist Approach [17, 15, 18, 19, 16], in order to treat the inherent diversity in judgments and perceptions as a valuable asset, and not as a problem to solve. In fact, it has been observed that the classical methods of majority voting used to create gold-standard corpora may result in the suppression of perfectly valid points of view, which may even add interesting information about a task [14].

2. Task Description

GSI:detect aims to evaluate systems' ability to detect and classify GSs across different kinds of short texts; it includes both a compulsory main task and an optional sub-task.

2.1. Main Task: Gender Stereotype Detection

GS value	Example
0	Non comprendo come si possano paragonare due fenomeni, gravissimi entrambi e concordo, come femminicidi e morti sul lavoro. *(I don't understand how one can compare two phenomena, both very serious and I agree, such as feminicides and deaths at work.)*
0.25	Tenete duro ancora qualche giorno e i vostri fidanzati partiranno in vacanza con le loro mogli. *(Hold on for a few more days and your boyfriends will be going on vacation with their wives.)*
0.50	Io rimango dell'idea che un figlio ha sempre bisogno della sua mamma, anche per dire buongiorno e buona notte. E la mamma idem. Soprattutto la mamma *(I still think an [adult] child always needs his/her mother, even to say good morning and good night. And the mother too. Especially the mother*
0.75	[Commento ad articolo di giornale dal titolo "Negli Usa quasi un manager su due è donna. In Italia meno di 1 su 3"] Infatti il Made usa va' peggio del Made italy *([Comment on a newspaper article titled "In the US, almost one in two managers is a woman. In Italy, less than one in three"] In fact, Made in USA is doing worse than Made in Italy)*
1	[Rivolto a una utente donna] fatevi voi una doccia e copritevi. Le donne vere si coprono. Gli animali vanno in giro nudi. *([Addressed to a female user] Take a shower and cover up. Real women cover up. Animals walk around naked.)*

Table 1
Sample texts annotated with GS values.

Given a short text, the main task requires systems to assign to it a numerical score, the GS value, that quantifies the degree to which the text contains or refers to a gender stereotype.

This is formulated as a regression task in which GS values are real numbers in the range [0,1], where 1 indicates the maximum degree of stereotypical content (the granularity is fixed at two decimal places). A few examples of short texts annotated with different GS values are reported in Table 1.

Notice that we consider two types of texts:

- NO CONTEXT: texts that can be understood without any additional contextual information (see the first three examples in Table 1);
- WITH CONTEXT: texts that are not self contained and are therefore enriched with contextual information in the form of standardized metadata (see the last two examples in Table 1).

2.2. Subtask: Gender Stereotype Classification

The subtask on GS classification is formulated as a multi-class classification task where, given a short text, systems are required to assign a GS category to that text; assignment must be performed for each

single text, independently of the GS value assigned by the systems in the main Task. Participation in the subtask was not compulsory; still, participants were strongly encouraged to submit their results to allow for a more comprehensive evaluation of the phenomenon.

Category	Example
Role	Cento uomini possono creare un accampamento, ma serve una donna per fare una casa. *(A hundred men can build a camp, but it takes a woman to make a home.)*
Personality	Sentivo qualcosa di speciale e sai, una donna non sbaglia mai le sensazioni. *(I felt something special and you know, a woman never mistakes her feelings.)*
Competence	[Commento ad articolo con titolo "La pilota della British Airways ubriaca in volo: cacciata dall'aereo, aggredisce pure i poliziotti"] Come si possono affidare le sorti di un aereo ad una donna? ....scherzo, naturalmente..." *([Comment on an article titled "British Airways pilot drunk on flight: kicked off plane, she even attacks police"] How can you trust a plane's fate to a woman? ....just kidding, of course...")*
Physical	"Oppure c'hanno le 5^, vanno in giro scollate come i manifesti messi d'inverno,e poi se rimani ""attirato"" dalle loro protuberanze ci rimangono male Povere cucciole." *("Or they are a size D, they walk around with low-cut clothes like winter posters, and then if you get ""attracted"" by their protuberances, they get upset. Poor little things.")*
Sexual	[Rivolto a una utente donna] fatevi voi una doccia e copritevi. Le donne vere si coprono. Gli animali vanno in giro nudi. *([Addressed to a female user] Take a shower and cover up. Real women cover up. Animals walk around naked.)*
Relational	[Commento a meme con testo "Aspettavo che mi mandassi tu un messaggio" e sotto l'immagine di un uomo vestito da principessa] Tipico post da zitella *([Comment on a meme with the text "I was waiting for you to text me" and underneath a picture of a man dressed as a princess] Typical spinster post)*

Table 2
Examples of texts assigned to the different GS categories.

The classification we propose, developed to capture the variety of ways in which stereotypes manifest in language and to support both linguistic analysis and automatic detection tasks, foresees the following GSs typologies (examples of texts assigned to the different categories are provided in Table 2)[2]:

- ROLE stereotypes: social and cultural expectations about what women and men should do and about how they should be;
- PERSONALITY stereotypes: emotional and behavioural traits assigned to men and women based on their gender;
- COMPETENCE stereotypes: generalized judgments of a person's abilities based on their gender;
- PHYSICAL stereotypes: expectations about the physical appearance of men and (especially) women, and all aspects of personal care in general;
- SEXUAL stereotypes: attitude and behaviour that men and women should have regarding sexuality;
- RELATIONAL stereotypes: the way in which women and men should behave in interpersonal/sentimental relations.

