Dialogue as Uncertainty Reduction: A Judge-Free Metric for Multi-Turn Dialogue Evaluation

Anonymous ACL submission

Abstract

Evaluating multi-turn dialogue systems remains challenging, as dialogue quality depends on how effectively an agent accumulates relevant information across turns. In this work, we propose a fast, information-theoretic metric for evaluating multi-turn dialogue based on uncertainty reduction over the course of a conversation in embedding space. Our approach admits a tractable Gaussian approximation and enjoys desirable theoretical properties, including monotonicity, telescoping over turns, and submodularity. Unlike recent approaches that rely on large language models as judges, our method is fully reference-free (no ground-truth answers, no gold references, no human annotations at evaluation time), deterministic, and computationally efficient. We show that the proposed metric remains effective even when instantiated with extremely lightweight embedding models under CPU-only execution, indicating that the evaluative signal does not require large model capacity or autoregressive inference. We evaluate the proposed metric on MT-Bench and Chatbot Arena, showing competitive and, on MT-Bench, improved agreement with human preferences compared to several LLM-as-a-judge baselines.

1 Introduction

Large Language Models are increasingly deployed as conversational agents for information-seeking tasks including question answering, and decision support (Li et al., 2023; Ye et al., 2023; Ma et al., 2025; Kamalloo et al., 2023). In these settings, interactions naturally unfold over multiple turns, with users refining their questions and agents progressively providing clarifications and evidence (Wu et al., 2023). Evaluating the quality of such multi-turn interactions is therefore central to understanding and improving modern dialogue systems.

Unlike single-turn tasks, the quality of a dialogue cannot be determined by inspecting individual responses in isolation (Deshpande et al., 2025; Kim et al., 2022). Prior work suggests that effective dialogues tend to exhibit meaningful progress across turns, while minimizing redundant or irrelevant information (Kim et al., 2022; Finch et al., 2023). As a result, dialogue evaluation must reason about the evolution of information over time rather than static response quality.

In many real-world conversational applications, particularly in task-oriented and information-seeking settings, effective information transfer is a central objective, alongside other aspects of dialogue quality such as fluency and style (Deriu et al., 2020; Beaver, 2022; Guan et al., 2025). In information-seeking dialogues and related interactive settings such as decision support, tutoring, and troubleshooting, user goals are often underspecified or evolve over time, making single-turn responses insufficient. As a result, effective systems must engage in multi-turn interactions to progressively elicit, refine, and convey relevant information (Deshpande et al., 2025; Piskala et al., 2025). In these settings, high-quality dialogues are those that progressively reduce uncertainty about the task at hand, avoiding unnecessary repetition or digressions while introducing new, relevant evidence. Measuring uncertainty reduction therefore captures a fundamental aspect of dialogue usefulness that is orthogonal to surface-level fluency or conversational style.

This perspective is particularly relevant in practical deployment scenarios where evaluation must be fast, reproducible, and scalable. For example, large-scale model comparison, regression testing during model development, and online monitoring of deployed systems all require lightweight evaluation signals that reflect meaningful dialogue progress (Deshpande et al., 2025; Li et al., 2025a; Guan et al., 2025). An information-theoretic metric provides a principled way to quantify this dimension directly, without relying on costly human annotation or heavyweight judge models.

We make the following contributions:

- **Information-theoretic formulation:** We formalize multi-turn dialogue evaluation as measuring uncertainty reduction over the course of a conversation and introduce an information-theoretic metric that captures dialogue-level progress independent of fluency or stylistic considerations.

- **Practical, training-free approximation:** We propose a fast, reference-free, and deterministic Gaussian approximation to information gain in embedding space, and establish theoretical properties including monotonicity, telescoping across turns, and diminishing returns for redundant information.

- **Empirical Validation:** We demonstrate that the proposed metric distinguishes dialogues of differing quality in controlled synthetic settings and achieves competitive agreement with human preferences on MT-Bench and Chatbot Arena (Zheng et al., 2023b), while being substantially more efficient than LLM-as-a-judge baselines (Li et al., 2025a; Zhang et al., 2024).

2 Related Work

Prior work on dialogue evaluation spans several complementary directions.

Task-Oriented and Structured Dialogue Evaluation. In structured and task-oriented settings, dialogue quality is often evaluated using a combination of turn-level and dialogue-level signals, sometimes augmented with learned or judge-based components. For example, TD-EVAL proposes a two-stage framework that integrates turn-level metrics with dialogue-level aggregation and LLM judgments, demonstrating improved alignment with human preferences on benchmarks such as MultiWOZ and Tau-Bench (Budzianowski et al., 2018; Yao et al., 2024; Acikgoz et al., 2025).These approaches focus on task-specific success criteria and typically rely on supervised models or external judges, whereas our metric abstracts multi-turn dialogue progress as uncertainty reduction in embedding space.

Dialogue-Level Coherence and Consistency Metrics. Another line of work proposes dialogue-level metrics based on coherence, consistency, or contextual appropriateness across turns (Dey et al., 2022; Ghazarian et al., 2022). Such methods assess whether a dialogue remains internally consistent or locally appropriate given its history, but they do not explicitly model dialogue-level progress or diminishing returns arising from redundant information.

Large Language Models as Dialogue Evaluators. A substantial body of work studies large language models themselves as dialogue evaluators, analyzing their agreement with human judgments as well as their sensitivity to prompting and evaluation protocols (Chen et al., 2023).While LLM-as-a-judge approaches have been shown to achieve strong alignment with human preferences, they incur significant computational and operational costs in practical evaluation pipelines (Jia et al., 2024). For instance, MT-Bench 101 demonstrates that with careful prompting and evaluation design, LLM-based judges such as GPT-4 can achieve very high agreement with human judgments on multi-turn dialogue quality (Bai et al., 2024). At the same time, these results rely on large models, extensive prompt engineering, and repeated inference, making such approaches expensive and difficult to deploy for large-scale model comparison, regression testing, or continuous monitoring.

Learned Evaluators and Judge Approximation. Closely related are learned evaluators that aim to approximate human judgments or LLM-based judges directly, including dialogue-level predictors, pairwise comparison models, and feature-based frameworks (Ou et al., 2024; Park et al., 2024; Zhou et al., 2024; Li et al., 2025b). Although effective, these methods require supervised training and inherit the opacity and distributional assumptions of the judgments they are trained to approximate.

Positioning of Our Work. Recent large-scale surveys and meta-evaluations conclude that LLM-as-a-judge approaches currently offer the strongest overall alignment with human dialogue evaluations across a wide range of datasets and quality dimensions, albeit at high computational cost and with limitations related to prompt sensitivity, reproducibility, and scalability (Li et al., 2025a; Zhang et al., 2024). Accordingly, we treat LLM judges as strong holistic baselines rather than targets to replace. Instead, our work proposes a complementary, training-free metric that captures multi-turn epistemic progress via uncertainty reduction, with explicit theoretical guarantees and substantially

lower computational overhead.

3 Our Method

3.1 Scope of Evaluation

The quality of a conversational agent is inherently multi-dimensional, encompassing factors such as correctness, helpfulness, safety, style, and creativity. No single scalar metric can capture all of these aspects simultaneously. In this work, we deliberately isolate a fundamental dimension: the ability of an agent to reduce relevant uncertainty over the course of a multi-turn dialogue. We focus on information gain as a measure of informativeness and dialogue progress, independent of fluency or stylistic considerations.

3.2 Preliminaries

Let Σ be a finite alphabet and Σ^* the set of all finite strings over Σ. A T-turn question–answer dialogue is a sequence

$$D_{1:T} \triangleq \big((q_1, a_1), \ldots, (q_T, a_T)\big) \tag{1}$$

where $(q_t, a_t) \in \Sigma^* \times \Sigma^*$ denotes question and answer in step t. Optionally, one may assume an abstract universe of atomic facts $\mathcal{U}$ and an extraction map $\mathcal{E} : \Sigma^* \times \Sigma^* \to 2^{\mathcal{U}}$, but our approximation in Section 3.4 does not require an explicit $\mathcal{U}$.

Problem Statement. Our objective is not to model pairwise human preference directly. Instead, we define a dialogue-level scoring function $\mathrm{IG} : \mathcal{D} \to \mathbb{R}$ that assigns a scalar score to a single multi-turn dialogue $D_{1:T} \in \mathcal{D}$. Pairwise preference datasets are used only for evaluation: given two dialogues $D^A_{1:T}$ and $D^B_{1:T}$, we compare $\mathrm{IG}(D^A_{1:T})$ and $\mathrm{IG}(D^B_{1:T})$ and measure agreement with human majority judgments. Our metric assigns a scalar score to a single dialogue in isolation. Pairwise preference benchmarks are used only to assess how well this score correlates with human judgments, rather than being the modeling objective itself.

