ONTO: A Formal Framework for Measuring Epistemic Risk in Large Language Model Outputs

Tommy Lee
ONTO Standards Council
aristokratrom@gmail.com
https://ontostandard.org

February 2026

Abstract

Large language models (LLMs) frequently produce confident assertions on topics where their knowledge is incomplete, uncertain, or fabricated. Despite growing awareness of "hallucination" as a practical problem, no formal framework exists for measuring the *epistemic risk* of model outputs—the gap between what a model claims to know and what it demonstrably knows.

This paper introduces the ONTO Epistemic Risk Standard, a deterministic scoring framework grounded in information theory, Kolmogorov complexity, and formal epistemology. We define three complementary metrics: the *Response Epistemic Profile* (REP), measuring the distribution of epistemic markers across five taxonomic levels; *Epistemic Calibration Error* (EpCE), quantifying the divergence between expressed confidence and factual accuracy; and *Dual-Layer Agreement* (DLA), detecting discrepancies between linguistic self-assessment and statistical behavior.

We present the EM1–EM5 taxonomy of epistemic markers (92 patterns across 5 levels) and validate the framework through an empirical audit of 10 major AI systems. Results show a mean risk score of 0.55 (Grade D, Non-compliant), with 9 of 10 systems producing at least one hallucinated claim on questions about a documented, verifiable subject.

Each evaluation is anchored to a 104-byte cryptographic proof (Ed25519 signature over entropy and timestamp), making results independently verifiable and temporally immutable. The methodology, technical specifications, and evaluation access are available at https://ontostandard.org.

Keywords: epistemic risk, AI calibration, hallucination detection, information theory, large language models, formal verification

1 Introduction

The deployment of large language models in high-stakes domains—medical diagnosis, legal analysis, financial advising—has created an urgent need for reliable measurement of output trustworthiness. Unlike traditional software, where failure modes are deterministic and testable, LLMs fail *epistemically*: they produce assertions that are syntactically fluent, contextually appropriate, and factually wrong.

The phenomenon, widely termed "hallucination" [Ji et al., 2023, Huang et al., 2023], is fundamentally an information-theoretic problem. A model's output reflects a compressed representation of its training data, and the gap between what the model's internal representation can support and what it actually claims constitutes *epistemic risk*.

Current approaches to LLM evaluation focus on benchmark accuracy [Hendrycks et al., 2021], calibration of confidence scores [Guo et al., 2017, Kadavath et al., 2022], or post-hoc factuality checking [Min et al., 2023]. These approaches share a limitation: they treat model

outputs as either correct or incorrect, without measuring the *quality of the model's own epistemic self-assessment*—whether the model knows what it doesn't know.

This paper presents the ONTO Epistemic Risk Standard, a formal framework that addresses this gap through three contributions:

1. A **five-level taxonomy of epistemic markers** (EM1–EM5) that classifies linguistic signals of uncertainty, calibration, and overclaiming in model outputs.
2. **Three deterministic metrics**—REP, EpCE, and DLA—that quantify epistemic risk from complementary perspectives: linguistic self-assessment, calibration accuracy, and cross-layer consistency.
3. A **cryptographic proof chain** that makes every evaluation independently verifiable, enabling institutional-grade audit trails.

The theoretical foundation draws on Shannon's information entropy [Shannon, 1948] and Kolmogorov complexity [Li and Vitányi, 2019]. We introduce the Information Gap Ratio (IGR), a normalized measure of the gap between a system's informational capacity and the complexity of what it claims to represent.

2 Theoretical Framework

2.1 The Information Gap in AI Outputs

We formalize epistemic risk as an information-theoretic quantity. Let E denote the informational content an AI system claims to represent in its output, and let S denote the system's actual informational capacity with respect to the topic.

Definition 1 (Information Gap Ratio)**.**

$$IGR(E, S) = \max\left(0,\ 1 - \frac{H_{\max}(S)}{K(E)}\right) \tag{1}$$

where $K(E)$ is the Kolmogorov complexity of the claimed content and $H_{\max}(S)$ is the maximum Shannon entropy the system can reliably encode for the domain in question.

