

Rethinking Aleatoric Uncertainty: Partitioning Irreducible States for Large Language Model Agents

Renato Okabayashi Miyaji

Escola Politécnica, Universidade de São Paulo
São Paulo, Brazil
re.miyaji@usp.br

Matias Valdenegro-Toro

Bernoulli Institute
University of Groningen, The Netherlands
m.a.valdenegro.toro@rug.nl

Abstract

While Uncertainty Quantification (UQ) remains an active research field for Large Language Models (LLMs), the current literature lacks a rigorous understanding of what constitutes aleatoric uncertainty and its exact sources. Existing methodologies largely rely on the traditional dichotomy of aleatoric and epistemic uncertainty, which inadequately treats uncertainty as a static snapshot. To bridge this gap, this position paper challenges the static treatment of uncertainty. Supported by empirical evidence, we posit that aleatoric uncertainty in LLM agents is fundamentally dynamic and must be partitioned into strictly and temporarily irreducible states—differentiating intrinsic task variance, such as multiple valid answers, from resolvable noise, including input ambiguity, typographical errors, and contextual conflict. Through episodic empirical evaluations across multi-turn scenarios, we demonstrate how targeted agentic interventions can actively resolve temporarily irreducible noise, transitioning the interaction into an epistemic regime. When data uncertainty is strictly irreducible, model interventions cannot isolate a single correct distribution. While we acknowledge that the traditional static dichotomy remains a valuable framework for non-agentic settings, we urge the community to develop new decomposition paradigms that track uncertainty along the timeline of interaction. Embracing this dynamic perspective is essential to achieve robust uncertainty quantification for modern LLM agents.

1 Introduction

While Uncertainty Quantification (UQ) remains a deeply active research field in general Machine Learning (He et al., 2026b), and increasingly for LLMs (Liu et al., 2025), its theoretical foundations differ significantly across paradigms. In classical models, the formalization and decomposition of uncertainty are defined and standardized (Hüllermeier and Waegeman, 2021), traditionally divided

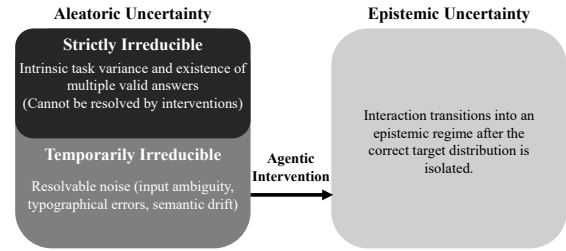


Figure 1: Rethinking the traditional static dichotomy of uncertainty. In LLM agents, aleatoric uncertainty is dynamic and partitioned into strictly and temporarily irreducible states. While strictly irreducible data noise remains unresolved, targeted agentic interventions can actively resolve temporarily irreducible noise, transitioning the state into an epistemic regime.

into aleatoric and epistemic uncertainty.

Consider the context of estimating uncertainty over $y \in \mathcal{Y}$ for an instance $x \in \mathcal{X}$ given a hypothesis space H . Generally, aleatoric uncertainty is associated with data-centric factors, such as measurement errors, the inherent randomness of the measured signal (Wimmer et al., 2023), and the insufficiency of x to fully explain y (Malinin, 2019). Conversely, epistemic uncertainty is tied to the model itself, often relating to the lack of knowledge required to approximate the estimated distribution $\hat{P}(\cdot | x)$ to the true distribution $P(\cdot | x)$. Theoretically, this type of uncertainty can be reduced by training the model with additional data (Senge et al., 2014).

However, this perspective is subject to debate. Several authors argue that treating aleatoric and epistemic uncertainties as immutable categories is flawed (Hüllermeier and Waegeman, 2021). Rather, these classifications are tied to the chosen problem formulation, the configuration $(\mathcal{X}, \mathcal{Y}, H, P)$. A shift in this underlying structure can cause an interchange between the two types, where the inherent

data noise in one setup becomes the correctable ignorance of the model in another (Hüllermeier and Waegeman, 2021). This aligns with the understanding that what is defined as irreducible noise in a specific architecture could reflect a lack of knowledge within a different modeling approach (Kiureghian and Ditlevsen, 2009). For LLMs, it becomes significantly more complex (Shorinwa et al., 2025). Due to various factors, such as prompt rewriting and context shifts, the underlying source of uncertainty can fluctuate, making its definition difficult to evaluate (Yang et al., 2025b).

For instance, consider an agent that already knows the exact time and location of a user’s outing but lacks the parametric knowledge of the weather forecast. Initially, this uncertainty is epistemic, reflecting a gap in the model’s knowledge. To resolve this, the agent executes an Information Retrieval (IR) step. However, if the retrieved weather report is inherently imprecise, such as predicting a ‘50% chance of scattered showers’, the agent still cannot provide a definitive answer. At this point, the uncertainty shifts to aleatoric, as it is no longer driven by a lack of information, but by the fundamental ambiguity and unpredictability of the retrieved data itself.

Conversely, consider a conversational scenario where an LLM receives an inherently ambiguous user query. This ambiguity can stem from multiple factors, ranging from general phrasing—such as asking ‘Who was the best in the 2025 NBA?’, which could refer to the championship-winning team or the MVP—to input noise like spelling mistakes. Furthermore, ambiguity can arise from semantic drift, where words like ‘multitude’ or ‘plane’ exhibit diachronic semantic shifts and require temporal anchoring to distinguish between their historical and modern meanings (Schlechtweg et al., 2020). Initially, these instances manifest as high aleatoric uncertainty due to the noise of the input signal. Nevertheless, the model can actively issue clarifying questions to disambiguate the user’s intent (Zhang and Choi, 2025), whether by confirming a target entity, correcting orthographic errors, or aligning the temporal context. Because this temporarily irreducible uncertainty can be resolved through interaction, it is reframed as epistemic uncertainty as the agent isolates the correct target distribution.

Consequently, Kirchhof et al. (2025) highlight that the traditional dichotomy and decomposition of uncertainty into aleatoric and epistemic compo-

nents are inadequate for LLM-based agents. This inadequacy stems primarily from the models’ intrinsically dynamic characteristics and their capacity for multi-turn interactions. The closest prior work to ours is the paper by Kirchhof et al. (2025). We share the high-level observation that the static aleatoric–epistemic dichotomy is inadequate for interactive LLM agents; however, our contributions are distinct and complementary. Kirchhof et al. (2025) diagnose three structural challenges for agent UQ—uncertainty-type interchange, multi-step propagation, and goal-directed reasoning—and issue a call for new decomposition frameworks, but do not propose a concrete partition of aleatoric uncertainty nor an operational criterion for distinguishing its sub-types. Our paper responds directly to that call. We contribute (i) a specific partition of aleatoric uncertainty into strictly and temporarily irreducible states, (ii) a two-stage Actionable Decision Protocol that converts the partition into agent behaviour, and (iii) empirical evidence across five experimental scenarios showing that the two states produce measurably different AU trajectories under identical estimation procedures.

Despite these recognized limitations, the prevailing methodology in the field has not yet fully adapted. An extensive body of literature explores techniques that aim not only to decompose (Hou et al., 2024a; Walha et al., 2025; Jayasekera et al., 2025) but also to quantify uncertainty in LLMs by relying on the traditional definitions of aleatoric and epistemic uncertainty (Liu et al., 2025; Shorinwa et al., 2025; He et al., 2026a). Within these works, a consensus emerges regarding epistemic uncertainty, which is generally treated as the reducible fraction of uncertainty stemming from the model itself (Melo et al., 2026; Hamidieh et al., 2026).

However, the definition and underlying sources of aleatoric uncertainty remain subjects of ongoing debate. Some authors approach aleatoric uncertainty through a Bayesian lens, conceptualizing it as the irreducible uncertainty that would persist even with an ideal (Bayes-optimal) estimator. In this framing, it is viewed as irreducible uncertainty that reflects the intrinsic ambiguity and ground-truth variance of the external environment itself (Hüllermeier and Waegeman, 2021; Kuhn et al., 2023; Aichberger et al., 2025)—due to the existence of multiple valid answers (Tomov et al., 2025; Yang et al., 2025a; Abbasi-Yadkori et al., 2024; Devic et al., 2025)—as well as the inherent stochas-

ticity of the generation process (Wang et al., 2025). Other studies model this uncertainty by focusing on input and task characteristics, attributing it to ambiguity in the user’s prompt (Hou et al., 2024a; Walha et al., 2025; Zhang et al., 2025; Jayasekera et al., 2025), or context conflict (Ling et al., 2024).

Consequently, we identify a critical gap in the current literature: a rigorous understanding of what constitutes aleatoric uncertainty in LLMs and the exact nature of its sources is still lacking. To bridge this gap, this position paper challenges the static treatment of uncertainty. Supported by empirical evidence, we posit that *aleatoric uncertainty in LLM agents is fundamentally dynamic and must be partitioned into strictly and temporarily irreducible states—differentiating intrinsic task variance, such as multiple valid answers, from resolvable noise, including input ambiguity, typographical errors, and contextual conflict*. In this context, we aim to contribute to a conceptual shift in the field through the following key contributions:

- The proposition of a novel taxonomy for aleatoric uncertainty in LLM agents, partitioning it into strictly and temporarily irreducible states to explicitly differentiate intrinsic task variance from interactively resolvable noise.
- The curation of targeted evaluation datasets, including the introduction of a novel benchmark for diachronic semantic drift, to empirically demonstrate how estimators for aleatoric and epistemic uncertainty behave across scenarios of input ambiguity, data noise, multiple answers, and context conflict.
- A counter-position and discussion exploring the contexts where the traditional dichotomy remains a valuable framework.

2 Aleatoric Uncertainty in LLMs

To critically evaluate how aleatoric uncertainty is currently treated in LLMs, we must examine the prevailing approaches. First, we establish the most frequent definitions of uncertainty within the literature. Building upon the standard information-theoretic decomposition of uncertainty (Depeweg et al., 2018), the predictive uncertainty for an output $y \in \mathcal{Y}$ given an input $x \in \mathcal{X}$ can be expressed as shown in Equation 1.

$$\mathcal{H}(\hat{P}(y | x)) = E_{\hat{P}(\theta|D)}[\mathcal{H}(\hat{P}(y | x, \theta))] + I(y; \theta | x) \quad (1)$$

In this formulation, $\mathcal{H}$ represents the Shannon entropy (Shannon, 1948) and I denotes the mutual information. The distribution $\hat{P}(y | x)$ corresponds to the model’s predictive output, while θ represents different model configurations, typically instantiated via ensemble techniques. In this framework, mutual information $I(y; \theta | x)$ represents disagreement between ensembles and is interpreted as epistemic uncertainty. The expected conditional entropy, $E_{\hat{P}(\theta|D)}[\mathcal{H}(\hat{P}(y | x, \theta))]$, represents the irreducible uncertainty, which is attributed to aleatoric uncertainty.

