Measuring Semantic Progress in Multi-turn Dialogue via Information Gain

Paul He[*]
NTU Singapore
paul005@ntu.edu.sg

Shiva Prasad Kasiviswanathan
Amazon
kasivisw@gmail.com

Dominik Janzing
Amazon
janzind@amazon.com

Abstract

Evaluating multi-turn dialogue is challenging because quality emerges across turns rather than within individual responses. We focus on a key dimension of information-seeking dialogue: *semantic progress*, defined as the accumulation of new, question-relevant, and non-redundant information over the course of a conversation. We formalize semantic progress as question-conditioned uncertainty reduction and introduce an information-theoretic metric that approximates it in embedding space. Our main estimator uses a tractable Gaussian formulation with closed-form updates, while a complementary maximum-entropy argument shows why log-determinant structure arises more broadly when only second-order embedding information is retained. This formulation yields desirable theoretical properties, including monotonicity, additive decomposition of total information gain across turns, and diminishing returns for redundant evidence. Unlike LLM-as-a-judge approaches, our metric requires no autoregressive inference at evaluation time and is fully reproducible for a fixed embedding model. Experiments on MT-Bench, Chatbot Arena, and UltraFeedback show that the proposed metric achieves competitive agreement with human judgments despite targeting only semantic progress, with improved alignment on MT-Bench and UltraFeedback compared to several LLM-based judges. Notably, the method remains effective with lightweight embedding models under CPU-only execution, indicating that semantic progress can be captured without reliance on large model capacity.

1 Introduction

Large Language Models (LLMs) are increasingly deployed as conversational agents across diverse settings. Some interactions are open-ended, creative, or social, while others are primarily information-seeking, such as question answering and decision support [1–4]. These use cases impose different requirements on evaluation, and no single metric can adequately capture all aspects of dialogue quality, including fluency, factuality, safety, style, user satisfaction, and task success.

In this work, we focus on a practically important subset: information-seeking dialogue. In such interactions, users iteratively refine their queries across turns, while agents provide clarifications, evidence, and partial answers that accumulate over the course of the conversation [5]. As a result, evaluating these dialogues requires more than assessing individual responses in isolation—it requires measuring how the conversation evolves over time [6, 7].

This temporal perspective is particularly important in real-world deployment. Developers often need to compare model variants, perform regression testing, monitor live systems, and analyze large volumes of user interactions for quality degradation. These workflows demand evaluation signals that are fast, reproducible, scalable, and informative at the level of entire dialogues [6, 8, 9].

[*]Work done while at Amazon Research, Tübingen, Germany

Preprint.

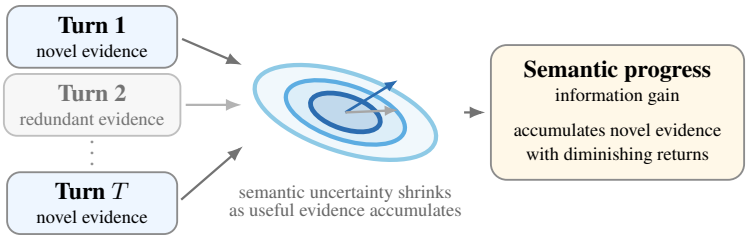


Figure 1: Conceptual view of *semantic progress* as captured by our information-gain metric. Useful turns introduce question-relevant evidence that reduces semantic uncertainty, while redundant turns yield diminishing marginal gains. Our metric approximates this uncertainty reduction in embedding space and accumulates it as dialogue-level information gain.

We focus on one such signal: whether an information-seeking dialogue accumulates new, question-relevant, and non-redundant information over time. We refer to this dimension as *semantic progress*. Effective information-seeking dialogues should exhibit consistent progress across turns while avoiding redundancy and irrelevance [7, 10]. Unlike holistic notions of dialogue quality, semantic progress does not aim to measure fluency, style, safety, or factual correctness. Instead, it captures whether the conversation reduces uncertainty about the information needed to address the user's request (see Figure 1).

We formalize semantic progress as *question-conditioned semantic uncertainty reduction*: as a dialogue unfolds, accumulated answer evidence should reduce uncertainty along directions relevant to the user's information need. In principle, this reduction could be defined over explicit facts, possible worlds, or structured answer states. However, such representations are typically unavailable in open-ended natural language settings. We therefore approximate this notion directly in embedding space, treating answer evidence as a sequence of observations that progressively shrink the covariance of a latent semantic state along question-relevant directions. The resulting information-gain score is deterministic, reference-free, and can be computed without autoregressive judge inference.

Our approach is complementary to, rather than a replacement for, holistic evaluators such as LLM-as-a-judge methods. While LLM judges capture broad aspects of dialogue quality, they are computationally expensive and sensitive to prompts, model versions, and stochastic inference. In contrast, our metric targets a specific dimension—semantic progress—while providing a fast, deterministic, and reproducible signal suitable for large-scale evaluation. We focus on semantic progress as one evaluation dimension; factuality, safety, coherence, and style are each studied by extensive evaluation literature [11, 12] and are complementary to the signal proposed here. These external signals can easily be incorporated through additional weighting schemes, but evaluating those combinations is outside the scope of this work. We make the following contributions:

- **Information-theoretic perspective.** We formulate multi-turn dialogue evaluation as question-conditioned semantic uncertainty reduction, giving a model-agnostic definition of semantic progress as the accumulation of new, relevant, non-redundant information.
- **Efficient, training-free estimator.** We introduce a deterministic, reference-free Gaussian approximation in embedding space with closed-form updates and guarantees of monotonicity, telescoping aggregation, and diminishing returns.
- **Empirical validation at scale.** On MT-Bench, Chatbot Arena, and UltraFeedback, GAUSSIANIG reaches 84.03%, 65.80%, and 71.64% agreement, respectively, while reducing evaluation time by 3–10x relative to hosted LLM judges and supporting CPU-only execution.

2 Related Work

Task-Oriented and Structured Dialogue Evaluation. In structured and task-oriented settings, dialogue quality is typically assessed using a combination of turn-level and dialogue-level signals, often augmented with learned evaluators or LLM-based judges. For instance, TD-EVAL employs a two-stage framework that combines turn-level metrics with dialogue-level aggregation and LLM judgments, achieving improved alignment with human preferences on benchmarks such as MultiWOZ and Tau-Bench [13–15]. These approaches are tailored to task-specific success criteria and generally

depend on supervised models or external evaluators. In contrast, our method abstracts multi-turn dialogue progress as question-conditioned uncertainty reduction, providing a task-agnostic, training-free signal defined directly in embedding space.

Dialogue-Level Coherence and Consistency Metrics. Another line of work evaluates dialogue quality using coherence, consistency, and contextual appropriateness across turns [16, 17]. These methods assess whether a dialogue remains internally consistent and contextually appropriate given its history. However, they do not explicitly model dialogue-level progress or account for diminishing returns from redundant information, which are central to our formulation.

Large Language Models as Dialogue Evaluators. A substantial body of work explores large language models as dialogue evaluators, examining their agreement with human judgments and sensitivity to prompting and evaluation protocols [18]. LLM-as-a-judge approaches can achieve strong alignment with human preferences, as demonstrated by benchmarks such as MT-Bench 101, where carefully prompted models (e.g., GPT-4) attain high agreement on multi-turn dialogue quality [19]. However, these methods incur significant computational and operational costs [20], relying on large models, prompt engineering, and repeated autoregressive inference. This makes them expensive and often impractical for large-scale model comparison, regression testing, or continuous monitoring.

Learned Evaluators and Judge Approximation. Closely related are learned evaluators that directly approximate human preferences or LLM-based judges, including dialogue-level predictors, pairwise comparison models, and feature-based frameworks [21–24]. While often effective, these approaches require supervised training and inherit the opacity, bias, and distributional assumptions of the signals they are trained to replicate.

Positioning of Our Work. Prior work spans task-specific evaluators for structured dialogue, coherence and consistency metrics, LLM-as-a-judge approaches, and learned evaluators that approximate human or model judgments. These methods either target holistic quality dimensions or rely on supervised training and prompt-dependent inference, often incurring substantial computational cost and limited reproducibility [8, 25, 24]. In contrast, we isolate a complementary dimension, *semantic progress*, and formalize it as question-conditioned uncertainty reduction. Our approach is task-agnostic, training-free, and avoids autoregressive inference, producing a fast, deterministic, and reference-free dialogue-level signal. It further provides explicit theoretical guarantees, including monotonicity and diminishing returns, enabling scalable and reliable evaluation pipelines while remaining complementary to holistic LLM-based judges.

3 Setting the Stage

Let $\mathcal{A}$ be a finite **alphabet** and $\mathcal{A}^*$ the set of all finite strings over $\mathcal{A}$. A T-turn **question–answer dialogue** is a sequence $D_{1:T} \triangleq \big((q_1, a_1), \ldots, (q_T, a_T)\big)$ where $(q_t, a_t) \in \mathcal{A}^* \times \mathcal{A}^*$ denotes question and answer in step $t \in [T]$. We assume each answer can be decomposed into a finite collection of smaller semantic units, which we refer to as **evidence items**. Let $\mathcal{E}$ denote the universe of all possible evidence items (e.g., semantic chunks of responses). Let $E_t \subseteq \mathcal{E}$ denote the **set of evidence** extracted from answer a_t. We define $A_t \triangleq \bigcup_{s=1}^{t} E_s$ as the **accumulated evidence** observed up to turn t. Let q denote the user's information need (e.g., dialogue context, question). We define a **relevance function** $\rho_q : \mathcal{E} \to \mathbb{R}_{\geq 0}$, where ρ_q measures how relevant an evidence item e is to q. We consider dialogue progress as a function of accumulated evidence. A **semantic progress measure** is a function $\mathcal{P}_q : 2^{\mathcal{E}} \to \mathbb{R}_{\geq 0}$ with marginal gain

$$\Delta_q(e \mid A) = \mathcal{P}_q(A \cup \{e\}) - \mathcal{P}_q(A) \tag{1}$$

A dialogue prefix $D_{1:t}$ is scored by first extracting its accumulated evidence A_t, then evaluating $\mathcal{P}_q(A_t)$ relative to the information need q. A useful measure of semantic progress in information-seeking dialogue should satisfy **irrelevance** (if $\rho_q(e) = 0$, then $\Delta_q(e \mid A) = 0$), **monotonicity** ($\Delta_q(e \mid A) \geq 0$), and **diminishing returns** (if $A \subseteq B$ then $\Delta_q(e \mid A) \geq \Delta_q(e \mid B)$). These capture that progress accumulates with relevant evidence, does not decrease when adding information, and discounts redundancy.

User Perspective and Agent-Induced Information Gain. Dialogue information gain can be viewed from two perspectives. A third-party observer may learn from both user questions and agent answers. From the user's perspective, however, the question is already known and should condition the current information need. Since we evaluate agent progress, we adopt the user perspective.