For a more accurate understanding of the task, participants were able to refer to the official guidelines for stereotype classification, which were followed during the manual annotation of the dataset. [3]

3. Dataset

The GSI:detect dataset[4] consists of 1,010 short written texts in Italian (for a total of 52,118 tokens), collected from social media and informative websites.

The texts have been manually collected from a diverse array of online spaces to provide a balanced representation of formal and informal written Italian, a variety that also allows us to explore the theme of GSs in different contexts. In fact, they include excerpts from information websites, as well as users' comments from Facebook, Instagram and Reddit pages and groups discussing both gender-related issues

[2]Note that this taxonomy is not intended to be fully exhaustive, as it is derived from an abstraction over the direct observation of the examples contained in the dataset.
[3]The annotation guidelines are available for download at this link.
[4]The GSI:detect dataset is distributed under a CC BY-NC-SA 4.0 Licence. The dataset is publicly available at this link. The distributed dataset includes, besides the GS values, also the individual, non-aggregated labels assigned by all annotators, in order to enable systems to learn from annotator disagreement [19].

and more generic topics (e.g. feminist influencers, pick-up artists, "mom influencers", parodic pages, math groups, gossip pages, major Italian newspapers, etc.),

Furthermore, the dataset includes GSs about not only women, but also men, or both, including texts that clearly express GSs as well as non-stereotypical examples, as well as cases where recognizing GSs may be a question of personal opinion and sensitivity.

3.1. Data Annotation

The GSI:detect dataset has been manually annotated by four expert annotators.

GS Value Annotation. Although all four annotators were expert and followed the annotation guidelines specifically created for GSI:detect (as mentioned in Section 2.2), the inherent subjectivity of the task inevitably introduced a certain level of disagreement. Following the perspectivist approach introduced in Section 1, we opted for merging all annotations into a numerical GS value, rather than selecting a binary label obtained through annotation aggregation on the basis of majority voting. This choice aligns with recent findings which indicate that leveraging disagreement is more convenient than effortlessly trying to eliminate it [15, 16].

The overall annotation procedure consists of two steps: (i) each annotator manually assigns, for each short text, a binary label *yes*/*no* indicating whether or not the text contains or refers to a GS; (ii) the final GS value is computed by combining the four individual annotations. The underlying assumption is that full IAA (all four annotators choose the label *no* or the label *yes*) corresponds to the endpoints of the continuum, while disagreement between annotators indicates intermediate GS values, such as 0.25 (three *no* labels and one *yes* label), 0.5 (two *yes* labels and two *no* labels), and 0.75 (three *yes* labels and one *no* label).

The overall IAA between the four annotators on the choice of the *yes* or *no* label is 0.61 (Fleiss' k), which is a moderate agreement, common in highly subjective tasks [20], as seen in Section 1.

GS Category Annotation. When annotators identify a text as containing or referring to a GS, they additionally assign it to one of the six GS categories, following the classification outlined in the annotation guidelines and summarized in Section 2.2. Due to the more explorative nature of GS category annotation compared to the preceding annotation level, we adopted a more conventional strategy: in case of disagreement between the annotators, a single category is determined by majority vote, with ties resolved by a GS expert acting as a super-judge (required for 6% of the dataset). Regarding the GS category annotation, the four experts scored moderate agreement [20], with a IAA of 0.61 (Fleiss' k).

3.2. Data Statistics: Test and Development Split

The complete dataset is divided as follows: 80% is allocated to the test set for the official evaluation and ranking of participant systems, while the remaining 20% constitutes the development data (dev set) (refer to Table 3 for more details). These proportions were chosen to balance the need for adequate data for model tuning with the goal of maintaining a larger and more representative test set for the final evaluation.

Table 3 reports also detailed information about the size of the dataset in terms of tokens[5]. In both the dev and test sets, approximately 58% of texts are WITH CONTEXT and 41% are NO CONTEXT (see Section 2.1). In the creation of the dev and test sets, particular care was taken to ensure a balanced distribution of examples across both subsets, as shown in Table 4a. Therefore, the

	Dev set	Test set	Total
WITH CONTEXT texts	82	323	405
NO CONTEXT texts	118	487	605
All Texts	200	810	1010
Tokens	10,055	42,063	52,118
Av. Length	50.27	51.93	51.6

Table 3: Dataset's statistics.

[5]The token count was computed using the Italian rule-based tokenizer included in the *spaCy* library (https://spacy.io, version 3.8.7) as part of the *it_core_news_sm* linguistic model. The average length of texts is 51.6 tokens.

GS value	Dev set	Test set	Total	Total%
0	60	242	302	29.90%
0.25	25	84	109	10.79%
0.50	27	85	112	11.09%
0.75	25	105	130	12.87%
1	63	294	357	35.35%
	200	810	1010	

(a) Dataset distribution by GS value.

Category	Dev set	Test set	Total	Total%
Role	30	107	137	13.56%
Personality	29	108	137	13.56%
Competence	34	120	154	15.25%
Physical	20	90	110	10.89%
Sexual	14	72	86	8.52%
Relational	13	71	84	8.32%
GS value = 0	60	242	302	29.90%
	200	810	1010	

(b) Dataset distribution by GS category.

