

SIFFI SINGH

siffisingh26@gmail.com | +1 6262989751 | Pasadena, US | [linkedin.com/in/siffi](https://www.linkedin.com/in/siffi) | siffi26.github.io

Research Areas: Generative AI • Multimodal AI • Natural Language Processing • Computer Vision • Speech Recognition • Information Extraction • Knowledge Graphs

EXPERIENCE

Applied Scientist II

May 2022 – Present

Amazon Web Services

Pasadena, US

- Led research and productionization of Automated Theme Detection for 4,000+ Amazon Connect customers. Designed novel evaluation methodology for unsupervised thematic clustering where no ground-truth labels exist, combining structured human evaluation with LLM-as-judge protocols to drive iterative improvement. Built end-to-end pipeline using semantic clustering, BART-based label generation, and LLM-based synthetic data augmentation, achieving 86% cluster coherence and reducing manual analysis time by 80%.
- Led deployment of Semantic Intent Retrieval and AI-Powered Summarization, replacing keyword search with dense vector embeddings and intent understanding. Validated through systematic cross-domain ablations across pooling strategies and similarity thresholds whether a single unified model could generalize across voice and chat domains without domain-specific fine-tuning, achieving 70%+ precision and reducing search time from hours to minutes.
- Researched and implemented Multilingual PII Redaction across 14 languages using XLM-RoBERTa, designing and validating language-family-specific tokenization strategies and LLM-generated synthetic data to bootstrap low-resource language coverage. A core research challenge was handling code-switching in mixed-locale contact scenarios (fr-CA/en-CA, es-US/en-US); achieved average macro F1 of 80% across 14 languages.
- Built real-time Automatic Speech Translation prototype for AWS Transcribe, eliminating cascaded (Automatic Speech Recognition) ASR+Machine Translation (MT) error propagation. Formulated the latency-accuracy tradeoff as a reinforcement learning problem using GRPO, achieving BLEU score of 60 on English-Spanish with planned expansion to 10+ languages including PFIS and CJK.

Applied Scientist

Sep 2019 – May 2022

Amazon Alexa

Cambridge, UK

- Built large-scale relation extraction and ontology alignment models from web tables for Alexa knowledge ingestion. Identified the lack of a domain-grounded benchmark as a core research gap; designed and released RedTab, a benchmark dataset with structured annotation guidelines later adopted for production evaluation of knowledge graph extraction models, with findings published at COLING 2022.
- Scaled Alexa's Evi Knowledge Graph by 130% (44.5M to 103M entities) by architecting distributed GRETL ingestion pipelines for 11M structured facts across IMDB Movies and People topics. Complemented growth with stale-fact removal to maintain answer freshness, directly improving Alexa response quality at scale.
- Developed multilingual question-answering by integrating ReFinED entity linking with a hybrid KGQA-LM reasoning architecture across English, French, Spanish, Japanese, and Hindi. A core research question was whether grounding LM reasoning in retrieved knowledge graph facts could reduce hallucinations on complex queries; validated through systematic evaluation achieving 83% performance improvement over LM-only baselines.

Deep Learning Research Intern

Jun 2017 – Jul 2019

JMicron Technology (in collaboration with National Chiao Tung University)

Hsinchu, Taiwan

- Researched and developed a privacy-preserving face clustering and recognition system using unsupervised deep learning in TensorFlow, designed for devices with constrained hardware. A core research challenge was handling real-world face variations (rotation, illumination, aging) without ground-truth labels; designed a progressive multi-stage training approach that improved cluster robustness iteratively, achieving 86% ARI accuracy and commercially launched as LivCore. Work published at IEEE ICCE 2021 (**Best Paper Award**).

Computer Vision Research Intern

Jun 2018 – Aug 2018

Industrial Technology Research Institute

Hsinchu, Taiwan

- Developed multimodal benchmark datasets for image, LiDAR, and video fusion to evaluate segmentation and tracking algorithms for autonomous vehicle perception.