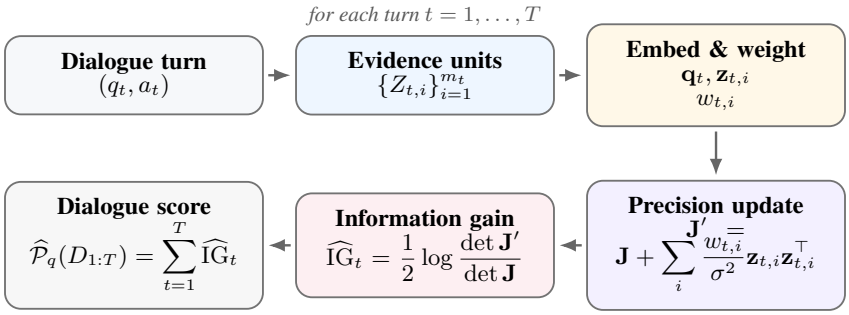


Figure 2: Algorithmic overview of Gaussian information-gain evaluation. Each answer is split into evidence units, embedded, weighted by question relevance, and incorporated through a weighted precision update. The resulting log-determinant change gives turn-level information gain, which is summed into the dialogue-level semantic progress score.

Let $C_{t-1}^A \triangleq (q_1, a_1, \ldots, q_{t-1}, a_{t-1})$ denote the context before the user's t-th turn. After the user asks q_t, define $C_t^U \triangleq (C_{t-1}^A, q_t)$, and after the agent answers a_t, define $C_t^A \triangleq (C_t^U, a_t)$. Thus C_t^U is the context available immediately before the agent answer, and C_t^A is the context after the full turn. Let W be a latent task-relevant state. By the chain rule for mutual information,

$$I(W; D_{1:T}) = \sum_{t=1}^{T} I(W; q_t \mid C_{t-1}^A) + \sum_{t=1}^{T} I(W; a_t \mid C_t^U). \tag{2}$$

The first term is user-induced task specification; the second is agent-induced semantic progress. The second term is not meant to equal the user's subjective belief update, since the user may have private knowledge not manifested in the transcript. Rather, we use it as a transcript-level proxy for agent-induced progress: the information added by the agent's answers after conditioning on the expressed user need. We therefore use $G^\star(D_{1:T}) \triangleq \sum_{t=1}^{T} I(W; a_t \mid C_t^U)$ as an idealized transcript-level proxy for agent-induced semantic progress. This decomposition allows q_t to depend on earlier answers, since C_{t-1}^A contains the full dialogue history. It also avoids penalizing the agent when the user makes the task broader or more ambiguous. Our estimator approximates $G^\star$ by treating agent answers as evidence weighted by relevance to the current user turn.

A Hypothesis. Let q denote a user information need, and let $D_{1:T}^A$ and $D_{1:T'}^B$ be two information seeking dialogues for the same q. Let A_T^A and A_T^B denote their accumulated evidence. Let W be a latent random variable representing the task-relevant state of the world (i.e., the information needed to resolve the user's request). We write $H(\cdot)$ for Shannon entropy of discrete variables and $h(\cdot)$ for differential entropy of continuous variables. We define the realized ideal semantic progress of an observed evidence accumulation a_T as

$$\mathcal{P}_q^\star(a_T) \triangleq H(W \mid q) - H(W \mid q, A_T = a_T) \tag{3}$$

where its expectation over the random evidence accumulation A_T is the conditional mutual information $\mathbb{E}_{A_T}[\mathcal{P}_q^\star(A_T)] = I(W; A_T \mid q)$. If human preference between the two dialogues is primarily driven by the accumulation of new, question relevant, non-redundant information, then the preferred dialogue is most likely to satisfy $\mathcal{P}_q^\star(a_T^A) > \mathcal{P}_q^\star(a_T^B)$ where a_T^A and a_T^B are the observed accumulated evidence sets. Since $\mathcal{P}_q^\star$ is not directly computable in open-ended dialogue, we test this hypothesis using a tractable estimator $\widehat{\mathcal{P}}_q$, predicting the dialogue with higher $\widehat{\mathcal{P}}_q$ as preferred and measuring agreement with human judgments.

4 Our Method

As an idealized symbolic analogy, suppose there is a finite hypothesis space of possible worlds, a fixed background theory, a uniform posterior over worlds consistent with the accumulated facts F_t, and deterministic evidence that only rules out inconsistent worlds. If $\mathcal{M}(F_t)$ denotes the surviving worlds after turn t, then the turn-level gain is $\widehat{\mathrm{IG}}_t = \log \frac{|\mathcal{M}(F_{t-1})|}{|\mathcal{M}(F_t)|}$. This captures uncertainty reduction,

telescoping progress, and diminishing returns for repeated evidence. Since natural-language does not provide explicit facts, a background theory, or tractable model counts, we approximate this idea with embeddings. Figure 2 summarizes the proposed pipeline, showing how dialogue turns are decomposed into evidence, embedded, and aggregated via precision updates. The full procedure is given in Algorithm 1 (Appendix H). The ideal score $\mathcal{P}_q^\star(a_T)$ measures information gained about a latent task-relevant state, but requires symbolic evidence, a posterior over possible worlds, and model counting. Our estimator replaces this symbolic uncertainty with Gaussian uncertainty in embedding space. The resulting score, denoted $\widehat{\mathcal{P}}_q(D_{1:T})$, is deterministic, reference-free, and computed from the dialogue text alone.

4.1 Gaussian Approximation to Information Gain

For each turn t, we extract a finite collection of evidence strings $\mathrm{Z}_t \triangleq \{\mathrm{Z}_{t,1}, \ldots, \mathrm{Z}_{t,m_t}\} \subseteq \mathcal{A}^*$ from the answer a_t. In our experiments, these evidence strings are sentence-level units (details later in the paper) though other deterministic segmentations can be used. This gives a concrete approximation to the abstract evidence set E_t. We write $\mathrm{enc} : \mathcal{A}^* \to \mathbb{R}^d$ for a fixed text embedding function, and denote

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

for the question and evidence embeddings. We introduce a latent semantic uncertainty variable $\mathbf{S} \in \mathbb{R}^d$, which represents unresolved semantic uncertainty in the embedding space. This variable is an abstract modeling device, not a literal world state or fact representation. We place an isotropic Gaussian prior

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

We justify the resulting log-determinant score through a maximum-entropy argument, formalized in Theorem 1. Among continuous distributions with fixed covariance, the Gaussian maximizes differential entropy. Since our estimator retains only second-order embedding statistics, Gaussian entropy gives a least-committal uncertainty surrogate, and entropy reduction becomes a log-determinant covariance ratio.

Each evidence embedding induces a scalar linear measurement of the **latent semantic state**:

$$y_{t,i} = \mathbf{z}_{t,i}^\top \mathbf{S} + \boldsymbol{\varepsilon}_{t,i}, \quad \varepsilon_{t,i} \sim \mathcal{N}(0, \sigma^2), \tag{4}$$

Appendix A provides intuition for viewing an embedding direction as a one-dimensional measurement of $\mathbf{S}$. Since the posterior covariance in Bayesian linear regression depends on the measurement directions and noise variances, but not on the observed scalar values, the specific values $y_{t,i}$ do not need to be observed or estimated; see Appendix B. To instantiate **question-conditioned relevance**, we assign each evidence unit a nonnegative weight $w_{t,i} = r_{t,i}^{\mathrm{rel}} r_{t,i}^{\mathrm{qual}}$ where

$$r_{t,i}^{\mathrm{rel}} = \max(0, \langle \widehat{\mathbf{q}}_t, \widehat{\mathbf{z}}_{t,i} \rangle)^\beta \tag{5}$$

where $\widehat{\mathbf{v}} = \mathbf{v}/\|\mathbf{v}\|_2$ and $r_{t,i}^{\mathrm{qual}} \in [0,1]$ is an optional external evidence-quality factor. In this work, set $r_{t,i}^{\mathrm{qual}} = 1$ and evaluate relevance-weighted semantic progress only. Optionally, we set $w_{t,i} = 0$ when $w_{t,i} < \eta$. The weight $w_{t,i}$ is the embedding-space instantiation of the relevance function $\rho_q(e)$: evidence weakly aligned with the question contributes little or no posterior update, while aligned evidence reduces uncertainty along its semantic direction.

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

$$\mathbf{J}_t = \mathbf{J}_{t-1} + \sum_{i=1}^{m_t} \frac{w_{t,i}}{\sigma^2} \mathbf{z}_{t,i} \mathbf{z}_{t,i}^\top. \tag{6}$$

This precision update is the information-form posterior covariance update for Bayesian linear regression with evidence embeddings as measurement directions. The relevance weights can be

equivalently interpreted as heteroscedastic observation noise, with $\varepsilon_{t,i} \sim \mathcal{N}(0, \sigma^2/w_{t,i})$ for $w_{t,i} > 0$, and $w_{t,i} = 0$ contributing no update. Appendix B gives the derivation. From now on, denote $\alpha_{t,i} = \frac{w_{t,i}}{\sigma^2}$. The differential entropy of a d-dimensional Gaussian $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is

$$h(\mathbf{S}) = \frac{d}{2}\log(2\pi e) + \frac{1}{2}\log\det(\boldsymbol{\Sigma}). \tag{7}$$

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

$$\widehat{\mathrm{IG}}_t \triangleq h(\mathbf{S} \mid \mathrm{Z}_{1:t-1}) - h(\mathbf{S} \mid \mathrm{Z}_{1:t}) = \frac{1}{2}\log\frac{\det(\boldsymbol{\Sigma}_{t-1})}{\det(\boldsymbol{\Sigma}_t)}. \tag{8}$$

which gives us the dialogue level estimator: $\widehat{\mathcal{P}}_q(D_{1:T}) \triangleq \sum_{t=1}^{T} \widehat{\mathrm{IG}}_t$.

The log-determinant score can be motivated independently of the Gaussian posterior model. In open-ended dialogue, we do not observe symbolic facts or a tractable posterior over possible worlds; instead, our estimator $\widehat{\mathrm{IG}}$ retains only second-order structure of the evidence embeddings. Under such a covariance-only description, the Gaussian distribution is the maximum-entropy distribution, so $\log\det\boldsymbol{\Sigma}$ is the least-committal surrogate for uncertainty volume. Theorem 1 formalizes this. It shows that, when evidence is represented only through its empirical covariance, the Gaussian log-determinant quantity upper-bounds the information that the evidence can reveal about the latent semantic state, up to a conditional-uncertainty term.

Theorem 1. *For a fixed turn $t \in [T]$, let $\mathbf{Z}$ be the random evidence vector corresponding to $\mathbf{z} = \langle \mathbf{z}_1, \ldots, \mathbf{z}_m \rangle \in \mathbb{R}^{md}$ with evidence components $\mathbf{Z}_i \in \mathbb{R}^d$. If $h(\mathbf{Z} \mid \mathbf{S}) \geq c$, where c is independent of the dialogue strategy, and the evidence has common empirical covariance $\boldsymbol{\Sigma}_{\mathbf{Z}} \succ \mathbf{0}$, then:*

$$I(\mathbf{Z}; \mathbf{S}) \leq m\big[\frac{d}{2}\log(2\pi e) + \frac{1}{2}\log\det\boldsymbol{\Sigma}_{\mathbf{Z}}\big] - c \tag{9}$$

We prove this in Appendix C. This reflects the information-theoretic principle that high-information evidence must be sufficiently non-compressible. In our setting, larger effective covariance volume corresponds to evidence that can potentially carry more information about the latent semantic state.