Table 4
Dataset distribution statistics.

split preserves the original distribution of GS
values, thereby guaranteeing a consistent representation of varying degrees of stereotypicality in both subsets. A comparable level of balance is also observed for GS categories (see Table 4b).

This careful selection ensures that both subsets are representative of the overall GSI:detect dataset, preventing unintended biases in the distribution of categories.

4. Evaluation

Evaluation in GSI:detect is designed to reflect the specific nature of both the main task on GS detection, formulated as a regression problem, and the subtask on GS classification, formulated as a multi-class classification problem. Accordingly, we adopt task-specific evaluation criteria to ensure meaningful comparison and reliable system ranking.

GS Detection. Participant systems' performance is assessed using a score derived from the *Mean Squared Error* (MSE), which measures the average squared distance between predicted and annotated GS values, penalizing larger deviations more heavily. To improve interpretability and comparability across systems, the MSE is normalized with respect to the variance of the target distribution (*Normalized Mean Squared Error*, NMSE). Since lower NMSE values indicate better performance, we further transform this quantity into a bounded score defined as $\frac{1}{1+\text{NMSE}}$ so that higher values correspond to better predictive accuracy. This formulation enables an intuitive ranking of systems while preserving the relative performance differences.
In addition to the scores described above, we also report the *Concordance Correlation Coefficient* (CCC) as a complementary measure of agreement between predictions and reference values, capturing both correlation and potential systematic bias.

GS Classification. GS classification is evaluated using the *F1* score, which combines precision and recall into a single performance indicator.[6] To account for possible class imbalance while maintaining sensitivity to per-class behaviour, we report both *Macro F1*, which weights all categories equally regardless of their frequency, and *Micro F1*, which reflects the overall performance at instance-. While both measures are reported for analysis purposes, *Micro F1* is adopted as the official metric for system ranking, as it provides an instance-level estimate of overall classification performance.

Baselines. For both GS detection and classification, participant systems' performance is compared against a set of four baselines[7]. They include a simple heuristic baseline obtained by assigning a constant

[6]As annotators assigned a GS category only to texts containing or referring to a stereotype, texts with GS value = 0 (i.e. no-no-no-no annotation) don't have a category. In the GS classification subtask, systems' performance has been therefore evaluated on 568 out of 810 texts.
[7]Baselines were computed only for the zero-shot track.

Team Name	GS Detection				GS Classification			
	Zero-shot	Few-shot	Fine-Tuning	Encoder-only	Zero-shot	Few-shot	Fine-Tuning	Encoder-only
DIAG-Sapienza [21]	1	1	-	4	1	1	-	4
Festa [22]	5	5	-	5	5	5	-	5
MINDS [23]	-	2	-	-	-	-	-	-
Prisma [24]	5	1	-	-	5	1	-	-
StereoBusters [24]	5	5	-	-	5	5	-	-
Tiz [25]	5	-	5	-	5	-	-	-
VellaAsta[8]	-	-	-	1	-	-	-	1
Total	21	14	5	10	21	12	0	10

Table 5
Number of runs submitted by each team for each track in the GS Detection and GS Classification tasks.

GS value of 0.5 and a random GS category prediction (B1 in Table 6), as well as the performance of two LLMs, namely *Qwen3-14B* (B2), and *GPT-5-nano-2025-08-07*. The performance of *GPT-5*, in particular, is evaluated under two prompting configurations: one where GS detection and classification are jointly addressed within a single prompt (B3), and one where the tasks are solved independently with two separate prompts (B4).

5. Participants

The GSI:detect shared-task attracted the participation of seven teams, coming both from academic and non-academic environments. Participant were allowed to submit multiple runs for each task, exploring different model architectures, prompting strategies, and technical configurations. The evaluation campaign included four different tracks for both the main task and the subtask: *zero-shot*, *few-shot*, *fine-tuning* of LLMs, and *encoder-only models*. Not all teams submitted runs to all tracks and tasks. Table 5 presents the participating teams and reports the number of runs submitted by each of them for each track in both GS Detection and GS Classification tasks. Further analysis of the impact of the different runs and of their behavior across the two tasks, leading to different performance trends, is presented in Section 6. For a detailed description of the individual system configurations and methodologies associated with each run, we refer the reader to the participants' reports, cited in Table 5.

6. Results and Discussion

The results are organized according to the four main tracks of the shared task, under which participants submitted multiple runs for both the main task and the subtask. This experimental setting results in eight distinct rankings, corresponding to the combination of the two tasks and the four tracks.

6.1. Zero-Shot Track

Tables 6 and 7 report the scores obtained by the participating systems on the main task and on the subtask, respectively, in the zero-shot setting. Overall, in both tasks several participant systems outperform the proposed baselines, which are positioned approximately in the middle of the ranking and therefore act as a rough boundary between higher- and lower-performing approaches in this track.

For our main task, the best-performing system is submitted by the DIAG-Sapienza team [21], based on *GPT-5* in a configuration that generates together the predicted GS value, the GS category plus a one-sentence explanation within the same model's call (see run 1 in Table 6 and 7). Interestingly, while this model's configuration achieves the top rank in GS Detection, its performance drops by seven positions in GS Classification, highlighting the increased difficulty of fine-grained stereotype categorization and

[8]As the VellaAsta team correctly and timely submitted the output of their system for official evaluation, we present their results even if they did not submit a report describing the system.

of the underlying reasoning process. Nevertheless, in the main task this configuration substantially outperforms our *GPT-5 nano* baseline (i.e., B4), whereas in the classification task the performance gap becomes much smaller (0.58 Micro F1 for DIAG-Sapienza vs. 0.53 for B4).

