CURRICULUM VITAE

PEPA ATANASOVA

Department of Computer Science,
University of Copenhagen
pepa@di.ku.dk

PERSONAL DETAILS

Pepa Atanasova
Øster Voldgade 3, 1350 Copenhagen, Denmark
Email: pepa@di.ku.dk
WWW: `apepa.github.io`
Twitter: @atanasovapepa

Github: `https://github.com/apepa`
LinkedIn: `https://www.linkedin.com/in/pepa-atanasova-65a2b417b/`
Google Scholar: `https://scholar.google.com/citations?user=CLOC3rEAAAAJ&hl=en`

RESEARCH INTERESTS

Natural Language Processing • Machine Learning • AI Explainability • Mechanistic Interpretability • Faithful Explainability Techniques • Explainability Diagnostics • Factuality • Knowledge Mechanisms • Accountability • Fairness

RESEARCH EXPERIENCE

2024 –	**Tenure-Track Assistant Professor** in Natural Language Processing (NLP) Section at the Department of Computer Science (DIKU), University of Copenhagen (UCPH)
2022 – 2024	**Postdoctoral Researcher** in Natural Language Processing (NLP) Section at the Department of Computer Science (DIKU), University of Copenhagen (UCPH)
2019 – 2022	**Ph.D. Student** in Natural Language Processing (NLP) Section at the Department of Computer Science (DIKU), University of Copenhagen (UCPH)
01 – 04/2022	**Research Intern** at Meta AI Research, Mountain View, USA
05 – 09/2020	**Research Intern** at Google, New York, USA
2017 – 2019	**Natural Language Processing Scientist** at Siteground, Bulgaria

ACADEMIC EDUCATION

2019 – 2022	Ph.D., Computer Science, **University of Copenhagen (*defense 8th November 2022*)**
2015 – 2017	M.S., Artificial Intelligence, **Sofia University "St. Kl. Ohridski", Bulgaria**
2011 – 2015	B.S., Computer Science, **Sofia University "St. Kl. Ohridski", Bulgaria**

GRANTS AND SCHOLARSHIPS

- University of Copenhagen Convergence Grant titled "DeceptionWeb: Foundations for Emergent Deception in Human-AI Interaction", 2026, PI, 3 836 644 DKK
- DFF-Research Project 2 titled "A Mechanistic Framework for Mitigating the Susceptibility of LLMs to Learning False Information", 2026, co-PI, 6 335 736 DKK
- Advancing Customer Experience AI copilots through tailored retrieval, context fusion and continuous adaption, Innovation Fund Denmark, Industrial PhD, 2026, primary supervisor.
- Pioneer Center for AI, Trustworthy AI Program grant, 2025, for collaborative, cross-disciplinary efforts within the Danish AI community, 50 000 DKK.
- DDSA Large Event Grant, 2025. Co-organising a local pre-ACL seminar to bring leading Natural Language Processing researchers to Denmark to foster international collaborations, 100 000 DKK.
- Informatics Europe (IE) best dissertation award, 2023, sponsored by Springer. PhD thesis published as a book in a dedicated Springer series.
- European Laboratory for Learning and Intelligent Systems (ELLIS) best dissertation award, 2023, sponsored by Kühborth Stiftung GmbH.
- PhD Fellowship under the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 801199.

TALKS

- *When Explanations Lie: Testing and Improving Faithfulness in Model Reasoning* Invited talk, Aurora Huawei Summit, Helsinki, March 2026
- *When Explanations Lie: Testing and Improving Faithfulness in Model Reasoning* Invited Talk, ELLIS Unconference, LLM Safety and Security Workshop, December 2025
- *Reality Check of LLM-driven Fact Verification: Retrieving and Utilising Evidence in the Wild.* Invited talk, Digital Tech Summit, November 2025.
- *The Challenge of Trustworthy AI Explanations.* Invited talk, Inria P16 Annual Conference, October 2025.