This estimator $\widehat{\mathrm{IG}}$ follows the same structure as the ideal symbolic formulation: evidence accumulates over turns, uncertainty decreases through posterior updates, and dialogue-level progress is the total reduction in uncertainty. The next results show that the estimator preserves the desired structural properties of semantic progress.

4.2 Structural Properties of the Estimator

The Gaussian estimator inherits the structural properties desired of a semantic progress measure. Because each evidence item contributes a nonnegative rank-one update to the precision matrix, posterior uncertainty can only decrease. This gives monotonicity and a telescoping dialogue-level score. Proofs of the theorems are provided in Appendix D, and Appendix E.

Theorem 2 (Monotonicity and Telescoping)**.** *Assume $\boldsymbol{\Sigma}_0 \succ \mathbf{0}$ and weights $\alpha_{t,i} \geq 0$. Under the precision update in Equation* (6)*, we have $\boldsymbol{\Sigma}_t \preceq \boldsymbol{\Sigma}_{t-1}$ (in PSD order) and hence $\widehat{\mathrm{IG}}_t \geq 0$ for all t. Moreover, the total gain telescopes:*

$$\sum_{t=1}^{T} \widehat{\mathrm{IG}}_t = \frac{1}{2}\log\frac{\det(\boldsymbol{\Sigma}_0)}{\det(\boldsymbol{\Sigma}_T)}, \tag{10}$$

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

Thus, adding evidence cannot reduce the score, and the dialogue-level value is exactly the total reduction in posterior uncertainty. We next show that the estimator also discounts redundancy. Let $\mathcal{X}$ be a finite set of evidence items, where each item $i \in \mathcal{X}$ has embedding $\mathbf{z}_i \in \mathbb{R}^d$ and weight $\alpha_i \geq 0$. For any subset $\mathcal{S} \subseteq \mathcal{X}$, define $\mathbf{J}(\mathcal{S}) = \mathbf{J}_0 + \sum_{i \in \mathcal{S}} \alpha_i \mathbf{z}_i \mathbf{z}_i^\top$ and $F(\mathcal{S}) = \log\det\mathbf{J}(\mathcal{S})$.

Theorem 3 (Diminishing Returns)**.** *Assume $\mathbf{J}_0 \succ 0$. The set function F is monotone submodular. That is, for all $A \subseteq B \subseteq \mathcal{X}$ and $x \notin B$,*

$$F(A \cup \{x\}) - F(A) \geq F(B \cup \{x\}) - F(B).$$

Since total Gaussian information gain is proportional to $\log\det \mathbf{J}(\mathcal{S})$ up to constants, Theorem 3 implies diminishing marginal returns: evidence contributes less when similar evidence has already been accumulated. This is the formal mechanism by which the estimator avoids rewarding redundancy or verbosity alone.

Implication. The estimator exhibits the desired properties for semantic-progress evaluation: information gains compose across turns, irrelevant evidence contributes negligibly, and redundant evidence yields diminishing marginal returns. Consequently, high scores arise only from sustained, question-relevant reduction in semantic uncertainty, rather than verbosity or repetition.

4.3 Local Behavior: Irrelevance and Repetition

The previous subsection establishes global properties. We now make the local behavior of a single evidence item explicit. For an evidence embedding $\mathbf{z}$ with weight $\alpha \geq 0$, define its marginal contribution at precision $\mathbf{J}$ as

$$\Delta(\mathbf{z}; \mathbf{J}) \triangleq \log\det(\mathbf{J} + \alpha \mathbf{z}\mathbf{z}^\top) - \log\det(\mathbf{J}) \tag{11}$$

$$= \log\big(1 + \alpha \mathbf{z}^\top \mathbf{J}^{-1} \mathbf{z}\big)\,, \tag{12}$$

where we used the matrix determinant lemma in the last equality. Thus, if the evidence is completely irrelevant under the question-conditioned weighting, then $\alpha = 0$ and its gain is exactly zero. If α is small, the gain is correspondingly bounded. Proofs of the lemmas are provided in Appendix F.1 and Appendix F.2.

Lemma 1 (Soft Irrelevance)**.** *Assume embeddings are norm bounded $\|\mathbf{z}\| \leq B$, and $\alpha \leq \varepsilon$. When $\mathbf{z}$ is irrelevant compared to the question $\mathbf{q}$, then information gain for that turn is bounded by* $\log(1 + \alpha \lambda_{\max}(\mathbf{J}_0^{-1}) B^2) = \mathcal{O}(\varepsilon)$.

Lemma 2 (Redundancy leads to diminishing returns)**.** *Consider repeatedly adding the same evidence vector z for k times with the same α. Then, $\Delta(\mathbf{z}; \mathbf{J}_k) \geq \Delta(\mathbf{z}; \mathbf{J}_{k+1})$*

Together, these lemmas show that weakly relevant evidence contributes little, while repeated evidence has decreasing marginal value. In Appendix G, we verify these behaviors in controlled synthetic dialogues, including cases where novel information is reintroduced after redundant or irrelevant turns.

Implication. These local properties are well-suited for semantic-progress evaluation: irrelevant content yields negligible gain, while repeated or paraphrased content is discounted even when on-topic. Consequently, high scores require evidence that is both question-relevant and genuinely informative relative to the dialogue history. Combined with the global properties, the estimator rewards sustained semantic progress, favoring evidence that continues to reduce uncertainty beyond what the dialogue has already established.

5 Experiments

5.1 Main Results

We evaluate Gaussian information gain as a signal for semantic progress in information-seeking dialogue. Our experiments first ask whether the score aligns with preference judgments and whether it offers operational speed advantages over LLM-as-a-judge evaluation. We then study robustness across embedding backends and controlled cases where semantic relevance and factual correctness diverge in Section 5.2. Additional experimental details are provided in Appendices H and J.

Setup. We evaluate on MT-Bench, Chatbot Arena [32], and UltraFeedback [33]. MT-Bench and Chatbot Arena provide paired dialogue comparisons with human preference votes. UltraFeedback provides scored responses across several quality dimensions. Following prior work, we evaluate only examples with non-tied preference labels. We do not filter by category, which makes the evaluation conservative with respect to our stated scope: GAUSSIANIG measures semantic progress rather than holistic quality, and is expected to be most meaningful for information-seeking interactions.

For each dialogue $D_{1:T}$, we compute the dialogue-level score $\widehat{\mathcal{P}}_q(D_{1:T}) = \sum_{t=1}^{T} \widehat{\mathrm{IG}}_t$. Given a pair $(D^A_{1:T}, D^B_{1:T})$, we predict the preferred dialogue by computing $\arg\max_{X \in \{A,B\}} \widehat{\mathcal{P}}_q(D^X_{1:T})$. We report agreement with the majority preference label and Kendall's τ. Unless otherwise stated, we use `Qwen3-Embedding-0.6B`; Section 5.2 studies other embedding backends and model sizes.

Table 1: Agreement and runtime comparison on MT-Bench, Chatbot Arena, and UltraFeedback. Agr. denotes agreement with the preference label, τ denotes Kendall's rank correlation, and time denotes end-to-end wall-clock seconds for $N = 100$ dialogue pairs averaged over $R = 5$ runs. LLM-as-a-judge baselines were executed via AWS Bedrock using temperature 0.

Method	**MT-Bench**			**Chatbot Arena**			**UltraFeedback**		
	Agr. ↑	τ ↑	Time ↓	Agr. ↑	τ ↑	Time ↓	Agr. ↑	τ ↑	Time ↓
Mistral Large 3 [26]	76.50	0.54	132	60.19	0.20	119	70.75	0.42	153
DeepSeek R1 [27]	80.96	0.61	633	66.14	0.32	480	71.43	0.43	1173
GPT OSS 120b [28]	72.27	0.44	568	65.27	0.31	328	70.75	0.42	408
Claude Sonnet 3.7 [29]	76.47	0.53	260	64.74	0.29	251	70.52	0.40	285
Claude Sonnet 4 [30]	81.93	0.64	220	65.58	0.31	175	69.84	0.38	535
Claude Sonnet 4.5 [31]	76.05	0.52	342	**66.42**	**0.33**	286	69.16	0.38	623
Ours (GAUSSIANIG)	**84.03**	**0.68**	**86**	65.80	0.32	**44**	**71.64**	**0.43**	**104**

We compare against LLM-as-a-judge baselines, which recent surveys identify as among the strongest automatic dialogue evaluators [8, 25]. These baselines are holistic evaluators, while our metric targets a narrower quantity: agent-induced semantic progress through the accumulation of relevant, non-redundant information. Runtime is measured on a fixed $N = 100$ dialogue-pair subset sampled from the same set with fixed ordering and averaged over $R = 5$ runs. Our method performs local embedding passes and closed-form Gaussian updates, while judge baselines call hosted inference APIs, so runtimes reflect operational deployment overhead such as network latency and service execution time.

Results. Table 1 reports agreement and runtime on three preference benchmarks. Since the LLM-as-a-judge baselines evaluate broad dialogue quality, while GAUSSIANIG measures only agent-induced semantic progress, the comparison is not intended as a holistic evaluator leaderboard. Instead, it tests whether semantic progress explains a substantial fraction of human preference judgments, including in mixed settings where creativity, tone, concision, safety, or style may also drive preferences. On MT-Bench, GAUSSIANIG obtains 84.03% agreement, comparable to or higher than the evaluated LLM-judge baselines; on UltraFeedback, it also achieves strong agreement, while on Chatbot Arena it remains competitive. These results support GAUSSIANIG as a complementary metric for relevance-weighted, non-redundant information gain. The dimension-level UltraFeedback analysis further shows weaker alignment with truthfulness and honesty than with helpfulness, reinforcing that factuality and honesty remain outside the scope of the metric.

The runtime results show the main operational advantage. GAUSSIANIG is substantially faster than hosted LLM judges across benchmarks: 86s versus 132–633s on MT-Bench, 44s versus 119–480s on Chatbot Arena, and 104s versus 153–1173s on UltraFeedback for $N = 100$ dialogue pairs. This efficiency comes from replacing autoregressive judge inference with embedding computation and closed-form rank-one covariance updates.

5.2 Ablations and Robustness

Embedding Model. Table 2 studies sensitivity to the embedding backend. The evaluated models span over three orders of magnitude in parameter count, from a ∼2M-parameter CPU model to a 4B-parameter embedding model.

Performance is stable across a wide range of embedding models. On MT-Bench, even the ∼2M-parameter CPU model achieves 80.25% agreement, while several small CPU models approach or exceed the default `Qwen3-Embedding-0.6B` configuration. Larger models provide only modest improvements at substantially higher computational cost. This suggests that the uncertainty-reduction signal saturates once the embedding space captures the dominant semantic directions, after which additional model capacity yields limited benefit for the Gaussian update. Appendix I discusses this saturation effect in more detail.