3.3 An Information-Theoretic Idealization

World-based view (ideal). Let Φ be a background theory (e.g., Horn clauses) and let $\mathcal{M}(F)$ be the set of possible worlds consistent with evidence F and Φ. With a uniform posterior over $\mathcal{M}(F)$, the entropy equals $\log |\mathcal{M}(F)|$ and the per-turn information gain is

$$\mathrm{IG}_t \triangleq \log \frac{|\mathcal{M}(F_{t-1})|}{|\mathcal{M}(F_t)|} \geq 0, \tag{2}$$

which telescopes over turns. This motivates our goal: compute a fast surrogate that preserves monotonicity and telescoping without explicit model counting.

3.4 Gaussian Approximation to Information Gain

The idealized formulation in Section 3.3 assumes access to an explicit universe of facts and model counting over possible worlds. Our approximation is universe-free in the sense that it bypasses any explicit symbolic universe or fact inventory, instead operating directly in continuous embedding space.

Evidence units. For each turn t, we deterministically extract a multiset of evidence strings from the answer[1],

$$\mathrm{Z}_t \triangleq \{\mathrm{Z}_{t,1}, \ldots, \mathrm{Z}_{t,m_t}\} \subseteq \Sigma^*, \tag{3}$$

e.g., sentences or semantic chunks of a_t. We write $\mathrm{enc} : \Sigma^* \to \mathbb{R}^d$ for a fixed text embedding function, and denote

$$\mathbf{q}_t \triangleq \mathrm{enc}(q_t) \in \mathbb{R}^d \quad \mathbf{z}_{t,i} \triangleq \mathrm{enc}(\mathrm{Z}_{t,i}) \in \mathbb{R}^d. \tag{4}$$

Latent semantic state. We model an unobserved latent vector $\mathbf{X} \in \mathbb{R}^d$ representing residual semantic uncertainty induced by the dialogue so far.[2] We place an isotropic Gaussian prior

$$\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_0), \qquad \mathbf{\Sigma}_0 \triangleq \sigma_0^2 \mathbf{I}. \tag{5}$$

Observation model. Each evidence embedding induces a scalar linear measurement of the latent semantic state:

$$y_{t,i} = \mathbf{z}_{t,i}^\top \mathbf{X} + \boldsymbol{\varepsilon}_{t,i}, \quad \varepsilon_{t,i} \sim \mathcal{N}(0, \sigma^2), \tag{6}$$

Since information gain depends only on the posterior covariance, the specific measurement values $y_{t,i}$ do not need to be observed or estimated.

Question-conditioned relevance weights. We downweight evidence unrelated to the question using cosine similarity in embedding space. Let $\hat{\mathbf{v}} \triangleq \mathbf{v}/\|\mathbf{v}\|_2$ denote ℓ_2 normalization. We define

$$w_{t,i} \triangleq \max\big(0, \langle \hat{\mathbf{q}}_t, \hat{\mathbf{z}}_{t,i} \rangle\big), \tag{7}$$

optionally with a hard cutoff $w_{t,i} \leftarrow 0$ if $w_{t,i} < \tau$ for some $\tau \in [0, 1]$.

[1] For our experiments, we simply just split them into sentences.

[2] $\mathbf{X}$ is an abstract modeling device rather than a literal world state or fact representation.

Precision-form update. Let $\mathbf{J}_t \triangleq \mathbf{\Sigma}_t^{-1}$ denote the precision matrix of the posterior distribution over the latent semantic uncertainty variable $\mathbf{X}$ after incorporating evidence up to turn t. For nonnegative evidence weights $w_{t,i} \geq 0$, the posterior precision updates additively:

$$\mathbf{J}_t = \mathbf{J}_{t-1} + \sum_{i=1}^{m_t} \frac{w_{t,i}}{\sigma^2} \mathbf{z}_{t,i} \mathbf{z}_{t,i}^\top. \tag{8}$$

This precision update corresponds to the information-form posterior covariance of Bayesian linear regression with linear measurements in embedding space. See Appendix A for a derivation. From now on, denote $\alpha_{t,i} = \frac{w_{t,i}}{\sigma^2}$.

Per-turn information gain. The differential entropy of a d-dimensional Gaussian $\mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma})$ is

$$\mathrm{H}(\mathbf{X}) = \frac{d}{2} \log(2\pi e) + \frac{1}{2} \log\det(\mathbf{\Sigma}). \tag{9}$$

Since the first term is constant in t, the per-turn information gain induced by turn t is

$$\mathrm{IG}_t \triangleq \mathrm{H}(\mathbf{X} \mid \mathrm{Z}_{1:t-1}) - \mathrm{H}(\mathbf{X} \mid \mathrm{Z}_{1:t}) \tag{10}$$

$$= \frac{1}{2} \log \frac{\det(\mathbf{\Sigma}_{t-1})}{\det(\mathbf{\Sigma}_t)}. \tag{11}$$

Theorem 3.1 (Monotonicity and Telescoping)**.** *Assume $\mathbf{\Sigma}_0 \succ \mathbf{0}$ and weights $\alpha_{t,i} \geq 0$. Under the precision update in Equation (8), we have $\mathbf{\Sigma}_t \preceq \mathbf{\Sigma}_{t-1}$ (in PSD order) and hence $\mathrm{IG}_t \geq 0$ for all t. Moreover, the total gain telescopes:*

$$\sum_{t=1}^{T} \mathrm{IG}_t = \frac{1}{2} \log \frac{\det(\mathbf{\Sigma}_0)}{\det(\mathbf{\Sigma}_T)}, \tag{12}$$

so it depends only on the initial and final posterior covariances.

Proof. Proved in Appendix B □

One important consequence of our formulation is that accumulating additional evidence yields diminishing returns: as evidence grows, posterior uncertainty shrinks, and new information can only reduce uncertainty by a smaller amount than it would have earlier. This property prevents verbosity and redundancy from being rewarded. We formalize this intuition using the notion of *submodularity*.

Definition (Submodularity) A set function $F : 2^{\mathcal{X}} \to \mathbb{R}$ is *submodular* if for any $\mathbf{x} \in \mathcal{X}$ and any $A \subseteq B \subseteq \mathcal{X}$,

$$F(A \cup \{\mathbf{x}\}) - F(A) \;\geq\; F(B \cup \{\mathbf{x}\}) - F(B). \tag{13}$$

Theorem 3.2 (Submodularity of Gaussian Information Gain)**.** *Let $\mathcal{X}$ be a finite set of evidence items. Each $\mathbf{x}_i \in \mathcal{X}$ is represented by an embedding $\mathbf{z}_i \in \mathbb{R}^d$ and a nonnegative weight $\alpha_i \geq 0$. Define the precision matrix induced by a subset $\mathcal{S} \subseteq \mathcal{X}$ as*

$$\mathbf{J}(\mathcal{S}) \;=\; \mathbf{J}_0 \;+\; \sum_{i \in \mathcal{S}} \alpha_i \, \mathbf{z}_i \mathbf{z}_i^\top, \tag{14}$$

where $\mathbf{J}_0 \succ 0$ is a fixed prior precision matrix. Define the set function

$$F(\mathcal{S}) \triangleq \log\det \mathbf{J}(\mathcal{S}). \tag{15}$$

Then F is a monotone submodular function.

Proof. Proved in Appendix C □

Since total information gain in our method is proportional to $\log\det \mathbf{J}(\mathcal{S})$ up to additive constants, this result implies that additional evidence exhibits diminishing returns. Consequently, longer or more verbose dialogues are not rewarded unless they contribute genuinely novel information.

3.5 Behavior Under Redundancy, Irrelevance, and Recovery

To illustrate how the proposed metric differentiates conversations of varying quality, we construct a controlled synthetic dialogue consisting of three phases: an initial informative phase, a middle phase dominated by redundant or irrelevant turns, and a final phase where novel information is reintroduced. All dialogues share the same length and structure, differing only in the informational content of their turns.

