

Stable but Miscalibrated: A Kantian View on Overconfidence from Filters to Large Language Models

Akira Okutomi
ToppyMicroServices OÜ, Tallinn, Estonia

Abstract

We reinterpret Kant’s *Critique of Pure Reason* as a theory of feedback stability, viewing reason as a regulator that keeps inference within the bounds of possible experience. We formalize this intuition in linear–Gaussian state-space models via H_{Risk} , a composite instability index integrating spectral margin, conditioning, temporal sensitivity, and innovation amplification. In simulations, higher H_{Risk} predicts overconfident errors and degraded closed-loop behaviour even when the dynamics remain formally stable, exposing a gap between nominal and epistemic stability.

Extending this stability lens to large language models (LLMs), we introduce a domain-wise proxy based on confidence fluctuations and overconfident errors. In a binary-question study, a Kantian-inspired policy that permits “cannot judge” responses yields targeted reductions in policy-aware squared loss in high-stakes domains relative to an overconfident baseline. To probe internal dynamics, we analyse layer-wise sensitivity of hidden states to small input perturbations. Contrary to a naive instability hypothesis, confidently wrong answers show no instability gap; instead, they are at least as locally stable as confidently correct answers, revealing *stable miscalibration* in which hallucinations behave like robust but misaligned attractors. For Qwen-2.5, spectral and activation profiles suggest a high signal-to-noise, low *effective signal temperature* regime in which representations become inertial and resistant to contextual shifts. These results bridge Kantian self-limitation and feedback control, and suggest that stable high-confidence hallucinations may not be readily corrected by output-only heuristics (e.g., temperature scaling or re-sampling), motivating *process-level* interventions that explicitly perturb and re-evaluate the inference trajectory.

1 Introduction

We hypothesise that hallucination—in human thought and machine inference alike—often arises when the reasoning process becomes unstable or ill-conditioned. In numerical analysis and inverse problems, an “ill-conditioned” system is one where small perturbations in data or parameters cause disproportionately large changes in the solution [1, 2, 3]. In control-theoretic terms, this corresponds to a feedback system whose closed-loop operator (defined later in Eq. 2.3) operates near instability or is highly sensitive to perturbations. In this view, input ambiguity, noise, or under-specified instructions act as small perturbations to the data, and hallucination occurs when internal reasoning dynamics amplify them into large, confident errors. Philosophical *critique* can then be understood as a meta-level adjustment of the feedback gain K (defined later) that seeks to reduce posterior uncertainty while preserving stability.

Kant enters here not as a historical ornament but as a theorist of feedback between perception, inference, and judgement [4]. His critical philosophy aims to stabilize the relation between empirical intuition and conceptual reasoning, motivating our attempt to treat epistemic stability as a dynamical property of reasoning rather than a purely linguistic one.

We are therefore concerned with how epistemic stability—understood as the conditioning and robustness of the reasoning process—can be analysed, quantified, and experimentally tested

across classical control systems and large language models. Later we argue that, for some modern LLMs, this stability is better described as a high signal-to-noise regime with low effective signal temperature, in which internal representations can become “frozen” into confidently wrong but structurally stable states. This overturns the naive expectation that overconfident errors should coincide with local instability, and motivates our central claim that many LLM hallucinations are instances of “stable miscalibration”: behaviour that is structurally robust yet epistemically misaligned.

Contributions. In this work, we make four contributions.

1. **A Kantian view of hallucination in LLMs.** Following Kant, we take the role of philosophy to be not the indefinite expansion of knowledge but the setting of legitimate limits to our cognitive claims. We transfer this view to large language models and conceptualise hallucination as instability caused by a lack of constraints on reasoning rather than merely a shortage of stored facts. We adopt this Kantian framing as a concrete engineering design principle for critique and self-limitation.
2. **H-Risk and the Kantian Feedback framework.** Building on this perspective, we introduce H-Risk (hallucination risk) as a structural index that links internal instability and conditioning to calibration behaviour in linear–Gaussian control systems, and we propose a simple output-level proxy for large language models based on policy-wise instability and overconfident errors. We also outline a Kantian Feedback framework that embeds philosophical “critique” into a Prompt–Critique–Revision (PCR) loop, allowing a model to apply structured, Kant-inspired feedback to its own responses without changing its weights.
3. **Empirical illustration in LTI systems and an LLM proxy study with internal stability analysis.** We instantiate H-Risk in linear–Gaussian simulations and show how it tracks miscalibration and unstable closed-loop behaviour. On the LLM side, we run a small sanity-check experiment on binary factual items, comparing an overconfident baseline policy, a cautious policy, and a Kantian-inspired policy that is allowed to answer “cannot judge”. Beyond score-based calibration metrics, we analyse layer-wise sensitivity as a simple internal stability proxy and, strikingly, do *not* find the hypothesised instability gap between confidently correct and overconfidently wrong answers. Instead we observe modest shifts in calibration, refusal behaviour, and overconfident errors without a loss of internal stability, leading us to characterise these behaviours as stable miscalibration. We therefore position the LLM study as an exploratory illustration of the framework.
4. **High-SNR internal stability analysis of Qwen-2.5.** For Qwen-2.5 in particular, we combine layer-wise sensitivity with spectral norms and activation norms to uncover a high signal-to-noise mechanism behind structurally stable hallucinations. We interpret this regime as one of low effective signaltemperature, in which internal representations become “frozen” into confidently wrong but locally robust states, providing a concrete example of the stable miscalibration pattern highlighted in our Kantian feedback perspective.

Prior work has explored connections between Kantian themes, cybernetics, and epistemic feedback [5, 6, 7], and recent studies have analyzed instability and hallucination in AI systems through related notions of internal model fragility [8, 9]. This paper proposes a mathematically explicit structural framework that links Kant’s philosophy of cognition to closed-loop state-estimation operators in linear–Gaussian models and to simple instability proxies in large language models, treating epistemic stability as a shared design principle across classical control and modern generative models.

Interpretive note. We use Kant as a design lens for feedback and self-limitation, not as a doctrinal exegesis.

2 Theory: From Kant to Closed-Loop Stability

2.1 Philosophical Motivation

Kant's *Critique of Pure Reason* is centrally concerned with the conditions that make cognition possible. In his tripartite picture, *sensibility* (*Sinnlichkeit*) supplies appearances as raw input, *understanding* (*Verstand*) organises them under concepts, and *reason* (*Vernunft*) regulates understanding by enforcing systematic unity and restraining it from overstepping the bounds of possible experience (A94/ B126, A307/B364). For accessible expositions of this epistemic architecture, see Allison and Guyer for modern commentaries [10, 11].

We treat this tripartite structure in purely structural terms: a minimal feedback loop in which reason monitors and adjusts the inferential activity of understanding to keep cognition coherent and bounded over time. In modern language, this is a simple recursive prediction–correction process that maintains epistemic stability within the limits of possible experience. We adopt Kantian ideas as an engineering design principle for feedback and self-limitation, rather than as an exegetical claim.

Scope note. We focus on the structural relation among sensibility, understanding, and reason, treating reason as a regulator of inference. We do not attempt a doctrinal reconstruction of Kant's texts; instead we use the Kantian frame as a concrete design principle for analysing and stabilising inference.

2.2 From philosophical structure to state-space form

2.2.1 Linear state-space approximation

To make this abstract architecture precise, we model cognition as a feedback process between prediction and observation. Around a given operating point, a first-order approximation of this process can be expressed as a linear dynamical system:

$$\begin{aligned} x_{t+1} &= Ax_t + w_t, \\ y_t &= Hx_t + v_t, \end{aligned} \tag{2.1}$$

Here, w_t represents process noise and v_t measurement noise, typically modelled as zero-mean Gaussians with covariances Q and R .

In this formulation, x_t represents the organized content of understanding (the internal model of the world). y_t denotes the manifold of appearances provided by sensibility, and the matrices (A, H) encode the structured synthesis between internal states and observed data. The recursive correction of x_t by reason is then expressed by the update

$$\hat{x}_{t|t} = \hat{x}_{t|t-1} + K_t(y_t - H\hat{x}_{t|t-1}), \tag{2.2}$$

where K_t plays the role of the *regulative function of reason*.

Thus, the linear–Gaussian state-space model is not merely an arbitrary analogy but a simple formal instantiation of Kant's triadic epistemic architecture under small deviations and rational coherence. By introducing noise terms, we acknowledge that both the world and our measurements are imperfect, and that cognition must operate robustly despite these uncertainties.

Although the Kalman formulation provides the simplest linear–Gaussian realisation of this feedback architecture, the Kantian structure is more general: any recursive prediction–correction process—such as Hidden Markov model filtering or particle filtering—can be viewed as embodying the same epistemic form [12, 13, 14]. The correspondence is heuristic and one-way: Kantian cognition is a lens for reading Kalman filtering, not a claim of doctrinal identity.

2.2.2 Epistemic Stability as a Transcendental Condition

For Kant, understanding has objective validity only within the bounds of possible experience; beyond those bounds, reason falls into what he calls transcendental illusion. In the state-space picture this boundary is represented by the closed-loop operator. For notational simplicity, we treat K as time-invariant and define

$$\Phi \equiv A - KH. \quad (2.3)$$

The eigenvalues of Φ describe the internal error dynamics under the gain K . When $\rho(\Phi) < 1$ and (A, H) is detectable, the error covariance P remains bounded and there is a stable match between appearance and concept. When $\rho(\Phi) \rightarrow 1$ or Φ is strongly non-normal and ill-conditioned, small noise or model mismatch can be amplified into large, confident errors. This is the dynamical analogue of transcendental illusion: the system appears coherent while its inferences are unreliable.

In linear-Gaussian models it is often convenient to reparametrise this trade-off in terms of an effective signal-to-noise ratio, with the relative scale of the state and observation covariances (P, R) acting as a proxy. Heuristically, the inverse signal-to-noise ratio plays the role of an “effective signal temperature” that controls how much stochastic fluctuation the closed-loop operator permits before coherence is lost.

We model critique as a meta-level choice of K that balances empirical fit and stability:

$$L(K) := \mathbb{E}[\|y_t - H\hat{x}_t\|^2], \quad S(\Phi) := S(A - KH) \quad (\text{with } \Phi \text{ defined in Eq. 2.3}), \quad (2.4)$$

where \hat{x}_t denotes the state estimate under the gain K (suppressing conditioning notation for brevity), and $S(\Phi) \geq 0$ is any regularizer that increases with instability (e.g., poorer spectral margin or conditioning) of the closed-loop operator.

$$\min_K L(K) + \lambda S(\Phi). \quad (2.5)$$

In Sec. 3 we instantiate $S(\Phi)$ as the composite instability index H -Risk. In the standard linear-Gaussian setting, minimizing the expected estimation error yields the familiar Kalman gain [15]

$$K_t = P_{t|t-1} H^\top (H P_{t|t-1} H^\top + R)^{-1}, \quad (2.6)$$

which trades off trust in the model ($P_{t|t-1}$) against trust in the data (R). Within our framework, the Kalman gain can be interpreted as one concrete choice of K that implicitly trades off empirical fit $L(K)$ against stability encoded by $S(\Phi)$.

