Curriculum vitae

PERSONAL INFORMATION

Giovanni Puccetti

📍 Lungarno Mediceo 10, Pisa (56127)

☎ +39 351 9443040

✉ giovanni.puccetti@isti.cnr.it / g.puccetti92@gmail.com

🌐 https://gpucce.github.io

⬤ https://github.com/gpucce

Gender Male | Date of birth 15/08/1992 | Nationality Italian

SUMMARY

I am a postdoctoral researcher at the Institute for Science and Technologies of Information "Alessandro Faedo" of the Italian National Council of Research (ISTI - CNR), in Pisa. I hold a B.Sc. and M.Sc. in Mathematics, followed by a PhD in Data Science. My current research focuses on the detection of texts generated by Large Language Models (LLMs), watermarking and fingerprinting. I also have ongoing research activities in mechanistic interpretability of LLMs and in the evaluation of LLMs in mathematical understanding.

PROFESSIONAL EXPERIENCE

16/09/2024 - Currently	Research fellowship ("Ricercatore III Livello a Tempo Determinato") at the Institute of Information Science and Technologies "Alessandro Faedo" of the Italian National Council of Research (ISTI-CNR). Atto della Presidente del CNR N. 0136502 8 Maggio 2023
01/01/2024 - 15/09/2024	Research fellowship ("Assegno di ricerca postdottorale") at the Institute of Information Science and Technologies "Alessandro Faedo" of the Italian National Council of Research (ISTI-CNR). Prot. ISTI 034/2022 - PI
01/01/2023 - 31/12/2023	Research fellowship ("Assegno di ricerca professionalizzante") at the Institute of Information Science and Technology "Alessandro Faedo" (ISTI-CNR). Prot. ISTI 034/2022 - PI
06/03/2019 - 5/06/2019	Research fellowship at the University of Pisa. Prot. 668 del 12.02.19 Rep.59.
	The focus of this research fellowship, which I have carried out under the supervision of Prof. Gualtiero Fantoni, has been the extraction of technical information from patents, with a focus on the use of pronouns to identify users of the invention described in the patent.
01/02/2018 - 01/04/2018	Teaching assistantship for the course "Experimental Design and Data Analysis", Vrije University, Amsterdam, NL. (Paid) Contract N. HRM201800034
	Tutoring a class of approximately 30 students, supervising exercise classes on statistics, machine learning algorithms, and their applications using the R language + Grading weekly assignments.
11/09/2018 - 31/10/2018	Teaching assistantship for the course "Statistics Simulation and Optimization", University of Amsterdam, Amsterdam, NL. (Paid) Contract N. PZ-2017/CoS/6134a
	Tutoring a class of approximately 60 students, supervising exercise classes on statistics, machine learning algorithms, and their applications using the R language + Grading weekly assignments + Final exam.

EDUCATION

07/11/2023	PhD in Data Science (with honors) from Scuola Normale Superiore, Pisa, IT. Supervisors: Felice Dell'Orletta, Gualtiero Fantoni, Fosca Giannotti.
	Thesis title: Transformer Models: from Model Inspection to Applications in Patents
01/11/2019-07/11/2023	Attended PhD programme in Data Science at Scuola Normale Superiore, Pisa, IT, under a scholarship from Scuola Normale Superiore.