SELECTED PUBLICATIONS

- [1] HP Zou, **S Singh**, Y Nian, et al. “**Active Generalized Category Discovery with Diverse LLM Feedback**,” *European Chapter of the Association for Computational Linguistics (EACL)*, 2026.
- [2] I Shalymov, H Su, **S Singh**, et al. “**Controllable Conversational Theme Detection Track at DSTC 12**,” *Dialog System Technology Challenge (DSTC 12)*, 2025.
- [3] Y Zhang, **S Singh**, S Sengupta, et al. “**Can Your Model Tell a Negation from an Implicature? Unravelling Challenges with Intent Encoders**,” *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.
- [4] H Aboutaleb, **S Singh**, et al. “**MAGID: An Automated Pipeline for Generating Synthetic Multi-Modal Datasets**,” *North American Chapter of the ACL (NAACL)*, 2024.
- [5] L Tang, I Shalymov, **S Singh**, et al. “**TofuEval: Evaluating Hallucinations of LLMs on Topic-Focused Dialogue Summarization**,” *North American Chapter of the ACL (NAACL)*, 2024.
- [6] H Song, I Shalymov, **S Singh**, et al. “**Enhancing Abstractiveness of Summarization Models through Calibrated Distillation**,” *Findings of Empirical Methods in NLP (EMNLP)*, 2023.
- [7] KH Huang, **S Singh**, X Ma, et al. “**SWING: Balancing Coverage and Faithfulness for Dialogue Summarization**,” *Findings of the European Chapter of the ACL (EACL)*, 2023.
- [8] **S Singh**, AF Aji, GS Tomar, C Christodoulopoulos. “**A Relation Extraction Dataset for Knowledge Extraction from Web Tables**,” *International Conference on Computational Linguistics (COLING)*, 2022.
- [9] Y Zhou, **S Singh**, C Christodoulopoulos. “**Tabular Data Concept Type Detection Using Star-Transformers**,” *ACM International Conference on Information and Knowledge Management (CIKM)*, 2021.
- [10] **S Singh**, HM Hang. “**A Continuous Learning System for Face Clustering and Recognition**,” *IEEE International Conference on Consumer Electronics (ICCE)*, 2021.

EDUCATION

National Chiao Tung University

M.S. Electrical Engineering and Computer Science, GPA: 4.0/4.0

Dr. A.P.J. Abdul Kalam Technical University

B.Tech Computer Science and Engineering, Ranked 1st among 10,000+ students

Hsinchu, Taiwan

2017 – 2019

Noida, India

2013 – 2017

TECHNICAL SKILLS

- **Programming:** Python, Scala, Java, SQL, C/C++
- **Generative AI / LLMs:** PyTorch, TensorFlow, HuggingFace Transformers, RAG, Fine-tuning, LLM Evaluation, Prompt Engineering, Semantic Search, Vector Embeddings
- **Agentic AI:** LangChain, LangGraph, Multi-Agent Systems, Agentic RAG, MCP, ReAct, Tool Calling, LLM Feedback Loops, Prompt Chaining
- **Cloud Infrastructure:** AWS SageMaker, Lambda, EC2, S3, DynamoDB, EMR, Glue, Docker, PySpark
- **Research Areas:** Generative AI, Multimodal AI, Natural Language Processing, Computer Vision, Speech Recognition, Information Extraction, Knowledge Graphs

HONORS AND LEADERSHIP

- **External Collaborator, California Institute of Technology** (Prof. Sergei Gukov) – Research on reinforcement learning and search algorithms for mathematical reasoning and combinatorial optimization, 2025–2026
- **External Collaborator, University of Cambridge** (Prof. Burigede Liu) – Research on material property prediction using LLMs and tabular deep learning models, 2024–2025
- **Organizer, DSTC 12 Controllable Conversational Theme Detection Track**, 2025
- **Research Mentor, Amazon Alexa & AWS**, 2019–2026 – Guided scientists and engineers whose work led to publications at ACL, EMNLP, and EACL
- **Best Paper Award (1st Place), IEEE ICCE**, 2021
- **Invited Speaker** - “Making Alexa More Knowledgeable,” AI & Big Data Expo and REWORK AI Summit London, 2022
- **Silver Medal, ACM ICPC Asia Jakarta Regionals**, 2017
- **Google Women Techmakers Scholar**, 2018