https://spacemanidol.com/ — Google Scholar **Daniel Campos**
daniel.campos@zipf.ai

EDUCATION

University of Illinois Urbana-Champaign(UIUC)- *PhD Computer Science* 2020-2023
University of Washington- *MS Computational Linguistics* 2018-2020
Rensselaer Polytechnic Institute- *BS Computer Science* 2011-2014

EXPERIENCE

Senior Research Scientist - Snowflake 05/23

- Leading a team of 5 researchers in building out retrieval, extraction, and generation for Snowflake Cortex Search. Led to the creation of Snowflakes embedding models, outperforming competitors such as Open AI and cohere with lower inference costs.
- Lead partnerships and collaborations for search, resulting in partnerships with NVIDIA and Voyage for embedding models.
- Started, recruited, and scaled Snowflake's research internships to 12 PhD Research interns and developed the Snowflake AI fellowship with the University of Waterloo.
- Enhancing language model capabilities as co-pilots for data analysts via retrieval augmentations and filtering to deliver conversation SQL generation, completion, and query optimization.
- Managed pricing, evaluation, and scalability of GPU usage across cloud platforms.

Senior Research Scientist - Neeva (Acquired by Snowflake) 12/22-05/23

- Training, compressing, and deploying summarization, retrieval, and comprehension LLMs for search.
- Compressed and optimized web summarization models using pruning, knowledge distillation, faster transformers, and pseudo labeling, bringing latency down from 7 seconds to 300ms on a batch of 10.
- Led a comprehensive transformation of the search stack, designing a multi-phase retrieval pipeline from the ground up. Implemented state-of-the-art modeling approaches, including multi-vector retrieval, layout-aware representations, and model and index cascading.

Research Scientist Consultant - Neural Magic 10/2020-12/22

- Developed novel compression methods for NLP systems inference around Question Answering, Summarization, and Information Retrieval using PyTorch, HuggingFace, and SparseML and leveraging quantization, structured pruning, unstructured pruning, and knowledge distillation.
- Compressed models, when combined with the DeepSparse inference engine, resulted in over 30x inference throughput on CPU.
- Compressed bi-encoder Dense Retrieval Model to deliver 4x speedup with no loss in accuracy.
- Developed method of transferring sparse model performance to domains like BioMedical with no additional hyperparameter tuning.

Applied Scientist Consultant- Walmart Labs 06/2022-12/2022

- Developed **C**onstrastive **A**lignment **Po**st **T**raining (CAPOT), a method for improving retrieval on noisy queries without index regeneration or retraining. CAPOT improved retrieval accuracy on noisy queries by 55 % while requiring less than an hour on a single GPU.
- Developed **K**ullback–Leibler **Al**ignment of **E**mbeddings (KALE), a post-training bi-encoder compression approach which leverages asymmetrical bi-encoders to deliver 4x the throughput with minimal losses in accuracy.

Applied Scientist Consultant- Qualtrics 03/2022-06/2022

- Scaled sentiment analysis, action ability, and emotion detection models to 12 languages and leveraged per task Knowledge Distillation and Quantization to decrease the model size by 90%, saving 44% on inference cost.

Research Scientist Consultant- Mendel AI 10/2021-03/2022

- Redesigned, built, and deployed Oncology Relation and Entity Extraction pipelines using Transformers and custom language models, delivering a 35% improvement in F1 over existing LSTM models with 10x faster inference

Senior AI Product Manager-Microsoft Research & AI, Bing 11/2017-10/2020

- Architected relevance metrics stack into a real-time streaming system through new human labeling tasks, novel data sampling methods, and metrics aggregation. Labeling scaled to $7m/year across 16 markets, generating more than 1,000,000 weekly judgments. The pipeline went from measuring user experience years old to seconds old. To scale the pipeline, created a novel multi-class-stratified sampling method using the Horvitz-Thompson estimator, work under review.
- Built family of MSMARCO Datasets and baselines including QnA, Passage Ranking, and Keyphrase Extraction, which collectively have been used by over 10000 Researchers. Papers Cited over 2000 times and competitive datasets with over 500 research submissions. Dataset creation using Pandas and baselines models built in PyTorch and TensorFlow.
- Research and analyzed search relevance and user experience for the executive team across various verticals, languages, and metrics using language embeddings, sampling methodologies, and classification methods. Analysis and visualization were done through tSNE, Pandas, Numpy, ANNOY, and Plotly.