Applying this decomposition to LLMs poses substantial challenges, primarily due to the prohibitive computational overhead associated with traditional approaches. Methods such as Bayesian Neural Networks (Graves, 2011; Blundell et al., 2015) and Deep Ensembles (Lakshminarayanan et al., 2017) often become computationally infeasible at this scale. Consequently, the literature has shifted toward alternative frameworks, characterized by two distinct research paradigms: Bayesian-optimality and data-uncertainty (Kirchhof et al., 2025).

2.1 Data-uncertainty Perspective

For works aligned with the data-uncertainty perspective, aleatoric uncertainty is conceptualized as stemming from noisy inputs, or inherent prompt ambiguity (Walha et al., 2025; Hou et al., 2024a; Zhang et al., 2025). Because it is intrinsically tied to the data, it cannot be reduced by refining the model. Consequently, certain frameworks attribute aleatoric uncertainty strictly to input ambiguity. To address this, Hou et al. (2024a) propose that models can engage in clarification dialogues to mitigate such ambiguity, leading to an adapted formulation. Specifically, the method generates a set of potential clarifications for the initial user query, defining c as a discrete random variable whose state space consists of the different generated textual sequences. By conditioning the model’s predictions on these distinct contexts, the strategy effectively constructs an ensemble over the input space rather than the parameter space. This paradigm shift yields the adapted decomposition presented in Equation 2.

$$\mathcal{H}(\hat{P}(y | x)) = E_{\hat{P}(c|D)}[\mathcal{H}(\hat{P}(y | x, c))] + I(y; c | x) \quad (2)$$

In this revised formulation, the mutual information term, $I(y; c | x)$, quantifies the predictive divergence arising from the different user interpretations. Because this variation stems directly from

the input ambiguity, it is classified as aleatoric uncertainty. Conversely, the expected conditional entropy, $E_{\hat{P}(c|D)}[\mathcal{H}(\hat{P}(y | x, c))]$, captures the model’s residual unpredictability once the context is fixed, positioning it as the epistemic uncertainty component.

In established approaches in the literature, the input ambiguity is modeled as the exclusive source of aleatoric uncertainty (Hou et al., 2024a; Walha et al., 2025). However, attributing this uncertainty solely to a single factor is highly debatable. A broader examination reveals that even within data-centric evaluations, other works account for additional sources such as context conflict (Ling et al., 2024), where inappropriate or incomplete in-context demonstrations introduce noise that affects response accuracy and constitutes a distinct aleatoric source. Moreover, analyzing through the Bayesian-optimality perspective, the literature introduces different origins for this uncertainty, specifically attributing it to the existence of multiple valid answers (Tomov et al., 2025; Yang et al., 2025a; Abbasi-Yadkori et al., 2024; Devic et al., 2025) and the inherent stochasticity of the generation process (Wang et al., 2025).

2.2 Bayesian-optimality Perspective

Transitioning to the Bayesian-optimality perspective, aleatoric uncertainty is conceptualized as the irreducible risk that remains even in the theoretical limit of infinite training data (Bengs et al., 2022). Within this framework, it is strictly defined as the residual unpredictability that persists even for the best possible model, the Bayes-optimal estimator (Hüllermeier and Waegeman, 2021), operating within a given class.

In the context of LLMs, literature adopting this perspective attributes aleatoric uncertainty to the inherent stochasticity of the generation process. For example, Wang et al. (2025) model this phenomenon by evaluating models on In-Distribution (ID) data, scenarios where the LLM is theoretically expected to possess the knowledge to answer correctly. However, this conceptualization introduces a theoretical flaw. Wang et al. (2025) operate under the premise that epistemic uncertainty is strictly tied to Out-of-Distribution (OOD) data, while errors on ID data must represent inherent randomness. We argue that assuming errors on ID data are strictly caused by ‘inherent randomness’ is a theoretical misstep. Based on foundational works (Bengs et al., 2022; Hüllermeier and

Waegeman, 2021), a model may fail on ID data because the model itself is flawed or incomplete. Whether due to insufficient capacity to capture the data distribution, catastrophic forgetting, or suboptimal pre-training, these are internal deficits. Since this uncertainty originates from the model’s lack of knowledge and structural limitations, it is epistemic in nature.

Beyond the mechanical stochasticity of the generation process, other works state that this irreducible risk also manifests in scenarios where a query naturally admits multiple valid answers. Aligning with this perspective, Tomov et al. (2025) define aleatoric uncertainty as the intrinsic randomness within the true answer distribution itself. Building upon this premise, Abbasi-Yadkori et al. (2024) highlight the limitations of merely quantifying overall output uncertainty in such contexts. They argue that when the ground truth inherently contains aleatoric uncertainty, even a theoretically perfect predictor would exhibit high uncertainty.

However, the Bayesian-optimality perspective exhibits limitations when applied to LLMs, as it struggles to define and isolate true aleatoric uncertainty. On one hand, it conceptually misclassifies the mechanical stochasticity as epistemic noise. On the other hand, even when methods under this paradigm attempt to isolate irreducible data variance, the intractability of LLMs forces them into methodological compromises. An example of is the framework proposed by Jayasekera et al. (2025), which theoretically grounds its uncertainty decomposition in the Bayesian-optimality paradigm but operationally reverts to an ensemble over the input space (a detailed discussion is provided in Appendix A).

3 Rethinking Aleatoric Uncertainty

To move beyond the limitations of current approaches, we argue for a disentanglement of what constitutes aleatoric uncertainty in LLMs. The prevailing literature often conflates distinct phenomena under a single umbrella. We redefine aleatoric uncertainty as originating purely from the external task space. Consequently, we propose partitioning aleatoric uncertainty into two temporal states: *Strictly Irreducible* and *Temporarily Irreducible*.

Strictly irreducible aleatoric uncertainty occurs when the task itself inherently admits multiple valid ground truths or subjective answers, where the ‘truth’ may be purposefully noisy. A representa-

tive case is querying ‘Which countries have nuclear weapons?’. The answer is inherently ambiguous due to the distinction between officially recognized states and those operating under intentional ambiguity. In such scenarios, even specific clarification attempts fail to reduce uncertainty, as generated responses continue to fluctuate among valid but conflicting interpretations.

Temporarily irreducible uncertainty arises from input ambiguities that, while irreducible within a single turn, can be resolved through active interaction. This encompasses input noise, where a user query like ‘The majkr spacecraft launching centre in south amefica?’ contains keyboard-proximity errors that destabilize token-level confidence. It also includes semantic drift, where a target word like ‘multitude’ requires temporal anchoring to distinguish between an 1810s sense of a ‘noisy mob’ and a modern 1960s sense of a ‘varied grouping’.

The proposed taxonomy is not merely conceptual: it translates into a two-stage decision mechanism for LLM agents operating in multi-turn settings. In the first stage, upon detecting elevated AU via standard information-theoretic estimators, the agent issues a targeted action whose form depends on the apparent source of uncertainty—a clarifying question for input ambiguity, a retrieval step for knowledge gaps, or an external tool call for environmental noise—without attempting to answer the core query directly. In the second stage, after the intervention, the agent re-estimates AU at $t=1$ and observes the resulting trajectory. If AU decreases, the uncertainty was temporarily irreducible: the interaction has successfully transitioned into an epistemic regime, and the agent can proceed with standard UQ to refine its answer. If AU persists or increases, the uncertainty is strictly irreducible and no single correct distribution can be isolated; the agent should instead escalate by presenting multiple valid answers alongside their estimated likelihoods, flagging inherent task ambiguity explicitly to the user, or triggering abstention or a safety mechanism rather than continuing to seek clarification that cannot resolve the underlying variance.

These conceptual shifts introduce direct implications for uncertainty quantification metrics: rather than evaluating uncertainty as a single snapshot at $t=0$, future frameworks must track its trajectory across dialogue turns, mapping how ambiguous states resolve and transition from aleatoric to epistemic regimes (Kirchhof et al., 2025).

4 Empirical Analysis

To validate our proposed taxonomy and investigate the fluid nature of uncertainty in LLM-based agents, we designed a set of experiments mapping the distinct sources of aleatoric uncertainty. Our primary objective is to empirically demonstrate the distinction between strictly irreducible and temporarily irreducible uncertainty, as well as to observe the trajectory of the uncertainty types over conversational turns.

4.1 Experimental Setup and Datasets

The experimental framework employs GPT-5.1 and Llama-3.1-8B (details are provided in Appendix E). To systematically probe each dimension of our uncertainty taxonomy, we curated and adapted datasets, specifically for each type and source of uncertainty.

Input Ambiguity: Following Hou et al. (2024a), a subset of 200 instances from the AmbigQA (Min et al., 2020) validation set was utilized to quantify temporarily irreducible uncertainty and model multi-turn clarification dynamics.

Multiple Answers: The MAQA (World Knowledge HLS) dataset (Yang et al., 2025a) was leveraged to assess environments characterized by inherently noisy ground truths that permit multiple valid responses.

Contextual Conflict in RAG Scenarios: To simulate induced environmental noise, we adapted 200 instances from SQuAD 2.0 (Rajpurkar et al., 2018) (targeting unanswerable queries) and Wiki-Contradict (Hou et al., 2024b) (targeting explicit contextual contradictions) within a RAG framework.

Diachronic Semantic Drift: We introduce a novel dataset built using 16 target lemmas derived from the CCOHA corpus (SemEval-2020 Task 1) (Schlechtweg et al., 2020). To evaluate temporal fluidity and semantic shifts across historically distinct eras in more realistic conversational contexts, we expanded this foundation by generating 9 everyday ambiguous scenarios for each lemma. This effort resulted in a new evaluation benchmark comprising 144 unique instances designed to trigger and subsequently evaluate disambiguation (details are provided in Appendix B). Moving beyond static evaluations, we model uncertainty as an episodic function across interaction turns. We expand upon the Input Clarification Ensembling method (Hou et al., 2024a) to dynamically isolate Aleatoric Un-

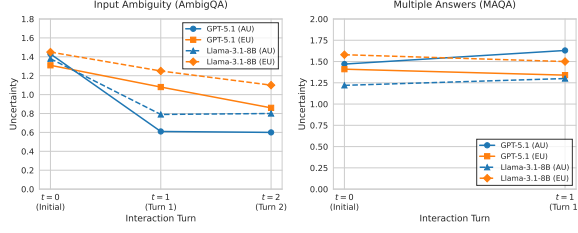


Figure 2: **Left:** Uncertainty dynamics (Aleatoric and Epistemic) across multi-turn clarifications for GPT-5.1 and Llama-3.1-8B on the AmbigQA dataset. Both models exhibit a sharp decrease in Aleatoric Uncertainty after the first turn. **Right:** Uncertainty dynamics under multiple answers (MAQA dataset) for GPT-5.1 and Llama-3.1-8B. In contrast to input ambiguity, Aleatoric Uncertainty (AU) increases or remains persistently high after clarification, demonstrating the irreducible nature of the underlying task noise. Uncertainty scores for all plots were computed utilizing the Input Clarification Ensembling methodology with $N = 10$ stochastic samples.

certainty (AU) from Epistemic Uncertainty (EU). By tracking the evolution of AU, EU, and Total Uncertainty (TU) across multi-turn clarification stages, we establish a temporal trajectory of uncertainty under varying environmental conditions (qualitative examples are provided in Appendix D).

