# Dayyán O'Brien

dayyanobrien.github.io

+44 7770 655794

dayyanobrien@icloud.com

## EDUCATION

**University of Edinburgh**  Edinburgh, UK

*PhD in Designing Responsible NLP*  *Sep 2025 – Aug 2029*

- **Supervision:** Dr. Emily Allaway
- Fully funded by the **G-Research PhD Scholarship** for outstanding academic merit.

*Master of Informatics* [transcript] (**First Class Honours**)  *Sep 2019 – May 2024*

- **Dissertation supervision**: Prof. Mirella Lapata
- **Courses:** Natural Language Processing, Machine Learning, Reinforcement Learning, Calculus, Linear Algebra

**Abbey Christian Brothers' Grammar School**  Newry, UK

*Secondary school* (AAA in A-Level Mathematics, Physics and Computer Science)  *Sep 2012 – May 2019*

## PUBLICATIONS

**MatheMagic: Generating Dynamic Mathematics Benchmarks Robust to Memorization**
**Dayyán O'Brien**, Pinzhen Chen, Barry Haddow, Emily Allaway
*arXiv:2510.05962. Under review for ACL Rolling Review (ARR).* 2025.
*MathNLP Workshop, Empirical Methods in Natural Language Processing (EMNLP) (non-archival).* 2025.

**DocHPLT: A Massively Multilingual Document-Level Translation Dataset**
**Dayyán O'Brien**, Bhavitvya Malik, Ona De Gibert Bonet, Pinzhen Chen, Barry Haddow, Jörg Tiedemann
*Proceedings of the Conference on Machine Translation (WMT).* 2025.

**An Expanded Massive Multilingual Dataset for High-Performance Language Technologies**
Laurie Burchell, Ona de Gibert, *et al.* (incl. **Dayyán O'Brien**)
*Proceedings of the Association for Computational Linguistics (ACL).* 2025.

**Mind the Gap: Diverse NMT Models for Resource-Constrained Environments**
Ona De Gibert Bonet, **Dayyán O'Brien**, Dušan Variš, Jörg Tiedemann
*Proceedings of the Nordic Conference on Computational Linguistics (NoDaLiDa).* 2025.

**EMMA-500: Enhancing Massively Multilingual Adaptation of Large Language Models**
Shaoxiong Ji, Zihao Li, *et al.* (incl. **Dayyán O'Brien**)
*arXiv:2409.17892. Under review for the Journal of Data-centric Machine Learning Research (DMLR).* 2024.

**Prompting Numerical Commonsense Reasoning across Languages**
**Dayyán O'Brien**
*Outstanding Honours Thesis, School of Informatics, University of Edinburgh.* 2024.

**Numerical Commonsense Reasoning across Languages**
**Dayyán O'Brien**
*Outstanding Honours Thesis, School of Informatics, University of Edinburgh.* 2023.

## RESEARCH EXPERIENCE

**University of Edinburgh**  Edinburgh, UK

*Doctoral Researcher, supervised by Emily Allaway*  *Sep 2025 – Present*

- Improving the compositionality of language models.

*Research Assistant, supervised by Barry Haddow*  *Aug 2024 – Aug 2025*

- Co-developed the HPLT 2.0 bitexting pipeline to mine parallel data.

*Junior Research Assistant, supervised by Pinzhen Chen*  *May 2024 – Jul 2024*

- Cleaned and processed data for EMMA-500, a 7B parameter LLM for over 500 languages.

*Junior Research Assistant, supervised by Mirella Lapata*  *Jun 2022 – Apr 2024*

- Curated and evaluated mNumersense: 36k+ Arabic, Chinese, Russian numeric commonsense sentences.

## Awards & Achievements

**G-Research PhD Scholarship (2025 – 2029)**

**Outstanding Honours Project (2023 & 2024)**

**Runner-up Best Coursework for Reasoning and Agents (2021)**

**Exemplary Project for Foundations of Data Science (2021)**

## Major Contributions & Open Source

**Leader, Document-Level HPLT Corpus:** Large-scale, document-level parallel corpus ([huggingface.co/datasets/HPLT/DocHPLT](huggingface.co/datasets/HPLT/DocHPLT)).

**Contributor, EMMA-500:** An open-source LLM for 500+ languages ([github.com/MaLA-LM/emma-500](github.com/MaLA-LM/emma-500)).

**Contributor, HPLT 2.0 Bitexting Pipeline:** Parallel data mining pipeline ([hplt-project.org/datasets/v2.0).](hplt-project.org/datasets/v2.0))

## Teaching Experience

| | |
|---|---|
| **University of Edinburgh** | Edinburgh, UK |
| *Demonstrator for Accelerated Natural Language Processing ([INFR11125](INFR11125))* | *Sep 2023 – Nov 2023* |
| *Tutor for Foundations of Natural Language Processing ([IINFR10078](IINFR10078))* | *Jan 2023 – May 2023* |
| *Tutor for Foundations of Data Science ([INFR08030](INFR08030))* | *Sep 2022 – May 2023* |
| *Demonstrator for Foundations of Natural Language Processing ([INFR10078](INFR10078))* | *Jan 2024 – Mar 2024* |
| **Self-employed** | Edinburgh, UK |
| *Private tutor* | *Sep 2020 – Mar 2024* |

## Industry Experience

| | |
|---|---|
| **Kainos** | Belfast, UK |
| *Software Engineer* | *Mar 2018* |
| **Bombardier** | Belfast, UK |
| *Engineer* | *Sep 2017* |
| **Computer Hospital** | Newry, UK |
| *Computer technician* | *Oct 2017* |

## Leadership & Outreach

| | |
|---|---|
| **G-Research Brand Ambassador** | Edinburgh, UK |
| *G-Research* | *Sep 2025 – Present* |
| **Committee Member, Community Events Planning** | Edinburgh, UK |
| *Edinburgh Bahá'í Community* | *Mar 2025 – Present* |
| **Primary School Micro:bit Programming Workshop Facilitator** | Bathgate, UK |
| *University of Edinburgh* | *Jan 2024 – Feb 2024* |

## Skills

**Programming:** Python, Java, Haskell, LaTeX, SQL

**Libraries:** PyTorch, Tensorflow, HuggingFace, NLTK, NumPy, Slurm, Kubernetes, Docker, MTurk, pandas, sklearn, statsmodels, Matplotlib, Festival, HTK, Tkinter, Google Cloud VM, Weather & Maps API, Seaborn, Kivy, JUnit, Maven

**Languages:** English (Native), German (Professional)