NL4AI 2026: Overview of the Ninth Workshop on Natural Language for Artificial Intelligence (NL4AI 2026)

Alessandro Bondielli[1], Giovanni Bonetta[2], Elisa Leonardelli[2] and Irene Siragusa[3]

[1]University of Pisa, Pisa, Italy
[2]Fondazione Bruno Kessler, Trento, Italy
[3]University of Palermo, Palermo, Italy

Abstract

The Ninth Workshop on Natural Language for Artificial Intelligence (NL4AI 2026) explores the interaction between Natural Language Processing and Artificial Intelligence, bringing together methodological research and applications. Held in Perugia, Italy, alongside AIxIA 2026, the workshop received 26 submissions from 80 distinct authors across nine countries. 3 papers have been withdrawn, while 14 papers were accepted, with an acceptance rate of 60.9%. The accepted contributions address Large Language Models (LLMs) evaluation, reasoning and interpretability, data construction, clinical applications, multilingual and low-resource processing, societal bias, annotation perspectives, text watermarking, and multi-agent workflows. This overview summarizes all accepted papers by research theme, highlighting their principal contributions and the scope of their reported findings. The collection illustrates the importance of reliable evaluation and task-specific evidence across written language, speech, and multimodal content.

The Natural Language for Artificial Intelligence (NL4AI) workshop brings together research on language technologies and their role in intelligent systems. Its scope includes the representations, resources, evaluation methods, and applications through which natural language contributes to learning, reasoning, and interaction. Supported by the Special Interest Group on Natural Language Processing (NLP) of the Italian Association for Artificial Intelligence (AIxIA) and the Italian Association of Computational Linguistics (AILC), NL4AI provides a forum for connecting these research perspectives. The ninth edition takes place in Perugia, Italy, on October 7–8, 2026, co-located with the 24th International Conference of the Italian Association for Artificial Intelligence (AIxIA 2026).

This year's edition attracted 26 submissions by 80 distinct authors from nine countries: Italy (42), France (12), India (7), Poland (6), Indonesia (5), the United States (4), Brazil (2), Turkey (1), and the United Kingdom (1). Three paper have been withdrawn, while after the peer-review process, 14 of the 23 submitted papers were accepted for inclusion in these proceedings (acceptance rate: 60.9%). The accepted papers involve 63 distinct authors and cover a broad range of topics related to the workshop. The selection process used single-blind peer review, with at least two reviews required for each paper undergoing review. The contributions address model evaluation and interpretability, data construction and domain-specific applications, multilingual and low-resource NLP, bias and annotation perspectives, text provenance, and multi-agent workflows. They span written language, clinical speech, and image–text content. The following overview groups the papers by their main research themes.

Several contributions examine how language models should be evaluated and what their outputs reveal about their capabilities. Ricciardi et al. [1] study the complementarity of LLM-based evaluation metrics for Retrieval-Augmented Generation across six languages. Their diagnostic protocol contrasts correctly paired inputs with shuffled variants that disrupt their semantic correspondence. Although Relevancy, Groundedness, and Correctness respond to this manipulation, Groundedness and Correctness provide strongly overlapping signals, whereas Relevancy is largely independent. The study therefore distinguishes a metric's sensitivity to meaning from its contribution to a broader evaluation framework. Godziszewski et al. [2] approach evaluation through social-choice methods: an LLM ranks multiple system outputs, and rankings obtained under different permutations and random seeds are aggregated

NL4AI 2026: Ninth Workshop on Natural Language for Artificial Intelligence, October 7–8, 2026, Perugia, Italy
alessandro.bondielli@unipi.it (A. Bondielli); gbonetta@fbk.eu (G. Bonetta); eleonardelli@fbk.eu (E. Leonardelli); irene.siragusa02@unipa.it (I. Siragusa)
© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

using voting rules. Their experiments suggest that this approach can improve stability and alignment with human judgments while reducing the effects of presentation order.

The relationship between observable behavior and internal representations is explored from two complementary perspectives. Auriemma et al. [3] analyze chain-of-thought traces in the Qwen3.5 model family on commonsense causal reasoning. They distinguish trace-level validity from local reasoning strategies and find that the relationship between valid reasoning and correct answers depends on model scale and the plausibility of the scenario. Larger thinking budgets mainly support completion of an existing strategy pattern rather than producing qualitatively different reasoning. Gjinika et al. [4] investigate sentence complexity through sparse autoencoder activations in two small neural language models. Across three datasets, their Number of Effective Features measure increases with complexity in middle-to-late layers, while the relationship reverses in early layers. These results connect linguistic variation to the structure of sparse representations without assuming that the same relationship holds throughout a model.