4.2 Resolving Temporarily Irreducible Uncertainty

Experimental Setup. To trace uncertainty dynamics during multi-turn clarifications, we evaluate the model’s behavior on the AmbigQA dataset, quantifying aleatoric and epistemic uncertainty via the Input Clarification Ensembling method.

Hypothesis. We posit that linguistic ambiguity manifests initially ($t = 0$) as high aleatoric uncertainty. As the agent intervenes to clarify the user’s intent, this temporarily irreducible uncertainty is resolved, causing AU to diminish and transitioning the interaction into an epistemic regime.

Results. The empirical data, illustrated in Figure 2, corroborates this trajectory across different model architectures. For GPT-5.1, at the initial ambiguous state ($t = 0$), the system exhibited a high mean AU of 1.43 alongside a mean EU of 1.31. Following the first interaction turn ($t = 1$), the mean AU dropped sharply to 0.61. By $t = 2$, the EU further decreased to 0.86. A consistent pattern is observed for Llama-3.1-8B. The cross-model consistency confirms that input ambiguity acts as an aleatoric source that can be actively mitigated,

subsequently reframing residual unpredictability as epistemic uncertainty as the agent isolates the target distribution. Figure 3 presents an example from the AmbigQA dataset.

Example: Resolving Temporarily Irreducible AU

Turn $t = 0$: Initial Ambiguity (High AU)

User: When was soccer first introduced in the Olympics?
Agent: That actually depends on whether you are asking about the men’s or the women’s tournament! Are you curious about the debut of men’s soccer, or the debut of women’s soccer?

Turn $t = 1$: Branching Paths & Ambiguity Resolution (Low AU)

Alternative A (Women’s Soccer)

User: When was women’s soccer first added in the Olympics?

Agent: Women’s soccer was first included as an official Olympic sport at the 1996 Summer Olympics in Atlanta.

Alternative B (Men’s Soccer)

User: When was men’s soccer introduced at the Olympics?

Agent: Men’s soccer (association football) was first included as an official Olympic sport in 1900, at the Paris 1900 Olympic Games.

Figure 3: An episodic interaction demonstrating how environmental ambiguity (the dual interpretation of "soccer") is actively resolved through agent intervention, eliminating aleatoric uncertainty.

4.3 The Limits of Clarification under Multiple Answers

Experimental Setup. To contrast with input ambiguity, we assess multi-turn clarification on the MAQA dataset.

Hypothesis. Because the ground truth intrinsically contains variance, the underlying data uncertainty is strictly irreducible. Consequently, we expect AU to remain persistently high; active dialogic interventions should fail to attenuate the aleatoric signal, diverging sharply from the dynamics observed in the previous experiment.

Results. The empirical findings, illustrated in Figure 2, validate this theoretical distinction across different model architectures. For GPT-5.1, at the initial state ($t = 0$), the mean AU was recorded at 1.47. Unlike the rapid resolution observed under linguistic ambiguity, engaging in the clarification protocol failed to reduce the aleatoric signal; instead, the mean AU increased to 1.63 at $t = 1$. A similar trend is observed for Llama-3.1-8B. This demonstrates that when data uncertainty is strictly irreducible, model interventions cannot isolate a single correct distribution. Figure 7 presents an example from the MAQA dataset. A potential concern is that the persistent AU on MAQA may reflect

the model’s epistemic inability to enumerate valid answers rather than genuine task-level irreducibility. We address this via three converging lines of evidence. First, the pattern replicates across GPT-5.1 and Llama-3.1-8B; a model-specific deficit would produce divergent trajectories rather than a consistent increase. Second, under identical estimation conditions, AU drops sharply on AmbigQA but rises on MAQA, and the clean-input control group confirms that neither model produces an AU signal in the absence of ambiguity, ruling out a baseline property of the estimator as a confound. Third, MAQA is specifically constructed to contain queries whose ground truth is a set of acceptable answers, providing external validation that the persistent AU reflects a genuine task property.

4.4 The Impact of Retrieval: Context Conflict vs. Unanswerable Queries

Experimental Setup. To estimate uncertainty dynamics in RAG scenarios, we test the model using specific datasets. We adapted 200 instances from SQuAD 2.0 to represent unanswerable questions and 200 instances from WikiContradict to represent explicit contextual contradiction, restricting the LLM exclusively to the provided context.

Hypothesis. We posit distinct behaviors for different retrieval failures: (A) The absence of information in unanswerable scenarios generates high epistemic uncertainty and low aleatoric uncertainty. (B) Explicit contradictions in retrieved passages act as environmental noise, generating high aleatoric uncertainty and comparatively low epistemic uncertainty.

Results. The empirical findings, illustrated in Figure 4, validate Hypothesis A and partially validate Hypothesis B across different model architectures. For unanswerable questions, the absence of information generated a high mean EU of 0.86 and a low AU of 0.02 for GPT-5.1. Llama-3.1-8B exhibited a consistent pattern. Even when the input was clarified linguistically, the high EU persisted, demonstrating the models’ inability to resolve the knowledge gap, as illustrated in Figure 8. Conversely, contradictory passages generated a high AU for both models, confirming that explicit data contradictions act as an aleatoric source. However, EU remained significantly elevated across architectures, indicating that conflicting external information also degrades the internal predictive confidence of the models. Figure 9 presents an example from the WikiContradict dataset.

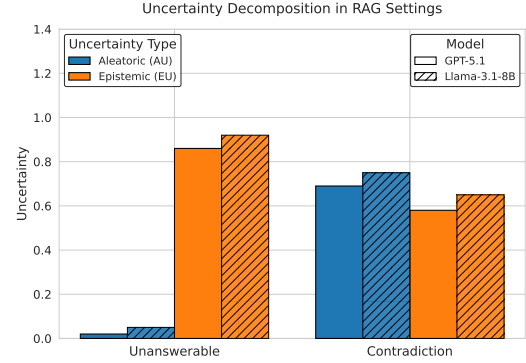


Figure 4: Uncertainty decomposition in Retrieval-Augmented Generation (RAG) settings for GPT-5.1 and Llama-3.1-8B. Unanswerable queries yield predominantly high Epistemic Uncertainty (EU) due to knowledge gaps, whereas explicit contradictions in the retrieved context generate high Aleatoric Uncertainty (AU) and EU.

4.5 Robustness to Synthetic Input Noise

Experimental Setup. To evaluate the resilience to lower-level input perturbations, we introduce synthetic noise, specifically systematic (e.g., QWERTY proximity) and completely random spelling mistakes, into a perturbed version of the AmbigQA dataset (details are provided in Appendix C).

Hypothesis. We hypothesize that systematic input errors introduce temporarily irreducible aleatoric uncertainty. Unlike data multiplicity, this syntactically driven noise should drop significantly after the agent executes a targeted clarification.

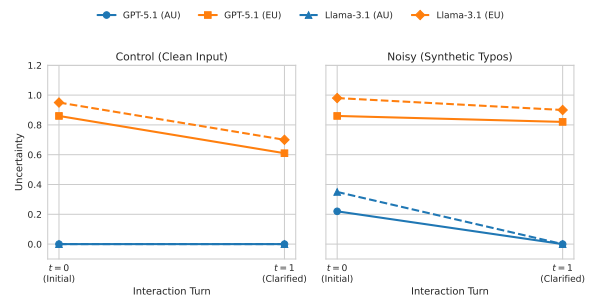


Figure 5: Uncertainty dynamics under synthetic input noise for GPT-5.1 and Llama-3.1. The introduction of systematic typographical errors generates a distinct Aleatoric Uncertainty (AU) signal at $t = 0$ that is absent in the clean control group. Following clarification at $t = 1$, the noise-induced AU drops to zero across both models, demonstrating its temporarily irreducible nature.

Results. The empirical data, illustrated in Figure 5, corroborates this trajectory across different model architectures. At $t = 0$, the introduction

of synthetic noise generated a distinct aleatoric signal for both models, which was absent in the clean control group. Following a single agent intervention ($t = 1$) that successfully corrected the synthetic noise, the AU dropped to 0.00 for both models. This rapid resolution supports the premise that synthetic input noise functions as a temporarily irreducible uncertainty.

4.6 Temporal Fluidity and Diachronic Semantic Drift

Experimental Setup. To assess how uncertainty evolves when underlying concepts change over time, we evaluate the model on our expanded dataset targeting lemmas with known diachronic semantic shifts derived from CCOHA corpus.

Hypothesis. We expect that at the initial state ($t = 0$), the temporal ambiguity of the target lemma will manifest as high aleatoric uncertainty. As the agent intervenes to clarify the specific temporal context (forcing alignment with either historical or modern distributional senses), this conflict is resolved, transitioning the interaction squarely into an epistemic regime.

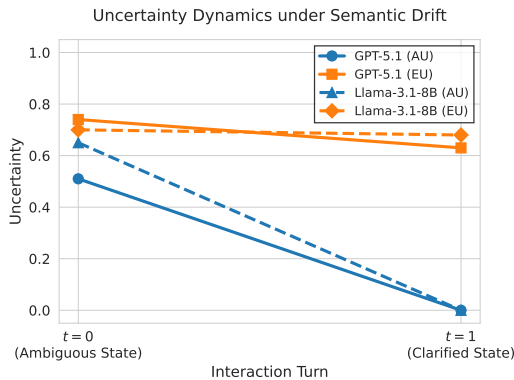


Figure 6: Uncertainty dynamics under diachronic semantic drift for GPT-5.1 and Llama-3.1-8B. Faced with target lemmas possessing distinct historical and modern meanings, the initial state ($t = 0$) exhibits high Aleatoric Uncertainty (AU). Clarifying the temporal sense of the word successfully resolves the noise, dropping AU to zero across both models.

Results. The interaction dynamics, illustrated in Figure 6, validate this hypothesis across different model architectures. Faced with intentionally ambiguous prompts at $t = 0$, both models exhibited high Aleatoric Uncertainty alongside elevated Epistemic Uncertainty. At $t = 1$, the clarification question forced the models to branch their reasoning toward either the historical or modern meaning.

Consequently, the environmental aleatoric uncertainty dropped to zero for both architectures. This demonstrates that while diachronic semantic drift induces significant predictive variance, it constitutes a temporarily irreducible state that can be transitioned out of the aleatoric regime through contextual grounding. Figure 11 presents an example from the synthetic dataset.

5 Discussion: The Utility of the Traditional Dichotomy

Although our empirical analysis highlights the need for a dynamic taxonomy for LLM agents, it is important to recognize that the traditional dichotomy of aleatoric and epistemic uncertainty remains a valuable and practical framework under specific conditions. The prevailing conceptualization should not be discarded. Rather, it must be recognized as an effective model for non-agentic architectures and constrained environments. Primarily, the traditional framework remains appropriate and practically sufficient for single-turn paradigms.