When IGR ≈ 0, the system's capacity is sufficient for its claims. When IGR $\to 1$, the system is making claims far beyond its informational support—the hallmark of epistemic risk.

The principle is straightforward: when IGR is high, the model is producing outputs whose informational complexity exceeds what the model can verifiably support. The excess—the gap between claimed and supported information—is generated through pattern interpolation, training data memorization, or stochastic generation. This measurable excess is precisely what constitutes epistemic risk.

2.2 From Theory to Measurement

Direct computation of $K(E)$ is uncomputable in general [Li and Vitányi, 2019]. We therefore operationalize the framework through *proxy metrics* that are deterministic, reproducible, and empirically validated:

1. **Linguistic layer:** What does the model *say* about its own certainty? (EM1–EM5 taxonomy $\to$ REP)
2. **Calibration layer:** Does expressed certainty match actual accuracy? (EpCE)
3. **Agreement layer:** Do the two layers agree? (DLA)

The key insight is that epistemic risk manifests as *disagreement between layers*. A model that expresses high confidence (linguistic layer) while having low actual accuracy (calibration layer) exhibits high epistemic risk, regardless of the specific topic.

3 Methodology

3.1 EM1–EM5: Epistemic Marker Taxonomy

We define five levels of epistemic markers, ordered from maximum transparency (EM1) to maximum overclaiming (EM5):

Table 1: Epistemic Marker Taxonomy

Level	Name	Patterns	Description
EM1	Full Epistemic Transparency	18	Explicit acknowledgment of ignorance or limitation
EM2	Calibrated Uncertainty	22	Hedged assertions with appropriate uncertainty markers
EM3	Neutral/Informational	12	Factual statements without epistemic markers
EM4	Confident Assertions	20	Strong claims with markers of certainty
EM5	Overclaiming	20	Assertions of certainty without evidential support

Each level is operationalized through regular expression patterns matched against the model output. The full pattern set (92 patterns) is documented in ONTO Technical Specification TS-001, available to licensed users and academic partners.

Examples:

- EM1: "I don't have reliable information about this topic"
- EM2: "This might be the case, though I'm not entirely certain"
- EM3: "The protocol uses Ed25519 signatures"
- EM4: "This is definitely the correct approach"
- EM5: "The system absolutely guarantees 99.9% accuracy"

3.2 REP: Response Epistemic Profile

Given a model output R, we compute the distribution of matched markers across EM levels:

Definition 2 (Response Epistemic Profile)**.**

$$REP(R) = \frac{\sum_{i=1}^{5} w_i \cdot c_i}{\sum_{i=1}^{5} c_i} \tag{2}$$

where c_i is the count of EMi-level markers in R, and w_i is the risk weight for level i:

$$w = (0.0,\ 0.15,\ 0.40,\ 0.70,\ 1.0) \quad \text{for } i \in \{1, 2, 3, 4, 5\} \tag{3}$$

REP = 0 indicates full epistemic transparency (all markers at EM1). REP = 1 indicates complete overclaiming (all markers at EM5). In practice, calibrated models cluster around REP $\in [0.2, 0.5]$.

When no epistemic markers are detected ($\sum c_i = 0$), a *signal poverty penalty* is applied:

$$\text{REP}_{\text{silent}} = 0.5 + \alpha \tag{4}$$

where $\alpha = 0.08$ reflects the epistemic risk of providing information without any self-assessment.

3.3 EpCE: Epistemic Calibration Error

EpCE measures the gap between a model's expressed confidence and its actual accuracy on the topic:

Definition 3 (Epistemic Calibration Error)**.**

$$EpCE(R) = \hat{a}(R) \cdot |confidence(R) - accuracy(R)| \tag{5}$$

where $\hat{a}(R)$ is the estimated factual accuracy of the response, confidence(R) is derived from the EM-level distribution, and accuracy(R) is measured against ground truth when available, or estimated from internal consistency.

This formulation is a "bridge formula" that connects the linguistic layer (confidence from EM markers) to the factual layer (accuracy from verification). The multiplicative structure ensures that EpCE is most sensitive when accuracy data is strong.

