Girishkumar Ponkiya

Data Scientist (Vice President) | Applied AI & NLP

Mumbai, India · girish.isical@gmail.com · +91-9004009268

girishp.in · linkedin.com/in/girishponkiya · github.com/girishponkiya · Google Scholar

PROFESSIONAL SUMMARY

Applied AI specialist and NLP practitioner with a Ph.D. from IIT Bombay and 6+ years building production AI systems in the financial industry. Builds LLM infrastructure at ISS STOXX — covering deployment, observability, and optimization — and serves as the organisation's go-to resource for NLP/LLM understanding and adoption, including a company-wide workshop series for software development teams. Published in ACL, EMNLP, and COLING; brings research-grade rigour to practical, scalable AI solutions.

TECHNICAL SKILLS

Applied NLP / LLMs: RAG pipelines, NER (spaCy fine-tuning), entity linking, knowledge graphs, text classification, word sense disambiguation

LLM Infrastructure: vLLM (deployment & monitoring), Telegraf, InfluxDB 2.0, Grafana, xGrammer (output consistency), Kubernetes, Apache NiFi, Vespa

ML / DL Frameworks: PyTorch, TensorFlow, Keras, scikit-learn, Hugging Face Transformers (BERT, RoBERTa, T5), ConvE

Languages & Data: Python, C/C++, Java, SQL; Apache Iceberg, JupyterLab, Pandas, Flask

PROFESSIONAL EXPERIENCE

Data Scientist (Vice President) | ISS STOXX, Technology Innovation Lab | *Mar 2021 – Present*

Mumbai, India | Individual contributor; cross-functional AI advisor across multiple engineering and product teams. Associate Vice President (Mar 2021 – Dec 2025) → Vice President (Jan 2026 – Present). Promoted in recognition of cross-team AI impact.

- Deployed and maintain multiple LLMs and generative AI models on internal GPU infrastructure using vLLM (including Gemma3:27B), applying FP8 quantization, kv-cache quantization, multi-GPU tensor parallelism, and optimal batch/context tuning; also diagnosed a CPU-level AVX2/AVX-512 bottleneck that led to a server hardware upgrade. Manager described this as a "break-through moment for my team and the company."
- Designed and currently delivering a company-wide "Introduction to LLM for Software Developers" workshop series (8 sessions planned; 3 delivered as of Mar 2026) using a flipped-classroom format covering architecture, prompt engineering, RAG, tool use, fine-tuning, and agents. ~70 participants per live session; additional reach via recordings.
- Grew the ESG news pipeline's business relevance rate from ~2% to ~35% — analysts now need ~3 reads per relevant article, down from ~50. Led a Japanese-language expansion that raised business relevance for Japanese content from ~9% to ~33%. The pipeline handles ~700K articles/day across 10+ languages.
- Improved the pipeline's relevance classifier from 82%+ to 85%+ accuracy by replacing a spaCy-based component with a tuned scikit-learn model (feature engineering + hyperparameter search).
- Evaluated 5 NER tools (open-source and commercial) against internal ESG data; selected and integrated 3 (spaCy, CoreNLP, Diffbot) into the production NiFi news filtering pipeline.
- Devised a novel threshold identification approach for the proprietary filtering rule that, for the first time, enabled mathematical confidence bounds on both the relevance rate and false positive rate — previously unmeasurable.
- Introduced entity linking into the ESG pipeline using a spaCy-based system; later redesigned it around a Memgraph knowledge graph to improve adaptability to a dynamic entity universe and index changes.
- Built and open-sourced vLLM Metrics Dashboard (github.com/iss-lab/vllm-dashboard) — a Telegraf → InfluxDB 2.0 → Grafana stack monitoring TTFT, throughput, KV-cache utilisation, HTTP responses, and Python GC across 3 vLLM inference instances.
- Built an AI-driven document processing pipeline for thematic index creation (PoC): PDF → Markdown (via Docling) → section filtering using heading matches, embeddings, and LLMs → in-house LLM summarisation to extract key business attributes from annual reports. Business stakeholders validated the output positively; pending productionisation.