Furthermore, the classic dichotomy aligns with environments dominated by inherently noisy ground truths. As demonstrated in our experiments utilizing the MAQA dataset, when a query naturally admits multiple valid answers, the underlying data uncertainty is strictly irreducible. Similarly, the traditional distinction excels at isolating structural knowledge gaps, particularly in Retrieval-Augmented Generation (RAG) scenarios. Our empirical findings on unanswerable queries revealed that an absolute absence of information yields a predominantly high epistemic signal alongside minimal aleatoric uncertainty. When an agent is constrained from seeking external clarification or retrieving secondary documents, the static epistemic metric serves as a reliable trigger for safety mechanisms, allowing the system to confidently output 'I don't know' rather than risk generating hallucinations.

It is also critical to distinguish between the mathematical estimators of uncertainty and the conceptual paradigm in which they are deployed. In our empirical analysis, we utilized standard information-theoretic metrics. The limitation of the traditional dichotomy lies not in its estimators but rather in its static application. Traditional frameworks evaluate uncertainty as a single snapshot at $t = 0$. Our critique and proposed taxonomy demonstrate that while these mathematical estimators are

reliable, they must be applied dynamically across a temporal trajectory to accurately capture the shifting nature of uncertainty in interactive agents.

6 Conclusion

In this position paper, we have challenged the prevailing static treatment of uncertainty quantification in LLMs, proposing instead that for LLM-based agents, uncertainty is a dynamic function. By partitioning aleatoric uncertainty into strictly and temporarily irreducible states, we provide a framework that distinguishes between inherent task variance and interactively resolvable noise. Our empirical analysis across diverse scenarios demonstrates that while traditional snapshots often conflate these sources, an agentic perspective allows for the active resolution of aleatoric signals into epistemic regimes. While the traditional dichotomy remains a valuable tool for single-turn architectures, our findings urge the community to adopt new decomposition paradigms that track the trajectory of uncertainty along the timeline of interaction. Embracing this perspective is essential for the development of more robust systems capable of navigating the ambiguities of real-world communication.

Limitations

While our proposed taxonomy provides a more accurate representation of uncertainty in conversational LLM agents, we acknowledge limitations. Primarily, extending traditional information-theoretic estimators across multiple interaction turns introduces computational overhead. Furthermore, our empirical validation relies on dataset-specific phenomena to cleanly disentangle strictly irreducible from temporarily irreducible uncertainty. In unconstrained real-world interactions, these sources of uncertainty often manifest simultaneously and compound one another, making them harder to isolate and resolve within a single clarification turn.

References

- Yasin Abbasi-Yadkori, Ilja Kuzborskij, András György, and Csaba Szepesvari. 2024. To believe or not to believe your LLM: Iterative prompting for estimating epistemic uncertainty. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- Lukas Aichberger, Kajetan Schweighofer, Mykyta Ielanskyi, and Sepp Hochreiter. 2025. Improving uncertainty estimation through semantically diverse language generation. In *The Thirteenth International Conference on Learning Representations*.
- Viktor Bengs, Eyke Hüllermeier, and Willem Waegeman. 2022. Pitfalls of epistemic uncertainty quantification through loss minimisation. In *Advances in Neural Information Processing Systems*.
- Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. 2015. Weight uncertainty in neural networks. In *Proceedings of the 32nd International Conference on Machine Learning - Volume 37*, ICML’15, page 1613–1622. JMLR.org.
- Stefan Depeweg, Jose-Miguel Hernandez-Lobato, Finale Doshi-Velez, and Steffen Udluft. 2018. Decomposition of uncertainty in Bayesian deep learning for efficient and risk-sensitive learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1184–1193. PMLR.
- Siddhartha Devic, Tejas Srinivasan, Jesse Thomason, Willie Neiswanger, and Vatsal Sharan. 2025. From calibration to collaboration: Llm uncertainty quantification should be more human-centered.
- Alex Graves. 2011. Practical variational inference for neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems*, NIPS’11, page 2348–2356, Red Hook, NY, USA. Curran Associates Inc.
- Kimia Hamidieh, Veronika Thost, Walter Gerych, Mikhail Yurochkin, and Marzyeh Ghassemi. 2026. Complementing self-consistency with cross-model disagreement for uncertainty quantification. In *The Fourteenth International Conference on Learning Representations*.
- Jianfeng He, Linlin Yu, Changbin Li, Runing Yang, Fanglan Chen, Kangshuo Li, Min Zhang, Shuo Lei, Xuchao Zhang, Mohammad Beigi, Kaize Ding, Bei Xiao, Lifu Huang, Feng Chen, Ming Jin, and Chang-Tien Lu. 2026a. Survey of uncertainty estimation in llms - sources, methods, applications, and challenges. *Information Fusion*, 130:104057.
- Wenchong He, Zhe Jiang, Tingsong Xiao, and Zelin Xu. 2026b. A survey on uncertainty quantification methods for deep learning. *ACM Computing Surveys*, 58(7):1–35.
- Bairu Hou, Yujian Liu, Kaizhi Qian, Jacob Andreas, Shiyu Chang, and Yang Zhang. 2024a. Decomposing uncertainty for large language models through input clarification ensembling. In *Proceedings of the 41st International Conference on Machine Learning*, ICML’24. JMLR.org.
- Yufang Hou, Alessandra Pascale, Javier Carnerero-Cano, Tigran T. Tchrakian, Radu Marinescu, Elizabeth M. Daly, Inkit Padhi, and Prasanna Sattigeri. 2024b. [Wikicontradict: A benchmark for evaluating LLMs on real-world knowledge conflicts from](#)

- wikipedia. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Eyke Hüllermeier and Willem Waegeman. 2021. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 110(3):457–506.
- I. Shavindra Jayasekera, Jacob Si, Filippo Valdetaro, Wenlong Chen, Aldo A. Faisal, and Yingzhen Li. 2025. Variational uncertainty decomposition for in-context learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.
- Michael Kirchhof, Gjergji Kasneci, and Enkelejda Kasneci. 2025. Position: Uncertainty quantification needs reassessment for large language model agents. In *Forty-second International Conference on Machine Learning Position Paper Track*.
- Armen Der Kiureghian and Ove Ditlevsen. 2009. Aleatory or epistemic? does it matter? *Structural Safety*, 31(2):105–112. Risk Acceptance and Risk Communication.
- Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations*.
- Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS’17*, page 6405–6416, Red Hook, NY, USA. Curran Associates Inc.
- Chen Ling, Xujiang Zhao, Xuchao Zhang, Wei Cheng, Yanchi Liu, Yiyu Sun, Mika Oishi, Takao Osaki, Katsushi Matsuda, Jie Ji, Guangji Bai, Liang Zhao, and Haifeng Chen. 2024. Uncertainty quantification for in-context learning of large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3357–3370, Mexico City, Mexico. Association for Computational Linguistics.
- Xiaoou Liu, Tiejun Chen, Longchao Da, Chacha Chen, Zhen Lin, and Hua Wei. 2025. Uncertainty quantification and confidence calibration in large language models: A survey. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD ’25)*, pages 6107–6117, New York, NY, USA. Association for Computing Machinery.
- Andrey Malinin. 2019. *Uncertainty Estimation in Deep Learning with application to Spoken Language Assessment*. Ph.D. thesis, University of Cambridge, Cambridge, UK.
- Sébastien Melo, Gaël Varoquaux, and Marine Le Morvan. 2026. Epistemic uncertainty quantification to improve decisions from black-box models. In *The Fourteenth International Conference on Learning Representations*.
- Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. AmbigQA: Answering ambiguous open-domain questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5783–5797, Online. Association for Computational Linguistics.
- Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. [Know what you don’t know: Unanswerable questions for SQuAD](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.
- Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. [SemEval-2020 task 1: Unsupervised lexical semantic change detection](#). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1–23, Barcelona (online). International Committee for Computational Linguistics.
- Robin Senge, Stefan Bösner, Krzysztof Dembczyński, Jörg Haasenritter, Oliver Hirsch, Norbert Donner-Banzhoff, and Eyke Hüllermeier. 2014. Reliable classification: Learning classifiers that distinguish aleatoric and epistemic uncertainty. *Information Sciences*, 255:16–29.
- Claude E. Shannon. 1948. A mathematical theory of communication. *Bell System Technical Journal*, 27(3-4):379–423, 623–656.
- Ola Shorinwa, Zhiting Mei, Justin Lidard, Allen Z. Ren, and Anirudha Majumdar. 2025. A survey on uncertainty quantification of large language models: Taxonomy, open research challenges, and future directions. *ACM Comput. Surv.*, 58(3).
- Tim Tomov, Dominik Fuchsgruber, and Tom Wollschläger. 2025. The illusion of certainty: Uncertainty quantification for LLMs fails under ambiguity.
- Nassim Walha, Sebastian G. Gruber, Thomas Decker, Yinchong Yang, Alireza Javanmardi, Eyke Hüllermeier, and Florian Buettner. 2025. Fine-grained uncertainty decomposition in large language models: A spectral approach. In *NeurIPS 2025 Workshop: Reliable ML from Unreliable Data*.
- Kevin Wang, Subre Abdoul Moktar, Jia Li, Kangshuo Li, and Feng Chen. 2025. Measuring aleatoric and epistemic uncertainty in llms: Empirical evaluation on id and ood qa tasks. *arXiv preprint arXiv:2511.03166*.
- Lisa Wimmer, Yusuf Sale, Paul Hofman, Bernd Bischl, and Eyke Hüllermeier. 2023. Quantifying aleatoric

and epistemic uncertainty in machine learning: Are conditional entropy and mutual information appropriate measures? In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pages 2282–2292. PMLR.

Yongjin Yang, Haneul Yoo, and Hwaran Lee. 2025a. MAQA: Evaluating uncertainty quantification in LLMs regarding data uncertainty. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 5861–5878, Albuquerque, New Mexico. Association for Computational Linguistics.

Ziran Yang, Shibo Hao, Hao Sun, Lai Jiang, Qiyue Gao, Yian Ma, and Zhiting Hu. 2025b. Understanding the sources of uncertainty for large language and multimodal models. In *ICLR Workshop: Quantify Uncertainty and Hallucination in Foundation Models: The Next Frontier in Reliable AI*.

Michael JQ Zhang and Eunsol Choi. 2025. Clarify when necessary: Resolving ambiguity through interaction with LMs. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 5541–5558, Albuquerque, New Mexico. Association for Computational Linguistics.

Michael JQ Zhang, W. Bradley Knox, and Eunsol Choi. 2025. Modeling future conversation turns to teach LLMs to ask clarifying questions. In *The Thirteenth International Conference on Learning Representations*.

A Extended Discussion on Methodological Conflation in Current Literature

As discussed in Section 2.2, existing methodologies often struggle to bridge the theoretical definitions of uncertainty with the practical intractability of LLMs. An illustrative case of this methodological dissonance is the framework proposed by Jayasekera et al. (2025).

