# Tiz at GSI:detect: Modeling Gender Stereotype Detection as Multi-Category Gender Stereotype Scoring

Tiziano Labruna

*University of Padua, 63 Via Trieste, Padova, 35121, Italy*

**Abstract**

Gender stereotypes in language contribute to the reinforcement of social inequalities and represent a critical challenge for natural language processing systems. The GSI:DETECT shared task addresses this issue by evaluating systems on their ability to detect and classify gender stereotypes in short Italian texts, combining a regression-based main task with a multi-class classification subtask. In this paper, we present an LLM-based approach inspired by strategy-aware reasoning methods, which explicitly models different types of gender stereotypes through a multi-step prompting framework. Our method decomposes the detection process by independently analyzing multiple stereotype categories and aggregating their contributions to estimate the overall degree of gender stereotype. The proposed approach is evaluated on both the main task of gender stereotype detection and the subtask of stereotype classification, showing the potential of structured reasoning and category-aware analysis for addressing nuanced forms of stereotype in text.

**Keywords**

Gender Stereotype Detection, Gender Stereotypes, Large Language Models, Prompting Strategies, Strategy-aware Reasoning, Regression-based Stereotype Modeling

## 1. Introduction

Gender stereotypes encoded in language play a key role in shaping and reinforcing social norms, influencing how roles, abilities, and behaviors are associated with different genders [1, 2]. As language technologies are increasingly deployed in real-world applications, the presence of such stereotypes has raised growing concerns regarding the fairness and social impact of NLP systems [3, 4]. Consequently, the automatic detection of gender stereotype has emerged as an important research problem, attracting sustained attention from the NLP community in the form of surveys, dedicated datasets, benchmarks, and shared tasks [5, 6, 7].

Most existing approaches to gender stereotype detection frame the task as a binary or categorical classification problem, typically relying on majority-voted annotations that collapse annotator disagreement into a single hard label [8, 9]. However, recent work has highlighted how this practice may obscure the inherently subjective and gradient nature of social bias, leading to an oversimplified representation of the phenomenon [10]. In response, several studies have advocated for perspectivist annotation frameworks that explicitly preserve annotator disagreement and model bias as a continuum rather than a binary property [11, 12]. Within this line of work, the GSI:DETECT shared task [13] at the EVALITA 2026 workshop [14] adopts a regression-based formulation, requiring systems to predict a continuous score reflecting the degree of gender stereotypical content in a text, alongside an auxiliary classification task identifying the stereotype category.

In parallel, Large Language Models (LLMs) have recently shown strong capabilities in a great number of different complex reasoning tasks. Despite these advances, the reliability of LLM-based judgments remains highly sensitive to the prompting strategy, making it crucial to explicitly assess their trustworthiness when no carefully designed prompting or supervision is applied [15, 16, 17]. Prior work has demonstrated that prompting strategies, in-context learning, and chain-of-thought reasoning can substantially improve LLM performance on tasks requiring structured reasoning [18, 19]. Building on these

*EVALITA 2026: 9th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Feb 26 – 27, Bari, IT*
tiziano.labruna@unipd.it (T. Labruna)
0000-0001-7713-7679 (T. Labruna)

advances, several approaches have explored the integration of LLM-based predictions with supervised models to improve robustness and alignment with human judgments in bias detection settings [20]. Furthermore, strategy-aware reasoning frameworks such as PCoT (Persuasion Chain-of-thought) [21] have shown that explicitly modeling high-level strategies can improve performance in related tasks like disinformation detection.

Inspired by these two lines of work, we propose a category-aware prompting approach tailored to gender stereotype detection. Unlike PCoT, our method does not consider persuasion strategies but instead focuses on the six gender stereotype categories defined in the GSI:DETECT task: *role*, *personality*, *competence*, *physical*, *sexual*, and *relational* stereotypes. Rather than injecting all categories simultaneously, we treat each category independently and perform a structured two-step prompting process for each of them.

In the first step, the LLM is prompted to generate an analytical paragraph assessing the presence of gender stereotype in the input text with respect to a single stereotype category. This step explicitly injects category-specific knowledge by including a description of the target stereotype type in the prompt, encouraging focused and interpretable reasoning. In the second step, the generated analysis is used to guide a decision or score generation, depending on the experimental setting.