Another team showing consistently strong performance in this track is StereoBusters [26], which evaluates several models of different sizes and configurations. In particular, *Llama 3.3* (70B) achieves competitive results in GS Detection, approaching the performance of the closed-source *GPT-5* model by DIAG-Sapienza (0.63 vs. 0.70). Moreover, *Llama 3.3* consistently maintains a high level of performance also in GS Classification (0.64 Micro F1), outperforming both proprietary models (*GPT-* and *Gemini*-based) and the systems submitted by the other teams. Additionally, the ensemble strategy adopted by this team (i.e., run 4), combining the predictions of four LLMs to determine the stereotype category, proves particularly effective in the classification task, achieving the highest Micro F1 score (0.646) in this subtask, with an improvement of roughly ten positions in the ranking compared to the main task.

The Festa team [22] explores multiple configurations of *Gemini 2.5 Flash*, outperforming most of our baselines in both tasks. For GS Detection, the best-performing configuration (i.e., run 4) relies on an English prompt, despite the Italian nature of the data, combined with negative constraints aimed at minimizing false positives. By contrast, the use of Chain-of-Thought prompting both in English and Italian (runs 1 and 3) appears less beneficial in this task, yielding scores approximately seven points lower than run 4 (0.62 vs. 0.55). In the classification task, however, Chain-of-Thought prompting proves more effective, with performance comparable to *Llama 3.3* (70B) by StereoBusters. This suggests that explicitly encouraging intermediate reasoning steps may support finer-grained category discrimination.

Finally, in this setting, the systems submitted by the Tiz [25] and Prisma [24] teams show substantially lower performance, consistently falling below the LLM-based baselines (i.e., B2, B3, and B4), with this performance gap remaining consistent across both the main task and the subtask. In particular, Prisma explores multiple configurations of *Claude 3.5 Sonnet* based on different annotator personas and their aggregation; however, these highly polarized configurations do not yield competitive results, possibly due to a mismatch between the induced persona biases in the configuration and the annotator perspectives underlying the dataset. This could also suggest that strong persona conditioning, when misaligned with the annotator distributions of the dataset, may introduce biases that increase the distance from the target judgments.

Overall, what emerges from this scenario is that, although *GPT-5* (DIAG-Sapienza) represents the best-performing system in the main GS Detection task, open-source models with different configurations and modeling strategies are able to closely approach, match, or even surpass closed-source systems across both tasks. In particular, StereoBusters' open-source models narrow the gap with *GPT-5* in the main task and outperform both proprietary systems and our baselines in GS Classification, regardless of whether large-scale (*Llama 3.3* 70B) or mid-sized (*Gemma 3* 12B) models are used. This trend suggests – and confirms – that stereotype categorization, which inherently involves subjective interpretation and nuanced reasoning, benefits less from pure model scale and more from diversified modeling choices and decision aggregation strategies.

6.2. Few-shot Track

Most of the teams participating in the zero-shot track also submitted systems to the few-shot track, except for Tiz team, which participated exclusively in the former setting. This semi-overlap allows for a direct comparison of model behaviour across the two settings, highlighting how the same architectures can exhibit substantially different performance when provided with in-context examples.

According to Table 8, for the GS Detection task, the DIAG-Sapienza team again achieves the best performance with *GPT-5* (i.e., run 1, four-shot), followed by the family of *Gemma 3* models (12B and 27B) evaluated by StereoBusters. However, the systems developed by both teams are not able to maintain the same trend in the subtask (see Table 9), dropping several positions in the ranking, except for *Gemma 3* (27B). This finding suggests that, within the same model family, model size may positively influence performance across both tasks.

Notably, in contrast with the zero-shot setting, the Prisma team shows a marked improvement: its

Team name	Model	Run id	**1/(1 + *NMSE*)** ↑	MSE ↓	NMSE ↓	CCC ↑
DIAG-Sapienza	GPT-5	1	0.70	0.077	0.43	0.78
StereoBusters	Llama_3.3_70B	1	0.63	0.11	0.60	0.60
StereoBusters	Llama_3.3_70B	2	0.62	0.11	0.61	0.59
Festa	Gemini_2.5_Flash	4	0.62	0.11	0.61	0.70
Festa	Gemini_2.5_Flash	5	0.62	0.11	0.62	0.69
StereoBusters	Gemma_3_12B	0	0.61	0.11	0.64	0.56
Festa	Gemini_2.5_Flash	2	0.61	0.11	0.64	0.69
BASELINE	GPT-5 nano	B4	0.61	0.11	0.64	0.60
Tiz	Gemma_3_12B	5	0.59	0.12	0.69	0.63
BASELINE	GPT-5 nano	B3	0.59	0.12	0.69	0.57
StereoBusters	Panel_4LLMS	4	0.57	0.13	0.75	0.62
StereoBusters	Panel_4LLMs	3	0.56	0.14	0.77	0.60
Festa	Gemini_2.5_Flash	1	0.56	0.14	0.79	0.63
Festa	Gemini_2.5_Flash	3	0.55	0.14	0.80	0.62
BASELINE	Qwen-3_14B	B2	0.54	0.15	0.84	0.46
BASELINE	N/A	B1	0.50	0.18	1.01	0
Tiz	Gemma_3_12B	3	0.48	0.19	1.06	0.55
Tiz	Gemma_3_12B	2	0.48	0.19	1.07	0.54
Tiz	Gemma_3_12B	1	0.47	0.20	1.12	0.52
Tiz	Gemma_3_12B	4	0.43	0.24	1.33	0.42
Prisma	Claude_3.5_Sonnet	2	0.33	0.36	2.04	0.15
Prisma	Claude_3.5_Sonnet	4	0.32	0.38	2.11	0.08
Prisma	Claude_3.5_Sonnet	1	0.32	0.38	2.11	0.09
Prisma	Claude_3.5_Sonnet	5	0.31	0.40	2.23	0.09
Prisma	Claude_3.5_Sonnet	3	0.31	0.40	2.24	0.08

