# Sourabrata **Mukherjee**

Postdoctoral Researcher · Microsoft Research India

I build generative models — and the measurements that decide whether they actually work.

**Ph.D. Charles University, Prague** | **6+ years production ML**

souro.github.io · github.com/souro · linkedin.com/in/souro

## WHAT I WORK ON

- **Generative AI**
  controllable generation, 10+ languages
- **Training & alignment**
  SFT · LoRA · DPO · PPO · GRPO
- **Evaluation & interpretability**
  LLM judges, agents, benchmarks
- **Where measurement fails**
  culture, low-resource languages, reward hacking

# One thread, four chapters

Ship ML systems → build generative models → learn to measure them → fix what the measurement breaks

**Industry**
2014–2019
Smart search, recommender systems

→

**Ph.D.**
2019–2025
Generative AI, controllable generation

→

**Internships**
2022–2025
Multilinguality, socio-cultural bias

→

**Microsoft Research**
2025–now
LLM-judge and agent evaluation

## WHAT I BUILD

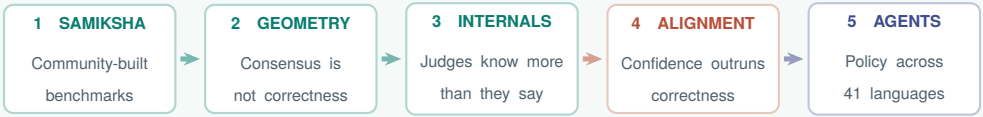
- **Generation** · style, politeness, detoxification
- **Training** · full FT · LoRA · DPO · PPO · GRPO
- **Multilingual** · 10+ languages, native annotators

## WHAT I MEASURE

- **Benchmarks** · built with the communities served
- **Judges** · when they agree, when they are wrong
- **Agents** · does a policy survive a new language?

# The MSR story in one line each

Each step is the question the previous step forced




## DIAGNOSE

- Where do trustworthy **labels** come from?
- Does judge **agreement** imply correctness?
- What do judges **compute but not report**?

## INTERVENE

- **Calibrate & adapt** · tune judges to human nuance
- **Align** · reward the answer, not the confidence
- **Extend** · to agent trajectories, not just answers

## THE STACK I RUN

**Train** — full FT & LoRA, distributed multi-GPU.

**Align** — SFT → DPO · PPO · GRPO.

**Open up** — logit lens, probes, calibration.

# Before research: six years shipping ML

Amdocs → o9 Solutions → Avaya → Tricon Infotech · 2014–2019

## TRACK RECORD

| YEARS | COMPANY | WHAT I OWNED |
|---|---|---|
| 2014–15 | Amdocs | Search & recommendation engine, e-commerce |
| 2015–16 | o9 | Recommender framework, enterprise planning |
| 2016–17 | Avaya | Real-time log pipeline at scale |
| 2017–19 | Tricon | Domain recommender; **AI Tech Lead** |

## RECOGNITION

- **Star Employee of the Year** — Amdocs
- **Star Employee of the Year** — Tricon
- **National Winner** — IBM competition

## HANDS-ON ENGINEERING

**Models**

PyTorch, gradient boosting, graph and sequence models.

**Search**

Query understanding, relevance ranking, boosting, A/B tests.

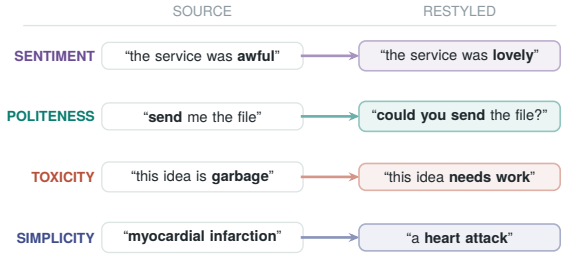
**Recommendation**

Preference models over user–item graphs, at scale.

# Ph.D. — controllable generation

Text Style Transfer · Charles University, Prague · advisor Ondřej Dušek

## ONE ATTRIBUTE MOVES. EVERYTHING ELSE MUST NOT.




## FIVE QUESTIONS, FIVE CONTRIBUTIONS

- **Data scarcity** — non-parallel training, style reward TSD’22 INLG’23
- **Multilinguality** — data + models, 10 languages INLG’24 ICON’24
- **Evaluation** — four metric tiers vs. humans NAACL’25
- **Application** — politeness in a dialogue system EACL’23
- **LLMs** — where prompting wins, and where not INLG’25

**Hard because:** no parallel data, **none outside English** — and no off-the-shelf answer to “did it work?”

# Ph.D. I — modelling without parallel data, and outside English

TSD 2022 · INLG 2023 · INLG 2024 · ICON 2024

## NO PARALLEL DATA TSD'22

- **Shared pre-trained encoder** · content
- **One decoder per style** · control is architectural
- **Sentiment masking** · denoising objective
- **SOTA style accuracy** at baseline content — not traded against it