Cosine-similarity Only. To isolate the contribution of the information-gain mechanism, we compare against a cosine-only ablation that sums per-turn cosine similarities between responses and the dialogue context, removing uncertainty accumulation and diminishing returns. On clear-majority

Table 2: Embedding model ablation across benchmarks of our method GAUSSIANIG. Here, Agr. denotes pairwise agreement. Device indicates whether embeddings were computed on CPU or GPU/MPS. Runtimes are end-to-end wall-clock seconds for $N = 100$ dialogue pairs, averaged over $R = 5$ runs.

Model	**Device**	**MT-Bench**		**Chatbot Arena**		**UltraFeedback**	
		Agr. ↑	Time ↓	Agr. ↑	Time ↓	Agr. ↑	Time ↓
Potion-2M [34]	CPU	80.25	7.1	66.05	0.4	71.20	0.8
MiniLM-23M [35]	CPU	83.61	17.9	66.27	6.77	68.93	21.5
Arctic-32M [36]	CPU	85.29	29.3	65.73	13.4	69.84	37.1
ModernBERT-0.1B [37]	GPU	83.19	28.5	65.75	15.5	68.03	57.0
MicroLlama-0.3B [38]	GPU	84.45	47.4	66.17	23.7	67.80	73.2
Qwen3-0.6B [39]	GPU	84.03	86.4	65.80	44.1	71.64	104.4
Qwen3-4B [39]	GPU	84.83	396.2	66.13	246.3	70.07	642.0

MT-Bench pairs, this baseline achieves only 44.5% agreement, below both a majority baseline (51.7%) and random guessing $(50.1\% \pm 3.2)$, and far below GAUSSIANIG (84.0%). This indicates that the gains are not explained by embedding similarity alone: the history-dependent covariance update and redundancy discounting are essential.

Length and Padding Robustness. To test whether $\widehat{\mathcal{P}}$ rewards length rather than semantic progress, we pad UltraFeedback responses with irrelevant, redundant, or novel relevant content. At $r = 2.0$, irrelevant and redundant padding have small effects (0.032 ± 0.010 and 0.111 ± 0.016), while novel relevant padding substantially increases the score (5.697 ± 0.408). This suggests $\widehat{\mathcal{P}}_q$ is sensitive to question-relevant evidence rather than length alone. See Appendix J.1 for details.

Hallucinations and Relevance Parameters. We evaluate whether the relevance weighting is sensitive to factual perturbations that preserve surface form and semantic uncertainty. We vary the relevance cutoff η and exponent β. With mild relevance weighting $\beta = 1, \eta = 0.05$ the metric already prefers the non-corrupted response above chance. Increasing β generally improves sensitivity to the factual perturbations when the cutoff η is not too aggressive. Several settings with $\beta \in \{3, 4, 5, 6\}$ achieve win rates between 0.85 and 0.93. However, performance drops sharply for large η, especially at high β, because the filter removes too much evidence. We provide more details in Appendix J.2.

6 Conclusion

This work develops an information-theoretic approach to multi-turn dialogue evaluation, framing semantic progress in information-seeking interactions as question-conditioned uncertainty reduction. Our Gaussian embedding-space estimator is reference-free, deterministic, and efficient, while preserving useful structural properties such as compositionality across turns and diminishing returns for redundant evidence. Across MT-Bench, Chatbot Arena, and UltraFeedback, it achieves meaningful agreement with human preferences while reducing the cost of evaluation relative to LLM-as-a-judge baselines. These results suggest that embedding models can support richer dialogue-level evaluation beyond static similarity scoring. When combined with relevance weighting, uncertainty updates, and redundancy discounting, even lightweight embeddings can yield useful dialogue-level evaluation signals. This makes the approach well suited for scalable model comparison, regression testing, and deployment monitoring. At the same time, semantic progress is only one dimension of dialogue quality; the metric does not directly assess factuality, safety, coherence, style, or user satisfaction, and depends on the chosen embedding space and evidence extraction procedure. Future work should combine this signal with complementary evaluators and study its applicability beyond information-seeking dialogue.

References

[1] Siheng Li, Cheng Yang, Yichun Yin, Xinyu Zhu, Zesen Cheng, Lifeng Shang, Xin Jiang, Qun Liu, and Yujiu Yang. AutoConv: Automatically generating information-seeking conversations with large language models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki

Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1751–1762, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-short.149. URL https://aclanthology.org/2023.acl-short.149/.

[2] Fanghua Ye, Meng Fang, Shenghui Li, and Emine Yilmaz. Enhancing conversational search: Large language model-aided informative query rewriting. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5985–6006, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.398. URL https://aclanthology.org/2023.findings-emnlp.398/.

[3] Chuangtao Ma, Yongrui Chen, Tianxing Wu, Arijit Khan, and Haofen Wang. Large language models meet knowledge graphs for question answering: Synthesis and opportunities. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24589–24608, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1249. URL https://aclanthology.org/2025.emnlp-main.1249/.

[4] Ehsan Kamalloo, Nouha Dziri, Charles Clarke, and Davood Rafiei. Evaluating open-domain question answering in the era of large language models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5591–5606, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.307. URL https://aclanthology.org/2023.acl-long.307/.

[5] Zeqiu Wu, Ryu Parish, Hao Cheng, Sewon Min, Prithviraj Ammanabrolu, Mari Ostendorf, and Hannaneh Hajishirzi. InSCIt: Information-seeking conversations with mixed-initiative interactions. *Transactions of the Association for Computational Linguistics*, 11:453–468, 2023. doi: 10.1162/tacl_a_00559. URL https://aclanthology.org/2023.tacl-1.27/.

[6] 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. MultiChallenge: A realistic multi-turn conversation evaluation benchmark challenging to frontier LLMs. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 18632–18702, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.958. URL https://aclanthology.org/2025.findings-acl.958/.

[7] Takyoung Kim, Hoonsang Yoon, Yukyung Lee, Pilsung Kang, and Misuk Kim. Mismatch between multi-turn dialogue and its evaluation metric in dialogue state tracking. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 297–309, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-short.33. URL https://aclanthology.org/2022.acl-short.33/.

[8] 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. From generation to judgment: Opportunities and challenges of LLM-as-a-judge. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2757–2791, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.138. URL https://aclanthology.org/2025.emnlp-main.138/.

[9] Shengyue Guan, Haoyi Xiong, Jindong Wang, Jiang Bian, Bin Zhu, and Jian guang Lou. Evaluating llm-based agents for multi-turn conversations: A survey, 2025. URL https://arxiv.org/abs/2503.22458.

[10] Sarah E. Finch, James D. Finch, and Jinho D. Choi. Don't forget your ABC's: Evaluating the state-of-the-art in chat-oriented dialogue systems. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15044–15071, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.839. URL https://aclanthology.org/2023.acl-long.839/.

[11] Zhaochen Hong, Haofei Yu, and Jiaxuan You. ConsistencyChecker: Tree-based evaluation of LLM generalization capabilities. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 33039–33075, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1585. URL https://aclanthology.org/2025.acl-long.1585/.

[12] Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. Llms-as-judges: A comprehensive survey on llm-based evaluation methods, 2024. URL https://arxiv.org/abs/2412.05579.

[13] Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1547. URL https://aclanthology.org/D18-1547/.

[14] Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains, 2024. URL https://arxiv.org/abs/2406.12045.

[15] Emre Can Acikgoz, Carl Guo, Suvodip Dey, Akul Datta, Takyoung Kim, Gokhan Tur, and Dilek Hakkani-Tur. TD-EVAL: Revisiting task-oriented dialogue evaluation by combining turn-level precision with dialogue-level comparisons. In Frédéric Béchet, Fabrice Lefèvre, Nicholas Asher, Seokhwan Kim, and Teva Merlin, editors, *Proceedings of the 26th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 113–132, Avignon, France, August 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.sigdial-1.7/.

[16] Suvodip Dey, Ramamohan Kummara, and Maunendra Desarkar. Towards fair evaluation of dialogue state tracking by flexible incorporation of turn-level performances. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 318–324, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-short.35. URL https://aclanthology.org/2022.acl-short.35/.

[17] Sarik Ghazarian, Behnam Hedayatnia, Alexandros Papangelis, Yang Liu, and Dilek Hakkani-Tur. What is wrong with you?: Leveraging user sentiment for automatic dialog evaluation. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Findings of the Association for Computational Linguistics: ACL 2022*, pages 4194–4204, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.331. URL https://aclanthology.org/2022.findings-acl.331/.

[18] Bao Chen, Yuanjie Wang, Zeming Liu, and Yuhang Guo. Automatic evaluate dialogue appropriateness by using dialogue act. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7361–7372, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.492. URL https://aclanthology.org/2023.findings-emnlp.492/.

[19] Ge Bai, Jie Liu, Xingyuan Bu, Yancheng He, Jiaheng Liu, Zhanhui Zhou, Zhuoran Lin, Wenbo Su, Tiezheng Ge, Bo Zheng, and Wanli Ouyang. 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, 2024. doi: 10.18653/v1/2024.acl-long.401. URL http://dx.doi.org/10.18653/v1/2024.acl-long.401.

[20] Jinghan Jia, Abi Komma, Timothy Leffel, Xujun Peng, Ajay Nagesh, Tamer Soliman, Aram Galstyan, and Anoop Kumar. Leveraging LLMs for dialogue quality measurement. In Yi Yang, Aida Davani, Avi Sil, and Anoop Kumar, editors, *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, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-industry.30. URL https://aclanthology.org/2024.naacl-industry.30/.

[21] Jiao Ou, Junda Lu, Che Liu, Yihong Tang, Fuzheng Zhang, Di Zhang, and Kun Gai. DialogBench: Evaluating LLMs as human-like dialogue systems. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *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, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.341. URL https://aclanthology.org/2024.naacl-long.341/.

[22] ChaeHun Park, Minseok Choi, Dohyun Lee, and Jaegul Choo. Paireval: Open-domain dialogue evaluation metric with pairwise comparisons. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=y6aGT625Lk.

[23] Lexin Zhou, Youmna Farag, and Andreas Vlachos. An LLM feature-based framework for dialogue constructiveness assessment. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5389–5409, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.308. URL https://aclanthology.org/2024.emnlp-main.308/.

[24] Qintong Li, Leyang Cui, Lingpeng Kong, and Wei Bi. Exploring the reliability of large language models as customized evaluators for diverse NLP tasks. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert, editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10325–10344, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.688/.

[25] Chen Zhang, Luis Fernando D'Haro, Yiming Chen, Malu Zhang, and Haizhou Li. 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, 2024. ISBN 978-1-57735-887-9. doi: 10.1609/aaai.v38i17.29923. URL https://doi.org/10.1609/aaai.v38i17.29923.

[26] Mistral AI. Introducing mistral 3. https://mistral.ai/news/mistral-3, 2025.

[27] 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, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun,

T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

[28] OpenAI. Introducing gpt-oss. https://openai.com/index/introducing-gpt-oss/, 2025.

[29] Anthropic. Claude 3.7 sonnet and claude code. https://www.anthropic.com/news/claude-3-7-sonnet, February 2025.

[30] Anthropic. Introducing claude 4. https://www.anthropic.com/news/claude-4, May 2025.

[31] Anthropic. Introducing claude sonnet 4.5. https://www.anthropic.com/news/claude-sonnet-4-5, Sep 2025.