As shown in Figure 1, per-turn information gain decreases during the redundant and irrelevant phase, reflecting diminishing returns when new evidence does not substantially reduce uncertainty. When novel information is reintroduced, the marginal gain increases again. While the Gaussian approximation does not exactly match an oracle uncertainty model, it consistently preserves the dialogue-level ordering: conversations that introduce novel information for a larger fraction of turns achieve higher cumulative information gain. This

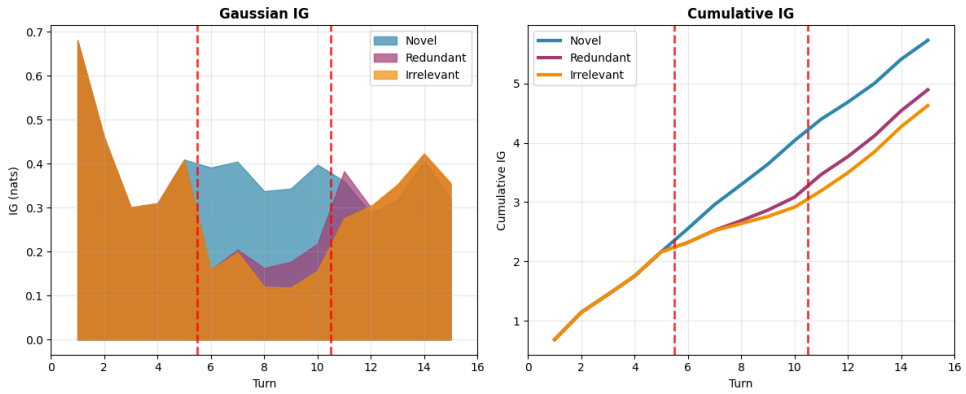

Figure 1: Information gain distinguishes dialogue quality in a 15-turn synthetic setting. We compare dialogues that introduce novel, redundant, or irrelevant information. (Left) Per-turn Gaussian information gain: redundant and irrelevant turns yield lower marginal gain than novel turns. (Right) Cumulative information gain: dialogues with a higher proportion of novel information achieve consistently higher total gain despite approximation noise.

demonstrates that the metric captures meaningful differences in long-horizon dialogue progress without relying on learned judges or supervision. We discuss more details and provide more examples in Appendix E.

We can also characterize these behaviors mathematically. Recall the precision update in Equation (8) and the per-turn information gain in Equation (10). For a single evidence embedding $\mathbf{z}$ with weight $\alpha \geq 0$, define the marginal information gain

$$\Delta(\mathbf{z}; \mathbf{J}) \triangleq \log\det(\mathbf{J} + \alpha \mathbf{z}\mathbf{z}^\top) - \log\det(\mathbf{J}). \tag{16}$$

By the matrix determinant lemma, this admits the closed form

$$\Delta(\mathbf{z}; \mathbf{J}) = \log\left(1 + \alpha\, \mathbf{z}^\top \mathbf{J}^{-1} \mathbf{z}\right). \tag{17}$$

In particular, if $\alpha = 0$ (corresponding to completely irrelevant evidence), the information gain is exactly zero.

Lemma 3.3 (Soft irrelevance yields a small upper bound)**.** *Assume embeddings are norm bounded, $\|\mathbf{z}\|_2 \leq B$, and that the evidence weight satisfies $\alpha \leq \varepsilon$. Then the information gain contributed by this evidence is bounded as*

$$\Delta(\mathbf{z}; \mathbf{J}) \;\leq\; \log\left(1 + \alpha\, \lambda_{\max}(\mathbf{J}_0^{-1}) B^2\right) = \mathcal{O}(\varepsilon). \tag{18}$$

Proof. Proved in Appendix D.1 □

Lemma 3.4 (Redundancy yields diminishing returns)**.** *Consider repeatedly adding the same evidence embedding $\mathbf{z}$ with fixed weight $\alpha > 0$. Let $\mathbf{J}_k = \mathbf{J}_0 + k\alpha\mathbf{z}\mathbf{z}^\top$. Then the marginal information gain is non-increasing:*

$$\Delta(\mathbf{z}; \mathbf{J}_k) \;\geq\; \Delta(\mathbf{z}; \mathbf{J}_{k+1}) \qquad \textit{for all } k \geq 0. \tag{19}$$

Proof. Proved in Appendix D.2. □

These results formally explain why irrelevant turns contribute little or no gain,and why repeated or redundant information exhibits diminishing returns, as illustrated in our synthetic experiments.

3.6 Length control via redundant filler

We consider an idealized append-only setting to analyze whether Gaussian information gain can be inflated by verbosity alone. Starting from a fixed dialogue history, suppose that additional turns introduce no new task-relevant information and contribute identical evidence embeddings. Formally, each filler insertion contributes the same embedding vector $\mathbf{f}$ with a fixed nonnegative weight, corresponding to repeated rank-one precision updates with $\mathbf{z} = \mathbf{f}$.

By Theorem 3.4, the marginal information gain from adding $\mathbf{f}$ is non-increasing as the precision matrix accumulates. Consequently, repeated redundant updates yield diminishing returns, and the cumulative information gain increases at most logarithmically before saturating. This guarantees that verbosity without introducing new evidence cannot substantially increase the total information gain, preventing inflation of the metric by response length alone.

Method	MT-Bench			Chatbot Arena		
	Acc. ↑	τ ↑	Time ↓	Acc. ↑	τ ↑	Time ↓
Mistral Large 3 (Mistral AI, 2025)	76.50	0.54	132	60.19	0.20	119
DeepSeek R1 (DeepSeek-AI et al., 2025)	80.96	0.61	633	66.14	0.32	480
GPT OSS 120b (OpenAI, 2025)	72.27	0.44	568	65.27	0.31	328
Claude Sonnet 3.7 (Anthropic, 2025a)	76.47	0.53	260	64.74	0.29	251
Claude Sonnet 4 (Anthropic, 2025b)	81.93	0.64	220	65.58	0.31	175
Claude Sonnet 4.5 (Anthropic, 2025c)	76.05	0.52	342	66.42	0.33	286
Ours (Gaussian IG)	84.03	0.68	86	65.80	0.32	44

Table 1: Agreement and runtime comparison on MT-Bench and Chatbot Arena. Time denotes end-to-end wall-clock seconds for $N = 100$ dialogue pairs (mean over $R = 5$ runs). LLM-as-a-judge methods are executed via hosted inference APIs.

4 Experiments

4.1 Setup

We evaluate alignment with human preferences on MT-Bench and Chatbot Arena (Zheng et al., 2023a). Each benchmark provides paired dialogues $(D^A_{1:T}, D^B_{1:T})$ with human preference votes; we follow prior work and evaluate only clear-majority (non-tie) cases. For each dialogue we compute a total score $\mathrm{IG}(D_{1:T})$ and predict the preferred dialogue by comparing $\mathrm{IG}(D^A_{1:T})$ and $\mathrm{IG}(D^B_{1:T})$. We report agreement accuracy with the majority label and Kendall's τ over pairwise rankings. Unless otherwise stated, we use `Qwen3-Embedding-0.6B`; Section 4.3 studies other embedding backends and sizes. We compare only against LLM-as-a-judge baselines, which are the strongest automatic evaluators in recent surveys (Li et al., 2025a; Zhang et al., 2024).

Runtime Evaluation. We measure end-to-end wall-clock time using identical pipelines, fixed example ordering, N=100 dialogue pairs per run, and R=5 runs (mean $\pm$ std). Our method performs local embedding passes and closed-form updates, while judge baselines call hosted inference APIs. Thus, runtimes reflect both computation and deployment overhead (e.g., network/service latency) and are intended as an operational comparison rather than a hardware-normalized benchmark.

4.2 Results

Agreement with Human Judges. Table 1 summarizes agreement with human preferences on MT-Bench and Chatbot Arena. On MT-Bench, our Gaussian information gain metric achieves higher accuracy and Kendall's τ than all LLM-as-a-judge baselines considered. Importantly, we obtain agreement levels comparable to those reported by (Zheng et al., 2023a) under the MT-Bench evaluation protocol. On Chatbot Arena, where preference labels are substantially noisier and majority votes are often absent, absolute agreement scores are lower for all methods. Nevertheless, our approach remains competitive with LLM-based judges while being significantly more computationally efficient.

Importantly, our method does not aim to replicate holistic human judgments, but rather to capture a specific dimension of dialogue quality: epistemic progress through uncertainty reduction. That the proposed metric matches or exceeds LLM judges on MT-Bench suggests that this dimension is a strong signal of human preference in information-seeking dialogues.

On Chatbot Arena, smaller performance gaps should be interpreted with caution due to higher label ambiguity; here, our results indicate that uncertainty reduction alone remains informative, though not sufficient to fully explain human preferences.

Runtime. Table 1 reports end-to-end wall-clock time. Our method reduces evaluation to embedding passes plus closed-form Gaussian updates, whereas LLM judges require hosted autoregressive inference. The similar relative speedups on MT-Bench and Chatbot Arena suggest that judge runtimes are driven not only by dialogue length, but also by substantial per-request overhead (e.g., network latency, service queueing, and provider-side execution). In contrast, our runtime is dominated by local embedding throughput and scales predictably with the amount of text to embed. We further show in Section 4.3 that this efficiency persists under strict CPU-only execution with lightweight embedding models.