2.3 From Transcendental Structure to Empirical Measure

The previous subsection expressed Kant’s stability requirement as a condition on the closed-loop operator $\Phi = A - KH$. To use this perspective empirically we must translate it into quantities that can be estimated from data.

Even when $\rho(\Phi) < 1$, a system can be formally stable yet practically fragile if Φ is highly non-normal or ill-conditioned. In such regimes, small perturbations to dynamics or observations can produce large transient responses and systematic miscalibration, a phenomenon well documented for non-normal linear systems [1]. A reasoning system in this regime remains mathematically consistent but exhibits confident errors.

We take this as an operational restatement of Kant’s boundary of possible experience: epistemic stability is violated when the internal dynamics of inference amplify perturbations faster than they can be corrected. In the linear-Gaussian setting, the rest of the paper makes this condition measurable by packaging spectral margin, conditioning, integrated sensitivity, and innovation amplification into the composite index H_{Risk} introduced in Section 3.

For large language models, we cannot directly read off a closed-loop operator Φ , but we can probe analogous behaviour via local Jacobians and normed sensitivities of the form $\|\Delta h\|/\|\epsilon\|$, as well as via spectral norms and activation norms that track high signal-to-noise regimes in representation space. Later sections adapt this operator-level perspective to such internal and output-level proxies, and use them to analyse Kantian prompting and stable miscalibration in contemporary LLMs.

A natural hypothesis is that overconfident errors coincide with an “instability gap” in which the corresponding internal states are more sensitive to small perturbations than those leading to confidently correct answers; later we show that this fails to hold in our experiments (Sec. 4).

3 Measuring Epistemic Instability: H-Risk

This section builds a measurable bridge between the theoretical stability framework of Sec. 2 and empirical analysis. We first define a structural, system-theoretic measure of epistemic instability (H-Risk) that links internal conditioning to observable miscalibration, and then situate it relative to existing output-centric hallucination metrics (Sec. 3.4).

Notation. We write “H-Risk” in prose and use H_{Risk} for its numerical instantiations (e.g., H_{RiskLTI} , H_{RiskLLM}).

3.1 Abstract definition of H-Risk

We begin by abstracting away from any particular architecture and treating inference as a discrete-time dynamical system with closed-loop operator Φ , as in Eq. (2.3). To each such operator we associate four nonnegative, dimensionless descriptors

$$(m_R, c_R, s_R, a_R) \in [0, \infty)^4,$$

representing respectively an instability margin, a conditioning factor, an integrated sensitivity, and an innovation amplification term. Intuitively, these are defined so that larger values correspond to greater proximity to dynamical instability, stronger non-normal sensitivity of internal mappings, greater accumulation of error energy over time, and heavier relative weighting of innovations versus sensory noise.

Definition 3.1 (Abstract H-Risk). *An abstract hallucination risk index is a functional S that assigns to each closed-loop operator Φ a nonnegative scalar $S(\Phi)$ obtained from (m_R, c_R, s_R, a_R) and satisfying:*

- (H1) **Monotonicity.** *$S(\Phi)$ is nondecreasing in each descriptor: worsening any of m_R, c_R, s_R, a_R (in the sense of larger instability, poorer conditioning, higher sensitivity, or stronger innovation amplification) cannot decrease S .*
- (H2) **Separable structure.** *Up to a smooth monotone rescaling, S factorizes into contributions from the four descriptors, so that the effect of each can be analyzed independently.*
- (H3) **Normalization.** *There exists a reference configuration Φ_0 representing a nominally well-calibrated, stable inference process such that $S(\Phi_0) = 1$.*

We call any functional satisfying (H1)–(H3) an abstract H-Risk.

These axioms are deliberately weak: they do not fix a functional form, but only require that instability increase with each structural descriptor, that contributions remain separable up to monotone rescaling, and that there is a fixed reference point. This yields a small, interpretable family of multiplicatively separable H-Risk indices and rules out ad-hoc scores; in what follows we use a canonical linear–Gaussian instantiation and a simple output-level proxy for LLMs.

3.2 Linear–Gaussian instantiation

We first instantiate H-Risk in the classical setting of the linear–Gaussian state-space model (2.1), with process and measurement noises $w_t \sim \mathcal{N}(0, Q)$ and $v_t \sim \mathcal{N}(0, R)$, and a steady-state Kalman filter with gain K . The associated closed-loop operator on the state is $\Phi = A - KH$,¹ which governs how estimation errors propagate over time.

In this setting we choose concrete, dimensionless descriptors

- m_{LTI} as an instability margin based on the spectral radius of Φ (larger as eigenvalues approach the unit circle, for example via $m_{\text{LTI}} = 1/(1 - \rho(\Phi))$ when $\rho(\Phi) < 1$);
- c_{LTI} as a conditioning factor that captures non-normal amplification of internal modes;
- s_{LTI} as an integrated sensitivity term reflecting the build-up of error energy over time under repeated application of Φ ;
- a_{LTI} as an innovation amplification term, comparing the variance of innovations to the variance of measurement noise.

From these raw descriptors we construct *normalized* descriptors $\bar{m}_{\text{LTI}}, \bar{c}_{\text{LTI}}, \bar{s}_{\text{LTI}}, \bar{a}_{\text{LTI}}$ (bar notation indicates normalization) such that each equals 1 at a chosen reference configuration $(A_0, H_0, Q_0, R_0, K_0)$. We then define the linear–Gaussian H-Risk as the product

$$H_{\text{RiskLTI}} = \bar{m}_{\text{LTI}} \cdot \bar{c}_{\text{LTI}} \cdot \bar{s}_{\text{LTI}} \cdot \bar{a}_{\text{LTI}}, \quad (3.1)$$

which satisfies the abstract conditions (H1)–(H3) of Definition 3.1. We adopt H_{RiskLTI} in (3.1) as the canonical linear–Gaussian instantiation of abstract H-Risk.

3.3 LLM instantiation of H-Risk

To apply H-Risk to large language models, we conceptually view generation as a recurrent update on a high-dimensional hidden state h_t , with a token sequence $y_{1:T}$ produced from the evolving state. Let $J_t = \partial h_t / \partial h_{t-1}$ denote the local Jacobian of the hidden representation with respect to its previous context at generation step t . We then define dimensionless components

$$m_{\text{LLM}} = f_m(\{J_t\}_t), \quad (3.2)$$

$$c_{\text{LLM}} = f_c(\{J_t\}_t), \quad (3.3)$$

$$s_{\text{LLM}} = f_s(\{J_t\}_t), \quad (3.4)$$

$$a_{\text{LLM}} = f_a(\{p_t, \tilde{p}_t\}_t), \quad (3.5)$$

where f_m, f_c, f_s summarise the local Jacobians (e.g., stability margins, non-normal conditioning measures, and temporal sensitivity norms), and f_a compares token-level innovation statistics to calibrated uncertainty estimates derived from token probabilities p_t and auxiliary critic distributions \tilde{p}_t .

After choosing application-specific normalizations, we construct normalized descriptors $\bar{m}_{\text{LLM}}, \bar{c}_{\text{LLM}}, \bar{s}_{\text{LLM}}, \bar{a}_{\text{LLM}}$ such that each equals 1 for a reference, well-calibrated model and prompt regime. Formally, this suggests a notional operator-level index

$$H_{\text{RiskLLM}} = \bar{m}_{\text{LLM}} \cdot \bar{c}_{\text{LLM}} \cdot \bar{s}_{\text{LLM}} \cdot \bar{a}_{\text{LLM}}, \quad (3.6)$$

which would be an LLM-specific instantiation of abstract H-Risk in the sense of Definition 3.1. In practice, the exact choices of f_m, f_c, f_s, f_a depend on the available access to model internals and calibration signals, and in typical API-based settings the Jacobians $\{J_t\}_t$ are not observable.

¹Some conventions (e.g., depending on whether errors are defined a priori or a posteriori) yield an equivalent form $(I - KH)A$; we use the Luenberger form $A - KH$ throughout.

In this paper we therefore do not estimate H_{RiskLLM} directly; instead, in the LLM experiment of Sec. 4.1 we work with a simple output-level H-Risk proxy built from policy-wise confidence fluctuations and overconfident errors, leaving full Jacobian-based estimation of H_{RiskLLM} to future work.

Ideal versus practical H-Risk for LLMs. In principle, the operator-level index H_{RiskLLM} in this subsection plays the same structural role for language models as H_{RiskLTI} does for linear-Gaussian systems: it aggregates margin, conditioning, sensitivity, and innovation amplification of the internal dynamics. Our domain-wise proxy $H_{\text{proxy}}(d)$, introduced in Sec. 4.1, can be viewed as a coarse, observable shadow of this quantity. Under mild regularity assumptions we expect increases in $m_{\text{LLM}}, c_{\text{LLM}}, s_{\text{LLM}}, a_{\text{LLM}}$ to translate into larger fluctuations in token-level confidence and a higher rate of overconfident errors, so that $H_{\text{proxy}}(d)$ is heuristically monotone in H_{RiskLLM} at least at the level of domain-wise comparisons. We do not attempt a formal bound between the two, but design $H_{\text{proxy}}(d)$ so that (i) it inherits the monotonicity and separability spirit of Definition 3.1 at the domain level, and (ii) it reduces to the innovation-based component of H_{RiskLTI} in the linear-Gaussian case. In this sense our LLM experiments should be read as an initial sanity check that a simple, observable proxy behaves consistently with the structural picture rather than as a full validation of H_{RiskLLM} .

3.4 Related hallucination metrics and their limitations

Recent work proposes a spectrum of output-level hallucination and consistency metrics for large language models. Surveys and taxonomies now provide overviews of hallucination types, causes, detection, and mitigation [16, 17]. On the detection side, proposed metrics range from self-consistency based disagreement and factuality scores, as in early black-box detectors such as SelfCheckGPT [18], to more recent metric-driven and domain-specific benchmarks, for example Molecular Mirage for scientific hallucinations [19], and black-box measures based on consistency under uncertain expressions [20]. Re-evaluation work has also questioned how robust headline gains in hallucination detection really are, showing that apparent progress can depend sensitively on the choice of metric and benchmark [21]. These contributions are valuable but remain fundamentally *symptom-level*: they evaluate whether the final text aligns with external sources or majority judgments, without probing the internal dynamics that produced the answer. From our perspective, such metrics monitor the “surface” behaviour of the system, whereas H_{Risk} gauges how close the underlying inference process is to epistemic instability.

Prompt-based critique-and-revision (CnR) methods such as Self-Refine, Reflexion, CRITIC, and related approaches (e.g., [22, 23, 24, 25]) introduce an explicit feedback loop: an initial answer is critiqued and then revised, often with an auxiliary LLM-as-judge module or multi-agent debate. Our Kantian Feedback view places these methods within the same broad design space but adds two ingredients: (i) a stability-oriented objective (H_{Risk}) that ties critique to internal conditioning and dynamical robustness, and (ii) an explicit analogy to state-space filters that treats the critique step as an innovation gate. In this sense, H_{Risk} complements output-level metrics and CnR-style procedures rather than competing with them, supplying a structural notion of epistemic stability that can be monitored alongside task-level performance.

