

Kaustubh D. Dhole

Emory University | BITS Pilani | Ex R&D Lead, Amelia.ai
+1-404-493-8690 | kdhole@emory.edu | [Publications](#) | [Website](#)

RESEARCH INTERESTS

My long-term research goal is to build autonomous systems that can reason with evidence and operate as trustworthy collaborators in high-stakes decision-making. I focus on methods that let artificial intelligence (AI) systems (1) retrieve and organize the external knowledge they rely on intelligently and autonomously, (2) provide evaluations that are interpretable, holistic, and optimizable, (3) act as accountable and goal-aligned decision partners for language, clinical support, scientific discovery, and policy rhetoric.

Natural Language Processing | Information Retrieval | Machine Learning | Reinforcement Learning

RELEVANT WORK EXPERIENCE

- *PhD Researcher*, Computer Science, **Emory University**
Atlanta, Georgia
July 2021 to 2026

Adviser: [Prof. Eugene Agichtein](#)

Current Projects: RAG Evaluation, Post-Training for non-verifiable rewards, Ranking Pretraining datasets for Math & Code Reasoning, Scaling Laws

Internships: Amazon AGI and Alexa Stores Foundational AI (Summers of 22, 23, 24, and 25)

- *Research Team Lead*, [Amelia R&D](#)
New York, USA & Bangalore, India
Jan 2018 to June 2021: **3.5 Years**

Conversational Agent: We experimented and productionized Amelia Agent (versions 2, 3, 4), a conversational agent that generates realistic and task-oriented responses through a combination of multiple responders (handling [user ambiguities](#), [unanswerable queries](#), [QA generation](#)). The NLU component extracts relevant semantic representations, followed by a parallel set of generative (semantic memory and episodic memory) and heuristic (based on VerbNet & AMR representations) responders, and business process networks (BPNs)

Agentic AI Dashboard: Designed and built an agent administrative framework for gathering annotations, supporting agent building data from generative capabilities, and AI based workflow designs (business process automations), and designed UI interfaces that can provide improved AI-human interaction through assisting in annotating dialogue flows.

Other additional facets: Whispering Agent, Paraphrastic Fallback Agent, BPN Design Agent

Adviser & Consultant: [Prof. Christopher D. Manning](#)

- *R&D Engineer*, [Amelia R&D](#)
New York, USA & Bangalore, India
July 2015 to Dec 2017: **3 Years**

Tested Different Hypothesis and performed rigorous benchmarking for Agent Episodic Memory (BiLSTM, Hierarchical RNN, etc), Semantic Memory (Doc Parsing to RAG): Tree Kernel Approaches, Spreading Activation, Using Lexical Answer Types, Clarification Question Generation.

Adviser & Consultant: [Prof. Christopher D. Manning](#)

EDUCATION

1. PhD Candidate (July 2021 - Current) Computer Science, Emory University
2. MS (July 2021 - Completed) Computer Science, Emory University

3. BE (Hons) Electrical and Electronics Engineering + MSc (Hons) Biological Sciences (Dual Degree)
BITS, Pilani, India, (July 2010-Jun 2015)
(1 year Thesis: **Tata Institute of Fundamental Research**, 2014-15: Queuing Theory, ML for proteins)

INTERNSHIPS

Amazon Stores Foundational AI: Applied Scientist II - Seattle, Summer 2025

- Reasoning Data Valuation for Pretraining and Midtraining
 - How do we rank useful datasets for pretraining and mid-training data that would improve downstream reasoning tasks like coding and math (Collaborators: David Mueller, Colin Lockard, Amazon Alexa)
 - Stabilizing Reinforcement Learning for Honesty Alignment in Language Models on Deductive Reasoning (Collaborators: Jiarui Liu and others)

Amazon Stores Foundational AI: Applied Scientist II - Seattle, Summer 2024

- LLM Judges and LLM based Annotations for Rerankers for Superlative Queries
- (Collaborators: Shervin Malmasi, Amazon Alexa)

