Team Prisma at GSI:detect: Comparing PB&J Persona-Based and Few-Shot Approaches to Gender Stereotype Detection at EVALITA 2026

Claudia Zaghi[1]

[1]*Independent Researcher*

Abstract

We present a comparative study of two approaches for automated gender stereotype detection in Italian text: persona-based modeling that simulates human annotator perspectives through psychological profiling, and few-shot learning using annotated examples. Our persona-based approach employs the Persona-based Judgment (PB&J) framework, which integrates established psychological constructs including Big Five personality traits, Schwartz values, and Primal World Beliefs to create synthetic annotator profiles. We submitted six runs: one aggregating four personas, four individual persona runs, and one few-shot baseline. Personas exhibited distinct detection behaviors (detection rates: 17–29%) but all under-detected stereotypes relative to the gold standard (70.1% stereotypical content). The few-shot approach outperformed persona runs on both the main task (intensity regression: NMSE 0.7041 vs. 2.04–2.24) and subtask (category classification: F1 Macro 0.6115 vs. 0.17–0.29). Persona modeling captures annotator variation but requires example-based calibration to match task conventions.

Keywords

gender stereotype detection, LLM personas, PB&J framework, few-shot learning, Italian NLP

1. Introduction

Gender stereotypes have shaped social discourse for centuries [1]. Social media has amplified stereotypical content [2], making automated detection both a social imperative and a computational challenge [3].

The Natural Language Processing (NLP) community has responded with corpora for probing bias in word embeddings [4], datasets for sexism detection [5], and evaluation frameworks for measuring stereotypical associations in language models [6]. Shared evaluation campaigns have emerged as key venues, including the NLPCC 2025 Gender Bias Mitigation Challenge for Chinese [7] and the GSI:detect task [8] at EVALITA 2026 [9] for Italian.

Gender stereotype detection is inherently subjective: legitimate disagreement among annotators reflects differences in interpretation, background, and sensitivity [10, 11]. Annotator characteristics significantly influence detection outcomes [12, 7], and GSI:detect embraces this by preserving individual judgments rather than collapsing them through majority voting.

We participated in GSI:detect with two approaches, motivated by the data harmonization literature emphasizing that consistency between sources is essential for valid predictions [13]:

1. **Persona-based zero-shot**: Constructing LLM personas using Persona-based Judgment (PB&J) framework [14] to match the demographic profile of the original annotators (three female, one male; ages 20–47) as much as possible.
2. **Few-shot learning**: Calibrating predictions using 35 stratified examples from the development data to align with annotation conventions.

Code and outputs are available at https://github.com/ClaZaghi/gsi_detect_evalita2026_prisma.

EVALITA 2026: 9th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Feb 26 – 27, Bari, IT

cla.zaghi@gmail.com (C. Zaghi)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

published 2026-04-16

CEUR-WS.org/Vol-4195/21.pdf

2. Task and Data

The GSI:detect task focuses on gender stereotype detection in Italian short texts, adopting a perspectivist approach that preserves annotator disagreement rather than collapsing it through majority voting [15]. The task is articulated in two subtasks:

Main task (GS Detection): A regression task requiring systems to predict a GS value in [0,1] quantifying the degree of stereotypical content. Values are computed from four independent binary annotations: complete agreement yields 0 or 1, while disagreement produces intermediate values (e.g., 0.25, 0.50, 0.75).

Subtask (GS Classification): A multi-class classification task requiring systems to assign texts to one of six stereotype categories: *role*, *personality*, *competence*, *physical*, *sexual*, and *relational*.

2.1. Dataset

The organisers provided approximately 1,000 Italian short texts collected from social media and informative websites, including both formal and informal language. The dataset is split as 20% development data and 80% test data. Each instance includes the text, individual annotations from four annotators, the aggregated GS value, the GS category, and a context flag indicating whether additional contextual information is provided. All four annotators are Italian native speakers, cisgender, and sensitive to gender-related issues; their demographic profile (three female, one male; ages 20–47; Master's or PhD level) is detailed in Section 3.1.

2.2. Additional Resources

We adopted a closed-task setting, using no external training data or additional resources beyond the provided development set. Our few-shot system (Run 6) used 35 stratified examples from the development data, while persona-based systems (Runs 1–5) operated in a zero-shot setting.

All prompts were written in Italian and incorporated the annotation guidelines provided by the task organizers (see Appendix C for complete prompts).