Embedding Model	Device	MT-Bench			Chatbot Arena		
		Acc. ↑	τ ↑	Time ↓	Acc. ↑	τ ↑	Time ↓
Minishlab-Potion-Base-2M (min, 2024)	CPU	80.25	0.61	7.1 s	66.05	0.32	0.4 s
all-MiniLM-L6-v2-23M (Wang et al., 2020)	CPU	83.61	0.67	17.9 s	66.27	0.33	6.77 s
Snowflake-Arctic-Embed-32M (Merrick et al., 2024)	CPU	85.29	0.70	29.3 s	65.73	0.31	13.4 s
modernbert-embed-base-0.1B (Nussbaum et al., 2024)	GPU	83.19	0.66	28.5 s	65.75	0.31	15.5 s
MicroLlama-text-embedding-0.3B (Wang, 2024)	GPU	84.45	0.69	47.4 s	66.17	0.32	23.7 s
Qwen3-Embedding-0.6B (Default) (Zhang et al., 2025)	GPU	84.03	0.68	86.4 s	65.80	0.32	44.1 s
Qwen3-Embedding-4B (Zhang et al., 2025)	GPU	84.83	0.69	396.2 s	66.13	0.32	246.3 s

Table 2: Embedding model ablation across benchmarks. "Device" indicates whether embeddings were computed on CPU-only or with hardware acceleration (GPU/MPS). Runtimes are end-to-end wall-clock time on $N = 100$ dialogue pairs (mean over $R = 5$ runs).

4.3 Ablation

Embedding Model. We study the sensitivity of the proposed Gaussian information gain metric to the choice of text embedding model. Table 2 reports results across embedding architectures spanning over three orders of magnitude in parameter count, from a ∼2M-parameter model to a 4B-parameter model.

Across both MT-Bench and Chatbot Arena, performance is remarkably stable: all embeddings achieve comparable agreement with human preferences, with differences in accuracy typically within 1–3 points. Notably, even a ∼2M-parameter embedding model run entirely on CPU attains over 80% accuracy on MT-Bench, despite being orders of magnitude smaller than commonly used embedding or judge models.

Larger embedding models, such as Qwen3-Embedding-4B, achieve slightly higher accuracy but at substantially increased computational cost, while smaller models offer dramatic speedups with only modest degradation in agreement. These results suggest that the evaluative signal captured by uncertainty reduction saturates at relatively small embedding capacities, and does not require large models or autoregressive inference.

Cosine-Similarity Only. To isolate the contribution of the information-gain mechanism from the underlying embedding representation, we conduct a mechanism ablation using cosine similarity alone. Specifically, we score dialogue pairs by summing per-turn cosine similarities between model responses and the dialogue context, removing any uncertainty accumulation or diminishing-returns effects. On clear-majority MT-Bench pairs, this cosine-based baseline achieves only 44.5% accuracy, performing worse than both a majority baseline (51.7%) and random guessing (50.1% ± 3.2), and far below the proposed Gaussian information gain metric (84.0%). This result shows that naive similarity aggregation is not merely weak but actively misleading in this setting, likely rewarding verbosity or redundancy rather than genuine information gain.

5 Discussions

Information acquisition and human judgment. Our approach begins from a normative view of information-seeking dialogue as uncertainty reduction. While humans do not explicitly compute information-theoretic quantities, a natural question is whether human judgments implicitly favor dialogues that resolve ambiguity and narrow plausible interpretations. Across benchmarks, we find that cumulative information gain correlates with human preferences, potentially suggesting that uncertainty reduction captures a factor that is subconsciously weighted in human evaluations of dialogue quality.

Why larger embedding models might not help. From Equation (17), the per-turn gain depends only on the scalar $\mathbf{z}^\top \mathbf{J}_{t-1}^{-1}\mathbf{z}$, measuring alignment with directions of remaining uncertainty. Diagonalizing $\mathbf{\Sigma}_{t-1} = \mathbf{J}_{t-1}^{-1} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$ and writing $\mathbf{z} = \mathbf{U}\mathbf{c}$ yields

$$\mathrm{IG}(\mathbf{z}) = \tfrac{1}{2}\log\left(1 + \alpha \sum_i \lambda_i c_i^2\right), \qquad (20)$$

where λ_i are the eigenvalues of $\mathbf{\Sigma}_{t-1}$. Crucially, information gain depends on the *effective rank* of remaining uncertainty rather than the raw dimensionality or parameter count of the embedding model. Once embeddings capture the dominant semantic axes relevant to the task, larger models primarily refine representations within directions where λ_i

is already small, yielding negligible additional volume reduction. As a result, increased model capacity does not necessarily translate into higher information gain.

Spectral interpretation of information gain. The log-determinant in Equation (10) admits a simple spectral interpretation. For any positive definite covariance matrix Σ, we have

$$\log\det(\Sigma) = \mathrm{tr}(\log\Sigma), \tag{21}$$

where $\log\Sigma$ denotes the matrix logarithm. Writing the eigendecomposition $\Sigma = U\Lambda U^\top$ with eigenvalues $\{\lambda_i\}_{i=1}^d$, this reduces to

$$\log\det(\Sigma) = \sum_{i=1}^{d} \log\lambda_i. \tag{22}$$

From this perspective, Gaussian information gain corresponds to the total logarithmic contraction of uncertainty along the principal semantic directions encoded by the eigenvectors of Σ. Each dialogue turn reduces uncertainty primarily in directions aligned with the evidence embeddings, shrinking a subset of eigenvalues while leaving others largely unchanged. As dialogue progresses, high-uncertainty directions are resolved first, so additional evidence aligned with already-constrained directions yields diminishing marginal gain. Conversely, evidence aligned with directions of remaining uncertainty produces larger reductions in $\sum_i \log\lambda_i$. This view highlights that the metric captures not just the accumulation of evidence, but the progressive concentration and contraction of uncertainty volume in embedding space.

Embedding bias. Like all representation-based methods, our approach inherits biases present in the underlying embedding model. These biases reflect properties of the training data and may affect the geometry of the embedding space. However, unlike LLM-based judges, such biases are static and reproducible, and their impact can be analyzed directly through the induced uncertainty geometry. Moreover, because information gain emphasizes reductions in uncertainty volume rather than absolute similarity, systematic biases that do not introduce new semantic directions tend to saturate quickly, limiting their influence on dialogue-level scores.

Applicability The proposed metric targets epistemic progress in multi-turn dialogue and is most appropriate for information-seeking, decision support, tutoring, and troubleshooting settings, where dialogue quality is closely tied to how efficiently uncertainty is reduced over time. In these domains, progress naturally corresponds to introducing new, task-relevant information while avoiding redundancy or irrelevance.

The method is not intended as a holistic evaluation of dialogue quality in settings where ambiguity, stylistic variation, or open-ended exploration are desirable. For example, in creative writing, dialogues may intentionally preserve or introduce ambiguity, which would be penalized by our metric.

Accordingly, we view information gain as complementary to holistic evaluators such as LLM-as-a-judge approaches. While learned judges can aggregate many aspects of quality, they are costly and opaque. Our metric isolates a single, well-defined dimension that is inexpensive, deterministic, and easy to apply, making it useful for large-scale model comparison, regression testing, and deployment monitoring.

6 Conclusion

We proposed an information-theoretic approach to multi-turn dialogue evaluation that measures dialogue quality through uncertainty reduction over time. Starting from an idealized notion of information gain, we derived a practical, reference-free approximation with closed-form updates and theoretical guarantees, including diminishing returns for redundant information.

Across MT-Bench and Chatbot Arena, the resulting metric achieves competitive performance—and on MT-Bench, improved agreement—with human preferences compared to LLM-as-a-judge baselines, while being substantially faster and cheaper to compute. These results suggest that epistemic progress captures an important dimension of dialogue quality in information-seeking settings.

Rather than replacing holistic evaluators, our metric isolates a single, interpretable, and scalable signal of dialogue progress. Its determinism and efficiency make it well suited for large-scale model comparison and deployment monitoring, and suggest that information-theoretic objectives can serve as effective complements to learned or judge-based evaluation methods.

Limitations

Our metric is designed to measure a single aspect of dialogue quality: how much uncertainty is reduced over the course of a multi-turn interaction. This makes it well suited for information-seeking and task-oriented settings, but it does not account for other important factors such as fluency, coherence, safety, factual accuracy, or stylistic quality. As a result, the score should not be interpreted as a complete assessment of dialogue quality, especially in open-ended or creative settings where ambiguity or exploration may be preferred.

The method relies on fixed text embeddings to represent dialogue content and therefore reflects the strengths and limitations of the chosen embedding model. While our experiments suggest that results are fairly stable across different embeddings, representational biases in the embedding space may still influence the measured information gain.

Our formulation models uncertainty reduction in a continuous embedding space rather than over explicit facts or symbolic world states. This abstraction enables fast and deterministic computation, but it may miss logical dependencies, contradictions, or fine-grained factual errors that are important to human evaluators.