Table 6
Results for *zero-shot* track [Main Task].

Team name	Model	Run id	**F1 Micro** ↑	F1 Macro ↑
StereoBusters	Panel_4LLMS	4	0.65	0.64
StereoBusters	Llama_3.3_70B	1	0.64	0.64
StereoBusters	Panel_4LLMs	3	0.64	0.63
StereoBusters	Llama_3.3_70B	2	0.63	0.63
Festa	Gemini_2.5_Flash	1	0.60	0.56
Festa	Gemini_2.5_Flash	5	0.60	0.55
Festa	Gemini_2.5_Flash	4	0.59	0.55
Festa	Gemini_2.5_Flash	2	0.59	0.54
DIAG-Sapienza	GPT-5	1	0.58	0.52
Festa	Gemini_2.5_Flash	3	0.57	0.52
StereoBusters	Gemma_3_12B	0	0.54	0.53
BASELINE	GPT-5 nano	B4	0.53	0.52
BASELINE	GPT-5 nano	B3	0.52	0.50
BASELINE	Qwen-3_14B	B2	0.39	0.39
Tiz	Gemma_3_12B	2	0.23	0.23
Tiz	Gemma_3_12B	3	0.23	0.23
Tiz	Gemma_3_12B	1	0.22	0.22
Prisma	Claude_3.5_Sonnet	1	0.22	0.29
Tiz	Gemma_3_12B	5	0.22	0.22
Tiz	Gemma_3_12B	4	0.21	0.21
BASELINE	N/A	B1	0.18	0.18
Prisma	Claude_3.5_Sonnet	4	0.18	0.23
Prisma	Claude_3.5_Sonnet	2	0.17	0.24
Prisma	Claude_3.5_Sonnet	3	0.13	0.18
Prisma	Claude_3.5_Sonnet	5	0.12	0.17

Table 7
Results for *zero-shot* track [Subtask].

system (i.e., run 6) benefits significantly from this configuration based on selecting 5 examples per category through stratified sampling. In this case, *Claude 3.5 Sonnet* appears to react positively both to the presence of in-context examples and to the inclusion of additional information derived from the annotation guidelines. This effect is even more evident in the GS Classification task (Table 9), where the same configuration outperforms all other submissions by approximately ten points (0.71).

Yet, *Gemini 2.5 Flash* exhibits a more heterogeneous behavior: while a 14-shot configuration leads to one of the worst performance in GS Detection, the presence of examples proves beneficial for reasoning on GS Classification, but the improvement over the zero-shot setting remains marginal (0.67 vs. 0.60).

A different trend is observed for the MINDS team [23], which evaluates *Qwen 2.5* (14B) exclusively in the few-shot setting for GS Detection. Two configurations (runs 1 and 2), each using 20 in-context

Team name	Model	Run id	Shots	$1/(1+NMSE)$ ↑	MSE ↓	NMSE ↓	CCC ↑
DIAG-Sapienza	GPT-5	1	4	0.68	0.08	0.46	0.76
StereoBusters	Gemma_3_27B	2	5	0.64	0.10	0.55	0.65
StereoBusters	Gemma_3_12B	0	5	0.63	0.10	0.59	0.60
StereoBusters	Gemma_3_12B	1	5	0.61	0.11	0.63	0.61
Prisma	Claude_3.5_Sonnet	6	35	0.59	0.12	0.70	0.63
StereoBusters	Panel_4LLMS	4	5	0.59	0.13	0.71	0.65
MINDS	Qwen2.5_14B	1	20	0.58	0.13	0.72	0.46
Festa	Gemini_2.5 Flash	5	non-fixed	0.57	0.13	0.75	0.63
Festa	Gemini_2.5 Flash	1	12	0.56	0.14	0.77	0.62
MINDS	Qwen2.5_14B	2	20	0.56	0.14	0.77	0.42
Festa	Gemini_2.5 Flash	2	10	0.56	0.14	0.78	0.62
Festa	Gemini_2.5 Flash	4	14	0.55	0.14	0.82	0.60
StereoBusters	Panel_4LLMs	3	5	0.55	0.15	0.82	0.56
Festa	Gemini_2.5 Flash	3	6	0.50	0.18	1.00	0.55

Table 8
Results for *few-shot* track [Main Task].