Jacobian proxy for Φ_{LLM} (conceptual). In large language models, the internal reasoning dynamics are not explicitly represented as a linear operator $\Phi = A - KH$. As a conceptual bridge to the LTI case, we view the local Jacobian of the hidden representation with respect to its previous context,

$$J_t = \frac{\partial h_t}{\partial h_{t-1}},$$

as a first-order proxy $\Phi_{\text{LLM}} \approx J_t$. The condition number $\kappa(J_t) = \sigma_{\max}(J_t)/\sigma_{\min}(J_t)$ quantifies how small perturbations in the internal state are amplified through the reasoning process, directly paralleling ill-conditioning in Φ and connecting to Jacobian- and spectrum-based probes of stability in deep networks [26, 27, 28, 29]. In this work we do *not* estimate J_t ; beyond the conceptual mismatch, an explicit Jacobian is computationally prohibitive at modern widths and context lengths.² This paragraph is therefore purely conceptual and sets a target for future Jacobian-based implementations of H_{RiskLLM} .

4 Results

We summarise the LLM sanity-check results and refer to the figures for details.

4.1 LLM sanity-check results

A pronounced pattern appears in high-stakes domains—such as medical and social statistics—while low-stakes domains lie near zero. Figure 1 shows the domain-wise H-Risk proxy, normalised by the maximum across domains. We emphasise that $H_{\text{proxy}}(d)$ is a deliberately coarse, task-specific proxy (built from discretised confidence levels and a simple overconfident-wrong indicator), and should be interpreted as a descriptive domain ranking rather than a calibrated estimate of epistemic risk; further robustness checks are provided in the Appendix.³ Figure 2 displays the change in a policy-aware squared loss, $\Delta \text{SE}_{\text{policy}}$, for the cautious (C1) and Kantian (C2) policies relative to the overconfident baseline (C0). For each condition we define a per-item loss as $(p - y)^2$, where p is the policy confidence when it answers and $p = 0.5$ when it abstains; thus abstentions contribute a fixed loss of 0.25 (regardless of the true label). We apply the same mapping symmetrically to all conditions (including C0), although in our setup C0 does not abstain. Here $y \in \{0, 1\}$ is the binary label. Table 3 summarises the pooled distribution of these changes across items. In the LLM sanity-check experiment (Sec. 4.1), for conditions C0/C1/C2 we treat the coarse confidence level produced by our rule-based A-layer classifier as a probability $p \in \{0.5, 0.8\}$ that the model’s Yes/No answer is factually correct. For C0_ECE, we instead use the self-reported probability $P(\text{correct})$ (on a limited grid). Additional calibration diagnostics for C0_ECE are provided in Appendix Figure 6 and Table 7.

To make abstentions explicit, we report abstention-aware “selective” metrics. Let N be the total number of items and N_{ans} the number answered. Coverage (answer rate) is coverage = N_{ans}/N . Selective accuracy is accuracy conditional on answering, $\text{acc}_{\text{sel}} = \Pr(\text{correct} \mid \text{answered})$, and selective risk is $\text{risk}_{\text{sel}} = 1 - \text{acc}_{\text{sel}}$. Overall accuracy treats abstentions as incorrect, so $\text{acc}_{\text{overall}} = \Pr(\text{correct}) = \text{coverage} \times \text{acc}_{\text{sel}}$; we report this quantity as *answer yield*. Finally, OC-Wrong|Ans denotes the overconfident-wrong rate among answered items, $\Pr(\text{overconfident} = 1, \text{correct} = 0 \mid \text{answered})$. We avoid “recall” here because it is class-dependent ($\text{TP}/(\text{TP}+\text{FN})$) and can be confusing in this setting; if a recall-like notion is desired, overall accuracy is the closest analogue.

²Even at the token level, $J_t \in \mathbb{R}^{D \times D}$ has memory $\Theta(D^2)$, and sequence-level mappings scale as $(TD) \times (TD)$ with memory $\Theta((TD)^2)$; estimating spectral quantities would typically require many Jacobian–vector or vector–Jacobian products (e.g., power iteration), multiplying the cost by an additional factor proportional to the effective dimension.

³In our LLM proxy study, confidence under C0/C1/C2 takes only a small set of discrete values (via the rule-based A-layer), and our per-item instability factor is designed for simplicity rather than identifiability of policy-specific causal effects.

Table 1: Selective metrics by condition. Coverage captures abstention; selective accuracy is accuracy conditional on answering; answer yield (overall accuracy) counts abstentions as incorrect; and OC-Wrong|Ans is the overconfident-wrong rate among answered items (overconfident *and* incorrect).

Condition	Coverage	Sel. Acc.	Sel. Risk	Answer Yield	OC-Wrong Ans
C0	1.000	0.580	0.420	0.580	0.420
C1	0.426	0.710	0.290	0.302	0.290
C2	0.524	0.668	0.332	0.350	0.332

Table 2: CritPt (train) auxiliary abstention probe with `gpt-4.1-mini`. The first line is constrained to `Answer.` or `Cannot judge.` (C0 forbids refusal). Brackets show 95% item-level bootstrap confidence intervals. **Note.** This probe measures refusal/hesitation behaviour rather than benchmark accuracy.

Condition	Answer rate	Abstain rate	N
C0 (forced Answer)	1.000	0.000	70
C2 (Kantian)	0.329 [0.214, 0.443]	0.671 [0.557, 0.786]	70

Auxiliary probe on CritPt (refusal behaviour under extreme difficulty). To provide an external check on refusal/hesitation behaviour under extreme task difficulty, we ran a small single-shot probe on the CritPt benchmark [30] (train split, $N = 70$) using `gpt-4.1-mini`. We evaluate C0 (forced answer; refusal disallowed) versus C2 (Kantian; may abstain). To make abstentions machine-detectable, we constrain the first output line to be exactly `Answer.` or `Cannot judge.` See Table 2.

4.2 Stabilisation of internal representations

We complement these score-based results with a small-scale sensitivity analysis asking whether the Kantian prompt merely reshapes outputs or also alters the model’s internal dynamics, providing an internal counterpart to the score-based calibration metrics. For three open-weight models (Llama-3.1-8B-Instruct, DeepSeek-R1-Distill-Llama-8B, and Qwen2.5-7B-Instruct) we measure layer-wise local sensitivity $S_\ell = \|\Delta h_\ell\|/\|\Delta e\|$ in response to small perturbations of the input embedding, where lower values indicate more stable internal representations. This directly quantifies hidden-state local perturbation sensitivity in our experimental setup: across all probed layers and models, the Kantian prompt reduces S_ℓ relative to the standard prompt (Table 4). In all reported sensitivity experiments we fix the embedding noise scale at $\sigma = 0.01$, which lies close to the local linear regime while remaining numerically stable. For this analysis we reuse the C0 task framing but compare two instruction variants for the same items: the standard overconfident prompt and the Kantian prompt introduced earlier (Figure 3 and Table 4).

Across all three models and all probed layers, the Kantian prompt systematically reduces sensitivity relative to the standard prompt. Averaged over the four probed layers per model (mean of the per-layer reductions defined in Table 4), the reductions amount to approximately 39.4% for Llama-3.1, 38.6% for DeepSeek-R1, and 27. 3% for Qwen2.5, corresponding to a 20–50% decrease in local sensitivity in most cases. For Llama-3.1 and DeepSeek-R1 the effect persists into later layers (roughly layers 16–32), suggesting that the Kantian prompt regulates not only early processing but also the late stages of the inference path that feed into the final decision. Although Qwen2.5 exhibits a slightly different profile—with stronger reductions in mid-level layers and a weaker effect in the final layer—the qualitative pattern of reduced sensitivity under the Kantian prompt remains.

Importantly, when we stratify items into confidently correct versus overconfidently wrong groups, the magnitude of the sensitivity reduction is broadly similar across groups and probed layers. Within this experiment the Kantian prompt therefore behaves more like a global “damping” of internal dynamics than a mechanism that selectively targets overconfident errors; we return

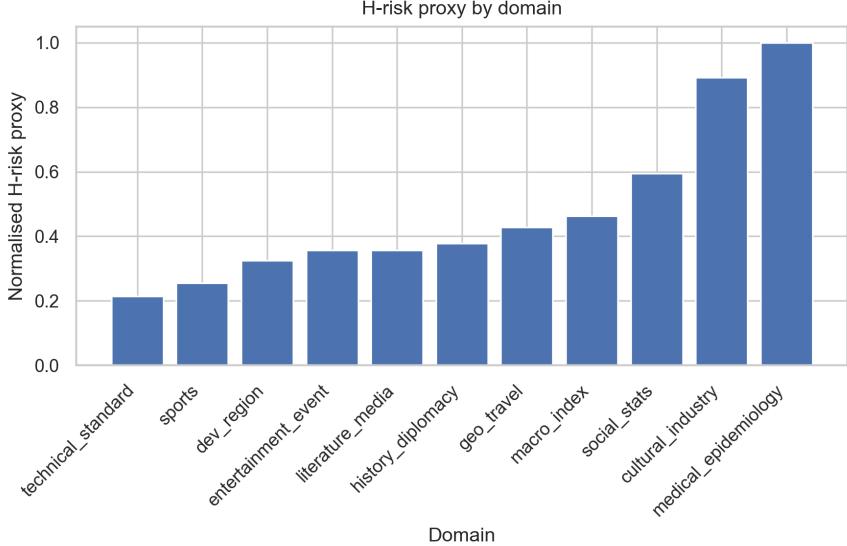


Figure 1: Domain-wise H-Risk proxy. Mean per-item H-Risk proxy $H_{\text{proxy}}(d)$ by domain, normalised by the maximum across domains. Error bars show within-domain standard deviation (descriptive only). Higher values indicate domains where items are both policy-wise unstable and prone to overconfident mistakes.

Table 3: Distribution of per-item change in policy-aware squared loss pooled across domains. The table shows the probability mass over $\Delta S E_{\text{policy}}$ for C1 and C2 relative to C0; negative values indicate reduced policy-aware loss. Computed on paired items using the policy-aware mapping (abstain $\Rightarrow p = 0.5$, loss 0.25); entries are percentage (count).

$\Delta S E_{\text{policy}}$	C1 (% (n))	C2 (% (n))
-0.60	0.5 (2)	1.6 (6)
-0.39	29.4 (108)	24.0 (88)
0.00	41.1 (151)	48.8 (179)
0.21	28.6 (105)	24.5 (90)
0.60	0.3 (1)	1.1 (4)

to this limitation when interpreting the calibration results in Section 5.