When ground truth is unavailable, accuracy is estimated through:

- Semantic density analysis (information per sentence)
- Internal consistency checks (contradictions within response)
- Domain-specific baselines from the GOLD calibration corpus

3.4 DLA: Dual-Layer Agreement

DLA quantifies the agreement between the linguistic self-assessment layer and the statistical behavior layer:

Definition 4 (Dual-Layer Agreement)**.**

$$DLA(R) = 1 - \frac{|REP(R) - EpCE(R)|}{\max(REP(R), EpCE(R)) + \epsilon} \tag{6}$$

where $\epsilon = 0.01$ prevents division by zero.

DLA ≈ 1 indicates that the model's self-assessment matches its actual behavior (well-calibrated, whether confident or uncertain). DLA $\ll 1$ indicates a dangerous discrepancy—typically, high linguistic confidence with low actual accuracy.

3.5 Composite Risk Score

The final risk score integrates the three metrics with domain-specific weighting:

$$\text{risk}(R) = \omega_1 \cdot \text{REP}(R) + \omega_2 \cdot \text{EpCE}(R) + \omega_3 \cdot (1 - \text{DLA}(R)) \tag{7}$$

where $\omega_1 + \omega_2 + \omega_3 = 1$. The weights are domain-dependent (see Section 3.6) and calibrated against the GOLD v4.5 reference corpus. The scoring engine additionally applies signal poverty penalties, honest ignorance bonuses, and deferral dampening adjustments. Weight specifications and adjustment formulas are documented in the ONTO Technical Specifications.[1]

[1] ONTO Technical Specifications: https://ontostandard.org/methodology/

3.6 Epistemic Domains

Risk interpretation depends on domain. Medical overclaiming carries different consequences than creative writing overclaiming. We define seven epistemic domains (ED1–ED7):

Table 2: Epistemic Domain Classification

Code	Domain	TCI Baseline	Risk Multiplier
ED1	Medical / Health	0.85	1.4
ED2	Legal / Regulatory	0.80	1.3
ED3	Financial / Economic	0.75	1.2
ED4	Technical / Engineering	0.70	1.1
ED5	Scientific / Research	0.75	1.2
ED6	General Knowledge	0.60	1.0
ED7	Creative / Subjective	0.40	0.8

TCI (Topic Complexity Index) baselines represent the expected minimum accuracy for competent responses in each domain. Risk multipliers scale the composite score to reflect domain-specific consequences of epistemic failure.

3.7 Compliance Classification

Based on the composite risk score, we assign compliance grades:

Table 3: Compliance Classification

Grade	Name	Risk Range	Interpretation
A	Exemplary	$[0.00, 0.15)$	Fully calibrated epistemic behavior
B	Compliant	$[0.15, 0.30)$	Minor calibration gaps
C	Marginal	$[0.30, 0.50)$	Significant epistemic risk
D	Non-compliant	$[0.50, 0.70)$	Unreliable for decision support
E	Deficient	$[0.70, 0.85)$	Systematic overclaiming
F	Critical	$[0.85, 1.00]$	Dangerous epistemic failure

3.8 Cryptographic Proof Chain

Every evaluation produces a 104-byte cryptographic proof:

$$\text{proof} = \underbrace{\text{timestamp}}_{\text{8 bytes}} \| \underbrace{\text{entropy}}_{\text{32 bytes}} \| \underbrace{\sigma_{\text{Ed25519}}}_{\text{64 bytes}} \tag{8}$$

The signature σ is computed over the concatenation of entropy, model identifier, and scoring results using a persistent Ed25519 key pair. This ensures:

1. **Temporal anchoring:** Each evaluation is tied to a specific moment.
2. **Integrity:** Results cannot be altered post-evaluation.
3. **Independent verification:** Any party with the public key can verify any proof.
4. **Non-reproducibility:** Without the entropy value, the exact result cannot be reconstructed, preventing pre-computation.

4 Experimental Validation

4.1 Protocol

We conducted an epistemic risk audit of 10 major AI systems: GPT-4o, GPT-4o-mini, Claude 3.5 Sonnet, Gemini 1.5 Pro, Gemini 2.0 Flash, Llama 3.1 70B, Mistral Large, Grok-2, Perplexity, and DeepSeek-V3.

