EVALITA 2026: Overview of the 9th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian

Francesco Cutugno[1], Alessio Miaschi[2], Alessio Palmero Aprosio[3], Giulia Rambelli[4], Lucia Siciliani[5] and Marco Antonio Stranisci[6,7]

[1]*Department of Electrical Engineering and Information Technology, University of Naples Federico II, Naples, Italy*
[2]*Institute for Computational Linguistics "A. Zampolli" (CNR-ILC) - ItaliaNLP Lab, Pisa, Italy*
[3]*University of Trento, Trento, Italy*
[4]*Aix-Marseille University | ILCB, Aix-en-Provence, France*
[5]*Department of Computer Science, University of Bari Aldo Moro, Bari, Italy*
[6]*Università degli Studi di Torino, Torino, Italy*
[7]*aequa-tech, Torino, Italy*

Abstract

EVALITA is a shared evaluation campaign designed to assess and compare Natural Language Processing (NLP) and Speech Technologies through tasks proposed by the Italian research community. It provides a common framework for addressing open linguistic challenges and real-world applications, with increasing attention to multilingual and multimodal settings. The 2026 edition included 10 tasks and attracted 54 participating groups from 13 countries, confirming the growing international interest in the initiative. The workshop presents the results of the evaluation campaign, highlighting the widespread adoption of Large Language Models and the organizers' effort in designing challenging tasks aimed at meaningfully evaluating and stress-testing such models.

Keywords

NLP, Evaluation, Italian, Language Models

1. Introduction

The Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA) is the biennial initiative aimed at promoting the development of language and speech technologies for the Italian language. EVALITA is promoted by the Italian Association of Computational Linguistics (AILC)[1] and it is endorsed by the Italian Association for Artificial Intelligence (AIxIA)[2] and the Italian Association for Speech Sciences (AISV)[3].

EVALITA provides a shared framework where different systems and approaches can be scientifically evaluated and compared across a diverse set of tasks, proposed and structured by the Italian research community. The tasks are designed to represent significant scientific challenges, allowing methods, resources, and systems to be tested against shared benchmarks that capture both linguistic open issues and real-world applications. In the current NLP landscape, increasingly shaped by the advent of Large Language Models (LLMs), the organization of these tasks offers a unique opportunity to investigate how traditional and modern approaches perform in a rapidly evolving domain. The collected datasets continue to provide valuable resources for researchers to explore both established and emerging

EVALITA 2026: 9th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Feb 26 – 27, Bari, IT
cutugno@unina.it (F. Cutugno); alessio.miaschi@ilc.cnr.it (A. Miaschi); a.palmeroaprosio@unitn.it (A. P. Aprosio); giulia.RAMBELLI@univ-amu.fr (G. Rambelli); lucia.siciliani@uniba.it (L. Siciliani); marcoantonio.stranisci@unito.it (M. A. Stranisci)
0000-0001-9457-6243 (F. Cutugno); 0000-0002-0736-5411 (A. Miaschi); 0000-0002-1484-0882 (A. P. Aprosio); 0000-0003-3379-6941 (G. Rambelli); 0000-0002-1438-280X (L. Siciliani); 0000-0001-9337-7250 (M. A. Stranisci)

[1]http://www.ai-lc.it
[2]http://www.aixia.it
[3]http://www.aisv.it

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-04-16

problems in Italian NLP, fostering the development of innovative solutions and enabling a community-wide discussion on the evolving capabilities and limitations of modern language technologies. While some tasks remain a core element of the evaluation campaign, new challenges are regularly introduced to reflect the changing research priorities and technological advances in the field.

This paper introduces the tasks proposed at EVALITA 2026 and provides an overview of the participants and systems whose descriptions and obtained results are reported in these Proceedings. The EVALITA 2026 final workshop, held in Bari on February 26-27th, counts 10 different tasks. In particular, the selected tasks are grouped into four research areas (tracks) according to their objective and characteristics, namely: (i) *Authorship Analysis*; (ii) *Computational Ethics*; (iii) *Emerging Tasks*; (iv) *Multimodality*; (v) *New Challenges in Long-Standing Tasks*; (vi) *Pragmatics*.