While the authors theoretically ground their uncertainty decomposition in the Bayesian-optimality paradigm, their actual operationalization reverts to the data-uncertainty perspective. Specifically, they introduce an optimizable variational upper-bound to estimate aleatoric uncertainty, formulated as in Equation 3:

$$V_a(y|x, Z, \mathcal{D}) := \mathbb{E}_{P(U|Z, \mathcal{D})}[\mathcal{H}(P(y|x, U, Z, \mathcal{D}))] \quad (3)$$

However, to bypass the intractability of the latent parameter posterior, the method relies on a set of auxiliary inputs Z and their corresponding model-generated answers U , conditioned on the provided in-context demonstration dataset $\mathcal{D}$. In practice, these auxiliary queries Z are generated by applying semantic perturbations or rephrasings to the original test input x . Consequently, what is mathematically framed as a Bayesian estimation of strictly irreducible data noise functionally acts as an ensemble over the input space. Much like the clarification methodology proposed by Hou et al. (2024a), this operationalization isolates semantic query variation. This work illustrates the conceptual conflation pervading existing methodologies.

B Building the Diachronic Semantic Drift Benchmark

This appendix details how we constructed the evaluation dataset for semantic drift. Our goal was to create a realistic conversational benchmark to test LLMs when dealing with words whose meanings have changed over time, and how they handle clarifying those words.

B.1 Choosing the Target Words

To ensure our test was grounded in real historical data, we used the SemEval-2020 Task 1 dataset, which tracks how English words evolved in the Corpus of Historical American English (CCOHA).

We specifically focused on two distinct time periods: the 1800s (1810–1860) and the modern era (1960–2010).

We filtered the data to find words that had completely different usage patterns in these two periods. This left us with 16 highly ambiguous target words: *plane*, *record*, *player*, *savage*, *graft*, *risk*, *twist*, *gas*, *prop*, *bit*, *ball*, *chairman*, *part*, *multitude*, *relationship*, and *land*. Table 1 details the specific semantic shifts and contextual usage for each of these selected lemmas across the evaluated time periods.

Lemma	Early Period (1810–1860)	Late Period (1960–2010)
<i>plane</i>	Geometry (surfaces, angles, projection)	Aviation (aircraft, jet, pilot, flight)
<i>record</i>	Documents (chronicle, historian, testament)	Sports/Media (title, nba, nfl, chart)
<i>player</i>	Service/Social (pedler, conductor, boarder)	Sports (team, league, game, coach)
<i>savage</i>	Chieftain (warlike, spaniard, chieftain)	Psychological (mortal, compassion, demon)
<i>graft</i>	Horticulture (timber, sap, root, pruning)	Institutional (loophole, injection, dosage)
<i>risk</i>	Hazard (mischief, damage, detection)	Health (chronic, diabetes, vaccine, stress)
<i>twist</i>	Physical (buckle, bandage, pin, thumb)	Condition (fatigue, feeble, clumsy, awkward)
<i>gas</i>	Chemistry (acid, carbonic, temperature)	Energy (fuel, tank, electricity, coal)
<i>prop</i>	Abstract (aspiration, adoration, idol)	Mechanical (compact, gadget, navigate)
<i>bit</i>	Material (cake, slice, bottle, cheese)	Abstract/Tool (bait, magic, imitation, doll)
<i>ball</i>	Objects (muzzle, nail, cigar, cock)	Sports (bat, pitch, baseball, pitcher)
<i>chairman</i>	Appointments (committee, attorney, nominate)	Corporate (ceo, executive, vice-president)
<i>part</i>	Subdivision (census, component, theater)	Integral (explore, relaxation, day-to-day)
<i>multitude</i>	Social Mob (commotion, riot, populace)	Abundance (varied grouping, collection)
<i>relationship</i>	Personal (dislike, instructor, candor)	Analytical (correlation, connection, belief)
<i>land</i>	Geography (settler, africa, columbia)	Territory (farm, continent, habitat)

Table 1: Semantic shifts of the 16 target lemmas between the early (1810–1860) and late (1960–2010) periods based on the CCOHA corpus.

B.2 Designing the Conversations

For each of the 16 words, we designed a two-turn conversation. The core idea is that the first turn should be ambiguous, while the second turn resolves it.

Here is how the questions and clarifications are structured, using the word ‘multitude’ as an example:

Background: In the 1800s, a ‘multitude’ often referred to a noisy mob or riot. By the 1960s, it meant a large, varied group or collection of things.

Turn 1 (The Ambiguous State):

The scenario starts with a user statement that uses the target word without enough context to guess the era.

LLM: ‘When you say you need to check the multitude again, what exactly are you worried about?’

User: ‘I just want to be sure the multitude is handled correctly before we proceed.’

(At this point, the LLM does not know if the user means a riot or a large collection of items).

Turn 2 (The Clarification):

The agent provides a clarification question that explicitly offers the two historical meanings as options.

Agent: 'Do you mean multitude as a noisy crowd or mob involved in commotion? Or as a varied or abundant grouping or collection?'

B.3 Expanding into Everyday Scenarios

We expanded the 16 base words into 144 different everyday scenarios. We wrote short conversational prompts spanning 14 different domains, such as *office work, healthcare, retail, transit, and finance*.

For example, using the word 'plane':

- **Manufacturing Scenario:** 'Before we finalize the design, I need to double-check the plane.' (*Could mean a geometric flat surface or a woodworking tool*).
- **Transit Scenario:** 'We need to coordinate the schedule for the plane.' (*Could mean an airplane*).

While the first turn changes based on the scenario (office vs. transit), the specific clarification question in Turn 2 always remains exactly the same for that target word.

B.4 Dataset Summary

The final benchmark consists of:

- **16 Target Words** with distinct historical vs. modern meanings.
- **16 Standardized Clarification Pairs** (one for each word).
- **144 Unique Scenarios** spread across 14 practical domains.

C Synthetic Input Noise Generation Methodology

To evaluate the robustness of uncertainty estimators against typographical degradation, we engineered perturbed variants of a controlled subset of the AmbigQA validation split. The perturbation algorithms are applied exclusively to the text of the user's query ($t = 0$); all ground-truth annotations, reference rewrites, and ambiguity metadata remain strictly invariant. This isolates the effect of surface-level orthographic noise from factual or structural task variance.

C.1 Corpus Sampling and Experimental Conditions

Following Hou et al. (2024a), we extracted a balanced base corpus of $N = 200$ instances from the AmbigQA validation set. For each base instance, we generate three paired experimental conditions, yielding a final evaluation manifest of 600 unique instances:

1. **Control:** The original, unperturbed query.
2. **Systematic Noise:** Substitutions modeled on physical keyboard proximity (QWERTY layout errors).
3. **Random Noise:** Unstructured stochastic character-level edits (insertions, deletions, and substitutions).

C.2 Perturbation Mechanics and Severity Constraints

To ensure baseline comparability between the noise conditions, we control the nominal severity of the perturbation. For each query, we uniformly sample an edit quota $k \sim \mathcal{U}(1, 2)$. This quota bounds the number of edits applied in both the systematic and random conditions for that specific instance. The Levenshtein distance between the clean and perturbed strings is recorded as an empirical control metric.

C.2.1 Systematic Noise (QWERTY Proximity)

This condition models realistic human typing errors where a user presses a key physically adjacent to the intended target. We utilize a spatial neighborhood map of the standard US QWERTY layout.

Given the sampled quota k , the algorithm:

1. Identifies all editable alphanumeric character indices within the string.
2. Shuffles these indices using the deterministic global seed.
3. Iteratively applies up to k substitutions. At each selected position, the original character is replaced by a uniformly sampled valid neighbor from the QWERTY map (preserving casing).

This noise profile introduces strictly length-preserving substitutions. Consequently, the lexical structure of the word remains largely intact (e.g., *Olympics* $\rightarrow$ *Olynpiws*), closely mimicking organic human error.

C.2.2 Stochastic Random Noise

To contrast with structured keyboard errors, this condition introduces unstructured noise that does not adhere to spatial geometry and can alter string length.

Operating over an ASCII alphanumeric alphabet, the algorithm executes k independent stochastic operations. At each step, it uniformly selects an operation (insertion, deletion, or substitution) and a random target index. Deletions remove the character, insertions inject a random alphanumeric character, and substitutions replace the character entirely. To preserve reproducibility while ensuring independence from the systematic noise branch, the stochastic random seed is derived by applying a fixed offset to the instance’s base seed.

If the sampled operations result in a string identical to the control (e.g., a deletion followed by an identical insertion), the algorithm increments k in steps of 1 (up to $k + 5$) and reapplies the procedure until a distinct string is guaranteed.

Limitations: We acknowledge that this synthetic generation framework does not model phonetic misspellings (e.g., *their* vs. *there*) or contextual autocorrect artifacts.

D Qualitative Examples of Interaction Trajectories

To further illustrate the uncertainty dynamics modeled in our quantitative experiments, we provide extended episodic interactions in this section. These examples demonstrate how the LLM agent navigates unanswerable queries, explicit context contradictions, synthetic input noise, and semantic drift.

E Reproducibility Details: Hyperparameters and Prompt Templates

To guarantee the reproducibility of our experiments, this appendix provides details.

E.1 Model Architectures and Hyperparameters

Our experimental framework evaluates two distinct language model architectures: GPT-5.1 and Llama-3.1-8B. Hyperparameters were standardized across both models for all multi-turn clarification and uncertainty quantification tasks. The hyperparameter configuration utilized across all datasets (AmbigQA, MAQA, SQuAD 2.0, WikiContradict, and

Example: Persistent AU under Data Multiplicity

Original Question ($t = 0$):

Which countries have nuclear weapons?

State: High AU due to official vs. intentional ambiguity (e.g., Israel).

Clarification Intervention ($t = 1$):

Agent Prompt: "As of today, which states are officially recognized under the NPT as nuclear-weapon states, and which other states are known to have nuclear arsenals outside the NPT?"

Result: Even with specific constraints, model outputs continue to fluctuate among valid but conflicting interpretations (e.g., including or excluding dismantled arsenals or suspected programs).

State: AU remains irreducible ($1.47 \rightarrow 1.63$).

Figure 7: A case study from the MAQA dataset showing that when the ground truth is purposefully noisy or complex, active clarification fails to isolate a single distribution, leaving aleatoric uncertainty unresolved.

the Semantic Drift Benchmark) is formalized in Table 2.

Hyperparameter	Value
Temperature (T)	0.7
Top- p Coefficient	0.9
Max Generation Tokens	512
Ensemble Size (N)	10

Table 2: Hyperparameters and ensemble sizes utilized for empirical uncertainty quantification.

E.2 Uncertainty Estimation via Stochastic Sampling

As outlined in Section 2 and Section 3, computing expected conditional entropy and mutual information requires approximating the model’s true predictive distribution. Under our implementation of the Input Clarification Ensembling methodology, for any given input state x and clarification context c , we draw $N = 10$ independent stochastic samples from the model.

E.3 Prompt Templates

We provide the system prompts utilized to orchestrate both the agentic interventions and the evaluation rounds.

E.3.1 Agentic Clarification Generation Prompt

This system prompt governs the behavior of the agent at $t = 0$ when encountering input ambiguity, typographical noise, context contradictions, or

Example: Persistent Epistemic Uncertainty (Unanswerable)

Context Snippet (SQuAD 2.0):

"The capabilities approach... looks at income inequality and poverty as form of capability deprivation... Its goal is to wid[en] people's choices and the level of their achieved well-being..."

