MINDS at GSI:detect: From Logits to Degrees of Agreement in Gender Stereotype Detection with LLMs

Flavio Giobergia

Politecnico di Torino. Turin, Italy

Abstract

We present our submission to the GSI:detect shared task at EVALITA, which focuses on detecting gender stereotypes in Italian texts. Unlike standard binary classification, the task requires predicting a continuous score representing the fraction of annotators who identify a text as containing a gender stereotype. We propose a two-step approach based on few-shot prompting of a large language model and a lightweight regression layer trained on model logits. This design allows us to approximate annotator uncertainty while relying on an open-source model and limited computational resources.

Keywords

annotator disagreement estimation, large language models, few-shot prompting, gender stereotypes

1. Introduction

The detection of gender stereotypes in online contents is an important problem at the intersection of natural language processing (NLP) and social bias analysis. Stereotypes can be expressed explicitly or implicitly, often through subtle linguistic cues, pragmatic assumptions, or culturally grounded generalizations. As a result, their identification is inherently subjective and sensitive to annotator background, values, and lived experience. This makes stereotype detection a particularly challenging task for automated systems, especially when compared to more clearly delimited phenomena such as offensive language or hate speech.

The GSI:detect shared task [1] at EVALITA [2] addresses this challenge for Italian by explicitly embracing subjectivity and disagreement. In this work, we addressed the main task of GSI:detect, and not the subtask of classifying gender stereotypes. Rather than framing stereotype detection as a binary classification problem, the task models it as a regression task, where the target represents the degree of agreement among annotators regarding the presence of gender stereotypes in a text. Each instance is annotated by four annotators, and the gold label – referred to as the *GS value* – corresponds to the fraction of annotators identifying stereotypical content, yielding values in the set $\{0, 0.25, 0.5, 0.75, 1\}$. While constrained by the number of annotators, this formulation is designed to generalize to settings with different annotation schemes and highlights the intrinsically subjective nature of stereotype detection.

This perspective aligns with prior work showing that stereotypes are deeply influenced by subjectivity and social positioning [3], motivating a shift toward perspectivist approaches in NLP [4]. In this view, disagreement among annotators is not noise to be minimized, but rather a signal that reflects how social meaning is constructed and negotiated through language.

Recent years have seen the rapid adoption of large language models (LLMs) across a wide range of NLP tasks, often yielding strong performance even in low-resource or few-shot settings. LLMs have been successfully applied to text classification problems in diverse domains, both in low-resource cases, where limited data is available, but sufficient to fine-tune language models on the smaller end of the scale (e.g., in Italian, for hate speech detection [5], dialect classification [6] [7]), as well as in few-shot scenarios, where only a handful of annotated examples are available: not enough to fine-tune a model,

EVALITA 2026: 9th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Feb 26 – 27, Bari, IT
flavio.giobergia@polito.it (F. Giobergia)
0000-0001-8806-7979 (F. Giobergia)