Product Manager 2-Microsoft Cloud + AI, Global Services 05/2016-11/2017

- Designed, built, and deployed Neural Machine Translation(NMT) into a software localization pipeline for all Cloud and Enterprise Microsoft products. Using the NMT models, supplier contracts were renegotiated, saving 20% $3m/year over the next three years.
- Optimized and scaled the NLP system for customer feedback translation, filtering, and categorization, decreasing team effort by 80% using NLTK, Tensorflow, and sci-kit-learn.
- Drove end-to-end localization and internationalization for Azure and Visual Studio products, including automation, bug triage, budgeting, vendor, and release management for $4m/year worth of translations.

Product Manager-Microsoft, Azure RemoteApp Summer 2014 & 08/2015-05/2015

- Designed scalable multi-cloud network architecture for cloud-based Remote Desktop Service and drove large-scale deployment to alpha customers and Microsoft.
- Created and scaled a predictive user demand modeling pipeline that optimized VM usage by 20%.

NLP Research Scientist-Basis Technology 01/2014-06/2014

- Grew the Rosette Entity Resolution pipeline to decrease document analysis time, increase accuracy, and increase throughput while expanding language support to Spanish, Russian, French, Farsi, and Chinese.
- Re-designed the system's machine-learned architecture to state-of-the-art neural methods, which improved multilingual accuracy to English parity.
- Formalized data creation, data ingestion, retrieval, formatting, and analysis pipeline using Bash, Java, and Python primarily. Optimized models were ANN, KNN, SVMs, LSTMs, and Decision Trees.

Software Engineer-Cisco Systems Summer 2013

- Developed JavaScript tooling for collecting user feedback, bug tracking, and HTML5-based screenshots used with alpha customers during initial product launches.
- Led a team of 12 other interns to develop a novel food ordering system for cafeterias that used containerized Node.js and MongoDB applications, allowing efficient scaling.

Co-Founder/Software Engineer- Gapelia 08/2013-05/2014

- Built an end-to-end digital multi-modal publishing platform that allowed non-technical users to create and publish beautiful digital magazines in minutes.
- Created REST API in Java, front end in Express, and deployed in AWS with Apache Tomcat, Maven, and MongoDB.
- Won the RPI 2014 Business Model Competition 1st place, 2014 Why not Change the World grant, Harvard iLab cultural entrepreneurship challenge, and was a member of Harvard's Innovation Lab 01/2014-05/2014.

Software Engineer-GE Capital Summer 2012

- Developed a reusable web system in Bootstrap, Javascript, and PHP, allowing GE employees to create websites that matched corporate guidelines around colors, fonts, icons, etc.

Publications

2024

- Synthetic Test Collections for Retrieval Evaluation - Hossein A. Rahmani, Nick Craswell, Emine Yilmaz, Bhaskar Mitra, **Daniel Campos** - In Proceedings of the 47rd International ACM SIGIR Conference on Research and Development in Information Retrieval
- Arctic-Embed: Scalable, Efficient, and Accurate Text Embedding Models - Luke Merrick, Danmei Xu, Gaurav Nuti, **Daniel Campos** - Arxiv Preprint

2023

- EFFICIENT AND ROBUST WEB SCALE LANGUAGE MODEL BASED RETRIEVAL, GENERATION, AND UNDERSTANDING - University of Illinois Urbana-Champaign Computer Science Doctoral Thesis
- Overview of the TREC 2023 Product Product Search Track - **Daniel Campos**, Surya Kallumadi, Corby Rosset, Cheng Xiang Zhai, Alessandro Magnani - TREC 2023
- Noise-Robust Dense Retrieval via Contrastive Alignment Post Training - **Daniel Campos**, ChengXiang Zhai, Alessandro Magnani - Arxiv Preprint
- Overview of the TREC 2022 deep learning track - Nick Craswell, Bhaskar Mitra, Emine Yilmaz, **Daniel Campos**, Jimmy Lin, Ellen M. Voorhees, and Ian Soboroff - TREC 2022
- To Asymmetry and Beyond: Structured Pruning of Sequence to Sequence Models for Improved Inference Efficiency **Daniel Campos**, ChengXiang Zhai - SustaiNLP 2023 Fourth Workshop on Simple and Efficient Natural Language Processing at ACL 2023