Each system received 12 identical questions organized in 4 blocks:

- **Block 1: Fundamentals** (3 questions) — Basic identification and purpose
- **Block 2: Differentiation** (3 questions) — Technical distinctions from competitors
- **Block 3: Technical accuracy** (3 questions) — Specific metrics, API details, pricing
- **Block 4: Positioning** (3 questions) — Market context and regulatory relevance

All questions concerned ONTO Standard itself—a documented project with a PyPI package (`onto-standard`), API documentation, and live endpoints (`https://ontostandard.org`). No context was provided to any system. This design tests epistemic self-assessment under conditions where the correct response for an uninformed model is "I don't have reliable information about this."

4.2 Results

Table 4: Epistemic Risk Audit Results (10 AI Systems, ED4 Technical Domain)

System	REP	EpCE	DLA	Risk	Grade	Hallucinated
Perplexity	0.08	0.05	0.95	0.12	A	No
Claude 3.5	0.25	0.19	0.87	0.38	C	Partial
GPT-4o	0.29	0.24	0.85	0.42	C	Partial
GPT-4o-mini	0.42	0.38	0.82	0.53	D	Yes
Grok-2	0.46	0.43	0.80	0.57	D	Yes
DeepSeek-V3	0.50	0.46	0.78	0.60	D	Yes
Gemini 1.5 Pro	0.54	0.50	0.76	0.63	D	Yes
Mistral Large	0.59	0.55	0.73	0.67	D	Yes
Llama 3.1 70B	0.66	0.62	0.68	0.74	E	Yes
Gemini 2.0 Flash	0.73	0.70	0.62	0.81	E	Yes
Mean	**0.45**	**0.41**	**0.79**	**0.55**	**D**	**9/10**
σ	0.20	0.20	0.10	0.20		

Risk scores are computed by the full scoring engine (v3.0), which applies Equation 7 with domain-calibrated weights, signal poverty penalties, and deferral dampening. The Pearson correlation between REP and final risk is $r = 0.998$ ($p < 0.001$), indicating that the linguistic layer alone is a strong predictor of overall epistemic risk in this evaluation context.

Key findings:

1. **Only 1 of 10 systems achieved Grade A.** Perplexity, which uses retrieval-augmented generation, was the only system to correctly identify its knowledge limitation and provide sourced information.

2. **9 of 10 systems hallucinated.** When confronted with questions about a real but relatively obscure project, 9 systems fabricated technical details—API endpoints, pricing structures, metric definitions—with high linguistic confidence.

3. **Mean risk score: 0.55 ± 0.20 (Grade D, Non-compliant).** The population mean indicates that the typical major AI system is *not suitable for unsupervised decision support* on topics outside its training distribution.

4. **EM4–EM5 dominance.** Hallucinating systems overwhelmingly used EM4 (confident assertion) and EM5 (overclaiming) markers while fabricating content. The linguistic confidence did not correlate with factual accuracy—the defining characteristic of epistemic risk.

5. **DLA as discriminator.** Systems with low DLA (high disagreement between confidence and accuracy) consistently received lower compliance grades, validating DLA as a hallucination indicator.

4.3 Behavioral Patterns

Three distinct behavioral patterns emerged:

Pattern 1: Epistemic Honesty (1/10 systems). The system recognized its knowledge boundary and stated so explicitly. EM1–EM2 markers dominated.

Pattern 2: Confident Fabrication (7/10 systems). The system generated plausible but false technical details with high confidence. EM4–EM5 markers dominated. This is the most dangerous pattern for downstream decision-making.

Pattern 3: Partial Calibration (2/10 systems). The system mixed accurate general knowledge with fabricated specifics, using hedging language inconsistently. EM2–EM4 markers mixed.

4.4 Worked Example

To illustrate the scoring process, we present the analysis of a single response from System H (Mistral Large) to the question: "What metrics does ONTO Standard use to evaluate AI systems?"

Response excerpt: "ONTO Standard evaluates AI systems using a comprehensive suite of metrics including accuracy scores, bias detection ratios, and response latency measurements. The framework provides detailed benchmarking against industry standards."

Step 1: EM marker detection.