[32] 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. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL https://openreview.net/forum?id=uccHPGDlao.

[33] Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, Zhiyuan Liu, and Maosong Sun. Ultrafeedback: boosting language models with scaled ai feedback. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

[34] Stephan Tulkens and Thomas van Dongen. Model2vec: Fast state-of-the-art static embeddings, 2024. URL https://github.com/MinishLab/model2vec.

[35] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers, 2020. URL https://arxiv.org/abs/2002.10957.

[36] Luke Merrick, Danmei Xu, Gaurav Nuti, and Daniel Campos. Arctic-embed: Scalable, efficient, and accurate text embedding models, 2024. URL https://arxiv.org/abs/2405.05374.

[37] Zach Nussbaum, John X. Morris, Brandon Duderstadt, and Andriy Mulyar. Nomic embed: Training a reproducible long context text embedder, 2024.

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

[39] 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. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025.

[40] T. Cover and J. Thomas. *Elements of Information Theory*. Wileys Series in Telecommunications, New York, 2 edition, 2006.

[41] OpenAI. Introducing GPT-5.4 mini and nano, March 2026. URL https://openai.com/index/introducing-gpt-5-4-mini-and-nano/.

Appendix

A Queries as One-dimensional Measurements

We model an embedding direction, such as a query or evidence embedding, as a one-dimensional (rank-one) measurement of the latent semantic state $\mathbf{S} \in \mathbb{R}^d$. A text span $x \in \mathcal{A}^*$ induces a direction defined by a vector $\mathbf{z}(x) \in \mathbb{R}^d$ in representation space. The corresponding response is a noisy linear measurement:

$$y_t = \mathbf{z}(x)^\top \mathbf{S} + \varepsilon \quad \varepsilon \sim \mathcal{N}(0, \sigma^2) \tag{13}$$

For intuition, the embedding direction $\mathbf{z}(x)$ specifies which direction of the latent semantic space $\mathbf{S}$ is probed. The scalar y reveals the component of $\mathbf{S}$ along that direction. This measurement admits an equivalent 1-dimensional operator view:

$$P_x \triangleq \mathbf{z}(x)\mathbf{z}(x)^\top \tag{14}$$

so that

$$P_x \mathbf{S} = \mathbf{z}(x)\Big(\mathbf{z}(x)^\top \mathbf{S}\Big) \tag{15}$$

Thus, although $P_x\mathbf{S}$ is a vector, it is determined solely based on the inner product $\mathbf{z}(x)^\top\mathbf{S}$, which is the quantity in Equation (13). In this sense, each question is a one-dimensional observation "projection like" measurement of $\mathbf{S}$.

If one wishes P_x to be an idempotent projection operator (i.e., $P_x^2 = P_x$) then $\mathbf{z}(x)$ must be normalized and the corresponding projector is:

$$\mathbf{\Pi}_x \triangleq \frac{\mathbf{z}(x)\mathbf{z}(x)^\top}{\|\mathbf{z}(x)\|^2} \tag{16}$$

In our setting however, we intentionally keep $\mathbf{z}(x)$ unnormalized (to preserve scale relevant for the bayesian update). In that case, we view Equation (14) as a scaled projector:

$$P_x \triangleq \alpha \mathbf{\Pi}_x \quad \alpha \triangleq \|\mathbf{z}(x)\|^2 \tag{17}$$

B Relation to Bayesian Linear Regression

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

$$y_i = \mathbf{w}^\top \mathbf{s}_i + \varepsilon_i. \tag{18}$$

Assume heteroscedastic Gaussian noise $\varepsilon_i \sim \mathcal{N}(0, \sigma_i^2)$ and a Gaussian prior $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \sigma_p^2\mathbf{I})$. Let $\mathbf{S} \in \mathbb{R}^{n\times d}$ be the design matrix whose rows are $\mathbf{s}_i^\top$, and let $\mathbf{y} \in \mathbb{R}^n$ collect the outputs y_i. Using Bayes' rule and independence of the observations, the log posterior is, up to an additive constant,

$$\log p(\mathbf{w} \mid \mathbf{S}, \mathbf{y}) = -\frac{1}{2}\left[\sigma_p^{-2}\|\mathbf{w}\|_2^2 + \sum_{i=1}^{n} \sigma_i^{-2}(y_i - \mathbf{w}^\top \mathbf{s}_i)^2\right] + c. \tag{19}$$

Let

$$\mathbf{R} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_n^2).$$

Then the same expression can be written as

$$\log p(\mathbf{w} \mid \mathbf{S}, \mathbf{y}) = -\frac{1}{2}\left[\sigma_p^{-2}\mathbf{w}^\top\mathbf{w} + (\mathbf{y} - \mathbf{S}\mathbf{w})^\top \mathbf{R}^{-1}(\mathbf{y} - \mathbf{S}\mathbf{w})\right] + c \tag{20}$$

$$= -\frac{1}{2}\left[\mathbf{w}^\top(\mathbf{S}^\top\mathbf{R}^{-1}\mathbf{S} + \sigma_p^{-2}\mathbf{I})\mathbf{w} - 2\mathbf{y}^\top\mathbf{R}^{-1}\mathbf{S}\mathbf{w}\right] + c. \tag{21}$$

Table 3: Correspondence between Bayesian linear regression and our weighted Gaussian information-gain formulation.

Bayesian Linear Regression	**Our Formulation**
Latent parameter $\mathbf{w}$	Latent semantic state $\mathbf{S}$
Input vector $\mathbf{s}_i$	Evidence embedding $\mathbf{z}_{t,i}$
Scalar observation y_i	Unobserved measurement value $y_{t,i}$
Noise variance σ_i^2	Effective evidence noise $\sigma^2/w_{t,i}$
Prior covariance $\sigma_p^2\mathbf{I}$	Initial covariance $\mathbf{\Sigma}_0$
Design matrix $\mathbf{S}$	Evidence embedding stack
Posterior covariance $\mathbf{\Sigma}$	Posterior semantic uncertainty
Precision update $\sigma_i^{-2}\mathbf{s}_i\mathbf{s}_i^\top$	Rank-one update $(w_{t,i}/\sigma^2)\mathbf{z}_{t,i}\mathbf{z}_{t,i}^\top$

Therefore the posterior covariance is

$$\mathbf{\Sigma} = \left(\mathbf{S}^\top\mathbf{R}^{-1}\mathbf{S} + \sigma_p^{-2}\mathbf{I}\right)^{-1}, \tag{22}$$

and hence the posterior precision is

$$\mathbf{\Sigma}^{-1} = \mathbf{S}^\top\mathbf{R}^{-1}\mathbf{S} + \sigma_p^{-2}\mathbf{I} = \sigma_p^{-2}\mathbf{I} + \sum_{i=1}^{n}\sigma_i^{-2}\mathbf{s}_i\mathbf{s}_i^\top. \tag{23}$$

In our formulation, the input vector $\mathbf{s}_i$ corresponds to an evidence embedding $\mathbf{z}_{t,i}$, and the prior precision $\sigma_p^{-2}\mathbf{I}$ corresponds to $\mathbf{J}_0 = \mathbf{\Sigma}_0^{-1}$. Setting

$$\sigma_i^2 = \frac{\sigma^2}{w_{t,i}} \quad \text{for } w_{t,i} > 0$$

gives

$$\mathbf{\Sigma}^{-1} = \mathbf{J}_0 + \sum_i \frac{w_{t,i}}{\sigma^2}\mathbf{z}_{t,i}\mathbf{z}_{t,i}^\top,$$

which is the weighted precision update in Equation (6). When $w_{t,i} = 0$, the corresponding observation has infinite effective noise and contributes no update.

C Proof of Theorem 1

Theorem. *For a fixed turn $t \in [T]$, let $\mathbf{Z}$ be the random evidence vector corresponding to $\mathbf{z} = \langle \mathbf{z}_1, \ldots, \mathbf{z}_m \rangle \in \mathbb{R}^{md}$ with evidence components $\mathbf{Z}_i \in \mathbb{R}^d$. If $h(\mathbf{Z} \mid \mathbf{S}) \geq c$, where c is independent of the dialogue strategy, and assume the evidence have common empirical covariance $\mathbf{\Sigma}_\mathbf{Z} \succ \mathbf{0}$, then:*

$$I(\mathbf{Z}; \mathbf{S}) \leq m\big[\frac{d}{2}\log(2\pi e) + \frac{1}{2}\log\det\mathbf{\Sigma}_\mathbf{Z}\big] - c \tag{24}$$

Proof. Let $t \in [T]$ be a fixed turn and $\mathbf{z} = \langle \mathbf{z}_1, \ldots, \mathbf{z}_m \rangle \in \mathbb{R}^{md}$ be one realization of a random vector $\mathbf{Z}$. We assume that a given strategy for generating a dialogue defines a conditional distribution $P(\mathbf{Z} \mid \mathbf{S})$. Together with a distribution $P(\mathbf{S})$ of semantic states, we thus obtain a joint distribution $P(\mathbf{Z}, \mathbf{S})$. The information $\mathbf{Z}$ reveals on average about $\mathbf{S}$ is given by the mutual information [40]:

$$I(\mathbf{Z}; \mathbf{S}) = h(\mathbf{Z}) - h(\mathbf{Z} \mid \mathbf{S}), \tag{25}$$

with the differential entropy h defined by $h(X) \triangleq -\int p(x)\log p(x)dx$, where $p(x)$ denotes a density some density in $\mathbb{R}^n$. For (discrete) Shannon entropy we have $H(\mathbf{Z} \mid \mathbf{S}) \geq 0$ so $I(\mathbf{Z}; \mathbf{S}) \leq H(\mathbf{Z})$. In the continuous case, entropy may be negative, so instead assume $h(\mathbf{Z} \mid \mathbf{S}) \geq c$ with c independent of the dialogue strategy. Then, $I(\mathbf{Z}; \mathbf{S}) \leq h(\mathbf{Z}) - c$. Now, discretize the evidence-embedding space into a finite alphabet $\mathcal{A}$ labeling cubic bins of size δ^d such that we can safely approximate $h(\mathbf{Z}) - h(\mathbf{Z} \mid \mathbf{S})$ with the corresponding discrete mutual information (which is at least guaranteed if the bins are small relative to the uncertainty of $P(\mathbf{Z} \mid \mathbf{S})$. The proof uses a source-coding perspective. Since entropy

characterizes the minimum achievable average code length, we can upper-bound the entropy of the evidence sequence by constructing an explicit code for it.

Let $\widehat{P}_{\mathbf{z}}$ be the empirical distribution of the observed sequence $\mathbf{z}$, to reconstruct the vector $\langle \mathbf{z}_1, \ldots, \mathbf{z}_m \rangle$ we can consider a two-part code.

First, encode the type $\widehat{P}_{\mathbf{z}}$. Since a type is determined by $|\mathcal{A}|$ integer counts summing to m, the number of possible types is at most $(m+1)^{|\mathcal{A}|}$ so encoding the type costs $|\mathcal{A}| \log(m+1)$ bits.