For the zero-shot variant, the model outputs a binary decision indicating whether gender stereotype is present for the given category. These decisions are then aggregated across categories to determine whether the text contains gender stereotypes. For the fine-tuned variant, the model produces a numerical score on a fixed scale for each category. The resulting six-dimensional score vector is subsequently used as input to a supervised regression model, implemented as a multi-layer perceptron, which is trained on the development data to predict the final GS value in the range $[0, 1]$.

We also participate in the gender stereotype classification subtask by leveraging the category-wise scores produced by the fine-tuned model. The predicted stereotype category is selected as the one associated with the highest score, with ties resolved randomly. While this strategy is simple, it allows us to assess the extent to which category-specific stereotype signals captured during the main task can be reused for stereotype classification.

The main contributions of this work can be summarized as follows:

- We propose a category-aware, multi-step prompting framework for gender stereotype detection that explicitly models the six stereotype categories defined in the GSI:DETECT task.
- We introduce a two-stage LLM-based reasoning process in which category-specific analytical explanations are generated and subsequently used to guide stereotype detection and scoring.
- We present both a zero-shot and a fine-tuned variant of the proposed approach, showing how LLM-generated category-wise signals can be aggregated to model gender stereotype as a continuous phenomenon.
- We leverage category-specific stereotype scores to address the gender stereotype classification subtask, enabling reuse of the same signals across multiple evaluation settings.

## 2. Data and Resources

### 2.1. Dataset

We base our experiments on the dataset released for the GSI:DETECT shared task [13], which targets the detection and classification of gender stereotypes in short Italian texts. The dataset, considering only the portion made available to participants for development, consists of 200 annotated texts.

Each instance is provided in `.jsonl` format and includes the following fields: a unique identifier, the textual content, a list of four independent binary annotations indicating the presence or absence of a gender stereotype, a continuous GS value in the range $[0, 1]$ derived from the aggregation of annotators' judgments, a gender stereotype category label, and a binary flag specifying whether additional contextual information is required to interpret the text. In cases where context is required, it is explicitly included in the textual content and precedes the main text.

The GS value reflects the degree of stereotypical content and is computed by combining the four binary annotations rather than applying majority voting. Consequently, GS values close to the extremes (0 or 1) correspond to high inter-annotator agreement, while intermediate values capture varying degrees of disagreement. For the classification subtask, each instance is assigned to one of six predefined gender stereotype categories, i.e. *role*, *personality*, *competence*, *physical*, *sexual*, and *relational*, using majority voting or adjudication by a super-judge when necessary.

### 2.2. Data Augmentation

Given the limited size of the development data and the need to train a supervised regression model for the fine-tuned variant of our approach, we performed a form of data augmentation at the score level. Starting from the original 200 instances, we leveraged the category-wise scores produced by the LLM (which will be more precisely described in the next sections) to generate additional training samples while preserving the relative relationships among stereotype categories.

For each instance, we consider the six category-specific scores (one per gender stereotype category) and generate augmented variants by applying small bounded perturbations to these scores. Augmented samples are retained only if they preserve the original ranking relationships among categories, thereby maintaining consistency in the relative strength of different stereotype signals. This constraint prevents unrealistic inversions while allowing controlled variability in the score space.

This process results in multiple augmented versions for each original instance, effectively increasing the amount of training data available for the supervised learner without altering the underlying semantic content of the texts.

### 2.3. Data Splits

For the fine-tuning experiments, we construct train, development, and test splits at the level of original instances, ensuring that all augmented variants derived from the same base text are assigned to the same split. This prevents information leakage across splits and preserves a clean evaluation setup. We adopt a $60\%$ / $20\%$ / $20\%$ split for training, development, and testing, respectively, using a fixed random seed for reproducibility. For each split, both the augmented score representations and the corresponding duplicated annotations are generated and aligned via consistent identifiers.

### 2.4. Computational Resources

All LLM-based experiments were conducted using the *Gemma-3.1-12B* model, accessed through the Hugging Face ecosystem without any task-specific fine-tuning. The model was employed exclusively in inference mode and used to generate both category-specific analyses and numerical stereotype scores through prompting. Inference was performed on a single NVIDIA A40 GPU, using mixed-precision computation to ensure efficiency and stability.