2. EVALITA 2026 Tracks and Tasks

In the 2026 edition of EVALITA, 11 different tasks were proposed, peer-reviewed, and accepted. Subsequently, one organizing team withdrew, and as a result, 10 tasks were ultimately presented at EVALITA. Data were produced by the task organizers and made available to the participants. For the future availability of this data, we are going to release them on GitHub[4], in accordance with the terms and conditions of the respective data sources. Such a repository will also reference alternative repositories managed by the task organizers. The tasks of EVALITA 2026 are grouped according to the following tracks corresponding to five broad research areas:

Authorship Analysis

DeSegMa-It: DEtection and SEgmentation of MAchine generated texts in Italian [1]. DeSegMa-It is the first shared task aimed at testing the robustness of machine-generated text (MGT) detectors by evaluating their performance under settings where the IID (i.e. independent and identically distributed) assumption does not hold. In particular, the shared task was organized around two subtasks: i) document-level detection of MGTs and ii) human-machine text segmentation.

Computational Ethics

GSI:detect – Detecting Gender Stereotypes in Italian [2]. GSI:detect is the shared task for the recognition and classification of gender stereotypes (GSs). The task is organized into two subtasks: i) GS detection, in which systems have to assign to a text a GS value and ii) GS classification, in which systems are required to assign a score to each input text from a set of six predefined categories (e.g. role, relational, etc.).

MultiPRIDE: Multilingual Automatic Detection of Reclamation of Slurs in the LGBTQ+ Context [3]. MultiPRIDE is the shared task focused on the detection of slur reclamation, a linguistic phenomenon in which derogatory terms are reappropriated by members of a community. The task is divided into two main tasks: i) binary detection of reclamation in social media posts and ii) contextual detection of reclamation considering users' biographies. Each task is further divided into language-specific subtasks (Italian, Spanish and English).

Emerging Tasks

Cruciverb-IT: Crossword Solving [4]. Cruciverb-IT is the first shared task on Italian crossword solving. The task comprises two subtasks: 1) *clue answering*, aimed at answering individual crossword clues given the expected answer length, and 2) *grid solving*, focused on autonomously solving complete crossword grids of varying sizes.

[4] https://github.com/evalita.

SVELA: Selective Verification of Erasure from LLM Answers [5]. SVELA is the shared task on machine unlearning verification in Large Language Models. Participants must determine whether specific fictional identities are retained, forgotten (targeted by unlearning), or never seen during training. The evaluation is multilingual (Italian, Spanish, French, German) and tests both entity-level and instance-level detection.

Multimodality

EVWSD-ITA: Enhanced Visual Word Sense Disambiguation for Italian [6]. EVWSD-ITA is the shared task on Visual Word Sense Disambiguation (VWSD). Specifically, the task involves selecting the most appropriate image for a target word within a given context, specifically designed to include "hard negatives" (i.e. co–hyponyms that share a common hyperonym but represent distinct concepts).

New Challenges in Long-Standing Tasks

ATE-IT: Automatic Term Extraction – Italian Testbed [7]. ATE-IT is the shared task on Automatic Term Extraction (ATE) for Italian, focused on the domain of municipal waste management. The task is organized into two subtasks: i) *Term extraction*, aiming at identifying domain-specific terms in instructional texts and ii) *Term Variants Clustering*, focusing on grouping morphological and semantic term variants.

PFB: Prometeia Financial Benchmark – Benchmarking Language Models in the Financial Domain [8]. The Prometeia Financial Benchmark task evaluates how well large and small language models understand financial-domain content. It introduces a multilingual benchmark of about 1,000 multiple-choice questions, each with a correct answer and difficulty level. Participants are asked to build or adapt systems to answer these domain-specific MCQs. Two tracks are available: (i) an Italian-only corpus; (ii) a multilingual corpus (Italian, English, Turkish).

