Gabriele Sarti

Postdoctoral Researcher, Northeastern University
Khoury College of Computer Sciences
02115 Boston, MA, USA
🏠 https://gsarti.com
✉ gabriele.sarti996@gmail.com

Current Position

Northeastern University Boston, MA, USA
Postdoctoral Researcher *Jan. 2026 – exp. Dec. 2026*
Advisor: David Bau. R&D for the NSF National Deep Inference Foundation (NDIF) project.

Education

University of Groningen (RUG) Groningen, Netherlands
Ph.D. in Natural Language Processing, Cum Laude *Sept. 2021 – Dec. 2025*
Advisors: Arianna Bisazza, Malvina Nissim, Grzegorz Chrupała
Thesis: "From Insight to Impact: Actionable Interpretability for Neural Machine Translation"

University of Trieste & SISSA Trieste, Italy
M.Sc. in Data Science and Scientific Computing, 110 cum laude *Oct. 2018 – Dec. 2020*
Advisors: Felice Dell'Orletta, Davide Crepaldi
Thesis: "Interpreting Neural Language Models for Linguistic Complexity Assessment"

Cégep de Saint-Hyacinthe Saint-Hyacinthe, QC, Canada
Collegial Studies Degree (DEC) in Management Informatics *Aug. 2015 – May 2018*
Valedictorian of the 2018 informatics cohort. Program taught in French.

Experience

Industrial Experience

Amazon Web Services AI Lab New York, NY, USA
Applied Scientist Intern (hosted by Georgiana Dinu, Maria Nadejde) *Jun. 2022 – Sept. 2022*
RAG-augmented LLM prompting for attribute-controlled translation.

Aindo Trieste, Italy
Research Scientist, Generative AI Systems *Nov. 2020 – Aug. 2021*
Structured clinical QA, recommender systems and sketch-to-image generation.

Skytech Communications Montréal, QC, Canada
Machine Learning Engineer Intern *Feb. 2018 – Jun. 2018*
Handwritten OCR detection and sentiment analysis for educational applications.

Academic Visiting

IRT Saint-Exupéry Toulouse, France
Visiting Researcher (hosted by Fanny Jourdan) *Feb. 2025 – Mar. 2025*
Collaboration on an open-source toolkit for concept-based interpretability of LLMs.

Institute of Computational Linguistics (ILC-CNR) Pisa, Italy
Visiting Research Assistant (hosted by Felice Dell'Orletta) *Sept. 2019 – Dec. 2019*
Multitask language model fine-tuning for gaze metrics prediction and complexity assessment.

Teaching and Advising

Classes

Advanced Topics in Natural Language Processing, RUG, MSc Inf. Sci. (Co-instructor), 2025

Natural Language Processing, RUG, MSc Inf. Sci. (Co-instructor), 2022-2024

Fundamentals of Machine Learning, RUG, Data Science Minor, 2024

Ph.D. Students

Sara Candussio, ADSAI Ph.D. at University of Trieste (co-sup. with Luca Bortolussi), 2024-

MSc. Students

Khondoker I. Islam, Project in Natural Language Processing, RUG, 2025
Samuele d'Avenia, Thesis in Data Science, University of Trieste, 2024
Sara Candussio, Thesis in Data Science, University of Trieste, 2024
Konstantin Chernyshev, Thesis in NLP, University of Saarland & RUG, 2024
Daniel Scalena, Project in Computer Science, University of Milano-Bicocca, 2023
Ludwig Sickert, Thesis in Artificial Intelligence, RUG, 2023
Qiankun Zheng, Project in Natural Language Processing, RUG, 2023

Awards

Best Paper, Italian Conference on Computational Linguistics (CLiC-it 2025)

Best Paper, Italian Conference on Computational Linguistics (CLiC-it 2023)

Best Paper runner-up, Cognitive Modeling and Computational Linguistics WS (CMCL 2021)

Best Master's Thesis Award, Italian Association for Computational Linguistics (AILC). 2022.

