Bangalore, India · +91 78919 47877
apoorvumang@gmail.com · apoorvumang.github.io
Google Scholar · github.com/apoorvumang · linkedin.com/in/apoorvumang

Apoorv Umang Saxena

Research scientist working on **LLM inference efficiency** and **grounded generation** — how to make models faster without losing quality, and how to verify that what they generate is supported by their inputs. Author of **Prompt Lookup Decoding**, a speculative decoding method upstreamed into Hugging Face Transformers and vLLM. **2,000+ citations, h-index 10, 25+ papers** at ACL/EMNLP/NAACL/ICLR/ICCV, **12+ patents filed**, with a track record of taking research from publication to production at Adobe and Inception.

EDUCATION

Indian Institute of Science, Bangalore — Ph.D., Computational and Data Sciences 2018 – 2022
Advisor: **Prof. Partha Talukdar** (now Google DeepMind) · Intel India PhD Fellowship, 2018–2022
Thesis: Leveraging KG Embeddings for Knowledge Graph Question Answering

BITS Pilani — B.E. (Hons.) Computer Science 2011 – 2015
GPA 8.6/10 · BITS merit scholarship, semesters 1–4 (top 1% of batch)

EXPERIENCE

Inception — Member of Technical Staff Sep 2025 – Present
Bangalore
Inception builds Mercury, a family of diffusion LLMs that decode in parallel — 1000+ tokens/sec, sub-300ms TTFT. Joined as one of the early members of the India team; work spans the training and inference stack.

- **Inference:** core algorithmic contribution to application-specific decoding performance for the Mercury series, extending my earlier public work on speculative decoding. **2× speedup on Mercury Edit.**
- **Post-training (project lead):** led post-training of Mercury 2 for a large enterprise customer's query-rewrite workload. Shipped a model at parity with their production Gemini baseline on quality, **2.5× faster**.
- **Data engineering:** owned tool-calling data and evaluation for the **Mercury 2 and Mercury 2.5 releases**, improving function-calling performance across the family.
- **Search:** worked with search and answer-engine customers on integrating Mercury into agentic retrieval pipelines. Authored *Mercury 2 for Search* — showed Mercury 2 is the fastest model at every pipeline step (2× vs Gemini 3.1 Flash Lite on query planning, 10× vs GPT-5 Mini) and lowest cost per correct answer on FRAMES and DeepSearchQA.
- Found and published a **benchmark-contamination result**: WideSearch scores largely measure closed-book recall rather than retrieval — one frontier model scored *better* with search disabled. Rebuilt the task set from post-cutoff events, where the ranking inverted to a tie. Harness and ablations open-sourced.

Adobe Research — Research Scientist II Jun 2022 – Aug 2025
Bangalore
Document Experiences Lab — research on LLMs for document understanding, generation and grounding, taken from idea to shipped product.

- **Shipped attribution into Acrobat's AI Assistant**, giving users sentence-level grounding for answers over contracts and long documents. Research basis published at ACL Findings 2024.
- Contributor to the **first Adobe Firefly image model**.
- Published **15+ papers** at ACL, EMNLP, NAACL, ICCV, EACL and INLG during tenure, spanning attribution, long-document QA, inference efficiency and multimodal generation. **12+ patents filed.**
- Mentored research interns and junior researchers, several of whom led first-author papers at ACL, NAACL and EMNLP under my supervision.

Google — Software Engineer III May 2016 – Mar 2017
Hyderabad

- Apps for Work (now Google Workspace) — Tools & Infrastructure for the Permissions Management mobile app.
- Built test automation that raised tester productivity and cut release turnaround time.