Pragmatics

FadeIT: Fallacy Detection in Italian Social Media Texts [9]. FadeIT is the first shared task on fallacy detection in social media texts in Italian. The shared task is articulated into two subtasks: i) *post-level fallacy detection*, aiming at predicting the fallacy types expressed in each input post, and ii) *span-level fallacy detection*, aiming at predicting all text segments expressing any given fallacy type in each input post.

IMPOLS: IMplicit contents in POLitical Speech [10]. IMPOLS is the shared task on automatic detection and classification of implicit contents in political speeches. The task is divided into three subtasks: i) implicit content detection; ii) implicit classification and iii) implicature type classification.

3. Participation

EVALITA 2026 attracted the interest of a large number of researchers from academia and industry, for a total of 54 single teams composed of about 127 individuals participating in one or more of the 10 proposed tasks. After the evaluation period, 57 system descriptions were submitted (reported in these proceedings), i.e., a 11.8% increase with respect to the previous EVALITA edition [11]. Moreover, task organizers allowed participants to submit more than one system result (called runs), for a total of 274 submitted runs.

Overall, EVALITA 2026 involved participants from 13 different countries. As expected, Italy was the most represented country, with a total of 44 teams having at least one Italian affiliation. Figure 1 shows the geographical distribution of participating teams excluding Italy, highlighting the international reach of the campaign. Contributions originated from several regions worldwide, with Vietnam being the most represented foreign country, accounting for 5 teams with at least one affiliated author. This

Track	Task	Teams	Runs
Authorship Analysis	DeSegMa-It	6	21
Computational Ethics	GSI:detect	6	91
	MultiPRIDE	18	83
Emerging Tasks	Cruciverb-IT	5	17
	SVELA	4	14
Multimodality	EVWSD-ITA	1	3
New Challenges in Long-Standing Tasks	ATE-IT	9	13
	PFB	2	4
Pragmatics	FadeIT	7	25
	IMPOLS	2	3

Table 1
Number of participating teams and number of runs organized by track and task. The data reported is an overestimation with respect to the systems described in the proceedings (e.g. teams participating in more than a task are counted according to the number of tasks they participated in).

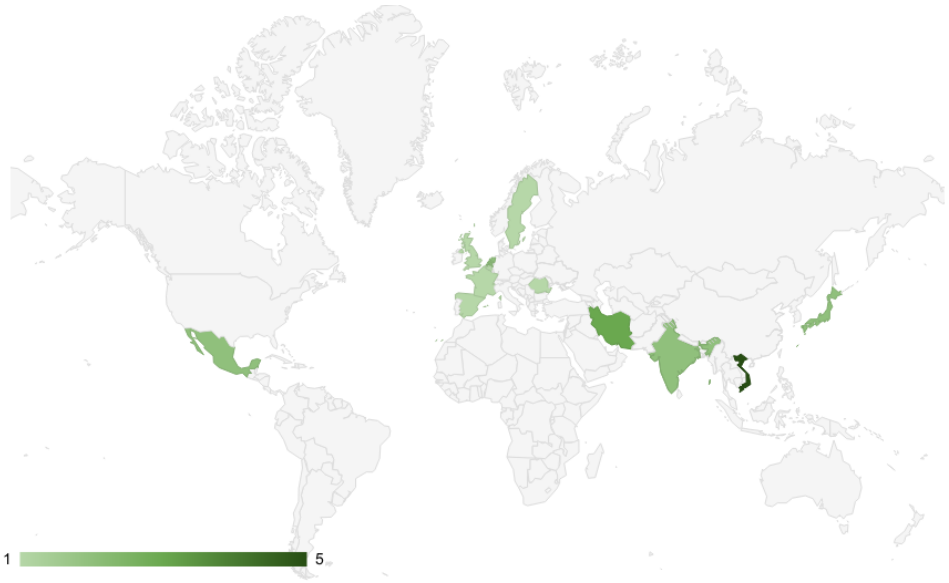

Figure 1: Number of teams per country (excluding Italy).

is followed by Iran, and by India, Japan, Mexico, and the Netherlands with 2 teams each, while all remaining countries are represented by a single team. Among the proposed tasks, MultiPRIDE stands out as the one attracting the largest international participation, with 8 teams from eight different non-Italian countries. This is followed by Fade-IT and GSI:detect, which also show a notable level of international engagement. From an institutional perspective, EVALITA 2026 involved a total of 35 affiliations, including 29 academic institutions and 6 industrial organizations, further confirming the predominantly academic nature of the participation, with a non-negligible contribution from industry.