3. System Description

3.1. Persona Construction with PB&J Framework

Our primary design goal was matching synthetic LLM personas to GSI:detect annotator demographics: four Italian native speakers (three female, one male; ages 20–40+; Master's or PhD level) [8]. Since annotator identities were anonymized, demographic matching served as a proxy for preserving annotator diversity [13].

We linked demographic attributes to expected behavioral patterns (e.g., older annotators with more conservative thresholds) and applied the PB&J [14] to structure personas around three psychological dimensions:

- **Big Five Personality Traits** [16]: Conservative annotators received medium openness and high conscientiousness; sensitive annotators received very high openness with medium-high conscientiousness.
- **Schwartz Values** [17]: Conservative annotators emphasized tradition, security, and conformity; sensitive annotators prioritized self-direction and universalism.
- **Primal World Beliefs** [18]: Conservative annotators held higher "safe world" beliefs (higher evidence thresholds); sensitive annotators held lower beliefs (increased vigilance toward implicit bias).

Each persona included an Italian-language annotation rationale and experiential narrative grounded in these psychological dimensions. Complete persona definitions are provided in Appendix B.

Table 1 summarizes the four personas.

Table 1
PB&J persona profiles for Runs 1–5.

Attribute	A	B	C	D
Type	Balanced	Traditional	Sensitive	Academic
Gender	Female	Male	Female	Female
Age	35	47	26	29
Education	Master's	Master's	Master's	PhD
Openness	Med-High	Medium	Very High	Very High
Conscientiousness	High	Very High	Med-High	High
Key Values	Universal.	Tradition	Self-dir.	Self-dir.
Safe World	Med-High	High	Med-Low	Low

3.2. Implementation

We submitted six runs using Claude 3.5 Sonnet via AWS Bedrock: five zero-shot persona-based runs and one few-shot baseline. All prompts were written in Italian and incorporated the annotation guidelines provided by task organizers in Italian (see Appendix A).

3.3. Run 1: Aggregated Personas (Zero-Shot)

All four personas defined in Appendix B annotated the test units independently, providing a binary judgment (stereotype present: yes/no). The final gs_value was computed as the proportion of personas detecting a stereotype, yielding values in {0.00, 0.25, 0.50, 0.75, 1.00}. Predictions were parallelized with 16 texts processed concurrently, each spawning 4 annotator threads.

3.4. Runs 2–5: Individual Personas (Zero-Shot)

Each run used a single persona predicting independently. For each text, the persona provided a binary stereotype judgment (yes/no), a category prediction, and an intensity estimate on a four-point scale for positive detections: 0.33 (slight/doubtful), 0.67 (moderate/clear), or 1.00 (strong/explicit). Negative detections received gs_value 0.00. Temperature parameters were calibrated to each persona's expected consistency, with lower temperatures for more conservative annotators:

- **Run 2**: Persona A (Balanced) – Temperature 0.5
- **Run 3**: Persona B (Traditional) – Temperature 0.3
- **Run 4**: Persona C (Sensitive) – Temperature 0.7
- **Run 5**: Persona D (Academic) – Temperature 0.8

Complete persona definitions are provided in Appendix B. We discuss potential confounds from temperature variation and output scale choice in Section "Limitations".

3.5. Run 6: Few-Shot Baseline

The baseline used standard few-shot prompting with the translated annotation guidelines and 35 stratified examples, but without persona framing. Examples were drawn from the development set (5 per category, including "no stereotype") using a fixed random seed, with order randomized for each prediction to reduce position bias. Temperature was set to 0.3. This run represents a closed-task setting with no external resources.

4. Results

4.1. Main Task: Stereotype Intensity Regression

The few-shot baseline outperformed all zero-shot persona runs on the regression task (NMSE 0.7041 vs. 2.04–2.24). Table 2 presents the main task results.

Table 2
Main task (regression) results on GSI:detect test set (810 samples).

Run	Approach	NMSE	MSE	Rank
Run 1	Aggregation (four-persona ensemble)	2.1152	0.3776	52/54
Run 2	Persona A	2.0382	0.3639	50/54
Run 3	Persona B	2.2409	0.4001	54/54
Run 4	Persona C	2.1134	0.3773	51/54
Run 5	Persona D	2.2322	0.3985	53/54
Run 6	**Few-shot**	**0.7041**	**0.1257**	**20/54**

Among zero-shot runs, results were mixed: Personas A and C slightly outperformed the aggregated four-persona ensemble (NMSE 2.0382 and 2.1134 vs. 2.1152), while Personas B and D underperformed (2.2409 and 2.2322). All zero-shot NMSE values exceeded 1.0, indicating performance worse than a baseline predicting the mean.