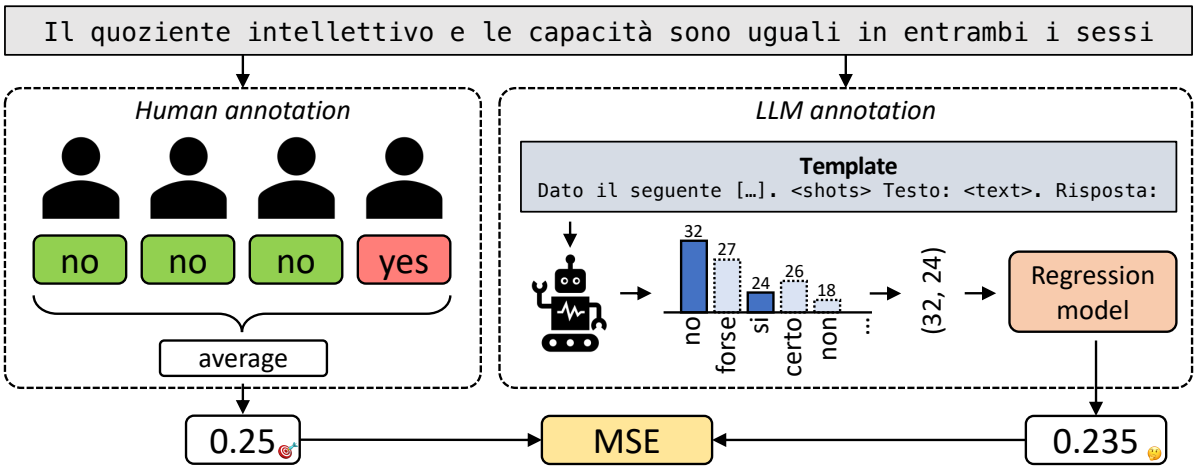

Figure 1: Summary of the human and LLM-based annotation of the target data. The input sentence (in English, "The intelligence quotient and the capabilities are the same for both genders") is annotated by four humans as either containing gender stereotypes (yes) or not (no). The GS value is computed as the average score (no = 0, yes = 1). The LLM is prompted to produce a similar annotation. The logits produced by the LLM are mapped, with a regression model, to the target GS value (using the limited training set available).

but sufficient to use larger models with an In-Context Learning approach [8]. This is particularly useful, for instance, when annotations are expensive or can only be carried out by specialized domain experts, as is the case for the verification of scientific claims [9], or the annotation of scientific works [10].

These applications suggest that LLMs capture rich semantic and pragmatic information, making them promising candidates for socially grounded tasks that require sensitivity to nuance and ambiguity. At the same time, most prior work focuses on categorical decisions, whereas fewer studies investigate whether LLMs can meaningfully model degrees of uncertainty or disagreement inherent in subjective annotations.

In this work, we propose a simple yet effective approach that leverages a large language model in a few-shot setting and exploits its internal confidence signals, as represented by predicted logits, to approximate annotator agreement. Rather than relying solely on discrete predictions, we treat these signals as proxies for uncertainty and calibrate them through a regression model. This allows us to capture both the presence of gender stereotypes and the noisy nature of annotator consensus.

2. Methodology

We summarize the annotation pipeline in Figure 1. The annotators' scores are summarized into their average. The LLM annotation is produced using a template that embeds a limited number of randomly sampled shots, and the text to be annotated. The output of the LLM is processed by (1) extracting the relevant logits and (2) building a regression model that maps logits to the final GS value. The rest of the section provides further details on the proposed approach.

2.1. Few-shot prompting

We employ a large language model in a few-shot learning setup. The model is prompted with a short description of the task, as well as 20 randomly selected examples illustrating gender-stereotypical and non-stereotypical content.

Importantly, the model is *not* asked to directly predict the numerical target value: it is well known that LLMs have limitations when it comes to handling numerical values, due to their operations at a token-level [11].

Instead, the model is prompted directly to produce a yes/no[1] outcome regarding the presence of

[1]Since the model is prompted in Italian, the actual question requires an annotation as si (yes) or no. The rest of the paper uses

stereotypical contents in the input text, and the 20 random shots provided in the prompt are binarized through a thresholding ('yes' if the probability is greater than or equal to 0.5, 'no' otherwise).

Upon generation of an answer, instead of directly using the generated answer, we extract the logits associated with the possible responses (namely, 'si' and 'no', since the model is prompted in Italian). We assume that the magnitudes associated with these two logits are representative of the model's internal confidence – which, we argue, is proportional to the agreement among annotators. In other words, for ambiguous text (i.e., GS value ≈ 0.5) the LLM is expected to assign similar probabilities to the 'yes' and 'no' tokens, whereas for a text for which annotators are more confident (i.e., GS value ≈ 0 or GS value ≈ 1), the LLM is expected to assign higher probability to the correct token.

2.2. Logit-based regression

Rather than relying on the model's direct output as the final prediction, we introduce a calibration step. The extracted logits for the 'yes' and 'no' outcomes are treated as separate features, used to train a lightweight regression model (e.g., linear regression or K-Nearest Neighbors) on the training set. The regression target is the fractional agreement score derived from the human annotations (i.e., the GS value).

This two-step process allows us to map the model's logits to a continuous prediction that reflects annotator consensus. Intuitively, larger differences between logits indicate higher certainty, whereas smaller differences correspond to ambiguous cases – mirroring the variability observed among human annotators.

2.3. Inference

At inference time, a new text is first processed by the language model under the same few-shot prompt. The resulting logits are then passed through the trained regression model to produce a final predicted score in $[0, 1]$. This score represents the estimated fraction of annotators who would identify the text as containing a gender stereotype.