- Developed an aspect-based sentiment analysis system (PoC) for the AI-driven sentiment index: assigns entity-level positive/negative/neutral sentiment per company from news article text, disambiguating sentiment when multiple companies appear in the same headline. Business validated outcome; ongoing.
- Leading development of AI-driven market/sector sentiment index for STOXX using E3 APIs and Vespa-backed vector embeddings.

Machine Learning Consultant | UnFound | *Dec 2018 – Jun 2019*

Mumbai, India | Pre-industry transition role.

- Re-defined question answering and stance detection pipelines using BERT; developed multi-task learning approach for stance detection.

Earlier Career

Research Intern, TRDDC / TCS Research, Pune (May–Jul 2014) — internship topic became Ph.D. thesis basis; mentor became co-supervisor. · Research Intern, Institute of Mathematical Sciences, Chennai (May–Jul 2012). · Trainer & Software Developer, Ishu InfoNet, Rajkot (2009–10). · Software Developer Trainee, HCL CDC, Ahmedabad (2009) — Best Trainee Award.

EDUCATION

Ph.D., Computer Science (Natural Language Processing)

Indian Institute of Technology Bombay (IIT Bombay), Mumbai

Thesis submitted: Mar 2021 · Degree awarded: 2023 | CPI: 7.7

Thesis: Noun Compound Interpretation — supervised by Prof. Pushpak Bhattacharyya (IIT Bombay) and Mr. Girish K. Palshikar (TRDDC). Best Poster (Technical), Research-Scholar Mela 2016.

M.Tech., Computer Science

Indian Statistical Institute (ISI), Kolkata | 2011 – 2013 | 78.54%

Dissertation: Priority Search Tree for Secondary Memory and its Application (Q-MER Problem). Nominated for Best Dissertation Award.

B.E., Computer Engineering

Saurashtra University, Rajkot | 2005 – 2009

PUBLICATIONS

~73 citations · h-index 5 · Google Scholar (Apr 2026)

[1] G. Ponkiya, D. Kanojia, P. Bhattacharyya, G. Palshikar. **FrameNet-assisted Noun Compound Interpretation.** *Findings of ACL-IJCNLP 2021.* [link]

[2] G. Ponkiya, R. Murthy, P. Bhattacharyya, G. K. Palshikar. **Looking inside Noun Compounds: Unsupervised Prepositional and Free Paraphrasing.** *Findings of EMNLP 2020.* [link]

[3] G. Ponkiya, K. Patel, P. Bhattacharyya, G. K. Palshikar. **Treat us like the Sequences we are: Prepositional Paraphrasing of Noun Compounds using LSTM.** *COLING 2018, Santa Fe, USA.* [link]

[4] G. Ponkiya, K. Patel, P. Bhattacharyya, G. K. Palshikar. **Towards a Standardized Dataset for Noun Compound Interpretation.** *LREC 2018, Miyazaki, Japan.* [link]

[5] G. Ponkiya, P. Bhattacharyya, G. K. Palshikar. **On Why Coarse Class Classification is a Bottleneck for Noun Compound Interpretation.** *ICON 2016, Varanasi, India.* [link]

[6] S. Pawar, S. Thombre, A. Mittal, G. Ponkiya, P. Bhattacharyya. **Tapping BERT for Preposition Sense Disambiguation.** *arXiv:2111.13972, 2021.* [link]

INVITED TALKS

- Entities and Relations — AICTE-ISTE STTP Program on NLP Novice-to-Pro, Shah & Anchor Kutchhi Engineering College, Mumbai. (Jan 2020)
- Noun Compound Interpretation — Workshop on NLP, VIVA Institute of Technology, Virar. (Jun 2016)

ACHIEVEMENTS

- GATE 2011: AIR 875 · GATE 2013: AIR 495, Percentile 99.78.
- Best Poster (Technical), Research-Scholar Mela 2016 / RISC 2016, CSE Dept., IIT Bombay.

LANGUAGES

English (Fluent) · Gujarati (Native) · Hindi (Advanced)