Catarina G Belem

https://pastelbelem8.github.io/

Last modified: June 21, 2026

catarina.garcia.belem@gmail.com

Research Focus

- **Uncertainty Quantification and Uncertainty Communication in LLMs**: Interested in better understanding how to quantify and estimate uncertainty in LLMs, as well as how to communicate uncertainty to humans or other LLMs (e.g., reasoning models, LM Agents, and Multi-Agent Systems).
- **LM Control**: Interested in exploring techniques for more effective, efficient, and robust decoding customization of LLMs.
- **LLM Evaluation**: Focused on the evaluation of hallucinations, fairness, safety, and robustness of LMs.

Education

- Ph.D., Computer Science | University of California, Irvine 2021-Present
  - **Advisors**: Sameer Singh, Padhraic Smyth
  - **Expected Graduation Date**: December 2026
- M.Sc., Computer Science and Engineering | Instituto Superior Tecnico, University of Lisbon 2016-2019
  - **Thesis**: Optimization of Time-Consuming Objective Functions: Derivative-Free Approaches and their Application in Architecture
- B.Sc., Computer Science and Engineering | Instituto Superior Tecnico, University of Lisbon 2013-2016

Publications (* denotes denotes joint authorship)

- **Journal**
  - M. Steyvers, H. Tejeda, A. Kumar, **C. Belem**, S. Karny, X. Hu, L. Mayer, P. Smyth, "What Large Language Models Know and What People Think They Know" *Nature Machine Intelligence*, https://doi.org/10.1038/s42256-024-00976-7, January 2025.
- **Conference**
  - Y. F. Bakman, S. Kang, Z. Huang, D. N. Yaldiz, **C. Belem**, C. Zhu, A. Kumar, A. Samuel, D. Liu, S. Avestimehr, s. P. Karimireddy, "Uncertainty as Feature Gaps: Epistemic Uncertainty Quantification of LLMs in Contextual Question-Answering" , The Fourteenth International Conference on Learning Representations (ICLR 2026).
  - **C. Belem**, P. Pezeshkpour, H. Iso, S. Maekawa, N. Bhutani, E. Hruschka, "From Single to Multi: How LLMs Hallucinate in Multi-Document Summarization" *Findings of the 2025 Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL)*, https://arxiv.org/abs/2410.13961, April 2025.
  - **C. Belem***, M. Kelly*, S. Singh, M. Steyvers, P. Smyth, "Perceptions of Linguistic Uncertainty by Language Models and Humans" *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp.8467–8502, 2024.
  - **C. Belem**, P. Seshadri, Y. Razeghi, S. Singh, "Are Models Biased on Text without Gender-related Language?" *Proceedings of the 12th International Conference on Learning Representations (ICLR)*, 2024.
  - A. F. Cruz, **C. Belem**, J. Bravo, P. Saleiro, P. Bizarro, "FairGBM: Gradient Boosting with Fairness Constraints" *Proceedings of the 11th International Conference on Learning Representations (ICLR)*, 2023.
  - A. F. Cruz, P. Saleiro, **C. Belem**, C. Soares, P. Bizarro, "Promoting fairness through hyperparameter optimization" *Proceedings of the 2021 IEEE International Conference on Data Mining (ICDM)*, pp.1036-1041, 2021.
  - S Jesus, **C Belem**, V. Balayan, J. Bento, P. Saleiro, P. Bizarro, and J. Gama, "How can I choose an explainer? An Application-grounded Evaluation of Post-hoc Explanations." *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*, 2021.
- **Workshop**
  - **C. Belem**, P. Glenn, A. Samuel, A. Kumar, D. Liu, "Readability Reconsidered: A Cross-Dataset Analysis of Reference-Free Metrics" EMNLP 2025 TSAR Workshop, 2025.
  - K. Ahmed*, **C. Belem***, P. Smyth, S. Singh, "Semantic Probabilistic Control of Language Models" NeurIPS'25 SPIGM Workshop, 2025.

