

DAVID ALFONSO-HERMELO

david.alfonso.hermelo@gmail.com | (+1) 438 399 05 66

 david-alfonso-hermelo-6646a1b1

 dahrs

 david-alfonso-hermelo

 dahrs

TL;DR

Senior Data Scientist at Huawei | 8+ years of experience | Masters in CS, NLP, and Ling. | Specialist in Big Data, AI, DL, Python.

SKILLS

EXPERTISE

• Large Language Models (LLM) • Data Science • Data analysis Natural Language Processing (NLP) • Deep Learning models • Data management • Natural Language Understanding • Linguistics

PROGRAMMING LANGUAGES

Python (9 years) • Bash (2 years) • Javascript (2 years) • Java (1 year) • Perl (2 years) • C# (4 months) • R (6 months)

MARKUP LANGUAGES

HTML • \LaTeX • XML • XLM • CSS • Markdown

QUERY LANGUAGES

SQL • SPARQL • XSLT • OWL

NATURAL LANGUAGES

Spanish (Native) • French (fluent, DALF C2) • English (fluent, Cambridge C1) • German (beginner, A2)

RELEVANT TOOLS

LLM APIs (2yrs) • Pytorch (4yrs) • HuggingFace (6yrs) • Selenium (6yrs) • Pandas (7 years) • AWS BedRock (1yr) • Azure AI (1yr)

RELEVANT EXPERIENCE

HUAWEI'S NOAH'S ARK LAB | NLP RESEARCHER, DATA SCIENTIST

December 2019 - To this day | Montreal, Quebec, Canada

- Led 6 annotation projects overseeing teams of annotators ranging in size from 2 to 32 members.
- Co-authored and published 9 papers, including 5 in A* conferences, with 2 additional papers currently in the submission process.
- Oversaw the assessment of a specialized fine-tuned LLM using citation-based QA that integrates web retrieval and KG extraction.
- Mentored the data science team for the successful release of 5 client products.
- Led the design of metrics and evaluation protocols for 4 projects.
- Conducted research in NLP tasks including Large Language Models, Information Retrieval, Natural Language Understanding, Machine Translation, Data Augmentation, and Named Entity Recognition.
- Mentored new team members on Data management, Agile methodologies, PEP-8 coding practices.

UNIVERSITY OF MONTREAL | RESEARCHER

January 2017 - November 2020 | Montreal, Quebec, Canada

- Successfully trained Machine Translation and Data Augmentation models that led to develop 2 impactful proofs of concept.
- Co-authored and published three significant research papers on Machine Translation, Taxonomy, and Knowledge Graph Population.
- Led the in-house research and development of a client's automatic skill ontology product.
- Directed the in-house research and development of a client's data extraction tool, seamlessly integrating it into its existing product.

RELEVANT EDUCATION

MASTER - NATURAL LANGUAGE PROCESSING | SORBONNE NOUVELLE UNIVERSITY

2015 - 2017 | Paris, France

MASTER - LANGUAGE SCIENCES | GRENOBLE-ALPES UNIVERSITY

2010 - 2012 | Grenoble, France

MASTER - APPLIED LINGUISTICS | UNIVERSITY OF HAVANA

2010 - 2012 | Havana, Cuba

RESEARCH PUBLICATIONS

2024

- [1] EUROPA: A Legal Multilingual Keyphrase Generation Dataset, ACL 2024.
- [2] CHARP: Conversation History Awareness Probing for Knowledge-grounded Dialogue Systems, ACL 2024.
- [3] EWEK-QA : Enhanced Web and Efficient Knowledge Graph Retrieval for Citation-based Question Answering Systems, ACL 2024.
- [4] Efficient Citer: Tuning Large Language Models for Enhanced Answer Quality and Verification, NAACL 2024.
- [5] NoMIRACL: Knowing When You Don't Know for Robust Multilingual Retrieval-Augmented Generation, arXiv preprint.

2023

- [6] MIRACL: A Multilingual Retrieval Dataset Covering 18 Diverse Languages, TACL 2023.
- [7] CIRAL at FIRE 2023: Cross-Lingual Information Retrieval for African Languages, FIRE'23.
- [8] Simple Yet Effective Neural Ranking and Reranking Baselines for Cross-Lingual Information Retrieval, arXiv preprint.
- [9] Evaluating Embedding APIs for Information Retrieval, arXiv preprint.
- [10] The state of OAI-PMH repositories in Canadian Universities, DCMI 2023.

2022

- [11] Making a MIRACL: Multilingual Information Retrieval Across a Continuum of Languages, arXiv preprint.
- [12] Refining an Almost Clean Translation Memory Helps Machine Translation, AMTA 2022.

2021

- [13] NATURE: Natural Auxiliary Text Utterances for Realistic Spoken Language Evaluation, NeurIPS 2021 Dataset Track.

2020

- [14] Human or Neural Translation?, ICCL 2020.

2019

- [15] Automatically learning a human-resource ontology from professional social-network data, Canadian AI 2019.

PERSONAL PROJECTS

- Developing a game in GDScript using Godot
- Developing a game in C# using Unity
- Digitally adapt the board game Samurai & Katana for Tabletop Simulator
- Created a daily news summarizer using LLMs
- Designed a real estate scraper
- Developing a bird song recognizer
- Implemented Instagram bot
- Contributed to open-source semantic mapping (IEML)
- Learned to use graphic design tools