Best Paper runner-up, Italian Conference on Computational Linguistics (CLiC-it 2020)

Winner of the AcCompl-it Shared Task, Evaluation of NLP Tools for Italian (EVALITA 2020)

Scholarships and Grants

Grant for Project Mentorship (7k$), Supervised Program for Alignment Research (SPAR), 2026

AInet Fellowship (visiting expenses), Deutscher Akademischer Austauschdienst (DAAD), 2025

Language Technology Grant (20k€) , Imminent Research Center, 2023

eScience Fellowship (2k€), Amsterdam Science Center, 2023

Data Science Excellence Fellowship (12k€), International School of Advanced Studies (SISSA), 2018

Informatics Merit Scholarship (1k€), Hydro-Québec, 2016

Publications

Journal Articles

[J4] Gabriele Sarti, Vilém Zouhar, Grzegorz Chrupała, Ana Guerberof-Arenas, Malvina Nissim, Arianna Bisazza. "QE4PE: Word-level Quality Estimation for Human Post-Editing". In *Transactions of the Association of Computational Linguistics (TACL)*. Issue 13: 1410–1435, 2025.

[J3] Lukas Edman, Gabriele Sarti, Antonio Toral, Gertjan van Noord, Arianna Bisazza. "Are Character-level Translations Worth the Wait? Comparing ByT5 and mT5 for Machine Translation". In *Transactions of the Association of Computational Linguistics (TACL)*. Issue 12: 392–410, 2024.

[J2] Alessio Miaschi, Gabriele Sarti, Dominique Brunato, Felice Dell'Orletta, Giulia Venturi. "Probing Linguistic Knowledge in Italian Neural Language Models across Language Varieties". In *Italian Journal of Computational Linguistics (IJCoL)*. Issue 8-1, 2022.

[J1] Ginevra Carbone, Gabriele Sarti. "ETC-NLG: End-to-end Topic-Conditioned Natural Language Generation". In *Italian Journal of Computational Linguistics (IJCoL)*. Issue 6-2, 2020.

Conference Proceedings

[C11] Daniel Scalena†, Gabriele Sarti†, Arianna Bisazza, Elisabetta Fersini, Malvina Nissim. "Steering Large Language Models for Machine Translation Personalization". In *Conference of the European Chapter of the Association for Computational Linguistics (EACL)*. 2026. († = Equal contribution)

[C10] Gabriele Sarti, Vilém Zouhar, Malvina Nissim, Arianna Bisazza. "Unsupervised Word-level Quality Estimation for Machine Translation Through the Lens of Annotators (Dis)agreement". In *Conference on Empirical Methods for Natural Language Processing (EMNLP)*. 2025. Top 15%.

[C9] Sara Candussio, Gaia Saveri, Gabriele Sarti, Luca Bortolussi. "Bridging Logic and Learning: Decoding Temporal Logic Embeddings via Transformers". In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD)*. 2025.

[C8] Mohammad Reza Ghasemi Madani, Aryo Pradipta Gema, Yu Zhao, Gabriele Sarti, Pasquale Minervini, Andrea Passerini. "Noiser: Bounded Input Perturbations for Attributing Large Language Models". In *Conference on Language Modeling (COLM)*. 2025.

[C7] Jirui Qi†, Gabriele Sarti†, Raquel Fernández, Arianna Bisazza. "Model Internals-based Answer Attribution for Trustworthy Retrieval-Augmented Generation". In *Conference on Empirical Methods for Natural Language Processing (EMNLP)*. 2024. († = Equal contribution)

[C6] Gabriele Sarti, Grzegorz Chrupała, Malvina Nissim, Arianna Bisazza. "Quantifying the Plausibility of Context Reliance in Neural Machine Translation". In *International Conference on Learning Representations (ICLR)*. 2024.

[C5] Anna Langedijk, Hosein Mohebbi, Gabriele Sarti, Willem Zuidema, Jaap Jumelet. "DecoderLens: Layerwise Interpretation of Encoder-Decoder Transformers". In *Findings of the Association for Computational Linguistics (NAACL)*. 2024.

[C4] Gabriele Sarti, Malvina Nissim. "IT5: Text-to-text Pretraining for Italian Language Understanding and Generation". In *Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*. 2024.

[C3] Gabriele Sarti, Nils Feldhus, Ludwig Sickert, Oskar van der Wal. "Inseq: An Interpretability Toolkit for Sequence Generation Models". In *Annual Meeting of the Association for Computational Linguistics: System Demonstrations (ACL Demo)*. 2023.