- R. Longjohn, S. Wu, S. Kher, **C. Belem**, P. Smyth, "Bayesian Evaluation of Blackbox LLM Behavior" NeurIPS'25 LLM Evaluations Workshop, 2025.
- **C. Belem***, M. Kelly*, S. Singh, M. Steyvers, P. Smyth, "Can LMs Interpret Verbalized Uncertainty?" TrustNLP Workshop in NAACL, 2024 (**Runner-Up Best Short Paper**).
- **C. Belem**, V. Balayan, P. Saleiro, P. Bizarro, "Weakly supervised multi-task learning for concept-based explainability" *Weakly Supervised Learning Workshop (WeaSul) at ICLR*, arxiv:2104.12459, 2021.
- V. Balayan, **C. Belem**, P. Saleiro, P. Bizarro, "Teaching the Machine to Explain Itself using Domain Knowledge " *Human And Machine in-the-Loop Evaluation and Learning Strategies (HAMLETS) Workshop at NeurIPS, arXiv:2012.01932*, 2020.

- **Under Review**
  - **C. Belem**, S. Wu, H. Yao, M. Steyvers, S Singh, P. Smyth, "From 'May' to 'Is': Certainty Distortion in Language Model Rewriting" , 2026.
  - S. Wu, H. Yao, **C. Belem**, S. Fu, M. Steyvers, P. Smyth, "The Impact of AI Usage and Informativeness on Skill Development in Logical Reasoning" arxiv.org/abs/2605.21695, 2026.
  - M. Steyvers, **C. Belem**, P. Smyth, "Improving Metacognition and Uncertainty Communication in Language Models" arXiv.org/abs/2510.05126, 2026.
  - K. Ahmed*, **C. Belem***, P. Smyth, S. Singh, "Semantic Probabilistic Control of Language Models" , 2025.

Research Experience

- CapitalOne, New York | Applied Research Intern — Jun 2025 - Sep 2025
  - **Customization of LLMs**: Investigated the controllability and style transfer of language models for production of readable and accurate outputs. Proposed and implement a reinforcement learning customization approach (GRPO algorithm using the TRL Python library). Authored 2 papers during the internship, including, a workshop paper accepted at the TSAR workshop at EMNLP 2025, focused on the systematic evaluation of readability metrics; and working towards a conference paper focused on benchmarking and proposing an RL-based approach for joint alignment of readability and accuracy.
- Megagon Labs, Mountain View | Research intern — Jun 2024 - Sep 2024
  - **Evaluating faithfulness in multi-document summarization (MDS)**: Analyzed hallucinations in multi-document summarization across 5 popular LLMs, proposing a taxonomy of error types through large-scale human annotation. Evaluated adversarial robustness and tested mitigation strategies (*e.g.*, NLI, LLM-as-a-judge). Resulting in a publication at NAACL-2025.
- University of California, Irvine | Graduate student researcher — 2021 - Present
  - **Uncertainty in LLMs**: Studied LLM calibration and linguistic uncertainty, including its effects on human-AI decision-making (Nature 2025) and alignment with human perceptions (EMNLP 2024). Currently exploring uncertainty quantification and communication on agentic LMs, reasoning models, and conversational settings.
  - **Constrained Decoding**: Proposed a method that leverages an attribute verifier's gradient information to efficiently reason over all generations that satisfy the target attribute (e.g., toxicity, sentiment, topic), enabling precise steering of LM generations by reweighing the next-token distribution.
  - **Fair NLP**: Developed an evaluation benchmark to uncover gender bias in non-stereotypical contexts, revealing significant biases in models (ICLR 2024). Interested in exploring inference-time and alignment techniques to mitigate biases or in evaluating social biases in LLMs and VLMs across downstream tasks.
  - **NLG Evaluation**: Used parameter efficient fine-tuning (PEFT) and in-context learning (ICL) techniques to perform automatic evaluation of generative LLMs with few labeled examples in machine translation, summarization, question-answering tasks.
- Feedzai, Lisbon, Portugal | Research data scientist — 2019 - 2021
  - **Algorithmic Fairness / Fair ML**: Investigated constrained optimization and hyperparameter selection techniques to develop bias mitigation methods for model selection and model training. Authored 2 patents and 2 papers to top tier conferences (ICDM 2021, ICLR 2023)
  - **Explainable AI**: Evaluated the impact of explanations on human decision-making through a set of user studies involving domain experts. Implemented deep neural networks to make predictions regarding fraud detection and provide concept-based explanations for domain experts. Authored 3 patents and 3 papers (WS at NeurIPS'20, FAccT'21, WS at ICLR'21).
- INESC-ID, Lisbon, Portugal | Graduate student researcher — 2017 - 2019
  - **Multi-Objective Optimization**: Investigated gradient-free methods, including genetic algorithms and ML algorithms, to perform multi-objective optimization of time-consuming objective functions in architectural design.