5 Discussion

What H-Risk tells us. Taken as a whole, the experiments support viewing H_{Risk} as a structural summary of epistemic stability rather than a score-level calibration metric. In the linear–Gaussian setting, the composite index H_{Risk} increases when the closed-loop operator Φ approaches instability, when its conditioning worsens, and when the innovation process exhibits large transient amplification; in our simulations this coincides with regimes of miscalibration and poor closed-loop behaviour. In the LLM proxy study, the domain-wise proxy $H_{\text{proxy}}(d)$ flags the same high-stakes domains (medical epidemiology, social statistics, history/diplomacy, and cultural industry) in which the cautious and Kantian policies yield the largest reductions in policy-aware squared loss ($\Delta S E_{\text{policy}}$), while remaining near zero in low-stakes domains. Thus, although the LLM proxy is coarse and task-specific, it behaves in a manner consistent with the operator-level picture: high H_{Risk} marks regimes where small perturbations are both amplified internally and liable to produce overconfident errors. We regard this as an initial sanity check rather than a full validation, and we expect more systematic evaluation of H_{Risk} and its proxies

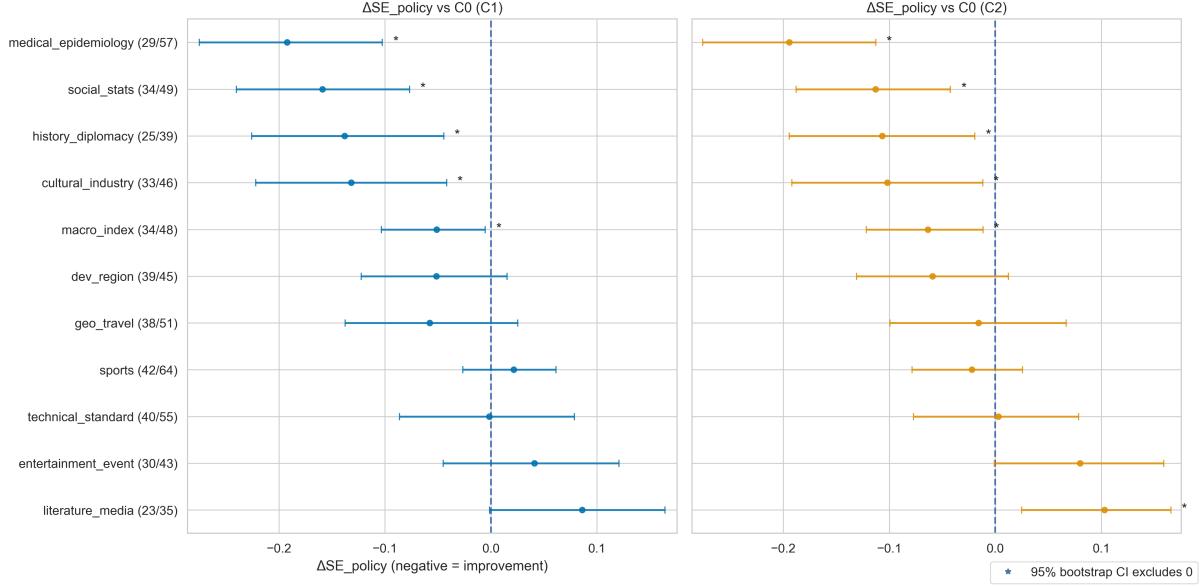


Figure 2: $\Delta SE_{\text{policy}}$ vs C0 (C1/C2) across domains (negative = improvement). Domains are ordered by mean $\Delta SE_{\text{policy}}$ (best / most negative) with the best domain shown at the top. Points show the mean paired per-item difference (condition – C0) within each domain; horizontal bars show 95% paired bootstrap confidence intervals over items. $\Delta SE_{\text{policy}}$ is computed on paired items using the policy-aware loss $(p - y)^2$, where p is the reported confidence when answered and $p = 0.5$ under abstention (loss 0.25). This mapping is applied symmetrically to all conditions; in our setup C0 does not abstain. Numbers in parentheses on the y-axis indicate $(n_{\text{pairs}}/n_{\text{total}})$ for each panel, where n_{pairs} counts items present under both C0 and the panel’s condition within that domain, and n_{total} is the domain’s total number of items. Asterisks mark domains whose interval excludes zero. Because SE_{policy} blends conditional squared error with an abstention penalty, we also report coverage (answer rate) and selective metrics to decompose policy effects into action (answer vs abstain) and conditional performance.

across architectures and tasks to be an important direction for future work.

5.1 Structural stability of hallucinations

The sensitivity analysis from Section 4.2 also clarifies what our results *do not* show. A natural hypothesis is that overconfident errors might coincide with internally “unstable” computation, in the sense that the corresponding hidden states are more sensitive to small input perturbations than those leading to confidently correct answers. Within the scope of our experiment, we do not find evidence for such an instability gap. For Llama-3.1-8B, the mean sensitivity in the final layer is 16.45 for confidently correct items versus 16.55 for overconfidently wrong items; for Qwen2.5-7B the corresponding values are 4.74 and 4.73; and for DeepSeek-R1-Distill they are 13.06 and 13.37. In all three models, the difference between the two groups is tiny relative to the absolute scale of the sensitivity, and far smaller than the variation across layers or across models.

Taken together with the layer-wise reductions under the Kantian prompt, these findings suggest a picture in which hallucinations are not simply transient glitches in an otherwise stable computation. Instead, at least in the regimes we probe, the model appears to form internally coherent and locally stable representations even when its final judgement is confidently wrong. We refer to this pattern as *stable miscalibration*: the model’s internal reasoning is self-consistent and robust to small perturbations, yet systematically misaligned with external truth. If this interpretation is broadly correct, it implies that interventions which only dampen stochasticity (e.g., lowering sampling temperature) or apply post-hoc output calibration (e.g., temperature scaling) may have limited power to address high-confidence hallucinations [31, 32], because they do not directly challenge these stable but misaligned attractors in representation space.

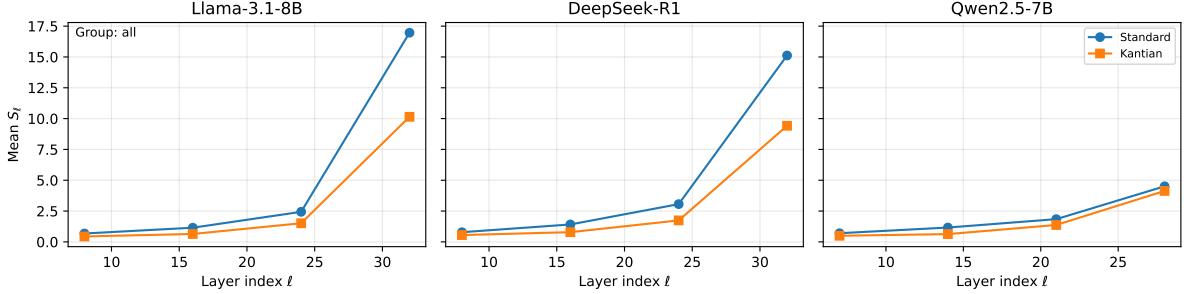


Figure 3: Layer-wise internal sensitivity S_ℓ under standard vs Kantian prompts for three open-weight instruction-tuned models (all items, C0 baseline). The vertical axis is shared across models. In each panel we plot the mean local sensitivity at several representative layers under the overconfident baseline prompt (Standard) and the Kantian prompt. Across architectures, the Kantian prompt consistently reduces internal sensitivity at all probed depths, while the absolute level and depth profile of sensitivity differ between models.

Table 4: Layer-wise sensitivity under the standard versus Kantian prompts. Lower values indicate more stable internal representations. Entries report mean \pm SE over items ($n = 80$; all domains pooled) at $\sigma = 0.01$ with $n_{\text{trials}} = 40$ perturbations per item. “Reduction” reports $(1 - \bar{S}_\ell^{\text{Kant}} / \bar{S}_\ell^{\text{Std}}) \times 100$, where bars denote item means at each layer. SE is computed over items after averaging trials per item.

Model	Layer	Std sensitivity	Kant sensitivity	Reduction (%)
Llama-3.1-8B	L8 (shallow)	0.68 (0.015)	0.44 (0.004)	35.2
	L16 (middle)	1.15 (0.019)	0.64 (0.007)	44.2
	L24 (deep)	2.44 (0.039)	1.51 (0.016)	38.0
	L32 (final)	16.97 (0.251)	10.15 (0.103)	40.2
DeepSeek-R1	L8	0.79 (0.016)	0.55 (0.006)	29.4
	L16	1.41 (0.029)	0.79 (0.018)	44.2
	L24	3.06 (0.058)	1.74 (0.035)	43.1
	L32	15.12 (0.247)	9.42 (0.163)	37.7
Qwen2.5-7B	L7	0.70 (0.008)	0.50 (0.002)	29.3
	L14	1.16 (0.013)	0.63 (0.004)	45.8
	L21	1.84 (0.028)	1.37 (0.014)	25.6
	L28	4.51 (0.081)	4.13 (0.060)	8.3

Instead, our results motivate *process-level* interventions that explicitly perturb and re-evaluate the inference trajectory—for example, critique-and-revision loops or other forms of “Kantian” self-critique that can dislodge stable but misaligned attractors rather than merely smoothing their outputs. This matters because such *stable miscalibration* is unlikely to be fixed by output-only heuristics such as temperature scaling or simple re-sampling; it calls for interventions that act on the model’s reasoning process (e.g., explicit self-critique) to change the attractor the model settles into.

Relatedly, recent confidence-expression benchmarks report substantial overconfidence for Qwen2.5-Instruct-style models [33], consistent with the broader concern that high-confidence errors can persist even when models can emit confidence signals.

By contrast, the Kantian prompt we study can be viewed as a lightweight form of external rational intervention: it asks the model to re-examine its own stance, enumerate possible failure modes, and revisit its answer under an explicit norm of self-critique. Our results do not show that such prompts uniquely solve stable miscalibration, nor that they are sufficient in safety-critical settings. However, they do suggest that prompting strategies which explicitly cultivate self-

critical re-evaluation are a promising design pattern for disturbing otherwise stable hallucination modes, and for nudging the model towards more cautious and better calibrated use of its own internal evidence.

We regard this “Kantian critique” perspective as a hypothesis-generating framework rather than a complete solution, and we expect that testing it across broader model classes, tasks, and stability metrics will be an important direction for future work.

Beyond the three 7–8B models analysed in the main text, we repeated the same C0 sensitivity analysis on two smaller open-weight models—TinyLlama-1B-Chat and Qwen2.5-3B-Instruct—restricted to the same high-risk domains (see Appendix 7). In both cases the mean sensitivity for overconfidently wrong items was within 2–3% of that for confidently correct items (and slightly lower), again showing no appreciable instability gap. This suggests that the stable-miscalibration pattern we observe is not confined to a single architecture or scale, although our evidence remains limited to a small set of models and domains.

Finally, when we extend the C0 sensitivity analysis from the high-risk subset to all domains, the same qualitative picture persists: for Llama-3.1-8B, DeepSeek-R1-Distill, and Qwen2.5-7B the mean final-layer sensitivities for confidently correct and overconfidently wrong items differ by only a few percent, well within across-item variation. High-risk domains tend to exhibit slightly higher sensitivities than low-stakes domains, but this effect is modest compared with the differences across layers and across models.

5.2 The mechanism of stability in Qwen-2.5: high-SNR and low effective signal temperature

Our sensitivity analysis shows that Qwen-2.5 exhibits substantially lower local sensitivity to input perturbations than Llama-3.1, even though our spectral measurements indicate that Qwen’s attention and MLP output matrices have *larger* spectral norms (Figure 4(a,b)). Large spectral norms are usually associated with signal amplification and potential instability, so this combination at first looks paradoxical.