Amazon Alexa AGI: Applied Scientist II - New York, Summer 2023

- Trained LLM based user simulators (prompting based+fine-tuned) to improve downstream performance of LLM based assistants
- Assisted other interns in exploring curriculum training strategies, exploring task-wise memorisation, and prompting strategies for reducing hallucination (Collaborators: Shikib Mehri, Jack FitzGerald)

Amazon Alexa AGI (prev DeepNU): Applied Scientist II - New York, Summer 2022

- A Multi-Encoder Frozen-Decoder Approach for NLP Tasks: Devising appropriate task taxonomies for multi-task learning for large language modeling across a variety of tasks - summarisation, data-to-text generation, paraphrasing, semantic parsing, (Multilingual AlexaTM)
- Analyzed the effects of freezing parts of encoder-decoder models and compared their effects on fine-tuning (Collaborators: Jack FitzGerald)

PUBLICATIONS

- 42) ***RubricRAG: Towards Interpretable and Reliable LLM Evaluation via Domain Knowledge Retrieval for Rubric Generation*** Dhole and Agichtein, SIGIR 2026
- 41) ***HieraRAG: How Fine-grained Should a RAG Benchmark Be? A Hierarchical Framework for Synthetic Question Generation*** Fensore, Dhole, Fan, Agichtein, Ho SIGIR 2026
- 40) ***Stabilizing Reinforcement Learning for Honesty Alignment in Language Models on Deductive Reasoning Logical and Symbolic Reasoning in Language Models @ AAI 2026*** J Liu, K Dhole, Y Wang, H Wen, S Zhang, H Mao, G Li, N Varshney, J Liu, X Pan, 2025
- 39) ***Scaling Laws for Reranking in Information Retrieval*** Seetharaman, Bansal, Zamani, Dhole
- 38) ***AdvERSEM: Adversarial Robustness Testing and Training of LLM-based Groundedness Evaluators via Semantic Structure Manipulation*** The 14th Joint Conference on Lexical and Computational Semantics 2025 K Dhole, R Chandradevan, E Agichtein

- 37) *Evaluating Hybrid Retrieval Augmented Generation using Dynamic Test Sets*, LiveRAG Workshop, C Fensore, K Dhole, JC Ho, E Agichtein
- 36) *InsertRank: LLMs can Reason over BM25 scores to Improve Listwise Reranking*, GenAIRecP @ WSDM 2026 Rahul S, K D. Dhole, Aman B
- 35) *ConQRet: Benchmarking Fine-Grained Evaluation of Retrieval Augmented Argumentation with LLM Judges*, KD Dhole, K Shu, E Agichtein, 2024, **NAACL 2025** Long Paper
- 34) *Generative Product Recommendation for Implicit Superlative Queries*, KD Dhole, N Vedula, Saar K, Giuseppe G, Eugene A, Shervin M, **NAACL 2025** Student Workshop
- 33) *BabyReasoningBench: Generating Developmentally-Inspired Reasoning Tasks for Evaluating Baby Language Models*, K Dhole, 2025
- 32) *SHADES: Towards a Multilingual Assessment of Stereotypes in Large Language Models*, Margaret Mitchell, .. KD Dhole,.. Zeerak Talat, **NAACL 2025** Long Paper
- 33) *Humanity's Last Exam, 2025*
- 31) *To Retrieve or Not to Retrieve? Uncertainty Detection for Dynamic Retrieval Augmented Generation*, KD Dhole, 2024, Quantify Uncertainty and Hallucination in Foundation Models Workshop, **ICLR 2025**
- 30) *A Multi-Encoder Frozen-Decoder Approach for Fine-Tuning Large Language Models*, KD Dhole, 2024
- 29) *CoRAGen: A Comparative User Interface for Multi-Aspect Retrieval Augmented Generation*, Atharva, KD Dhole, [Demo](#)
- 28) *LLM Judges for Retrieval Augmented Argumentation*, K Dhole, E Agichtein, Task Focused IR in the Era of Generative AI, Microsoft Research, Redmond
- 27) *PyTerrier-GenRank: The PyTerrier Plugin for Reranking with Large Language Models*, K Dhole
- 26) *Chatting with Logs: An exploratory study on Finetuning LLMs for LogQL*, Vishwanath Seshagiri, Siddharth Balyan, Vaastav Anand, Kaustubh Dhole, Ishan Sharma, Avani Wildani, José Cambronero, Andreas Züfle, 2024
- 25) *DUQGen: Effective Unsupervised Domain Adaptation of Neural Rankers by Diversifying Synthetic Query Generation* Ramraj, Kaustubh D, Eugene A, **NAACL 2024**
- 24) *QueryExplorer: An Interactive Query Generation Assistant for Search and Exploration*, Kaustubh D, Shivam B, Ramraj C, Eugene A. **NAACL Demo 2024**
- 23) *Kaucus: Knowledge Augmented User Simulators for Training Language Model Assistants*, Kaustubh D, *Simulation of Conversational Intelligence in Chat*, **EACL 2024**
- 22) *GenQREnsemble : Zero-Shot LLM Ensemble Prompting for Generative Query Reformulation*, Kaustubh D, Eugene A. **ECIR 2024**

