Memory Conditioned Semantic Entropy for Multi-turn Dialogue Systems

Anonymous ACL submission

Abstract

Large language models have been increasingly deployed in multi-turn dialogue settings, where hallucinations arise not from isolated errors, but from cross-turn inconsistency and semantic drift. Existing uncertainty estimation methods, typically operate at the single-turn level and ignore temporal dependencies introduced by dialogue history, limiting their ability to characterize cross-turn hallucinations.

We introduce Memory-Conditioned Semantic Entropy (MC-SE), a framework that generalizes semantic entropy by incorporating an external dialogue memory into its uncertainty estimates. MC-SE combines multiple sampled responses through semantic clustering and penalizes outputs that conflict with earlier statements in the dialogue using natural language inference. In doing so, it produces uncertainty estimates that explicitly account for cross-turn consistency constraints.

Using controlled experiments on synthetic multi-turn question-answering dialogues, we analyze how memory conditioning systematically reshapes uncertainty across dialogue stages and consistency regimes.Our results show that memory-aware uncertainty reveals cross-turn inconsistencies that remain invisible to turn-local measures, highlighting the importance of memory-aware analysis for understanding hallucination behavior in multi-turn dialogue systems.

1 Introduction

In multi turn dialogue settings, the model outputs are not at all independent across turns and the responses at later turns are conditioned on the previously generated content as well. Recent work in this domain has made progress in hallucination detection especially for single turn settings but majority of these uncertainty measuring techniques ignores the dependencies introduced by dialogue system. (Maynez et al., 2020; Dziri et al., 2022) The history in a dialogue system accumulates over time and introduces dependencies on the further generation of response. As a result those measures fail to capture the reason for failure or the failure mode of multi turn systems: hallucination that arises from multi turn inconsistency and semantic drift.

Semantic entropy method estimates the uncertainty for single turn setting by sampling multiple generations and then by clustering them based on their semantic meaning instead of token matching or surface form similarity. Semantic meaning acts as a more faithful measurement. However this method is fundamentally only for single turn locally independent samples. (Kuhn et al., 2023) Each input output is treated independently and unrelated to each other regardless of the input. The method assumes that uncertainty arises solely from the current prompt. However this is not the case in majority of the Large language model systems. A LLM may produce a locally confident response that is inaccurate with its previous generations, leading to hallucinations that are invisible to single turn uncertainty methods.

From an individual-utterance perspective, a model's answer may seem assured and without ambiguity; however, when viewed within the context of a larger conversation (a "dialogue"), this same answer is clearly inconsistent with the previous correct answer. Thus, detecting such inconsistencies, requires an understanding of the model's previous answers as well as an understanding of how much the model is uncertain about its current answer. Therefore, it is important to have memory-aware uncertainty estimation methods in multi-turn conversations.

We introduce in this work the Memory-Conditioned Semantic Entropy (MC-SE) framework, which extends semantic entropy to condition uncertainty estimates on an external dialogue memory. MCSE maintains a memory of previous model

assertions during a dialogue and evaluates subsequent generations for semantic consistency with this memory using natural language inference. At estimation time, sampled responses which contradict previous dialogue assertions are penalized, and uncertainty is calculated over semantic clusters after memory-aware aggregation. Importantly, the memory is never modified during uncertainty estimation but only updated at the end of each turn, thus ensuring strict temporal causality.

Instead of suggesting a corrective mechanism for hallucinations, MC-SE is framed as a diagnostic signal that captures the evolution of uncertainty with accumulating dialogue context. MCSE uncovers failure modes unreachable to turn-local methods by conditioning uncertainty on dialogue memory: belief drift, contradictions that are delayed rather than immediate, and inconsistency across repeated inquiries. This framing allows uncertainty estimates to reveal not only ambiguity in the current input but also compatibility with the model's prior commitments.

We evaluate MC-SE in the context of multiturn, document-grounded QA, where dialogue facts are time-invariant, and contradictions are a reliable indicator of inconsistency. In the absence of standardized benchmarks for memory-aware hallucination detection, the evaluation focuses on controlled internal analyses rather than direct benchmarking to existing methods. Concretely, we investigate MC-SE's behavior across dialogue stages and regimes of consistency, isolating the contribution of memory conditioning to uncertainty patterns. Results show that memory conditioning systematically affects uncertainty estimates and reveals cross-turn failure modes that cannot be detected by turn-local uncertainty measures.

