POLICY BRIEF

POLLICINO & PARTNERS
AI DVISORY

The Epistemic Gap in AI Governance

From Probabilistic Models to Stochastic Organisational Reliance

Alberto Stazzone
Senior Advisor

Executive Summary

Large language models present a governance problem that looks simpler than it is. The model holds a distribution over possible continuations; the organisation receives one fluent sample, and that handoff does not automatically preserve or communicate the uncertainty, evidence and limitations needed to determine permitted reliance. The difficulty is not that every output is wrong, but that uncertainty becomes hard to see, hard to allocate and hard to govern. The legally relevant object is therefore not the token distribution but organisational reliance in a concrete process: a hallucinated citation, or a commitment made by a customer-facing chatbot, becomes salient only when the organisation treats the sample as a basis for action. What follows establishes why no output or interface signal, taken alone, can settle that question.

Executive Takeaway

1. The gap is not between deterministic and random systems, but between the uncertainty the model handles and the confidence the organisation sees.

2. Token probability, calibration, semantic uncertainty and legal reliability are different questions, and the answer carries a decision-ready account of none of them.

3. Where material harm is foreseeable and proportionate controls are available, the response lies in disclosure, management of residual uncertainty and control of reliance.

I. Two Vantage Points

LLMs are trained to predict language; organisations use them to produce information. The model answers a probabilistic question: what continuation is likely here? The deployer faces a legal one: can this output be relied upon, by this person, for this purpose? Close enough to be confused; different enough to create liability.

The distinction should be kept precise. “Probabilistic” names the model-side vantage point: uncertainty represented internally as distributions over tokens—a token being the small unit of text, like a tile in a mosaic, over which the model scores continuations. “Stochastic” names the organisation-side vantage point: variable outputs whose reliability cannot be read from their form. The pairing is this brief’s shorthand rather than a settled technical taxonomy: “stochastic” here describes organisational consequences, not stochastic decoding. The gap persists under deterministic decoding, where variability is a symptom rather than the cause.

info@orestepollicino.eu
https://pollicinoaidvisory.eu/

The cause is the loss of epistemic context at the interface: a conclusion moves outward while the uncertainty that produced it stays behind. A lawyer sees a citation in a draft; a compliance officer sees a risk classification. The legal object is not the model's computation but the act of treating its output as usable.

Four things travel under the word uncertainty, and the governance problem is that they are not equivalent. Predictive uncertainty is dispersion across possible continuations; it does not measure the truth of a claim. Semantic uncertainty is dispersion across incompatible meanings the system produces; it is not automatically factual error. Factual uncertainty is doubt about correspondence with the world, and is not necessarily readable from the model's probabilities. Evidential uncertainty is the absence or weakness of sources allowing a claim to be checked, and does not coincide with anything internal to the model. Legal reliability is a fifth question: whether the output is fit to ground a specific action in a specific context. Each use of "uncertainty" below belongs to one of these.

Executive Takeaway

The deployer should not ask whether an output is fluent or statistically plausible. It should ask what kind of reliance the process permits.

II. The Lossy Pipeline

Lossy compression is a governance heuristic, not a theorem about what an LLM is. Delétang et al. show that language models can act as compressors; they do not establish that a model compresses the world.[1] Chiang's "blurry JPEG" is likewise a metaphor: models generalise, they do not reconstruct training items.[2] Both still track what a workflow carries forward: training uses a partial, dated corpus; generation returns a continuation rather than an auditable copy of its sources; and an output can supply detail absent from the user's evidence.

The pipeline has two consequential handoffs. Distribution-to-output reduces many candidates to one decoded answer; output-to-workflow then discards prompt context, caveats, sources and review status unless the deployer preserves them. Later stages cannot reconstruct the discarded context from the sampled output alone; verification must reintroduce evidence from an independent or authoritative source.

Figure 1 — The epistemic pipeline and the point of reliance

Stage	Information carried forward and governance significance
World	Facts, law, context, and events. The world is sampled only partially.
Training corpus	A partial and dated representation; the first material loss occurs as the world is represented selectively.
Model weights	The learned statistical model state derived from the corpus.
Output distribution	Candidate continuations and token probabilities; multiple continuations remain possible here.
Sampled output	One decoded answer. The distribution is reduced to one output, and available context may not travel with it.
Organisational record	A file note, advice, classification, or decision input. This is the point of reliance, where controls must match residual uncertainty.

[1] Delétang et al., *Language Modeling Is Compression* (ICLR 2024).
[2] Chiang, *ChatGPT Is a Blurry JPEG of the Web* (2023).

Executive Takeaway

Governance capacity has to sit at the point of reliance: that is where a generated sample becomes an organisational act.

III. Probability Is Not Reliability

Token probability is not legal reliability. Three findings in the technical literature explain why, and each matters to deployer-side counsel.

First, token-level confidence is poorly calibrated to factual correctness. Calibration is the difference between a forecast and a guess: if a system says "90 per cent", nine in ten comparable claims should be right. LLMs have often been observed not to behave that way, though the degree varies by model, task and elicitation method, and reinforcement learning from human feedback may worsen overconfidence.[3] Seeing token probabilities would not, on its own, give the deployer a legal reliability score.

Second, the legally relevant uncertainty often lives at the level of meaning, not tokens. Semantic entropy asks whether multiple answers say the same thing or scatter across incompatible meanings; Farquhar, Kossen, Kuhn and co-authors show it detects confabulations that token-level signals miss.[4]

Third, the sample does not say what kind of uncertainty it contains. Some variability is harmless linguistic variety; other uncertainty reflects ignorance about the world—whether the case exists, whether the policy is current. Machine-learning theory distinguishes aleatoric from epistemic uncertainty.[5] The sample arrives without a label saying which dominates.