- 21) *Generative Query Reformulation Using Ensemble Prompting, Document Fusion, and Relevance Feedback*, Kaustubh D, Ramraj C, Eugene A.
- 20) *An Interactive Query Generation Assistant using LLM-based Prompt Modification and User Feedback* (Kaustubh D, Ramraj, Eugene A), Demo Day, IARPA BETTER (*Publications on US government website*)
- 19) *Large Language Models as SocioTechnical Systems* (Kaustubh Dhole), Accepted at *EMNLP 2023 Workshop BigPicture*
- 18) *Contextual Response Interpretation for Automated Structured Interviews: A Case Study in Market Research*, (S Harshita, D Kaustubh, P Ankur, V Venugopal, A Eugene) *WWW 2023*
- 17) *NusaCrowd: Open Source Initiative for Indonesian NLP Resources*, *ACL 2023*
- 16) *Lessons from Digital India for the Right to Internet Access*, *Situating Network Infrastructure with People, Practices, and Beyond*, *CSCW 2022*
- 15) *A Bird's-Eye Tutorial of Graph Attention Architectures*, Kaustubh D, Carl Yang 2022
- 14) *Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models, 2022 (Core Contributor) Transactions of Machine Learning, 2023*
- 13) *GEMv2: Multilingual NLG Benchmarking in a Single Line of Code*, *EMNLP 2022* (Sebastian Gehrmann .. Kaustubh D...)
- 12) *NL-Augmenter* 🌱 → 🐸 *A Framework for Task-Sensitive Natural Language Augmentation* (KD Dhole et al) *GitHub NEJLT, 2023*
- 11) *Automatic Construction of Evaluation Suites for Natural Language Generation Datasets* (S Mille, KD Dhole, S Mahamood, L Perez-Beltrachini, V Gangal, M Kale, S Gerhmann) *NeurIPS 2021*
- 10) 🕯️ *CANDLE: Decomposing Conditional and Conjunctive Queries for Task-Oriented Dialogue Systems* (Aadesh Gupta, Kaustubh D. Dhole, Rahul Tarway, Swetha Prabhakar, Ashish Shrivastava)
- 9) *The GEM Benchmark: Natural Language Generation, its Evaluation and Metrics* (Sebastian Gehrmann .. Kaustubh D...) *ACL Workshop 2021*
- 8) *Saying No is An Art: Contextualized Fallback Responses for Unanswerable Dialogue Queries* (A Shrivastava, KD Dhole , A Bhatt , S Raghunath) *ACL 2021*
- 7) *Resolving Intent Ambiguities in Task Oriented Dialogue Systems by Retrieving Discriminative Questions*
- 6) *Benchmarking BioRelEx for Entity Tagging and Relation Extraction* (Abhinav Bhatt, Kaustubh D. Dhole)
- 5) *Document categorization using semantic relatedness and Anaphora resolution: A discussion* (K Dhole, H Kohli) *Research in Computational Intelligence and Communication Networks, 2015*
- 4) *Differential regulation of cryptic genetic variation shapes the genetic interactome underlying complex traits* (A Yadav, K Dhole, H Sinha) *Genome Biology and Evolution, 2016*