In summary, this paper makes the following contributions:

- We formalize memory-conditioned semantic entropy, extending semantic entropy to multi-turn dialogue by incorporating external dialogue memory and entailment-based consistency checks.
- We present a temporally causal workflow for memory-aware uncertainty estimation that operates at the semantic level rather than surface form.
- We provide a detailed behavioral analysis of MC-SE in multi-turn question answering, highlighting how memory conditioning exposes hallucination patterns that are not captured by turn-local uncertainty estimates.

Together, these contributions underscore the importance of memory-aware uncertainty estimation for understanding and diagnosing hallucinations in multi-turn dialogue systems.

2 Related Work

Hallucination detection has received significant attention in recent years, particularly in the context of large language models deployed for question answering and text generation (Maynez et al., 2020; **?**). Prior work has explored a range of strategies, including confidence estimation (Kuhn et al., 2023), factual verification against external sources (Thorne and Vlachos, 2018), and post-hoc consistency checks (Krishna et al., 2021). These approaches have proven effective in identifying unsupported or fabricated content in single-turn generation settings. However, most existing hallucination detection methods operate on isolated model outputs and do not explicitly account for dependencies introduced by prior dialogue turns (Dziri et al., 2022). As a result, they are ill-suited to detect hallucinations that arise from cross-turn inconsistency, where a response may be locally plausible yet incompatible with earlier model assertions.

Several methods attempt to mitigate hallucinations through model alignment or correction mechanisms, such as reinforcement learning, constrained decoding, or retrieval-based grounding (Lewis et al., 2020). While these approaches aim to reduce hallucination frequency, they differ fundamentally from our objective. Rather than correcting or suppressing hallucinations, our work focuses on characterizing uncertainty in multi-turn dialogue and identifying failure modes that emerge as dialogue context accumulates.

3 Background

3.1 Uncertainty Estimation for Free-Form Generation

Uncertainty estimation for language generation has traditionally been studied in single-turn settings, where each input-output pair is treated independently.(Gal and Ghahramani, 2016) In such settings, uncertainty arises primarily from ambiguity in the prompt or from variability among plausible responses. This formulation has proven effective

for isolated generation tasks, where outputs can be evaluated without reference to prior model behavior.

However, this assumption does not hold in multi-turn dialogue. In interactive settings, model outputs are temporally dependent, and responses generated at earlier turns influence the interpretation and validity of future generations. As a result, uncertainty estimation methods that treat each turn independently fail to capture dependencies introduced by dialogue history.

3.2 Semantic Entropy

Semantic entropy has been proposed as a meaning-level alternative to token-level uncertainty estimation (Farquhar et al., 2024; Kuhn et al., 2023). Rather than measuring uncertainty over surface forms, semantic entropy samples multiple model generations and clusters them according to semantic equivalence. This approach provides a more faithful estimate of uncertainty in free-form text generation, particularly in the presence of paraphrastic variation.

Semantic entropy is then computed as the entropy of the probability distribution over the semantic clusters. Let $\mathcal{Y} = \{y_1, \ldots, y_N\}$ denote a set of sampled generations, and let $\mathcal{C} = \{c_1, \ldots, c_K\}$ be the resulting set of semantic clusters. The probability of a cluster is estimated as the fraction of samples assigned to it. Semantic entropy is then defined as:

$$H_{\text{sem}} = -\sum_{k=1}^{K} p(c_k) \log p(c_k), \qquad (1)$$

where $p(c_k)$ denotes the empirical probability of cluster c_k. Low semantic entropy indicates agreement over meaning across generations, while high semantic entropy suggests semantic disagreement and increased risk of hallucination. This observation motivates the need for memory-aware uncertainty estimation methods that explicitly model cross-turn consistency. In the following section, we introduce Memory-Conditioned Semantic Entropy (MC-SE), which extends semantic entropy by conditioning uncertainty on an external dialogue memory and evaluating semantic consistency across turns.