Finally, our evaluation is based on existing human preference benchmarks such as MT-Bench and Chatbot Arena, which contain annotation noise and ambiguous cases. In particular, many Chatbot Arena comparisons lack a clear majority preference, making small differences in agreement difficult to interpret.

References

2024. Model2vec: Turn any sentence transformer into a small fast model.

Emre Can Acikgoz, Carl Guo, Suvodip Dey, Akul Datta, Takyoung Kim, Gokhan Tur, and Dilek Hakkani-Tur. 2025. TD-EVAL: Revisiting task-oriented dialogue evaluation by combining turn-level precision with dialogue-level comparisons. In *Proceedings of the 26th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 113–132, Avignon, France. Association for Computational Linguistics.

Anthropic. 2025a. Claude 3.7 sonnet and claude code. https://www.anthropic.com/news/claude-3-7-sonnet.

Anthropic. 2025b. Introducing claude 4. https://www.anthropic.com/news/claude-4.

Anthropic. 2025c. Introducing claude sonnet 4.5. https://www.anthropic.com/news/claude-sonnet-4-5.

Ge Bai, Jie Liu, Xingyuan Bu, Yancheng He, Jiaheng Liu, Zhanhui Zhou, Zhuoran Lin, Wenbo Su, Tiezheng Ge, Bo Zheng, and Wanli Ouyang. 2024. Mt-bench-101: A fine-grained benchmark for evaluating large language models in multi-turn dialogues. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 7421–7454. Association for Computational Linguistics.

Ian Beaver. 2022. The success of conversational ai and the ai evaluation challenge it reveals. *AI Magazine*, 43(1):139–141.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium. Association for Computational Linguistics.

Bao Chen, Yuanjie Wang, Zeming Liu, and Yuhang Guo. 2023. Automatic evaluate dialogue appropriateness by using dialogue act. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7361–7372, Singapore. Association for Computational Linguistics.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Jan Deriu, Alvaro Rodrigo, Arantxa Otegi, Guillermo Echegoyen, Sophie Rosset, Eneko Agirre, and Mark Cieliebak. 2020. Survey on evaluation methods for dialogue systems. *Artificial Intelligence Review*, 54(1):755–810.

Kaustubh Deshpande, Ved Sirdeshmukh, Johannes Baptist Mols, Lifeng Jin, Ed-Yeremai Hernandez-Cardona, Dean Lee, Jeremy Kritz, Willow E. Primack, Summer Yue, and Chen Xing. 2025. MultiChallenge: A realistic multi-turn conversation evaluation benchmark challenging to frontier LLMs. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 18632–18702, Vienna, Austria. Association for Computational Linguistics.

Suvodip Dey, Ramamohan Kummara, and Maunendra Desarkar. 2022. Towards fair evaluation of dialogue state tracking by flexible incorporation of turn-level performances. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 318–324,

Dublin, Ireland. Association for Computational Linguistics.

Sarah E. Finch, James D. Finch, and Jinho D. Choi. 2023. Don't forget your ABC's: Evaluating the state-of-the-art in chat-oriented dialogue systems. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15044–15071, Toronto, Canada. Association for Computational Linguistics.

Sarik Ghazarian, Behnam Hedayatnia, Alexandros Papangelis, Yang Liu, and Dilek Hakkani-Tur. 2022. What is wrong with you?: Leveraging user sentiment for automatic dialog evaluation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 4194–4204, Dublin, Ireland. Association for Computational Linguistics.

Shengyue Guan, Haoyi Xiong, Jindong Wang, Jiang Bian, Bin Zhu, and Jian guang Lou. 2025. Evaluating llm-based agents for multi-turn conversations: A survey. *Preprint*, arXiv:2503.22458.

Jinghan Jia, Abi Komma, Timothy Leffel, Xujun Peng, Ajay Nagesh, Tamer Soliman, Aram Galstyan, and Anoop Kumar. 2024. Leveraging LLMs for dialogue quality measurement. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, pages 359–367, Mexico City, Mexico. Association for Computational Linguistics.

Ehsan Kamalloo, Nouha Dziri, Charles Clarke, and Davood Rafiei. 2023. Evaluating open-domain question answering in the era of large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5591–5606, Toronto, Canada. Association for Computational Linguistics.

Takyoung Kim, Hoonsang Yoon, Yukyung Lee, Pilsung Kang, and Misuk Kim. 2022. Mismatch between multi-turn dialogue and its evaluation metric in dialogue state tracking. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 297–309, Dublin, Ireland. Association for Computational Linguistics.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. 2025a. From generation to judgment: Opportunities and challenges of LLM-as-a-judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2757–2791, Suzhou, China. Association for Computational Linguistics.

Qintong Li, Leyang Cui, Lingpeng Kong, and Wei Bi. 2025b. Exploring the reliability of large language models as customized evaluators for diverse NLP tasks. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10325–10344, Abu Dhabi, UAE. Association for Computational Linguistics.

Siheng Li, Cheng Yang, Yichun Yin, Xinyu Zhu, Zesen Cheng, Lifeng Shang, Xin Jiang, Qun Liu, and Yujiu Yang. 2023. AutoConv: Automatically generating information-seeking conversations with large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1751–1762, Toronto, Canada. Association for Computational Linguistics.

Chuangtao Ma, Yongrui Chen, Tianxing Wu, Arijit Khan, and Haofen Wang. 2025. Large language models meet knowledge graphs for question answering: Synthesis and opportunities. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24589–24608, Suzhou, China. Association for Computational Linguistics.

Luke Merrick, Danmei Xu, Gaurav Nuti, and Daniel Campos. 2024. Arctic-embed: Scalable, efficient, and accurate text embedding models. *Preprint*, arXiv:2405.05374.

Mistral AI. 2025. Introducing mistral 3. `https://mistral.ai/news/mistral-3`.

Zach Nussbaum, John X. Morris, Brandon Duderstadt, and Andriy Mulyar. 2024. Nomic embed: Training a reproducible long context text embedder. *Preprint*, arXiv:2402.01613.

OpenAI. 2025. Introducing gpt-oss. `https://openai.com/index/introducing-gpt-oss/`.

Jiao Ou, Junda Lu, Che Liu, Yihong Tang, Fuzheng Zhang, Di Zhang, and Kun Gai. 2024. DialogBench: Evaluating LLMs as human-like dialogue systems. 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 6137–6170, Mexico City, Mexico. Association for Computational Linguistics.

ChaeHun Park, Minseok Choi, Dohyun Lee, and Jaegul Choo. 2024. Paireval: Open-domain dialogue evaluation metric with pairwise comparisons. In *First Conference on Language Modeling*.

Deepak Babu Piskala, Sharlene Chen, Udita Patel, Parul Kalra, and Rafael Castrillo. 2025. Mind the goal: Data-efficient goal-oriented evaluation of conversational agents and chatbots using teacher models. *Preprint*, arXiv:2510.03696.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Preprint*, arXiv:2002.10957.

Zixiao Ken Wang. 2024. Microllama: A 300m-parameter language model trained from scratch. https://github.com/keeeeenw/MicroLlama, https://huggingface.co/keeeeenw/MicroLlama. GitHub and Hugging Face repositories.

Zeqiu Wu, Ryu Parish, Hao Cheng, Sewon Min, Prithviraj Ammanabrolu, Mari Ostendorf, and Hannaneh Hajishirzi. 2023. InSCIt: Information-seeking conversations with mixed-initiative interactions. *Transactions of the Association for Computational Linguistics*, 11:453–468.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. 2024. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. *Preprint*, arXiv:2406.12045.

Fanghua Ye, Meng Fang, Shenghui Li, and Emine Yilmaz. 2023. Enhancing conversational search: Large language model-aided informative query rewriting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5985–6006, Singapore. Association for Computational Linguistics.

Chen Zhang, Luis Fernando D'Haro, Yiming Chen, Malu Zhang, and Haizhou Li. 2024. A comprehensive analysis of the effectiveness of large language models as automatic dialogue evaluators. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023a. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023b. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

Lexin Zhou, Youmna Farag, and Andreas Vlachos. 2024. An LLM feature-based framework for dialogue constructiveness assessment. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5389–5409, Miami, Florida, USA. Association for Computational Linguistics.

