Siva Rajesh Kasa

Curriculum Vitae

✉ *firstname.lastname@gmail.com*
kasakh.github.io
Google Scholar: 207 citations · h-index 8

Summary

Applied ML researcher and practitioner with 4+ years of industry experience shipping production-grade LLM systems for search relevance and ranking in e-commerce. Research interests span text classification and generation, inference acceleration, ML privacy, and distribution shifts. First-author publications at NeurIPS, TMLR, **EMNLP (Outstanding Paper)**, AISTATS, ACL, KDD, Bioinformatics, Econometrics & Statistics, and Scientific Reports. **Active open-source contributor**: upstreamed relaxed speculative decoding into HuggingFace `transformers` and author/maintainer of the `Mixture Models` (PyPI) and `quadmatrix` (CRAN) packages.

Industry Research Experience

Senior Applied Scientist, Amazon — Designed and Deployed production-grade Text generation and Classification systems for Search Relevance and Recommendation systems, covering multimodal calibration, long-tail ranking under sparse behavioral signal, counterfactual ranking analysis, and the A/B measurement infrastructure adopted by every internal search team. Owned the custom training and inference stacks behind these systems.

Education

2018-2022 **Doctor of Philosophy**, *School of Computing*, National University of Singapore.
Dissertation Title: Dependency Modeling with Mixture Models and Copulas, Supervisor: Dr. Vaibhav Rajan

2016-2018 **MBA**, *Decision Sciences*, Indian Institute of Management Bangalore.

2008–2012 **Bachelor of Technology**, *Mechanical Engineering*, Indian Institute of Technology Madras.

Publications

AR - Autoregressive models, DF - Diffusion models, IA - Inference Acceleration, HD - High-Dimensionality, RL - Reinforcement Learning, TS - Time Series, DM - Dependency Modeling, MM - Mixture Models, CLA - Classification, CLU - Clustering, PRI - ML Privacy

Conferences

DF, IA — Fast-dLLM++: Fréchet Profile Decoding for Faster Diffusion LLM Inference, **Siva Rajesh Kasa** et. al. **NeurIPS 2026 & SPGIM workshop at ICML 2026**

AR, DF, PRI — The Hidden Cost of Modeling $P(X)$: Membership Inference Attacks in Generative Text Classifiers, Owais Makroo and **Siva Rajesh Kasa** et. al. (joint first author) **TMLR 2026 (full version); Regulatable ML workshop at NeurIPS 2025 (abridged version)**

AR, IA, RL — DIVERSED: Relaxed Speculative Decoding via Dynamic Ensemble Verification, Ziyi Wang and **Siva Rajesh Kasa** et. al. (joint first author) **AISTATS 2026 Main Track and Efficient Reasoning workshop at NeurIPS 2025**

AR, DF, CLA — Generative or Discriminative? Revisting Text Classification in the era of Transformers, **Siva Rajesh Kasa** et. al. **EMNLP 2025 Main Track - Outstanding Paper Award**

AR, PRI, RL — Exposing Privacy Gaps: Membership Inference Attack on Preference Data for LLM Alignment, Qizhang Feng and **Siva Rajesh Kasa** et. al. (joint first author) **AISTATS 2025 Main Track**

AR, DF, CLA — Generative vs Discriminative? Revisiting the Shortcut Learning Debate in Text Classification, **Siva Rajesh Kasa** et. al. **Combining Theory and Benchmarks (CTB) workshop at ICML 2026**

DM, CLA — Leveraging Structural Information in Tree Ensembles for Table Representation Learning, Nikhil Pattisapu, **Siva Rajesh Kasa** et. al. **International World Wide Web conference WWW 2025 (Shortpaper)**

AR, CLA — Exploring Ordinality in Text Classification: A Comparative Study of Explicit and Implicit Techniques, **Siva Rajesh Kasa** et. al. **Findings of ACL 2024**

CLA — Tackling Concept Shift in Text Classification using Entailment-style modeling, Sumegh Roychowdhury, Karan Gupta, **Siva Rajesh Kasa**, et. al. (joint first author) **KDD 2024, Applied Data Science Track**

AR, CLA How robust are LLMs to in-context majority label bias?, Karan Gupta, Sumegh Roychowdhury, **Siva Rajesh Kasa**, et. al. (joint first author) **Proceedings of Responsible Language Modeling workshop, AAAI 2024**