Team name	Model	Run id	Shot	F1 Micro ↑	F1 Macro ↑
Prisma	Claude_3.5_Sonnet	6	35	0.71	0.61
Festa	Gemini_2.5 Flash	4	14	0.67	0.67
StereoBusters	Gemma_3_27B	2	5	0.67	0.66
Festa	Gemini_2.5 Flash	5	non-fixed	0.67	0.66
StereoBusters	Panel_4LLMs	3	5	0.66	0.65
StereoBusters	Panel_4LLMS	4	5	0.66	0.65
Festa	Gemini_2.5 Flash	1	12	0.66	0.65
Festa	Gemini_2.5 Flash	2	10	0.65	0.63
Festa	Gemini_2.5 Flash	3	6	0.64	0.63
DIAG-Sapienza	GPT-5	1	4	0.61	0.55
StereoBusters	Gemma_3_12B	0	5	0.57	0.57
StereoBusters	Gemma_3_12B	1	5	0.56	0.56

Table 9
Results for *few-shot* track [Subtask].

examples, extract logits from the model and feed them into downstream statistical predictors (i.e., Linear Regression and KNN) to estimate the numerical GS value. This hybrid approach, however, remains substantially poor, with a gap of nearly ten points compared to the top model in this track (0.58 vs. 0.68).

Moreover, the trend shown here closely mirrors the zero-shot setting one, further confirming that performance gains are driven less by model scale alone and more by how in-context examples are selected, structured, and integrated into the reasoning process. In support of this observation, we can observe that i. *GPT-5* achieves the highest performance in GS Detection, yet its performance drops sharply in GS Classification (ten positions), despite the use of a four-shot prompt, ii. the perspectivist approach adopted by Prisma in the GS Classification subtask, combined with a careful selection of representative examples for each category, proves particularly effective, iii. in the main task, an open-source mid-sized model such as *Gemma 3* (27B) by StereoBusters is able to achieve performance comparable to that of *GPT-5* (DIAG-Sapienza).

6.3. Fine-tuning

Although potentially interesting for comparison with the other tracks, this setting was explored only by the Tiz team, exclusively through different configurations of the same model (i.e., *Gemma 3* 12B), and only for the main task. Nevertheless, the results obtained in this setup are highly promising, as the best-performing system (i.e., run 2) achieves a score of 0.64, ranking behind the best zero-shot and few-shot systems by only 6 and 2 points, respectively. Moreover, it is worth emphasizing that these fine-tuning results were obtained using an open model, which here is nearly matching the performance of a state-of-the-art proprietary model (i.e., *GPT-5*) under both of the aforementioned settings.

Team name	Model	Run id	$1/(1+NMSE)$ ↑	MSE ↓	NMSE ↓	CCC ↑
Tiz	Gemma_3_12B	2	0.64	0.10	0.57	0.63
Tiz	Gemma_3_12B	4	0.62	0.11	0.60	0.58
Tiz	Gemma_3_12B	1	0.62	0.11	0.61	0.61
Tiz	Gemma_3_12B	5	0.62	0.11	0.61	0.57
Tiz	Gemma_3_12B	3	0.60	0.12	0.67	0.52

Table 10
Results for *fine-tuning (LLMs)* track [Main Task].

Team name	Model	Run id	$1/(1+NMSE)$ ↑	MSE ↓	NMSE ↓	CCC ↑
Festa	UmBERTo	1	0.56	0.14	0.78	0.31
Festa	UmBERTo	2	0.56	0.14	0.78	0.31
Festa	UmBERTo	5	0.56	0.14	0.78	0.31
Festa	UmBERTo	4	0.55	0.14	0.81	0.36
Festa	UmBERTo	3	0.54	0.15	0.85	0.32
VellaAsta	tum-nlp/bertweet-sexism	1	0.49	0.19	1.05	0.24
DIAG-Sapienza	RoBERTa	4	0.47	0.20	1.13	0.37
DIAG-Sapienza	RoBERTa	3	0.46	0.21	1.16	0.37
DIAG-Sapienza	RoBERTa	1	0.45	0.21	1.21	0.36
DIAG-Sapienza	RoBERTa	2	0.45	0.22	1.24	0.36

Table 11
Results for *encoder-only models* track [Main Task].

6.4. Encoder-only models

The use of encoder-only models represents a valuable reference for analyzing discrimination-oriented tasks. Although such systems can often achieve performance comparable to or even better than LLMs, especially when fine-tuned on the target task [27], this trend does not emerge in our track as proved by the Festa, DIAG-Sapienza, and VellaAsta teams, reported in Tables 11 and 12. Overall, their results fall below our baselines on the main task and consistently find themselves in the lowest positions in the GS classification subtask. When comparing models within the same track, the BERT-based *UmBERTo* (Festa) achieves the best performance, with a clear margin over *RoBERTa* model (DIAG-Sapienza) and *bertweet-sexism*[9] model (VellaAsta) likely due to its Italian-specific tokenization and masking strategies. Conversely, the *tum-nlp/bertweet-sexism* model outperforms *RoBERTa* in the GS detection task, as it has been fine-tuned for sexism detection on Twitter data. However, this specialization does not appear sufficient to achieve competitive performance on the more fine-grained GS classification task.

Overall, the consistently lower performance of encoder-only models across both tasks suggests that the limitations observed are not merely due to model capacity, but rather to architectural constraints. Unlike LLM-based systems, encoder-only models lack an explicit generative and reasoning component, which appears crucial for capturing the contextual, interpretative, and often implicit nature of gender stereotypes. While task-specific pre-training or fine-tuning (e.g., sexism detection on social media) can provide advantages in coarse-grained detection, as observed for *bertweet-sexism* model (VellaAsta Team) in the main task, such specialization does not translate into robust performance on fine-grained GS classification. This highlights the difficulty for encoder-only architectures to model subjective and socially grounded distinctions without access to richer contextual reasoning mechanisms.