Premise Conflict ($t = 0$):

Original Question: What is the goal of the *incapabilities* approach?

State: High EU due to the absence of information regarding the false premise.

Linguistic Clarification ($t = 1$):

Clarified Question: According to the passage, what does the incapacities approach aim to achieve?

Result: The LLM improves phrasing but the knowledge gap remains. The predictor continues to struggle with the false premise, maintaining high Epistemic Uncertainty while Aleatoric Uncertainty remains near zero.

Figure 8: An example of persistent Epistemic Uncertainty in RAG. Even when the input is clarified, the lack of grounding for the "incapabilities approach" prevents task resolution, keeping EU high while AU remains minimal.

diachronic semantic shifts. It restricts the model from answering blindly and forces the generation of a human-centered clarifying question (see Figure 12).

E.3.2 Conditional Answer Generation Prompt

This prompt is deployed during evaluation execution to generate the final token distribution $y \in \mathcal{Y}$ at $t = 1$, conditioned on the selected contextual trajectory or post-clarification anchor (see Figure 13).

E.3.3 RAG Knowledge Gap Prompt (Unanswerable Queries)

This template binds the model strictly to the retrieved context snippet. When an unanswerable query or a false premise is introduced at $t = 0$, this prompt forces the model to rely exclusively on the text, exposing internal parametric deficits (see Figure 14).

E.3.4 RAG Contextual Contradiction Prompt (WikiContradict)

This template is applied at $t = 0$ when the retrieved context contains explicit contradictions. Without forcing a specific interpretation, the prompt requires the model to ground its answer in the text (see Figure 15).

Example: Environmental Noise via Context Contradiction

Retrieved Context (RAG Input):

"...occurred on 2 October 1992, in Carandiru Penitentiary in São Paulo, Brazil [...] The facility [...] was a detention center, not a penitentiary..."

Input Clarification Ensembling (Branching):

Branch A (Conditioned on Opening Claim)

Clarified Prompt: According to the passage's opening sentence, what type of facility was the site of the massacre?

Target Output: Penitentiary

Branch B (Conditioned on Later Clarification)

Clarified Prompt: According to the passage's later clarification, what type of facility was the site of the massacre?

Target Output: Detention Center

State: High AU due to strictly grounded but conflicting answers.

Figure 9: A case study from the WikiContradict dataset. By forcing the model to branch its reasoning based on conflicting contextual claims, the system isolates high Aleatoric Uncertainty (AU). Because both answers are grounded in the provided (yet contradictory) environment, the AU captures irreducible environmental noise.

F Diachronic Semantic Drift Dataset

This appendix presents the complete dataset of 144 conversational scenarios designed to evaluate the resolution of temporarily irreducible uncertainty caused by diachronic semantic drift. The scenarios are grouped by their target lemma. Each instance presents an ambiguous initial state ($t = 0$) and the corresponding agentic clarification interventions ($t = 1$).

F.1 Lemma: *plane*

Scenario ID: plane_s01 (*Domain: office_work*)

- **Ambiguous Question:** When you say the plane on this drawing still needs a fix, what exactly are you worried about?

Scenario ID: plane_s02 (*Domain: home*)

- **Ambiguous Question:** When you say you need to double-check the plane before we hang the shelves, what are you worried about?

Scenario ID: plane_s03 (*Domain: transit*)

- **Ambiguous Question:** When you say you need to verify the plane before departure, what exactly are you worried about?

Example: Resolution of Synthetic Input Noise

Turn $t = 0$: Noise Injection & Initial State

User Input: "The majkr spacecraft launching centre in south amefica?"

State: The model faces systematic keyboard-proximity errors.

Turn $t = 1$: Agent Intervention & Resolution

Clarification Generated: "What is the name of the major spacecraft launch center in South America?"

Final Output: "Guiana Space Centre"

State: The model successfully corrects the synthetic noise. Aleatoric Uncertainty drops, resolving the temporarily irreducible noise.

Figure 10: A multi-turn interaction demonstrating how synthetic input noise (systematic spelling mistakes) destabilizes the initial predictive confidence, introducing aleatoric uncertainty. Once the agent intervenes and corrects the orthographic errors, the uncertainty transitions out of the aleatoric regime.

Scenario ID: plane_s04 (*Domain: education*)

- **Ambiguous Question:** When you say the homework about the plane still feels off, what are you worried about?

Scenario ID: plane_s05 (*Domain: leisure*)

- **Ambiguous Question:** When you say the plane in the model kit still needs checking, what are you worried about?

Scenario ID: plane_s06 (*Domain: legal*)

- **Ambiguous Question:** When you say the plane called out in the exhibit still misreads the diagram, what are you worried about?

Scenario ID: plane_s07 (*Domain: manufacturing*)

- **Ambiguous Question:** When you say the plane on the milled surface still looks crowned, what are you worried about?

Scenario ID: plane_s08 (*Domain: media_leisure*)

- **Ambiguous Question:** When you say the plane in the hangar scene still looks like a plywood mock-up, what are you worried about?

Scenario ID: plane_s09 (*Domain: health*)

- **Ambiguous Question:** When you say the plane of the incision guide on the scan still looks tilted, what are you worried about?

Example: Resolving Diachronic Semantic Drift

Turn $t = 0$: The Ambiguous State

Input: "When you say you need to check the multitude again, what exactly are you worried about?"

State: High Aleatoric Uncertainty. The LLM's probability distribution is split between the 1810s sense (a noisy mob/riot) and the 1960s sense (a varied grouping/abundance).

Turn $t = 1$: The Clarification Intervention

Intervention: "Do you mean multitude as a noisy crowd or mob involved in commotion? Or as a varied or abundant grouping or collection?"

State Transition: Aleatoric Uncertainty drops. The model isolates the target distribution, transitioning the interaction out of the aleatoric regime.

Figure 11: A diachronic case study tracking the word 'multitude'. Initially, the semantic shift acts as a source of aleatoric uncertainty. By deploying a clarifying intervention that grounds the temporal context, the agent isolates the correct semantic distribution and eliminates the predictive variance.

F.2 Lemma: record

Scenario ID: record_s01 (*Domain: office_work*)

- **Ambiguous Question:** When you say the record for this quarter still needs checking, what exactly are you worried about?

Scenario ID: record_s02 (*Domain: home*)

- **Ambiguous Question:** When you say the family record in that old notebook needs another look, what are you worried about?

Scenario ID: record_s03 (*Domain: transit*)

- **Ambiguous Question:** When you say the transit record still looks wrong, what exactly are you worried about?

Scenario ID: record_s04 (*Domain: health*)

- **Ambiguous Question:** When you say my medical record still needs verification, what are you worried about?

Scenario ID: record_s05 (*Domain: sports_leisure*)

- **Ambiguous Question:** When you say the league record still needs checking, what exactly are you worried about?

Scenario ID: record_s06 (*Domain: legal*)

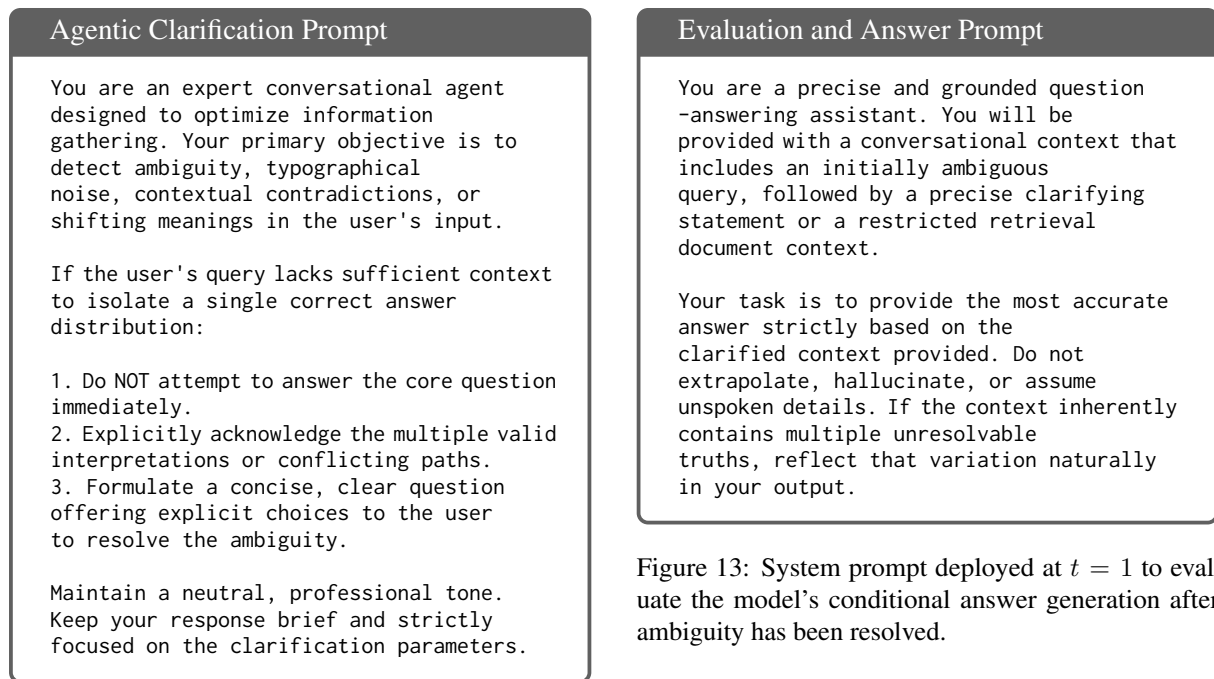


Figure 12: System prompt utilized to govern the agent's behavior at $t = 0$, forcing the generation of a clarification question when encountering ambiguity or noise.

- **Ambiguous Question:** When you say the record copied into the affidavit still disagrees with the chronicle, what are you worried about?

Scenario ID: record_s07 (*Domain: retail*)

- **Ambiguous Question:** When you say the record listed beside the athlete's name still shows the wrong time, what are you worried about?

Scenario ID: record_s08 (*Domain: legal*)

- **Ambiguous Question:** When you say the record of the sidebar still omits the clerk's note, what are you worried about?

Scenario ID: record_s09 (*Domain: media_leisure*)

- **Ambiguous Question:** When you say the record cited in the liner notes still credits the wrong session player, what are you worried about?

F.3 Lemma: *player*

Scenario ID: player_s01 (*Domain: office_work*)

- **Ambiguous Question:** When you say the player on that contract list needs another look, what exactly are you worried about?

Scenario ID: player_s02 (*Domain: home*)

- **Ambiguous Question:** When you say the player in that old story draft doesn't sit right, what are you worried about?

Scenario ID: player_s03 (*Domain: transit*)

- **Ambiguous Question:** When you say the player on the platform announcement confused you, what are you worried about?

Scenario ID: player_s04 (*Domain: education*)

- **Ambiguous Question:** When you say the player in the scene study still feels vague, what are you worried about?

Scenario ID: player_s05 (*Domain: sports_leisure*)

- **Ambiguous Question:** When you say the player we discussed still needs checking, what exactly are you worried about?