- 3) i) *Unraveling the genetic architecture of cryptic genetic variation* (A Yadav, K Dhole, H Sinha) bioRxiv, 2016
 ii) *Genetic regulation of phenotypic plasticity and canalisation in yeast growth* (A Yadav, K Dhole, H Sinha) *PloS one*, 2016
- 2) *Sequence-based prediction of protein–protein interaction sites with L1- logreg classifier* (K Dhole, G Singh, PP Pai, S Mondal) *Journal of Theoretical Biology*, 2014
- 1) **SPRINGS: Prediction Of Protein-Protein Interaction Sites Using Neural Networks** (G Singh, K Dhole, PP Pai, S Monda) *Journal of Proteomics and Computational Biology*, 2014

RECENT WORKSHOPS CONDUCTED

- **Mentored** for the **LOGML workshop** (London Geometry & Machine Learning Summer School, 2022): Guided 5 MS/PhD students on Efficient variations of the transformer architecture and its implementation in [Graph Attention Networks](#)
- **Co-organizer, NL-Augmenter:** Conducted a large collaborative workshop on accumulating the largest set of natural language transformations (paraphrasers as well as task-specific augmenters). Received and reviewed >160 data augmentation techniques, evaluated them to check for robustness and merged >117 of them. [Organization](#), [Link to Paper](#)
- **Co-organized** the first four LLM Generation, Evaluation & Metrics Workshops: [GEM \(2021\)](#), [GEM \(2022\)](#), [GEM \(2023\)](#), [GEM \(2025\)](#)

MASTERS THESIS

- *Research Intern, Operations Research : Modeling Multiple Classes of customers and Random volumes of Arrivals with known probability distributions*,
 Dec 2014 to Jun 2015
[School of Technology and Computer Science, Tata Institute of Fundamental Research, Mumbai](#)
 Advisor: Prof. Sandeep Juneja
- *Research Intern, Gene Interactions and Biostatistics*
 May 2014 to Dec 2014
[Department of Biological Sciences, Tata Institute of Fundamental Research, Mumbai](#)
 Advisor: Prof. Himanshu Sinha

RECENT INVITED TALKS

- Department of Psychology, Stanford University by Prof. Michael C. Frank for a talk on Retrieval Augmented Generation and Evaluation (virtual)
- Was invited to Department of Medicine, Stanford University by Prof. Nigam Shah for a talk on Retrieval Augmented Generation and Evaluation
- Was invited to Department of Biomedical DataScience, Stanford University by Prof. Jason Fries for a talk on Retrieval Augmented Generation and Data Valuation
- Was invited to the University of Edinburgh by Prof. Mirelle Lapata for a talk on Retrieval Augmented Generation [[Video](#) of similar lecture at Emory University]
- Was invited to Vidyalkar Institute of Technology, Mumbai for giving a talk on the progress of Natural Language Processing, and designing a Machine Learning based curriculum for engineering undergraduates, 2017 and 2021 [[2021](#) Talk at ICAITR 2021]

RESEARCH LEADERSHIP & MENTORSHIP:

- Mentoring and/or directly managing several great individuals, particularly on Conversational AI projects at Amelia R&D. Most of these projects were as short as 1 month to as long as 1 year. Some are listed here, in no particular order:
 - R&D/Senior R&D Engineers (at Amelia.ai):
 - Krishna Mohan Barakam (Vice President at JPMorgan Chase & Co),
 - Manjunath Hegde (Senior Applied Scientist, Goldman Sachs)
 - Anurag Kashyap (Applied Scientist, Amazon SFAI)
 - Aadesh Gupta (Principal Data Scientist, Walmart),
 - Ashish Srivastava (Senior Data Scientist, Walmart),
 - Vishwa Teja Ramavtar Malav (Amazon Pricing)
 - Venkatesh Magham (Software Engineer, Meta)
 - Kaustav Dutta (Graduate Teaching Assistant at UC San Diego)
 - Priyank Soni (Senior Data Scientist, IBM)
 - Mohit Rohatgi (Scientist, Adobe)
 - Abhinav Bhatt (Software Engineer, SAP), Arpan Kulshreshtha, Roopesh Mangal,, Rohit Kalra
 - R&D Interns (at Amelia.ai): Bhargav Sagiraju, Chandra Reddy, Pranav Kamojjhala
- Recent (2025):
 - Atharv Negi (Emory University)
 - Chase Fensore (Emory University)
 - Jairui Liu (Carnegie Mellon University)
 - Rahul Seetharaman (UMass Amherst/MLE @ LinkedIn)
 - Aman Bansal (UMass Amherst/ Nutanix)
 - Edward Yuon (Emory University)

PROGRAMMING SKILLS:

Python, PyTorch, TensorFlow, Java (Spring, Hibernate), C

CERTIFICATIONS/WORKSHOPS ATTENDED

- DeepLearning.ai: Building Code Agents with Hugging Face smolagents, Reasoning with o1, Quantization Fundamentals with Hugging Face, Building and Evaluating Advanced RAG
- Neural Networks for Machine Learning (Geoffrey Hinton), Machine Learning (Andrew Ng), Deep Learning 5-course Specialization (Andrew Ng), Tensorflow for Deep Learning, Data Science, Financial Mathematics, Statistical Inference, Security Analysis and Portfolio Management, Financial Management, R Programming, Cloud Computing on Windows Azure (Certifications available on request),
- [Tutorial and Workshop on Learning and Related Applications](#), 2015, TIFR
- Informal: Knowledge Representation and Reasoning, NPTEL, IIT-Madras, Deep Learning for NLP, Deep Reinforcement Learning, Generative Adversarial Networks (Andrew Ng)
- Initiated and taught the [first course of Machine Learning](#) as a part of the Centre for Technical Development, BITS Pilani covering Supervised & Unsupervised Learning, Reinforcement Learning and Genetic Algorithms using Matlab assignments, 2014.
- Actively working as a part of the founding team for the project working on conducting the [Generation Evaluation and Metrics \(GEM\)](#) workshop at ACL, 2021 (Generation Evaluation Metrics) onwards.
- CS Ethics: Presented arguments on the [Right to Internet Access](#) at Situating Network Infrastructure with People, Practices, and Beyond Workshop, CSCW 2022.

AWARDS

Key Scientific Article for Excellence in Biomedical Research: *Article was featured on Global Medical Discovery [ISSN 1929-8536] as a Key Scientific Article contributing to excellence in biomedical research. Article: <http://dx.doi.org/10.1016/j.jtbi.2014.01.028>*

OTHER PROJECTS (Development, Research & Others)

RL based Instance Selection for CommonSense Choice Classification (Course Project of AI, 2021), Kaustubh D.

- Created an RL-based instance selector to choose high-quality paraphrases from NLAugmenter augmentations for a commonsense choice dataset.
- Trained and evaluated a downstream classifier using the selected instances.

Self Evaluation Courtroom of Language Models (Link)

- Designed a multi-agent “courtroom” setup with three LM advocates and a fourth LM as judge in a few-shot setting.
- Created four LM evaluation tasks for BIG-BENCH.

Improved QA-GNN formulation for Common Sense Question Answering (Course Project of Data Mining, 2022)

- Enhanced QA-GNN with dynamic attention; implemented and compared variants for commonsense QA.

Question Generation (QG)

- Developed rule-driven QG using dependency parsing, SRL, VerbNet/PropBank, and implicative constructions.
- Trained and analyzed T5 and BART models to study question diversity.