[C2] Gabriele Sarti, Phu Mon Htut, Xing Niu, Benjamin Hsu, Anna Currey, Georgiana Dinu, Maria Nadejde. "RAMP: Retrieval and Attribute-Marking Enhanced Prompting for Attribute-Controlled

Translation". In *Annual Meeting of the Association for Computational Linguistics (ACL)*. 2023.

[C1] Gabriele Sarti, Arianna Bisazza, Ana Guerberof-Arenas, Antonio Toral. "DivEMT: Neural Machine Translation Post-Editing Effort Across Typologically Diverse Languages". In *Conference on Empirical Methods for Natural Language Processing (EMNLP)*. 2022.

Workshop Proceedings

[W10] Dana Arad, Yonatan Belinkov, Hanjie Chen, Najoung Kim, Hosein Mohebbi, Aaron Mueller, Gabriele Sarti, Martin Tutek. "Findings of the BlackboxNLP 2025 Shared Task: Localizing Circuits and Causal Variables in Language Models". In *Proceedings of the 8th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*. 2025.

[W9] Cristiano Ciaccio, Gabriele Sarti, Alessio Miaschi, Felice Dell'Orletta. "Crossword Space: Latent Manifold Learning for Italian Crosswords and Beyond". In *Proceedings of the 11th Italian Conference on Computational Linguistics (CLiC-it)*. 2025. Best paper award.

[W8] Gabriele Sarti, Tommaso Caselli, Malvina Nissim, Arianna Bisazza. "Non Verbis, Sed Rebus: Large Language Models Are Weak Solvers of Italian Rebuses". In *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it)*. 2024.

[W7] Daniel Scalena, Gabriele Sarti, Malvina Nissim. "Multi-property Steering of Large Language Models with Dynamic Activation Composition". In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*. 2024.

[W6] Gabriele Sarti, Nils Feldhus, Jirui Qi, Malvina Nissim, Arianna Bisazza. "Democratizing Advanced Attribution Analyses of Generative Language Models with the Inseq Toolkit". In *Proceedings of the 2nd World Conference on Explainable AI (xAI): System Demonstrations*. 2024.

[W5] Federico Bianchi, Giuseppe Attanasio, Raphael Pisoni, Silvia Terragni, Gabriele Sarti, Dario Balestri. "Contrastive Language-Image Pre-training for the Italian Language". In *Proceedings of the 9th Italian Conference on Computational Linguistics (CLiC-it)*. 2023. Best paper award.

[W4] Gabriele Sarti, Dominique Brunato, Felice Dell'Orletta. "That Looks Hard: Characterizing Linguistic Complexity in Humans and Language Models". In *Workshop on Cognitive Modeling and Computational Linguistics (CMCL)*. 2021. Best paper award runner-up.

[W3] Ludovica Pannitto, Lucia Busso, Claudia Roberta Combei, Lucio Messina, Alessio Miaschi, Gabriele Sarti, Malvina Nissim. "Teaching NLP with Bracelets and Restaurant Menus: An Interactive Workshop for Italian Students". In *Proceedings of the Fifth Workshop on Teaching NLP*. 2021.

[W2] Alessio Miaschi, Gabriele Sarti, Dominique Brunato, Felice Dell'Orletta, Giulia Venturi. "Italian Transformers Under the Linguistic Lens". In *Proceedings of the 7th Italian Conference on Computational Linguistics (CLiC-it)*. 2020. Best paper award runner-up.

[W1] Gabriele Sarti. "Improving Complexity and Acceptability Prediction with Multi-task Learning on Self-Supervised Annotations". In *AcCompl-it Shared Task at EVALITA*. 2020. Best system award.

Preprints

[P4] Antonin Poché†, Thomas Mullor, Gabriele Sarti et al., Fanny Jourdan†. "Interpreto: An Explainability Library for Transformers". In *CoRR/2512.09730*. 2025. († = Equal contribution)

[P3] Malvina Nissim†, Danilo Croce†, Viviana Patti†, Pierpaolo Basile† et al. (incl. Gabriele Sarti). "Challenging the Abilities of Large Language Models in Italian: a Community Initiative". In *CoRR/2512.04759*. 2025. († = Equal contribution)

[P2] Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Summer Yue, Alexandr Wang, Dan Hendrycks et al. (incl. Gabriele Sarti). "Humanity's Last Exam". In *CoRR/2501.14249*. 2025.