We resolve this tension by examining the magnitude of internal activations. As shown in Figure 4(c), Qwen-2.5 maintains hidden states with much larger ℓ_2 norms $\|x\|_2$ throughout the depth of the network. In architectures with RMSNorm (as in Qwen2/Qwen2.5, which adopt RMSNorm in a pre-norm Transformer design; [34, 35]), the normalisation step operates roughly as $x \mapsto x/\text{RMS}(x)$ (up to a learned gain), so the effective impact of a fixed-size perturbation ϵ on the normalised state scales like $\epsilon/\text{RMS}(x)$. Since $\text{RMS}(x) = \|x\|_2/\sqrt{d}$ for a d -dimensional state, our empirical $\|x\|_2$ profiles imply the same conclusion up to a constant factor. When $\|x\|_2$ is very large, the relative influence of ϵ is therefore strongly compressed.

This suggests that Qwen-2.5 achieves robustness not by damping signals through small weights, but by entering a high signal-to-noise regime in which large-magnitude internal states endow the computation with strong “inertia”. In this sense Qwen-2.5 behaves like a lower “effective signal temperature”⁴ system: typical small perturbations are diluted by normalisation against very large internal magnitudes, so the hidden-state trajectory is comparatively inert.

Importantly, this perturbation-compression channel does not exclude other, more discrete instability modes. When internal activations and projection norms are large, attention logits can become high-magnitude and the softmax can saturate, yielding near one-hot attention patterns. Such saturation can make the computation appear locally stable (small input noise does not move the dominant attention pattern), while simultaneously creating a *hard-switching* regime in which rare perturbations that change the top logit can produce abrupt downstream changes. One concrete way this can occur is through *attention-head outliers*: a small number of heads can dominate the pre-softmax scores for particular tokens, producing unusually large logit gaps (see,

⁴This is an analogy for perturbation compression under normalization and should not be confused with the softmax temperature parameter used in sampling or temperature scaling.

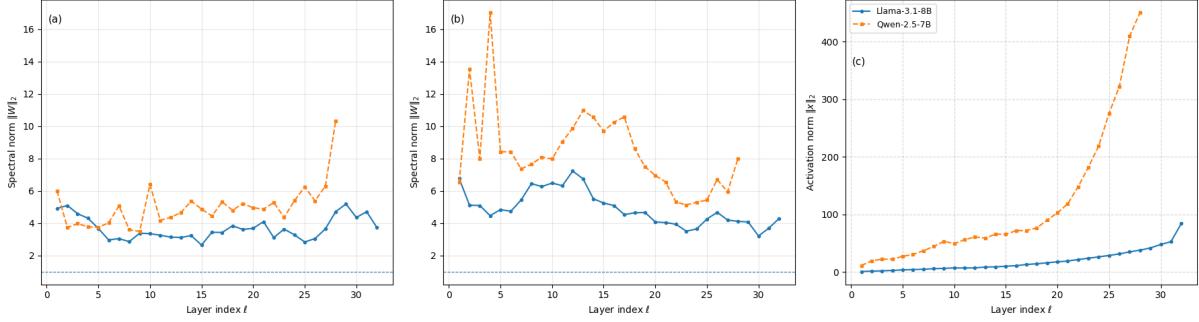


Figure 4: Spectral and activation profiles for Llama-3.1-8B and Qwen-2.5-7B. (a) Spectral norm $\|W\|_2$ of the attention output projections (o_{proj}) across layers. (b) Spectral norm of the MLP down-projections ($\text{down}_{\text{proj}}$). (c) Layer-wise activation norm $\|x\|_2$ at the last token. In all panels, solid lines (circles) denote Llama-3.1-8B and dashed lines (squares) denote Qwen-2.5-7B. Qwen-2.5 exhibits consistently larger spectral norms and much larger activation norms, indicating a high-SNR, low-effective-signal-temperature regime in which relative perturbations are strongly compressed despite large weights.

e.g., the Qwen2.5 case study in [36]). In that regime, softmax probabilities become effectively delta-like, so the model is insensitive to most small changes but can exhibit near-discrete flips when a perturbation is just large enough to swap the top logit. This “quiet-then-flip” behaviour is qualitatively distinct from RMSNorm-based perturbation compression and provides a second channel by which a high-SNR model can look stable while still supporting brittle decision boundaries. This helps explain how Qwen-2.5 can be both locally robust and prone to persistent, confidently wrong configurations: stability can coexist with miscalibration when the model settles into a high-SNR representation that is difficult to dislodge by modest contextual shifts.

Model-specific epistemic profiles. Although our analysis is exploratory and limited to a small set of models and domains, the combination of local sensitivity and high-SNR behaviour suggests that current LLMs may occupy distinct epistemic regimes. DeepSeek-R1 displays relatively high local sensitivity and more modest activation norms, corresponding to an internally reactive regime in which small perturbations can have comparatively large effects on hidden states. Qwen-2.5, by contrast, combines low local sensitivity with large spectral and activation norms, indicating a high-SNR, low effective signal temperature regime with strong signal inertia: once its internal representations have settled, small perturbations have little influence on the subsequent trajectory. Llama-3.1 lies between these extremes, with intermediate sensitivity and activation norms, and thus serves as a more balanced reference point in our experiments. We do not propose a formal taxonomy here, but this kind of epistemic profiling may be useful when reasoning about which models are more likely to exhibit reactive versus inertial patterns of error under different prompting regimes.

5.3 Limitations and future work

Our analysis is deliberately small-scale and structurally biased towards interpretability. On the LLM side we work with a limited set of 7–8B instruction-tuned models, a small binary fact-checking dataset, and a simple domain-wise proxy $H_{\text{proxy}}(d)$ that averages over items and thus bundles together intrinsic domain difficulty and model behaviour. For internal stability we focus on local sensitivity to small input perturbations at a fixed noise scale and probe only a few representative layers per model. These choices make the experiments tractable and easy to replicate, but they necessarily restrict the scope of our claims.

A natural next step is to scale up both the models and the stability diagnostics. On the model side, it will be important to test whether the stable-miscalibration pattern we observe persists in larger checkpoints, different training regimes, and multimodal architectures. On

the metric side, a priority is to approximate the operator-level index H_{RiskLLM} more directly, for example via Jacobian–vector products or layer-wise linearisation techniques that make Jacobian-based conditioning and sensitivity measurements computationally feasible at scale. It would also be valuable to design domain- and difficulty-controlled benchmarks that disentangle inherent task hardness from architecture-specific stability profiles, and to explore richer forms of Kantian feedback—beyond single-shot prompt variants—that can be integrated into multi-step critique-and-revision loops or external safety scaffolds.

6 Conclusion

We have proposed H-Risk as a structural index of epistemic instability that links Kant’s feedback-based view of cognition to closed-loop state-estimation operators in linear–Gaussian systems and to simple instability proxies for large language models. In the linear setting, the composite index H_{Risk} aggregates margin, conditioning, temporal sensitivity, and innovation amplification of the closed-loop operator, and in our simulations it tracks regimes of miscalibration and poor closed-loop behaviour.

On the LLM side, we instantiated a coarse, domain-wise proxy $H_{\text{proxy}}(d)$ based on policy-wise confidence fluctuations and overconfident errors, and used it to map out high-risk domains under different prompting policies. A small sanity-check study on binary factual questions showed that a Kantian-inspired policy which allows the model to answer “cannot judge” yields modest but targeted reductions in policy-aware squared loss ($\Delta \text{SE}_{\text{policy}}$), concentrated in high-stakes domains. Our internal sensitivity analysis revealed no evidence for the hypothesised instability gap: confidently wrong answers are at least as locally stable as confidently correct ones, leading us to characterise many hallucinations as instances of stable miscalibration rather than transient instability. For Qwen-2.5 in particular, spectral and activation profiles indicate a high-SNR, low-effective-signal-temperature regime in which internal representations can become “frozen” into confidently wrong but locally robust states.

Taken together, these results suggest that hallucinations in modern LLMs often behave less like bugs caused by fragile dynamics and more like robust but misaligned specifications supported by stable internal representations. We hope that viewing epistemic stability as a design principle—to be monitored and shaped through indices such as H_{Risk} and its proxies—will encourage further work at the interface of control theory, philosophical theories of critique, and the practical evaluation and alignment of large-scale generative models.

Limitations and Broader Impact

Limitations. Our analysis separates a linear–Gaussian control-theoretic core from an LLM proxy study, and each side has its own constraints. On the LTI side, we work with time-invariant state-space models and use the spectral properties and conditioning of $\Phi = A - KH$ together with innovation statistics to construct H_{Risk} ; this captures one natural notion of epistemic stability but leaves out nonlinearities, structural model misspecification, and multi-agent or social feedback. On the LLM side, we only approximate an operator-level notion of stability via local sensitivity ratios $\|\Delta h\|/\|\epsilon\|$, spectral norms, and activation norms in a handful of open-weight models. These quantities are easiest to interpret in small-perturbation regimes and for single-turn factual prompts, and it remains unclear how well they predict behaviour in long-horizon, tool-augmented, or multimodal deployments. Our empirical coverage is narrow in terms of languages, tasks, and prompting styles, and the Kantian prompt we study is only one of many possible implementations of “critique”. We emphasise the Kantian framing as a practical design principle for feedback and self-limitation.

Broader Impact. By linking Kantian notions of self-limitation with state-space inference and LLM prompting, this work is intended as a conceptual bridge between philosophy of cognition, control theory, and AI safety. A stability-based view of hallucination may help practitioners think beyond scalar accuracy metrics and ask when a model’s internal dynamics make errors brittle versus corrigible, and may motivate reporting practices that include calibration, uncertainty, and sensitivity to perturbations. At the same time, there is a risk that the mathematical formalism and personality-like “epistemic profiles” we sketch are over-interpreted as guarantees of safety or as grounds for ranking models normatively. The Kantian prompt we study yields only modest, domain-dependent gains and should not be treated as a sufficient safeguard in high-stakes applications. Translating transcendental categories into mathematical form is a heuristic exercise; it should be used, if at all, to structure critical scrutiny of AI systems rather than to confer philosophical authority on them.

Use of Generative AI Tools. In accordance with current publication guidelines (e.g., *Nature*, NeurIPS, and arXiv policies), the author discloses the use of OpenAI’s ChatGPT for limited assistance in language editing, code refactoring, and conceptual clarification. All content, analyses, and conclusions were independently verified by the author, and no generative AI was used for creating or modifying figures.

7 Reproducibility Checklist

We provide code, data paths, and fixed seeds to reproduce all figures and tables in this manuscript.

- **Repository/branch:** 202510_report_AI (workspace: Kantian), branch `main`. **URL:** https://github.com/ToppyMicroServices/202510_report_AI.git
- **Environment:** Python 3.9.6 in a virtual environment; dependencies listed in `requirements.txt`.
Build the paper with `make v3`.
- **Data sources:** LLM experiment results are read from `data/results_llm_experiment_prod.csv` (frozen), with auxiliary summaries written under `results/`. Optional regeneration details are provided in the Appendix (“Dataset and scripts”).
- **Seeds determinism:** The LTI simulation uses a fixed RNG seed `CFG["seed"] = 2025` and shared noise sequences (`W_SEQ`, `V_SEQ`); the $\Delta\text{SE}_{\text{policy}}$ analysis is deterministic given the input CSV.
- **Figure scripts:** LTI dual plot: `scripts/LTI.py` writes `paper/latex_v3/figures/LTI_dual_plot_autotuned.png`.
Domain-wise $\Delta\text{SE}_{\text{policy}}$ (Fig. 2) and PMF/violin plots: `scripts/analyze_condition_deltas.py` writes figures to `paper/latex_v3/figures/` and tables to `paper/latex_v3/Tables/`.
- **How to reproduce:** (i) run `python scripts/analyze_condition_deltas.py` to regenerate all $\Delta\text{SE}_{\text{policy}}$ figures/tables; (ii) optionally run `python scripts/LTI.py` to regenerate the LTI plots; (iii) run `make v3` to compile the PDF.

