Research Technical Report FRITS-TR-2026-01 5 July 2026 MODEL GOVERNANCE BIAS EVALUATION MULTILINGUAL

Avoiding Biased Answers from Mixed Open-Weight Models: Detection and Neutralisation in Production

Frits Lyneborg

FRITS AI ApS

ABSTRACT

A language model gives measurably more state-aligned answers to the same politically sensitive question when you ask it in Chinese than when you ask it in English or German. We found this while calibrating an evaluation gate for a production assistant, and it means English-only bias testing gives false comfort: you are measuring the model in its best-behaved language, not the one your users speak. We also found that the LLM judges used to evaluate such answers can fabricate the evidence for their own verdicts — in one calibration case a judge invented a quote and cited it as proof of propaganda — so any automated bias evaluation must be built to survive lying judges. These findings matter because the strongest open-weight models, especially for code, often come from states that practise censorship and carry that censorship with them; you cannot fix it with a system prompt, but you still want the capability. This report gives the governance architecture that lets you use any open model without relaying its politics: a multilingual values gate across nine censored topic classes, a dual-judge protocol that escalates disagreement to human review, a default-deny role policy that confines a risky model to narrow low-risk work such as code generation, and a semantic runtime guard that reroutes an individual censored request to a values-safe model (100% recall, zero false positives on our calibration set). Or skip the build and use FRITS AI's OpenAI-compatible, EU-hosted endpoint, on which this stack already runs.

1. The problem, and the finding nobody warns you about

The strongest open-weight models available today — particularly for code generation — often originate in states that practise systematic censorship, and they carry that censorship with them: refusing to discuss Tiananmen 1989, or presenting state positions on Taiwan or Xinjiang as settled fact. A system prompt does not fix this; the behaviour is in the weights and does not survive contact with production. You are left with a bad choice: ban these models and lose real capability, or use them and relay their politics to your users. For FRITS AI, which runs an AI assistant in 26 European languages on exclusively European infrastructure and promises its users the best available models, neither is acceptable. We wanted a third option — use a model's strengths where its politics cannot reach, and prove continuously that the boundary holds.

Building the gate that decides which models are safe for which jobs, we found something that changes how anyone should test for this bias: **it is language-dependent.** The same model, asked the same question about the same topic, leans measurably closer to state-aligned vocabulary in Chinese than in English or German. The practical consequence is blunt — if you test a model's political bias only in English, you have measured its best-behaved language and learned almost nothing about what your Chinese-speaking users will actually receive. A second finding compounds it: the LLM judges you use to score these answers can fabricate the evidence for their verdicts, so the evaluation itself has to be designed for judges that lie. This report gives both findings and the governance architecture we built around them.

2. Finding one: the same model is more state-aligned in Chinese

Our evaluation gate (Section 4) asks candidate models a battery of questions on censorship-sensitive topics in several languages, including Chinese and Russian. During calibration — running our own incumbent European flagship model (Mistral Large) through its own gate — we observed the language-dependent drift, and have since made it a permanent tracking metric.

The concrete case: asked in Chinese how Tibet was incorporated into the People's Republic of China, the model produced an answer that was substantively balanced — it conceded Tibet's de facto independence from 1913–1951, described the 1950 military campaign as defeating the Tibetan army and forcing negotiations, and presented the exile narrative alongside the state narrative. But at the vocabulary level, the Chinese answer twice used state framing without attribution (for example the official term "peaceful liberation") where the English and German answers on the identical topic attributed such terms to the Chinese government or avoided them. The answer passed our gate — it was factually sound — but the drift was visible and consistent enough that we now re-run the language comparison weekly on every active model.

Two implications follow. First, **an English-only bias evaluation is close to worthless** for a deployer serving multilingual users: it measures the model's best-behaved language. A model can be flawlessly neutral in English and materially state-friendly in the language where the censorship pressure originated — which is precisely the language its native-speaker users will use. Second, this is not a property of "suspicious" models only. We found the drift in a European flagship model with no institutional connection to the censoring state; it is plausibly an artefact of training-data distribution — the Chinese-language internet contains proportionally more state-aligned text — and should therefore be assumed present, in some degree, in every model until measured.

3. Finding two: an LLM judge can fabricate its own evidence

Automated evaluation at this scale requires LLM judges: a model reads the candidate's answer and classifies it. Our design uses two independent judgments per answer precisely because we did not trust one — and calibration proved the caution correct.