Earlier: PayPal — Software Engineer I, Chennai · Flipkart — Intern, Bangalore · Bhabha Atomic Research Centre — Summer Intern, Mumbai 2013 – 2015

SELECTED PUBLICATIONS

2,000+ citations · h-index 10 · i10-index 10 · full list on Google Scholar

Inference efficiency and decoding

1. **A. Saxena**. *Prompt Lookup Decoding.* 2023. Speculative decoding via n-gram lookup over the prompt. Upstreamed into **Hugging Face Transformers** and **vLLM**; 617 GitHub stars.
2. D. J. Bajpai, S. Agarwal, **A. Saxena**, K. Kulkarni, S. Mitra, M. K. Hanawal. *FlowCast: Trajectory Forecasting for Scalable Zero-Cost Speculative Flow Matching.* **ICLR 2026.**
3. S. Somasundaram, A. Phukan, **A. Saxena**. *PLD+: Accelerating LLM Inference by Leveraging Language Model Artifacts.* **Findings of NAACL 2025.**
4. S. Shekhar, T. Dubey, K. Mukherjee, **A. Saxena**, A. Tyagi, N. Kotla. *Towards Optimizing the Costs of LLM Usage.* 2024. (91 citations)

Attribution, grounding and faithfulness

5. A. Phukan, S. Somasundaram, **A. Saxena**, K. Goswami, B. V. Srinivasan. *Peering into the Mind of Language Models: An Approach for Attribution in Contextual Question Answering.* **Findings of ACL 2024.** *(shipped in Adobe Acrobat AI Assistant)*
6. A. Phukan, H. K. Morj, **A. Saxena**, K. Goswami. *Beyond Logit Lens: Contextual Embeddings for Robust Hallucination Detection & Grounding in VLMs.* **NAACL 2025.** (30 citations)
7. N. Anand, S. Somasundaram, A. Phukan, **A. Saxena**, K. Mukherjee. *ContextFocus: Activation Steering for Contextual Faithfulness in Large Language Models.* 2026.
8. P. Ramu, K. Goswami, **A. Saxena**, B. V. Srinivasan. *Enhancing Post-hoc Attributions in Long Document Comprehension via Coarse Grained Answer Decomposition.* **EMNLP 2024.**
9. A. Rani, A. Garimella, **A. Saxena**, B. V. Srinivasan, P. P. Liang. *RADAR: A Reasoning-Guided Attribution Framework for Explainable Visual Data Analysis.* **Findings of EACL 2026.**

Knowledge graphs and question answering *(PhD work)*

10. **A. Saxena**, A. Tripathi, P. Talukdar. *Improving Multi-hop Question Answering over Knowledge Graphs using Knowledge Base Embeddings.* **ACL 2020.** (877 citations)
11. **A. Saxena**, A. Kochsiek, R. Gemulla. *Sequence-to-Sequence Knowledge Graph Completion and Question Answering.* **ACL 2022.** (336 citations)
12. **A. Saxena**, S. Chakrabarti, P. Talukdar. *Question Answering over Temporal Knowledge Graphs.* **ACL 2021.** (276 citations)
13. A. Sharma, **A. Saxena**, C. Gupta, M. Kazemi, P. Talukdar, S. Chakrabarti. *TwiRGCN: Temporally Weighted Graph Convolution for QA over Temporal Knowledge Graphs.* **EACL 2023.** (48 citations)
14. A. Kochsiek, **A. Saxena**, I. Nair, R. Gemulla. *Friendly Neighbors: Contextualized Sequence-to-Sequence Link Prediction.* **RepL4NLP @ ACL 2023.**

Documents, retrieval and multimodal generation

15. A. Agarwal, S. Karanam, K. J. Joseph, **A. Saxena**, K. Goswami, B. V. Srinivasan. *A-STAR: Test-time Attention Segregation and Retention for Text-to-Image Synthesis.* **ICCV 2023.** (108 citations)
16. A. Goyal, K. Mukherjee, **A. Saxena**, A. Phukan, E. Chandrasekharan et al. *Masking or Mitigating? Deconstructing the Impact of Query Rewriting on Retriever Biases in RAG.* **Findings of ACL 2026.**
17. A. Jain, P. Ramu, A. Garimella, **A. Saxena**. *Doc2Chart: Intent-Driven Zero-Shot Chart Generation from Documents.* **EMNLP 2025.**
18. I. Nair, S. Somasundaram, **A. Saxena**, K. Goswami. *Drilling Down into the Discourse Structure with LLMs for Long Document Question Answering.* **Findings of EMNLP 2023.**
19. S. Bandyopadhyay, H. Maheshwari, A. Natarajan, **A. Saxena**. *Enhancing Presentation Slide Generation by LLMs with a Multi-Staged End-to-End Approach.* **INLG 2024.** (35 citations)
20. T. Santosh, N. Modani, **A. Saxena**. *A Tale of Two Revisions: Summarizing Changes Across Document Versions.* **Findings of ACL 2024.**