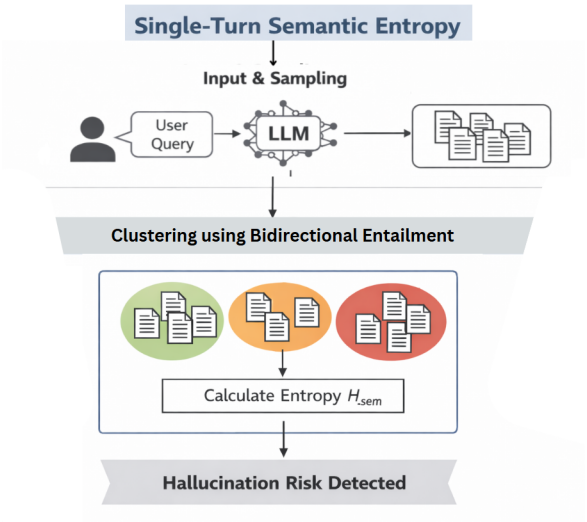

Figure 1: Workflow of single-turn semantic entropy computation. Given a user query, multiple responses are sampled from the language model, grouped into semantic clusters, and used to compute semantic entropy as a measure of hallucination risk.

4 Memory-Conditioned Semantic Entropy

We now introduce Memory-Conditioned Semantic Entropy (MC-SE), a framework for uncertainty estimation in multi-turn dialogue that explicitly conditions semantic uncertainty on prior dialogue history.

4.1 Dialogue Memory

MC-SE extends semantic entropy by conditioning uncertainty estimates on an external dialogue memory that stores prior model outputs within a dialogue. It serves as a record of the model's prior commitments rather than as a verified knowledge base. Memory is scoped at the dialogue level and reset between dialogues.

At turn t, the memory state $M_{1:t-1}$ consists of all finalized model responses generated in turns 1 through $t-1$. Importantly, memory is read-only during uncertainty computation at turn t and is updated only after uncertainty has been fully computed, enforcing strict temporal causality.

4.2 Sampling and Semantic Clustering

Given an input x_t, we sample multiple candidate responses $\{s_1, \ldots, s_N\}$ from the language model using stochastic decoding. Each sample is associated with an approximate log-likelihood computed

as the sum of per-token log probabilities:

$$\log p(s_i \mid x_t) \approx \sum_j \log p(w_{i,j} \mid w_{i,<j}, x_t). \quad (2)$$

Following prior work on semantic entropy, we cluster sampled responses based on semantic equivalence rather than surface-form similarity. In our implementation, semantic equivalence is assessed using bidirectional entailment, grouping responses that mutually entail one another into the same semantic cluster. This yields a set of semantic clusters $\{C_t^{(1)}, \ldots, C_t^{(K_t)}\}$ representing distinct meanings expressed by the model.

4.3 Memory Consistency Weighting

To incorporate dialogue history, semantic clusters are evaluated for consistency with retrieved dialogue memory. Using a natural language inference (NLI) model, we assess whether each cluster's representative responses are entailed by, neutral with respect to, or contradictory to prior memory facts.

Memory-consistency weights are then assigned at the cluster level, allowing MC-SE to reduce the contribution of interpretations that violate earlier dialogue commitments without discarding them entirely. Clusters that remain consistent with memory retain full weight, while contradictory clusters contribute reduced probability mass during entropy computation.

4.4 Memory-Conditioned Semantic Entropy

Memory consistency weights are incorporated during cluster probability aggregation. Specifically, log-likelihoods are combined with the logarithm of memory weights before aggregating probability mass at the cluster level. The resulting memory-conditioned cluster distribution reflects both semantic diversity and cross-turn consistency constraints.

The memory-conditioned semantic entropy at turn t is then computed as:

$$H_{\mathrm{MC}}(x_t, M_{1:t-1}) = -\sum_{k=1}^{K_t} \pi_t(k) \log \pi_t(k),$$

where $\pi_t(k)$ denotes the normalized probability of cluster $C_t^{(k)}$ after memory-aware aggregation.

4.5 Scope of the Framework

While MC-SE defines a broader composite risk formulation that may incorporate additional signals such as contradiction frequency or temporal drift, in this work we focus on the entropy component to isolate the effect of memory conditioning on uncertainty estimation.

5 Dataset Construction

To study memory effects and long-form uncertainty in dialogue settings, we require conversational data with controlled factual grounding. Due to the lack of established benchmarks for memory-aware hallucination detection, we construct a synthetic multi-turn question answering dataset based on SQuAD-style passages. (Rajpurkar et al., 2016) Each dialogue consists of multiple turns that reference the same underlying document, including repeated and paraphrased questions designed to expose cross-turn inconsistency. This setup enables precise control over dialogue structure and consistency regimes.