[P1] Javier Ferrando, Gabriele Sarti, Arianna Bisazza, Marta R. Costa-jussà. "A Primer on the Inner Workings of Transformer-based Language Models". In *CoRR/2405.00208*. 2024.

Manuscripts

[M2] Gabriele Sarti. "From Insights to Impact: Actionable Interpretability for Neural Machine Translation". *Ph.D. Thesis in Natural Language Processing, University of Groningen.* 2025. Cum Laude.

[M1] Gabriele Sarti. "Interpreting Neural Language Models for Linguistic Complexity Assessment". *MSc. Thesis in Data Science, University of Trieste.* 2020. AILC Best master's thesis award.

Edited Works

[E1] Yonatan Belinkov, Aaron Mueller, Najoung Kim, Hosein Mohebbi, Hanjie Chen, Dana Arad, Gabriele Sarti. "Proceedings of the 8th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP". 2025.

Professional Service

Organizing Committee

EVALITA Shared Task on Automatic Italian Crossword Solving (Cruciverb-IT), 2026
EMNLP Workshop on Analyzing and Interpreting Neural Networks for NLP (BlackboxNLP), 2025-

Area Chair

Annual Meeting of the Association for Computational Linguistics (ACL, via ARR), 2025-
Empirical Methods for Natural Language Processing (EMNLP, via ARR), 2025-
Italian Conference on Computational Linguistics (CLiC-it), 2025-

Peer Reviewer

Transactions on Machine Learning Research (TMLR), 2025-
International Conference on Learning Representations (ICLR), 2025-2026
EMNLP Workshop on Analyzing and Interpreting Neural Networks for NLP (BlackboxNLP), 2023-2025
Annual Conference of the European Chapter of ACL (EACL, via ARR), 2025
Annual Conference of the Asia-Pacific Chapter of ACL (AACL, via ARR), 2025
ICML/NeurIPS Workshop on Mechanistic Interpretability, 2024
Conference on Language Modeling (COLM), 2024
Annual Conference of the Nations of the Americas Chapter of ACL (NAACL, via ARR), 2024
Empirical Methods for Natural Language Processing (EMNLP, via ARR), 2022-2024
Italian Conference on Computational Linguistics (CLiC-it), 2021, 2024
Annual Meeting of the Association for Computational Linguistics (ACL, via ARR), 2023
International Conference on Language Resources and Evaluation (LREC), 2022

Outstanding Reviewer Mention

Empirical Methods for Natural Language Processing (EMNLP) 2024

Dissemination Activities

Invited Lectures

University of Trieste, *Interpretability for Language Models: Current Trends and Applications*, MSc. Data Science and AI Course on Explainable and Neurosymbolic AI (hosted by Luca Bortolussi), 2025.

University of Groningen, *Interpretability for Language Models: Current Trends and Applications*, MSc. AI Course on Trustworthy and Explainable AI (hosted by Marco Zullich), 2025.

Polytechnical University of Turin, *Interpreting Context Usage in Generative Language Models*, MSc. course on Explainable and Trustworthy AI (hosted by Eliana Pastor), 2024.

Sapienza University of Rome, *Interpretability for Language Models: Current Trends and Applications*, Ph.D. Course on Explainable AI (hosted by Simone Scardapane), 2024.

University of Pisa, *Interpretability of linguistic knowledge in neural language models*, AILC Lectures on Computational Linguistics (co-instructor with Alessio Miaschi), 2023.

