Omar Khattab

✉ *okhattab@mit.edu*
🖱 *omarkhattab.com*

Education

Sep 2019 – Jan 2025 **Stanford University**, Stanford, CA
Ph.D. in Computer Science, Advised by Christopher Potts & Matei Zaharia
Thesis: *Building more reliable and scalable AI systems with foundation model programming*

Aug 2015 – May 2019 **Carnegie Mellon University** in Qatar (CMU-Q), Doha, Qatar
B.S. in Computer Science (Minor: Math), Advised by *Mohammad Hammoud*

Employment

Jul 2025 – **Assistant Professor at MIT EECS.**
Affiliated with CSAIL and EECS AI+D & CS.

Jun 2024 – Jun 2025 **Research Scientist at Databricks.**
Working on NLP, IR, and ML Systems.

Jun–Sep 2022 **Research Ph.D. Intern at Apple** via a "PhD Scholar in AI/ML" fellowship.
Advised by *Dr. Ruoming Pang, Distinguished Engineer at Apple AI/ML.*

Fellowships, Awards & Honors

2025 SIGIR 2025 "Best Paper Award" for WARP (last author).

2025 "2025 Tartans on the Rise" Honoree, Carnegie Mellon.

2022–24 "Apple PhD Scholar in AI/ML" Fellowship.

2022–23 HAI–Azure Cloud Credits Grant ($50,000) for "Reliable Generation with Retrieval-Based Models" with David Hall. Lead PI: Chris Potts.

2021–24 Industry academic research grants (over $300,000) from research teams at IBM & Oracle. PIs: Chris Potts & Matei Zaharia.

2021 "Top-5% of Stanford CS Course Assistants": Stanford CS244u Spring 2021.

2021 Nominated by Stanford CS (one of four) to apply for the Microsoft PhD fellowship.

2019–20 The Eltoukhy Family Graduate Fellowship. Stanford CS Department.

2019 "Alumni Award for Undergraduate Excellence" for Senior Thesis. CMU (Pittsburgh).

2019 "Outstanding Academic Achievement in Computer Science". CMU-Qatar.

2019 "Best Project", Annual Meeting of the Minds". CMU-Qatar.

2018 "Qatar Campus Scholar". CMU-Qatar.

2016–19 Full-Tuition Merit Scholarship. Qatar Foundation.

2016–19 Dean's List Awards. CMU-Qatar.

2015 "Top in the World" Outstanding Learner. Cambridge International Exams: A.S. Computing.

Papers & Manuscripts

∗ indicates co-first author

SIGIR 2025 WARP: An Efficient Engine for Multi-Vector Retrieval (Best Paper Award).
JL Scheerer, M Zaharia, C Potts, G Alonso, **O Khattab**.

ReNeuIR 2025 Drowning in Documents: Consequences of Scaling Reranker Inference
M Jacob, E Lindgren, M Zaharia, M Carbin, **O Khattab**, A Drozdov.

ICLR 2025 Grounding by trying: LLMs with reinforcement learning-enhanced retrieval.
S Hsu, **O Khattab**, C Finn, A Sharma.

NAACL 2025 PAPILLON: PrivAcy Preservation from Internet-based and Local Language MOdel ENsembles.
L Siyan, VC Raghuram, **O Khattab**, J Hirschberg, Z Yu.

ArXiv 2024 Drowning in Documents: Consequences of Scaling Reranker Inference.
M Jacob, E Lindgren, M Zaharia, M Carbin, **O Khattab**, A Drozdov. arXiv preprint arXiv:2411.11767.

ArXiv 2024 Prompts as Auto-Optimized Training Hyperparameters: Training Best-in-Class IR Models from Scratch with 10 Gold Labels.
J Saad-Falcon, **O Khattab**, K Santhanam, R Florian, M Franz, S Roukos, A Sil, M Sultan, C Potts.

EMNLP 2024 Fine-tuning and prompt optimization: Two great steps that work better together.
D Soylu, C Potts, **O Khattab**.