Second, given the type $\widehat{P}_{\mathbf{z}}$, encode which ordered sequence within the type class occurred. The type class has cardinality:

$$|T(\widehat{P}_{\mathbf{z}})| = \frac{m!}{\prod_{a \in \mathcal{A}} (m \cdot \widehat{P}_{\mathbf{z}}(a))!} \tag{26}$$

which is bounded by $|T(\widehat{P}_{\mathbf{z}})| \leq 2^{mH(\widehat{P}_{\mathbf{z}})}$ following Theorem 3.1.2 in [40] with $\epsilon = 0$. Therefore, identifying the ordered realization within the type class requires at most $m \cdot H(\widehat{P}_{\mathbf{z}})$ bits. Hence, $\mathbf{z}$ can be described in $m \cdot H(\widehat{P}_{\mathbf{z}}) + |\mathcal{A}| \log(m+1)$ many bits. Due to Shannon's source coding theorem (Theorem 5.4.1 in [40]), the entropy cannot be larger than the number of bits needed on the average, we thus have:

$$H(\mathbf{Z}) = H(\mathbf{Z}_1, \ldots, \mathbf{Z}_m) \leq m \cdot \mathbb{E}[H(\widehat{P}_{\mathbf{z}})] + |\mathcal{A}| \log(m+1) \tag{27}$$

Next, suppose we retain only the empirical second moment matrix

$$\widehat{\mathbf{\Sigma}}_{\mathbf{z}} \triangleq \frac{1}{m} \sum_{i=1}^{m} \mathbf{z}_i \mathbf{z}_i^\top \tag{28}$$

We now interpret the discrete distribution as density $\widehat{p}_{\mathbf{z}}$ in $\mathbb{R}^d$ that is constant within one bin. Its differential entropy $h(\widehat{p}_{\mathbf{z}})$ cannot be larger than the entropy of a Gaussian with zero mean and covariance $\widehat{\mathbf{\Sigma}}_{\mathbf{z}}$, see [40], Theorem 8.6.5. Thus, $h(\widehat{P}_{\mathbf{z}}) \leq \frac{d}{2} \log(2\pi e) + \frac{1}{2} \log\det \widehat{\mathbf{\Sigma}}_{\mathbf{z}}$. The discrete Shannon entropy $H(\widehat{P}_{\mathbf{z}})$ is thus bounded from above by $\frac{d}{2} \log(2\pi e) + \frac{1}{2} \log\det \widehat{\mathbf{\Sigma}}_{\mathbf{z}} + d \log \delta$, where the extra summand comes from multiplying the density with δ^d to get the probability mass function.

To obtain a lower bound on the discretized version of $P(\mathbf{Z} \mid \mathbf{S})$ we observe that averaging the density over bins can only increase entropy and thus $H(\mathbf{Z} \mid \mathbf{S}) \geq c + dm \cdot \log \delta$ (with a similar extra term from dm-dimensional cubes). Obviously, these extra terms cancel for the mutual information.

Therefore, for a fixed turn, the expression $m\big[\frac{d}{2} \log(2\pi e) + \frac{1}{2} \log\det \widehat{\mathbf{\Sigma}}_{\mathbf{z}}\big]$ is the empirical Gaussian maximum-entropy surrogate for the entropy of the evidence vector, up to the type-encoding term $|\mathcal{A}| \log(m+1)$, and hence up to the conditional uncertainty term, and this lower-order coding cost, it gives an upper-bound surrogate for the information the turn's evidence can reveal about the semantic state. □

Remark: When $|\mathcal{A}|$ is fixed, the normalized error term $|\mathcal{A}| \log(m+1)/m$ vanishes as $m \to \infty$. Thus, the Gaussian log-determinant quantity is a large-m upper-bound surrogate.

D Proof of Theorem 2

Theorem (Monotonicity and Telescoping)**.** *Assume $\mathbf{\Sigma}_0 \succ \mathbf{0}$ and weights $\alpha_{t,i} \geq 0$. Under the precision update in Equation* (6)*, we have $\mathbf{\Sigma}_t \preceq \mathbf{\Sigma}_{t-1}$ (in PSD order) and hence $\widehat{\mathrm{IG}}_t \geq 0$ for all t. Moreover, the total gain telescopes:*

$$\sum_{t=1}^{T} \widehat{\mathrm{IG}}_t = \frac{1}{2} \log \frac{\det(\mathbf{\Sigma}_0)}{\det(\mathbf{\Sigma}_T)}, \tag{29}$$

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

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

$$\mathbf{J}_t = \mathbf{J}_{t-1} + \sum_{i=1}^{m_t} \frac{w_{t,i}}{\sigma^2} \mathbf{z}_{t,i} \mathbf{z}_{t,i}^\top. \tag{30}$$

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

$$\mathbf{A}_t \triangleq \sum_{i=1}^{m_t} \frac{w_{t,i}}{\sigma^2} \mathbf{z}_{t,i} \mathbf{z}_{t,i}^\top \tag{31}$$

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

$$\mathbf{J}_t = \mathbf{J}_{t-1} + \mathbf{A}_t \tag{32}$$

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 $\mathbf{J}_t^{-1} \preceq \mathbf{J}_{t-1}^{-1}$ meaning $\mathbf{\Sigma}_t \preceq \mathbf{\Sigma}_{t-1}$. Thus, posterior covariance is monotone non-increasing in PSD order.

Next, recall

$$\widehat{\mathrm{IG}}_t = \frac{1}{2} \log \frac{\det(\mathbf{\Sigma}_{t-1})}{\det(\mathbf{\Sigma}_t)}. \tag{33}$$

Since, $\mathbf{\Sigma}_t \preceq \mathbf{\Sigma}_{t-1}$, determinant monotonicity on the positive-definite cone gives $\det(\mathbf{\Sigma}_t) \leq \det(\mathbf{\Sigma}_{t-1})$, and hence $\widehat{\mathrm{IG}}_t \geq 0$. Finally, telescoping follows alternating sums

$$\sum_t \widehat{\mathrm{IG}}_t = \frac{1}{2} \sum_t \log\det(\mathbf{\Sigma}_{t-1}) - \log\det(\mathbf{\Sigma}_t) \tag{34}$$

$$= \frac{1}{2} \log\det(\mathbf{\Sigma}_0) - \log\det(\mathbf{\Sigma}_T) \tag{35}$$

$$= \frac{1}{2} \log \frac{\det \mathbf{\Sigma}_0}{\det \mathbf{\Sigma}_T} \tag{36}$$

This proves both monotonicity and telescoping. □

E Proof of Theorem 3

Theorem (Diminishing Returns)**.** *Assume $\mathbf{J}_0 \succ 0$. The set function F is monotone submodular. That is, for all $A \subseteq B \subseteq \mathcal{X}$and $x \notin B$,*

$$F(A \cup \{x\}) - F(A) \geq F(B \cup \{x\}) - F(B).$$

Let

$$\mathbf{J}(\mathcal{S}) \;=\; \mathbf{J}_0 \;+\; \sum_{i \in \mathcal{S}} \alpha_i \, \mathbf{z}_i \mathbf{z}_i^\top, \quad \mathbf{J}_0 \succ 0, \alpha_i \geq 0 \tag{37}$$

and

$$F(\mathcal{S}) \triangleq \log\det \mathbf{J}(\mathcal{S}). \tag{38}$$

We first show F is monotone:

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

$$\mathbf{J}(\mathcal{S} \cup \{e\}) = \mathbf{J}(\mathcal{S}) + \alpha_e \, \mathbf{z}_e \mathbf{z}_e^\top \succeq \mathbf{J}(\mathcal{S}) \tag{39}$$

Since $\log\det(\cdot)$ is increasing over the positive-definite 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:

$$\Delta_e(\mathcal{S}) \triangleq F(\mathcal{S} \cup \{e\}) - F(\mathcal{S}) \tag{40}$$

$$= \log \frac{\det(\mathbf{J}(\mathcal{S}) + \alpha_e \, \mathbf{z}_e \mathbf{z}_e^\top)}{\det(\mathbf{J}(\mathcal{S}))} \tag{41}$$

using the matrix determinant lemma

$$\det(\mathbf{J}(\mathcal{S}) + \alpha_e \, \mathbf{z}_e \mathbf{z}_e^\top) = \det(\mathbf{J}(\mathcal{S}))(1 + \alpha_e \, \mathbf{z}_e^\top \mathbf{J}(\mathcal{S})^{-1} \mathbf{z}_e) \tag{42}$$

Therefore

$$\Delta_e(\mathcal{S}) = \log(1 + \alpha_e \mathbf{z}_e^\top \mathbf{J}(\mathcal{S})^{-1} \mathbf{z}_e) \tag{43}$$

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:

$$\alpha_e \mathbf{z}_e^\top \mathbf{J}(\mathcal{A})^{-1} \mathbf{z}_e \geq \alpha_e \mathbf{z}_e^\top \mathbf{J}(\mathcal{B})^{-1} \mathbf{z}_e \tag{44}$$

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:

$$F(\mathcal{A} \cup \{e\}) - F(\mathcal{A}) \;\geq\; F(\mathcal{B} \cup \{e\}) - F(\mathcal{B}). \tag{45}$$

□

F 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 $\mathbf{M} \succeq \mathbf{0}$ we have $\mathbf{z}^\top \mathbf{M} \mathbf{z} \leq \lambda_{\max}(\mathbf{M}) \|\mathbf{z}\|_2^2$.

F.1 Proof of Lemma 1

Lemma (Soft Irrelevance)**.** *Assume embeddings are norm bounded $\|\mathbf{z}\| \leq B$, and $\alpha \leq \varepsilon$. When $\mathbf{z}$ is irrelevant compared to the question $\mathbf{q}$, then information gain for that turn is bounded by $\log(1 + \alpha \lambda_{\max}(\mathbf{J}_0^{-1}) B^2) = \mathcal{O}(\varepsilon)$.*

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

$$\Delta(\mathbf{z}; \mathbf{J}) = \log\big(1 + \alpha \, \mathbf{z}^\top \mathbf{J}^{-1} \mathbf{z}\big) \, . \tag{46}$$

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

$$\mathbf{z}^\top \mathbf{J}^{-1} \mathbf{z} \leq \mathbf{z}^\top \mathbf{J}_0^{-1} \mathbf{z} \tag{47}$$

Since $\mathbf{J}_0$ is a positive-definite precision matrix, $\mathbf{J}_\mathbf{0}^{-\mathbf{1}}$ is a symmetric positive-definite covariance matrix. Applying the Rayleigh quotient bound, we get

$$\mathbf{z}^\top \mathbf{J}_0^{-1} \mathbf{z} \leq \lambda_{\max}(\mathbf{J}_0^{-1}) \|\mathbf{z}\|_2^2 \leq \lambda_{\max}(\mathbf{J}_0^{-1}) B^2 \tag{48}$$

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

$$\Delta(\mathbf{z}; \mathbf{J}) = \log\big(1 + \alpha \, \mathbf{z}^\top \mathbf{J}^{-1} \mathbf{z}\big) \tag{49}$$

$$\leq \log(1 + \alpha \lambda_{\max}(\mathbf{J}_0^{-1}) B^2) \tag{50}$$

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

F.2 Proof of Lemma 2

Lemma (Redundancy leads to diminishing returns)**.** *Consider repeatedly adding the same $\mathbf{z}$ k times with the same α. Then, $\Delta(\mathbf{z}; \mathbf{J}_k) \geq \Delta(\mathbf{z}; \mathbf{J}_{k+1})$*