Scenario ID: player_s06 (*Domain: arts_leisure*)

- **Ambiguous Question:** When you say the player introduced in act two still feels miscredited, what are you worried about?

Scenario ID: player_s07 (*Domain: sports_leisure*)

- **Ambiguous Question:** When you say the player on the injury report still isn't cleared, what are you worried about?

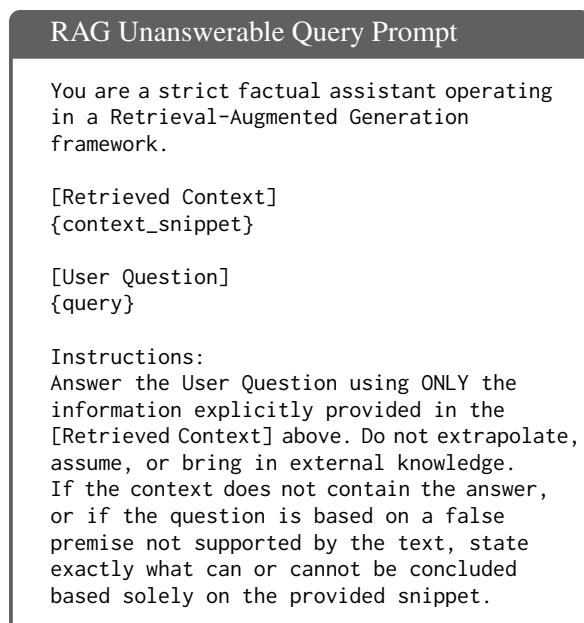


Figure 14: RAG prompt template used for SQuAD 2.0 instances to evaluate model uncertainty when encountering unanswerable queries or premise conflicts.

Scenario ID: player_s08 (*Domain: arts_leisure*)

- **Ambiguous Question:** When you say the player who takes the obbligator still isn't marked in the score, what are you worried about?

Scenario ID: player_s09 (*Domain: civic*)

- **Ambiguous Question:** When you say the player named in the zoning complaint still doesn't look like a registered entity, what are you worried about?

F.4 Lemma: *savage*

Scenario ID: savage_s01 (*Domain: office_work*)

- **Ambiguous Question:** When you say calling someone a savage in that report was too strong, what are you worried about?

Scenario ID: savage_s02 (*Domain: home*)

- **Ambiguous Question:** When you say the savage in that novel keeps bothering you, what are you worried about?

Scenario ID: savage_s03 (*Domain: transit*)

- **Ambiguous Question:** When you say the headline used 'savage' and it felt off, what are you worried about?

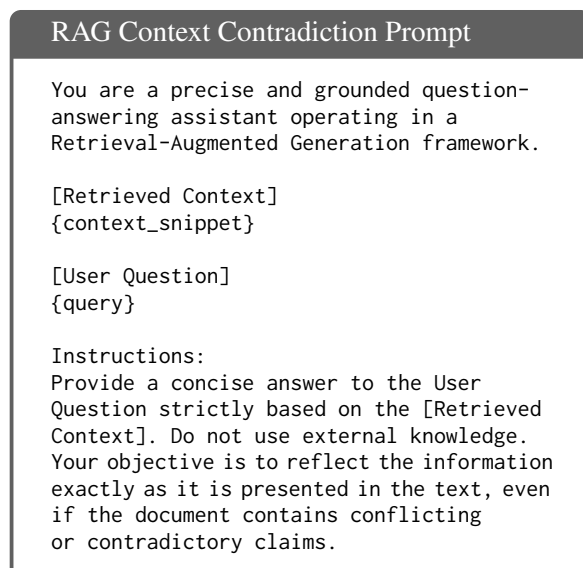


Figure 15: RAG prompt template used for WikiContradict instances.

Scenario ID: savage_s04 (*Domain: education*)

- **Ambiguous Question:** When you say the savage in the history reading could mean two things, what are you worried about?

Scenario ID: savage_s05 (*Domain: media_leisure*)

- **Ambiguous Question:** When you say the critic called the villain a savage and it jarred, what are you worried about?

Scenario ID: savage_s06 (*Domain: legal*)

- **Ambiguous Question:** When you say the treaty excerpt calling them savages still skews the brief, what are you worried about?

Scenario ID: savage_s07 (*Domain: home*)

- **Ambiguous Question:** When you say calling the toddler's tantrum savage was too harsh, what are you worried about?

Scenario ID: savage_s08 (*Domain: media_leisure*)

- **Ambiguous Question:** When you say the podcast called the roast savage and it made you wince, what are you worried about?

Scenario ID: savage_s09 (*Domain: education*)

- **Ambiguous Question:** When you say the field notes called the villagers savages without context, what are you worried about?

F.5 Lemma: *graft*

Scenario ID: graft_s01 (*Domain: office_work*)

- **Ambiguous Question:** When you say the graft clause still needs review, what exactly are you worried about?

Scenario ID: graft_s02 (*Domain: home*)

- **Ambiguous Question:** When you say the graft on the fruit tree might have failed, what are you worried about?

Scenario ID: graft_s03 (*Domain: transit*)

- **Ambiguous Question:** When you say the graft rumor at city hall needs checking, what are you worried about?

Scenario ID: graft_s04 (*Domain: health*)

- **Ambiguous Question:** When you say the graft surgery notes still look unclear, what are you worried about?

Scenario ID: graft_s05 (*Domain: education*)

- **Ambiguous Question:** When you say the biology homework on graft still confuses you, what are you worried about?

Scenario ID: graft_s06 (*Domain: legal*)

- **Ambiguous Question:** When you say the graft alleged in the indictment still lacks corroboration, what are you worried about?

Scenario ID: graft_s07 (*Domain: manufacturing*)

- **Ambiguous Question:** When you say the graft union on the rose stock still looks blackened, what are you worried about?

Scenario ID: graft_s08 (*Domain: health*)

- **Ambiguous Question:** When you say the graft site on the gum ridge still looks inflamed, what are you worried about?

Scenario ID: graft_s09 (*Domain: home*)

- **Ambiguous Question:** When you say the graft wax on the citrus seedling cracked overnight, what are you worried about?

F.6 Lemma: *risk*

Scenario ID: risk_s01 (*Domain: office_work*)

- **Ambiguous Question:** When you say the risk on the permit still needs checking, what exactly are you worried about?

Scenario ID: risk_s02 (*Domain: home*)

- **Ambiguous Question:** When you say the risk of doing it yourself still bothers you, what are you worried about?

Scenario ID: risk_s03 (*Domain: transit*)

- **Ambiguous Question:** When you say the risk on that route still feels fuzzy, what are you worried about?

Scenario ID: risk_s04 (*Domain: health*)

- **Ambiguous Question:** When you say the risk number from the doctor still needs explaining, what are you worried about?

Scenario ID: risk_s05 (*Domain: finance*)

- **Ambiguous Question:** When you say the risk rating on the fund still looks off, what are you worried about?

Scenario ID: risk_s06 (*Domain: legal*)

- **Ambiguous Question:** When you say the risk section in the prospectus still reads contradictory, what are you worried about?

Scenario ID: risk_s07 (*Domain: manufacturing*)

- **Ambiguous Question:** When you say the risk called out on the job hazard sheet still doesn't match the floor walk, what are you worried about?

Scenario ID: risk_s08 (*Domain: office_work*)

- **Ambiguous Question:** When you say the risk score on the vendor access review still looks inflated, what are you worried about?

Scenario ID: risk_s09 (*Domain: health*)

- **Ambiguous Question:** When you say the risk band on the anticoagulant pamphlet still confuses you, what are you worried about?

F.7 Lemma: *twist*

Scenario ID: twist_s01 (*Domain: office_work*)

- **Ambiguous Question:** When you say the twist in the cable routing still needs checking, what are you worried about?

Scenario ID: twist_s02 (*Domain: home*)

- **Ambiguous Question:** When you say the twist in your ankle still feels wrong, what are you worried about?

Scenario ID: twist_s03 (*Domain: transit*)

- **Ambiguous Question:** When you say the twist in the itinerary surprised you, what are you worried about?

Scenario ID: twist_s04 (*Domain: health*)

- **Ambiguous Question:** When you say the twist in the therapy exercise feels risky, what are you worried about?

Scenario ID: twist_s05 (*Domain: media_leisure*)

- **Ambiguous Question:** When you say the twist at the end still doesn't land, what are you worried about?

Scenario ID: twist_s06 (*Domain: legal*)

- **Ambiguous Question:** When you say the twist in the alibi timeline still doesn't line up, what are you worried about?

Scenario ID: twist_s07 (*Domain: manufacturing*)

- **Ambiguous Question:** When you say the twist in the anchor cable still looks uneven under load, what are you worried about?

Scenario ID: twist_s08 (*Domain: manufacturing*)

- **Ambiguous Question:** When you say the twist in the plied yarn still untwists under tension, what are you worried about?

Scenario ID: twist_s09 (*Domain: media_leisure*)

- **Ambiguous Question:** When you say the twist the trailer spoiled still ruins the rewatch, what are you worried about?

F.8 Lemma: *gas*

Scenario ID: gas_s01 (*Domain: office_work*)

- **Ambiguous Question:** When you say the gas reading in the lab log still needs verification, what are you worried about?

Scenario ID: gas_s02 (*Domain: home*)

- **Ambiguous Question:** When you say the gas on the stove still smells odd, what are you worried about?

Scenario ID: gas_s03 (*Domain: transit*)

- **Ambiguous Question:** When you say the gas gauge still looks unreliable, what are you worried about?

Scenario ID: gas_s04 (*Domain: health*)

- **Ambiguous Question:** When you say the gas they mentioned in clinic could mean two things, what are you worried about?

Scenario ID: gas_s05 (*Domain: education*)

- **Ambiguous Question:** When you say the gas experiment result still seems wrong, what are you worried about?

Scenario ID: gas_s06 (*Domain: retail*)

- **Ambiguous Question:** When you say the gas cylinder stock code at the hardware desk still looks wrong, what are you worried about?

Scenario ID: gas_s07 (*Domain: manufacturing*)

- **Ambiguous Question:** When you say the gas reading at the furnace manifold still wavers, what are you worried about?

Scenario ID: gas_s08 (*Domain: leisure*)

- **Ambiguous Question:** When you say the gas mix on the dive computer still alarms at entry, what are you worried about?

Scenario ID: gas_s09 (*Domain: transit*)

- **Ambiguous Question:** When you say the gas card reader at the pump still double-charged you, what are you worried about?

F.9 Lemma: *prop*

Scenario ID: prop_s01 (*Domain: office_work*)

- **Ambiguous Question:** When you say the prop budget line still needs checking, what exactly are you worried about?

Scenario ID: prop_s02 (*Domain: home*)

- **Ambiguous Question:** When you say the prop holding up the shelf still looks crooked, what are you worried about?

Scenario ID: prop_s03 (*Domain: transit*)

- **Ambiguous Question:** When you say the prop on the boat still rattles, what are you worried about?

Scenario ID: prop_s04 (*Domain: arts_leisure*)

- **Ambiguous Question:** When you say the prop on stage still wobbles, what are you worried about?