3. Experimental Results

In this section, we report the experimental results obtained with the proposed system. We selected Qwen 2.5 14B [12] as a trade-off between representational capacity and computational constraints (single-GPU inference, no fine-tuning), observing consistent gains over smaller open models. As such, we report all results with the specified Qwen model.

The dataset available included 200 annotated samples, and 810 test samples used for the final evaluation. As already discussed, we used 10% of the development set (i.e., 20 randomly chosen samples) for the generation of the prompt, and we adopted the remaining 180 samples for the training of the regression model that predicts the final agreement, based on the yes/no logits.

Table 1 reports the official results of the GSI:detect shared task, in terms of Mean Squared Error (MSE), Normalized MSE (NMSE) and the inverse of 1 + NMSE (for a *higher-is-better* metric). Our system ranks approximately in the middle of the leaderboard. While it does not outperform all other approaches, it achieves competitive performance when compared against similar baselines, such as the baseline model using Qwen 3 (14B), and even outperforms, in some cases, commercial baselines (e.g., based on GPT-5 nano).

Prompt details. We note that the runs that were submitted for the competition contained a misleading prompt, focused on misogyny rather than gender stereotype detection (entries marked as *(mys)* in Table 1). This experiment was carried out as a part of the prompt engineering process. We additionally report the (non official) results, obtained with the revised prompt (marked as *(GS)* in Table 1). This small, but

yes/no or si/no interchangeably.

Table 1
Results overview with ranking, models, and evaluation metrics. Given the large number of participants, only the top-3 solutions submitted by other participants (DIAG-Sapienza, StereoBusters) are reported. An asterisk is used to highlight non-official solutions (i.e., solutions that have *not* been submitted for the competition). *(GS)* marks solutions with a gender stereotype-focused prompt, whereas *(mys)* marks solutions with a misogyny-focused one.

Rank	Team	Model	Approach	MSE ↓	NMSE ↓	$\frac{1}{1+\text{NMSE}}$ ↑
1	DIAG-Sapienza	GPT-5	zero-shot	0.0766	0.4292	0.700
2	DIAG-Sapienza	GPT-5	few-shot (4)	0.0820	0.4594	0.685
3	StereoBusters	Gemma 3 (27B)	few-shot (5)	0.0983	0.5508	0.645
-	**MINDS (GS)*	Qwen 2.5 (14B) + KNN	few-shot (20)	0.1104	0.6184	0.618
16	BASELINE	GPT-5 nano (split prompt)	zero-shot	0.1150	0.6430	0.609
-	**MINDS (GS)*	Qwen 2.5 (14B) + LR	few-shot (20)	0.1192	0.6676	0.600
19	BASELINE	GPT-5 nano (unified prompt)	zero-shot	0.1240	0.6950	0.590
22	*MINDS (mys)*	Qwen 2.5 (14B) + KNN	few-shot (20)	0.1288	0.7213	0.581
-	**MINDS (GS)*	Llama 3 (8B) + LR	few-shot (20)	0.1329	0.7446	0.573
-	**MINDS (GS)*	Llama 3 (8B) + KNN	few-shot (20)	0.1363	0.7636	0.567
27	*MINDS (mys)*	Qwen 2.5 (14B) + LR	few-shot (20)	0.1379	0.7725	0.564
37	BASELINE	Qwen 3 (14B)	zero-shot	0.1500	0.8420	0.543
40	BASELINE	prediction=0.5	-	0.1800	1.0083	0.498

important prompt detail shows a relevant change in performance (for the best-performing model, a 14% improvement in terms of MSE, from 0.1288 down to 0.1104).

Additional models. We include the performance obtained with a smaller model (Llama 3 8B [13]) to show that the larger Qwen model consistently achieves better performance w.r.t. smaller versions (when considering the *"GS"* solutions): a potentially larger model, therefore, could be expected to achieve even higher performance.

Regression approach. As for the regression model, we include two alternatives: a simple linear regression (LR), and KNN (with K = 15, selected following a hyperparameter tuning step). We show that both regression models produce similar results, with slight variations. KNN provides the best performance for Qwen, although this is not necessarily always the case, as confirmed in Section 3.2.

