COLM 2026

Conference on Language Modeling | San Francisco, October 6–9, 2026

Actions Speak Louder than Words

Measuring Cross-Lingual Policy Retention in Tool-Using Agents

Sourabrata Mukherjee Kalika Bali Sunayana Sitaram

Microsoft Research, India

8	6	41	5,776	2.38M
models	benchmarks	languages	tasks	rollouts

arXiv:2608.11110

THE PROBLEM • A REAL EXAMPLE

One question, six languages, six different routes

EN Who won Super Bowl XLIX?

HI सुपर बाउल XLIX किसने जीता?

Translate · Search · Finish

English	Finish	✓ 1 step
हिन्दी Hindi	Translate, Translate, Finish	✓ 3 steps
中文 Chinese	Search, Finish	✓ 2 steps
العربية Arabic	Translate, Translate, Translate, Finish	✓ 4 steps
Ελληνικά Greek	Search, Search, Search, Search, Search, Search, Search, Search, Finish	✓ 9 steps
Deutsch German	Search, Search, Search, Search, Search, Search, Search, Search, Search	✗ none

1 → 9

steps for the same right answer, English versus Greek

Thinks in English

Hindi translates twice, then answers in English

Qwen3-235B, XQuAD item 8, one real rollout per language, T=0.5

Same answer, different behaviour. Evaluation today only checks the answer.

© Sourabrata Mukherjee

THE PROBLEM • WHY IT MATTERS

For an agent, the route is the product

Cost

Extra steps change the price in each language.

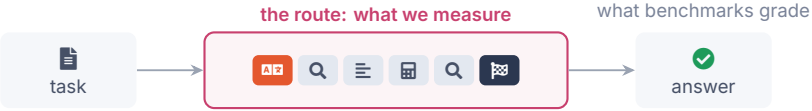
Failure

Steps off the English path fail where tests never look.

Governance

Permissions and audits are written against actions, not answers.

THE BLIND SPOT



Our question: do agents run the same action policy in every language?

© Sourabrata Mukherjee

METHOD • WHY IT IS HARD

The obvious measurement fails for five reasons

No baseline

0.63–0.80

agreement between two runs in one language

✔ Run every cell twice

Trace length

$R^2 = 0.67$

of raw agreement explained by length

✔ Length-match both ways

Empty traces

0.79 → 0.02

one model, once blank pairs are dropped

✔ Drop empty pairs

Ceiling

$r = +0.97$

gap against self-consistency, T=0

✔ Divide by the ceiling

Chance floor

0.56

already scored by traces of different tasks

✔ Measure it by permutation

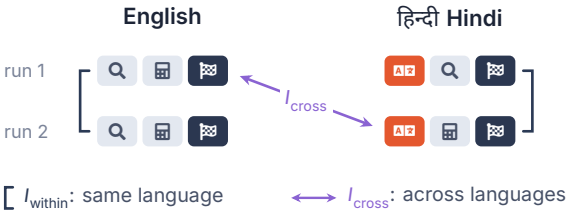
On our own data, four of the five reverse a sign or a ranking.

© Sourabrata Mukherjee

METHOD • WHAT WE MEASURE

Run every cell twice, then divide

TWO RUNS PER CELL



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

POLICY RETENTION

$$\tilde{I} = \frac{I_{cross}}{I_{within}}$$

The share of a model's own consistency that survives a change of language.

- ✔ Length-matched
- ✔ Empty pairs dropped
- ✔ Task-level bootstrap
- ✔ Chance floor measured

© Sourabrata Mukherjee

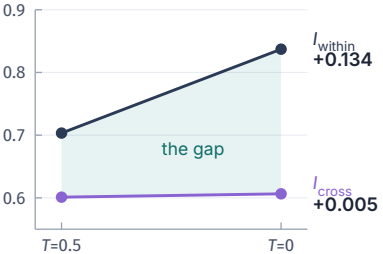
RESULTS • FINDING 1

The gap is real, and sampling was hiding it

Every correction made it larger

length-matched gap, pooled over frontier models

Remove sampling: one line moves



At T=0 the gap is positive in all 24 cells, every interval above zero.

© Sourabrata Mukherjee

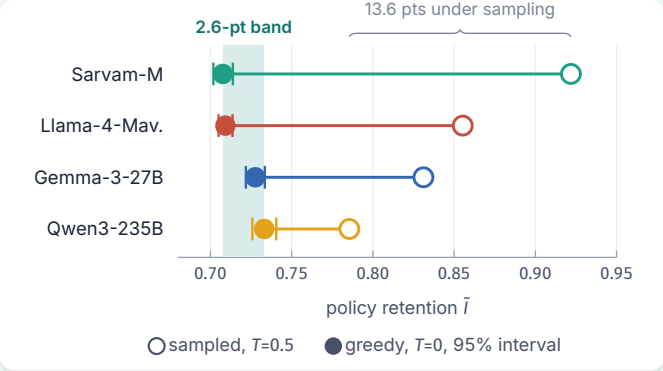
RESULTS • FINDING 2

Frontier agents keep 71–73% of their action policy

71–73%

of a model's own consistency survives a change of language

four frontier models, greedy decoding (T=0)



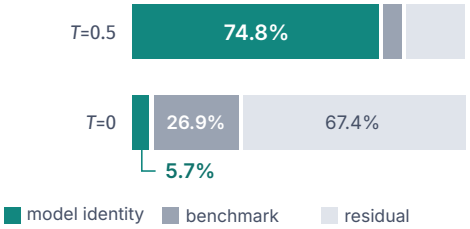
Four very different models land within 2.6 points of each other.