- Quick Dense Retrievers Consume KALE: Post Training Kullback Leibler Alignment of Embeddings for Asymmetrical dual encoders - **Daniel Campos**, ChengXiang Zhai, Alessandro Magnani - SustaiNLP 2023 Fourth Workshop on Simple and Efficient Natural Language Processing at ACL 2023
- Dense Sparse Retrieval: Using Sparse Language Models for Inference Efficient Dense Retrieval - **Daniel Campos**, ChengXiang Zhai - Arxiv Preprint
- oBERTa: Improving Sparse Transfer Learning via improved initialization, distillation, and pruning regimes - **Daniel Campos**, Alexandre Marques, Tuan Nguyen, Mark Kurtz, ChengXiang Zhai - SustaiNLP 2023 Fourth Workshop on Simple and Efficient Natural Language Processing at ACL 2023
- Compressing Cross-Lingual Multi-task Models at Qualtrics - **Daniel Campos**, Daniel Perry, Samir Joshi, Yashmeet Gambhir, Wei Du, Zhengzheng Xing, and Aaron Colak - The Thirty-Fifth Annual Conference on Innovative Applications of Artificial Intelligence (IAAI-23)

2022

- The Optimal BERT Surgeon: Scalable and Accurate Second-Order Pruning for Large Language Models - Eldar Kurtic, **Daniel Campos**, Tuan Nguyen, Elias Frantar, Mark Kurtz, Benjamin Fineran, Michael Goin, Dan Alistarh - EMNLP 2022
- Compressing Cross-Lingual Multi-task Models at Qualtrics - **Daniel Campos**, Daniel Perry, Samir Joshi, Yashmeet Gambhir, Wei Du, Zhengzheng Xing, and Aaron Colak - Arxiv Preprint
- Sparse*BERT: Sparse Models are Robust - **Daniel Campos**, Alexandre Marques, Tuan Nguyen, Mark Kurtz, ChengXiang Zhai - Sparsity in Neural Networks Workshop at ICML 2022
- Fostering Coopetition While Plugging Leaks: The Design and Implementation of the MS MARCO Leaderboards - Jimmy Lin, **Daniel Campos**, Nick Craswell, Bhaskar Mitra, Emine Yilmaz - SIGIR 2022
- The Optimal BERT Surgeon: Scalable and Accurate Second-Order Pruning for Large Language Models - Eldar Kurtic, **Daniel Campos**, Tuan Nguyen, Elias Frantar, Mark Kurtz, Benjamin Fineran, Michael Goin, Dan Alistarh - Arxiv Preprint

2021

- Overview of the TREC 2021 deep learning track - Nick Craswell, Bhaskar Mitra, Emine Yilmaz, **Daniel Campos** - TREC 2021
- IMG2SMI: Translating Molecular Structure Images to Simplified Molecular-input Line-entry System - **Daniel Campos**, Heng Ji - Arxiv Preprint
- Curriculum learning for language modeling - **Daniel Campos** - Arxiv Preprint
- TREC Deep Learning Track: Reusable Test Collections in the Large Data Regime - Nick Craswell, Bhaskar Mitra, Emine Yilmaz, **Daniel Campos**, Ellen Voorhees and Ian Soboroff - SIGIR 2021
- MS MARCO: Benchmarking Ranking Models in the Large-Data Regime - Nick Craswell, Bhaskar Mitra, Emine Yilmaz, **Daniel Campos**, Jimmy Lin - SIGIR 2021

- Significant Improvements over the State of the Art? A Case Study of the MS MARCO Document Ranking Leaderboard - Jimmy Lin, **Daniel Campos**, Nick Craswell, Bhaskar Mitra, Emine Yilmaz - SIGIR 2021
- Overview of the TREC 2020 deep learning track - Nick Craswell, Bhaskar Mitra, Emine Yilmaz, **Daniel Campos** - TREC 2020

2020

- Explorations In Curriculum Learning Methods For Training Language Models - University of Washington Computational Linguistics Master's Thesis
- GAIA at SM-KBP 2020 - A Dockerized Multi-media Multi-lingual Knowledge Extraction, Clustering, Temporal Tracking and Hypothesis Generation System - Manling Li, Ying Lin, Tuan Manh Lai, Xiaoman Pan, Haoyang Wen, Sha Li, Zhenhailong Wang, Pengfei Yu, Lifu Huang, Di Lu, Qingyun Wang, Haoran Zhang, Qi Zeng, Chi Han, Zixuan Zhang, Yujia Qin, Xiaodan Hu, Nikolaus Parulian, **Daniel Campos**, Heng Ji, Brian Chen, Xudong Lin, Alireza Zareian, Amith Ananthram, Emily Allaway, Shih-Fu Chang, Kathleen McKeown, Yixiang Yao, Michael Spector, Mitchell DeHaven, Daniel Napierski, Marjorie Freedman, Pedro Szekely, Haidong Zhu, Ram Nevatia, Yang Bai, Yifan Wang, Ali Sadeghian, Haodi Ma, Daisy Zhe Wang - Proceedings of Thirteenth Text Analysis Conference 2020
- ORCAS: 18 Million Clicked Query-Document Pairs for Analyzing Search - Nick Craswell, **Daniel Campos**, Bhaskar Mitra, Emine Yilmaz, Bodo Billerbeck- CIKM 2020
- XGLUE: A New Benchmark Dataset for Cross-lingual Pre-training, Understanding and Generation - Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, **Daniel Campos**, Rangan Majumder, Ming Zhou - EMNLP 2020
- On the Reliability of Test Collections to Evaluating Systems of Different Types - Emine Yilmaz, Nick Craswell, Bhaskar Mitra and **Daniel Campos** - SIGIR
- Overview of the TREC 2019 deep learning track Nick Craswell, Bhaskar Mitra, Emine Yilmaz, **Daniel Campos**, Ellen M. Voorhees, - TREC 2019
- Leading Conversational Search by Suggesting Useful Questions Corbin Rosset, Chenyan Xiong, Xia Song, **Daniel Campos**, Nick Craswell, Saurabh Tiwary and Paul Bennett - WWW 2020