3.1. Predictive capabilities of logits

We provide a qualitative result that supports the choice of using the yes/no logits for the prediction of the agreement. In particular, we report in Figure 2 the scatter plot of the yes and no logits for the 180 samples with known ground truth, as well as for the 810 unlabelled samples. The labelled samples are colored according to the ground truth agreement: it can be observed that high-confidence results (i.e., large logit values for one class and small logit for the other one) are associated with higher agreement (either close to 0, or close to 1 depending on the outcome).

3.2. Number of shots

One aspect of interest for the choice of In-Context Learning approaches is the quantity of shots to be used for the prompt. In general, a large number of shots (e.g., as many as can fit the context window) can be used. However, for this specific system, there is a trade-off to be made. The limited number of labelled points available is used both for the preparation of the prompt, as well as for the calibration through the training of the regression model. So, using an excessive number of examples in the prompt would not leave sufficient samples for the second step. We measure the performance of the solution for a varying number of shots provided in the context (from 0 to 32). Figure 3 shows the evolution, using as heads both a linear regression, and a KNN regressor, over 5 runs with different randomly sampled shots.

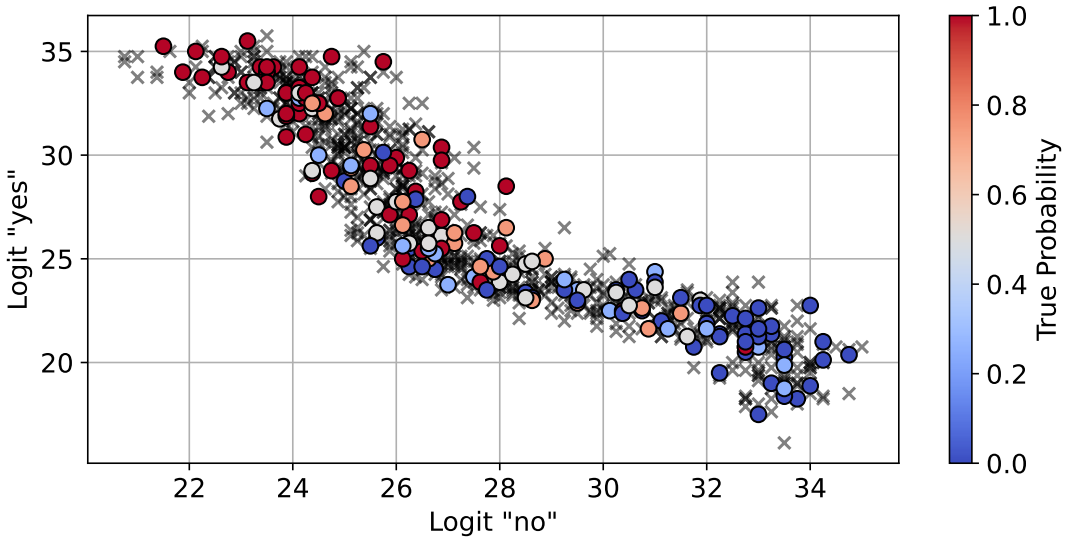

Figure 2: Scatter plot of the no/yes logits produced by the LLM when asked to annotate each sample. The color is proportional to the ground truth value for the "train" samples, whereas the 'x' marker represents "test" samples.