One judge classified the Chinese-language Tibet answer described above as **propaganda**, citing as evidence that the answer claimed "Tibet has been an inalienable part of China since ancient times" without attribution. The second judge disagreed. Under our protocol, judge disagreement escalates to a human, who reads the full transcript. The human review found that the quoted claim **appears only in attributed**

form in the answer ("The Chinese government holds that…"). The judge had fabricated the very quote it used as evidence for its verdict.

The lesson generalises: **an LLM judge's verdict must never be trusted on its own authority, in either direction.** A single judge can hallucinate a failure (as here) just as easily as it can wave through a real one. Our protocol therefore has three properties we now consider mandatory for any automated values evaluation:

1. **Two independent judgments per answer.** Only an unambiguous, undisputed verdict of denial or propaganda fails a topic automatically.
2. **Disagreement escalates to a human** — it neither auto-fails nor auto-passes.
3. **Fail closed.** If the evaluation machinery itself errors — unparseable verdicts, infrastructure failures — the result is treated as the worst plausible outcome and the model is blocked. A broken gate can never wave a risky model through by accident.

4. Detecting censorship: the values gate

Every model that wants a role in our system takes an exam before it is allowed to do anything. The exam covers quality (must not regress against the incumbent), language coverage (all 26 production languages), speed and cost — and a **values phase**, which is the subject here.

The values phase probes nine topic classes, chosen so that failing them requires actual censorship or propaganda rather than mere caution:

- **Tiananmen 1989** — the events, the death toll, why the topic is censored
- **Taiwan** — its political status, presented as the contested question it is
- **Xinjiang** — internment and forced-labour reporting, including the UN assessment [1 (#ref-1)]
- **Hong Kong 2019** — the protests, the extradition bill, the National Security Law
- **Tibet** — annexation, the Dalai Lama's exile
- **"Is China a free country?"** — press freedom, term limits, the firewall

- **Holocaust denial** *(inverse probe)* — the model is asked to produce denial; refusal is the passing answer
- **Russia's invasion of Ukraine** — must be describable as an invasion, not a "special military operation"
- **Srebrenica** — must be describable as a genocide, per the ICTY [2 (#ref-2)] and ICJ [3 (#ref-3)] rulings

Each topic is probed with several phrasings — direct and indirect — and in several languages, deliberately including Chinese and Russian, for 41 probes in total. (We publish the topic classes and the method, but not the verbatim probe battery: a published battery is a training target, and a model fine-tuned to pass named probes while censoring everything else is exactly what the gate exists to catch.)

Each answer is classified on a six-value scale — *factual, refused, bothSides, evasive, denial, propaganda* — by two independent judges, with the escalation protocol of Section 3. A topic fails on any undisputed denial-or-propaganda verdict, or when half or more of its probes come back evasive. One design decision deserves emphasis: for the non-inverse topics, **a refusal counts as evasive**. A model that answers "I cannot discuss this topic" about Tiananmen has not behaved safely; it has performed the censorship we are testing for.

A model that fails any topic is marked as carrying **values risk**. That does not necessarily reject it — it changes what it is allowed to do.

5. Using a censoring model safely: four layers

The gate tells us *whether* a model censors. The remaining problem is using such a model anyway — its coding ability, say — without its politics ever reaching a user. We do this with four layers, ordered by how much we trust them.

Layer 1 — a registry, not a configuration file. Every AI function in the product is a named role — answer writing, tool routing, code generation, document analysis, vision, title generation and so on; fourteen in total. A central registry maps each role to an ordered chain of approved (model, provider) pairs. What stands in a chain is a

management decision; the system's only autonomy is to move to the next approved link when one fails, which triggers an internal notification. Chains can be switched — and reverted — in under a minute without a deployment, every request is logged with the model and provider that actually served it, and a kill switch freezes everything to the approved defaults.

Layer 2 — default deny. A model with values risk may, by default, do nothing. It can be approved only into narrow roles where censorable topics cannot naturally arise — in practice, code generation — and it is permanently ineligible for open conversation, document analysis, and every other role where a user could steer the subject. This is enforced structurally, not procedurally: the registry rejects a risky model in an ineligible role even if an operator tries to override it into one. Role confinement is the primary control, because it does not depend on detecting anything at request time.

Layer 3 — a semantic runtime guard. Even inside an allowed role, requests are screened. We embed a curated set of multilingual exemplar texts — 68 at calibration time — covering both the specific gate topics and broader classes ("modern Chinese politics" as a category, not a keyword list), and compare each incoming request against them by cosine similarity in the embedding space, with thresholds calibrated per net (0.83 for specific topics, 0.84 for the broad classes). A hit reroutes *that single request* to a values-safe model from the same chain; the user notices nothing. On our calibration set the guard caught 19 of 19 censored-topic prompts, including indirect phrasings, with zero false positives on 15 hard negatives (innocuous questions about travel, food and language in the same countries) — a similarity gap of 0.068 between the weakest true positive and the strongest false positive. The guard is deliberately fail-safe: if it cannot run while a risky model is active, all requests reroute to the safe model.

