

# Gabriele Sarti

PhD Student, University of Groningen  
Center for Language and Cognition (CLCG)  
9712EK Groningen, The Netherlands  
🏠 <https://gsarti.com>  
✉ [gabriele.sarti996@gmail.com](mailto:gabriele.sarti996@gmail.com)

## CURRENT POSITION

---

<b>University of Groningen (RUG)</b> Doctoral Researcher in Natural Language Processing	Groningen, the Netherlands <i>Sept. 2021 – Present</i>
--	---

## EDUCATION

---

<b>University of Groningen (RUG)</b> Ph.D. in Natural Language Processing Advisors: Arianna Bisazza, Malvina Nissim, Grzegorz Chrupala Thesis: “ <a href="#">From Insight to Impact: Actionable Interpretability for Neural Machine Translation</a> ”	Groningen, Netherlands <i>Sept. 2021 – Present</i>
--	---

<b>University of Trieste &amp; SISSA</b> M.Sc. in Data Science and Scientific Computing, 110 cum laude Advisors: Felice Dell’Orletta, Davide Crepaldi Thesis: “ <a href="#">Interpreting Neural Language Models for Linguistic Complexity Assessment</a> ”	Trieste, Italy <i>Oct. 2018 – Dec. 2020</i>
---	--

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

## EXPERIENCE

---

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

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

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

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

<b>Skytech Communications</b> Machine Learning Engineer Intern Handwritten OCR detection and sentiment analysis for educational applications.	Montréal, QC, Canada <i>Feb. 2018 – Jun. 2018</i>
---	--

<b>CTT Group</b> Research Assistant Intern	Saint-Hyacinthe, QC, Canada <i>Oct. 2015 – Aug. 2017</i>
---	---

## 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-supervised with Luca Bortolussi)

### MSc. STUDENTS

Khondoker I. Islam, LCT 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 Natural Language Processing, 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, LCT Project in Natural Language Processing, RUG, 2023

### INVITED LECTURES

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

### AWARDS

---

**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

---

**Language Technology Grant**, Imminent Research Center, 2023

**eScience Fellowship**, Amsterdam Science Center, 2023

**Data Science Excellence Fellowship**, International School of Advanced Studies (SISSA), 2018

**Informatics Perseverance Scholarship**, Hydro-Québec, 2016

**AFS Italy Exchange Year Full-ride Scholarship**, Telecom Italia Mobile, 2013

### Journal Articles

- [J4] Gabriele Sarti, Vilém Zouhar, Grzegorz Chrupała, Ana Guerberofo-Arenas, Malvina Nissim, Arianna Bisazza. “[QE4PE: Word-level Quality Estimation for Human Post-Editing](#)”. In *Transactions of the Association of Computational Linguistics (TACL)*. 2025. (To appear)
- [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

- [C13] 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% of accepted papers.** (To appear)
- [C12] 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. (To appear)
- [C11] 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.
- [C10] Jirui Qi<sup>†</sup>, Gabriele Sarti<sup>†</sup>, 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. (<sup>†</sup> = Equal contribution)
- [C9] 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.
- [C8] 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.
- [C7] 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.
- [C6] 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.
- [C5] 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.

- [C4] 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.
- [C3] 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**.
- [C2] Gabriele Sarti, Arianna Bisazza, Ana Guerbero-f-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.
- [C1] 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**.

## Workshop Proceedings

- [W5] 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.
- [W4] 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.
- [W3] 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**.
- [W2] 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.
- [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

- [P3] Daniel Scalena<sup>†</sup>, Gabriele Sarti<sup>†</sup>, Arianna Bisazza, Elisabetta Fersini, Malvina Nissim. “[Steering Large Language Models for Machine Translation Personalization](#)”. In *CoRR/2505.16612*. 2025. (<sup>†</sup> = 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

- [M1] Gabriele Sarti. “[Interpreting Neural Language Models for Linguistic Complexity Assessment](#)”. In *MSc. Thesis in Data Science, University of Trieste*. 2020. **AILC Best master’s thesis award**.

### Organizing Committee

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

### 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 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) 2025

### INVITED TALKS

---

#### ACADEMIC PRESENTATION

**University of Groningen, InCLoW**, Interpreting Latent Features in Large Language Models. Groningen (NL), 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 the Plausibility of Context Reliance in Neural Machine Translation. Trieste (IT), 2024.

**University of Sheffield**, Post-hoc Interpretability for Generative Language Models. Online, 2024.

**University of Amsterdam**, Explaining Foundation Models with Inseq. Amsterdam (NL), 2023.

**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**, Inseq: An Interpretability Toolkit for Sequence Generation 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 Workshop**, Empowering Human Translators via Interpretable Interactive Neural Machine Translation. Online, 2021.

**Bocconi University, MilaNLP**, Characterizing 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

**Artificial Intelligence Student Society (AI2S)**, Aprire la scatola nera dei modelli del linguaggio: rischi e opportunità (with Andrea Gasparin). Trieste, 2024.

**Trieste Next, AI-Italo Svevo: Lettere da un'intelligenza artificiale** (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.

PRESS COVERAGE

---

**Slator**, [Does Word-Level Quality Estimation Really Improve AI Translation Post-Editing?](#) 2025.

**MarkTechPost**, [Deciphering Transformer 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.

SOCIAL ENGAGEMENTS

---

2022 – 2025	Research and Upskilling Advisor, <a href="#">AI Safety Initiative Groningen (AISIG)</a>
2021	Team Mentor, <a href="#">PiCampus School of AI W21 Program</a>
2020	Co-founder and President, <a href="#">Artificial Intelligence Student Society (AI2S)</a>
2020 – 2021	SciComm Team Member, <a href="#">Italian Association for Computational Linguistics (AILC)</a>

ACADEMIC REFERENCES

---

**Malvina Nissim**, Full Professor, University of Groningen — [m.nissim@rug.nl](mailto:m.nissim@rug.nl)

**Arianna Bisazza**, Associate Professor, University of Groningen — [a.bisazza@rug.nl](mailto:a.bisazza@rug.nl)

**Felice Dell'Orletta**, Senior Researcher, ILC-CNR — [felice.dellorletta@ilc.cnr.it](mailto:felice.dellorletta@ilc.cnr.it)

**Luca Bortolussi**, Full Professor, University of Trieste — [lbortolussi@units.it](mailto:lbortolussi@units.it)