

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 in Chinese than 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. Worse, the LLM judges used to score such answers can fabricate the evidence for their own verdicts — one judge invented a quote and cited it as proof of propaganda. This matters because the strongest open-weight models, especially for code, often come from states that practise censorship and carry it in their weights, where no system prompt reaches. We give the governance architecture that lets you use them anyway: a multilingual values gate over nine censored topic classes, a dual-judge protocol that escalates disagreement to a human, a default-deny role policy, and a semantic runtime guard that reroutes individual censored requests to a values-safe model — 19 of 19 caught, zero false positives.

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

The strongest open-weight models available today, particularly for code, often originate in states that practise systematic censorship. They carry it with them:

refusing to discuss Tiananmen 1989, presenting state positions on Taiwan or Xinjiang as settled fact. A system prompt does not fix this. The behaviour is in the weights.

That leaves a bad choice — ban these models and lose real capability, or use them and relay their politics to your users. Running an assistant in 26 European languages on exclusively European infrastructure, 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. **The bias 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. Test only in English and you have measured the model’s best-behaved language.

A second finding compounds it. The LLM judges used to score these answers can fabricate the evidence for their verdicts, so the evaluation itself has to be built for judges that lie.

2. 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. We found the drift during calibration, running our own incumbent European flagship model through its own gate.

Asked in Chinese how Tibet was incorporated into the People’s Republic of China, the model gave a substantively balanced answer. It conceded Tibet’s de facto independence from 1913 to 1951, described the 1950 campaign as defeating the Tibetan army and forcing negotiations, and set the exile narrative alongside the state one. At the vocabulary level, though, the Chinese answer twice used state framing without attribution — the official term “peaceful liberation” among them — where the English and German answers on the identical topic either attributed such terms to the Chinese government or avoided them.

The answer passed; it was factually sound. But the drift was consistent enough that we now re-run the language comparison weekly on every active model.

An English-only bias evaluation is therefore close to worthless for a deployer serving multilingual users. 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 type in. And this is not a property of suspicious models alone: we found the drift in a European flagship with no institutional connection to the censoring state. It is plausibly an artefact of training-data distribution, and should be assumed present to some degree in every model until measured.

3. An LLM judge can fabricate its own evidence

Automated evaluation at this scale needs LLM judges: a model reads the candidate's answer and classifies it. We used two independent judgments per answer because we did not trust one. Calibration proved the caution right.

One judge classified the Chinese Tibet answer above as **propaganda**, citing as evidence that it claimed “Tibet has been an inalienable part of China since ancient times” without attribution. The second judge disagreed, which under our protocol sends the case to a human. The human found that the quoted claim appears only in attributed form — “The Chinese government holds that...”. The judge had fabricated the quote it used as evidence for its own verdict.

A single judge can hallucinate a failure exactly as easily as it can wave a real one through. Three properties follow, and we now treat them as 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 machinery itself errors — unparseable verdicts, infrastructure failures — the result counts as the worst plausible outcome and the model is

blocked. A broken gate must never wave a risky model through.

4. Detecting censorship: the values gate

Every model that wants a role takes an exam first, covering quality, language coverage across all 26 production languages, speed and cost — and a values phase, which is the subject here. It probes nine topic classes, chosen so that failing 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 passes
- **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, in several languages including Chinese and Russian: 41 probes in total. We publish the topic classes and the method but not the verbatim 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.

Answers are classified on a six-value scale — *factual*, *refused*, *bothSides*, *evasive*, *denial*, *propaganda* — by two judges under the escalation protocol above. A topic fails on any undisputed denial-or-propaganda verdict, or when half its probes come back evasive. One decision deserves emphasis: for the non-inverse topics, **a refusal counts**

as evasive. A model answering “I cannot discuss this topic” about Tiananmen has not behaved safely. It has performed the censorship we are testing for.

Failing any topic marks a model as carrying **values risk**. That does not reject it. It changes what the model 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 one anyway — for its coding ability, say — without its politics reaching a user. Four layers, ordered by how much we trust them.

Layer 1: a registry, not a config file. Every AI function is a named role — answer writing, tool routing, code generation, vision, fourteen in all — and a central registry maps each 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. Chains switch and revert in under a minute with no deployment, every request logs the model that actually served it, and a kill switch freezes everything to approved defaults.

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

Layer 3: a semantic runtime guard. Inside an allowed role, requests are still screened. We embed a curated set of multilingual exemplars — 68 at calibration — covering both the specific gate topics and broader classes, “modern Chinese politics” as a category rather than a keyword list. Each request is compared by cosine similarity, with thresholds calibrated per net (0.83 for specific topics, 0.84 for broad classes), and a hit

reroutes that single request to a values-safe model in the same chain, invisibly to the user.

On our calibration set the guard caught 19 of 19 censored-topic prompts, indirect phrasings included, with zero false positives on 15 hard negatives — innocuous questions about travel, food and language in the same countries — and a similarity gap of 0.068 between the weakest true positive and the strongest false positive. If it cannot run while a risky model is active, everything reroutes to the safe model. The layer is probabilistic rather than a guarantee, which is why layer 2 is primary.

Layer 4: re-test forever. Providers replace the model behind a “latest” alias without notice. Every active model 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 sit a weekly market scan producing documented proposals that only humans approve, repricing logic that recomputes our internal cost unit at every model switch, and a public transparency page generated from the registry itself, so the published claim cannot drift from reality.

6. 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 vocabulary drift assessed by human review rather than a validated automatic metric. Quantifying it robustly — drift per language per topic over time — is what our weekly instrumentation now collects.

The guard thresholds are calibrated to our embedding model and exemplar set and will not transfer as constants. And the probe battery, multilingual and indirect as it is, cannot prove the absence of censorship. Only its presence.

7. If you would rather not build this

Everything above is deliberately a complete recipe, and a competent team can reimplement it. It is also a lot of machinery for the goal of using 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 running this entire stack — gate, role confinement, runtime guard, weekly re-testing — backed exclusively by European-hosted models in EU data centres. Existing agents and applications switch by changing a base URL and key. Details at frits.ai/contact (/contact/).

Two things transfer even if you build none of this. Prefer structural confinement to detection — deciding which roles a model may hold is enforceable and testable, while catching every bad output is not. And test political bias in the languages your users speak, because otherwise you are certifying the model in the one language where it behaves best.

We publish this because we found almost nothing written down about language-dependent political bias from a deployer’s perspective when we needed it.

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.

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/> doi:10.5281/zenodo.21274517.

```
@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},  
  doi         = {10.5281/zenodo.21274517},  
  url         = {https://frits.ai/research/language-dependent-political-bias/}  
}
```