Sai Chetan Chinthakindi

+1-630-822-5026 | scc8@illinois.edu | linkedin.com/in/sai-chetan-c

Education

University of Illinois Urbana-Champaign — Champaign, United States
Master of Science in Computer Science; GPA: 4.00/4.00 — *Aug. 2021 – Dec. 2022*

Indian Institute Of Technology Bombay — Mumbai, India
Bachelor of Technology in Computer Science; GPA: 8.65/10.00 — *Jul. 2014 – May. 2018*

Interests

Natural Language Processing, **Question Answering**, Optimization, Algorithms, Artificial Intelligence and Machine Learning

Technical Skills

Languages: C/C++, Python, Scala, Java, Javascript, R, SQL
Tools and Libraries: Tensorflow, PyTorch, keras, transformers, scikit-learn

Publications

- NeurQuRI: Neural question requirement inspector for answerability prediction in machine reading comprehension in **ICLR 2020** PDF
- Keep Learning: Self-supervised Meta-learning for Learning from Inference in **EACL 2021** PDF
- Learning to Generate Questions by Learning to Recover Answer-containing Sentences in **ACL-IJCNLP 2021 Findings** PDF
- Beyond Reptile: Meta-Learned Dot-Product Maximization between Gradients for Improved Single-Task Regularization in **EMNLP 2021 Findings** PDF
- NewsClaims: A New Benchmark for Claim Detection from News with Background Knowledge in **EMNLP 2022** PDF
- A Zero-Shot Claim Detection Framework using Question Answering in **COLING 2022** PDF

Scholastic Achievements

- Secured **All India Rank 2** in **IIT-JEE (Advanced) - 2014** among 150,000 students
- Secured **All India Rank 14** in **IIT-JEE (Mains) - 2014** among 1.3M students
- Among **top 35** students in India in **INAO** (Indian National Astronomy Olympiad)

Work Experience

Apple Inc. — Cupertino, California
Machine Learning Engineer, Apple Maps — *Jan. 2023 – present*

- Researched and engineered advanced neural ranking systems to improve search relevance and semantic understanding in Apple Maps
- Designed and deployed a neural re-ranking model using transformer-based language models (e.g., BERT and its variants), significantly improving top-K ranking precision across key map search intents such as POIs, navigation queries and addresses
- Integrated semantic search capabilities into the recall layer by embedding user queries and indexed entities into a shared vector space using dense retrieval techniques (e.g., dual encoders), enabling the system to match queries with high contextual alignment
- Contributed to multiple components of the search stack, including neural rankers, query rewriting (e.g., intent disambiguation and entity expansion), and spelling correction using character-level sequence models and edit-distance heuristics
- Leveraged large language models (LLMs) and instruction-tuned variants to enhance ranking and retrieval workflows, incorporating zero-shot and few-shot capabilities for handling tail queries and edge-case user intents
- Addressed critical production challenges such as inference latency (via model distillation and quantization), recall trade-offs, and error rate reduction, ensuring low-latency and high-availability performance for real-time search
- Analyzed loss trends and failure modes during model experimentation and A/B testing, contributing to diagnostics tooling for ranking stack evaluation and iteration
- Tech stack included Python, Scala, and ML frameworks such as TensorFlow, PyTorch, with deployment in distributed search infrastructure

Samsung Research — Seoul, South Korea
Research Engineer, Speech Recognition & Natural Language Processing — *Sep. 2018 – Jul. 2021*