7. Conclusions

In this paper, we presented the EVALITA shared task GSI:detect, which focuses on the automatic identification and classification of gender stereotypes in Italian. The task was structured into two subtasks: (i) a main task targeting the detection of GSs, and (ii) a fine-grained subtask aimed at classifying the type of GS expressed in the text. The dataset was constructed by explicitly leveraging disagreement among human annotators, to preserve the intrinsic subjectivity of the phenomenon rather than enforcing a single, fully convergent label. This design choice allows the benchmark to better

[9]This model can be found in the Huggingface Hub here.

Team name	Model	Run id	**F1 Micro ↑**	F1 Macro ↑
Festa	UmBERTo	1	0.52	0.50
Festa	UmBERTo	2	0.49	0.47
Festa	UmBERTo	3	0.48	0.46
Festa	UmBERTo	5	0.47	0.45
Festa	UmBERTo	4	0.45	0.43
DIAG-Sapienza	RoBERTa	4	0.37	0.29
DIAG-Sapienza	RoBERTa	3	0.36	0.29
DIAG-Sapienza	RoBERTa	2	0.35	0.29
DIAG-Sapienza	RoBERTa	1	0.35	0.29
VellaAsta	tum-nlp/bertweet-sexism	1	0.10	0.11

Table 12
Results for *encoder-only models* track [Subtask].

reflect the variability of human judgments in sensitive and socially grounded tasks.
The results obtained by the seven participating teams on the main task, and by a subset of them on the subtask, highlight how different model architectures and configurations have diverse responses to a task in which subjectivity is a core component. Across zero-shot, few-shot, and encoder-only settings, we observe that performance is not solely determined by model scale, but is strongly influenced by architectural choices, prompting strategies, and the way contextual information is integrated into the reasoning process. In particular, for such fine-grained and inherently subjective settings, the performance gap between open-source and state-of-the-art proprietary systems consistently narrows, and in several cases reverses, especially when open models are combined with diversified perspectives or carefully selected in-context examples. Conversely, encoder-only architectures struggle to achieve competitive performance, suggesting that generative and reasoning capabilities play a crucial role in modeling socially grounded and interpretative distinctions.

These findings suggest that model behavior cannot be fully characterized only in terms of raw accuracy, but should also be analysed in relation to how models implicitly encode and reproduce subjective judgments. More broadly, the results highlight the importance of evaluation frameworks that explicitly account for subjectivity, disagreement, and modeling diversity when assessing AI systems on sensitive social phenomena. For this reason, investigating the interplay between model and human subjectivity represents a promising research direction for explaining some of the observed dynamics. Future work will focus on socio-profiling the models to relate their prediction patterns to specific socio-demographic groups, shedding light on the sources and structure of their biases and sensitivities.

Acknowledgments

This paper and the GSI:detect Task are the result of collaboration between all authors. Gloria Comandini wrote 1 and 3; Manuela Speranza wrote 2; Sofia Brenna wrote 4; Davide Testa wrote 5, 6 and 7.
This work was carried out while Davide Testa was enrolled in the Italian National Doctorate on Artificial Intelligence run by Sapienza University of Rome together with Fondazione Bruno Kessler (FBK).

Declaration on Generative AI

Authors used ChatGPT (GPT-5.2) to refine the manuscript's writing style and to support parts of the evaluation code. The authors reviewed and edited all outputs and take full responsibility for the content.

References

[1] 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.