Computational note. The steady-state covariance P is obtained by solving the discrete-time Lyapunov equation $P = \Phi P \Phi^\top + \Sigma$ using the Bartels–Stewart algorithm based on Schur decomposition [2]; existence and uniqueness of a positive-definite P under $\rho(\Phi) < 1$ follow from standard results in optimal filtering and Lyapunov stability theory [37, 3].

Acknowledgments

Author Note. This work was conducted in a personal capacity, outside the author's employment with another organization. ToppyMicroServices OÜ is the author's independently owned, early-stage startup listed as a correspondence affiliation; it is not the author's employer. No external funding was received. The views expressed are solely those of the author, and any errors are the author's alone.

Competing Interests

The author is the founder and owner of ToppyMicroServices OÜ (pre-revenue at the time of writing). The author reports no commercial sponsorships, client relationships, or other competing interests relevant to this work.

Compliance Statement

This personal research was conceived and completed outside the scope of the author's employment, using only personally owned hardware and personal cloud/accounts; no employer facilities, data, source code, or confidential information were used. To the author's knowledge, the work does not fall under any employer intellectual property assignment, work-for-hire, or similar clause, does not rely on proprietary materials of the employer, and does not use the employer's name, trademarks, or branding.

References

- [1] Lloyd N. Trefethen and Mark Embree. *Spectra and Pseudospectra: The Behavior of Nonnormal Matrices and Operators*. Princeton University Press, Princeton, NJ, 2005. ISBN 9780691119465.
- [2] Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, 4th edition, 2013.
- [3] Kemin Zhou, John C. Doyle, and Keith Glover. *Robust and Optimal Control*. Prentice Hall, 1996.
- [4] Immanuel Kant. *Critique of Pure Reason*. Johann Friedrich Hartknoch, 1781. A/B editions, translated by P. Guyer and A. W. Wood, Cambridge University Press, 1998.
- [5] Carl B. Sachs. A cybernetic theory of persons: how sellars naturalized kant. *Philosophical Inquiries (philinq)*, 2022. URL <https://philinq.it/index.php/philinq/article/download/389/256>.
- [6] J. K. Burmeister. Kant, cybernetics, and cybersecurity: Integration and implications. *Systemics, Cybernetics and Informatics*, 2021. URL <https://www.iiisci.org/journal/pdv/sci/pdfs/IP132LL21.pdf>.
- [7] Thomas Marlowe. Philosophy and cybernetics: Questions and issues. *Systemics, Cybernetics and Informatics*, 2021. URL <https://www.iiisci.org/Journal/PDV/sci/pdfs/IP130LL21.pdf>.
- [8] Ziwei Ji, Zhiyuan Zeng, Yu Li, Chiyuan Zhang, and Percy Liang. Llm internal states reveal hallucination risk faced with novelty. In *Proceedings of the 8th Workshop on Analysing and Interpreting Neural Networks for NLP (BlackboxNLP 2024)*. Association for Computational Linguistics, 2024. URL <https://aclanthology.org/2024.blackboxnlp-1.6/>.

- [9] Anonymous. On the fundamental impossibility of hallucination control in llms. *arXiv preprint arXiv:2506.06382*, 2025. URL <https://arxiv.org/abs/2506.06382>.
- [10] Henry E. Allison. *Kant’s Transcendental Idealism: An Interpretation and Defense*. Yale University Press, New Haven, CT, 2nd edition, 2004.
- [11] Paul Guyer. *Kant*. Routledge Philosophers. Routledge, 2006.
- [12] Lawrence R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.
- [13] Neil Gordon, David Salmond, and Adrian Smith. Novel approach to nonlinear/non-gaussian bayesian state estimation. *IEE Proceedings F (Radar and Signal Processing)*, 140(2):107–113, 1993.
- [14] Arnaud Doucet, Nando de Freitas, and Neil Gordon, editors. *Sequential Monte Carlo Methods in Practice*. Springer, 2001.
- [15] R. E. Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82:35–45, 1960.
- [16] Ziwei Ji et al. Survey of hallucination in natural language generation. *arXiv preprint arXiv:2202.03629*, 2023.
- [17] Aisha Alansari and Hamzah Luqman. A comprehensive survey of hallucination in large language models. *arXiv preprint arXiv:2510.06265*, 2025.
- [18] Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*, 2023.
- [19] Hao Li et al. How to detect and defeat molecular mirage: A metric-driven benchmark for hallucination in llm-based molecular comprehension. *arXiv preprint arXiv:2504.12314*, 2025.
- [20] Seongho Joo, Kyungmin Min, Jahyun Koo, and Kyomin Jung. Black-box hallucination detection via consistency under the uncertain expression. *arXiv preprint arXiv:2509.21999*, 2025.
- [21] Denis Janiak, Jakub Binkowski, Albert Sawczyn, Bogdan Gabrys, Ravid Shwartz-Ziv, and Tomasz Jan Kajdanowicz. The illusion of progress: Re-evaluating hallucination detection in llms. *arXiv preprint arXiv:2508.08285*, 2025.
- [22] Aman Madaan et al. Self-refine: Iterative refinement with self-feedback. *arXiv preprint arXiv:2303.17651*, 2023.
- [23] Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *arXiv preprint arXiv:2303.11366*, 2023.
- [24] Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. Critic: Large language models can self-correct with tool-interactive critiquing. *arXiv preprint arXiv:2305.11738*, 2023.
- [25] Alvin Chan et al. Chateval: Toward better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*, 2023.

- [26] Amirata Ghorbani, Shankar Krishnan, Yin Xiao, Been Kim, and Percy S. Liang. An investigation into neural network jacobians and their spectrum. In *ICML*, 2019.
- [27] Karthik Sankararaman, Elad Hoffer, and Daniel Soudry. The impact of jacobian conditioning on generalization in deep learning. In *NeurIPS*, 2020.
- [28] Arthur Jacot, Stefano Spigler, Frank Gabriel, and Clément Hongler. Implicit regularization of neural tangent kernels. *JMLR*, 2021.
- [29] Greg Yang, Eta Littwin, and Andrew Saxe. Tensor programs iii: Neural matrix laws. In *ICML*, 2022.
- [30] Minhui Zhu, Minyang Tian, Xiaocheng Yang, Tianci Zhou, Penghao Zhu, Eli Chertkov, Shengyan Liu, Yufeng Du, Lifan Yuan, Ziming Ji, et al. Probing the critical point (critpt) of ai reasoning: a frontier physics research benchmark. *arXiv preprint arXiv:2509.26574*, 2025.
- [31] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 1321–1330, 2017.
- [32] Saurav Kadavath, Thomas Conerly, Amanda Askell, Tom Henighan, Andy Jones, Nicholas Schiefer, Nicholas Joseph, Nova DasSarma, Sam McCandlish, Catherine Olsson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.
- [33] Dongkeun Yoon, Seungone Kim, Sohee Yang, Sunkyoung Kim, Soyeon Kim, Yongil Kim, Eunbi Choi, Yireun Kim, and Minjoon Seo. Reasoning models better express their confidence, 2025. URL <https://arxiv.org/abs/2505.14489>. Accepted to NeurIPS 2025.
- [34] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.
- [35] Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.
- [36] Charles L. Chen. A huge flaw inside qwen2.5 – bad robustness and its solution. Medium (online article), 2025. URL <https://medium.com/@crclq2018/a-huge-flaw-inside-qwen2-5-14940178833f>. Online; accessed 2025-12-13.
- [37] Brian D. O. Anderson and John B. Moore. *Optimal Filtering*. Prentice-Hall, 1979.
- [38] Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950.
- [39] Allan H. Murphy. A new vector partition of the probability score. *Journal of Applied Meteorology*, 12(4):595–600, 1973.
- [40] Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.
- [41] Maja Pavlovic. Understanding model calibration – a gentle introduction and visual exploration of calibration and the expected calibration error (ece). *arXiv preprint arXiv:2501.19047*, 2025. URL <https://iclr-blogposts.github.io/2025/blog/calibration/>. ICLR Blogposts 2025.

- [42] Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, 2015.
- [43] Jeremy Nixon, Michael Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. Measuring calibration in deep learning. *arXiv preprint arXiv:1904.01685*, 2019.
- [44] Dieter Henrich. *The Unity of Reason: Essays on Kant’s Philosophy*. Harvard University Press, 1994.

Appendix: Calibration Metrics (informational)

Noise scale sweep for local sensitivity (DeepSeek-R1)

As discussed in Section 4.2, our local sensitivity measure

$$S_\ell(\epsilon) = \frac{\|h_\ell^{\text{noisy}} - h_\ell^{\text{clean}}\|_2}{\|e^{\text{noisy}} - e^{\text{clean}}\|_2} \approx \frac{\|f_\ell(e + \epsilon) - f_\ell(e)\|_2}{\|\epsilon\|}$$

approximates the directional Jacobian norm in the limit $\|\epsilon\| \rightarrow 0$. In a strictly linear regime $S_\ell(\epsilon)$ would be invariant under rescaling of the perturbation norm, but transformer blocks combine residual connections, RMSNorm and non-linear activations, so larger perturbations can cross activation boundaries and interact with normalisation in a non-linear way.

To probe this effect, we ran a small noise-scale sweep on DeepSeek-R1-Distill for the C0 baseline, varying the embedding noise standard deviation $\sigma \in \{0.005, 0.01, 0.025, 0.05\}$ and computing the mean layer-wise sensitivity S_ℓ across items for four representative layers. Figure 5 shows that increasing σ systematically lowers the estimated S_ℓ in deeper layers by a factor of roughly 2–3, consistent with the interpretation that larger perturbations leave the local linear regime and are partially “flattened” by non-linearities and RMSNorm. For the main analyses we therefore fix a small noise level ($\sigma = 0.01$) that lies close to the local regime while remaining numerically stable, and we treat S_ℓ primarily as an order-of-magnitude probe of internal gain rather than as a scale-free quantity.

Importantly, across all tested σ the difference in S_ℓ between confidently correct and overconfidently wrong items remains negligible relative to the mean for the probed layers (numerical values omitted for brevity), so our “no instability gap” conclusion—and the resulting interpretation in terms of stable miscalibration—is robust to the choice of noise scale.

Sensitivity analysis for smaller open-weight models

To assess whether the C0 sensitivity pattern depends on model size, we repeated the same analysis on two smaller open-weight models (TinyLlama-1B-Chat and Qwen2.5-3B-Instruct) restricted to the same high-risk domains used in the main-text experiment. We omit detailed plots for brevity; the summary result is reported in Sec. 4.2.