4.2. Subtask: Stereotype Category Classification

The few-shot baseline dominated the classification subtask, achieving first place among the 12 participating teams (F1 Macro 0.6115). Table 3 presents classification results.

Table 3
Subtask (classification) results on GSI:detect test set.

Run	Approach	F1 Macro	Rank
Run 1	Aggregation (four-persona ensemble)	0.2896	39/47
Run 2	Persona A	0.2358	44/47
Run 3	Persona B	0.1778	45/47
Run 4	Persona C	0.2291	43/47
Run 5	Persona D	0.1711	46/47
Run 6	**Few-shot**	**0.6115**	**1/47**

The aggregated four-persona ensemble outperformed all single-persona runs on classification (F1 0.2896 vs. 0.1711–0.2358). Unlike the main regression task, where outputs from single personas proved more effective, classification benefits from pooling diverse viewpoints to disambiguate overlapping categories.

Category detection difficulty varied with linguistic explicitness. Table 4 shows per-category F1 for Run 6 (few-shot), showing that concrete categories (e.g., physical) were easier to detect than implicit ones (e.g., relational). Table 4 shows per-category F1 for the best-performing run (Run 6).

5. Discussion

5.1. Detection Rate Analysis

We define detection rate as the percentage of texts a system classifies as containing stereotypes (i.e., gs_value > 0). Table 5 shows that all persona runs, both single and aggregated, systematically under-

Table 4
Per-category F1 scores for Run 6 (Few-shot).

Category	F1	Precision	Recall
physical	0.8404	0.8061	0.8778
role	0.7580	0.7411	0.7757
competence	0.7232	0.7788	0.6750
sexual	0.7143	0.7353	0.6944
personality	0.6757	0.6579	0.6944
relational	0.5690	0.7333	0.4648

detected stereotypes relative to the test data. The few-shot approach achieved 82.6% detection, substantially closer to the test data (70.1% stereotypical content), demonstrating effective threshold calibration through in-context examples.

Table 5
Stereotype detection rates across runs.

Run	No Stereotype	Stereotype	Detection Rate
Run 5: Persona D	669	141	17.4%
Run 3: Persona B	645	165	20.4%
Run 2: Persona A	629	181	22.3%
Run 1: Aggregation (four-persona ensemble)	607	203	25.1%
Run 4: Persona C	575	235	29.0%
Gold Standard	242	568	70.1%
Run 6: Few-shot	**141**	**669**	**82.6%**

Despite poor absolute performance, personas demonstrated behavioral differentiation. Persona B (conservative, high safe-world beliefs) achieved 20.4% detection, while Persona C (sensitive, low safe-world beliefs) achieved the highest rate at 29.0%. Persona D's low detection rate (17.4%) despite a sensitive profile may reflect multiple factors: (1) its high temperature setting (0.8) increased output variability; (2) sequential processing meant Persona D ran last, potentially encountering rate limiting; and (3) post-processing rules defaulted unparseable JSON responses to non-detection (gs_value = 0).

5.2. Limitations

Our study has several limitations that should be considered when interpreting results.

- **Output scale mismatch.** Individual persona runs used a four-point scale ({0, 0.33, 0.67, 1.0}) while gold standard values, derived from four binary annotations, can only take values in {0, 0.25, 0.50, 0.75, 1.0}. This mismatch introduced irreducible error. However, this effect was minor relative to under-detection errors.
- **Temperature confounds.** Varying temperature settings across personas (0.3–0.8) confound interpretation of behavioral differences. Observed variation in detection rates may reflect temperature effects rather than psychological profile differences. Future work should isolate these factors by testing personas at fixed temperatures.
- **Persona D anomaly.** Persona D achieved the lowest detection rate (17.4%) despite having the most sensitive psychological profile (low safe-world beliefs, very high openness). This unexpected result may stem from its high temperature setting (0.8) introducing excessive randomness, or from rate limiting during sequential API calls causing degraded responses. We cannot distinguish between these explanations with current data.
- **Persona validation.** Mappings from demographics to psychological profiles were based on design assumptions, not empirically validated against actual annotator behavior.