No external fine-tuning of the language model was performed. The only supervised component of our system is the multi-layer perceptron used in the fine-tuned variant to map category-wise LLM scores to a final GS value in the $[0, 1]$ range.

## 3. Methodology

Our approach models gender stereotype as a multidimensional phenomenon by decomposing the detection process across the six gender stereotype categories defined in the GSI:DETECT task: *role*, *personality*, *competence*, *physical*, *sexual*, and *relational*. Instead of prompting the model with all categories simultaneously, we adopt a category-aware, one-task-at-a-time strategy inspired by strategy-aware reasoning frameworks. Each category is analyzed independently through a structured two-step prompting procedure, and the resulting signals are subsequently aggregated to produce the final prediction.

### 3.1. Overview of the Framework

Given an input text, the system performs the following steps:

1. Independently analyze the text with respect to each gender stereotype category using a category-specific prompt.
2. Generate an explicit analytical explanation describing whether and how the stereotype category appears in the text.
3. Use the generated analysis to guide a second prompt that produces either a binary decision or a numerical score.
4. Aggregate category-level outputs to obtain a final gender stereotype score and category prediction.

This design encourages focused reasoning, reduces interference between categories, and makes intermediate model outputs interpretable.

### 3.2. Step 1: Category-Specific Analysis (Knowledge Infusion)

For each input text, we run six independent prompts, one for each gender stereotype category. Each prompt injects structured knowledge derived from the official task guidelines, explicitly defining the stereotype category, providing examples, and describing its social implications.

The system prompt establishes the model as an expert in a specific stereotype category, while the user prompt asks for a conservative assessment of whether that category is present in the text.

An example of a category-specific prompt for *[ROLE_STEREOTYPES]* is reported below:

> ***System prompt:***
> *You are an assistant who detects gender stereotypes in text. Gender stereotypes are rigid and generalized beliefs about the roles, behaviors, abilities, or characteristics that men and women "should" have based on their sex or gender.*
> *Your expertise and focus is on **[ROLE_STEREOTYPES]**, meaning social and cultural expectations that define which family or social roles are "appropriate" for men and women.*
> *Men: breadwinners, heads of the household.*
> *Women: mothers, housewives, responsible for domestic work.*
> *You are the expert who detects **[ROLE_STEREOTYPES]**.*
>
> ***User prompt:***
> *Given a certain text, critically evaluate its gender stereotype potential. Additionally, analyze whether the text employs the **[ROLE_STEREOTYPES]** category. Explain how **[ROLE_STEREOTYPES]** appears or does not appear in the text. Be conservative in your assessment: if you are not fully certain that the category is used, assume it is not present.*

Analogous prompts are used for the remaining five categories, with the placeholder **[CATEGORY]** replaced by the corresponding stereotype definition and examples. The output of this step is a short analytical paragraph describing the presence or absence of the target stereotype category.

### 3.3. Step 2: Decision or Scoring Based on the Analysis

The analysis generated in Step 1 is then passed to a second prompt, together with the original text. This second step uses a general gender stereotype detection instruction and leverages the previously generated analysis as explicit reasoning input. We consider two variants for this step.

**Binary (Zero-Shot) Variant.** In the zero-shot setup, the model is asked to produce a binary decision indicating whether gender stereotype is present with respect to the analyzed category.

> **System prompt:** You are a Gender Stereotype Detector. Your goal is to detect if a text presents gender stereotypes. Gender stereotypes are rigid and generalized beliefs about the roles, behaviors, abilities, or characteristics that men and women "should" have based on their sex or gender. You will respond with either Yes or No.
>
> **User prompt:** Given a text and an analysis of potential gender stereotypes or biased expressions it may contain, you must respond with ONLY Yes or No based on if you think that there is gender stereotype in the text or not.

The output is mapped to a binary value, where *Yes* corresponds to 1 and *No* to 0.

**Score-Based Variant.** In the score-based setup, used for the fine-tuned approach, the model produces a numerical score reflecting the degree of gender stereotype.