© Sourabrata Mukherjee

Which model it is barely matters

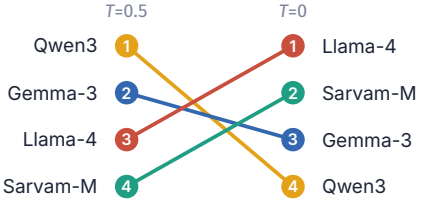
What explains the variation in $\tilde{I}$?

share of variance across 24 cells



Ranked by the raw gap

rank 1 = largest cross-lingual gap



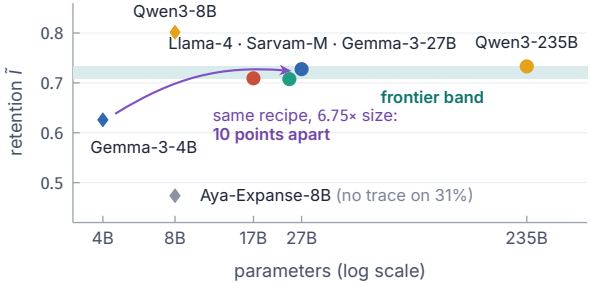
$\rho = -0.80$

A ranking on the raw gap measures determinism, not multilinguality.

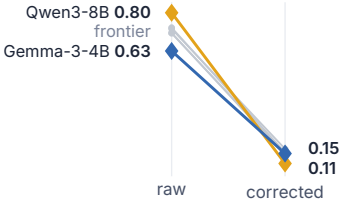
© Sourabrata Mukherjee

Below roughly 10B, the constant breaks

Retention by model size



The chance floor flips it



Corrected, Qwen3-8B falls below the 4B model.

Below about 10B, only the chance-corrected scale can order models.

© Sourabrata Mukherjee

ANALYSIS • WHY IT HAPPENS

Agents plan in English, and will not stop

We observe it

#1

Translate is the most-used tool

≈99%

of reasoning is ASCII, even on हिन्दी, தமிழ், ଓଡ଼ିଆ tasks

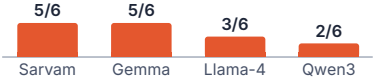
We intervene

Remove Translate

Length-matched, agreement **drops**

Mandate Translate first

Gains track head-room, as predicted **in advance**



We instruct

"Reason in the task's language."

Gemma-3-27B 0.79%

Sarvam-M 0.08%

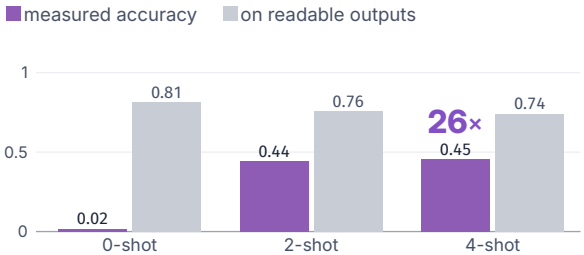
<1% comply

The English pivot is load-bearing, and prompts cannot switch it off.

© Sourabrata Mukherjee

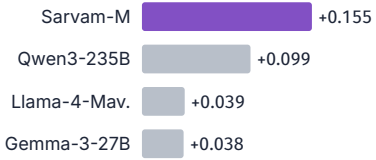
A regex, not the model, faked a multilingual failure

Parser expects `Action: Translate(text=…)`
GPT-OSS writes **"We will use Translate."** ✕ no action



English still wins on accuracy

English advantage, four gold benchmarks



The Indic-specialised Sarvam-M has the largest English edge.

Report parse-failure rates; above about 20%, leave the model unranked.

© Sourabrata Mukherjee

CONCLUSION

Answer agreement is not behavioural agreement

WHAT WE SHOWED

1. A ceiling-corrected estimand that prices five confounds
2. Frontier agents keep 71–73% of their action policy under greedy decoding
3. Below about 10B, the regularity breaks
4. An English pivot drives it, and prompts cannot remove it

WHEN YOU EVALUATE AN AGENT

- Log the **action trace**
- Run every condition **twice**
- Rank on **retention** $\tilde{I}$
- **Measure** the chance floor
- Report **parse failures** and T

> **An agent equally accurate in two languages is not thereby equally trustworthy in both.**

© Sourabrata Mukherjee

Thank you

Microsoft Research, India

Questions and discussion welcome

connect.souro@gmail.com

aka.ms/multilingual_agents_evaluation

Code, prompts, both benchmark suites and all 2.38M traces

Paper

Website

LinkedIn