Susmit Das

Privacy Research Scientist (AI) | LLM Reasoning, Auditable AI & Agentic Systems

susmitdas@alum.iisc.ac.in | GitHub: susmitds | Google Scholar | ORCID: 0000-0001-6837-1121

Website: susmitds.github.io | OpenReview: ~Susmit_Das1 | LinkedIn: susmitds

GRE General (Sep 2026): Q **170**, V **163**, AW **4.5** | **TOEFL iBT (Nov 2025): 112/120** (R29, L29, S27, W27)

Research Interests

LLM reasoning and post-training; adaptive thinking; reasoning evaluation and output auditability; trustworthy language agents; ML reproducibility.

Peer-Reviewed Research

TIME: Temporally Intelligent Meta-reasoning Engine for Context-Triggered Explicit Reasoning

FINDINGS OF ACL 2026 | SOLE AUTHOR | INDEPENDENT RESEARCH

Susmit Das. *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 38872–38906.

- **Framework:** Introduced TIME, a behavioral-alignment framework that treats explicit reasoning as a *context-triggered control policy* rather than a fixed thinking/no-thinking mode. Uses temporal discontinuities as cues to latent conversational state change, with timestamped turns, silent tick events and optional, locally placeable reasoning traces.
- **Alignment:** Designed a four-phase supervised fine-tuning curriculum for Qwen3 (4B–32B): structural seeding, temporal exposure, contextual modulation, then a maximally diverse 128-conversation full-batch stage designed to isolate the target reasoning policy across varying topics and styles.
- **Evaluation:** Introduced TIMEBench, a model-agnostic diagnostic of reasoning from temporal cues rather than date recall (77 scenarios; seven categories; 770 trials/model). TIME-32B scored **64.81 vs. 37.40** against the original Qwen3-32B in thinking mode, with approximately **84 vs. 911 explicit reasoning tokens/run**; gains were observed across all four tested sizes.

Paper: aclanthology.org/2026.findings-acl.1936/ **DOI:** 10.18653/v1/2026.findings-acl.1936

Short explanation: susmitds.github.io/time.html | **Code/data:** GitHub **Benchmark:** TIMEBench

PERAI: Privacy Enhancing & Responsible AI Framework for Data & AI Lifecycle

IEEE ISoPE 2025 | FIRST AUTHOR | INDUSTRY RESEARCH, PRIVASAPIEN

Susmit Das, Abilash Soundararajan, Sahil Suresh Kamble, Sayanil Ghosh, Ankith Udupa, Romit Bhaumik, Siddharth Desai, Nitesh Chachriya. *2025 IEEE Symposium on Privacy Expectations*, pp. 32–37.

Role: Led the seven-person R&D team; developed a five-stage privacy-by-design framework spanning risk assessment, privacy-enhancing methods, model evaluation, inference-time controls and agent governance.

Paper: ieeexplore.ieee.org/document/11486340/ **DOI:** 10.1109/ISoPE67098.2025.00026

Current Manuscripts

Ephemeral Models, Durable Claims: Closed-Model Dependence and the Preservation of AI Research

SUBMITTED TO ICLR 2027 | UNDER REVIEW

- **Scope:** Audited **53,538 final archival papers** from ACL, EMNLP, ICLR, ICML and NeurIPS (2023–2026), separating four scientific roles (direct experimental use, baseline, judge and data generation) from ancillary assistance.
- **Method:** Built a frozen measurement pipeline combining full-text retrieval, deterministic evidence rules and local LLM adjudication of ambiguous cases; blinded validation on 200 random papers: **0.967 F1** for scientific-use detection and **0.888 micro-F1** across the four roles. Code, processed data and validation artifacts accompany the submission.
- **Findings: 17,841 papers (33.32%)** have a covered closed-model scientific dependency; **7 of the 20** most-used normalized models had retired from first-party APIs by 24 Sep 2026. Characterizes retirement, drift and under-specified identifiers as distinct preservation risks, not a measured irreproducibility rate.

OpenReview: openreview.net/forum?id=b4dstEWp9v | **Short explanation:** susmitds.github.io/ephemeral-models.html

Walk Me Through Your Work: Explainability in LLMs Lies in the Auditability of Outputs

SUBMITTED TO ACM FAccT 2027

- **Position:** Argues that explainability for reliance concerns whether an LLM's public output *justifies its claims*, rather than whether its rationale reproduces hidden computation; distinguishes output auditability from mechanistic interpretability and chain-of-thought faithfulness.
- **Framework:** Develops a semantic account of replayable justification with statement roles (fact, assumption, inference, claim), recoverable support sets, local entailment and first-failure localization; illustrates the framework through worked cases.