> **System prompt:** You are a Gender Stereotype Detector. Your goal is to evaluate the degree of gender stereotypes present in a text. Gender stereotypes are rigid and generalized beliefs about the roles, behaviors, abilities, or characteristics that men and women s̈houldḧave based on their sex or gender. You will provide a score from 1 to 10.
>
> **User prompt:** Given a text and an analysis of potential gender stereotypes or biased expressions it may contain, you must respond with a number from 1 to 10 based on the degree of gender stereotype in the text, preceded by a brief explanation of your reasoning. Give your answer in the form of a dictionary: {“explanation”: “Brief explanation of the reasoning that led you to this score.”, “response”: “Score from 1 to 10.”}

Each category thus produces either a binary indicator or a numerical score.

### 3.4. Zero-Shot Aggregation

In the binary setting, each text produces six binary outputs, one per category. We experiment with multiple aggregation strategies to derive a final prediction. The specific strategies will be presented in the next section. For category prediction, the stereotype category associated with the highest score is selected, with ties broken randomly.

### 3.5. Supervised Aggregation via Multi-Layer Perceptron

In the fine-tuned variant, the category-wise scores produced by the LLM are further combined using a supervised learning component. Specifically, we train a Multi-Layer Perceptron (MLP) to map the six-dimensional vector of category scores to the final gender stereotype signal required by the main task.

For each text, the MLP receives as input the six scores generated by the LLM, one for each gender stereotype category. In addition to these raw scores, we optionally augment the input with simple descriptive statistics computed over the six values. Specific experimental setting details can be found in the next section.

**Model Architecture.** The aggregation model is implemented as a feed-forward neural network composed of:

- an input layer matching the feature dimensionality,
- one or two hidden layers with ReLU activations,
- a final linear layer producing four outputs.

The model outputs correspond to the four binary annotations provided by the human annotators. This design choice allows the network to learn from annotator-level disagreement rather than directly regressing to a single aggregated label.

# 4. Experimental Setup

This section describes the experimental configurations adopted for both the zero-shot and the fine-tuned variants of our approach.

## 4.1. Zero-Shot Aggregation Strategies

In the zero-shot setting, the methodology described in Section 3 is applied without using any task-specific training data. For the main task, the LLM produces a binary score (0 or 1) for each of the six gender stereotype categories. As a result, each input text is associated with a six-dimensional binary vector.

Since the task requires a single continuous score in the $[0, 1]$ range, we explore multiple heuristic aggregation strategies to combine the six binary signals into a final prediction. The following five strategies are considered:

- **Max-One**: the final score is set to 0 if at most one category score is 1, and to 1 otherwise. (*run_1*)
- **Max-Two**: the final score is set to 0 if at most two category scores are 1, and to 1 otherwise. (*run_2*)
- **Max-Three**: the final score is set to 0 if at most three category scores are 1, and to 1 otherwise. (*run_3*)
- **All Zeros**: the final score is set to 0 if all six category scores are 0 (i.e., no category is flagged as biased), and to 1 otherwise. (*run_4*)
- **Average**: the final score is computed as the average of the six binary values, rounded to two decimal places. (*run_5*)

As this is a purely zero-shot approach, no development data is used to select or tune these strategies. Consequently, the chosen aggregation rules are not optimized on the task data, and other untested strategies might potentially yield better performance.

## 4.2. Fine-Tuned Aggregation Experiments

The fine-tuned variant represents the primary focus of this work. In this setting, the aggregation of category-wise signals is learned via a supervised Multi-Layer Perceptron, as described in Section 3.5.

Unlike the zero-shot setup, the MLP operates on the six continuous scores in the $[1, 10]$ range generated by the LLM for each category. We conduct an extensive set of experiments to identify the most effective configuration. Specifically:

- **add_scores** indicates that the six raw category scores are used as input features.
- **add_stats** indicates that three additional statistics computed over the six scores (mean, variance, entropy) are included.
- **threshold** denotes that the annotator-level probabilities predicted by the MLP are binarized using the specified threshold before aggregation.
- **augmented** indicates that the MLP is trained on the augmented version of the training and development sets, generated as described in Section 3.5.