A Relation to Bayesian Linear Regression

Let $y_i \in \mathbb{R}$ be the output of a linear function of the input $\mathbf{x} \in \mathbb{R}^d$:

$$y_i = \mathbf{w}^\top \mathbf{x}_i + \varepsilon_i \tag{23}$$

and assume $\varepsilon_i \sim \mathcal{N}(0, \sigma_n^2)$ is homoscedastic Gaussian noise and assume a gaussian prior $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \sigma_p^2 \mathbf{I})$. Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ whose rows are $\mathbf{x}_i^\top$, and let $\mathbf{y} \in \mathbb{R}^n$ collect the outputs y_i. Using bayes rule we get the log posterior

$$\log p(\mathbf{w} \mid \mathbf{X}, \mathbf{y}) = \log p(\mathbf{w}) + \tag{24}$$

$$\log p(\mathbf{y} \mid \mathbf{X}, \mathbf{w}) + \text{const}, \tag{25}$$

then, via independence of the samples this becomes

$$= \log p(\mathbf{w}) + \sum_{i=1}^{n} \log p(y_i \mid \mathbf{x}_i, \mathbf{w}) + \text{const}, \tag{26}$$

using the gaussian prior this becomes

$$= -\frac{1}{2}\Big[\sigma_p^{-2}\|\mathbf{w}\|_2^2 + \sigma_n^{-2}\sum_{i=1}^{n}(y_i - \mathbf{w}^\top \mathbf{x}_i)^2\Big] + \text{const} \tag{27}$$

$$= -\frac{1}{2}\Big[\sigma_p^{-2}\|\mathbf{w}\|_2^2 + \sigma_n^{-2}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2\Big] + \text{const} \tag{28}$$

$$= -\frac{1}{2}\Big[\sigma_p^{-2}\mathbf{w}^\top\mathbf{w} + \sigma_n^{-2}\big(\mathbf{w}^\top\mathbf{X}^\top\mathbf{X}\mathbf{w} - 2\mathbf{y}^\top\mathbf{X}\mathbf{w} + \mathbf{y}^\top\mathbf{y}\big)\Big] + \text{const} \tag{29}$$

$$= -\frac{1}{2}\Big[\mathbf{w}^\top(\sigma_n^{-2}\mathbf{X}^\top\mathbf{X} + \sigma_p^{-2}\mathbf{I})\mathbf{w} - 2\sigma_n^{-2}\mathbf{y}^\top\mathbf{X}\mathbf{w}\Big] + \text{const} \tag{30}$$

we can see

$$\mathbf{\Sigma} = \Big(\sigma_n^{-2}\mathbf{X}^\top\mathbf{X} + \sigma_p^{-2}\mathbf{I}\Big)^{-1} \tag{31}$$

and hence

$$\mathbf{\Sigma}^{-1} = \sigma_n^{-2}\mathbf{X}^\top\mathbf{X} + \sigma_p^{-2}\mathbf{I} \tag{32}$$

B Proof of Theorem 3.1

Proof. Assume $\mathbf{\Sigma}_0 \succ \mathbf{0}$ with weights $w_{t,i} > 0$, recall

$$\mathbf{J}_t = \mathbf{J}_{t-1} + \sum_{i=1}^{m_t} \frac{w_{t,i}}{\sigma^2}\, \mathbf{z}_{t,i}\mathbf{z}_{t,i}^\top. \tag{33}$$

Bayesian Linear Regression	Our Formulation
Latent parameter $\mathbf{w}$	Latent semantic state $\mathbf{X}$
Input vector $\mathbf{x}_i$	Evidence embedding $\mathbf{z}_{t,i}$
Scalar observation y_i	Unobserved measurement value
Noise variance σ_n^2	Evidence noise σ^2
Prior covariance $\sigma_p^2\mathbf{I}$	Initial covariance $\mathbf{\Sigma}_0$
Design matrix $\mathbf{X}$	Stack of evidence embeddings
Posterior covariance $\mathbf{\Sigma}$	Posterior semantic uncertainty
Precision update $\mathbf{x}_i\mathbf{x}_i^\top$	Rank-one update $\mathbf{z}_{t,i}\mathbf{z}_{t,i}^\top$

Table 3: Correspondence between Bayesian linear regression and our information-theoretic formulation.

which means $\mathbf{\Sigma}_t \preceq \mathbf{\Sigma}_{t-1}$. Let

$$\mathbf{A}_t \triangleq \sum_{i=1}^{m_t} \frac{w_{t,i}}{\sigma^2}\, \mathbf{z}_{t,i}\mathbf{z}_{t,i}^\top \tag{34}$$

Since $w_{t,i} > 0$, and $\mathbf{z}\mathbf{z}^\top \succeq \mathbf{0}$, each term is PSD, and therefore $\mathbf{A}_t \succeq \mathbf{0}$. The update is

$$\mathbf{J}_t = \mathbf{J}_{t-1} + \mathbf{A}_t \tag{35}$$

which implies $\mathbf{J}_t \succeq \mathbf{J_{t-1}}$. Because $\mathbf{\Sigma}_0 \succ \mathbf{0}$, and $\mathbf{J}_0 \succeq \mathbf{0}$, by induction $\mathbf{J}_t \succ \mathbf{0}$ for all t, so $\mathbf{\Sigma}_t^{-1} = \mathbf{J}_t$ is well-defined. Using the fact for any $\mathbf{\Lambda} \succ \mathbf{0}, \mathbf{\Gamma} \succ \mathbf{0}$, and $\mathbf{\Lambda} \succeq \mathbf{\Gamma}$, we have $\mathbf{\Lambda}^{-1} \preceq \mathbf{\Gamma}^{-1}$. Applying to $\mathbf{\Lambda} = \mathbf{J}_t$ and $\mathbf{\Gamma} = \mathbf{J}_{t-1}$, we get $\mathbf{J}_t^{-1} \prec \mathbf{J}_{t-1}^{-1}$ meaning $\mathbf{\Sigma}_t \preceq \mathbf{\Sigma}_{t-1}$. Thus, posterior covariance is monotone non-increasing in PSD order.

Next, recall

$$\mathrm{IG}_t = \frac{1}{2}\log\frac{\det(\mathbf{\Sigma}_{t-1})}{\det(\mathbf{\Sigma}_t)}. \tag{36}$$

Since, $\mathbf{\Sigma}_t \preceq \mathbf{\Sigma}_{t-1}$ the ratio is at least 1 and hence $\mathrm{IG}_t \geq 0$. Finally, telescoping follows alternating sums

$$\sum_t \mathrm{IG}_t = \frac{1}{2}\sum_t \log\det(\Sigma_{t-1}) - \log\det(\Sigma_t) \tag{37}$$

$$= \frac{1}{2}\log\det(\Sigma_0) - \log\det(\Sigma_T) \tag{38}$$

$$= \frac{1}{2}\log\frac{\det\Sigma_0}{\det\Sigma_T} \tag{39}$$

This proves both monotonicity and telescoping. □

C Proof of Theorem 3.2

Let

$$\mathbf{J}(\mathcal{S}) \;=\; \mathbf{J}_0 \;+\; \sum_{i\in\mathcal{S}} \alpha_i\, \mathbf{z}_i\mathbf{z}_i^\top, \quad J_0 \succ 0, \alpha_i > 0 \tag{40}$$

and

$$F(\mathcal{S}) \triangleq \log\det \mathbf{J}(\mathcal{S}). \tag{41}$$

We first show F is monotone:

Proof. First, to show monotonicity: for any $\mathcal{S}$ and element $e \notin \mathcal{S}$:

$$\mathbf{J}(\mathcal{S}\cup\{e\}) = \mathbf{J}(\mathcal{S}) + \alpha_e\, \mathbf{z}_e\mathbf{z}_e^\top \succeq \mathbf{J}(\mathcal{S}) \tag{42}$$

Since $\log\det(\cdot)$ is increasing over the PD cone* under PSD increments, $F(\mathcal{S}\cup\{e\}) \geq F(\mathcal{S})$ □

Next, we show F is submodular:

Proof. Consider the marginal gain of adding an element e:

$$\Delta_e(\mathcal{S}) \triangleq F(\mathcal{S}\cup\{e\}) - F(\mathcal{S}) \tag{43}$$

$$= \log\frac{\det(\mathbf{J}(\mathcal{S}) + \alpha_e\, \mathbf{z}_e\mathbf{z}_e^\top)}{\det(\mathbf{J}(\mathcal{S}))} \tag{44}$$

using the matrix determinant lemma

$$\det(\mathbf{J}(\mathcal{S}) + \alpha_e\, \mathbf{z}_e\mathbf{z}_e^\top) = \det(\mathbf{J})(1 + \alpha_e\, \mathbf{z}^\top\mathbf{J}^{-1}\mathbf{z}) \tag{45}$$

Therefore

$$\Delta_e(\mathcal{S}) = \log(1 + \alpha_e\mathbf{z}_e^\top\mathbf{J}(\mathcal{S})^{-1}\mathbf{z}_e) \tag{46}$$

Now to show submodularity, take $\mathcal{A} \subseteq \mathcal{B}$ which gives us $\mathbf{J}(\mathcal{A}) \preceq \mathbf{J}(\mathcal{B})$ from the earlier proof, implying $\mathbf{J}(\mathcal{A})^{-1} \succeq \mathbf{J}(\mathcal{B})^{-1}$. Then, the following must hold:

$$\alpha_e\mathbf{z}_e^\top\mathbf{J}(\mathcal{A})^{-1}\mathbf{z}_e \geq \alpha_e\mathbf{z}_e^\top\mathbf{J}(\mathcal{B})^{-1}\mathbf{z}_e \tag{47}$$

because $x \mapsto \log(1+\alpha_e x)$ is increasing for $\alpha_e \geq 0$. Therefore, $\Delta_e(\mathcal{A}) \geq \Delta_e(\mathcal{B})$ substituting back we get the conditions for submodularity:

$$F(\mathcal{A}\cup\{e\}) - F(\mathcal{A}) \;\geq\; F(\mathcal{B}\cup\{e\}) - F(\mathcal{B}). \tag{48}$$

□

D Proofs for Irrelevance and Redundancy Lemmas

We will be using two important facts:

Order Reversal under Inversion For any $\mathbf{A} \succ \mathbf{0}, \mathbf{B} \succ \mathbf{0}$, if $\mathbf{A} \prec \mathbf{B}$ then $\mathbf{A}^{-1} \succ \mathbf{B}^{-1}$.