EMNLP 2024 Optimizing instructions and demonstrations for multi-stage language model programs.
K Opsahl-Ong, M Ryan, J Purtell, D Broman, C Potts, M Zaharia, **O Khattab**.

EMNLP 2024 Findings Problem-Oriented Segmentation and Retrieval: Case Study on Tutoring Conversations.
R E Wang, P Wirawarn, K Lam, **O Khattab**, D Demszky.

NAACL 2024 Assisting in Writing Wikipedia-like Articles From Scratch with Large Language Models.
Y Shao, Y Jiang, T Kanell, P Xu, **O Khattab**, M Lam.

NAACL 2024 ARES: An Automated Evaluation Framework for Retrieval-Augmented Generation Systems.
J Saad-Falcon, **O Khattab**, C Potts, M Zaharia.

EACL 2024 Findings Backtracing: Retrieving the Cause of the Query.
R Wang, P Wirawarn, **O Khattab**, N Goodman, D Demszky.

ICLR 2024 DSPy: Compiling Declarative Language Model Calls into Self-Improving Pipelines.
O Khattab, A Singhvi, P Maheshwari, Z Zhang, K Santhanam, S Vardhamanan, S Haq, A Sharma, T Joshi, H Moazam, H Miller, M Zaharia, C Potts. Spotlight paper: top 5% of submissions.

Digital Discovery 2024 Image and Data Mining in Reticular Chemistry Using GPT-4V.
Z Zheng, Z He, **O Khattab**, N Rampal, M Zaharia, C Borgs, J T Chayes, O M Yaghi.

ArXiv 2024 In-Context Learning for Extreme Multi-Label Classification.
K D'Oosterlinck, **O Khattab**, F Remy, T Demeester, C Develder, C Potts.

KB Systems 2024 Building Efficient and Effective OpenQA Systems for Low-Resource Languages.
E Budur, R Özçelik, D Soylu, **O Khattab**, T Güngör, C Potts.

ArXiv 2023 DSPy Assertions: Computational Constraints for Self-Refining Language Model Pipelines.
A Singhvi, M Shetty, S Tan, C Potts, K Sen, M Zaharia, **O Khattab**.

EMNLP 2023 UDAPDR: Unsupervised Domain Adaptation via LLM Prompting and Distillation of Rerankers.
J Saad-Falcon, **O Khattab**, K Santhanam, R Florian, M Franz, S Roukos, A Sil, M Sultan, C Potts.

REML 2023 Resources and Evaluations for Multi-Distribution Dense Information Retrieval.
S Chatterjee, **O Khattab**, S Arora.

ACL 2023 Findings Moving Beyond Downstream Task Accuracy for Information Retrieval Benchmarking.
K Santhanam, J Saad-Falcon, M Franz, **O Khattab**, A Sil, R Florian, S Roukos, A Sil, M Sultan, M Zaharia, C Potts.

TMLR 2023 Holistic evaluation of language models.
P Liang, R Bommasani, T Lee, D Tsipras, D Soylu, ..., **O Khattab**, ..., Y Zhang, Y Koreeda. Transactions on Machine Learning Research (TMLR) 2023. This is a multi-component, 50-author project. I directed the Information Retrieval evaluation.

ArXiv 2022 Demonstrate-Search-Predict: Composing retrieval and language models for knowledge-intensive NLP.
O Khattab, K Santhanam, XL Li, D Hall, P Liang, C Potts, M Zaharia. ArXiv preprint arXiv:2212.14024.

CIKM 2022 PLAID: An Efficient Engine for Late Interaction Retrieval.
K Santhanam*, **O Khattab***, C Potts, M Zaharia. Proceedings of the 31st ACM International Conference on Information and Knowledge Management (CIKM).

CIKM 2022 Introducing Neural Bag of Whole-Words with ColBERTer: Contextualized Late Interactions using Enhanced Reduction.
S Hofstätter, **O Khattab**, S Althammer, M Sertkan, A Hanbury. Proceedings of the 31st ACM International Conference on Information and Knowledge Management (CIKM).

NAACL 2022 ColBERTv2: Effective and Efficient Retrieval via Lightweight Late Interaction.