03/10/2022-13/12/2022	Period abroad: RIKEN, High Performance Artificial Intelligence Systems Research Team, Tokyo, JP. Working with Dr. Aleksandr Drozd.
	Theme: Training Transformer-based Image-Text Alignment Models
	The focus of this internship has been the pre-training (from scratch) of transformer models by learning joint representations of text and images, with the goal of making the training of these models more efficient.
01/02/2022-31/05/2022	Period abroad: University of Copenhagen, Copenhagen Center for Social Data Science (SO-DAS), Copenhagen, DK. Working with Dr. Anna Rogers.
	Theme: Improving the mechanistic understanding of transformer-based language models
	The focus of this visit has been on outliers, specific parameters present in language models, that when set to zero have a disrupting effect on model performance, unlike other parameters. To better understand this phenomenon, we have fine-tuned a large number of models on a wide variety of tasks to understand which properties of the models are more seriously affected; this work has led to publication [14].
14/11/2018	Obtained MSc Degree in Mathematics from the University of Amsterdam, NL. (Score: 7.99/10)
	Thesis Title: Average Nearest Neighbor Degree in Geometric Inhomogeneous Random Graphs date 14/11/2018.
2016-2018	Attended the MSc program in Mathematics at the University of Amsterdam, NL.
11/03/2016	Obtained BSc Degree in Mathematics from the University of Pisa, IT. (Score: 105/110)
	Thesis Title: Calcolo del Terzo Gruppo di Omotopia della Due-Sfera
	Diploma n. 303760, date 11/03/2016. University of Pisa, IT.
2011/2012-2014/2015	Attended the course in Mathematics at University of Pisa.
10/07/2011	Obtained Maturità classica. (86/100)
2006-2011	Attended High School "Liceo Classico Enea Silvio Piccolomini", Siena, IT.

COLLABORATIONS

13/11/2023 - 24/11/2023	Visiting Researcher at Mohamed bin Zayed University for Artificial Intelligence, Abu Dhabi, UAE, invited by Prof. Preslav Nakov.
	The focus of this visit has been on contributing text generated by language models to the SemEval 2024 Shared Task 8 on the Detection of Synthetic Text `https://github.com/mbzuai-nlp/SemEval2024-task8`.

INVITED TALKS

02/12/2024	I have presented the work REVERINO: Regesta generation versus Latin summarization at the workshop Digital Humanities, Intelligenza Artificale e Studi Religiosi: initial research perspectives
11/07/2024	I have presented current trends in the detection and watermarking of machine generated text at the US - EU Dialogue on Synthetic Content
17/10/2023	I have presented my PhD work at the Natural Language Processing Department of Mohammed Bin Zayed University for Artificial Intelligence, invited by Prof. Preslav Nakov

POSTERS

19/02/2024	Synthetic Text may be Coming in your Language: a Case Study in Italian. Presented at the Future Artificial Intelligence Research (FAIR) Workshop, 19-20 February 2024, Pisa, IT.
03/10/2023	MULTI-Fake-DetectIVE: Detecting fake news with visual and textual content. Presented at the Artificial Intelligence for Media (AI4Media) Workshop, 3-4 October 2023, Pisa.
07/09/2023	AIMH @ MULTI-Fake-DetectIVE. Presented at Final Workshop of the 8th Evaluation Campaign for NLP and Speech Tools for Italian (EVALITA 2023), 7-8 September 2023, Parma, IT.

PEER REVIEWING ACTIVITY

Journals	I have reviewed papers for several scientific journals, including Nature Machine Intelligence (Springer Nature), Natural Language Engineering (Cambridge University Press), Engineering Applications of Artificial Intelligence (Elsevier), IEEE Transactions on Engineering Management (IEEE), and Expert Systems with Applications (Elsevier).

Conferences | I have been a member of the Program Committee of the EMNLP 2022 conference, the ACL ARR 2024 initiative, the LREC-COLING 2024 conference, and the 2022 Workshop on Insights from Negative Results.

PROJECTS

ITSERR (PNRR project B53C22001770006, with Dr. Carlo Meghini (ISTI-CNR) as PI) | Within this project I currently facilitate the annotation of texts by domain experts, and I am in charge of deploying the Inception tagging tool. After the conclusion of the manual tagging, I will focus on deploying machine-learning based methods to extract domain information from new texts.

WEMB (PRIN 2022 project 2022EPTPJ9, with Dr. Fabrizio Sebastiani (ISTI-CNR) as PI) | Within this project I am studying the performance of Large Language Models in assessing the abstraction and inclusiveness of words in context, in order to understand if these models can cope with this complex linguistic property of words. My work complements the study of these concepts carried out by linguists, by implementing computational Language Modelling techniques.