## A LITTLE PARALLEL DATA INLG'23

- **Low-resource MT recipe** · augmentation, filtered self-training
- **Style-reward term** · the evaluated property enters the loss
- **Best content preservation**; competitive with GPT-3.5 in English

## MULTILINGUAL INLG 2024 · ICON 2024

- **1,000 parallel pairs** per language, native annotators, **10 languages**
- **The multilingual gap was a data gap, not a capability gap**

# Ph.D. II — the result that redirected my research

Meta-evaluation of style-transfer metrics · NAACL 2025

### THE PROBLEM UNDERNEATH

- Style scored by a **classifier**
- Content scored by **BLEU**
- Both correlate **poorly with humans**

### FOUR TIERS OF METRIC, ONE SET OF HUMAN RATINGS

- **Existing** — classifier accuracy · BLEU · cosine
- **Adapted** — EMD, classifier confidence, BERTScore, BLEURT
- **Novel** — **AMR-graph** and **syntax-tree** similarity
- **Hybrid + LLM judges** — learned ensemble; GPT-4, Llama-3.1

**Ordering:** adapted > existing · hybrid > both · **LLM judges correlate best with humans**

I had that from *one* correlation, *one* task, *one* language. **Testing it properly became my postdoc.**

# Research internships — multilinguality and social norms

MBZUAI, Abu Dhabi · EMNLP 2025 Main: honorific usage in Bengali and Hindi Wikipedia

## THE QUESTION

- Hindi and Bengali mark respect **grammatically**
- Wikipedia has **no guideline** for it
- A norm is already applied — **whose, and to whom?**

## DESIGN

- **10k** articles per language
- GPT-4o labels + **6 attributes**
- Native validation **85%** / **93%**

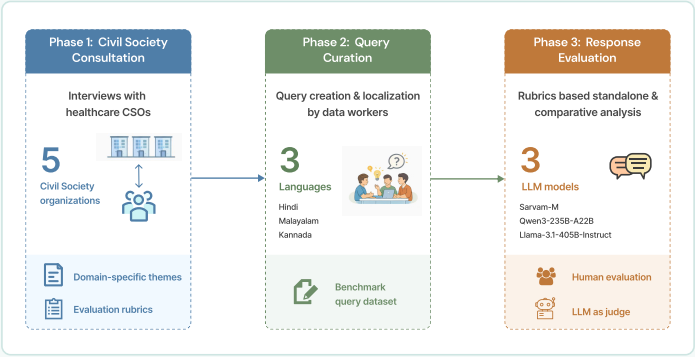
## WHO GETS HONORIFICS $\chi^2$, ALL $p < 0.05$; $+$ = HONORIFIC

| | BN | HI | | BN | HI |
|---|---|---|---|---|---|
| famous | $+1.11$ | $+1.10$ | infamous | $-0.97$ | $-1.42$ |
| older | $+1.22$ | $+0.63$ | juvenile | $-0.89$ | $-0.66$ |
| male | $+0.42$ | $+0.57$ | female | $+0.22$ | $\mathbf{-0.15}$ |

**The control that makes it a bias claim:** native HI speakers are balanced ($+0.12$ vs. $+0.09$); HI Wikipedia is not. **An editorial convention, not a property of the language** — and every model trained on that corpus inherits it.

# MSR I — Samiksha: build the human labels first

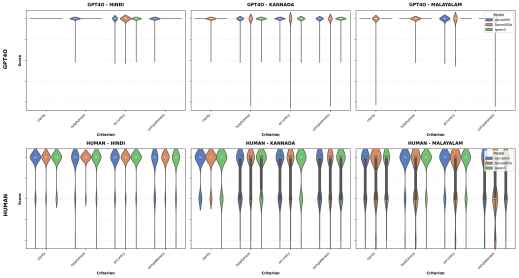
CHI 2026 · Honourable Mention · community-centred benchmarking with civil-society organisations across India



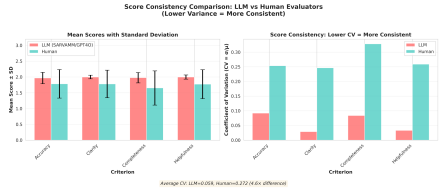

11 languages × 6 domains | sub-domains co-created with NGOs | two annotator tiers | usability outranks correctness

# MSR I — what the automatic evaluator did to those labels

Judges pin every criterion at ceiling; the community annotators use the whole scale




GPT-4o judge (top) vs. community annotators (bottom).



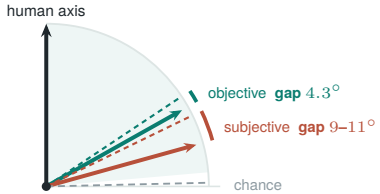

Score variance: humans spread, judges do not.