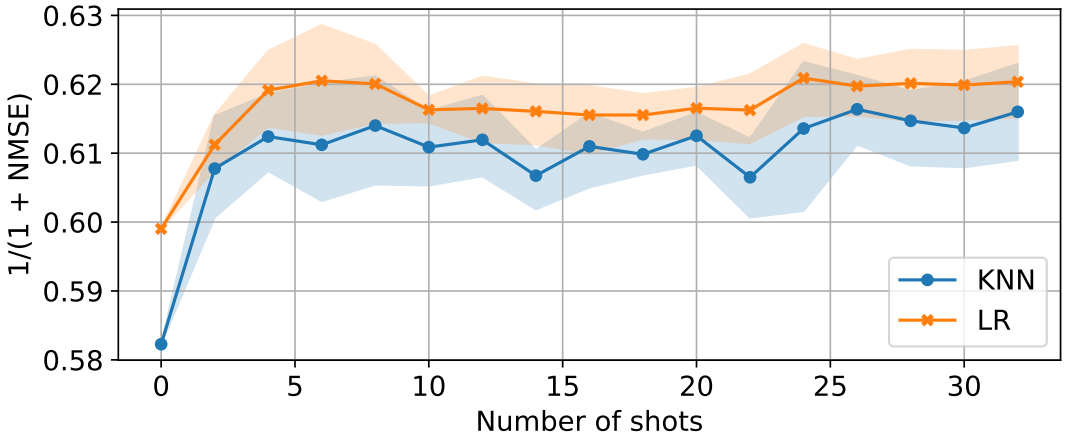

Figure 3: Change in performance, measured as 1/(1+NMSE), as a function of the number of shots provided. The error bars represents the 95% confidence intervals over 5 runs, with randomly samples shots.

Two interesting aspects emerge. First, the 95% confidence interval overlap between KNN and LR, although the mean performance observed for LR w.r.t. KNN suggests consistent superiority of LR, without strong statistical separation. Second, the measured performance improves when passing a non-zero number of shots, but it does not improve significantly when using more than 4-8 shots. The choice of using 20 shots for the submissions, therefore, may not have been ideal. However, these are considerations that can only be made in hindsight: during the competition, the limited labelled dataset available was used to run a similar experiment, producing different results – which seemed to indicate that a larger number of shots could be beneficial. However, the limited validation set size used and the execution of a single run may have produced less reliable results.

4. Conclusions

We presented a two-step approach for the GSI:detect shared task that combines few-shot prompting of a large language model with a regression-based calibration layer. By leveraging model logits as a proxy for confidence, our method captures the graded nature of annotator agreement and provides a nuanced

prediction of gender stereotype presence.

Our experiments show that this approach outperforms direct score prediction and achieves competitive results using an open-source model at a fraction of the cost of large proprietary systems. Future work could explore more advanced calibration techniques, alternative uncertainty-aware objectives, and extensions to other bias-related tasks and languages.

Acknowledgments

This study was carried out within the FAIR - Future Artificial Intelligence Research and received funding from the European Union Next-GenerationEU (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR) – MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.3 – D.D. 1555 11/10/2022, PE00000013). This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT (GPT 5.2) in order to: Drafting content. After using this tool, the author reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] 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.

[2] 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.

[3] M. Sanguinetti, F. Poletto, C. Bosco, V. Patti, M. A. Stranisci, An italian twitter corpus of hate speech against immigrants, in: Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018), 2018.

[4] F. Cabitza, A. Campagner, V. Basile, Toward a perspectivist turn in ground truthing for predictive computing, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 37, 2023, pp. 6860–6868.

[5] M. Sanguinetti, G. Comandini, E. Di 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, Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (2020).

[6] A. Ramponi, C. Casula, Geolingit at evalita 2023: Overview of the geolocation of linguistic variation in italy task, in: EVALITA Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian Final Workshop: Parma, Italy, September 7-8th, 2023, Accademia University Press, 2024, p. 109.

[7] A. Koudounas, F. Giobergia, I. Benedetto, S. Monaco, L. Cagliero, D. Apiletti, E. Baralis, et al., baρtti at geolingit: Beyond boundaries, enhancing geolocation prediction and dialect classification on social media in italy, in: CEUR workshop proceedings, CEUR, 2023.

[8] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Advances in neural information processing systems 33 (2020) 1877–1901.

[9] C. Alvarez, M. Bennett, L. L. Wang, Zero-shot scientific claim verification using llms and citation text, in: Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024), 2024, pp. 269–276.

[10] F. Giobergia, A. Koudounas, E. Baralis, Large language models-aided literature reviews: A study on few-shot relevance classification, in: 2024 IEEE 18th International Conference on Application of Information and Communication Technologies (AICT), IEEE, 2024, pp. 1–5.

[11] H. Li, X. Chen, Z. Xu, D. Li, N. Hu, F. Teng, Y. Li, L. Qiu, C. J. Zhang, L. Qing, et al., Exposing numeracy gaps: A benchmark to evaluate fundamental numerical abilities in large language models, in: Findings of the Association for Computational Linguistics: ACL 2025, 2025, pp. 20004–20026.

[12] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, et al., Qwen3 technical report, arXiv preprint arXiv:2505.09388 (2025).

[13] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, et al., The llama 3 herd of models, arXiv preprint arXiv:2407.21783 (2024).