Proof. Let $\mathbf{z} \in \mathbb{R}^d$, $\alpha \geq 0$, 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

$$\mathbf{J}_{k+1} = \mathbf{J}_k + \alpha \, \mathbf{z}\mathbf{z}^\top \succeq \mathbf{J}_k \succ \mathbf{0}. \tag{51}$$

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

$$\mathbf{z}^\top \mathbf{J}_{k+1}^{-1} \mathbf{z} \leq \mathbf{z}^\top \mathbf{J}_k^{-1} \mathbf{z} \tag{52}$$

Because $x \mapsto \log(1 + \alpha x)$ is nondecreasing for $\alpha \geq 0$, we conclude

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

$$\leq \log\big(1 + \alpha \, \mathbf{z}^\top \mathbf{J}_k^{-1} \mathbf{z}\big) = \Delta(\mathbf{z}; \mathbf{J}_k). \tag{54}$$

□

G 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):

$$p_0(u) = \frac{1}{N}, \quad u \in U, \tag{55}$$

$$\mathrm{H}_0 = -\sum_{u \in U} p_0(u) \log p_0(u) = \log N. \tag{56}$$

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:

$$S_0 = U, \tag{57}$$

$$S_t = \Big\{ u \in S_{t-1} \;\Big|\; u \text{ is consistent with } e_t \Big\}. \tag{58}$$

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

$$\mathrm{H}_t = \log |S_t|. \tag{59}$$

The per-turn oracle information gain is then

$$\mathrm{IG}_t^{\text{oracle}} = \mathrm{H}_{t-1} - \mathrm{H}_t = \log \frac{|S_{t-1}|}{|S_t|} \;\geq\; 0, \tag{60}$$

and the cumulative oracle gain telescopes:

$$\sum_{t=1}^{T} \mathrm{IG}_t^{\text{oracle}} = \log \frac{|S_0|}{|S_T|} = \log \frac{N}{|S_T|}. \tag{61}$$

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

Comparison to GAUSSIANIG. For the same dialogues, we compute our Gaussian information gain using only question–answer embeddings and closed-form covariance updates (Equation (6)). 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 3, Figure 4, Figure 5, Figure 6 for results. Note that the user and LLM here are synthetic. We can see 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 demonstrates that the metric captures meaningful differences in long-horizon dialogue progress without relying on learned judges or supervision.

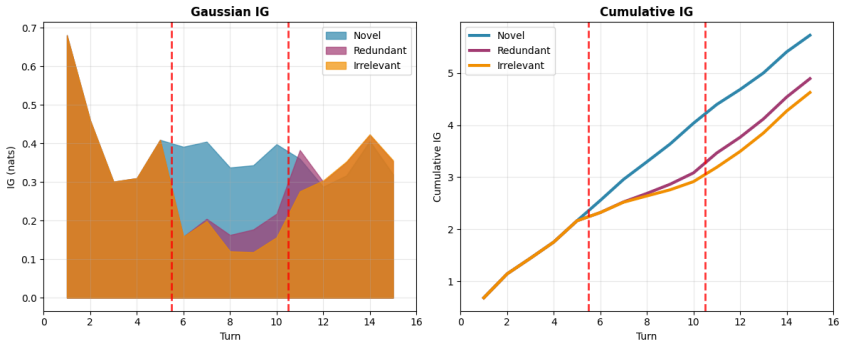


Figure 3: Information gain distinguishes dialogue quality in a 15-turn synthetic setting. We compare dialogues that introduce novel, redundant, or irrelevant information. (Left) Per-turn Gaussian information gain: redundant and irrelevant turns yield lower marginal gain than novel turns. (Right) Cumulative information gain: dialogues with a higher proportion of novel information achieve consistently higher total gain despite approximation noise.

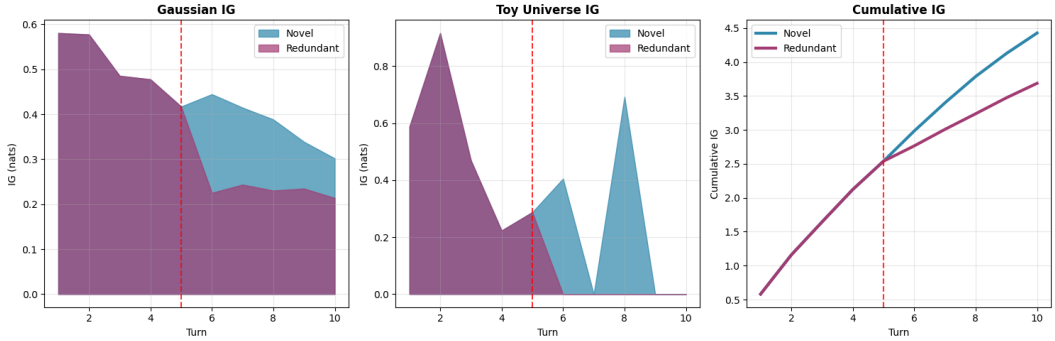


Figure 4: Toy example with the city guessing universe. The LLM tries to guess a city by asking questions and the user provides Yes/No to the questions. We compare the case where at turn 5, the LLM starts asking redundant questions to if it kept asking novel questions, in an ideal universe we know $S_t = S_{t-1}$ 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.

H Experiment Details

Algorithm and Parameters. Algorithm 1 presents the estimator using an explicit evidence-level loop for readability. In implementation, the per-turn rank-one updates can be vectorized as $\mathbf{J}' = \mathbf{J} + \sigma^{-2}\mathbf{Z}_t^\top \operatorname{diag}(\mathbf{w}_t)\mathbf{Z}_t$, where rows of $\mathbf{Z}_t$ are evidence embeddings. In Table 4, we report the parameters we used.

Existing assets and licenses. We use existing public benchmarks and models only for research evaluation. MT-Bench and Chatbot Arena are used following their public release terms [32]; UltraFeedback is used following its public release terms [33]. Embedding models and LLM baselines are credited through their original papers or model cards. We do not redistribute any existing datasets, model checkpoints, or generated annotations.

LLM Scoring. We use pointwise scalar judging without explanations for all LLM baselines to standardize API outputs and runtime measurement. Pairwise prompts may improve these baselines, especially for reasoning-specialized models. Thus, these baselines should be interpreted as operational judge configurations rather than the best possible performance of each underlying LLM. Since the evaluation set excludes human-tie cases, predicted ties are counted as incorrect: when the two candidate dialogues receive the same 1–10 score, the judge fails to identify the preferred dialogue.

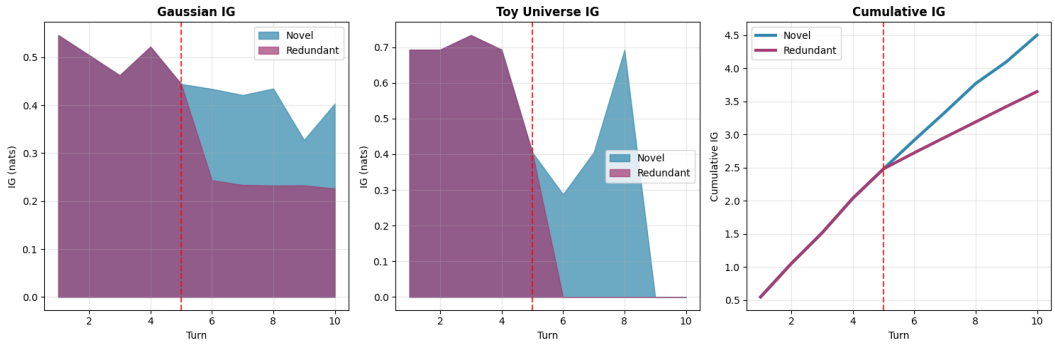


Figure 5: Toy example with the movie guessing universe. The LLM tries to guess a movie by asking questions and the user provides Yes/No to the questions. We compare the case where at turn 5, the LLM starts asking redundant questions to if it kept asking novel questions. The trends observed are similar to Figure 4.

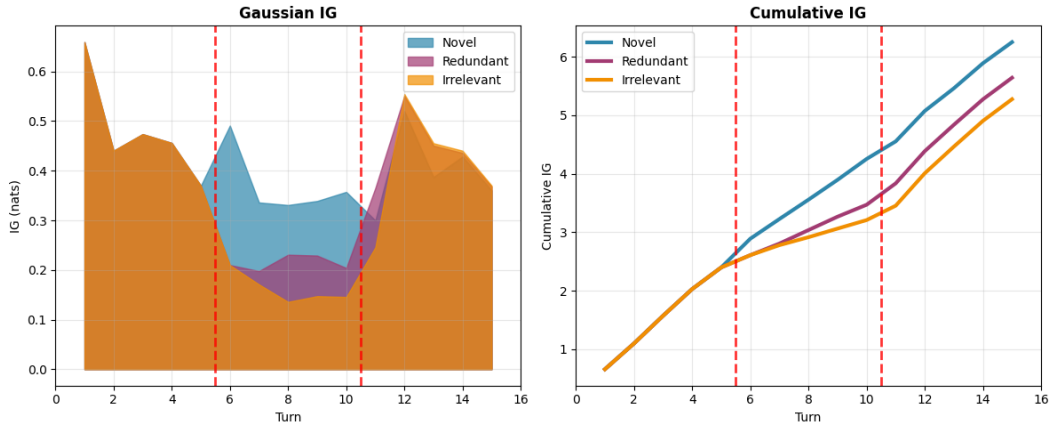


Figure 6: Toy example where the user asks open questions for advice related to gardening. The LLM tries to answer the question. We compare the cases where between turns 5 and 10, where LLM answers are either novel, redundant, or irrelevant. We observe similar trends as Figure 4, Figure 5, and also that information gain recovers after the LLM starts asking novel questions again.

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 5.