K Santhanam*, **O Khattab***, J Saad-Falcon, C Potts, M Zaharia. Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL).

ICLR 2022 Hindsight: Posterior-guided training of retrievers for improved open-ended generation.

A Paranjape, **O Khattab**, C Potts, M Zaharia, CD Manning. Proceedings of The Tenth International Conference on Learning Representations (ICLR).

NeurIPS 2021 Baleen: Robust Multi-Hop Reasoning at Scale via Condensed Retrieval.

O Khattab, C Potts, M Zaharia. Thirty-Fifth Conference on Neural Information Processing Systems (NeurIPS'21). Spotlight paper: top 3% of submissions.

ArXiv 2021 On the opportunities and risks of foundation models.

R Bommasani, DA Hudson, E Adeli, R Altman, S Arora, S von Arx, ..., **Omar Khattab**, ..., P Liang. ArXiv preprint arXiv:2108.07258. This is a multi-component, 114-author project. I co-authored the sections on Modeling and on Systems.

SIGIR 2021 Learning passage impacts for inverted indexes.

A Mallia, **O Khattab**, T Suel, N Tonellotto. Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. Short paper.

TACL 2021 Relevance-guided Supervision for OpenQA with ColBERT.

O Khattab, C Potts, M Zaharia.

SIGIR 2020 ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT.

O Khattab, M Zaharia. Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval.

SIGIR 2020 Finding the best of both worlds: Faster and more robust top-k document retrieval.

O Khattab, M Hammoud, T Elsayed. Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval.

CIKM 2018 PolyHJ: A polymorphic main-memory hash join paradigm for multi-core machines.

O Khattab, M Hammoud, O Shekfeh. Proceedings of the 27th ACM International Conference on Information and Knowledge Management.

VLDB 2018 LA3: A scalable link-and locality-aware linear algebra-based graph analytics system.

Y Ahmad, **O Khattab**, A Malik, A Musleh, M Hammoud, M Kutlu, M Shehata, T Elsayed. Proceedings of the VLDB Endowment 11 (8), 920-933.

Technical Blog Posts

SAIL 2021 Building Scalable, Explainable, and Adaptive NLP Models with Retrieval.
O Khattab, M Zaharia, C Potts. Stanford AI Lab (SAIL) blog.

HAI 2021 A moderate proposal for radically better AI-powered Web search.
O Khattab, M Zaharia, C Potts. Stanford HAI blog.

Invited Talks, Keynotes & Conference Talks

Talk Title	**Building More Reliable and Scalable AI Systems with Language Model Programming**
	— *delivered at* —
2024 Apr	**University of Waterloo.**
2024 Apr	**Purdue University.**
2024 Apr	**Georgia Tech.**
2024 Mar	**Carnegie Mellon in Qatar.**
2024 Mar	**MIT.**
2024 Mar	**Caltech.**
2024 Mar	**University of Maryland.**
2024 Feb	**Yale.**
2024 Jan	**Carnegie Mellon.**

Talk Title	**"DSPy: Compiling Declarative Language Model Calls into Self-Improving Pipelines."**
	— *delivered at* —
2023 Nov	**ScaleByTheBay.**
2023 Oct	**UC Berkeley**. Guest Lecture.
2023 Sep	**Google X.**
2023 Sep	**Apple.**
2023 Aug	**Qatar University** (QU). BigIR group.
2023 Aug	**Avey**. Doha, Qatar.
2023 July	**SIGIR REML**. Invited talk at the REML workshop.