Table 1 shows the different tracks and tasks along with the number of participating teams and submitted runs. The data reported in the table is based on information provided by the task organizers at the end of the evaluation process. Such data represents an overestimation with respect to the systems described in the proceedings. Participation was quite imbalanced across different tracks and tasks, as reported in Figure 2: each rectangle represents a task whose size reflects the number of participants, while the color indicates the corresponding track. The *Computational Ethics* track emerges as the most attended overall, with MultiPRIDE and GSI:detect standing out as the two most participated tasks of the campaign. While MultiPRIDE attracted the largest number of teams, GSI:detect resulted in a higher number of submitted runs, indicating a particularly strong engagement from participating groups. This interest can be linked to the societal relevance of the addressed phenomena, as well as to

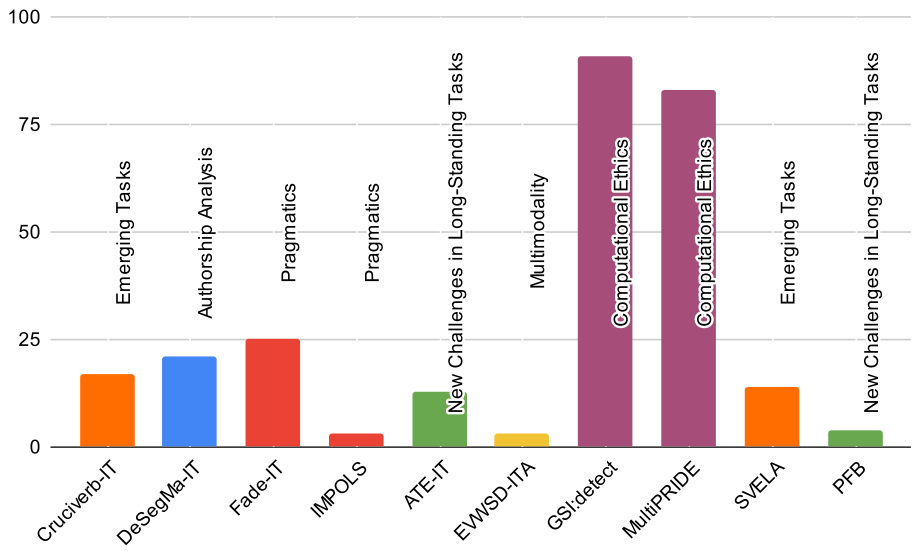

Figure 2: Number of participating teams organized by track (color) and task.

the multilingual and classification-oriented nature of the tasks. A substantial level of participation is observed in the Pragmatics track, where FadeIT attracted a relatively large number of teams, reflecting the strong interest in challenges related to fallacy detection and the analysis of pragmatic phenomena in social media. In the same track, IMPOLS shows a more limited participation, possibly reflecting both the higher complexity of modeling implicit meaning in political discourse and the additional challenges posed by its multimodal setting, which involves the joint processing of speech audio and textual transcriptions. The Authorship Analysis track, which in this edition includes DeSegMa-It, still exhibits a solid level of participation, confirming the relevance of research on machine-generated text detection within the Italian NLP community. Among the *Emerging Tasks*, Cruciverb-IT and SVELA exhibit different participation patterns. Cruciverb-IT attracted a moderate number of teams, likely due to the novelty and specificity of crossword solving in Italian, while SVELA reflects a growing interest in machine unlearning and verification in Large Language Models. The *Multimodality* track, represented by EVWSD-ITA, shows a more contained participation, consistent with the higher entry barriers typically associated with multimodal evaluation settings. Finally, tasks belonging to the *New Challenges in Long-Standing Tasks* track display heterogeneous participation. ATE-IT continues to attract a solid number of teams, confirming sustained interest in automatic term extraction, whereas PFB, despite its industrial relevance and multilingual design, involves a smaller number of participants, likely due to the domain-specific nature of the benchmark. In line with the effort to foster a more international and inclusive evaluation campaign, task organizers were encouraged to propose multilingual tasks, providing datasets in Italian and in one or more additional languages. In the 2026 edition, two tasks explicitly adopted a multilingual setting, namely MultiPRIDE, which includes Italian, Spanish and English, and PFB, which covers Italian, English and Turkish. Although the number of multilingual tasks is still limited, this represents a further step towards strengthening the international profile of EVALITA and broadening its impact beyond the Italian-speaking research community.

