IJCNLP-AACL 2025 Tutorial Proposal

📌 Title

Beyond Guardrails: Advanced Safety for Large Language Models — Monolingual, Multilingual and Multimodal Frontiers

*By **Somnath Banerjee**, **Rima Hazra** and **Animesh Mukherjee***

📝 Abstract

*LLMs are now embedded in workflows that span languages, modalities, and tools. This raises safety challenges that outpace conventional "guardrails": jailbreaks and prompt injections, attributional safety failures under code-mixing, multimodal bypass via typography and icons, activation-level manipulation, and agentic risks from tool use. This tutorial synthesizes **the newest advances (2023–2025)** and lays out **open research questions** around (i) failure modes in monolingual / multilingual / multimodal settings, (ii) training-time and inference-time defenses (rejection SFT, RLHF/RLAIF, decoding-time safety, parameter/activation steering), and (iii) evaluation and red-teaming pipelines balancing safety and utility. We anchor the tutorial with recent results including our safety related papers published at top tier conferences, and connect them to emerging best practices from recent safety tutorials. The target audience is researchers/engineers with basic NLP knowledge who want the latest techniques and a research roadmap; format is half-day with short demos and Q&A.*

📖 Introduction

*LLM safety has matured from monolithic policy filters to a research field spanning **robustness, value alignment, and secure tool use**. Yet even frontier models can be compromised by **context switching, euphemisms, persona modulation, and multilingual shifts**; multimodal systems can be jailbroken through **typographic prompts, iconography, and cross-modal indirection**; The community now needs a consolidated view that goes beyond "refuse-lists" to **mechanistic defenses**, **language-aware alignment**, and **measurable safety–utility trade-offs**.*

*This tutorial is timely for IJCNLP-AACL because it centers around **linguistic diversity** (code-mixing, low-resource languages), **multimodal safety**, themes that directly affect real deployments and research agendas. We build on the structure and harms taxonomy popularized in recent safety tutorials while extending it with multilingual and agentic emphases, further distilling the lessons into concrete evaluation playbooks and open problems.*

Computational Linguistics: Tutorial Abstract, pages 25–33
December 20-24, 2025 ©2025 Association for Computational Linguistics

🎯 Target Audience

- *NLP researchers, applied scientists, industry professionals, and practitioners working with LLMs.*
- *Red-teamers / robustness and security researchers.*
- *Policy/governance professionals seeking a technical grounding in LLM safety.*

Prior knowledge:

- *Transformers, fine-tuning, and basic evaluation; Python proficiency is helpful; prior experience in AI Safety is not required.*
- *Optional but useful: familiarity with multilingual NLP and/or basic computer vision.*

Expected participants: *60–90, based on recent safety tutorials' attendance and the cross-disciplinary interest in multilingual and agentic safety.*

🗂️ Outline

Format & duration: *Half-day (≈3.5–4 hours total, including a 30-minute coffee break). Short lectures with focused demos and Q&A. The flow mirrors classic NLP tutorial structuring (background → cases/defences → evaluation/roadmap) while concentrating on 2023–2025 advances.*

Session 1 (80 minutes): How safety fails in 2025

1. LLM safety primer (15')

- *Scope: harmlessness/helpfulness/honesty; harms taxonomy (bias/toxicity, hallucination, privacy leakage, disinformation, unsafe assistance).*
- *What's new since 2023: multilingual/code-mixed failures; multimodal typographic attacks; agentic failure modes in tool-augmented systems.*

2. Failure modes (25’)

- ***Prompt-level jailbreaks & evasions:*** *euphemisms; context switching; role-play/virtual simulation; task-format asymmetries (e.g., summarization vs QA).*

- ***Decoding/activation manipulation:*** *when sampling, logits, or activation steering open “glide paths” around alignment.*

- ***Multimodal jailbreaks:*** *typography and icons; flowchart-style logic images; cross-modal indirection and weak compositionality in VLMs.*
- ***Model Editing:*** *stealth safety erasure, localized backdoors/trigger phrases, capability insertion in specific languages/modalities*

3. Deep-dive Monolingual Solutions and Case studies (35’)

- *[**Language Models are Homer Simpson! Safety Re-Alignment of Fine-tuned Language Models through Task Arithmetic**](https://aclanthology.org/2024.acl-long.762/) (Bhardwaj et al., ACL 2024)*

- *[**SafeInfer: Context Adaptive Decoding Time Safety Alignment for Large Language Models**] AAAI 2025, 39, 27188-27196. (Banerjee et al., AAAI 2025)*