Rayleigh Quotient Bound For a symmetric $\mathbf{M} \succeq \mathbf{0}$,

$$\mathbf{z}^\top\mathbf{M}\mathbf{z} \leq \lambda_{\max}(\mathbf{M})\|\mathbf{z}\|_2^2 \tag{49}$$

D.1 Proof of Theorem 3.3

Proof. Assume $\|\mathbf{z}\| < B$, we have

$$\Delta(\mathbf{z}; \mathbf{J}) = \log\Big(1 + \alpha\, \mathbf{z}^\top \mathbf{J}^{-1} \mathbf{z}\Big). \tag{50}$$

Since $\mathbf{J} \succeq \mathbf{J}_0 \succ \mathbf{0}$, by order reversal under inversion, we know $\mathbf{J}^{-1} \preceq \mathbf{J}_0^{-1}$ and hence

$$\mathbf{z}^\top \mathbf{J}^{-1} \mathbf{z} < \mathbf{z}^\top \mathbf{J}_0^{-1} \mathbf{z} \tag{51}$$

Since $\mathbf{J}_0$ is a covariance matrix, that means they are always symmetric. Applying the Rayleigh quotient bound, we get

$$\mathbf{z}^\top \mathbf{J}_0^{-1} \mathbf{z} \le \lambda_{\max}(\mathbf{J}_0^{-1}) \|\mathbf{z}\|_2^2 \le \lambda_{\max}(\mathbf{J}_0^{-1}) B^2 \tag{52}$$

substituting back into $\Delta(\mathbf{z}; \mathbf{J})$ we get

$$\Delta(\mathbf{z}; \mathbf{J}) = \log\Big(1 + \alpha\, \mathbf{z}^\top \mathbf{J}^{-1} \mathbf{z}\Big) \tag{53}$$

$$\le \log(1 + \alpha \lambda_{\max}(\mathbf{J}_0^{-1}) B^2) \tag{54}$$

Finally, if $\alpha \le \varepsilon$, then $\Delta(\mathbf{z}; \mathbf{J}) \le \log(1 + C\varepsilon) = \mathcal{O}(\varepsilon)$ where $C = \lambda_{\max}(\mathbf{J}_0^{-1}) B^2$. □

D.2 Proof of Theorem 3.4

Proof. Let $\mathbf{z} \in \mathbb{R}^d$, $\alpha > 0$, define $\mathbf{J}_k = \mathbf{J}_0 + k\alpha \mathbf{z}\mathbf{z}^\top$ with $\mathbf{J}_0 \succ \mathbf{0}$. By definition, we know

$$\mathbf{J}_{k+1} = \mathbf{J}_k + \alpha\, \mathbf{z}^\top \mathbf{z} \succ \mathbf{0}. \tag{55}$$

, by order reversal under inversion we know $\mathbf{J}_{k+1}^{-1} \preceq \mathbf{J}_k^{-1}$, and hence

$$\mathbf{z}^\top \mathbf{J}_{k+1}^{-1} \mathbf{z} \le \mathbf{z}^\top \mathbf{J}_k^{-1} \mathbf{z} \tag{56}$$

And because $x \mapsto \log(1 + \alpha x)$ is increasing for $\alpha > 0$, we conclude

$$\Delta(\mathbf{z}; \mathbf{J}_{k+1}) = \log\Big(1 + \alpha\, \mathbf{z}^\top \mathbf{J}_{k+1}^{-1} \mathbf{z}\Big) \tag{57}$$

$$\le \log\Big(1 + \alpha\, \mathbf{z}^\top \mathbf{J}_k^{-1} \mathbf{z}\Big) = \Delta(\mathbf{z}; \mathbf{J}_k). \tag{58}$$

□

E Toy Examples

Goal. We construct a controlled information-seeking dialogue where *oracle* uncertainty can be computed exactly using a finite hypothesis universe. This allows a direct comparison between (i) oracle information gain computed by eliminating inconsistent hypotheses, and (ii) our Gaussian information gain computed purely in embedding space.

Universe and prior. Let U be a finite set of hypotheses. In our instantiations, U is a fixed list of N cities (city-guessing), a fixed list of N movies (movie-guessing), and more generally QAs related to gardening or workouts, with $N = 100$. We initialize either a uniform prior or a simple non-uniform prior (e.g., population-based city priors):

$$p_0(u) = \frac{1}{N}, \quad u \in U, \tag{59}$$

$$\mathrm{H}_0 = -\sum_{u \in U} p_0(u) \log p_0(u) = \log N. \tag{60}$$

Dialogue and evidence. We generate a multi-turn dialogue $\{(q_t, a_t)\}_{t=1}^T$ consisting of yes/no questions for city-guessing and movie-guessing, as well as general responses to gardening and workout advice questions. We compare two variants of equal length: (i) **Novel**, where each turn introduces new discriminative information, and (ii) **Redundant**, where later turns repeat earlier information (and thus should yield diminishing or zero marginal gain under the oracle).

Oracle consistency update. Let e_t denote the evidence revealed at turn t (e.g., the semantic constraint implied by (q_t, a_t)). We maintain a feasible set $S_t \subseteq U$ of hypotheses consistent with all evidence so far:

$$S_0 = U, \tag{61}$$

$$S_t = \Big\{ u \in S_{t-1} \;\Big|\; u \text{ is consistent with } e_t \Big\}. \tag{62}$$

In our implementation, since we know the universe, we can manually construct the consistency predicates at each turn given the question and answers, mapping dialogue history to a subset of surviving hypotheses, and we enforce $S_t \subseteq S_{t-1}$ when the turns are not irrelevant or redundant.

Oracle entropy and information gain. Under the uniform posterior over S_t, the oracle entropy is

$$\mathrm{H}_t = \log |S_t|. \tag{63}$$

The per-turn oracle information gain is then

$$\mathrm{IG}_t^{\text{oracle}} = \mathrm{H}_{t-1} - \mathrm{H}_t = \log \frac{|S_{t-1}|}{|S_t|} \ge 0, \tag{64}$$

and the cumulative oracle gain telescopes:

$$\sum_{t=1}^{T} \mathrm{IG}_t^{\text{oracle}} = \log \frac{|S_0|}{|S_T|} = \log \frac{N}{|S_T|}. \tag{65}$$

For redundant turns that add no new constraint, we have $S_t = S_{t-1}$ and thus $\mathrm{IG}_t^{\text{oracle}} = 0$.

Method	MT-Bench	Chatbot Arena
Mistral Large 3	131.8 ± 11.8	118.9 ± 2.0
DeepSeek R1	633.1 ± 9.2	479.6 ± 3.2
GPT OSS 120b	568.0 ± 198.0	328.3 ± 94.7
Claude Sonnet 3.7	259.9 ± 13.1	250.9 ± 23.9
Claude Sonnet 4	219.6 ± 12.7	175.4 ± 8.4
Claude Sonnet 4.5	341.9 ± 16.8	286.3 ± 12.8
Ours (Gaussian IG)	86.4 ± 6.8	44.1 ± 5.3

Table 4: Runtime variability across $R = 5$ runs. Values report mean $\pm$ standard deviation of end-to-end wall-clock time (seconds) for evaluating $N = 100$ dialogue pairs.

Comparison to Gaussian IG. For the same dialogues, we compute our Gaussian information gain using only question–answer embeddings and closed-form covariance updates (Equation (8)). We plot per-turn gain and cumulative gain for both the oracle and Gaussian metrics; across controlled settings, dialogues that introduce novel constraints for more turns achieve larger cumulative gain, while redundant variants exhibit diminished marginal gain. Refer to Figure 1, Figure 2, Figure 3, Figure 4 for results. Note that the user and LLM here are synthetic.