PDF based FAQ system

- Designed a document-level QA system for PDFs, including document ranking and text extraction.
- Implemented hybrid matching: rule-based (Dependency+VerbNet), embedding-based (FAISS/BLAS), and transformer semantic matching.
- Built a transformer-based entailment model for answer validation.

Probing Neural Architectures for Syntax

- Tested syntactic and word-order invariance (active↔passive, topicalization, extraposition, etc.) in modern LMs.
- Converted natural-language instructions to structured forms; mapped dependencies to detect conditional expressions.
- Benchmarked multiple transformer approaches for semantic parsing and reviewed SEMPRES.

Dialogue Modeling

- **Next Utterance Retrieval:** Benchmarked and deployed top-k response retrieval (skip-thought, paragram, contextual/ hierarchical LSTMs, transformers) with co-attention and intra-attention + CNN features (DSTC-7, Ubuntu).
- **Intent Classification & Entity Tagging:** Developed intent/NER frameworks (LSTM and transformer); built zero-shot IC via label descriptions; benchmarked span-based NER and relation extraction on domain datasets.
- **VerbNet-based Semantic Mapping:** Mapped senses and thematic roles to support IC/ET.
- **Clarification Question Generation:** Built rule-based and neural methods to resolve ambiguity.
- **PropBank/VerbNet Response Generator:** Implemented a SpanBERT-style architecture for structured response generation.
- **Contextual “Don’t Know” Generation:** Designed a T5-based generator; created a DoNotKnowDataset via Quora + syntactic rules using a novel technique of cross paraphrasing through T5.
- **Catenation:** Implemented utterance concatenation/substitution to update conversational context.

Affective & Others

- Built sentiment and emotion classifiers for English and other European languages.
- Designed a rule+neural toxicity detector.
- Integrated RoBERTa with OCC/PAD models to capture emotion, mood, and personality for users and bots.

Question Answering (QA)

- Implemented pre-SQuAD TREC-QA systems using tree-kernel (constituency/dependency), spreading activation, and lexical answer types.
- Built KB-QA with VerbNet predicates and backward chaining.
- Improved a multi-pointer-generator decoder for CoQA using a copy mechanism over context, question, and vocabulary.

Social Chat Bot Developed a chit-chat responder by training character-level response generator

Fine-Grained NER & Weak Labels

- Created weak labels from Wikipedia to train fine-grained NER over 16k entities.
- Explored data programming (Snorkel), semantic parsing (SEMPRE: domain-specific & Freebase), sentence gapping, and neural architecture search.

Face Recognition - Integrated MTCNN-based face detection and recognition; ran experiments on face-spoof identification.

Annotation Framework

- Led development of an administrative platform for intent/entity annotation, model training/tracking, and visual labeling (akin to wit.ai/dialogflow/luis).
- Designed a neural pipeline mapping imperative instructions to dialog responses

The Concert Queuing Game (Operations Research and Game Theory)

- Modeled multi-class customer queues with random arrival volumes using known distributions. (*Advisor: Prof. Sandeep Juneja*)

Opportunistic Spectrum Allocation in Radio Networks Using Game Theory (2014)

- Achieved near-global optimization for distributed channel selection via local-interaction games. (Advisor: Dr. Nitin Sharma)

Exploring Disulphide Connectivity in Animal Toxins using SVMs (2013)

- Predicted disulphide connectivity from protein sequences with 82% accuracy on the Animal Toxin Database. (Advisor: Dr. Sukanta Mondal)

Measuring Hardness of Training Data Extraction Attacks from Fine-Tuned Language Models (Privacy-Preserving ML, 2022) Kaustubh D., Shivam B.

- Evaluated whether post-fine-tuning attacks reduce leakage from pretraining/fine-tuning data using four paraphrase-based membership-inference metrics (based on Carlini et al., 2021).

RECOMMENDATIONS:

Professor Eugene Agichtein (Emory University)

Uday Chinta (Managing Director, [Amelia.ai](#))

James Gung (Amazon)

Simon Mille (ADAPT Research Centre)

Professor Christopher D. Manning (Stanford University)