5.1 Dialogue Generation

Each SQuAD question-answer pair is transformed into a multi-turn conversational interaction by prompting a large language model to generate follow-up questions and answers conditioned on the evolving dialogue history. The initial turn corresponds to the original SQuAD question, while subsequent turns are generated to reference earlier responses, encouraging temporal continuity across the dialogue.

5.2 Evaluation Protocol

Dialogues are processed sequentially, with memory initialized at the start of each dialogue and updated incrementally across turns. At each turn, MC-SE is computed using sampled model responses and the dialogue memory accumulated from prior turns. Memory is accessed in a read-only manner during uncertainty computation and updated only after uncertainty has been fully computed, ensuring strict temporal causality.

Rather than benchmarking against existing uncertainty measures, our evaluation focuses on characterizing the behavior of MC-SE across dialogue stages and consistency regimes, isolating the effect of memory conditioning on uncertainty estimates.

6 Workflow of Memory-Conditioned Semantic Entropy

While Section 4 defines MC-SE formally, this section describes the end-to-end workflow of Memory-Conditioned Semantic Entropy (MC-SE) as applied

to multi-turn dialogue. We focus on the operational pipeline that computes uncertainty at each dialogue turn while enforcing dialogue-level consistency.

MC-SE operates at the level of complete dialogues rather than individual turns. Each dialogue is processed sequentially, and a fresh external memory is initialized at the beginning of each dialogue. Turns are handled strictly in chronological order, ensuring that uncertainty estimation at turn t depends only on information available up to turn $t-1$.

Given an input at each turn, x_t, MC-SE samples a set of candidate responses from the language model $\mathcal{S}_t = \{s_1, \ldots, s_N\}$ using stochastic decoding. Each sample is associated with an approximate log-likelihood computed as the sum of per-token log probabilities. Following the semantic clustering procedure defined in Section 4, sampled responses are grouped into clusters corresponding to distinct meanings.

Dialogue memory is then consulted, following the consistency mechanism defined in Section 4.3, to assess each sampled interpretation. Samples that contradict any stored memory assertion are down-weighted according to the memory-consistency mechanism defined in Section 4, while memory-consistent interpretations retain full weight.

The resulting memory-conditioned distribution over semantic clusters is used to compute the MC-SE uncertainty score for the current turn. After uncertainty computation is complete, the finalized model response is appended to dialogue memory and becomes available for subsequent turns.

This inference workflow enables MC-SE to capture uncertainty arising from both prompt ambiguity and cross-turn inconsistency, while preserving temporal causality throughout the dialogue.

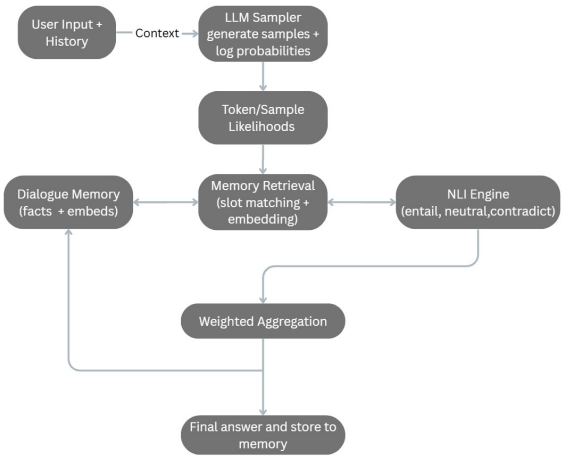

Figure 2: Memory-Conditioned Semantic Entropy (MC-SE) workflow. At each dialogue turn, multiple responses are sampled and evaluated against an external dialogue memory using an NLI model. Memory-consistency weights are applied during cluster-level aggregation to form a memory-conditioned distribution over semantic interpretations. MC-SE is computed as the entropy of this distribution, and the selected response is appended to memory for subsequent turns.

7 Experimental Setup

The experiments are designed to evaluate the effectiveness of memory-conditioned semantic entropy as a *diagnostic signal* for hallucination risk in multi-turn dialogue, rather than as a task-specific performance metric. Accordingly, we focus on controlled generation settings and analysis-driven evaluation.

7.1 Models Evaluated

All experiments are conducted using a single large language model (Mistral-7B-Instruct-v0.1), with fixed parameters to isolate the effect of entropy-based uncertainty measures. The model is used in a frozen setting without fine-tuning. For each dialogue turn, multiple responses are sampled under identical conditions to estimate semantic uncertainty.