GRANTS OBTAINED

GELATINO (ISCRA C project HP10CQRW2J with Dr. Andrea Esuli (ISTI-CNR) as PI) | Period: 12 September, 2023 to 12 June, 2024. Project summary: I have collaborated in writing a project proposal that has been granted with 10k GPU hours on CINECA's Leonardo supercomputer to train small Language Models.

TRAVEL (ISCRA C project HP10CY9V7K with Dr. Alejandro Moreo (ISTI-CNR) as PI) | Period: 18 October, 2023 to 18 July, 2024. Project summary: I have collaborated in writing a project proposal that has been granted with 50k GPU hours on CINECA's Leonardo supercomputer to fine-tune Large Language Models on Italian and other European Languages.

EvaLMM (within AI4Media project European Commission H2020 n. 951911 with Dr. Andrea Esuli (ISTI-CNR)) | I have collaborated in writing a project proposal, answering an open call from the AI4media European Project. Our proposal, called "EvaLMM - Evaluating LMMs on common sense and factuality", was awarded € 45,000 (3 out of 7 awarded).

Grant for Young Mobility (GYM) | Period: 1 May, 2024 to 31 May, 2024. Project summary: 3,500 Euro Mobility Grant for a collaborative project with Associate Professor Anna Rogers at IT University of Copenhagen (on detection of synthetic text)

AWARDS

Area Chair Award of Association for Computational Linguistics for the paper *AI 'News' Content Farms are Easy to Make and Hard to Detect: A Case Study in Italian*.

BIBLIOMETRICS

Google Scholar: Citations: 229, h-index: 8

Scopus: Citations: 98, h-index: 5

PEER REVIEWED PUBLICATIONS

[1] Luca Moroni, **Giovanni Puccetti**, Pere-Luis Huguet Cabot, Andrej Stefan Bejgu, Edoardo Barba, Alessio Miaschi, Felice Dell'Orletta, Andrea Esuli, and Roberto Navigli. "Optimizing LLMs for Italian: Reducing Token Fertility and Enhancing Efficiency Through Vocabulary Adaptation". In: *Findings of the Association for Computational Linguistics: NAACL 2025*. Albuquerque, USA, 2025.

[2] **Giovanni Puccetti**, Marianna Marcella Bolognesi, and Andrea Esuli. "Wordnet and Word Ladders: Climbing the abstraction taxonomy with LLMs". In: *13th Global Wordnet Conference*. Pavia, IT, 2025.

[3] **Giovanni Puccetti**, Maria Cassese, and Andrea Esuli. "The Invalsi Benchmarks: measuring Linguistic and Mathematical understanding of Large Language Models in Italian". In: *31st International Conference on Computational Linguistics*. Abu Dhabi, UAE, 2025.

[4] Yuxia Wang, Artem Shelmanov, Jonibek Mansurov, Akim Tsvigun, Vladislav Mikhailov, Rui Xing, Zhuohan Xie, Jiahui Geng, **Giovanni Puccetti**, Ekaterina Artemova, Jinyan Su, Minh Ngoc Ta, Mervat Abassy, Kareem Ashraf Elozeiri, Saad El Dine Ahmed El Etter, Maiya Goloburda, Tarek Mahmoud, Raj Vardhan Tomar, Nurkhan Laiyk, Osama Mohammed Afzal, Ryuto Koike, Masahiro Kaneko, Alham Fikri Aji, Nizar Habash, Iryna Gurevych, and Preslav Nakov. "GenAI Content Detection Task 1: English and Multilingual Machine-Generated Text Detection: AI vs. Human". In: *Proceedings of the 1st Workshop on GenAI Content Detection (GenAIDetect)*. Ed. by Firoj Alam, Preslav Nakov, Nizar Habash, Iryna Gurevych, Shammur Chowdhury, Artem Shelmanov, Yuxia Wang, Ekaterina Artemova, Mucahid Kutlu, and George Mikros. Abu Dhabi, UAE: International Conference on Computational Linguistics, Jan. 2025, pp. 244–261. URL: https://aclanthology.org/2025.genaidetect-1.27/.

[5] Andrea Esuli, Fabrizio Falchi, Marco Malvaldi, and **Giovanni Puccetti**. "You Write like a GPT". In: *10th Italian Conference on Computational Linguistics*. Pisa, IT, 2024.