Judge Prompt:

> You are an expert evaluator of LLM responses.
>
> Evaluate the quality of the following response with respect to the dimension: `{DIMENSION_DEFINITION}`.
>
> Conversation:
>
> `{CONVERSATION}`

Algorithm 1 GAUSSIANIG: Gaussian Information Gain Evaluation

Input: Dialogue turns $(q_t, a_t)_{t=1}^{T}$, encoder enc, parameters $\sigma_0, \sigma, \beta, \eta$
Output: Dialogue score $\widehat{\mathcal{P}}_q(D_{1:T})$

$\mathbf{J} \leftarrow \sigma_0^{-2}\mathbf{I}, \quad \widehat{\mathcal{P}} \leftarrow 0$ ▷ *Initialize prior precision*
for $t = 1, \ldots, T$ **do**
  $\{Z_{t,i}\}_{i=1}^{m_t} \leftarrow \mathrm{Split}(a_t)$ ▷ *Extract answer evidence*
  $\mathbf{q}_t \leftarrow \mathrm{enc}(q_t), \quad \widehat{\mathbf{q}}_t \leftarrow \mathbf{q}_t / \|\mathbf{q}_t\|_2$ ▷ *Embed current user need*
  for $i = 1, \ldots, m_t$ **do**
    $\mathbf{z}_{t,i} \leftarrow \mathrm{enc}(Z_{t,i}), \quad \widehat{\mathbf{z}}_{t,i} \leftarrow \mathbf{z}_{t,i} / \|\mathbf{z}_{t,i}\|_2$ ▷ *Embed evidence unit*
    $w_{t,i} \leftarrow \max(0, \langle \widehat{\mathbf{q}}_t, \widehat{\mathbf{z}}_{t,i} \rangle)^{\beta}$ ▷ *Question relevance*
    if $w_{t,i} < \eta$ **then** $w_{t,i} \leftarrow 0$ ▷ *Discard weak evidence*
  end for
  $\mathbf{J}' \leftarrow \mathbf{J} + \sum_{i=1}^{m_t} \frac{w_{t,i}}{\sigma^2} \mathbf{z}_{t,i} \mathbf{z}_{t,i}^{\top}$ ▷ *Precision update*
  $\widehat{\mathrm{IG}}_t \leftarrow \frac{1}{2} \left(\log\det \mathbf{J}' - \log\det \mathbf{J} \right)$ ▷ *Turn-level information gain*
  $\mathbf{J} \leftarrow \mathbf{J}', \quad \widehat{\mathcal{P}} \leftarrow \widehat{\mathcal{P}} + \widehat{\mathrm{IG}}_t$ ▷ *Accumulate progress*
end for
return $\widehat{\mathcal{P}}$

Table 4: Default experimental settings for Gaussian information-gain evaluation.

Setting	Value
Default encoder	`Qwen/Qwen3-Embedding-0.6B` via `SentenceTransformer`
Hardware	Apple M1 Pro, 10-core CPU, 16-core GPU, 16GB RAM
Device backend	MPS
Batch size	32
Encoder mode	Evaluation mode
Evidence segmentation	Deterministic sentence/bullet chunking
Minimum chunk length	12 characters
Embedding normalization	L2-normalized before relevance scoring
Relevance weight	$r_{t,i}^{\mathrm{rel}} = \max(0, \langle \widehat{\mathbf{q}}_t, \widehat{\mathbf{z}}_{t,i} \rangle)^{\beta}$
Relevance exponent	$\beta = 1.0$
Relevance cutoff	$\eta = 0.05$
Quality factor	$r_{t,i}^{\mathrm{qual}} = 1$
Projection dimension	$k = d$; no random projection
Prior variance	$\sigma_0^2 = 1.0$
Observation variance	$\sigma^2 = 0.25$
Log-determinant computation	`numpy.linalg.slogdet`
Numerical jitter	10^{-12}, only if log-det sign is nonpositive
Runtime measurement	End-to-end wall-clock time including embedding and Gaussian updates

> Rate the response on a scale from 1 to 10,
> Do not reward answers simply for being longer. Focus on quality relative to the specified dimension.
> Return ONLY a single integer from 1 to 10. Do not include any explanation or additional text.

The dimension definitions injected into the prompt are:

- **Helpfulness:** The response should directly and usefully address the user's request, provide relevant information, and avoid unnecessary or unhelpful content.
- **Instruction following:** The response should follow the user's explicit and implicit instructions, including requested format, scope, constraints, and task requirements.
- **Truthfulness:** The response should avoid factual errors, unsupported claims, and misleading statements.
- **Honesty:** The response should accurately represent uncertainty, limitations, and whether the requested information is known or inferable.

Table 5: 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.

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 (GAUSSIANIG)	86.4 ± 6.8	44.1 ± 5.3

I Why Larger Embedding Models Might Not Help

From Equation (11), 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

$$\widehat{\mathrm{IG}}(\mathbf{z}) = \frac{1}{2}\log(1 + \alpha \sum_i \lambda_i c_i^2), \tag{62}$$

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.

J Synthetic Experiment Details

J.1 Padding

Table 6: Length and padding robustness on UltraFeedback on embedding model [34]. Each entry reports the mean absolute score change $\Delta_{c,r}$, with the relative change shown in parentheses.

r	Irrelevant	Redundant	Novel relevant
0.00	0.000 (0.00%)	0.000 (0.00%)	0.000 (0.00%)
0.25	0.000 (0.00%)	0.049 (0.76%)	1.076 (25.27%)
0.50	0.000 (0.00%)	0.060 (1.04%)	1.860 (39.64%)
1.00	0.001 (0.02%)	0.066 (1.53%)	3.234 (64.85%)
2.00	0.012 (0.19%)	0.068 (1.77%)	5.184 (103.78%)

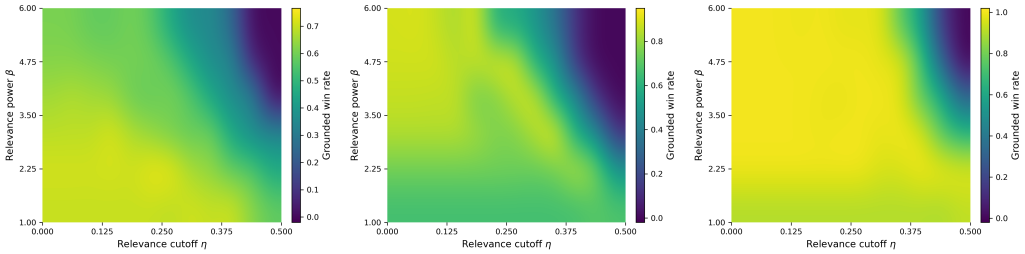
Table 7: Length and padding robustness on UltraFeedback on embedding model [35]. Each entry reports the mean absolute score change $\Delta_{c,r}$, with the relative change shown in parentheses.

r	Irrelevant	Redundant	Novel relevant
0.00	0.000 (0.00%)	0.000 (0.00%)	0.000 (0.00%)
0.25	0.000 (0.00%)	0.083 (1.19%)	1.130 (23.62%)
0.50	0.004 (0.05%)	0.100 (1.54%)	1.978 (38.13%)
1.00	0.010 (0.15%)	0.106 (2.10%)	3.583 (66.50%)
2.00	0.032 (0.48%)	0.111 (2.45%)	5.697 (105.07%)

A potential concern is that $\widehat{\mathcal{P}}$ may reward response length rather than semantic progress. To isolate the effect of added content, we construct controlled perturbations of real UltraFeedback prompt-response

Table 8: Length and padding robustness on UltraFeedback on embedding model [39]. Each entry reports the mean absolute score change $\Delta_{c,r}$, with the relative change shown in parentheses.

r	Irrelevant	Redundant	Novel relevant
0.00	0.000 (0.00%)	0.000 (0.00%)	0.000 (0.00%)
0.25	0.037 (0.82%)	0.050 (0.47%)	1.183 (23.69%)
0.50	0.037 (0.82%)	0.067 (0.73%)	2.127 (39.58%)
1.00	0.036 (0.81%)	0.055 (0.51%)	3.714 (68.25%)
2.00	0.062 (1.19%)	0.069 (0.75%)	6.589 (122.46%)



(a) potion-base-2M (b) Qwen3-Embedding-0.6B (c) Qwen3-Embedding-4B

Figure 7: Sensitivity of grounded win rate to the relevance cutoff η and weighting exponent β in the synthetic hallucination stress test.

pairs. For each base pair (q, a), we form a padded response

$$a^{(c,r)} = a \oplus p^{(c,r)},$$

where $\oplus$ denotes string concatenation, c denotes the padding condition, and r is the target padding ratio. The padding text is chosen so that

$$|p^{(c,r)}| \approx r|a|,$$

where $|\cdot|$ denotes token length. We then measure the score change

$$\Delta_{c,r} = \widehat{\mathcal{P}}_q(a^{(c,r)}) - \widehat{\mathcal{P}}_q(a).$$

We consider three padding conditions. *Irrelevant padding* is sampled from responses to unrelated prompts and filtered to have low embedding similarity to q. *Redundant padding* repeats evidence units already present in the original response a. *Novel relevant padding* is sampled from other responses to the same prompt and filtered to be relevant to q. The filtering is used only to construct controlled perturbation sets; the scoring procedure itself is unchanged.

This setup separates three possible effects of added text: generic length increase, repetition of existing evidence, and addition of new question-relevant evidence. If $\widehat{\mathcal{P}}_q$ were primarily a length proxy, all three padding conditions would increase the score similarly as r grows. Instead, we find that irrelevant and redundant padding produce only small score changes, while novel relevant padding produces a substantially larger increase. These results indicate that $\widehat{\mathcal{P}}_q$ is sensitive to whether added content contributes new question-relevant evidence, rather than simply rewarding longer responses. Tables 6, 7, 8, show the results with $r \in \{0.00, 0.25, 0.50, 1.00, 2.00\}$. We evaluate it on $n = 100$ examples, yielding a total of 1500 dialogues.

J.2 Hallucination

Let each domain define a set of valid entity pairs $(x, y) \in \mathcal{F}_d$ where x is the query entity and y is the correct answer entity. For example in the country-capital domain, $(\text{Germany}, \text{Berlin}) \in \mathcal{F}_d$.

For each pair (x, y), we instantiate a factual response from a fixed template $T_d(x, y)$. We then construct a hallucinated counterpart by replacing y with an entity $y' \in \mathcal{Y}_d$ where $y' \neq y$ and $(x, y') \notin \mathcal{F}_d$. This would give us a matched pair

$$T_d(x, y) \quad \text{vs.} \quad T_d(x, y')$$

where both responses share the same query entity and surface structure, but only the first contains the correct factual relation.

To increase linguistic diversity, we then use LLMs OpenAI's GPT 5.4 [41] only as paraphrasers, the entities x, y, y' are fixed, and the model is instructed not to alter them. Thus the correctness is determined by the manually specified factual relation set $\mathcal{F}_d$, not by the LLM.

For each matched pair, we compute GAUSSIANIG for the factual and corrupted responses under a sweep of relevance cutoffs η and relevance exponents β. We report grounded win rate, defined as the fraction of pairs for which the factual response receives a higher semantic-progress score than the corrupted response:

$$\mathrm{WinRate} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\left[\widehat{P}_q(D_i^{\mathrm{fact}}) > \widehat{P}_q(D_i^{\mathrm{corr}}) \right]. \tag{63}$$

Ties are counted as non-wins. Figure 7 shows that sensitivity to factual perturbations depends on both the embedding model and the relevance parameters. For Qwen3-Embedding-0.6B, $\beta = 1$ gives a stable win rate around 0.66, while moderate sharpening with larger β raises performance above 0.85 and peaks at 0.93. However, large cutoffs can sharply reduce performance, indicating that overly aggressive filtering removes useful evidence.

Qwen3-Embedding-4B is more robust, reaching near-perfect win rates across many settings, while Potion-2M is weaker and degrades under high cutoffs or large exponents. Overall, the stress test shows that GAUSSIANIG can prefer grounded evidence over semantically similar corrupted evidence, but that this behavior depends on embedding quality and parameter calibration. We therefore treat factuality evaluation as complementary to semantic-progress evaluation.

K 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.