- *[**Safety Arithmetic: A Framework for Test-time Safety Alignment of Language Models by Steering Parameters and Activations**](https://aclanthology.org/2024.emnlp-main.1212/) (Hazra et al., EMNLP 2024)*

Coffee break (30 minutes)

Session 2 (80 minutes): How to defend and how to measure

4. Deep-dive Multilingual and Multimodal Solutions and Case studies (40’)

- *[**Soteria: Language-Specific Functional Parameter Steering for Multilingual Safety Alignment**](https://arxiv.org/abs/2502.11244) (Banerjee et al., 2025)*

- *[**Navigating the Cultural Kaleidoscope: A Hitchhiker's Guide to Sensitivity in Large Language Models**](https://aclanthology.org/2025.naacl-long.388/) (Banerjee et al., NAACL 2025)*

- *[**MemeSense: An Adaptive In-Context Framework for Social Commonsense Driven Meme Moderation**](https://arxiv.org/abs/2502.11246) (Adak et al., 2025)*

- *[**Immune: Improving Safety Against Jailbreaks in Multi-modal LLMs via Inference-Time Alignment**](https://arxiv.org/abs/2411.18688) (Ghosal et al., CVPR 2025)*

5. Evaluation & red-teaming (25')

- ***Metrics:** obedience/rejection, relevance/fluency, harmfulness/toxicity, overkill (helpfulness loss).*

- ***Testbeds:** Do-Not-Answer; adversarial prompt generation (e.g., prompt-attack styles); multilingual suites; multimodal safety probes.*

- ***Pipelines:** automated red-team harnesses; measuring safety–utility balance; reporting standards.*

6. Open problems & roadmap (20')

- ***Affordance (risk-aware alignment):** As models gain tools (code exec, web, files, images), even benign prompts can unlock risky actions the user didn't intend. Safety is needed to detect those latent capabilities and gate or steer them in real time so small requests don't escalate into harmful operations.*
- ***Pluralistic alignment:** Real users span empathy, sensitivity, and values; a single global policy either over-blocks or harms specific groups. Safety is needed to respect legitimate differences—selecting the right norms per context—while keeping outputs lawful, fair, and still useful.*
- ***Implicit-math evasion:** Harmful requests can be disguised as innocent arithmetic or optimization (e.g., splitting quantities across sources, unit conversions, route planning). Safety is needed to track restricted entities and totals across steps, normalize units, and block aggregation-based procurement even when posed as "just math."*
- ***Domain-specific safeguards:** Use domain-aware tagging and global constraints (hazard classes, thresholds, legality checks) that persist across turns, not just*

per-message screening; if a restricted entity is detected anywhere in the chain, enforce refusal and pivot to safe, educational alternatives.

🌍 Diversity Considerations

1. ***Content & fairness:*** *The tutorial centers* ***multilingual/code-mixed*** *safety and cultural sensitivity—directly improving fairness for under-represented languages/communities and aligning with IJCNLP's regional strengths.*

2. ***Audience impact:*** *Particularly relevant for researchers and practitioners serving* ***low-resource*** *and* ***code-mixed*** *user populations, and for moderators handling* ***culturally coded multimodal content****.*

📚 Reading List

Reading List — Attacks, Jailbreaks, and Prompt Injection

- *LinkPrompt: Natural and Universal Adversarial Attacks on Prompt-based Language Models — arXiv:2403.16432 — https://arxiv.org/abs/2403.16432*

- *Neural Exec: Learning (and Learning from) Execution Triggers for Prompt Injection Attacks — arXiv:2403.03792 — https://arxiv.org/abs/2403.03792*

- *Jailbreaking Leading Safety-Aligned LLMs with Simple Adaptive Attacks — arXiv:2404.02151 — https://arxiv.org/abs/2404.02151*

- *Rapid Optimization for Jailbreaking LLMs via Subconscious Exploitation and Echopraxia — arXiv:2402.05467 — https://arxiv.org/abs/2402.05467*

- *AutoDAN: Interpretable Gradient-Based Adversarial Attacks on Large Language Models — arXiv:2310.1514 — https://arxiv.org/abs/2310.1514*

- *Universal and Transferable Adversarial Attacks on Aligned Language Models — arXiv:2307.15043 — https://arxiv.org/abs/2307.15043*

- *Soft-prompt Tuning for Large Language Models to Evaluate Bias — arXiv:2306.04735 — https://arxiv.org/abs/2306.04735*

- *TrojLLM: A Black-box Trojan Prompt Attack on Large Language Models — arXiv:2306.06815 — https://arxiv.org/abs/2306.06815*