[6] **Giovanni Puccetti**, Maria Cassese, and Andrea Esuli. "INVALSI: Abstraction and Inclusiveness in Context: A CALAMITA Challenge". In: *10th Italian Conference on Computational Linguistics*. Pisa, IT, 2024.

[7] **Giovanni Puccetti**, Claudia Collacciani, Andrea Amelio Ravelli, Marianna Marcella Bolognesi, and Andrea Esuli. "ABRICOT: Abstraction and Inclusiveness in Context: A CALAMITA Challenge". In: *10th Italian Conference on Computational Linguistics*. Pisa, IT, 2024.

[8] **Giovanni Puccetti**, Anna Rogers, Chiara Alzetta, Felice Dell'Orletta, and Andrea Esuli. "AI 'News' Content Farms Are Easy to Make and Hard to Detect: A Case Study in Italian". In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Lun-Wei Ku, Andre Martins, and Vivek Srikumar. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 15312–15338. URL: https://aclanthology.org/2024.acl-long.817.

[9] Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Osama Mohammed Afzal, Tarek Mahmoud, **Giovanni Puccetti**, and Thomas Arnold. "SemEval-2024 Task 8: Multidomain, Multimodel and Multilingual Machine-Generated Text Detection". In: *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*. Ed. by Atul Kr. Ojha, A. Seza Doğruöz, Harish Tayyar Madabushi, Giovanni Da San Martino, Sara Rosenthal, and Aiala Rosá. Mexico City, Mexico: Association for Computational Linguistics, June 2024, pp. 2057–2079. DOI: 10.18653/v1/2024.semeval-1.279. URL: https://aclanthology.org/2024.semeval-1.279.

[10] Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Osama Mohammed Afzal, Tarek Mahmoud, **Giovanni Puccetti**, Thomas Arnold, Alham Aji, Nizar Habash, Iryna Gurevych, and Preslav Nakov. "M4GT-Bench: Evaluation Benchmark for Black-Box Machine-Generated Text Detection". In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Lun-Wei Ku, Andre Martins, and Vivek Srikumar. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 3964–3992. URL: https://aclanthology.org/2024.acl-long.218.

[11] Vito Giordano, **Giovanni Puccetti**, Filippo Chiarello, Tommaso Pavanello, and Gualtiero Fantoni. "Unveiling the inventive process from patents by extracting problems, solutions and advantages with natural language processing". In: *Expert Systems with Applications* 229 (2023), p. 120499. ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2023.120499.

[12] **Giovanni Puccetti** and Andrea Esuli. "AIMH at MULTI-Fake-DetectIVE: System Report". In: *Proceedings of the 8th Workshop on the Evaluation of Natural Language and Speech Tools for Italian (EVALITA 2023)*. Parma, IT, 2023. URL: https://api.semanticscholar.org/CorpusID:261529509.

[13] **Giovanni Puccetti**, Vito Giordano, Irene Spada, Filippo Chiarello, and Gualtiero Fantoni. "Technology identification from patent texts: A novel named entity recognition method". In: *Technological Forecasting and Social Change* 186 (2023), p. 122160. ISSN: 0040-1625. DOI: https://doi.org/10.1016/j.techfore.2022.122160.

[14] **Giovanni Puccetti**, Anna Rogers, Aleksandr Drozd, and Felice Dell'Orletta. "Outlier Dimensions that Disrupt Transformers are Driven by Frequency". In: *Findings of the Association for Computational Linguistics: EMNLP 2022*. Ed. by Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 1286–1304. DOI: 10.18653/v1/2022.findings-emnlp.93.

[15] **Giovanni Puccetti**, Filippo Chiarello, and Gualtiero Fantoni. "A simple and fast method for Named Entity context extraction from patents". In: *Expert Systems with Applications* 184 (2021), p. 115570. ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2021.115570.