Other papers investigate how data selection and task adaptation affect performance in practical settings. Alfarano et al. [5] compare strategies for constructing pseudo-labeled news datasets under a limited LLM annotation budget. Their retrieval-and-verification pipeline selects relevant candidates before querying an LLM. On the two news benchmarks considered, classifiers trained on pseudo-labels outperform direct LLM inference on Macro F1; the retrieval-based strategy is particularly effective on metrics sensitive to minority categories. Carefully specified category descriptions also make unsupervised alternatives competitive. Pucci et al. [6] compare model scale, specialization, and architecture for Italian fact verification on FEVER-IT. Adapting Llama-3.1-8B-Instruct with LoRA yields the strongest observed result, while smaller encoder models offer a different accuracy–parameter-count trade-off. The authors emphasize that the adapted model uses labeled training data unavailable to the few-shot baselines, so the result supports task-specific model selection rather than an intrinsic superiority claim based on model size.

Clinical applications illustrate the importance of matching computational representations to the information present in the source material. Siragusa and Pirrone [7] address alignment between free-text disease descriptions in MedPix 2.0 and the ICD-9, ICD-10, and ICD-11 coding systems. Their pipeline combines exact matching, semantic retrieval, and BERT-F1 reranking. Definition-based queries provide the best overall retrieval performance, and the study finds closer lexical alignment between clinical descriptions and ICD-11 than with earlier revisions. Lawand et al. [8] focus on information that ordinary transcripts can lose, analyzing silence and filled pauses in 39 French psychiatric interviews from APAISE+. Their manual annotation protocol and computational pipeline reveal differences in silence duration across clinical groups, but not in filled-pause rate. The exploratory findings also show why longer silences cannot be interpreted as a straightforward marker of clinical status and motivate preserving prosodic information in clinical NLP.

Multilingual and low-resource settings raise related questions about access to linguistic and factual knowledge. Colace et al. [9] compare five LLMs for translation from Old Occitan into Italian, combining COMET-based evaluation with qualitative analysis. GPT-5 performs best among the tested models, but errors persist, and few-shot prompting brings only moderate improvements. Their analysis also identifies limitations in applying automatic translation evaluation to this historical language pair. Pamungkas et al. [10] study cross-lingual few-shot transfer for Indonesian misogynistic meme detection. Across demonstrations drawn from six source languages, static example selection is the most consistent strategy, while textual and multimodal retrieval are preferable in particular settings. Their results support transfer without translating the target memes, while showing that no selection strategy is uniformly best.

Responsible language technology also requires attention to what models predict, whose perspective annotations represent, and how generated text can be identified. Manno and Miliani [11] introduce Beliefs, an English–Italian dataset of minimally different, gender-counterfactual biographies for occupation prediction. Their analysis exposes gender-dependent occupational predictions that aggregate accuracy can obscure, including cases in which a change to a female referent leads to a less prestigious predicted occupation. D'Avenia et al. [12] distinguish Speaker, Perceived Speaker, and Observer annotation roles

in emotion recognition. In a prompting study across five models, matching the prompt's role to the dataset's annotation perspective produces no consistent performance benefit, indicating that perspective alignment requires more than a change in prompt framing. Opreni et al. [13] investigate the resilience of three text-watermarking approaches across five Italian-centric LLMs under back-translation attacks. Detectability deteriorates across the tested approaches, including those based on semantic representations, highlighting the need to assess provenance mechanisms under cross-lingual transformations.

Finally, Ismayilov [14] presents MAWGE, a library for generating and simulating synthetic multi-agent workflows. Its pipeline combines a generator, serializer, event-driven simulator, and visualizer, supporting six reasoning-pattern topologies and agent configurations drawn from a collection of 75 LLMs. By modeling execution characteristics such as latency, throughput, and cost, the library provides a controlled setting for studying workflow structure and scheduling. Validation against realistic workflow traces remains a direction for future work, distinguishing its synthetic benchmarking role from a direct evaluation of deployed multi-agent systems.

Taken together, the papers emphasize the value of explicit evaluation assumptions, carefully constructed resources, and analyses that go beyond a single aggregate score. Across their different tasks, they distinguish model scale from task suitability, knowledge availability from retrievability, and observable output from the processes or perspectives it is intended to represent. This combination of methodological analysis and concrete applications characterizes the contribution of NL4AI 2026 to the dialogue between Natural Language Processing and Artificial Intelligence.

References

[1] Salvatore Ricciardi, Simone Scaboro, Beatrice Turano, Cristina Giannone, Andrea Favalli, and Raniero Romagnoli. Who Judges the Judges? A Multilingual Probe for Complementary LLM Evaluation Metrics. In: *Proceedings of the Ninth Workshop on Natural Language for Artificial Intelligence (NL4AI 2026)*, 2026. This volume.