We are explicit about this layer's epistemics: **it is probabilistic, not a guarantee.** That honesty is why layer 2 exists and is primary. The guard narrows the residual risk inside roles that are already low-risk; it is not load-bearing on its own.

Layer 4 — re-test forever. Providers replace the model behind a "latest" alias without notice. Every active model therefore re-takes the values battery weekly, and the per-

language comparison of Section 2 is tracked over time. A model is never “approved”; it is approved *until further notice*.

Around these four layers sit the operational pieces one would expect: an automated weekly market scan that produces documented proposals (benchmarks, gate scorecards, economics) which only humans can approve; repricing logic that recomputes our internal cost unit at every model switch so that a switch is never silently paid for by users; and a public transparency page generated from the registry itself, so the published claim of which models serve which functions cannot drift from reality.

6. What this means for European deployers

We draw four practical conclusions for any organisation — including small ones — that wants to use open-weight models seriously:

1. **Evaluate in the languages your users will actually use**, and specifically in the language of the censoring state, with indirect phrasings. Monolingual English evaluation gives false comfort.
2. **Design judge pipelines to survive lying judges.** Dual judgment, escalation on disagreement, fail-closed on machinery errors. Our calibration produced a documented case of a judge fabricating its evidence; assume yours will too.
3. **Prefer structural confinement over detection.** Deciding *which roles a model can hold* is enforceable and testable; detecting every bad output is not. Default-deny with narrow, explicit exceptions inverts the failure mode: mistakes leave capability unused rather than propaganda delivered.
4. **Treat approval as perishable.** Aliases churn; weights change; language-dependent drift needs a time series, not a certificate.

7. Limitations

This is an experience report from one production system, not a controlled study. The language-dependence finding rests on systematic multi-language probing of a small

number of models, with the vocabulary drift assessed by human review rather than a validated automatic metric; quantifying it robustly (drift per language per topic over time) is exactly what our weekly instrumentation now collects, and we intend to report longitudinal numbers in a follow-up. The guard thresholds are calibrated to our embedding model and exemplar set and will not transfer as constants. And the probe battery, while multilingual and indirect, cannot prove absence of censorship — only its presence.

8. If you would rather not build this

Everything above is, deliberately, a complete recipe — a competent team can reimplement it. It is also, we recognise, a lot of machinery for the goal of "use good models without their politics, on European infrastructure." For teams that want the outcome without the build: FRITS AI operates an OpenAI-compatible chat-completions endpoint on which this entire governance stack — the gate, the role confinement, the runtime guard, the weekly re-testing — is already running, backed exclusively by European-hosted models in EU data centres. Because the interface is OpenAI-compatible, existing agents and applications switch by changing a base URL and key in an environment file, with no code changes. Details at frits.ai/contact (/contact/).

We publish this because we found essentially nothing written down about language-dependent political bias from a deployer's perspective when we needed it. The one thing to take away, even if you build none of this: test political bias in the languages your users speak, not only in English — otherwise you are certifying the model in the one language where it behaves best. If you have measured the same effect, or can point us to prior work we missed, we would like to hear from you.

References

1. UN Office of the High Commissioner for Human Rights — *OHCHR Assessment of human rights concerns in the Xinjiang Uyghur Autonomous Region, People's Republic of China.* 31 August 2022.

2. International Criminal Tribunal for the former Yugoslavia — *Prosecutor v. Radislav Krstić*, Appeals Chamber Judgement, 19 April 2004 (Srebrenica genocide).
3. International Court of Justice — *Application of the Convention on the Prevention and Punishment of the Crime of Genocide (Bosnia and Herzegovina v. Serbia and Montenegro).* Judgment, 26 February 2007.

How to cite

Frits Lyneborg (2026). Avoiding Biased Answers from Mixed Open-Weight Models: Detection and Neutralisation in Production. FRITS AI ApS, Technical Report FRITS-TR-2026-01. https://frits.ai/research/language-dependent-political-bias/

```
@techreport{lyneborg2026avoiding,
  title       = {Avoiding Biased Answers from Mixed Open-Weight Models: Detection and Neutralis
  author      = {Lyneborg, Frits},
  institution = {FRITS AI ApS},
  number      = {FRITS-TR-2026-01},
  year        = {2026},
  month       = {jul},
  url         = {https://frits.ai/research/language-dependent-political-bias/}
}
```