[16] **Giovanni Puccetti**, Alessio Miaschi, and Felice Dell'Orletta. "How Do BERT Embeddings Organize Linguistic Knowledge?" In: *Proceedings of Deep Learning Inside Out (DeeLIO): The 2nd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*. Online: Association for Computational Linguistics, June 2021, pp. 48–57. DOI: 10.18653/v1/2021.deelio-1.6.

[17] **Giovanni Puccetti**, Luis Bolanos, Filippo Chiarello, and Gualtiero Fantoni. "B4DS @ PRELEARN: Ensemble Method for Prerequisite Learning (short paper)". In: *Proceedings of the 7th Workshop on the Evaluation of Natural Language and Speech Tools for Italian (EVALITA 2020)*. Vol. 2765. CEUR Workshop Proceedings. CEUR-WS.org, 2020. URL: http://ceur-ws.org/Vol-2765/paper107.pdf.

LANGUAGES

During my studies I attended different courses of language learning: Malta (English Language Academy), Galles, Canada (English as a second language). Paris (Alliance Francaise), Bruxelles.

Mother tongue Italian

Other languages

UNDERSTANDING		SPEAKING		WRITING
Listening	Reading	Spoken interaction	Spoken production	
B2	B2	B2	B2	B2
B1	B1	B1	B1	B1
C1	C1	C1	C1	C1

French — French row, Dutch — Dutch row, English — English row

Levels: A1 and A2: Basic user – B1 and B2: Independent user – C1 and C2: Proficient user
Common European Framework of Reference for Languages

IT skills **Python language**. Main packages: pytorch, pytorch-lightning, huggingface-transformers, huggingface-datasets, scikit-learn, spacy, nltk.
Julia language. Main packages: Javis.jl, Pluto.jl, DataFrames.jl, Statistics.jl, GLM.jl, MLJ.jl.
The use of scientific computing software, moderate ability in Matlab and R.

Software

Contrastive Captioners I was the main contributor on adding Contrastive Captioners models to https://github.com/mlfoundations/open_clip. Main pull request: https://github.com/mlfoundations/open_clip/pull/308 and more following ones adding generation. Blogpost with details https://laion.ai/blog/coca/.

Outliers Software for the study of outlier dimensions: `https://github.com/gpucce/outliersvsfreq`. This repository allows to reproduce all the experiments in [14] and more in general to measure the effect of outliers removal on transformer models. It is mostly based on PyTorch and HuggingFace-transformers.

GSOC Participate as a mentor in 2022 Google Summer of Code

Open source contributor `https://github.com/gpucce/`

Event Organization

Semeval24 Contributed to the organization of **SemEval-2024 Task 8: Multigenerator, Multidomain, and Multilingual Black-Box Machine-Generated Text Detection** (`https://github.com/mbzuai-nlp/SemEval2024-task8?tab=readme-ov-file#organizers`). I contributed to build the dataset of artificial news used in the task, generated using LLMs that I trained on Italian.

ENG4GED22 Participated in organizing the webinar series **ENG4GED** (`http://b4ds.unipi.it/eng4ged/`). This 6-months cycle of weekly webinars, partially funded by the University of Pisa, focused on Green, Education and Dual Use of Defence and involved presentations from academics and industry experts on topics related to these themes, ranging from the use of innovative technologies for online education, to circular economy, and to the analysis of novel cyber-security threats. I personally organized 3 webinars, one with a speaker from Leonardo explaining the company's innovation cycle, one about Language Modeling and Computational Linguistics, and a third about Gamification and Education.

UniPi19 Organized the event **A Glimpse of Complexity: The Lens of AI** `https://www.eventbrite.it/e/biglietti-a-glimpse-of-complexity-the-lens-of-ai-83581621881`, where Davide Bacciu, Professor of Computer Science at the University of Pisa, and Antonio Latela, CTO of AI Technologies, discussed the influence of novel applications of deep learning to several fields. The first speaker focused on the more recent developments of graph neural networks, while the second gave insights on the application of these technologies in industry.

Extra Curricular activities

At `https://gpucce.github.io/side_projects` I have a blog where I explain mathematics through animations.

Sports: Basketball, Rugby.

From 2000 to 2008 I was a boy scout in Siena.