Turning to the types of systems submitted by participants, several interesting patterns emerge from the analysis. As shown in Figure 3, decoder-only approaches are not the most widely adopted, despite their prominence in recent NLP research. Encoder-only architectures remain the predominant choice, while a non-negligible number of teams also rely on non–LLM-based, symbolic, and encoder–decoder approaches. This distribution indicates that participants did not simply apply Large Language Models in inference-only settings, but instead explored task-specific modeling strategies aimed at building effective systems. In turn, this reflects the fact that the proposed tasks were, on average, sufficiently challenging and, in many cases, not readily solvable by directly applying Large Language Models without additional

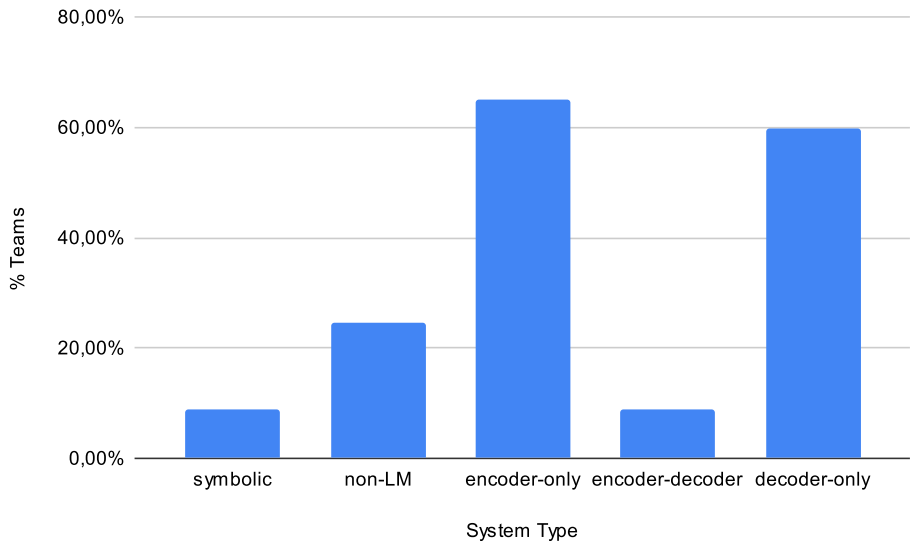

Figure 3: Percentage of the systems submitted to EVALITA 2026 according to the five categories of: symbolic, non-LM, encoder-only, encoder-decoder, decoder-only.

modeling choices. In this respect, it is worth noting that several teams still adopted traditional machine learning pipelines based on feature extraction, often leveraging linguistically motivated or task-specific features, rather than relying exclusively on end-to-end neural architectures. Moreover, a substantial number of submissions made use of data augmentation strategies, frequently involving the generation of synthetic data to enrich the training material. This practice represents an emerging trend in the campaign and is in line with recent work exploring the use of synthetic data to improve robustness and generalization in NLP systems.

As a final remark, we highlight the case of a team that participated in multiple tasks (i.e. DeSegMa, MultiPRIDE, IMPOLS and FadeIT), which submitted systems for four different tasks while adopting a unified modeling approach. In particular, the team relied on the same core architecture across all submissions, leveraging a BabyLM model pre-trained on a dedicated Italian dataset.