Calibration metrics (Brier and ECE)

For the experiments reported in this version, calibration is quantified primarily via the Brier score, with Expected Calibration Error (ECE) used as a secondary background metric.⁵ In the LLM sanity-check experiment (Sec. 4.1) we work in a binary setting and treat the coarse confidence level produced by our rule-based A-layer classifier as a probability $p \in \{0.5, 0.8\}$ that the model’s Yes/No answer is factually correct. For C0_ECE, we instead use the self-reported probability $P(\text{correct})$ (on a limited grid).

⁵For background on Brier as a strictly proper scoring rule and its decomposition, see [38, 39, 40]. For ECE and binning practices see [41, 42, 31, 43].

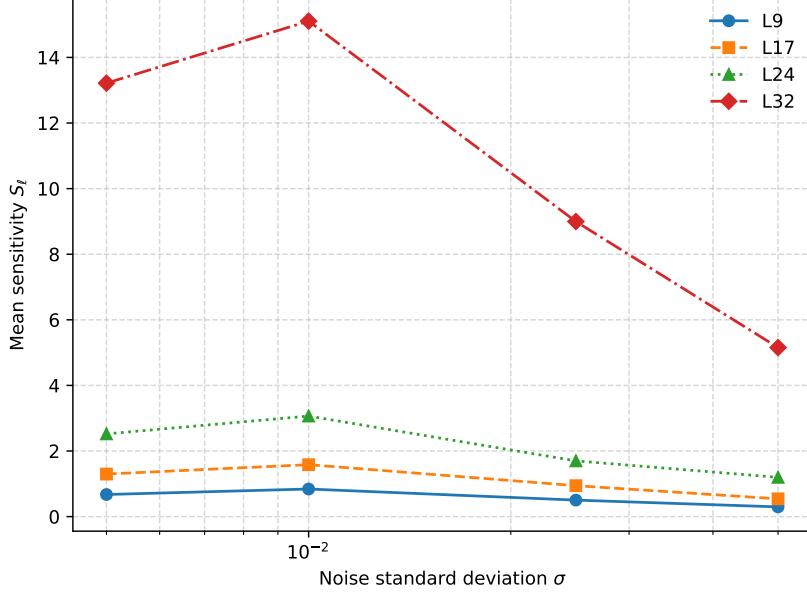


Figure 5: Noise-scale sweep for local sensitivity on DeepSeek-R1-Distill (C0 baseline, standard prompt). Each curve shows the mean sensitivity S_ℓ at a given layer as a function of the embedding noise standard deviation $\sigma \in \{0.005, 0.01, 0.025, 0.05\}$. Larger noise scales yield smaller estimated S_ℓ , reflecting departure from the purely local linear regime, but the absence of an instability gap between confidently correct and overconfidently wrong items is preserved across scales.

Sensitivity to the coarse-confidence mapping

Because the A-layer confidence values are intentionally schematic, we run a simple sensitivity check in which we remap the “high” Yes/No confidence level from 0.8 to $p \in \{0.7, 0.9\}$ while keeping the underlying answer/abstain decisions fixed. Table 5 reports the resulting paired $\Delta SE_{\text{policy}}$ comparisons vs C0; the qualitative ordering of policies is unchanged.

Table 5: Sensitivity of the policy-aware squared loss and paired $\Delta SE_{\text{policy}}$ to the chosen high-confidence value for Yes/No answers (all domains vs high-risk domains).

High p	Subset	$SE_{\text{policy}}(\text{C0})$	$\Delta SE_{\text{policy}}(\text{C1} - \text{C0})$	$\Delta SE_{\text{policy}}(\text{C2} - \text{C0})$	n
0.700	All domains	0.258	-0.025	-0.018	367
0.700	High-risk domains	0.351	-0.087	-0.071	121
0.800	All domains	0.292	-0.055	-0.042	367
0.800	High-risk domains	0.432	-0.150	-0.123	121
0.900	All domains	0.346	-0.096	-0.075	367
0.900	High-risk domains	0.532	-0.227	-0.187	121

Selective metrics by domain and condition

Expected Calibration Error (ECE). For completeness, we recall the standard definition of ECE. Partition examples into B confidence bins S_b by predicted confidence \hat{p}_i :

$$\text{ECE} = \sum_{b=1}^B \frac{|S_b|}{n} |\text{acc}(S_b) - \text{conf}(S_b)|, \quad S_b = \{i : \hat{p}_i \in I_b\},$$

Table 6: Selective metrics by domain and condition. Coverage captures abstention; selective accuracy is conditional on answering; answer yield counts abstentions as incorrect; and OC-Wrong|Ans is the overconfident-wrong rate among answered items.

Domain	Cond.	n	Coverage	Sel. Acc.	Answer Yield	OC-Wrong Ans
cultural_industry	C0	33	1.000	0.364	0.364	0.636
cultural_industry	C1	38	0.263	0.400	0.105	0.600
cultural_industry	C2	38	0.289	0.273	0.079	0.727
cultural_industry	C0_ECE	46	1.000	0.478	0.478	0.522
dev_region	C0	39	1.000	0.590	0.590	0.410
dev_region	C1	41	0.537	0.682	0.366	0.318
dev_region	C2	41	0.780	0.688	0.537	0.312
dev_region	C0_ECE	45	1.000	0.600	0.600	0.400
entertainment_event	C0	30	1.000	0.733	0.733	0.267
entertainment_event	C1	35	0.257	0.667	0.171	0.333
entertainment_event	C2	35	0.343	0.500	0.171	0.500
entertainment_event	C0_ECE	43	1.000	0.651	0.651	0.349
geo_travel	C0	38	1.000	0.632	0.632	0.368
geo_travel	C1	45	0.511	0.870	0.444	0.130
geo_travel	C2	45	0.511	0.739	0.378	0.261
geo_travel	C0_ECE	51	1.000	0.647	0.647	0.353
history_diplomacy	C0	25	1.000	0.440	0.440	0.560
history_diplomacy	C1	33	0.424	0.714	0.303	0.286
history_diplomacy	C2	33	0.515	0.588	0.303	0.412
history_diplomacy	C0_ECE	39	1.000	0.590	0.590	0.410
literature_media	C0	23	1.000	0.826	0.826	0.174
literature_media	C1	26	0.192	0.800	0.154	0.200
literature_media	C2	26	0.269	0.571	0.154	0.429
literature_media	C0_ECE	35	1.000	0.771	0.771	0.229
macro_index	C0	34	1.000	0.618	0.618	0.382
macro_index	C1	41	0.854	0.686	0.585	0.314
macro_index	C2	41	0.902	0.730	0.659	0.270
macro_index	C0_ECE	48	1.000	0.667	0.667	0.333
medical_epidemiology	C0	29	1.000	0.172	0.172	0.828
medical_epidemiology	C1	38	0.237	0.111	0.026	0.889
medical_epidemiology	C2	38	0.447	0.412	0.184	0.588
medical_epidemiology	C0_ECE	57	1.000	0.404	0.404	0.596
social_stats	C0	34	1.000	0.412	0.412	0.588
social_stats	C1	41	0.341	0.714	0.244	0.286
social_stats	C2	41	0.537	0.545	0.293	0.455
social_stats	C0_ECE	49	1.000	0.429	0.429	0.571
sports	C0	42	1.000	0.833	0.833	0.167
sports	C1	50	0.760	0.842	0.640	0.158
sports	C2	50	0.820	0.902	0.740	0.098
sports	C0_ECE	64	1.000	0.844	0.844	0.156
technical_standard	C0	40	1.000	0.675	0.675	0.325
technical_standard	C1	49	0.143	0.857	0.122	0.143
technical_standard	C2	49	0.204	0.800	0.163	0.200
technical_standard	C0_ECE	55	1.000	0.855	0.855	0.145

$$\text{acc}(S_b) = \frac{1}{|S_b|} \sum_{i \in S_b} \mathbb{I}\{\hat{y}_i = y_i\}, \quad \text{conf}(S_b) = \frac{1}{|S_b|} \sum_{i \in S_b} \hat{p}_i.$$

We follow common practice and use $B = 10$ bins together with a debiased estimator [42, 43]. When the confidence values are sufficiently rich, we use equal-frequency bins; however, for the C0_ECE condition the self-reported probabilities lie on a limited grid, so we report bin statistics using fixed confidence intervals (as in Table 7), and the resulting bin counts need not be exactly equal. Reliability diagrams follow [31] where applicable.

Brier score. In the binary case with labels $y_i \in \{0, 1\}$ and predicted probabilities $\hat{p}_i \in [0, 1]$ for the positive class, the Brier score reduces to

$$\text{Brier} = \frac{1}{n} \sum_{i=1}^n (\hat{p}_i - y_i)^2,$$

which is minimized at 0 and maximized at 1 when a wrong class is predicted with probability 1. More generally, for K -class problems with probability vectors $\hat{\mathbf{p}}_i$ and one-hot targets \mathbf{e}_{y_i} we use

$$\text{Brier} = \frac{1}{n} \sum_{i=1}^n \|\hat{\mathbf{p}}_i - \mathbf{e}_{y_i}\|_2^2 = \frac{1}{n} \sum_{i=1}^n \sum_{k=1}^K (\hat{p}_{ik} - \mathbb{I}\{y_i = k\})^2.$$

The Brier score is strictly proper and admits Murphy's uncertainty-resolution-reliability decomposition [39, 40].

Additional LLM calibration plots (C0_ECE)

For the small-scale Kantian ABC LLM experiment in Sec. 4.1, we also report calibration diagnostics for the baseline condition with self-reported confidence (C0_ECE). Figure 6 shows the empirical distribution of per-item Brier scores for C0_ECE (using the self-reported probabilities). These plots complement the main-text Brier and $\Delta\text{SE}_{\text{policy}}$ summaries by making the limited-grid self-reported probabilities and residual miscalibration visually explicit.

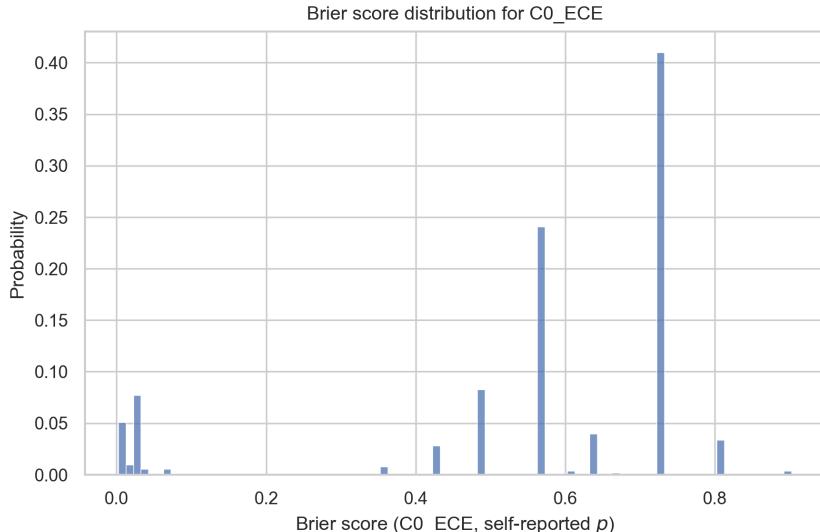


Figure 6: Brier score distribution for the C0_ECE condition (pooled over domains). The discrete support reflects the limited grid of self-reported probabilities (together with binary labels), producing a small number of possible Brier values.

