Alessio **Cocchieri**

PhD student at the University of Bologna

Department: Computer Science and Engineering | Research Group: UniboNLP | Supervisor: Prof. Gianluca Moro

About Me

I am a last-year NLP PhD candidate (ending Nov 2026) at UniboNLP, with 9+ publications at ACL, EMNLP, NAACL, and EACL. I specialize in LLM evaluation and knowledge distillation for low-resource NLP, with applications to high-stakes domains like medicine. I also serve as an active reviewer for ARR and top-tier ML conferences like NeurIPS.

Selected Publications (* Equal Contribution)

ACL 2026 (Main) — **Cocchieri**, et al. *LLMs (Almost) Never Abstain Under Medical Uncertainty.* (to appear)
TL;DR. We introduce MedQAbstain, a benchmark for medical abstention under uncertainty, revealing that state-of-the-art LLMs systematically overcommit, rarely abstaining even when the question itself is hidden.

EACL 2026 (Main) — **Cocchieri**, et al. *ReMedQA: Are We Done With Medical Multiple-Choice Benchmarks?*
TL;DR. We show that high LLM accuracy in medical MCQA masks severe inconsistency. We propose novel metrics to evaluate true reliability across MCQA formats.

ACL 2025 (Main) — **Cocchieri**, et al. *What do you call a dog that is incontrovertibly true? Dogma: Testing LLM Generalization through Humor.*
TL;DR. We introduce Phunny, a novel benchmark using uncontaminated English puns, revealing that LLMs struggle with generalization even on simple tasks, consistently underperforming the human baseline.

EMNLP 2025 (Main) — **Cocchieri**, et al. *Can Large Language Models Win the International Mathematical Games?*
TL;DR. We introduce MathGames, a novel multimodal benchmark of age-graded math problems from an international competition, showing that frontier LLMs underperform compared to humans, including 11-year-olds.

NAACL 2025 (Findings) w/ IBM Research — **Cocchieri**, et al. *OpenBioNER: Lightweight Open-Domain Biomedical NER Through Entity Type Description*
TL;DR. We introduce a 110M BERT model that leverages descriptions for zero-shot Biomedical NER, outperforming GPT-4o, specialized LLMs, and GLiNER by up to 10% F1.

ACL 2024 (Main) — Frisoni*, **Cocchieri***, et al. *To Generate or to Retrieve? On the Effectiveness of Artificial Contexts for Medical Open-Domain Question Answering*
TL;DR. We introduce MedGENIE, the first *generate-then-read* framework for open-domain medical QA, demonstrating the effectiveness of generated over retrieved contexts and significantly improving LLM RAG performance.

Work Experience

IBM Research — **Research Scientist Internship** (PhD Intern) Dublin (IR) | March 2026 — May 2026
Natural Language Processing — LLM Factuality
- Created a novel benchmark for multimodal scientific fact-checking, targeting LLM long-form generation
- Coordinated a multi-team human annotation pipeline, overseeing annotation guidelines, quality control, and IAA
- Conducted systematic analysis exposing critical factuality fragilities in multimodal LLMs across scientific domains

IBM Research — **Research Scientist Internship** (Master Thesis) Dublin (IR) | March 2023 — May 2023
Natural Language Processing — Named Entity Recognition
- Leveraged LLM distillation to improve smaller-sized models for zero-shot NER
- Produced two peer-reviewed publications: OpenBioNER (NAACL 2025) and ZeroNER (ACL 2025)
- Contributed to the IBM zshot library for zero-shot NER model inference (#82, #84)

Education

Doctor of Philosophy (PhD) — **University of Bologna** Bologna (IT) | November 2023 — Now
Natural Language Processing | LLMs, RAG, Information Extraction, Benchmarking, Low-resource

Master of Science (MSc) — **University of Bologna** Bologna (IT) | September 2021 — December 2023
Grade: 110 with honors / 110 | Artificial Intelligence

Bachelor of Science (BSc) — **University of Bologna** Bologna (IT) | September 2018 — July 2021
Grade: 110 with honors / 110 | Computer Science

Skills

- **Hard:** Strong: Python, PyTorch, Hugging Face, vLLM, Scikit-Learn, Git, NumPy, Pandas, NLTK, SpaCy; Proficient: Java, Docker, MySQL, MongoDB, TensorFlow, LangChain, Object-Oriented Programming
- **Soft:** Communication, Problem-Solving, Teamwork, Quick Learning, Leadership, Adaptability, Public Speaking
- **Languages:** Italian (native), English (professional fluency), Spanish (basic proficiency)