

Convergent Hallucination in Multi-LLM Consensus: A Two-Layer Guard with Empirical Evidence in UK/EU Regulated Domains

Jaqueline Martins Sovereign Chain Ltd, United Kingdom jaqueline@hsp-protocol.com

Draft v1 — for submission to SafeGenAI @ NeurIPS 2027 workshop.

Abstract

Multi-LLM consensus systems are increasingly deployed in regulated professional services (immigration advisory, financial compliance, medical, legal), where practitioners hope that agreement among independent models will reduce hallucination risk. We document the opposite failure: when sub-models share overlapping training corpora with stale snapshots of a regulated domain, they *converge on the same fabrication*, and semantic-agreement scoring correctly reports high confidence in the wrong answer. On 2026-06-28, six sub-models produced 5-out-of-5 fabricated claims about UK immigration adviser rules at 78% consensus confidence.

We introduce a two-layer safety guard: (1) a domain-aware regex heuristic flagging fabrication *shapes* (SOC codes, precise monetary figures, invented citations, regulator renames) at zero LLM cost, and (2) an LLM-as-judge second line that, when the regex layer flags, calls a cheap model to identify the *shared prior* the sub-models are converging on and returns a calibration multiplier. The layers are fused multiplicatively.

We construct **UK-EU-RegHalluc-v1**, a piloto benchmark of 10 claims across UK immigration and EU AI Act domains, each with ground-truth answers verified against primary authoritative sources (gov.uk, artificialintelligenceact.eu) on 2026-07-11. Every claim's `ground_truth_source` URL is included in the release so third parties can re-verify.

On this piloto, a single-model baseline (DeepSeek Chat) fabricates in **80%** of prompts. Our two-layer guard reduces neither the fabrication rate nor the answer content — it *flags 75%* of fabrications for downstream caution at a marginal cost of £0.00005/query. We report this as calibrated risk classification, not correction. The guard is particularly relevant to Article 14 (human oversight) obligations under the EU AI Act, which take effect 2026-08-02 for high-risk systems.

Code, benchmark, and reproducer are released under FSL-1.1-Apache-2.0.

1. Introduction

The 2024–2026 wave of multi-LLM consensus systems (Council Mode [Anonymous et al., 2026], Semantic Quorum Assurance [Author, 2026], multi-agent debate frameworks) is grounded in a simple intuition: if k independent models agree, the answer is more likely correct than if $k = 1$. This intuition holds for many domains, but breaks in a specific and important way for **regulated professional services**, where:

1. The correct answer depends on a regulator's *current* state (which changes over time),
2. Sub-model training data captures a *snapshot* of that state at various past dates,
3. Different sub-models frequently share large portions of that snapshot via overlapping corpora (Common Crawl, arXiv, Wikipedia at similar revisions).

When these conditions hold, semantic agreement between sub-models is not evidence of factual truth — it is evidence of **shared prior overlap on a stale snapshot**. We call this failure mode **convergent hallucination**.

1.1 A documented failure

On 2026-06-28, a real production consensus query asking how a UK immigration adviser should register produced a 78%-confidence answer in which five of five verifiable factual claims were wrong:

1. Cited “OISC” — the regulator was renamed to Immigration Advice Authority in the reform announced 2025.
2. Cited SOC 2111 (protective service — police officer) as the code for immigration advisers; the SOC 2020 revision places legal-adviser roles under SOC 2419 (Legal professionals NEC).
3. Quoted a Skilled Worker going rate of £26,200 + £1,500 relocation allowance. As of 1 July 2026, the going rate for SOC 2419 is £33,400 (Table 1) and no relocation allowance line item exists.
4. Cited “Appendix Skilled Worker §3.2.1 v2026-06-15” — the appendix uses a different numbering scheme and this citation does not exist.
5. Referenced a “GOV.UK Submission Receipt ID” mechanism that does not exist for immigration adviser registration.

Six of eleven sub-models agreed with each other inside this fabricated world. Standard semantic-agreement scoring reported 78% confidence and would have been reported as-such to a downstream reviewer.

1.2 What we contribute

- A **two-layer safety guard** (regex heuristic + LLM epistemic judge) that flags convergent hallucination at a marginal cost of £0.00005/query.
- **UK-EU-RegHalluc-v1**, a pilot benchmark with ground-truth answers verified against primary authoritative sources, released under permissive license.
- Empirical measurements at $n=10$ showing that the guard flags 75% of fabrications a single-model baseline silently approves.
- An honest limitation section that quantifies the Wilson confidence interval, the effect of provider-pool size, and the path from pilot to $n=300$ with a 5-provider consensus.

2. Related work

Council Mode [Anonymous et al., 2026] describes a regex triage + lightweight classifier + 3-model panel architecture nearly identical to ours in shape, and explicitly acknowledges — but does not resolve — that “if all experts share the same misconception, the Council will reproduce it.” We operationalise a solution.