Executive Takeaway

A deployer cannot convert model confidence into legal confidence without additional controls. Probability, calibration, meaning, source-grounding, and permitted reliance are different questions.

IV. The Legal Hinge: Foreseeable Error and Reliance

Error should not be treated as an exceptional defect awaiting a final patch, but the argument has to be kept in registers. Xu, Jain and Kankanhalli give a formal, computability-style result for LLMs as general problem-solvers; the proof reaches computable systems generally, so it does not by itself establish an LLM-specific legal

[3] Zhu, Chiwei, Benfeng Xu, Quan Wang, Yongdong Zhang and Zhendong Mao, *On the Calibration of Large Language Models and Alignment* (2023); Leng, Jixuan, Chengsong Huang, Banghua Zhu and Jiaxin Huang, *Taming Overconfidence in LLMs: Reward Calibration in RLHF* (2024).

[4] Farquhar, Kossen, Kuhn et al., *Detecting hallucinations in large language models using semantic entropy*, *Nature* 630, 625–630 (2024).

[5] Hüllermeier and Waegeman, *Aleatoric and Epistemic Uncertainty in Machine Learning*, *Machine Learning* 110, 457–506 (2021).

conclusion.[6] Kalai, Nachum, Vempala and Zhang give a statistical account: pre-training produces factual errors, and accuracy-centred evaluation rewards guessing over abstention.[7] Empirical work has separately profiled material hallucination rates in legal tasks. The organisational inference drawn here is narrower than any of the three: in certain uses a class of material error is foreseeable in advance, even though no particular output can be predicted.[8]

Foreseeability alone does not establish breach. The governance case becomes stronger where the potential harm is material, reliance is causally relevant, and proportionate controls are reasonably available. Where those conditions hold, the response cannot be to eliminate error in every output; it attaches to disclosing known limitations, managing residual uncertainty and controlling reliance. The first two are matters of candour and record; the third is a question of design. For negligence analysis, the relevant organisational failure may lie not in the existence of an erroneous output, but in permitting reliance without controls proportionate to a foreseeable risk.

This is why "use AI, but check it" is too weak: it does not say what must be checked, by whom, against which source, before which kind of reliance, with what record. Repeatedly acceptable outputs also weaken vigilance. The deployer needs an architecture, not a slogan.

Executive Takeaway

The foreseeable failure is not in the model. It sits in the workflow that lets a confident sample pass into advice, assessment or decision.

V. Governance as Uncertainty Management

Organisation theory does not prove the duty; it supplies design heuristics for locating and sizing controls. March and Simon called one mechanism "uncertainty absorption": inferences are drawn from evidence, and the inference rather than the evidence is communicated—the memo arrives without the file.[9] An LLM industrialises it. The training data, the distribution, retrieved sources, the prompt path and the uncertainty signals stay behind; what travels is a finished paragraph. The actor who absorbs uncertainty also shifts risk downstream, so responsibility is already shaped by the information flow.

Galbraith's information-processing view adds the next step. Uncertainty is the gap between what a task requires and what the organisation has; organisations respond by reducing that need or by increasing the capacity to meet it.[10] Every usable control belongs to one family or the other: use-case restrictions, templates and grounding reduce the burden; review, logging and escalation increase capacity.

---

[6] Xu, Jain and Kankanhalli, *Hallucination is Inevitable: An Innate Limitation of Large Language Models* (2024).

[7] Kalai, Nachum, Vempala and Zhang, *Why Language Models Hallucinate* (OpenAI, 2025).

[8] Dahl, Magesh, Suzgun and Ho, *Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models*, *Journal of Legal Analysis* 16(1) (2024).

[9] March and Simon, *Organizations* (1958).

[10] Galbraith, *Organization Design: An Information Processing View*, *Interfaces* 4(3) (1974).

Ashby's law of requisite variety supplies an adequacy test rather than a mandate for any particular architecture: a regulator must possess sufficient variety to absorb the variety of the regulated system.[11] Here "variety" means the space of possible erroneous outputs a use case can generate, not run-to-run randomness, so the test holds under deterministic decoding. Deployment architecture is therefore variety engineering: grounding, retrieval and intended-use constraints attenuate the failure space; review, sampling and sign-off amplify capacity.

Governance does not eliminate error; it structures reliance around the remaining failure space. The operational question is whether regulatory capacity matches residual variety at the point of reliance. A chatbot used for brainstorming and an LLM embedded in customer communications do not need the same controls. And the question can be put to a concrete use case today, without any settled taxonomy, by asking four things: what errors this use can produce; what impact they would have; which controls would prevent or detect them; and what risk remains once those controls are in place.

VI. Conclusions

The chain is now explicit. Training learns statistical structure from a partial and dated corpus. Generation returns one continuation drawn from a distribution of possible ones. Organisational flow strips what remains of the context: the memo arrives without the file. Reliance then occurs at a specific point in a specific process, and the question there is not whether the model was right, but whether residual uncertainty is matched by the capacity to absorb it.

What this brief delivers is that question, not an architecture: organisations will answer it with different controls. The duty it implies has three limbs—disclose what is known about the system's limits, manage what remains uncertain, and decide in advance what the output may be relied upon to do. Three decisions can be taken before any taxonomy is adopted: identify the point at which the output is relied upon; classify the potential harm at that point; and assign an owner and a control before the use begins. The deployer needs no metaphysics of language models in order to act.

The Intuition in One Line

AI governance is the discipline of preventing a lossy generated sample from becoming an unexamined organisational fact.

---

[11] Ashby, *An Introduction to Cybernetics* (1956).