[2] Michał Tomasz Godziszewski, Mateusz Galiński, Rzyszard Tuora, Mateusz Czyżnikiewicz, Tomasz Ziętkiewicz, Adam Kozakiewicz, and Cristina Cornelio. Election-Based Evaluation Methods for NLG-Systems. In: *Proceedings of the Ninth Workshop on Natural Language for Artificial Intelligence (NL4AI 2026)*, 2026. This volume.

[3] Serena Auriemma, Martina Miliani, Alessandro Bondielli, Lucia Passaro, and Alessandro Lenci. Right for the Right Reasons? A Chain-of-Thought Analysis of Commonsense Causal Reasoning. In: *Proceedings of the Ninth Workshop on Natural Language for Artificial Intelligence (NL4AI 2026)*, 2026. This volume.

[4] Ejdis Gjinika, Luca Putelli, Ivan Serina, and Alfonso Emilio Gerevini. Exploring the Relationship between Sparse Autoencoder Activations and Sentence Complexity. In: *Proceedings of the Ninth Workshop on Natural Language for Artificial Intelligence (NL4AI 2026)*, 2026. This volume.

[5] Giulio Alfarano, Stefano Scotta, and Alberto Messina. Comparing Generated Pseudo-Labeled News Datasets: A Retrieval-and-Verification Approach. In: *Proceedings of the Ninth Workshop on Natural Language for Artificial Intelligence (NL4AI 2026)*, 2026. This volume.

[6] Alessandro Pucci, Danilo Croce, and Roberto Basili. Scale or Specialization? Evaluating Language Models for Italian Fact Verification. In: *Proceedings of the Ninth Workshop on Natural Language for Artificial Intelligence (NL4AI 2026)*, 2026. This volume.

[7] Irene Siragusa and Roberto Pirrone. Evaluating Clinical Coding Alignment Across ICD-9, ICD-10, and ICD-11 with a Semantic Retrieval Approach. In: *Proceedings of the Ninth Workshop on Natural Language for Artificial Intelligence (NL4AI 2026)*, 2026. This volume.

[8] Claire Lawand, Haeeul Hwang, Jalal Al-Tamimi, Motasem Alrahabi, Mahshid Hozhabr, Deok-Hee Kim-Dufor, Christophe Lemey, Julien Desclés, and Valeria Lucarini. The Prosody of Clinical Speech: Annotation and Computational Analysis of Psychiatric Interviews. In: *Proceedings of the Ninth Workshop on Natural Language for Artificial Intelligence (NL4AI 2026)*, 2026. This volume.

[9] Francesco Colace, Sabrina Galano, Angelo Lorusso, Domenico Santaniello, Giusy Strollo, and Annapia Tanzola. Machine Translation of Low-Resource Historical Languages: A Comparative Study of LLMs on Old Occitan. In: *Proceedings of the Ninth Workshop on Natural Language for Artificial Intelligence (NL4AI 2026)*, 2026. This volume.

[10] Endang Wahyu Pamungkas, Maryam, Rona Rizkhy Bunga Chasana, Dade Nurjanah, and Ikhlasul Amal. Static or Dynamic? Exploring Cross-Lingual Few-Shot Transfer for Indonesian Misogynistic Meme Detection. In: *Proceedings of the Ninth Workshop on Natural Language for Artificial Intelligence (NL4AI 2026)*, 2026. This volume.

[11] Chiara Manno and Martina Miliani. Same Job, Different Gender: Uncovering LLM Stereotypes via Counterfactual Biographies. In: *Proceedings of the Ninth Workshop on Natural Language for Artificial Intelligence (NL4AI 2026)*, 2026. This volume.

[12] Samuele D'Avenia, Timothy Meinert, Marta Marchiori Manerba, Valerio Basile, and Anna Koufakou. Does Perspective Matter? Annotation Role Alignment in Emotion Recognition. In: *Proceedings of the Ninth Workshop on Natural Language for Artificial Intelligence (NL4AI 2026)*, 2026. This volume.

[13] Elena Opreni, Alice Pietrasanta, Davide Altieri, Giulia Rizzi, and Elisabetta Fersini. Evaluating Watermark Resilience in Italian Large Language Models Under Back-Translation Attacks. In: *Proceedings of the Ninth Workshop on Natural Language for Artificial Intelligence (NL4AI 2026)*, 2026. This volume.

[14] Goshgar C. Ismayilov. MAWGE: A Parametric and Synthetic Multi-Agentic Workflow Generator Library. In: *Proceedings of the Ninth Workshop on Natural Language for Artificial Intelligence (NL4AI 2026)*, 2026. This volume.