- **Single model and language.** All experiments used Claude 3.5 Sonnet; results may not generalize to other LLMs. Additionally, the PB&J framework and psychological instruments were developed in English-language contexts, potentially limiting effectiveness for Italian annotation conventions.

6. Conclusion

We presented Team Prisma's systems for the GSI:detect task at EVALITA 2026, comparing psychologically-grounded personas constructed with the PB&J framework against a few-shot baseline.

The few-shot system substantially outperformed all persona-based runs, ranking 20th of 54 on regression (NMSE 0.7041) and first among the 12 participating teams on classification (F1 Macro 0.6115). Among zero-shot approaches, we observed task-dependent trade-offs: single personas achieved lower regression error through continuous intensity outputs, while aggregating multiple personas improved classification by pooling diverse perspectives.

All persona runs systematically under-detected stereotypes (17.4%–29.0% vs. 70.1% gold standard), confirming that demographic alignment alone cannot replace task-specific calibration. Nevertheless, the 11.6% point spread between personas validates the PB&J framework's ability to induce distinct behavioral patterns aligned with psychological profiles.

Future work will explore hybrid approaches combining persona diversity with few-shot calibration, aiming to capture legitimate annotator disagreement while achieving competitive absolute performance.

Acknowledgments

The author thanks the GSI:detect task organizers for providing the dataset and evaluation infrastructure.

Declaration on Generative AI

During the preparation of this work, the author used Claude 3.5 Sonnet (Anthropic) via AWS Bedrock for:

- **Writing assistance**: Editing manuscript text
- **Code development**: Assistance with Python for debugging and implementation

The author reviewed and edited all content and takes full responsibility for the publication.

References

[1] L. Arcuri, M. R. Cadinu, Gli Stereotipi. Dinamiche psicologiche e contesto delle relazioni sociali, Il Mulino, Bologna, 1998.

[2] M. R. Costa-jussà, An analysis of gender bias studies in natural language processing, Nature Machine Intelligence 1 (2019) 495–496.

[3] S. L. Blodgett, S. Barocas, H. Daumé III, H. Wallach, Language (technology) is power: A critical survey of "bias" in NLP, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 5454–5476. URL: https://aclanthology.org/2020.acl-main.485/. doi:10.18653/v1/2020.acl-main.485.

[4] T. Bolukbasi, K. Chang, J. Y. Zou, V. Saligrama, A. Kalai, Man is to computer programmer as woman is to homemaker? debiasing word embeddings, CoRR abs/1607.06520 (2016). URL: http://arxiv.org/abs/1607.06520. arXiv:1607.06520.

[5] P. Chiril, V. Moriceau, F. Benamara, A. Mari, G. Origgi, M. Coulomb-Gully, An annotated corpus for sexism detection in French tweets, in: N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis

(Eds.), Proceedings of the Twelfth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020, pp. 1397–1403. URL: https://aclanthology.org/2020.lrec-1.175/.

[6] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, K.-W. Chang, Gender bias in coreference resolution: Evaluation and debiasing methods, in: M. Walker, H. Ji, A. Stent (Eds.), Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 15–20. URL: https://aclanthology.org/N18-2003/. doi:10.18653/v1/N18-2003.

[7] Y. Li, G. Zhang, H. Hong, Y. Wang, C. Lin, Overview of the nlpcc 2025 shared task: Gender bias mitigation challenge, in: X.-L. Mao, Z. Ren, M. Yang (Eds.), Natural Language Processing and Chinese Computing, Springer Nature Singapore, Singapore, 2026, pp. 453–464.

[8] 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.

[9] 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.

[10] L. Aroyo, C. Welty, Truth is a lie: Crowd truth and the seven myths of human annotation, AI Magazine 36 (2015) 15–24. URL: https://doi.org/10.1609/aimag.v36i1.2564. doi:10.1609/aimag.v36i1.2564.

[11] A. Uma, T. Fornaciari, A. Dumitrache, T. Miller, J. Chamberlain, B. Plank, E. Simpson, M. Poesio, Learning from disagreement: A survey, Journal of Artificial Intelligence Research 72 (2021) 1385–1470.

[12] Z. Waseem, Are you a racist or am I seeing things? Annotator influence on hate speech detection on Twitter, in: Proceedings of the First Workshop on NLP and Computational Social Science, Austin, Texas, 2016, pp. 138–142.