PATENTS

12+ filed during tenure at Adobe Research. Public to date:

- **US 12,633,000 (granted)** — *Text-to-image synthesis utilizing diffusion models with test-time attention segregation and retention optimization.*
- US App. 18/924,398 — *Generating draft sequence rankings for speculative decoding using large language model hidden states.*
- US App. 18/750,243 — *Thematic summary generation of digital document differences.*
- US App. 18/671,265 — *Determining large language model effectiveness utilizing deep learning.*
- US App. 18/593,834 — *Summary page generation using documents.*

TALKS, TEACHING AND OPEN SOURCE

- **Guest lecturer**, *Introduction to Natural Language Processing (DS 207)*, Indian Institute of Science — lecture and notebook on **LLM decoding**, invited two years running (2025, 2026). Course by Prof. Danish Pruthi.
- **Reviewer** for *ACL conferences (ACL / EMNLP / NAACL), through ACL 2025. Presented at ACL 2020, 2021 and 2022 and subsequent *ACL venues.
- **Prompt Lookup Decoding** — 617★, integrated in Hugging Face Transformers and vLLM (`ngram` speculative decoding).
- **KGT5** — 105★, ACL 2022 reference implementation and HuggingFace Spaces demo. **CronKGQA** — 98★, ACL 2021 temporal KGQA.
- **retrieval-vs-recall** — open harness and ablations behind the WideSearch contamination finding. **LLM Knowledge Cutoff Benchmark** — measures models' *actual* knowledge cutoff versus the date labs report; live leaderboard and open code.

SELECTED PROJECTS

TokenPath · tokenpath.ai · 2025 – present, co-founded with one collaborator

Token-level attribution API: traces any LLM answer back to the exact source spans behind it by measuring attention over the document. Model-agnostic, post-hoc, no regeneration.

- **0.815 citation F1 on LongBench-Cite** — matches Anthropic's Citations API (0.812) and lands within 0.04 of a prompted frontier LLM (0.851), at **~7× lower cost** ($0.013 vs ~$0.09 per query) and **~5–6× lower latency** (p50 1.6s vs 8–10s).
- Scales to million-token documents via overlapping windows; validated across 23 languages plus all 22 scheduled Indian languages, including cross-lingual attribution.
- Built end to end: API, GPU serving on open-weight models (on-prem / air-gapped capable), docs and cookbook, and a Chrome extension for grounded chat over any page, PDF or video.

Everything 3D · everything3dindia.com · 2023 – present

Custom 3D-printing store — design, storefront and fulfillment.

AWARDS

- Intel India PhD Fellowship — 2018–2022
- **4th place, Teen-size league, RoboCup 2013** (Netherlands), representing India with team Acyut — India's first indigenous humanoid robot, sponsored by the Govt. of India (DeitY)
- 1st runner-up, image processing and robotics events, IIT Bombay Techfest 2013 and 2014 · 1st place, DesertHack 2012, BITS Pilani
- AIEEE rank 301 (top 0.05%) · IIT-JEE rank 2,423 (top 0.49%) — 2011

SKILLS

- **ML / inference:** PyTorch, Hugging Face Transformers, vLLM, speculative decoding, diffusion LLMs, post-training (SFT / preference optimization), attention analysis, KG embeddings
- **Evaluation:** agentic-search benchmarking, contamination and ablation design, LLM-as-judge harnesses
- **Languages & systems:** Python, C++, Java, C, JavaScript/TypeScript · GPU serving, API infrastructure, Node, React