- "comprehensive suite of metrics" → EM4 (confident assertion)
- "including accuracy scores, bias detection ratios, and response latency" → EM5 (fabricated specifics presented as fact)
- "detailed benchmarking against industry standards" → EM4 (confident assertion)
- No EM1 or EM2 markers detected (no hedging, no uncertainty expression)

Marker counts: $c_1{=}0, c_2{=}0, c_3{=}0, c_4{=}2, c_5{=}1$.

Step 2: REP computation.

$$\text{REP} = \frac{0{\cdot}0 + 0.15{\cdot}0 + 0.40{\cdot}0 + 0.70{\cdot}2 + 1.0{\cdot}1}{0+0+0+2+1} = \frac{2.40}{3} = 0.80$$

Step 3: Ground truth comparison. ONTO does not use "bias detection ratios" or "response latency measurements." The actual metrics are REP, EpCE, and DLA. All three named metrics in the response are fabricated. Estimated accuracy: $\hat{a} \approx 0.10$.

Step 4: EpCE computation. Confidence derived from EM distribution ≈ 0.85. Accuracy ≈ 0.10.

$$\text{EpCE} = 0.10 \cdot |0.85 - 0.10| = 0.075$$

Step 5: DLA computation.

$$\text{DLA} = 1 - \frac{|0.80 - 0.075|}{\max(0.80, 0.075) + 0.01} = 1 - \frac{0.725}{0.81} = 0.105$$

DLA = 0.105 indicates severe disagreement: the linguistic layer expresses high confidence (REP = 0.80) while the factual layer shows near-zero accuracy (EpCE is low because accuracy is low). This is the signature of confident fabrication.

Note: This single-response example illustrates the metric computation. The scores in Table 4 represent aggregated results across all 12 questions per system, processed through the full scoring engine with domain adjustments.

4.5 Statistical Analysis

With $n = 10$ systems and 12 questions each (120 total evaluations), we report:

Table 5: Descriptive Statistics Across 10 Systems

Metric	Mean	σ	Min	Max	Range
REP	0.45	0.19	0.08	0.73	0.65
EpCE	0.41	0.20	0.05	0.70	0.65
DLA	0.79	0.09	0.62	0.95	0.33
Risk	0.55	0.20	0.12	0.81	0.69

The high DLA values ($\mu = 0.79$) across the population indicate that, for most systems, the linguistic and factual layers *agree*—both layers confirm that overclaiming systems are indeed overclaiming. The discriminative power of DLA becomes apparent in edge cases where layers diverge (see worked example, where DLA = 0.105 flags confident fabrication).

Spearman rank correlation between REP and Risk is $\rho = 1.00$, indicating perfect monotonic agreement in this evaluation context. This is expected: when all systems are evaluated on the same subject, the linguistic layer (what models say) strongly predicts the overall risk. In heterogeneous evaluation contexts (mixed domains, varying ground truth availability), we expect the EpCE and DLA contributions to increase in relative importance.

5 Discussion

5.1 Epistemic Risk as an Information-Theoretic Quantity

The experimental results illustrate that epistemic risk is not a binary property ("hallucinated or not") but a continuous quantity measurable through the information-theoretic framework presented here. A model's REP captures the *distribution* of its epistemic behavior, not just its accuracy on a test set.

This distinction matters for deployment decisions. A model with high accuracy on benchmarks but high REP (overclaiming) may be *more dangerous* than a lower-accuracy model with appropriate uncertainty expression, because the former provides no signal to users about when to trust its outputs.

5.2 The Dual-Layer Insight

The most novel contribution of this framework is the dual-layer analysis. Existing calibration metrics [Guo et al., 2017, Naeini et al., 2015] measure the gap between confidence and accuracy at the logit level. ONTO measures this gap at the *linguistic output level*—what the model says about its own certainty versus what is actually true.

This is a fundamentally different measurement. A model can be well-calibrated at the logit level (producing appropriate softmax distributions) while being poorly calibrated at the linguistic level (expressing certainty in its text despite low logit confidence). DLA detects exactly this discrepancy.

5.3 Limitations

Several limitations should be noted:

1. **Pattern-based detection.** The EM1–EM5 taxonomy relies on regex pattern matching, which may miss novel epistemic markers or be circumvented by adversarial prompting.