[13] M. Y. Guan, V. Gulshan, A. M. Dai, G. E. Hinton, Who said what: Modeling individual labelers improves classification, 2018. URL: https://arxiv.org/abs/1703.08774. arXiv:1703.08774.

[14] B. Joshi, X. Ren, S. Swayamdipta, R. Koncel-Kedziorski, T. Paek, Improving language model personas via rationalization with psychological scaffolds, in: C. Christodoulopoulos, T. Chakraborty, C. Rose, V. Peng (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2025, Association for Computational Linguistics, Suzhou, China, 2025, pp. 21747–21770. URL: https://aclanthology.org/2025.findings-emnlp.1187/. doi:10.18653/v1/2025.findings-emnlp.1187.

[15] 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. URL: http://dx.doi.org/10.1609/aaai.v37i6.25840. doi:10.1609/aaai.v37i6.25840.

[16] L. R. Goldberg, The structure of phenotypic personality traits, American Psychologist 48 (1993) 26–34.

[17] S. H. Schwartz, Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries, Advances in Experimental Social Psychology 25 (1992) 1–65.

[18] J. D. Clifton, J. D. Baker, C. L. Park, et al., Primal world beliefs, Psychological Assessment 31 (2019) 82–99.

A. Annotation Guidelines

All prompts incorporated the following Italian annotation guidelines, adapted from the task organizers' original guidelines. These were included verbatim in all persona-based and few-shot prompts.

Listing 1: Italian annotation guidelines included in all prompts

```
LINEE GUIDA - Categorie di Stereotipi

Gli stereotipi di genere sono credenze rigide e generalizzate riguardo
ai ruoli, comportamenti, abilita o caratteristiche che uomini e donne
"dovrebbero" avere in base al loro sesso o genere.

Ogni esempio deve essere assegnato a una sola categoria, anche se nello
stesso testo compaiono piu stereotipi.

Regola generale: se nello stesso esempio compaiono diverse categorie di
stereotipi, gli annotatori devono selezionare la categoria che appare
per prima nel testo.

Esempio:
"Le donne sono troppo emotive per essere buone leader, e gli uomini
dovrebbero sempre mantenere la famiglia."
-> Contiene sia stereotipo di competenza che di ruolo, ma poiche lo
stereotipo di competenza appare per primo, deve essere assegnato alla
categoria "Competenza".

NB: Nel caso dei commenti, e importante considerare se il commento e
pertinente al contesto.
- Se qualcuno commenta l'aspetto fisico di una donna in un contesto in
  cui e irrilevante (es. durante un discorso politico) -> stereotipo
- Se l'aspetto e contestualmente rilevante (es. commento sull'outfit
  in una sfilata di moda) -> non stereotipo

---

1. Stereotipi di Ruolo (role)
Aspettative sociali e culturali che definiscono quali ruoli familiari
o sociali sono "appropriati" per uomini e donne.
- Uomini: capifamiglia, responsabili del sostegno economico, della
  protezione e delle decisioni familiari.
- Donne: madri, casalinghe, responsabili della cura dei figli e del
  lavoro domestico.

2. Stereotipi di Personalita (personality)
Attribuzione di tratti emotivi o caratteriali basati sul genere.
Include insulti: "femminuccia", "checca isterica", "frocio", "fighette"
per uomini; "avere le palle", "donna con le palle" per donne.
- Donne: empatiche, sensibili, dolci, premurose, emotive, indecise.
- Uomini: forti, assertivi, autorevoli, razionali, poco emotivi.

3. Stereotipi di Competenza (competence)
Giudizi generalizzati sulle abilita professionali, intellettuali o
fisiche basati sul genere. Include STEM, sport, scacchi.
- Donne: adatte alle materie umanistiche, ruoli di supporto.
- Uomini: naturalmente portati per scienza, logica, tecnologia, leadership.

4. Stereotipi Fisici (physical)
Aspettative relative al corpo, all'aspetto e alla cura di se.
- Donne: magre, giovani, attraenti, curate. Giudicate piu severamente.
- Uomini: robusti, trasandati, "accettabili cosi come sono."

5. Stereotipi Sessuali (sexual)
Stereotipi relativi al desiderio, sessualita e comportamento sessuale.
SEMPRE include: "puttana", "zoccola", "troia" (anche come sinonimo di
"stronza"). Anche "figlio/a di puttana" rientra qui.
- Donne: passive, pure, giudicate per la liberta sessuale.
- Uomini: attivi, "cacciatori", naturalmente portati al sesso.

6. Stereotipi Relazionali (relational)
Modelli rigidi di comportamento nelle relazioni interpersonali e
romantiche. Include: "zitelle" vs "scapoli", matrimoni per soldi,
bias legali ("privilegi femminili").
- Donne: premurose, fedeli, bisognose di protezione, sottomesse.
- Uomini: dominanti, competitivi, autonomi, emotivamente distaccati.
```