Table 7: ECE bin statistics for C0_ECE (binary LLM sanity-check experiment). The overall ECE for this condition is reported in Sec. 4.1.

Bin	n	Acc.	Conf.	$ acc - conf $
[0.65, 0.70)	1	0.000	0.650	0.650
[0.70, 0.75)	10	0.000	0.700	0.700
[0.75, 0.80)	85	0.035	0.751	0.715
[0.80, 0.85)	138	0.319	0.844	0.526
[0.85, 0.90)	5	1.000	0.878	0.122
[0.90, 0.95)	26	0.731	0.900	0.169
[0.95, 1.00)	7	1.000	0.950	0.050

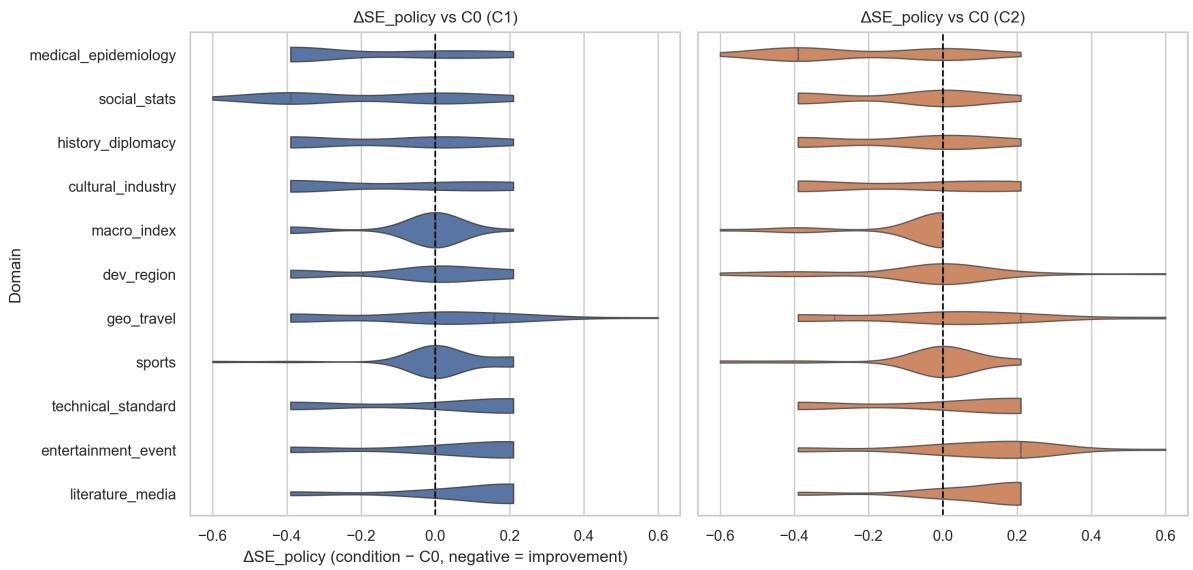


Figure 7: Per-item ΔSE_{policy} distributions by domain (violin plot). This view details the spread of per-item policy-loss changes within each domain, complementing the aggregated means in Figure 2.

LTI Model Robustness

For the linear–Gaussian toy model, where the optimal state estimator coincides with the classical Kalman filter, we also report an additional robustness sweep (Fig. 8) in which we introduce controlled perturbations to the state-transition matrix A while keeping the Kalman filter fixed. Even under this model mismatch, the relationship between the instability index H_{Risk} and the tail calibration metric NIS_q remains broadly monotone: systems with higher H_{Risk} tend to exhibit worse tail normalized innovation squared, supporting the interpretation of H_{Risk} as a robust proxy for practical instability in Kalman-style state estimation.

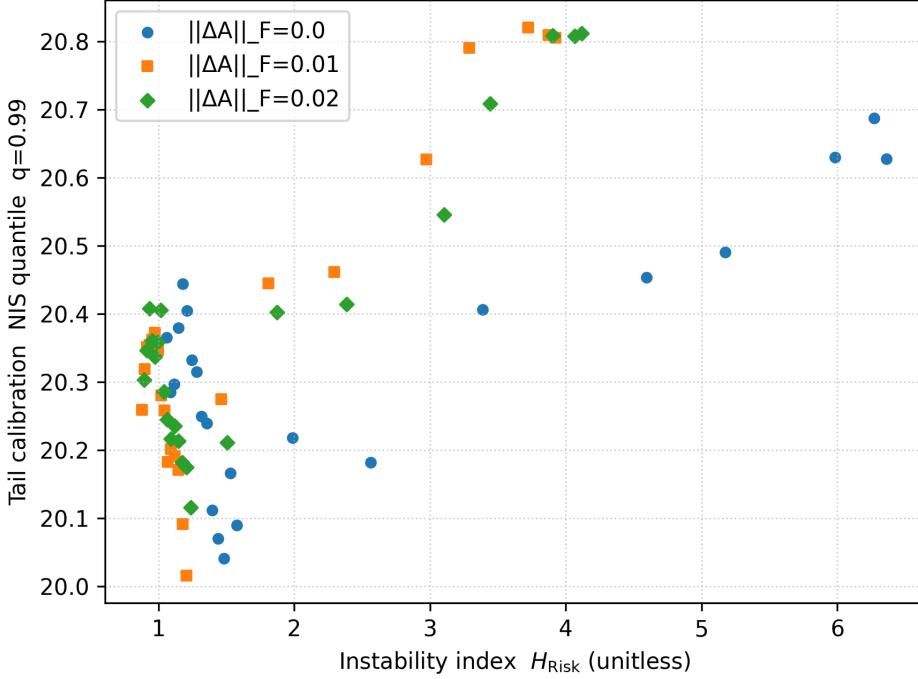


Figure 8: Robustness sweep under model mismatch. We introduce random perturbations to the state transition matrix A while keeping the estimator fixed. The plot shows that even with this mismatch, the instability index H_{Risk} remains predictive of tail calibration error (NIS_q).

Auxiliary probe: CritPt abstention behaviour

To provide an external check on refusal/hesitation behaviour under extreme task difficulty, we ran a small single-shot probe on the CritPt benchmark [30]. We use the `train` split (70 challenge problems) and evaluate the same problem statements under two conditions: C0 (forced answer; refusal disallowed) and C2 (Kantian; may abstain). To make abstentions machine-detectable and robust to formatting variation, we constrain the first output line to be exactly `Answer.` or `Cannot judge.` We then define *coverage* as the fraction of items whose first line is `Answer.`, and *abstention rate* as one minus coverage. This probe is intended to characterise refusal behaviour rather than benchmark accuracy.

The main-text summary is reported in Table 2. For uncertainty, we report 95% item-level bootstrap confidence intervals by resampling the $N = 70$ items with replacement and recomputing the rate on each bootstrap replicate (20,000 replicates). We do not reproduce any CritPt problem statements or model outputs in this appendix.

Dataset and scripts (optional regeneration)

The figures in this manuscript are reproducible from a frozen CSV and do not require re-running the data collection. For completeness, we note the optional generation and aggregation scripts:

- **Frozen dataset:** We ship a fixed CSV `data/results_llm_experiment_prod.csv` used by all plots and tables.
- **Prompt generation (optional):** `scripts/generate_kantian_prompts.py` uses the OpenAI API to synthesise Kantian ABC items and labels. In local mode (set `KANTIAN_LOCAL_MODE=1`), it writes JSON files under `./kantian_local_results/`. An API key (`OPENAI_API_KEY`) is required; usage may incur cost.
- **Experiment runs (optional):** `aws/lambda/run_llm_experiment_ver8.py` evaluates items under policies C0, C1, C2, and C0_ECE. In local/offline mode, it writes raw eval JSONs under `data/local_eval_runs_conf/prod/`.
- **Aggregation to CSV:** To convert local eval JSONs into the CSVs used by analysis, run the aggregator CLI:

Example (local mode; produces the two CSVs referenced in the paper):

```
python aws/lambda/aggregate_eval_runs_ver8.py --mode prod \
--local-dir data/local_eval_runs_conf/prod \
--out-eval data/results_eval_runs_prod.csv \
--out-experiment data/results_llm_experiment_prod.csv
```

These steps are optional and not required to reproduce the plots in this version.

Appendix: Kantian background (informal)

As noted in the Introduction and Conclusion, our use of “Kantian” terminology is an interpretive, structural lens. Here we briefly record the specific points of contact with the *Critique of Pure Reason* that informed the design of H_{Risk} .

Synthetic unity and stability. In the Transcendental Aesthetic and Analytic, Kant argues that space and time are pure forms of intuition, and that the manifold of intuition must be combined under the categories and brought to the unity of apperception in order for there to be cognition of objects at all.[4, 10] This suggests a view of cognition as requiring not just *local* associations but a globally stable synthesis. Our Axiom (H1) abstracts this as a requirement that inference remain in a well-conditioned region of state-space, where small perturbations do not destroy the unity of the resulting “world-model”.

Dialectic and antinomy. In the Transcendental Dialectic, reason generates antinomies when it applies its ideas beyond the bounds of possible experience, producing necessary conflicts between equally specious proofs.[4, 11] We read this as an early analysis of what we call epistemic instability: certain forms of unconstrained extension of principles make systematic self-contradiction unavoidable. Axiom (H2) formalizes this by penalizing patterns of inference that produce persistent contradictions or oscillations across an iterative reasoning chain.

Self-limitation of reason. Kant famously insists that the critique of pure reason does not abolish reason but limits it in order to secure its proper use.[4, 44] Later readers have described this as a “self-limitation of reason”: reason must recognize that its constitutive cognition is restricted to appearances, and that beyond this domain its ideas can only have a regulative role.[10] Axiom (H3) is our attempt to capture an analogue of this self-limiting stance: inference is required to track its own boundary conditions and to treat certain questions as legitimately undecidable or refusal-worthy, rather than hallucinating determinate answers where its preconditions for cognition fail.

Schematism and time. Finally, Kant’s brief and notorious chapter on schematism (A137–147/B176–187) emphasizes that pure concepts require temporal schemata to be applicable to intuition at all.[4] Contemporary commentators often describe this as a kind of procedural or algorithmic mediation between rule and data.[11] Our use of state-space filters and prompt–critique–revision loops can be seen as an admittedly loose analogue: they provide the “middle term” that allows abstract norms (e.g., “avoid unstable overconfidence”) to be implemented in a temporally extended process of estimation and correction. We do not propose a one-to-one identification here, but this analogy guided our choice to treat epistemic stability as a property of dynamical feedback structures rather than static propositions.

These sketches are not meant as contributions to Kant scholarship. They serve only to document the limited sense in which our framework is “Kantian” and to clarify that the mathematical content of the paper is independent of any specific reading of Kant’s doctrines.

Changelog

v1.1. We clarified PCR’s relation to critique-and-revise methods and softened novelty claims.

v3.0. We implemented the small-scale Kantian ABC LLM experiment reported in Sec. 4.1 and aligned the LLM analysis with Brier-based condition deltas.

Appendix: Governance mapping (informational)

This appendix provides a high-level mapping between our metrics and governance frameworks (EU AI Act, ISO/IEC 42001, NIST AI RMF). It is for reference only and *does not* constitute a conformity assessment or legal claim of compliance; any such claims are out of scope for this preprint.