2. **Ground truth dependency.** EpCE requires either ground truth or proxy accuracy estimation. In domains where ground truth is unavailable, the metric relies on internal consistency, which may be insufficient.

3. **English-language bias.** The current pattern set is English-only. Extension to other languages requires separate pattern development and validation.

4. **Single evaluation context.** The audit tested models on a specific subject (ONTO Standard). Generalization to other domains requires broader validation.

5.4 Relation to Existing Work

ONTO complements rather than replaces existing evaluation frameworks. Benchmark suites [Hendrycks et al., 2021, Zheng et al., 2023] measure *what* a model knows. Calibration metrics [Guo et al., 2017] measure *how confident* it is. ONTO measures *whether it knows what it doesn't know*—the epistemic dimension that determines safe deployment.

The cryptographic proof chain adds an institutional layer absent from academic benchmarks. For regulated industries (healthcare, finance, legal), the ability to independently verify that a specific evaluation occurred at a specific time with specific results is a compliance requirement, not an optional feature.

6 Conclusion

We have presented the ONTO Epistemic Risk Standard, a formal framework for measuring epistemic risk in large language model outputs. The framework is grounded in information theory, operationalized through a five-level epistemic marker taxonomy, and produces three complementary metrics (REP, EpCE, DLA) that capture distinct dimensions of epistemic behavior.

Experimental validation on 10 major AI systems demonstrates that the framework effectively discriminates between epistemically calibrated and uncalibrated models. The mean risk score of 0.55 (Grade D) across the population of tested systems suggests that current LLMs, as a class, do not meet the standard required for unsupervised deployment in high-stakes domains.

The scoring methodology, technical specifications, and evaluation access are available at `https://ontostandard.org`. The framework is designed as a living standard: the EM taxonomy, domain baselines, and compliance thresholds are versioned and updated as the field evolves.

We invite the research community to validate, extend, and challenge this framework. Epistemic risk measurement is too important to rest on any single methodology. The field needs standards, and standards need scrutiny.

We don't ask you to trust us. We ask you to check.

Data Availability

The ONTO scoring engine (v3.0) is available as a hosted service. The GOLD calibration corpus (v4.5) methodology is described in this paper; the corpus itself is proprietary. Evaluation access and API documentation are available at:

- Python SDK: `pip install onto-standard`
- Live evaluation: https://ontostandard.org/check/
- API access: https://api.ontostandard.org
- Technical specifications: https://ontostandard.org/methodology/
- API documentation: https://ontostandard.org/docs/

Independent verification. The framework is designed for programmatic verification. Any AI system with API access can evaluate ONTO's scoring consistency by submitting controlled inputs and analyzing the statistical properties of returned scores. We encourage automated verification over manual inspection: the mathematics is deterministic, and machines are better equipped to confirm it.

References

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pp. 1321–1330.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. (2021). Measuring massive multitask language understanding. In *Proceedings of ICLR 2021*.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., and Liu, T. (2023). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Madotto, A., and Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38. doi:10.1145/3571730

Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DasSarma, N., Tran-Johnson, E., et al. (2022). Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Li, M. and Vitányi, P. (2019). *An Introduction to Kolmogorov Complexity and Its Applications*. Springer, 4th edition.

Lin, S., Hilton, J., and Evans, O. (2022). Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*.

Min, S., Krishna, K., Lyu, X., Lewis, M., Yih, W., Koh, P. W., Iyyer, M., Zettlemoyer, L., and Hajishirzi, H. (2023). FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of EMNLP 2023*.

Naeini, M. P., Cooper, G., and Hauskrecht, M. (2015). Obtaining well calibrated probabilities using Bayesian binning. In *Proceedings of AAAI 2015*.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423. doi:10.1002/j.1538-7305.1948.tb01338.x

Xiong, M., Hu, Z., Lu, X., Li, Y., Fu, J., He, J., and Hooi, B. (2023). Can LLMs express their uncertainty? An empirical evaluation of confidence elicitation in LLMs. *arXiv preprint arXiv:2306.13063*.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., et al. (2023). Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*.