*Hindi:* “knees hurt on police duty” → a **Nordic track**. **Right physiology, wrong country** — scored at ceiling.

# MSR II — The Geometry of LLM-as-Judge

EMNLP 2026 · 42 judges · 170,890 judge calls · 8 languages · 2 community benchmarks

## CONSENSUS IS NOT ALIGNMENT




Solid = judges · dashed = *matched* human floor · wedge = the gap.

## WHAT THE ANGLE SETTLES

- **Claim under test** — *judges agree* ⇒ *judges are right*
- Shared competence and shared blind spots give **identical agreement** ⇒ **non-identifiable**
- **Fix** — score columns as **vectors**, read the **angle** against a *reference-matched* human floor
- judge↔judge **0.356** ≈ human↔human **0.357**; judge↔human **0.207**
- A **humans-only** statistic predicts the regime **before any LLM runs**; 6 tuning recipes widen spread but not angle — **calibration, not orientation**

Human annotation can be cut where the geometry passes — **and nowhere else**.

# MSR III — Opening the Judge: interpretability of the evaluator

EACL 2026, under review · 64 models, 0.36B–405B · 41 causally intervened · 14 datasets

## METHOD

- **Read-out** · every layer, through the model's **own** LayerNorm and unembedding
- **Debias** · order-average $(A, B)$ and $(B, A)$ — no gradient step
- **Probe** · $L_2$ logistic, 5-fold, every number out-of-fold

## LLMBAR-ADVERSARIAL 50 JUDGES

| | |
|---|---|
| emitted verdict, order-debiased | **0.456** |
| lens probe (41 params) | **0.758** |
| residual-stream probe | **0.846** |
| shuffled-label control | 0.500 |

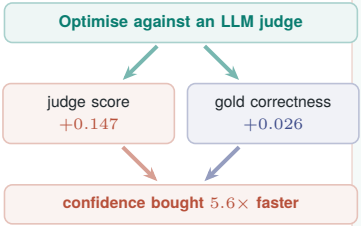
**The correct preference is inside the judge even when its output is below chance** — positive for **100%** of judges.

**Cheap triage:** a **length-only** probe predicts the recovery over 8 benchmarks at $\rho = \mathbf{0.90}$ — **decide before you spend**.

# MSR IV — Evaluation Is Training: when the judge becomes the reward

EACL 2026, under review · DPO · PPO · GRPO · best-of-$N$ · 11 verifiable multilingual benchmarks

## WHAT OPTIMISATION BUYS




## WHY IT IS A MECHANISM, NOT AN ANECDOTE

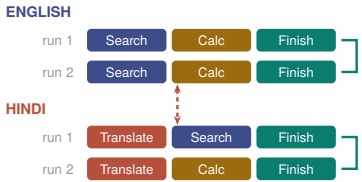
- **Transfers** · ∼**85%** of the gain reaches judges never trained against
- **Is the evaluator** · against a generic reward model the effect **vanishes**
- **Is rank-based** · the judge scores the **correct answer below the wrong one it picks** ⇒ ensembling does not help
- **Hits the tail** · largest in **low-resource languages**; re-orders shipped leaderboards

**The fix is affordable:** verify the ∼**25%** of items judges most disagree on — $1.3\times$ the correctness recovery of random verification.

# MSR V — Actions Speak Louder than Words: agents across languages

COLM 2026 · Reviewers' Spotlight · 8 models · 41 languages · 505 cells · 2.38M rollouts

## WHAT WE MEASURE




$$\tilde{I} = I_{\text{cross}} \,/\, I_{\text{within}}$$

Both legs compare run 1 with run 2, so only the language differs.

## THE NAIVE NUMBER IS NOT AN ESTIMAND

- **C1 no baseline** · decoding noise read as a language effect
- **C2 trace length** · $R^2 = 0.67$, **flips two signs**
- **C3 empty traces** · non-adherence *rewarded*
- **C4 ceiling** · **the ranking ranks determinism**
- **C5 no zero** · permutation floor $c \approx 0.56$

**Every correction made the effect larger:** $+0.063 \rightarrow +0.086 \rightarrow +0.207$.

# In summary

**BUILD**

- Controllable generation
- Multilingual data & models
- Full FT · LoRA · adapters
- DPO · PPO · GRPO

**MEASURE**

- Community-built benchmarks
- Matched human floors
- Judge internals, causally
- Agent trajectories, 41 languages

**FIX**

- Calibration & adapter tuning
- Correctness-first alignment
- Cheap triage before the spend
- Released, reproducible artifacts

**How I work** — predict first · ship the falsification · release the artifact · measure before you spend

**Thank you** — happy to go deeper on any of these.