Talk Title	**"Demonstrate–Search–Predict: Composing retrieval and language models for knowledge-intensive NLP."**
	— *delivered at* —
2023 April	**UC Berkeley.** Research talk.
2023 April	**Databricks.**
2023 April	**Meta.**
2023 March	**Neeva.**
2023 March	**Netflix ML.**
2023 March	**Vectara.**
2023 March	**Google.**
2023 Jan	**Oracle Labs.**
2023 Jan	**Apple.**

Talk Title **“ColBERTv2: Effective and Efficient Retrieval via Lightweight Late Interaction.”**

— *delivered at* —

2023 Nov **Stanford.** Guest Lecture.

2023 Sep **UC Berkeley.** Guest Lecture.

2023 Aug **Carnegie Mellon University Qatar** (CMU-Qatar).

2023 July **SIGIR ReNeuIR**. *Keynote.*

2022 Aug **Apple.**

2022 July **NAACL.** Conference Talk on ColBERTv2.

2022 May **Etsy.** Retrieval Research Group.

Talk Title **“Baleen: Robust Multi-Hop Reasoning at Scale via Condensed Retrieval.”**

— *delivered at* —

2021 Dec **NeurIPS.** Conference talk.

2021 Feb **Stanford NLP.**

Talk Title **ColBERT & “Relevance-guided Supervision for OpenQA with ColBERT.”**

— *delivered at* —

2021 Oct **Google Research.** IR Group.

2021 Aug **ACL.** Conference talk.

2021 Jun **Google Research.** N2Formal RG.

2021 Mar **IBM Research.** NLP RG.

2020 Nov **Univ. of Glasgow.**

Talk Title **“ColBERT: Efficient and Effective Search via Contextualized Late Interaction over BERT.”**

— *delivered at* —

2020 Jul **SIGIR.** Conference talk.

2020 Jun **DAWN Retreat.**

2020 Jul **“Finding the Best of Two Worlds: Faster and More Robust Top-k Document Retrieval.”** Conference talk at **SIGIR**.

2019 Apr **“IRg: A Distributed Graph-based Framework for Information Retrieval.”** Senior Thesis Public Talk at **CMU-Qatar**.

2018 Oct **“PolyHJ: A Polymorphic Main-Memory Hash Join Paradigm for Multi-Core Machines.”** Conference talk at **CIKM**.

2018 Aug **“LA3: A Scalable Link- and Locality-Aware Linear Algebra-Based Graph Analytics System.”** Conference talk at **VLDB**.

Panels, Podcast Episodes, & Tutorials

2023 Oct	**UC Berkeley SkyCamp 2023**. Tutorial.
2023 Oct	**MLOps Community Podcast**.
2023 Oct	**DemandBase**. Webinar.
2023 Oct	**MLOps Learners.** Webinar.
2023 Oct	**LlamaIndex Webinar.**
2023 July	**SIGIR GenIR.** Panel on LM behavior.
2023 June	**Vertex.** Podcast episode with hosts Sandeep Bhadra & Chase Roberts.
2023 June	**Cohere4AI.** Webinar.
2023 June	**LlamaIndex Webinar.**
2023 May	**LangChain Panel.**
2023 May	**DataBrew.** Podcast episode.
2023 Jan	**Samaya.** Webinar.
2022 May	**Stanford CS224U Guest Podcasts.** Podcast episode with host Prof. Christopher Potts.

Teaching Assistant Experience

Spring 2022	Stanford CS224u: Natural Language Understanding (Prof. Christopher Potts)
Spring 2021	Stanford CS224u: Natural Language Understanding (Prof. Christopher Potts)
Spring 2019	CMU 15-210: Parallel & Sequential Data Structures & Algorithms (Prof. Kemal Oflazer)
Fall 2018	CMU 11-785: Introduction to Deep Learning (Prof. Bhiksha Raj)
Fall 2018	CMU 15-451: Algorithms Design & Analysis (Prof. Christos Kapoutsis)
Spring 2018	CMU 15-210: Parallel & Sequential Data Structures & Algorithms (Prof. Kemal Oflazer)

Service

Review	2024: ACL ARR (Oct) Senior Area Chair, ICLR, ICML, and other venues
	2023: NeurIPS, ICLR, ACL ARR, and other venues
	2022: TOIS, IP&M (Top IR journals)
	2021: TOIS, IP&M (Top IR journals)
Committee	Jan 2023: Stanford CS PhD admissions in NLP
	Jan 2022: CSLI Summer Internships
	Jan 2021: CSLI Summer Internships
	Nov 2020: Student-Applicant Support Program