- *AutoDAN: Generating Stealthy Jailbreak Prompts on Aligned Large Language Models — arXiv:2310.04451 — https://arxiv.org/abs/2310.04451*

- *Gradient Cuff: Detecting Jailbreak Attacks on Large Language Models by Exploring Refusal Loss Landscapes — arXiv:2403.00867 — https://arxiv.org/abs/2403.00867*

- *Token Highlighter: Inspecting and Mitigating Jailbreak Prompts for Large Language Models — arXiv:2412.18171 — https://arxiv.org/abs/2412.18171*

Reading List — Guardrails, Safety Tooling, and Benchmarks

- *RigorLLM: Resilient Guardrails for Large Language Models against Undesired Content — arXiv:2403.13031 — https://arxiv.org/abs/2403.13031*

- *NeMo Guardrails: A Toolkit for Controllable and Safe LLM Applications with Programmable Rails — arXiv:2310.10501 — https://arxiv.org/abs/2310.10501*

- *Llama Guard: LLM-based Input-Output Safeguard for Human-AI Conversations — arXiv:2312.06674 — https://arxiv.org/abs/2312.06674*

- *SPML: A DSL for Defending Language Models Against Prompt Attacks — arXiv:2402.11755 — https://arxiv.org/abs/2402.11755*

- *Robust Safety Classifier for Large Language Models: Adversarial Prompt Shield — arXiv:2311.00172 — https://arxiv.org/abs/2311.00172*

- *AI Control: Improving Safety Despite Intentional Subversion — arXiv:2312.06942 — https://arxiv.org/abs/2312.06942*

- *Maatphor: Automated Variant Analysis for Prompt Injection Attacks — arXiv:2312.11513 — https://arxiv.org/abs/2312.11513*

- *Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned — arXiv:2209.07858 — https://arxiv.org/abs/2209.07858*

- *DICES Dataset: Diversity in Conversational AI Evaluation for Safety — arXiv:2306.11247 — https://arxiv.org/abs/2306.11247*

- *Latent Jailbreak: A Benchmark for Evaluating Text Safety and Output Robustness of Large Language Models — arXiv:2307.08487 — https://arxiv.org/abs/2307.08487*

- *Tensor Trust: Interpretable Prompt Injection Attacks from an Online Game — arXiv:2311.01011 —* [*https://arxiv.org/abs/2311.01011*](https://arxiv.org/abs/2311.01011)

- *Can LLMs Follow Simple Rules? — arXiv:2311.04235 —* [*https://arxiv.org/abs/2311.04235*](https://arxiv.org/abs/2311.04235)

- *SimpleSafetyTests: A Test Suite for Identifying Critical Safety Risks in Large Language Models — arXiv:2311.0837 —* [*https://arxiv.org/abs/2311.0837*](https://arxiv.org/abs/2311.0837)

- *Benchmarking and Defending Against Indirect Prompt Injection Attacks on Large Language Models — arXiv:2312.14197 —* [*https://arxiv.org/abs/2312.14197*](https://arxiv.org/abs/2312.14197)

- *SC-Safety: A Multi-round Open-ended Question Adversarial Safety Benchmark for Large Language Models in Chinese — arXiv:2310.05818 —* [*https://arxiv.org/abs/2310.05818*](https://arxiv.org/abs/2310.05818)

- *Walking a Tightrope — Evaluating Large Language Models in High-Risk Domains — arXiv:2311.14966 —* [*https://arxiv.org/abs/2311.14966*](https://arxiv.org/abs/2311.14966)

👥 Presenters

***Somnath Banerjee** is currently Technical Leader at Cisco Systems. He has submitted his PhD thesis from the Department of Computer Science and Engineering, IIT Kharagpur, on "TUTORING LARGE LANGUAGE MODELS TO BE DOMAIN-ADAPTIVE, PRECISE AND SAFE" in May 2025. Earlier, he received his M.Tech from the same department in 2018 from IIT(ISM) Dhanbad. Mr. Banerjee received the prestigious university gold medal for academic excellence for his masters journey.*
Mr. Banerjee's research interests include large language models safety, evolution and change, NLP for resource-poor languages and domain adaptation. He has more than 20 publications in prestigious CORE A* and A conferences such as NeurIPS, AAAI, ACL, EMNLP, NAACL, COLING, ECML - PKDD, IEEE Bigdata, ASONAM. *A complete list of his publications can be found at his webpage:* [*https://scholar.google.com/citations?user=X5Zh5BwAAAAJ&hl=en*](https://scholar.google.com/citations?user=X5Zh5BwAAAAJ&hl=en)*.*