- Researched and developed deep learning methods for Natural Language Processing, with a core focus on Machine Reading Comprehension (MRC) and generalization under limited supervision
- Developed NeurQuRI, a neural MRC framework for handling unanswerable questions by enforcing a structured checklist of answerability conditions across candidate spans, enabling robust rejection of invalid answers; published at ICLR
- Proposed novel pretraining objectives for MRC that move beyond standard language modeling by formulating supervised pretext tasks aligned with answer span prediction, improving sample efficiency and task alignment; published in ACL Findings
- Designed a first-order meta-learning method that enables MRC models to generalize from indirect supervision by learning from inference signals over unlabeled corpora, mimicking few-shot adaptation without explicit task supervision; published in EACL
- Introduced a gradient alignment strategy based on maximizing the dot product of gradient updates across batches, improving single-task regularization and optimization stability in deep networks; published in EMNLP Findings
- Techniques spanned transformer-based encoders (e.g., BERT, RoBERTa), meta-learning algorithms (e.g., MAML, Reptile-inspired variants), and differentiable scoring mechanisms for span selection and confidence modeling
- Codebase primarily implemented in PyTorch, with training pipelines built on HuggingFace Transformers

Internship

Apple Inc. Cupertino, California

SWE Intern, Apple Maps *May. 2022 – Aug. 2022*

- Designed and implemented a neural language model–based re-ranking system to improve search result relevance in a production-scale search pipeline
- Built a neural re-ranker leveraging transformer-based architectures (e.g., BERT, RoBERTa) to model semantic similarity between user queries and candidate search results, enabling context-aware relevance scoring beyond lexical matching
- Integrated the re-ranker into the second-pass ranking layer, allowing it to refine coarse results returned by the initial retrieval stage (BM25 or dense retrievers), leading to significant gains in precision@K and user engagement metrics
- Fine-tuned the model on task-specific relevance-labeled datasets using pairwise and listwise ranking losses (e.g., margin ranking loss, Softmax loss), with careful sampling of hard negatives for improved discriminative ability
- Optimized for low-latency inference using model distillation and quantization techniques to ensure real-time responsiveness in production environments

Samsung Research Seoul, South Korea

Research Intern, Speech Recognition & Natural Language Processing *May. 2017 – Jul. 2017*

- Developed and analyzed deep neural language models with a focus on linguistic generalization, memory augmentation, and cross-lingual representation learning
- Conducted controlled experiments to evaluate the efficacy of various language model components, including embedding layers, attention mechanisms, and contextual encoding strategies, across different NLP tasks
- Investigated the effectiveness of morpheme-level embeddings for agglutinative languages (e.g., Korean), demonstrating improved generalization in downstream tasks by capturing subword-level semantics and reducing vocabulary sparsity
- Implemented a Key-Value Memory Network architecture for factoid question answering, enabling the model to retrieve and reason over structured memory slots for accurate span-level answer selection
- Explored the impact of explicit memory representations and subword modeling in enhancing performance for both monolingual and multilingual QA datasets
- Used frameworks such as PyTorch and TensorFlow, and trained models with custom tokenization pipelines leveraging SentencePiece and mecab-ko for Korean morpheme analysis

Electronics for Imaging Bengaluru, India

Software Intern *May. 2016 – Jul. 2016*

- Designed and developed a modular web-based rule engine platform with RESTful APIs, enabling dynamic creation and management of custom business rules via an intuitive user interface
- Built a REST API framework using Node.js to expose endpoints for rule creation, update, and deletion, ensuring seamless backend integration with the rule engine logic
- Engineered a responsive and user-friendly web interface using HTML/CSS and JavaScript, allowing non-technical users to author, visualize, and manage complex rule hierarchies without writing code
- Integrated a MySQL relational database to persist rule metadata and execution states, with robust query handling for CRUD operations and transactional consistency
- Implemented comprehensive unit tests using the Mocha testing framework, to validate API behavior, edge cases, and error handling, contributing to high code reliability and maintainability
- Emphasized modular design, error resilience, and input validation throughout the system to support scalable deployment across multiple business domains

Teaching

- Graduate Teaching Assistant, University of Illinois Urbana-Champaign (Fall 2021, Spring 2022, Fall 2022) for **Intro to Data Structures and Algorithms with C++**
- Undergraduate Teaching Assistant, IIT Bombay (Spring 2018) for **Artificial Intelligence**
- Undergraduate Teaching Assistant, IIT Bombay (Spring 2017) for **Differential Equations**