4. "Federico Sangati" Best System Across Tasks Awards

In line with the previous edition, we confirmed the award for the best system across-task. Starting from this year, the award is dedicated to the memory of Federico Sangati. The award was introduced with the aim of fostering student participation in the evaluation campaign and in the workshop. A committee of 4 members (Dominique Brunato, Camilla Casula, Alessandro Mazzei and Marco Polignano) was asked to choose the best system across tasks. The composition of the committee is balanced with respect to the level of seniority as well as to their academic background (computer science-oriented vs. humanities-oriented). In order to select a short list of candidates, the task organizers were invited to propose candidate system participating in their tasks during the reviewing process (not necessarily top-ranking). The committee was provided with the list of candidate systems and the criteria for eligibility, based on:

- *novelty* with respect to the state of the art;
- *originality*, in terms of identification of new linguistic resources, identification of linguistically motivated features, and implementation of a theoretical framework grounded in linguistics;
- *critical insight*, paving the way to future challenges (deep error analysis, discussion on the limits of the proposed system, discussion of the inherent challenges of the task);
- *technical soundness* and *methodological rigor*.

We collected 13 system nominations from the organizers of 9 tasks from across all tracks. The candidate systems are authored by 45 authors, among whom 24 are students, either at the master's or PhD level. The award recipient(s) will be announced during the final EVALITA workshop, during the "Awards Session" plenary session.

5. Final Remarks

The outcomes of the EVALITA 2026 campaign provide a timely snapshot of how the Italian NLP community is currently positioning itself with respect to the rapid evolution of language technologies. While Large Language Models continue to play a central role in contemporary research, an important finding of this edition is that decoder-only architectures are not the dominant modeling choice among participating systems. Instead, encoder-only models remain the most widely adopted approach, and a substantial number of teams still rely on encoder–decoder, non–LLM-based and symbolic pipelines. This distribution clearly indicates that, rather than simply applying large generative models in inference-only settings, participants favored task-oriented modeling strategies tailored to the specific linguistic and computational challenges posed by each task. This trend highlights an important shift with respect to recent editions: the proposed tasks appear, on average, not to be directly solvable by off-the-shelf large language models without additional modeling choices. As a consequence, many submissions combine neural architectures with task-specific components and linguistically motivated features, and several systems still follow more traditional machine learning pipelines. In this sense, EVALITA 2026 confirms the continued relevance of hybrid and feature-based approaches, especially when dealing with fine-grained phenomena such as fallacy detection, implicit content classification, gender stereotypes, and term variation.

Another emerging pattern concerns the increasing use of data augmentation strategies, often based on the generation of synthetic training material. A non-negligible number of systems exploit automatically generated data to compensate for limited annotated resources and to improve robustness and generalization. This practice reflects a growing interest in exploiting generative models as data producers rather than solely as end-to-end predictors, and represents a promising direction for future editions of the campaign.

From a thematic perspective, participation across tracks remains heterogeneous. The Computational Ethics track stands out as the most attended, with MultiPRIDE and GSI:detect attracting the largest number of teams and runs. This result confirms the strong interest of the community in socially relevant NLP applications, such as stereotype recognition and the analysis of reclaimed slurs in online communication. A similarly high level of engagement is observed for tasks dealing with machine-generated text and fallacies in social media, which directly address pressing challenges related to content authenticity and online discourse. In contrast, tasks requiring deeper modeling of implicit meaning or multimodal reasoning show lower participation, likely reflecting both the higher conceptual complexity of the phenomena and the higher technical barriers associated with these settings. EVALITA 2026 also strengthens the role of the campaign as a testbed for novel and emerging research directions. The introduction of tasks such as Italian crossword solving and selective verification of unlearning in LLMs demonstrates a growing interest in moving beyond traditional classification benchmarks, towards evaluation scenarios that explicitly target reasoning, memorization, and controllability of language models.