Rima Hazra is a senior postdoc at Eindhoven University of Technology (TU\e), Netherlands. Earlier she was a Postdoctoral Researcher at the Singapore University of Technology and Design, working in the area of AI safety alignment, natural language processing, and LLM reasoning. She earned her Ph.D. from the Indian Institute of Technology, Kharagpur, where she explored the area of Information retrieval, NLP and graph learning. With experience in information retrieval, NLP and graph learning, Dr. Hazra has published several papers in prestigious CORE A* and A conferences such as AAAI, ACL, EMNLP, NAACL, ECIR, ECML-PKDD and JCDL. She has also received the prestigious Microsoft Academic Partnership Grant

(MAPG) and the PaliGemma Academic Program award from Google for her work in AI safety alignment.

Animesh Mukherjee

Presently, Animesh Mukherjee is a Full Professor in the Department of Computer Science and Engineering, Indian Institute of Technology, Kharagpur. He is also a Distinguished Member of ACM. His main research interests center includes safe and responsible AI including red teaming and alignment of LLMs/VLMs. He regularly publishes in all top CS conferences including AAAI, IJCAI, ACL, NAACL, EMNLP, The Web Conference, CSCW, etc. He has received many notable awards and fellowships including the Facebook ethics for AI research award, India, Google course award for the course AI and Ethics, IBM faculty award, Humboldt Fellowship for Experienced Researchers, Simons Associateship, ICTP, to name a few.

ℹ️ Other Information

Expected attendees: *60–90, extrapolating from recent safety tutorials and workshops; interest is high in* ***multilingual****.*

Special equipment/requirements: *Projector; stable internet for* ***sandboxed demos*** *(with pre-recorded backups if connectivity is limited); no external privileged systems.*

Unique features:

- *A* ***linguistically grounded*** *safety tutorial (code-mixing, cultural context) rarely covered in depth.*
- ***Hands-on*** *mini-evaluations and take-home red-teaming checklists.*
- *Clear* ***research roadmap*** *for 2025–2027 with concrete problem statements.*

⚖️ Ethics Statement

*We adopt **responsible disclosure** and **do-no-harm** norms throughout. Demos use **sanitized** prompts and **non-actionable** examples; no personal data are used. We explicitly discuss: (i) privacy and data leakage risks; (ii) bias/fairness, especially under multilingual/code-mixed settings; (iii) dual-use concerns in red-teaming and how to publish findings that advance defence without enabling misuse; and (iv) community reporting practices and documentation that balance transparency with safety.*

Some recent publications in this area by the instructors

1. *[**SafeInfer: Context Adaptive Decoding Time Safety Alignment for Large Language Models**] AAAI 2025, 39, 27188-27196. (Banerjee et al., **AAAI 2025**)*
2. *[**Safety Arithmetic: A Framework for Test-time Safety Alignment of Language Models by Steering Parameters and Activations**](https://aclanthology.org/2024.emnlp-main.1212/) (Hazra et al., **EMNLP 2024**)*
3. *[**Soteria: Language-Specific Functional Parameter Steering for Multilingual Safety Alignment**](https://arxiv.org/abs/2502.11244) (Banerjee et al., 2025)*
4. *[**Navigating the Cultural Kaleidoscope: A Hitchhiker's Guide to Sensitivity in Large Language Models**](https://aclanthology.org/2025.naacl-long.388/) (Banerjee et al., **NAACL 2025**)*
5. *[**MemeSense: An Adaptive In-Context Framework for Social Commonsense Driven Meme Moderation**](https://arxiv.org/abs/2502.11246) (Adak et al., 2025)*
6. *[**Sowing the Wind, Reaping the Whirlwind: The Impact of Editing Language Models**](https://aclanthology.org/2024.findings-acl.960/) (Hazra et al., **ACL 2024**)*
7. *[**AURA: Affordance-Understanding and Risk-aware Alignment Technique for Large Language Models**](https://arxiv.org/abs/2508.06124) (Adak et al., 2025)*
8. *[**Attributional Safety Failures in Large Language Models under Code-Mixed Perturbations**](https://arxiv.org/abs/2508.06124) (Banerjee et al., 2025)*
9. *[**How (Un)ethical Are Instruction-Centric Responses of LLMs? Unveiling the Vulnerabilities of Safety Guardrails to Harmful Queries**] ICWSM 2025. (Banerjee et al., **ICWSM 2025**)*
10. *[**InfFeed: Influence Functions as a Feedback to Improve the Performance of Subjective Tasks**](https://aclanthology.org/2024.lrec-main.794/) (Banerjee et al., **COLING 2024**)*