For all experiments, hyperparameters are selected via grid search on the development set using the following configuration:

```
hidden_dim: {32, 64, 128}
hidden_layers: {1, 2}
lr: {1e-3, 1e-4}
batch_size: {64, 128}
```

Performance is evaluated using the 1/1+MSE (Mean Squared Error) metric, in line with the official evaluation metric of the task. Table 1 reports the scores obtained across all fine-tuning configurations.

| **Configuration** | $1/(1+\textbf{MSE})$ |
|---|---|
| base + add_scores (*run_1*) | 0.8137 |
| base + add_stats (*run_2*) | 0.8317 |
| base + add_scores + add_stats (*run_3*) | 0.8316 |
| base + threshold 0.4 | 0.7407 |
| base + threshold 0.5 | 0.7767 |
| base + threshold 0.6 | 0.7767 |
| base + threshold 0.7 | 0.7881 |
| base + threshold 0.8 | 0.8000 |
| base + threshold 0.9 | 0.7111 |
| base + add_scores + threshold 0.5 | 0.7729 |
| base + add_stats + threshold 0.5 | 0.7729 |
| base + add_stats + threshold 0.8 | 0.7729 |
| base + add_scores + add_stats + threshold 0.5 | 0.7767 |
| augmented + add_scores | 0.8215 |
| augmented + add_stats | 0.8169 |
| augmented + add_scores + add_stats | 0.8211 |
| augmented + threshold 0.4 | 0.7148 |
| augmented + threshold 0.5 | 0.7431 |
| augmented + threshold 0.6 | 0.7645 |
| augmented + threshold 0.7 | 0.7541 |
| augmented + threshold 0.8 | 0.7556 |
| augmented + add_scores + threshold 0.5 | 0.7568 |
| augmented + add_stats + threshold 0.5 | 0.7622 |
| augmented + add_stats + threshold 0.6 | 0.7741 |
| augmented + add_scores + add_stats + threshold 0.5 | 0.7332 |

**Table 1**
Performance scores computed as $1/(1+\text{MSE})$. Higher values indicate better performance. Cell color intensity increases with performance, from white (lower score) to darker green (higher score).

## 5. Discussion

The experimental results highlight several insights regarding both the effectiveness of the proposed aggregation strategies and the broader implications of modeling gender stereotype as a category-structured phenomenon.

Unlike the zero-shot aggregations strategies, which provide a simple and intuitive way to map category-wise binary signals to a final stereotype score, the fine-tuned aggregation detects gender stereotype by learning how to combine category-wise signals.

Across both base and augmented settings, configurations that rely on continuous category scores without thresholding tend to perform better, suggesting that preserving the full range of LLM-generated scores allows the model to capture more nuanced signals of stereotype, rather than losing information through aggressive binarization or thresholding.

The use of augmented data produces mixed results. While augmentation slightly improves performance in some configurations, it does not consistently outperform the best base models. This suggests that, although augmentation increases data diversity, the quality and alignment of augmented examples with real annotator behavior remain critical factors.

Beyond raw performance, the results support the underlying theoretical motivation of this work. By decomposing gender stereotype into interpretable categories, the model is encouraged to reason explicitly about different manifestations of stereotypes rather than collapsing them into a single undifferentiated signal. The fact that supervised aggregation over these category-wise signals yields strong results indicates that such a decomposition is not only conceptually meaningful, but also practically useful.

Importantly, the learned aggregation weights implicitly reflect the relative contribution of different stereotype categories as perceived by human annotators. This aligns with the idea that gender stereotype

is a multidimensional and graded phenomenon, where different dimensions interact and contribute unevenly to the final judgment.

## 6. Conclusions

This work investigated gender stereotype detection through a structured and category-aware framework that combines LLM-based reasoning with supervised aggregation. Instead of modeling gender stereotype as a single monolithic signal, the proposed approach decomposes it into six interpretable stereotype categories and studies how these dimensions contribute to human judgments of stereotype.

Simple zero-shot aggregation strategies provide a straightforward and interpretable way to combine category-level scores, offering an intuitive baseline. Building on this, we hypothesize that there may be a more nuanced interplay between the individual scores generated by the model, and that learning how these different dimensions interact could lead to more accurate and informative assessments. To explore this idea, we experiment with a lightweight supervised aggregation model that takes as input the raw category scores, allowing the system to learn patterns in how different stereotype dimensions jointly contribute to the overall assessment.