Academic Seminars

University of Padua, Interpretability for Language Models: Trends and Applications. Padua, 2025.
Fondazione Bruno Kessler (FBK), Interpreting Context Usage in Language Models. Online, 2025.
University of Groningen, Interpreting Latent Features in Large Language Models. Groningen, 2025.
DFKI Saarbrücken, QE4PE: Word-level Quality Estimation for Human Post-Editing. Online, 2025.
IRT Saint Exupéry, Interpreting Context Usage in Generative Language Models. Toulouse (FR), 2025.
CIS LMU Munich, Interpreting Context Usage in Generative Language Models. Munich (DE), 2024.
Area Science Park, Quantifying Context Reliance in Neural Machine Translation. Trieste (IT), 2024.
University of Sheffield, Post-hoc Interpretability for Generative Language Models. Online, 2024.
Netherlands eScience Center, Post-hoc Interpretability for Language Models. Amsterdam, 2023.
Universitat Pompeu Fabra, REST-CL, Post-hoc Interpretability for NLG. Tarragona (ES), 2023.
University of Trieste, AI-Lab, Post-hoc Interpretability for Neural Language Models. Trieste, 2023.
University of Groningen, Post-hoc Interpretability for Neural Language Models. Groningen, 2023.
Sapienza University of Rome, An Interpretability Toolkit for Language Models. Rome (IT), 2023.
Radboud University, Advanced Interpretability for Language Models. Nijmegen (NL), 2023.
Translated S.r.l., Towards User-centric Interpretability of NLP Models. Online, 2022.
NeurIPS XAI4Debugging WS, Interpretable Interactive Neural Machine Translation. Online, 2021.
Bocconi University, MilaNLP, Linguistic Complexity in Humans and Language Models. Online, 2021.
University of Trieste, StaTalk, Neural Language Models: the New Frontier of Natural Language Understanding. Trieste, 2019.
38th Symposium of the Québec Association for Collegial Pedagogy (AQPC), The Educational Impact of Artificial Intelligence (with Alexandre Brunet). Saint-Hyacinthe (CA), 2018.

Science Communication

The Inquisitive Mind Magazine, Peer reviewer. 2025.
Trieste Next, Inside the Algorithm: Transparency and Impact of Generative AI (panel). Trieste, 2025.
Artificial Intelligence Student Society (AI2S), Aprire la scatola nera dei modelli del linguaggio: rischi e opportunità. Trieste, 2024.
Trieste Next, AI-Italo Svevo: Lettere da un'intelligenza artificiale (live demo with Cristina Fenu and Eric Medvet). Trieste, 2019.
Trieste Science+Fiction Festival, The Literary Ordnance: When the Writer is an AI (with Felice dell'Orletta and Cristina Fenu). Trieste, 2019.

Industrial Masterclasses

InDeep Masterclass, Interpreting and Understanding LLMs and Other Deep Learning Models. Presentation & hands-on tutorial. Deloitte Amsterdam, 2025.

InDeep Masterclass, Explaining Foundation Models. Presentation & hands-on tutorial. University of Amsterdam, 2023.

Press Coverage

Slator, Does Word-Level Quality Estimation Really Improve AI Translation Post-Editing? 2025.
MarkTechPost, Deciphering Language Models: Advances in Interpretability Research. 2024.
Imminent Research Center, Can Word-level Quality Estimation Inform and Improve Machine Translation Post-editing?. 2024 (guest post).
Bocconi University Hub News, L'AI che racconta le immagini in italiano. 2021.
Poliflash, CLIP-Italian: un nuovo modello di AI per collegare immagini e testi in italiano. 2021.

Mentorship and Social Engagements

Spring 2026	Mentor, Implicit Personalization project, Supervised Program for Alignment (SPAR)
Fall 2025	Co-mentor for Project Telos, Supervised Program for Alignment (SPAR)
2022 – 2025	Research and Upskilling Advisor, AI Safety Initiative Groningen (AISIG)
Winter 2021	Team Mentor, PiCampus School of AI
2020	Co-founder and President, Artificial Intelligence Student Society (AI2S)
2020 – 2021	SciComm Team Member, Italian Association for Computational Linguistics (AILC)

Academic References

Malvina Nissim, Full Professor, University of Groningen — m.nissim@rug.nl

Arianna Bisazza, Associate Professor, University of Groningen — a.bisazza@rug.nl

Felice Dell'Orletta, Senior Researcher, ILC-CNR — felice.dellorletta@ilc.cnr.it

Luca Bortolussi, Full Professor, University of Trieste — lbortolussi@units.it