With respect to the international dimension, the 2026 edition confirms the growing involvement of research groups from outside Italy, with contributions from thirteen different countries. Although the majority of teams still include at least one Italian affiliation, the geographical distribution of participants and the strong international participation in some tasks indicate a steadily increasing visibility of the EVALITA initiative within the broader NLP community.

Overall, the results of EVALITA 2026 underline a mature and diversified research landscape, in which LLMs coexist with more traditional approaches, and where linguistic insight, task-specific modeling and evaluation-driven design continue to play a crucial role. The campaign confirms the importance

of shared evaluation initiatives as a means to foster methodological rigor, stimulate the exploration of emerging challenges, and support the development of robust and socially responsible language technologies for Italian and beyond.

Acknowledgments

We would like to thank our sponsors: AI2B[5], Aptus.ai[6] and Prometeia[7]. Our gratitude goes also to the University of Bari Aldo Moro for hosting the event. In addition, we sincerely thank the Best System award committee for providing their expertise and experience. Moreover, we acknowledge the AILC Board members for their trust and support. We warmly thank our invited speaker, Paolo Rosso, for sharing his knowledge and insights with his talk. Last but not least, we would like to thank all the task organizers and participants who made this edition special with their enthusiasm and creativity.

Declaration on Generative AI

During the preparation of this work, the author used GPT-5 and Grammarly to conduct grammar and spelling checking. The author reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] G. Puccetti, A. Pedrotti, A. Esuli, DeSegMa-IT at EVALITA 2026: Overview of the "Detection and Segmentation of Machine generated texts in Italian" Task, 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] 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), Bari, Italy, 2026.

[3] C. Ferrando, L. Draetta, M. Madeddu, M. Sosto, V. Patti, P. Rosso, C. Bosco, J. Mata, E. Gualda, MultiPRIDE at EVALITA 2026: Overview of the Multilingual Automatic Detection of Slur Reclamation in the LGBTQ+ Context Task, in: Proceedings of the Ninth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2026), Bari, Italy, 2026.

[4] C. Ciaccio, G. Sarti, A. Miaschi, F. Dell'Orletta, M. Nissim, Cruciverb-IT @ EVALITA 2026: Overview of the Crossword Solving in Italian Task, 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.

[5] C. Savelli, M. La Quatra, A. Koudounas, F. Giobergia, SVELA at EVALITA 2026: Overview of the Selective Verification of Erasure from LLM Answers Task, in: Proceedings of the Ninth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2026), Bari, Italy, 2026.

[6] E. Musacchio, L. Siciliani, P. Basile, G. Semeraro, EVWSD-ITA at EVALITA 2026: Overview of the Enhanced Visual Word Sense Disambiguation for Italian Task, in: Proceedings of the Ninth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2026), Bari, Italy, 2026.

[7] N. Cirillo, G. M. Di Nunzio, F. Vezzani, ATE-IT at EVALITA 2026: Overview of the Automatic Term Extraction Italian Testbed Task, in: Proceedings of the Ninth Evaluation Campaign of Natural

[5] https://www.ai2b.it/

[6] https://aptus.ai/it/

[7] https://www.prometeia.com/it

Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2026), Bari, Italy, 2026.

[8] A. P. Bardelli, T. Çekiç, I. Demirtaş, M. Filannino, S. Scala, A. Galassi, G. Pappacoda, P. Torroni, PFB at EVALITA 2026: Overview of the Prometeia Financial Benchmark, in: Proceedings of the Ninth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2026), Bari, Italy, 2026.

[9] A. Ramponi, S. Tonelli, FadeIT at EVALITA 2026: Overview of the Fallacy Detection in Italian Social Media Texts Task, in: Proceedings of the Ninth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2026), Bari, Italy, 2026.

[10] L. Gregori, W. Paci, V. Saccone, IMPOLS at EVALITA 2026: Overview of the IMPOLS task, in: Proceedings of the Ninth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2026), Bari, Italy, 2026.

[11] M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, Evalita 2023: Overview of the 8th evaluation campaign of natural language processing and speech tools for italian, 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. 3.