Skills

- **Programming Languages**: Python, Java, Julia
- **Machine Learning Frameworks**: scikit-learn, Pytorch, Apache Spark, HuggingFace Transformers
- **Data analysis tools**: Numpy, Pandas
- **Other**: Docker, PostgreSQL, SLURM
- **Courses**: Natural language processing (NLP), Machine Learning, Probabilistic learning, Deep generative models, Graphical models

Awards

- ICS Steckler Family Endowed Fellowship — Sep 2024 - Jun 2025
- Fulbright Scholar — Sep 2021 - Jun 2025
- Grace Hopper Celebration scholarship — 2022
- CS Department Excellence Fellowship — 2021
- Maria de Lourdes Pintasilgo Award - Young Alumna — 2019
- Teaching Excellency Award — 2019

Invited Talks

- **C. Belem**, "Perceptions of Linguistic Uncertainty by Language Models and Humans" Mila/McGill NLP reading group, April 2025.
- **C. Belem**, "Can Language Models Perceive Verbalized Uncertainty?" Runner-Up for Best Short Paper at TrustNLP Workshop at NAACL 2024, June 2024.
- **C. Belem**, "An Introduction to RLHF (preference learning)" Cognitive Science Department at University of California Irvine, February 2024.
- **C. Belem**, "On the Calibration of Generative Question-Answering Models" Priberam Machine Learning Lunch Seminars, May 2022.
- V. Balayan & **C. Belem**, "Concept-based Explainability: Challenges and Applications to Fraud Detection" Deep Learning Sessions Portugal Meetup, July 2021.

Community Involvement

- Mentored an Undergraduate Honors Thesis on measuring gender-occupation bias amplification — Sep 2023 - Jun 2024
- Mentorship in Machine Learning to a high school student — Jun 2023 - Sep 2023
- Jury at World Data League Competition — 2021,2022
- Organizer and host at Deep Learning Sessions Portugal — Mar 2021 - Jun 2023

Teaching Assistant Experience

- Projects in AI - University of California, Irvine — Winter 2026
  - **Responsibilities:**: Hosted office hours; Proposed and graded NLP-oriented; Mentored students in their project development.
- Statistical NLP - University of California, Irvine — Spring 2023
  - **Responsibilities:**: Hosted office hours; Graded reports of students; Developed homework assignments, namely implemented a retrieval-augmented generation (RAG) pipeline for biomedical question answering and pretraining LLM and implemented decoding algorithms.
- Machine Learning for NLP - University of California, Irvine — Aug 2022
- Advanced Programming - Instituto Superior Tecnico, University of Lisbon — Feb 2018 - Jul 2019
- Programming Languages, Instituto Superior Tecnico, University of Lisbon — Feb 2018 - Jul 2018

Service (* denotes denotes recognition as Top Reviewer)

- **2026 - Conferences**: ICML*, ICLR, ARR (Jan, Mar, May), NeurIPS
- **2026 - Journal**: CHBAH
- **2025 - Conferences**: ICLR, NeurIPS*
- **2024 - Conferences**: CoLM, ARR (Jun, Aug*, Oct*, Dec*)
- **2024 - Journal**: IEEE TNNLS (Dec)
- **2024 - Workshops**: XAI (NeurIPS), RBFM (NeurIPS), SeT LLM (ICLR)