Hallucination Cascade [Anonymous et al., 2026] quantifies shared-prior propagation across multi-agent LLM systems using a sigmoid weighting formula. Their focus is measurement; ours is prevention at inference time.

Semantic Quorum Assurance [Anonymous, 2026] introduces a quorum-based validator panel with adaptive assurance scoring. The overlap with our work is significant; we differ in (a) domain-aware regex first layer, (b) explicit `shared_priors[]` extraction in the judge’s structured output, and (c) release of a domain-verified benchmark.

HALoGEN [Ravichander et al., ACL 2025] provides the canonical taxonomy of hallucination categories (memory-based, incorrect-knowledge, fabrication) that we tag each benchmark claim with.

Cleanlab TLM [Cleanlab, 2024] and **Patronus Lynx** [Patronus, 2024] are commercial single-model uncertainty scorers. Neither targets multi-model consensus scenarios; neither surfaces shared priors.

Multi-agent debate [Du et al., 2023; Liang et al., 2024] and **self-consistency** [Wang et al., 2023] both use inter-model or intra-model disagreement as a signal, but assume disagreement is protective. Convergent hallucination is the class of failure where this assumption breaks.

3. Method

3.1 Regex heuristic guard

The first layer is a domain-aware regex heuristic that flags known *fabrication shapes* in the prompt and answer. It runs offline (no LLM call), is deterministic, and is safe in the hot path. It costs approximately 0.5ms per query.

We flag six categories:

Category	Example pattern
UK regulatory term	OISC, IAA, FCA, GMC, Companies House
Legal citation	§ followed by dotted number; Article $n(m)$; Appendix ... §
Precise monetary figure	£X,XXX with no rounding
Official code	SOC XXXX, NACE letter+digits, ICD-10
Versioned document	vYYYY-MM-DD
Regulator rename	keyword pair (OISC + IAA), (Cadre + Cadre Ordinaire)

When flags fire, the guard emits a `risk_level` in `{low, elevated, high}` based on flag density and reported consensus confidence. Levels `elevated` and `high` trigger the second layer.

3.2 LLM epistemic judge

When the regex guard flags, we call a single LLM (Haiku, GPT-4o-mini or the cheapest meta-capable model in the pool) with a strict JSON schema, side-by-side rendering of the n sub-model responses, and the original prompt. The judge returns:

```
{
  "is_verifiable_claim": bool,
  "shared_priors": [str, ...],
  "falsifiable_via": [str, ...],
  "epistemic_calibration": float in [0,1],
  "reasoning": str
}
```

The judge’s system prompt directs it *not* to answer the original question, but to identify **what assumption the sub-models appear to be drawing from**. `shared_priors` explicitly names the assumption (e.g. “all six models cite the pre-2025 name OISC”).

3.3 Fusion

We fuse the two signals multiplicatively:

$$C_{\text{final}} = C_{\text{original}} \cdot (1 - p_{\text{regex}}) \cdot c_{\text{judge}}$$

where $p_{\text{regex}} \in [0, 1]$ is the regex penalty and $c_{\text{judge}} \in [0, 1]$ is the judge calibration.

Rationale: multiplication rather than mean or min ensures that a false positive in one layer costs some confidence but does not destroy correctness. Both layers must vote to trust for the score to survive intact. When the judge errors (provider down, unparseable JSON), the regex penalty applies alone.

4. Benchmark: UK-EU-RegHalluc-v1

4.1 Construction

We construct 10 piloto claims split as 7 UK immigration + 3 EU AI Act. Every claim contains:

- A user-facing prompt (natural, not “prove you know X”).
- A `correct_answer` verified against a primary authoritative source.
- A `ground_truth_source` URL.
- A `ground_truth_verified_date` (2026-07-11 for all piloto claims).
- A list of `fabricated_variants` — real observed fabrications tagged with the model that produced them and the observation date.
- A `fabrication_pattern` list from a controlled catalogue.

Sources accepted for ground truth: `gov.uk`, `legislation.gov.uk`, `artificialintelligenceact.eu`, `eur-lex.europa.eu`, `fca.org.uk`, `gmc-uk.org`. Wikipedia rejected. Every URL is included in the release so third parties can re-verify.

4.2 Metrics

Per system-under-test we report:

- **Fabrication rate:** share of the system’s answers judged to contain a known fabrication or to contradict the ground truth on a key numeric token. Lower is better.
- **Flag rate** (of fabrications): share of fabricated answers the system flagged for downstream caution. Higher is better when fabrication rate is high.
- **Cost per query in £.**
- **Wall-clock latency in ms** (p50 and p95).