DM, RL Dependency Modeling with Copulas in Multi-armed Bandits, **Siva Rajesh Kasa** and Vaibhav Rajan **International Conference on Information Systems (ICIS) 2021**

Preprints

AR, IA, DM CopRoute: Dependence-Aware Conformal Routing for Language Models, **Siva Rajesh Kasa**, Santhosh Kumar Kasa (joint first author), Karan Gupta, *preprint, under review*

DF SatisDive: Traversing the Satisfaction–Diversity Frontier in Text-to-Image Diffusion, Kevin Zhai and **Siva Rajesh Kasa**, et. al., *preprint, under review*

RL Correcting the Optimizer's Curse in World-Model Planning, **Siva Rajesh Kasa**, *preprint, under review*

Journals

MM, CLU Avoiding inferior clusterings with misspecified Gaussian mixture models, **Siva Rajesh Kasa**, Vaibhav Rajan, **Scientific Reports**, Volume 13, Issue 1, November 2023, **Nature**, **Impact Factor: 4.6**

HD, MM, CLU Gaussian Mixture Copulas for High-Dimensional Clustering and Dependency-based Subtyping, **Siva Rajesh Kasa**, Sakyajit Bhattacharya, Vaibhav Rajan, **Bioinformatics**, Volume 36, Issue 2, 15 January 2020, Pages 621–628, **Oxford University Press**, **Impact Factor: 6.9**

DM, MM, CLU Improved Inference for Gaussian Mixture Copula Models for Clustering and Reproducibility Analysis using Automatic Differentiation, **Siva Rajesh Kasa** and Vaibhav Rajan, **Econometrics and Statistics**, **Impact Factor: 1.5**

DM, TS A Statistical Test for Detecting Dependency Breakdown in Financial Markets, **Siva Rajesh Kasa** and Malay Bhattacharyya, **SN Computer Science**, Volume 2, Issue 4, Pages 322, 2021, **Impact Factor: 3.7**

Manuscripts under Preparation

MM, HD Mixture-Models: a one-stop Python Library for Model-based Clustering using various Mixture Models, **Siva Rajesh Kasa**, Hu Yijie, Santhosh Kumar Kasa, Vaibhav Rajan, *available on Arxiv*

MM, HD Mixture-Models: a one-stop Python Library for Model-based Clustering using various Mixture Models, **Siva Rajesh Kasa**, Hu Yijie, Santhosh Kumar Kasa, Vaibhav Rajan, *available on Arxiv*

DM, TS Dependency Modeling in Multi-armed Bandits, **Siva Rajesh Kasa** and Vaibhav Rajan, *in preparation for submission to Management Science*

DM, DL Multi-way Clustering of Heterogeneous Data through Deep Collective Matrix Tri-Factorization, Ragunathan Mariappan, **Siva Rajesh Kasa**, Vaibhav Rajan, *manuscript under preparation*

Miscellaneous

Pre MBA Ex Assistant Manager (2012-2016), Mechanical Engineering, NTPC LTD, *Responsible for installation and commissioning of high-pressure piping of 800MW super-critical power plant* .

Open Source Creator and maintainer of open source statistical computing packages.
`Mixture Models` (PIP), `quadmatrix` (CRAN)

Open Source Upstreamed DIVERSED relaxed speculative decoding (`assistant_ensemble_weight`) into HuggingFace `transformers`, PR #45979.

PC Member International Conference on Information Systems (ICIS) 2021.

Session Chair Resources and Evaluation Track, Empirical Methods in Natural Language Processing 2025.

Reviewer NeurIPS 2026, 2020, AISTATS 2021, EMNLP 2023, Econometrics and Statistics, Journal of Statistical Software, AISTATS 2025, Journal of Computational Statistics, Biometrika, Computational Statistics and Data Analysis, Scientific Reports.

Mentor YiChao (Final Year Undergrad @ NUS, Inverse RL for healthcare), Yijie (Final Year Masters @ NUS, High-dimensional Mixtures), Aniket Goel (Intern @ Amazon, Ordinal Text Classification), Qizhang Feng (Intern @ Amazon, Privacy in LLMs), Ziyi Wang (Intern @ Amazon, Final Year Ph.D. student at Purdue University).