- *Facts Unveiled: Navigating Factuality in the Era of Generative Models.* Invited talk, Sheffield NLP Group, November 2024.
- *Navigating the Right to Explanation in AI: Methods and Technical Challenges* Keynote talk, Romanian AI days 2024, September 2024.
- *Facts Unveiled: Navigating Factuality in the Era of Generative Models.* Opening Keynote, iGeLU 2024, September 2024.
- *Exploring the Explainability Landscape: Testing and Enhancing Explainability Techniques.* Invited talk, Chalmers University, November 2023.
- *From Opacity to Clarity: Embracing Transparent and Accountable Fact Verification.* Invited conference talk, MISDOOM, November 2023.
- *Faithfulness Tests for Natural Language Explanations.* Talk at Nordic AI Meet, September 2023.
- *Explainable and Accountable Fact Checking.* Invited talk at The University of Massachusetts' NLP group, April 2023.
- *Methods for Accountable and Explainable Complex Reasoning Tasks.* Invited talk at the Responsible Data Science and AI Speaker Series at the University of Illinois at Urbana-Champaign, October 2022.
- *When Research Goes Wrong: Deepfakes!.* Invited for a panel as a part of the Legal Tech Research Talks at the University of Copenhagen's Faculty of Law, March 2022.
- *Explainable and Accountable Automatic Fact Checking.* Invited talk for the NLP group at Oxford, February 2022.
- *Explaining Automated Fact Checking Predictions and Current Vulnerabilities.* Invited talk at FAIR's AI and Society talk series, September 2021.
- *Check-worthiness of Claims in Political Debates.* Invited talk at Data Science Society Meetup, Sofia, September 2018.
- *Leveraging Expert Annotations for Fact-Checking.* Invited talk at DataBeers Copenhagen, May 2019.
- *Check-worthiness of Claims in Political Debates.* Invited talk at Information Retrieval Workshop invited speaker, RANLP, September 2017.
- *Finding the Right Articles – A Supervised Approach to Search.* Talk at PyData London, April 2017.

LEADING ROLES AND ORGANISING COMMITTEES

- ACL Special Interest Group on Representation Learning (SIGREP) Secretary, 2026-2028
- The European Chapter of the ACL (EACL) 2027, Publication Track chair
- The 2026 Conference on Empirical Methods in Natural Language Processing (EMNLP 2026), Industry Track chair

- *ACL Rolling Review (ARR)* - Area Chair/Action Editor, since 2023
- Workshop on Representation Learning for NLP (Repl4NLP) co-located with ACL, 2024, Organiser
- Conference of the European Chapter of the Association for Computational Linguistics (EACL), 2023, Website Chair
- SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval), 2020, Organiser
- SemEval-2019 task 8: Fact checking in community question answering forums, 2019, Organiser
- CLEF-2019 CheckThat Lab on Automatic Identification and Verification of Claims, 2019 Organiser
- CLEF-2018 CheckThat! Lab on automatic identification and verification of political claims, 2018, Organiser
- Recent Advances in Natural Language Processing (RANLP), Student Research Workshop, 2017, Organiser

TEACHING

- Continuing Education Data Science Course Development. University of Copenhagen. 2025-. Steering committee, course organiser.
- Fairness and Transparency in Machine Learning. Master's course, University of Copenhagen, 2025-. Course organiser; developed and delivered lab and lecture materials, and conducted examinations.
- Guest lectures at the IT-University of Copenhagen, graduate level, "Explainability and Explainability Evaluations". Master's course, 2023. Developed and delivered lecture materials for 30 students.
- Fairness and Transparency in Machine Learning. Master's course, University of Copenhagen, 2022-2023. Developed and delivered lab and lecture materials for 28 students, and participated in examinations.
- Introduction to Natural Language Processing. Master's course, University of Copenhagen, 2019-2022. Developed and delivered mainly lab materials for up to 50 students, supplemented by lecturing materials in the last year, and participated in examinations.
- Tutorial at the Advanced Language Processing Winter School (ALPS), "Explainability and Explainability Evaluations". Graduate level, 2021. Developed and delivered lab materials for 50 students.
- Information Retrieval. Master's course, Sofia University. Developed and delivered lab materials for 20 students.
- Natural Language Processing. Master's course, Sofia University, 2018-2019. Developed and delivered lab materials for 20 students.
- Data Mining. Master's course, Sofia University, 2016-2017. Developed and delivered lab materials for 20 students.

- Artificial Intelligence. Bachelor's course, Sofia University, 2016-2017. Developed and delivered lab materials for 20 students, and participated in examinations.
- Object-Oriented Programming. Bachelor's course, Sofia University, 2012-2013. Developed and delivered lab materials for 20 students.
- Introduction to Programming. Bachelor's course, Sofia University, 2012-2013. Developed and delivered lab materials for 20 students.
- Workshop "Introduction to Machine Learning" for industry practitioners. 2019 Developed and delivered lecture and lab materials for 20 practitioners.