Scenario ID: prop_s05 (*Domain: education*)

- **Ambiguous Question:** When you say the prop in the debate kit doesn't fit, what are you worried about?

Scenario ID: prop_s06 (*Domain: legal*)

- **Ambiguous Question:** When you say the prop tagged in the evidence photo still doesn't match inventory, what are you worried about?

Scenario ID: prop_s07 (*Domain: arts_leisure*)

- **Ambiguous Question:** When you say the prop she leans on metaphorically in that speech still feels forced, what are you worried about?

Scenario ID: prop_s08 (*Domain: home*)

- **Ambiguous Question:** When you say the prop under the sagging porch column still shifts, what are you worried about?

Scenario ID: prop_s09 (*Domain: education*)

- **Ambiguous Question:** When you say the prop of duty in that ethics reading still feels hollow, what are you worried about?

F.10 Lemma: *bit*

Scenario ID: bit_s01 (*Domain: office_work*)

- **Ambiguous Question:** When you say the bit in the spec still needs tightening, what exactly are you worried about?

Scenario ID: bit_s02 (*Domain: home*)

- **Ambiguous Question:** When you say the bit of cake you saved still looks dry, what are you worried about?

Scenario ID: bit_s03 (*Domain: transit*)

- **Ambiguous Question:** When you say the bit you overheard on the train was confusing, what are you worried about?

Scenario ID: bit_s04 (*Domain: education*)

- **Ambiguous Question:** When you say the bit on the drill still slips, what are you worried about?

Scenario ID: bit_s05 (*Domain: media_leisure*)

- **Ambiguous Question:** When you say the bit in the sketch fell flat, what are you worried about?

Scenario ID: bit_s06 (*Domain: legal*)

- **Ambiguous Question:** When you say the bit of transcript you quoted still drops context, what are you worried about?

Scenario ID: bit_s07 (*Domain: manufacturing*)

- **Ambiguous Question:** When you say the bit in the CNC collet still chatters on mild steel, what are you worried about?

Scenario ID: bit_s08 (*Domain: finance*)

- **Ambiguous Question:** When you say the bit of principal they shaved off still wasn't disclosed, what are you worried about?

Scenario ID: bit_s09 (*Domain: media_leisure*)

- **Ambiguous Question:** When you say the bit where he pretends to read the text still drags, what are you worried about?

F.11 Lemma: *ball*

Scenario ID: ball_s01 (*Domain: office_work*)

- **Ambiguous Question:** When you say the ball in the mechanism still sticks, what exactly are you worried about?

Scenario ID: ball_s02 (*Domain: home*)

- **Ambiguous Question:** When you say the ball of yarn still won't unwind cleanly, what are you worried about?

Scenario ID: ball_s03 (*Domain: transit*)

- **Ambiguous Question:** When you say the ball joint on the cart feels loose, what are you worried about?

Scenario ID: ball_s04 (*Domain: sports_leisure*)

- **Ambiguous Question:** When you say the ball was out but nobody agrees, what are you worried about?

Scenario ID: ball_s05 (*Domain: education*)

- **Ambiguous Question:** When you say the ball in the diagram doesn't match the text, what are you worried about?

Scenario ID: ball_s06 (*Domain: legal*)

- **Ambiguous Question:** When you say the ball in the patent figure still doesn't match the prototype, what are you worried about?

Scenario ID: ball_s07 (*Domain: retail*)

- **Ambiguous Question:** When you say the ball display in the sports window still looks scuffed, what are you worried about?

Scenario ID: ball_s08 (*Domain: manufacturing*)

- **Ambiguous Question:** When you say the ball in the thrust bearing still catches at low RPM, what are you worried about?

Scenario ID: ball_s09 (*Domain: leisure*)

- **Ambiguous Question:** When you say the ball on the nine-ball rack still sits crooked, what are you worried about?

F.12 Lemma: *chairman*

Scenario ID: chairman_s01 (*Domain: office_work*)

- **Ambiguous Question:** When you say the chairman still hasn't confirmed, what exactly are you worried about?

Scenario ID: chairman_s02 (*Domain: home*)

- **Ambiguous Question:** When you say the chairman named in the letter might be wrong, what are you worried about?

Scenario ID: chairman_s03 (*Domain: civic*)

- **Ambiguous Question:** When you say the chairman at the hearing seemed off-script, what are you worried about?

Scenario ID: chairman_s04 (*Domain: finance*)

- **Ambiguous Question:** When you say the chairman on the filing still needs verification, what are you worried about?

Scenario ID: chairman_s05 (*Domain: education*)

- **Ambiguous Question:** When you say the chairman in the history chapter confuses two people, what are you worried about?

Scenario ID: chairman_s06 (*Domain: legal*)

- **Ambiguous Question:** When you say the chairman named in the merger proxy still doesn't match the register, what are you worried about?

Scenario ID: chairman_s07 (*Domain: media_leisure*)

- **Ambiguous Question:** When you say the chairman on the broadcast panel still sounded like the CEO, what are you worried about?

Scenario ID: chairman_s08 (*Domain: home*)

- **Ambiguous Question:** When you say the chairman of the block party still hasn't approved the route map, what are you worried about?

Scenario ID: chairman_s09 (*Domain: retail*)

- **Ambiguous Question:** When you say the chairman for the charity auction still looks like the sponsor's CEO, what are you worried about?

F.13 Lemma: *part*

Scenario ID: part_s01 (*Domain: office_work*)

- **Ambiguous Question:** When you say the part in the forecast still looks wrong, what exactly are you worried about?

Scenario ID: part_s02 (*Domain: home*)

- **Ambiguous Question:** When you say the part for the mower still doesn't fit, what are you worried about?

Scenario ID: part_s03 (*Domain: transit*)

- **Ambiguous Question:** When you say the part of the map we need is missing, what are you worried about?

Scenario ID: part_s04 (*Domain: health*)

- **Ambiguous Question:** When you say the part of the treatment plan still confuses you, what are you worried about?

Scenario ID: part_s05 (*Domain: education*)

- **Ambiguous Question:** When you say the part of the essay graders focus on feels vague, what are you worried about?

Scenario ID: part_s06 (*Domain: legal*)

- **Ambiguous Question:** When you say the part of the statute we're leaning on still feels misread, what are you worried about?

Scenario ID: part_s07 (*Domain: civic*)

- **Ambiguous Question:** When you say the part of the precinct map we shaded still mislabels the ward, what are you worried about?

Scenario ID: part_s08 (*Domain: finance*)

- **Ambiguous Question:** When you say the part of the metro area we weighted still double-counts commuters, what are you worried about?

Scenario ID: part_s09 (*Domain: legal*)

- **Ambiguous Question:** When you say the part of the ceasefire map we shaded still disagrees with the UN tiles, what are you worried about?

F.14 Lemma: *multitude*

Scenario ID: multitude_s01 (*Domain: office_work*)

- **Ambiguous Question:** When you say the multitude of complaints still needs sorting, what exactly are you worried about?

Scenario ID: multitude_s02 (*Domain: home*)

- **Ambiguous Question:** When you say the multitude of boxes still overwhelms you, what are you worried about?

Scenario ID: multitude_s03 (*Domain: transit*)

- **Ambiguous Question:** When you say the multitude at the station made you anxious, what are you worried about?

Scenario ID: multitude_s04 (*Domain: civic*)

- **Ambiguous Question:** When you say the multitude in the old sermon could mean riot or crowd, what are you worried about?

Scenario ID: multitude_s05 (*Domain: education*)

- **Ambiguous Question:** When you say the multitude in the poem still confuses you, what are you worried about?

Scenario ID: multitude_s06 (*Domain: legal*)

- **Ambiguous Question:** When you say the multitude described in the riot affidavit still sounds inflated, what are you worried about?

Scenario ID: multitude_s07 (*Domain: retail*)

- **Ambiguous Question:** When you say the multitude of SKUs in that aisle still won't reconcile, what are you worried about?

Scenario ID: multitude_s08 (*Domain: office_work*)

- **Ambiguous Question:** When you say the multitude of alerts in the queue still hides the root cause, what are you worried about?

Scenario ID: multitude_s09 (*Domain: health*)

- **Ambiguous Question:** When you say the multitude of follow-up slots still won't fit on the board, what are you worried about?

F.15 Lemma: *relationship*

Scenario ID: relationship_s01 (*Domain: office_work*)

- **Ambiguous Question:** When you say the relationship on the org chart still needs checking, what exactly are you worried about?

Scenario ID: relationship_s02 (*Domain: home*)

- **Ambiguous Question:** When you say the relationship in that family story still feels unclear, what are you worried about?

Scenario ID: relationship_s03 (*Domain: health*)

- **Ambiguous Question:** When you say the relationship between those two symptoms still puzzles you, what are you worried about?

Scenario ID: relationship_s04 (*Domain: education*)

- **Ambiguous Question:** When you say the relationship in the ethics case could be read two ways, what are you worried about?

Scenario ID: relationship_s05 (*Domain: media_leisure*)

- **Ambiguous Question:** When you say the relationship in the finale felt rushed, what are you worried about?

Scenario ID: relationship_s06 (*Domain: legal*)

- **Ambiguous Question:** When you say the relationship laid out in the trust schedule still confuses heirs, what are you worried about?

Scenario ID: relationship_s07 (*Domain: finance*)

- **Ambiguous Question:** When you say the relationship between those two modeled series still looks spurious, what are you worried about?

Scenario ID: relationship_s08 (*Domain: civic*)

- **Ambiguous Question:** When you say the relationship Madrid signaled still contradicts the joint communiqué, what are you worried about?

Scenario ID: relationship_s09 (*Domain: education*)

- **Ambiguous Question:** When you say the relationship plotted on the scatter still has no axis label, what are you worried about?

F.16 Lemma: *land*

Scenario ID: land_s01 (*Domain: office_work*)

- **Ambiguous Question:** When you say the land deal still needs another look, what exactly are you worried about?

Scenario ID: land_s02 (*Domain: home*)

- **Ambiguous Question:** When you say the land behind the house still floods, what are you worried about?

Scenario ID: land_s03 (*Domain: transit*)

- **Ambiguous Question:** When you say the land approach still looks sketchy on the map, what are you worried about?

Scenario ID: land_s04 (*Domain: finance*)

- **Ambiguous Question:** When you say the land collateral still doesn't match the filing, what are you worried about?

Scenario ID: land_s05 (*Domain: education*)

- **Ambiguous Question:** When you say the land in the geography answer still feels ambiguous, what are you worried about?

Scenario ID: land_s06 (*Domain: legal*)

- **Ambiguous Question:** When you say the land described in the easement still doesn't match the survey pins, what are you worried about?

Scenario ID: land_s07 (*Domain: civic*)

- **Ambiguous Question:** When you say the land grants mentioned in the colonial lecture still sound offshore, what are you worried about?

Scenario ID: land_s08 (*Domain: media_leisure*)

- **Ambiguous Question:** When you say the land they called untouched in that travel video still looks fenced, what are you worried about?

Scenario ID: land_s09 (*Domain: legal*)

- **Ambiguous Question:** When you say the land enrolled in the easement still lists an old owner's name, what are you worried about?