The findings support the theoretical view that gender stereotype is multidimensional, context-dependent, and unevenly distributed across different stereotype categories. By explicitly modeling these dimensions and learning their aggregation from data, the proposed framework offers a flexible and interpretable alternative to end-to-end stereotype classifiers. This design allows the system to adapt to different annotation schemes or definitions of stereotype, while maintaining transparency in how individual stereotype dimensions influence the final outcome.

## Declaration on Generative AI

During the preparation of this work, the author used GPT-5 and Gemini-3 in order to conduct grammar and spelling check. After using these tools, the author reviewed and edited the content as needed and takes full responsibility for the publication's content.

## References


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

[2] S. Cavagnoli, F. Dragotto, et al., Sessismo, Mondadori Education, 2021.

[3] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, A. T. Kalai, Man is to computer programmer as woman is to homemaker? debiasing word embeddings, Advances in neural information processing systems 29 (2016).

[4] A. Caliskan, J. J. Bryson, A. Narayanan, Semantics derived automatically from language corpora contain human-like biases, Science 356 (2017) 183–186.

[5] K. Stanczak, I. Augenstein, A survey on gender bias in natural language processing, arXiv preprint arXiv:2112.14168 (2021).

[6] A. T. Cignarella, A. Giachanou, E. Lefever, A survey on stereotype detection in natural language processing, ACM Computing Surveys 58 (2025) 1–33.

[7] Y. Li, G. Zhang, H. Hong, Y. Wang, C. Lin, Overview of the nlpcc 2025 shared task: Gender bias mitigation challenge, in: CCF International Conference on Natural Language Processing and Chinese Computing, Springer, 2025, pp. 453–464.

[8] T. Davidson, D. Warmsley, M. Macy, I. Weber, Automated hate speech detection and the problem of offensive language, in: Proceedings of the international AAAI conference on web and social media, volume 11, 2017, pp. 512–515.

[9] M. Sap, D. Card, S. Gabriel, Y. Choi, N. A. Smith, The risk of racial bias in hate speech detection, in: Proceedings of the 57th annual meeting of the association for computational linguistics, 2019, pp. 1668–1678.
[10] M. Zakizadeh, M. T. Pilehvar, Blind men and the elephant: Diverse perspectives on gender stereotypes in benchmark datasets, arXiv preprint arXiv:2501.01168 (2025).
[11] S. Frenda, G. Abercrombie, V. Basile, A. Pedrani, R. Panizzon, A. T. Cignarella, C. Marco, D. Bernardi, Perspectivist approaches to natural language processing: a survey, Language Resources and Evaluation 59 (2025) 1719–1746.
[12] B. Chulvi, L. Fontanella, R. Labadie-Tamayo, P. Rosso, et al., Social or individual disagreement? perspectivism in the annotation of sexist jokes, in: CEUR Workshop Proceedings, volume 3494, 2023.
[13] 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.
[14] 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.
[15] T. Labruna, A. Lavelli, B. Magnini, Testing chatgpt for stability and reasoning: A case study using italian medical specialty tests, in: Proceedings of the Ninth Italian Conference on Computational Linguistics (CLiC-it 2023), 2023.
[16] T. Labruna, S. Brenna, G. Bonetta, B. Magnini, Are you a good assistant? assessing llm trustability in task-oriented dialogues, in: Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024), 2024, pp. 470–477.
[17] T. Labruna, S. Gallo, G. Da San Martino, Positional bias in binary question answering: How uncertainty shapes model preferences, in: Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025), 2025, pp. 550–560.
[18] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al., Chain-of-thought prompting elicits reasoning in large language models, Advances in neural information processing systems 35 (2022) 24824–24837.
[19] T. Labruna, G. Bonetta, B. Magnini, Task-oriented dialogue systems through function calling, RANLP 2025 (2025) 614.
[20] J. Tavarez-Rodríguez, F. Sánchez-Vega, A. Rosales-Pérez, A. P. López-Monroy, Better together: Llm and neural classification transformers to detect sexism, Working Notes of CLEF (2024).
[21] A. Modzelewski, W. Sosnowski, T. Labruna, A. Wierzbicki, G. Da San Martino, Pcot: Persuasion-augmented chain of thought for detecting fake news and social media disinformation, in: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2025, pp. 24959–24983.