PhD AND POSTDOC SUPERVISION

- Lenka Tetkova (2026 -), DDSA Postdoc (co-supervisor with Georgios Arvanitidis), geometry of trust: a unified framework for convex latent steering and causal concept alignment
- Cristiana Lazar (2026 -), Industrial PhD (co-supervised with Maria Maistro, Zendesk), advancing customer experience AI copilots through tailored retrieval, context fusion and continuous adaption.
- Katarzyna Lyczek (2026 -), PhD (main supervisor, co-supervised with Isabelle Augenstein), theoretical foundations of mechanistic interpretability methods.
- Jean Seo (2026 -), PhD (co-supervised with Isabelle Augenstein), mechanistic interpretability of sycophancy in LLMs.
- Sekh Manuil Islam (2024 -), PhD (co-supervised with Isabelle Augenstein), mechanistic interpretability of training dynamics for LLMs.
- Haeun Yu (2023 -), PhD (co-supervised with Isabelle Augenstein), mechanistic interpretability of parametric knowledge and context utilisation for LLMs.
- Jingyi Sun (2023 -), PhD (co-supervised with Isabelle Augenstein), evaluation and advancement of explainability techniques.

PUBLICATIONS

Publications: 47; Citations: 3364; h-index: 24; i10-index: 29 (source Google Scholar, May, 2026)

SELECTED PUBLICATIONS

- Yingming Wang, **Pepa Atanasova**. *Self-Critique and Refinement for Faithful Natural Language Explanations.* In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, Long Papers.
- Shuzhou Yuan, Jingyi Sun, Ran Zhang, Michael Farber, Steffen Eger, **Pepa Atanasova**, Isabelle Augenstein. *Graph-guided textual explanation generation framework.* In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, Long Papers.

- Sara Vera Marjanovic, Haeun Yu, **Pepa Atanasova**, Maria Maistro, Christina Lioma, Isabelle Augenstein *DYNAMICQA: Tracing Internal Knowledge Conflicts in Language Models.* Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2024), November 2024.
- Haeun Yu, **Pepa Atanasova**, Isabelle Augenstein. *Revealing the Parametric Knowledge of Language Models: A Unified Framework for Attribution Methods.* Under review at ACL Rolling Review for the Annual Meeting of the Association for Computational Linguistics (ACL), Long Papers, 2024.
- **Pepa Atanasova**, Oana-Maria Camburu, Christina Lioma, Thomas Lukasiewicz, Jakob Grue Simonsen, and Isabelle Augenstein. *Faithfulness Tests for Natural Language Explanations.* Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), July 2023.

PAPERS IN CONFERENCE PROCEEDINGS

- Yingming Wang, **Pepa Atanasova**. *Self-Critique and Refinement for Faithful Natural Language Explanations.* In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, Long Papers.
- Shuzhou Yuan, Jingyi Sun, Ran Zhang, Michael Farber, Steffen Eger, **Pepa Atanasova**, Isabelle Augenstein. *Graph-guided textual explanation generation framework.* In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, Long Papers.
- Lucie-Aimee Kaffee, **Pepa Atanasova**, Anna Rogers. *Local Differences, Global Lessons: Insights from Organisation Policies for International Legislation.* The 3rd Workshop on Regulatable ML @NeurIPS2025.
- Lovisa Hagstrom, Sara Vera Marjanovic, Haeun Yu, Arnav Arora, Christina Lioma, Maria Maistro, **Pepa Atanasova**, Isabelle Augenstein. *A Reality Check on Context Utilisation for Retrieval-Augmented Generation.* In Proceedings of the Association for Computational Linguistics 2025, Long Papers.
- Jingyi Sun, **Pepa Atanasova**, Isabelle Augenstein. *From Tokens to Span Interactions: a Multi-level Comparative Framework for Highlight-based Explanations.* To be published in NAACL 2025, Long Papers.
- Sara Vera Marjanovic, Haeun Yu, **Pepa Atanasova**, Maria Maistro, Christina Lioma, Isabelle Augenstein *DYNAMICQA: Tracing Internal Knowledge Conflicts in Language Models.* Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2024), November 2024.
- Haeun Yu, **Pepa Atanasova**, Isabelle Augenstein. *Revealing the Parametric Knowledge of Language Models: A Unified Framework for Attribution Methods.* Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), Long Papers, 2024.
- Sagnik Ray Choudhury, **Pepa Atanasova**, and Isabelle Augenstein. *Explaining Interactions Between Text Spans.* Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP), December 2023.
- **Pepa Atanasova**, Oana-Maria Camburu, Christina Lioma, Thomas Lukasiewicz, Jakob Grue Simonsen, and Isabelle Augenstein. *Faithfulness Tests for Natural Language Explanations.* Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), July 2023.