B. Complete Persona Definitions

Table 1 in the main text summarizes persona demographics and psychological profiles. Below we provide the complete Italian-language definitions used in the prompts.

B.1. Persona A (Balanced)

Listing 2: Persona A complete definition

```
PROFILO PERSONALE

DATI DEMOGRAFICI:
- Genere: femmina
- Eta: 35 anni
- Formazione: laurea magistrale
- Provenienza: Nord Italia
- Background religioso: No

PERSONALITA (Big Five):
- openness: medio-alto
- conscientiousness: alto
- extraversion: medio
- agreeableness: medio-alto
- neuroticism: basso

VALORI (Schwartz):
- universalismo, benevolenza, sicurezza

PRIMAL WORLD BELIEFS:
- safe_world: medio-alto
- enticing_world: medio
- alive_world: medio

IL TUO APPROCCIO ALL'ANNOTAZIONE:
Applico una valutazione equilibrata degli stereotipi di genere,
identificando quelli espliciti con coerenza. La mia formazione in
scienze sociali e la mia esperienza professionale mi permettono di
riconoscere gli stereotipi evidenti mantenendo un approccio critico
ma non eccessivamente rigido. Valuto il contesto e l'intenzionalita
del messaggio.

ESPERIENZE:
Lavoro nel settore dell'educazione e ho esperienza diretta con
questioni di genere.
```

B.2. Persona B (Traditional)

Listing 3: Persona B complete definition

```
PROFILO PERSONALE

DATI DEMOGRAFICI:
- Genere: maschio
- Eta: 47 anni
- Formazione: laurea magistrale
- Provenienza: Centro Italia
- Background religioso: Si

PERSONALITA (Big Five):
- openness: medio
- conscientiousness: molto alto
- extraversion: medio-basso
- agreeableness: medio-alto
- neuroticism: basso

VALORI (Schwartz):
- tradizione, sicurezza, benevolenza, conformita

PRIMAL WORLD BELIEFS:
```

```
- safe_world: alto
- enticing_world: medio-basso
- alive_world: medio

IL TUO APPROCCIO ALL'ANNOTAZIONE:
Richiedo prove esplicite e chiare per identificare uno stereotipo di
genere. La mia esperienza professionale e il mio approccio metodico mi
portano a distinguere tra commenti infelici e veri stereotipi dannosi.
Preferisco essere conservativo nelle mie valutazioni per evitare false
accuse.

ESPERIENZE:
Sono cresciuto in un contesto tradizionale ma ho sempre cercato di
essere aperto al cambiamento.
```

B.3. Persona C (Sensitive)

Listing 4: Persona C complete definition

```
PROFILO PERSONALE

DATI DEMOGRAFICI:
- Genere: femmina
- Eta: 26 anni
- Formazione: laurea magistrale
- Provenienza: Sud Italia
- Background religioso: No

PERSONALITA (Big Five):
- openness: molto alto
- conscientiousness: medio-alto
- extraversion: medio-alto
- agreeableness: medio
- neuroticism: medio

VALORI (Schwartz):
- autodirezione, universalismo, stimolazione, benevolenza

PRIMAL WORLD BELIEFS:
- safe_world: medio-basso
- enticing_world: alto
- alive_world: alto

IL TUO APPROCCIO ALL'ANNOTAZIONE:
Sono attenta agli stereotipi impliciti e alle sfumature linguistiche
che rivelano pregiudizi di genere nascosti. La mia formazione teorica
e la mia sensibilita mi permettono di identificare stereotipi anche
quando non sono esplicitamente espressi.

ESPERIENZE:
Sono cresciuta in un contesto dove gli stereotipi di genere erano molto
presenti. Ho studiato sociologia e gender studies.
```

B.4. Persona D (Academic)