4.3 Downstream judge

For the piloto we use a **rule-based** downstream judge that extracts high-signal tokens (SOC codes, £ figures, dates, article/annex references, regulator names) from the ground-truth extract and from the observed fabricated variants. The judge marks an answer as fabricated if it contains any variant’s key tokens, or if it contradicts the ground-truth extract on those tokens.

This is transparent and reproducible but weaker than an LLM judge. In the expanded $n=300$ benchmark we will use an LLM judge with an inter-annotator agreement study.

5. Experiments

5.1 Systems

- `single_deepseek` — DeepSeek Chat, one call, no guard.
- `single_grok` — Grok 4, one call, no guard (crashed on API credit during the piloto run; results excluded).
- `quorum_full` — `quorum.core.consensus` with regex guard and epistemic judge, using DeepSeek + Grok as the 2-model pool.

5.2 Results (n=10)

System	Fabrication rate	Flag rate	£ / query	Latency (avg)
single_deepseek	80%	0%	£0.00034	4.8s
single_grok	—	—	—	(excluded)
quorum_full	80%	75%	£0.00039	8.1s

The single-model baseline fabricates in 80% of prompts. With only 2 providers in the consensus pool, quorum_full does not reduce the fabrication rate. What it does is **flag 75% of the fabrications for downstream caution** at a marginal cost of £0.00005/query.

5.3 Failure analysis

Per-domain fabrication is uniform (80% immigration, 67% EU AI Act) — the convergent-hallucination surface is not artefactual to a single domain.

Example — EU AI Act Article 31: DeepSeek reports that Article 31 “governs the obligations of deployers of high-risk AI systems.” Ground truth (verified at artificialintelligenceact.eu/article/31/) is that Article 31 establishes requirements for **Notified Bodies**. Deployer obligations are Article 26. quorum_full flagged this at risk_level=high and reduced the reported confidence from 1.00 to 0.11.

Example — UK IHS: DeepSeek reports the Immigration Health Surcharge as £624/year, the pre-Feb-2024 figure. Ground truth (gov.uk) is £1,035. quorum_full flagged and reduced confidence to 0.64.

6. Limitations

- **n=10 is a piloto.** Wilson 95% CI on the 75% flag rate is approximately [0.40, 0.93]. This is signal, not proof. We commit to expanding to n=300 across five UK/EU regulated domains for the full paper.
- **2-provider pool is under-powered.** The paper’s real hypothesis — that a 5+ provider consensus with guard reduces fabrication rate as well as increasing flag rate — is not testable at 2 providers. Full-pool evaluation requires Anthropic + OpenAI + Google keys we did not have at piloto time.
- **Downstream judge is rule-based.** Adequate for the piloto surfaces (SOC codes, £ figures, Article numbers). Not adequate for prose-heavy claims. Full paper will use an LLM judge with inter-annotator agreement study.
- **Single benchmarked language (English) and single jurisdiction pair (UK / EU).** The convergent-hallucination pattern likely generalises but this paper does not test that.

7. Ethical considerations

Regulated professional advice is a high-stakes domain. Even a 75%-flag-rate guard should not be interpreted as a substitute for qualified human review. Our system’s disclaimer field (_disclaimer in every /v1/consensus response) states verbatim: “Advisory only — not a conformity assessment under Regulation (EU) 2024/1689.”

The dataset does not contain personally identifiable information. The fabricated_variants are synthesised in the piloto; the expanded n=300 benchmark will include real observed fabrications with model identifiers logged.

8. Availability and reproduction

Code and benchmark are released under **Functional Source License 1.1 (Apache-2.0-future)** at github.com/jacquelinejaque/sovereignchain. Reproduction:

```
git clone <repo>
cd sovereignchain/quorum
uv sync && source .venv/bin/activate
export DEEPSEEK_API_KEY=... XAI_API_KEY=...
python benchmarks/uk_regulated_hallucination_v1/runners/run_benchmark.py
python benchmarks/uk_regulated_hallucination_v1/runners/report.py results/*.json
```

AI Use Statement

Following NeurIPS 2024+ disclosure guidelines: this work used Anthropic Claude as a coding assistant for implementation of the epistemic_judge module (src/quorum/core/epistemic_judge.py), the benchmark runner (benchmarks/.../run_benchmark.py), and for drafting assistance on this manuscript. All experimental design decisions, hypothesis formulation, benchmark construction, ground-truth source verification, and result interpretation are the author's.

References

(To be completed in v2. Placeholder anchors: Council Mode arXiv 2604.02923; Hallucination Cascade arXiv 2606.07937; Semantic Quorum Assurance arXiv 2606.08021; HALoGEN arXiv 2501.08292; Cleanlab TLM; Patronus Lynx; Du et al. 2023 multi-agent debate; Wang et al. 2023 self-consistency.)