- Momchil Hardalov, **Pepa Atanasova**, Todor Mihaylov, Galia Angelova, Kiril Simov, Petya Osenova, Veselin Stoyanov, Ivan Koychev, Preslav Nakov, and Dragomir Radev. *bgGLUE: A Bulgarian General Language Understanding Evaluation Benchmark.* Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL), July 2023.
- **Pepa Atanasova**, Jakob Grue Simonsen, Christina Lioma, Isabelle Augenstein. *Diagnostics-Guided Explanation Generation.* In Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence (AAAI), February 2022.
- Wojciech Ostrowski, Arnav Arora, **Pepa Atanasova**, Isabelle Augenstein. *Multi-Hop Fact Checking of Political Claims.* Proceedings of the 30th International Joint Conference on Artificial Intelligence (IJCAI), August 2021.
- Sara Rosenthal, **Pepa Atanasova**, Georgi Karadzhov, Marcos Zampieri, Preslav Nakov. *SOLID: A Large-Scale Semi-Supervised Dataset for Offensive Language Identification.* Findings of the Association for Computational Linguistics (ACL-IJCNLP 2021), August 2021.
- **Pepa Atanasova**, Jakob Grue Simonsen, Christina Lioma, Isabelle Augenstein. *Generating Fact Checking Explanations.* Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), July 2020.
- **Pepa Atanasova**, Jakob Grue Simonsen, Christina Lioma, Isabelle Augenstein (2020). *A Diagnostic Study of Explainability Techniques for Text Classification.* Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), November 2020.
- **Pepa Atanasova**, Dustin Wright, Isabelle Augenstein. *Generating Label Cohesive and Well-Formed Adversarial Claims.* Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2020), November 2020.
- **Pepa Atanasova**, Jakob Grue Simonsen, Christina Lioma, Isabelle Augenstein. *Generating Fact Checking Explanations.* Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), July 2020.
- Slavena Vasileva, **Pepa Atanasova**, Lluis Márquez, Alberto Barrón-Cedeño, Preslav Nakov. *It Takes Nine to Smell a Rat: Neural Multi-Task Learning for Check-Worthiness Prediction.* Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP), September 2019.
- **Pepa Atanasova**, Preslav Nakov, Georgi Karadzhov, Mitra Mohtarami, Giovanni Da San Martino. *Overview of the CLEF-2019 CheckThat! Lab: Automatic Identification and Verification of Claims. Task 1: Check-Worthiness.* International Conference of the Cross-Language Evaluation Forum for European Languages, August 2019
- **Pepa Atanasova**, Georgi Karadzhov, Yasen Kiprov, Preslav Nakov, Fabrizio Sebastiani. *Evaluating Variable-Length Multiple-Option Lists in Chatbots and Mobile Search.* Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR). July 2019.
- Tamer Elsayed, Preslav Nakov, Alberto Barrón-Cedeño, Maram Hasanain, Reem Suwaileh, Giovanni Da San Martino, **Pepa Atanasova**. *CheckThat! at CLEF 2019: Automatic identification and verification of claims.* In European Conference on Information Retrieval (ECIR), April 2019.
- Preslav Nakov, Alberto Barrón-Cedeño, Tamer Elsayed, Reem Suwaileh, Lluis Márquez, Wajdi Zaghouani, **Pepa Atanasova**, Spas Kyuchukov, Giovanni Da San Martino. *Overview of the CLEF-2018 CheckThat! Lab on automatic identification and verification of political*

claims. International conference of the Cross-Language Evaluation Forum for European languages (CLEF), August 2018.

- Israa Jaradat, **Pepa Atanasova**, Alberto Barrón-Cedeño, Lluis Márquez, Preslav Nakov. *ClaimRank: Detecting Check-Worthy Claims in Arabic and English*. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations (NAACL), June 2018.
- Georgi Karadzhov, **Pepa Atanasova**, Preslav Nakov, Ivan Koychev. *We Built a Fake News / Click Bait Filter: What Happened Next Will Blow Your Mind!* In Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP), September 2017.
- **Pepa Atanasova**, Preslav Nakov, Lluis Márquez, AAlberto Barrón-Cedeño, and Ivan Koychev. *A Context-Aware Approach for Detecting Worth-Checking Claims in Political Debates*. In Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP), September 2017.
- Yasen Kiprov, **Pepa Atanasova**, Ivan Koychev. *Generating Labeled Datasets of Twitter Users*. In Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization (UMAP), July 2017.