Short explanation: susmitds.github.io/wmtyw.html

Education

Indian Institute of Science (IISc) 2022–2024

Master of Management (M.Mgt.) — Business Analytics | CGPA: 8.70/10

Graduate training in deep learning, NLP and machine learning through institute-wide electives, alongside the programme's quantitative and data-science core.

Graded institute electives: Deep Learning for NLP (CSA, A); Machine Learning (CSA, B+); Bayesian Learning (RBCCPS, A); Topics in AI (ECE, A); Foundations of Secure Computation (CSA, A).

Thesis: *Enhancing Privacy in Large Language Models* (A+; advisor: Prof. Parthasarathy Ramachandran). Compared seven NER architectures (best: DeBERTa-v3-large+CRF, **0.9536 F1**) and evaluated pseudonymization privacy–utility trade-offs across five domains and six LLMs.

RCC Institute of Information Technology 2017–2021

B.Tech., Computer Science and Engineering | 8.14/10

Research Experience

PrivaSapien Technologies — Privacy Research Scientist (AI) Jul 2024–Present

Early technical-team member; previously a research intern (May–Jul 2023 and Jan–Jun 2024).

- **Multi-agent DPIA system:** Delivered a multi-agent system for regulatory question mapping, interactive assessments, answer validation, scoring and recommendations; v2 added conditional question trees, policy-clause provenance and revalidation after answer edits.
- **Document-grounded multi-agent assistant:** Built a copilot with task-scoped retrieval agents, evidence reconciliation and source-backed answer suggestions; users retain approval within the assessment workflow.
- **Privacy threat modeling:** Developed sensitive-entity detection (DeBERTaV3/GLiNER, pattern matching and lookup) and privacy-exposure measures integrated into assessments.
- **Privacy-preserving inference:** Built reversible pseudonymization for LLM/RAG prompts; extended FF3-1-based format preservation for mixed-format text using custom alphabets, chunking and reassembly.
- **Agent tool evaluation:** Co-developed pre-connection screening of tool descriptions and schemas, with a 100-case benign/adversarial evaluation set.
- **Intern mentoring:** Trained and assessed interns arriving with introductory ML experience but little privacy-engineering background, including domain onboarding and reviewing their technical work.
- **Client and cross-sector engagement:** Met regularly with client DPOs and CISOs to discuss privacy risks in prevalent data and AI practices, explain technical concerns and learn their operational challenges; participated in company initiatives such as Safe Human Future.

Personal Lab & Tinkering

Outside work, I run large open-weight models on a self-funded four-GPU workstation for fun, experimentation and my own coding agents. Keeping recent open-weight releases working locally has led me to maintain custom llama.cpp forks for architecture compatibility on my mixed setup (including custom kernels and low-bit quantization ports), develop mixed-precision quantizations, and write a constrained MoE expert-mix optimizer. The same hands-on skills carry into my independent research: I trained and evaluated TIME and designed, ran, validated and analyzed the Ephemeral Models study alongside full-time employment.

Academic Service

Reviewer: AAAI 2026, NeurIPS 2026, AAAI 2027.

Earlier Peer-Reviewed Book Chapters

S. Das, S. Karmakar, H. N. Saha. *Secure Digital Health Data Management in Internet of Things Using Blockchain and Machine Learning.* CRC Press, 2022, pp. 23–45. DOI: 10.1201/9781003247128-2.

S. Das, S. Karmakar, H. N. Saha. *An Efficient Hash-Selection-Based Blockchain Architecture for Industrial IoT (IIoT).* Springer LNDECT 112, 2022, pp. 237–265. DOI: 10.1007/978-981-16-9260-4_10.

Technical Scope

Research: LLM training/post-training, curriculum and data design, information extraction, benchmarking, corpus-scale measurement and statistical validation.

Systems/tooling: Multi-agent systems, RAG, model serving, GGML/GGUF, mixed-precision quantization; Python, C/C++, PyTorch, vLLM, llama.cpp, Flask/FastAPI, Docker, Linux, Git, GCP.