2019

- Open Domain Web Keyphrase Extraction Beyond Language Modeling - Lee Xiong, Chuan Hu, Chenyan Xiong, **Daniel Campos**, Arnold Overwijk and Xiayu Huang - EMNLP 2019
- Overview of the TREC 2019 deep learning track Nick Craswell, Bhaskar Mitra, Emine Yilmaz, **Daniel Campos**, Ellen M. Voorhees, - TREC 2019

2018

- MS MARCO: A Human Generated MAchine Reading COmprehension Dataset Payal Bajaj, **Daniel Campos**, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, Tong Wang

Awards and Research Activities

Fellowships

- Ripple X Fellow Summer 2022
- Z Fellow Jan 2022
- Gene Golub FELLOWSHIP at UIUC 2020-2021
- Summer Predoctoral Institute Fellow at UIUC 2020

Patents

- ENHANCED SEARCHING USING FINE-TUNED MACHINE LEARNING MODELS - Filed 03/21/2024
- ENHANCED SEARCH RESULT GENERATION USING MULTI-DOCUMENT SUMMARIZATION - Filed 03/21/2024
- Using a Multi-Task-Trained Neural Network to Guide Interaction with a Query-Processing System via Useful Suggestions- 408364-US-NP - Filed 4/16/2020 - Granted 2023-12-26
- Keyphrase Extraction Beyond Language Modeling - U.S. Appln. No. 16/460,853 - Filed July 2nd, 2019 - Granted 2023-05-23

Activity and Invited Talks/Lectures

- Invited Talk: Benchmarking End to End Product Retrieval at The 2023 SIGIR Workshop On eCommerce
- Invited Talk: Making LLM Inference Affordable at the 2023 LLMs in Production Conference II
- Invited Lecture on Unstructured Pruning at UT Austin's VITA
- Guest Lecture on Conversational Information Retrieval at UIUC's Advanced Methods of Information Retrieval
- Guest Lecture on Neural Information Retrieval at UIUC's Advanced Methods of Information Retrieval
- Invited Panelist on Cost Optimization and Performance at the 2023 LLMs in Production Conference
- CS 510: Advanced Information Retrieval-Fall 2022 Teaching Assistant
- NIST TREC 2023 Product Search-Principal Track Coordinator
- NIST TREC 2023 Deep Learning-Track Coordinator
- CS 410: Information Retrieval-Fall 2022 Teaching Assistant
- CS 124: Introduction to Computer Science-Spring 2022 Teaching Assistant
- NIST TREC 2022 Deep Learning-Track Coordinator

- NIST TREC 2021 Deep Learning-Track Coordinator
- NIST TREC 2020 Deep Learning-Track Coordinator
- NIST TREC 2019 Deep Learning-Track Coordinator
- Scaling Language Model Inference to Web-Scale Workloads - Invited Talk at Neeva - Aug 2022
- Efficient Language Model Inference - Invited Talk at Walmart Tech - Jan 2022
- Efficient Language Model Training and Inference - Invited Talk at Qualtrics Research - Nov 2021
- Efficient Language Model Training and Inference - Invited Talk at You.com - Oct 2021
- ACM SIGIR/SIGKDD Africa Summer School on Machine Learning for Data Mining and Search 2019-Invited Lecturer for Deep Learning in Search
- AACM SIGIR/SIGKDD Africa Summer School on Machine Learning for Data Mining and Search 2020-Invited Lecturer for Deep Learning in Search