JOURNAL ARTICLES

- Jingyi Sun, **Pepa Atanasova**, Sagnik Ray Choudhury, Sekh Mainul Islam, Isabelle Augenstein. *Evaluation Framework for Highlight Explanations of Context Utilisation in Language Models*. In Computational Linguistics, 2026.
- **Pepa Atanasova**, Jakob Grue Simonsen, Christina Lioma, Isabelle Augenstein. *Fact Checking with Insufficient Evidence*. Transactions of the Association for Computational Linguistics (TACL), Vol 10 (2022).
- Shailza Jolly, **Pepa Atanasova**, Isabelle Augenstein. *Generating Fluent Fact Checking Explanations with Unsupervised Post-Editing*. Information, Vol 13 (2022).
- Luna De Bruyne, **Pepa Atanasova**, Isabelle Augenstein. *Joint Emotion Label Space Modelling for Affect Lexica*. Computer Speech & Language, Volume 71, 2022.
- **Pepa Atanasova**, Preslav Nakov, Lluis Márquez, Alberto Barrón-Cedeño, Georgi Karadzhov, Tsvetomila Mihaylova, Mitra Mohtarami, James Glass. *Automatic Fact-Checking Using Context and Discourse Information*. J. Data and Information Quality 11, (JDIQ), 2019.

PAPERS IN WORKSHOP PROCEEDINGS

- Marcos Zampieri, Preslav Nakov, Sara Rosenthal, **Pepa Atanasova**, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, Çağrı Çöltekin. *SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020)*. Proceedings of the Fourteenth Workshop on Semantic Evaluation (SemEval), December 2020.
- Tsvetomila Mihaylova, Georgi Karadzhov, **Pepa Atanasova**, Ramy Baly, Mitra Mohtarami, Preslav Nakov. *SemEval-2019 Task 8: Fact Checking in Community Question Answering Forums*. Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval), June 2019.

- **Pepa Atanasova**, Preslav Nakov, Georgi Karadzhov, Mitra Mohtarami, Giovanni Da San Martino. *Overview of the CLEF-2019 CheckThat! Lab: Automatic Identification and Verification of Claims. Task 1: Check-Worthiness.* Sun SITE Central Europe Workshop (CEUR-WS), June 2018.
- Alberto Barrón-Cedeño, Tamer Elsayed, Reem Suwaileh, Lluis Márquez, **Pepa Atanasova**, Wajdi Zaghouani, Spas Kyuchukov, Giovanni Da San Martino, Preslav Nakov. *Overview of the CLEF-2019 CheckThat! Lab: Automatic Identification and Verification of Claims. Task 2: Factuality.* Sun SITE Central Europe Workshop (CEUR-WS), June 2018.
- Tsvetomila Mihaylova, **Pepa Atanasova**, Martin Boyanov, Ivana Yovcheva, Todor Mihaylov, Momchil Hardalov, Yasen Kiprov, Daniel Balchev, Ivan Koychev, Preslav Nakov, Ivelina Nikolova Galia Angelova. *SUper Team at SemEval-2016 Task 3: Building a Feature-Rich System for Community Question Answering.* In Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval), June 2016.
- **Pepa Atanasova**, Martin Boyanov, Elena Deneva, Preslav Nakov, Yasen Kiprov, Ivan Koychev, and Georgi Georgiev. *PANcakes team: A Composite System of Genre-Agnostic Features for Author Profiling.* In CEUR Workshop Proceedings (CEUR-WS), June 2016.

BOOKS AND EDITED VOLUMES

- Chen Zhao, Marius Mosbach, **Pepa Atanasova**, Seraphina Goldfarb-Tarrent, Peter Hase, Arian Hosseini, Maha Elbayad, Sandro Pezzelle, Maximilian Mozes. *Proceedings of the 9th Workshop on Representation Learning for NLP (RepL4NLP-2024), August 2024.*
- **Pepa Atanasova**. *Accountable and Explainable Methods for Complex Reasoning over Text.* Last publishing stage of the PhD Thesis as a Springer book in a dedicated series, as part of the Informatics Europe best PhD Dissertation award.
- Venelin Kovatchev, Irina Temnikova, **Pepa Atanasova**, Yasen Kiprov, Ivelina Nikolova. *Proceedings of the Student Research Workshop Associated with RANLP 2017.* September 2017.