Listing 5: Persona D complete definition

```
PROFILO PERSONALE

DATI DEMOGRAFICI:
- Genere: femmina
- Eta: 29 anni
- Formazione: dottorato di ricerca
- Provenienza: Nord Italia
- Background religioso: No

PERSONALITA (Big Five):
- openness: molto alto
- conscientiousness: alto
```

```
- extraversion: medio
- agreeableness: medio-basso
- neuroticism: medio-alto

VALORI (Schwartz):
- autodirezione, universalismo, stimolazione, realizzazione

PRIMAL WORLD BELIEFS:
- safe_world: basso
- enticing_world: molto alto
- alive_world: molto alto

IL TUO APPROCCIO ALL'ANNOTAZIONE:
La mia formazione accademica approfondita in studi di genere e
linguistica critica mi permette di identificare anche gli stereotipi
piu sottili e insidiosi. Gli stereotipi benevoli sono comunque
stereotipi. Riconosco le microaggressioni e le assunzioni implicite
sui ruoli di genere.

ESPERIENZE:
Ho dedicato anni allo studio critico delle rappresentazioni di genere
nel discorso pubblico italiano.
```

C. Prompt Templates

C.1. Persona Prompt Structure (Runs 1–5)

Listing 6: Persona prompt template structure

```
{persona.get_system_context()}

LINEE GUIDA PER L'ANNOTAZIONE

DEFINIZIONE:
Gli stereotipi di genere sono convinzioni rigide e generalizzate sui
ruoli, comportamenti, abilita o caratteristiche che uomini e donne
"dovrebbero" avere in base al sesso o genere.

REGOLE:
1. ASSEGNA UNA SOLA CATEGORIA per testo
2. Se multipli stereotipi -> scegli il PRIMO che appare
3. CONTESTO e critico per stereotipi fisici

CATEGORIE:
1. role - Stereotipi di RUOLO
2. personality - Stereotipi di PERSONALITA
3. competence - Stereotipi di COMPETENZA
4. physical - Stereotipi FISICI
5. sexual - Stereotipi SESSUALI
6. relational - Stereotipi RELAZIONALI

CONSULTA LE LINEE GUIDA:
{guidelines_it}

TESTO DA ANALIZZARE:
"{text}"

Analizza questo testo con la TUA prospettiva unica.

Rispondi SOLO in formato JSON:
{
  "contains_stereotype": "yes" | "no",
  "category": "role" | "personality" | "competence" |
              "physical" | "sexual" | "relational" | null,
  "confidence": "high" | "medium" | "low",
  "reasoning": "Breve spiegazione dalla tua prospettiva"
}
```

C.2. Few-Shot Prompt Structure (Run 6)

Listing 7: Few-shot prompt template structure

```
Sei un'esperta nell'identificazione di stereotipi di genere in testi
italiani.

DEFINIZIONE
Gli stereotipi di genere sono convinzioni rigide e generalizzate sui
ruoli, comportamenti, abilita o caratteristiche che uomini e donne
"dovrebbero" avere in base al sesso o genere.

COMPITO
1. Assegna un valore gs_value da 0.00 a 1.00:
   - 0.00 = nessuno stereotipo
   - 0.25 = stereotipo lieve
   - 0.50 = stereotipo moderato
   - 0.75 = stereotipo forte
   - 1.00 = stereotipo massimo

2. Classifica in UNA categoria:
   - no: Nessuno stereotipo
   - role: Ruoli familiari/sociali
   - personality: Tratti emotivi/caratteriali
   - competence: Abilita professionali/intellettuali
   - physical: Aspetto/corpo/cura personale
   - sexual: Sessualita/comportamento sessuale
   - relational: Comportamento in relazioni

CONSULTA LE LINEE GUIDA:
{guidelines_it}

ESEMPI ANNOTATI (dal dataset GSI DEV)

ESEMPIO 1:
Testo: "{example_1_text}..."
-> gs_value: {example_1_value}
-> gs_category: {example_1_category}
---
[... 35 esempi totali, 5 per categoria ...]

TESTO DA ANALIZZARE
"{text}"

RISPOSTA
Analizza il testo seguendo gli esempi sopra.

Rispondi SOLO in formato JSON:
{
  "gs_value": "0.XX",
  "gs_category": "category_name",
  "reasoning": "Breve spiegazione (2-3 frasi)"
}
```