[2] A. P. Association, Guidelines for psychological practice with transgender and gender nonconforming people, American Psychologist 70 (2015) 832–864.
[3] E. Fersini, D. Nozza, P. Rosso, Overview of the evalita 2018 task on automatic misogyny identification (ami), in: T. Caselli, N. Novielli, V. Patti, P. Rosso (Eds.), EVALITA Evaluation of NLP and Speech Tools for Italian. Proceedings of the Final Workshop, Accademia University Press, Torino, 2018.
[4] H. Kirk, W. Yin, B. Vidgen, P. Röttger, Semeval-2023 task 10: Explainable detection of online sexism, in: A. Ojha, A. Doğruöz, G. Da San Martino, H. Tayyar Madabushi, R. Kumar, E. Sartori (Eds.), Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), Association for Computational Linguistics, Stroudsbourg, 2023, p. 2193–2210.
[5] L. Plaza, J. Carrillo-de Albornoz, R. Morante, E. Amigó, J. Gonzalo, D. Spina, P. Rosso, Overview of exist 2023: sexism identification in social networks, in: J. Kamps, L. Goeuriot, F. Crestani, M. Maistro, H. Joho, B. Davis, C. Gurrin, U. Kruschwitz, A. Caputo (Eds.), Advances in Information Retrieval: 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2–6, 2023, Proceedings, Part III, Springer, Cham, 2023, pp. 593–599.
[6] Y. T. Cao, A. Sotnikova, H. Daumé III, R. Rudinger, L. Zou, Theorygrounded measurement of u.s. social stereotypes in english language models, in: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Seattle, 2022, p. 1276–1295.
[7] A. Ovalle, P. Goyal, J. Dhamala, Z. Jaggers, K.-W. Chang, A. Galstyan, R. Zemel, R. Gupta, "i'm fully who i am": Towards centering transgender and non-binary voices to measure biases in open language generation., in: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery, New York, 2023, p. 1246–1266.
[8] B. Savoldi, J. Bastings, L. Bentivogli, E. Vanmassenhove, A decade of gender bias in machine translation, Patterns 6 (2025) 101257.
[9] M. Sanguinetti, F. Poletto, C. Bosco, V. Patti, S. M., An italian twitter corpus of hate speech against immigrants, in: N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, T. Tokunaga (Eds.), Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), European Language Resources Association (ELRA), Paris, 2018, pp. 2798–2805.
[10] M. Wojatzki, T. Horsmann, D. Gold, T. Zesch, Do women perceive hate differently: Examining the relationship between hate speech, gender, and agreement judgement, in: A. Barbaresi, H. Biber, F. Neubarth, R. Osswald (Eds.), Proceedings of the 14th conference on Natural Language Processing (KONVENS 2018), 2018, pp. 110–120.
[11] L. Krusic, Constructing a sentiment-annotated corpus of austrian historical newspapers: Challenges, tools, and annotator experience, in: M. Hämäläinen, E. Öhman, S. Miyagawa, K. Alnajjar, Y. Bizzoni (Eds.), Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities, Association for Computational Linguistics, 2024, pp. 51–62.
[12] G. Comandini, «frena la locomotiva d'europa»: il 'linguaggio della ferrovia' per raccontare la crisi economica tedesca. analisi linguistica sul corpus locomin, Studi Germanici 27 (2025) 127–155.
[13] R. Sprugnoli, F. Mambrini, M. Passarotti, G. Moretti, The sentiment of latin poetry. annotation and automatic analysis of the odes of horace, Italian Journal of Computational Linguistics 9 (2023).
[14] M. Klenner, A. G¨ohring, M. Amsler, Harmonization sometimes harms, in: S. Ebling, D. Tuggener, M. Hürlimann, M. Cieliebak, M. Volk (Eds.), Proceedings of the 5th Swiss Text Analytics Conference (SwissText) 16th Conference on Natural Language Processing (KONVENS), CEUR Workshop Proceedings, 2020.
[15] V. Basile, F. Cabitza, A. Campagner, Toward a perspectivist turn in ground truthing for predictive computing., in: Toward a Perspectivist Turn in Ground Truthing for Predictive Computing, Association for the Advancement of Artificial Intelligence, Washington DC, 2023, pp. 6860–6868.
[16] B. Muscato, C. Sree Mala, M. Marchiori Manerba, G. Gezici, F. Giannotti, An overview of recent approaches to enable diversity in large language models through aligning with human perspectives, in: G. Abercrombie, V. Basile, D. Bernadi, S. Dudy, S. Frenda, L. Havens, S. Tonelli (Eds.), Proceedings

of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024, ELRA and ICCL, 2024, p. 49–55.
[17] V. Basile, 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, in: G. Vizzari, M. Palmonari, A. Orlandini (Eds.), Proceedings of the AIxIA 2020 Discussion Papers Workshop, CEUR Workshop Proceedings, 2020, pp. 31–40.
[18] G. Rizos, B. Schuller, Average jane, where art thou? – recent avenues in efficient machine learning under subjectivity uncertainty, in: M. Lesot, S. Vieira, M. Reformat, J. Carvalho, A. Wilbik, B. Bouchon-Meunier, R. Yager (Eds.), Information Processing and Management of Uncertainty in Knowledge-Based Systems, Springer, Cham, 2020, pp. 42–55.
[19] M. Madeddu, S. Frenda, M. Lai, V. Patti, V. Basile, Disaggreghate it corpus: A disaggregated italian dataset of hate speech, in: F. Boschetti, G. E. Lebani, B. Magnini, N. Novielli (Eds.), Proceedings of the 9th Italian Conference on Computational Linguistics (CLiC-it 2023), CEUR Workshop Proceedings, 2023, pp. 243–250.
[20] R. Artstein, Inter-annotator agreement, in: N. Ide, J. Pustejovsky (Eds.), Handbook of Linguistic Annotation, Springer, Dordrecht, 2017, pp. 297–313.
[21] O. E. Sorokoletova, E. Musumeci, D. Nardi, Diag-sapienza at gsi:detect: A technical report, 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.
[22] D. Festa, Festa at gsi:detect: In-context learning vs. fine-tuning for gender stereotype detection, 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.
[23] F. Giobergia, Minds at gsi:detect: From logits to degrees of agreement in gender stereotype detection with llms, 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.
[24] C. Zaghi, Prisma at gsi:detect: Comparing pbj persona-based and few-shot approaches to gender stereotype detection, 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.
[25] T. Labruna, Tiz at gsi:detect: Modeling gender stereotypes detection as multi-category gender stereotype scoring, 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.
[26] S. Greco, M. La Quatra, M. Marchiori Manerba, R. Muñoz Sánchez, A. T. Cignarella, Stereobusters at gsi:detect: Llm-based detection and human qualitative analysis of gender stereotypes in italian short texts, 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.
[27] P. Göttfert, R. Huber, F. Mariani, NLPeace@GermEval shared task 2025: Fine-tuned BERT vs. prompted LLMs for German hate speech detection, in: Proceedings of the 21st Conference on Natural Language Processing (KONVENS 2025): Workshops, HsH Applied Academics, Hannover, Germany, 2025.