F Experiment Details

Timing Protocol. Let $\mathcal{D} = \{(D_i^A, D_i^B)\}_{i=1}^N$ denote a fixed ordered subset of MT-Bench dialogue pairs. For each method, we measure runtime using Python wall-clock timers (`time.perf_counter`) as follows:

1. A fixed random seed is used to select and order evaluation examples.
2. Timing starts immediately before the first scoring call.
3. Each dialogue pair is scored sequentially.
4. Timing stops after the final prediction is produced.

We report total elapsed time and normalized runtime (seconds per dialogue pair), averaged over $R = 5$ runs. We further verify that overhead from data access and prediction logic is negligible relative to scoring time. The standard deviations for the runtimes are reported in Table 4.

G Frequently Asked Questions

What does reference-free mean in this work? Reference-free means that the metric does not require human-written ground-truth answers, gold references, or human annotations at evaluation time. The score is computed solely from the dialogue history using fixed embeddings and closed-form updates.

Why assume a Gaussian model in embedding space? The Gaussian assumption is a modeling choice that enables a simple, closed-form approximation to information gain with strong theoretical guarantees. Under a Gaussian posterior, information gain reduces to a log-determinant of the covariance matrix, yielding monotonicity, telescoping across turns, and submodularity under rank-one updates.

We do not claim that semantic uncertainty is truly Gaussian; rather, the Gaussian serves as a tractable surrogate that captures relative uncertainty volume in embedding space.

Empirically, we find that this approximation is sufficient to produce a stable and informative evaluative signal across embedding models of widely varying capacity.

Why does uncertainty always decrease under your Bayesian update? Our formulation fixes the observation noise variance and updates only the posterior uncertainty over a fixed latent semantic state. Each update adds a positive semi-definite rank-one matrix to the precision matrix, which guarantees that posterior covariance is non-increasing in Loewner order. This differs from hierarchical Bayesian models in which noise variance or model parameters are inferred and posterior uncertainty may increase.

Can Bayesian updates ever increase uncertainty in general? Yes. In Bayesian models that infer observation noise or higher-level hyperparameters, posterior uncertainty can increase when observations are inconsistent with prior assumptions. Our model intentionally excludes this behavior by fixing the noise variance, enabling a fast, monotone, and tractable approximation to information gain.

Does higher information gain imply factual correctness? No. The metric measures epistemic progress rather than correctness. Confident but incorrect responses may still reduce uncertainty in embedding space. We view this limitation as shared with many automatic evaluators and consider factuality checks a complementary signal.

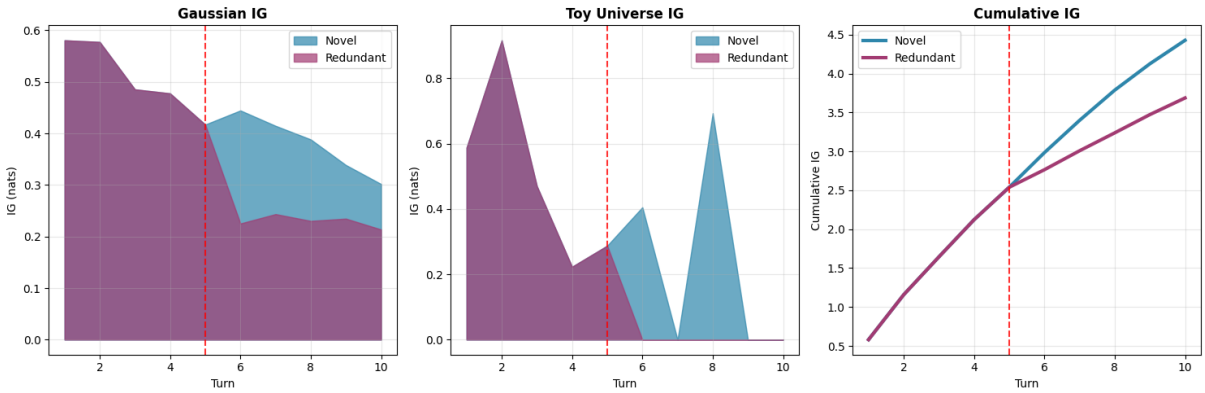

Figure 2: Toy example with the city guessing universe. The LLM tries to guess a city by asking questions and the user provides Yes/No to the questions. We compare the case where at turn 5, the LLM starts asking redundant questions to if it kept asking novel questions, in an ideal universe we know $S_t = S_{t-1}$ so information gain is 0, in our Gaussian Approximation, although information gain does not drop to 0, we see a clear difference between if the questions asked was novel or redundant.

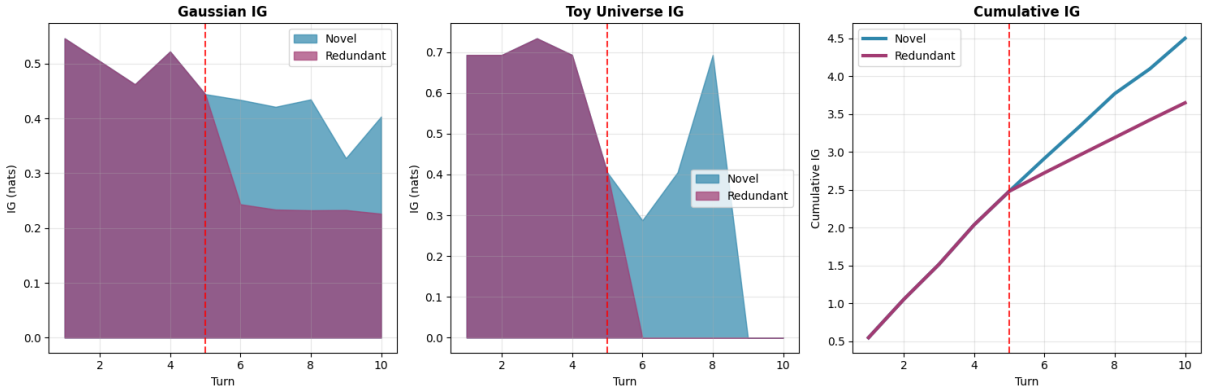

Figure 3: Toy example with the movie guessing universe. The LLM tries to guess a movie by asking questions and the user provides Yes/No to the questions. We compare the case where at turn 5, the LLM starts asking redundant questions to if it kept asking novel questions. The trends observed are similar to Figure 2.

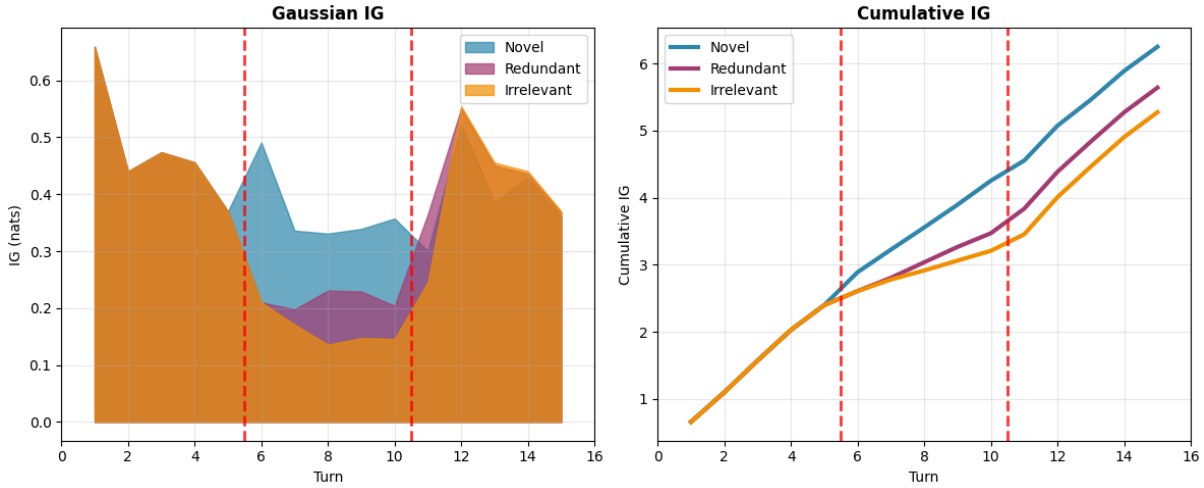

Figure 4: Toy example where the user asks open questions for advice related to gardening. The LLM tries to answer the question. We compare the cases where between turns 5 and 10, where LLM answers are either novel, redundant, or irrelevant. We observe similar trends as Figure 2, Figure 3, and also that information gain recovers after the LLM starts asking novel questions again.