7.2 Metrics

We evaluate uncertainty using both standard single-turn semantic entropy and the proposed memory-conditioned semantic entropy (MC-SE). Hallucination behavior is assessed via semantic correctness and cross-turn consistency signals, and we report correlations between entropy values and hallucination indicators across dialogue turns. Analysis is preferred over threshold-based accuracy to reflect the continuous nature of uncertainty.

7.3 Evaluation Goals

Our evaluation aims to determine whether memory conditioning improves the sensitivity of semantic entropy to long-horizon inconsistencies and inaccuracies in multi-turn dialogue. Rather than optimizing for state-of-the-art detection performance, we focus on characterizing when and why memory-aware uncertainty provides additional signal, particularly across dialogue stages and consistency regimes.

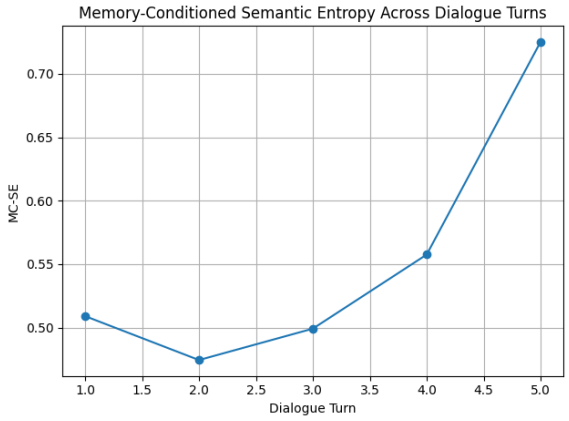

Figure 3: Memory-conditioned semantic entropy (MC-SE) as a function of dialogue turn. MC-SE diverges from turn-local uncertainty as dialogue memory accumulates, reflecting cross-turn consistency constraints.

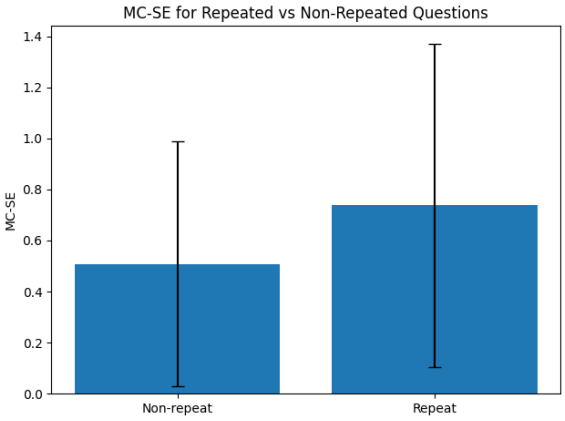

Figure 4: Memory-conditioned semantic entropy for repeated and non-repeated questions. Repeated questions exhibit higher uncertainty, indicating sensitivity to dialogue-level consistency rather than surface-form variation.

8 Results and Behavioral Analysis

8.1 Evaluation Overview

We present a behavioral analysis of Memory-Conditioned Semantic Entropy (MC-SE) in controlled multi-turn dialogue settings. Rather than benchmarking against task-specific performance metrics, our analysis focuses on how memory conditioning alters uncertainty estimates across dialogue stages and consistency regimes. We compare MC-SE to standard semantic entropy to isolate the effect of incorporating dialogue memory.

8.2 MC-SE across Dialogue Stages

We first examine how uncertainty evolves as dialogue context accumulates. Figure 3 shows the average memory-conditioned semantic entropy as a function of dialogue turn index.

At early turns, where dialogue memory is empty, MC-SE closely matches standard semantic entropy, reflecting uncertainty driven primarily by prompt ambiguity. As dialogues progress, MC-SE diverges from turn-local semantic entropy, indicating that memory conditioning increasingly influences uncertainty estimates.

Notably, MC-SE does not exhibit a strictly monotonic trend across turns. In some dialogues, uncertainty decreases as earlier ambiguities are resolved by additional contextual constraints, rather than reduced model variability, while in others it increases due to accumulating cross-turn constraints. This behavior suggests that MC-SE captures interactions between prior commitments and current interpretations rather than reflecting dialogue length alone.

8.3 Effect of Memory Conditioning

To isolate the effect of memory conditioning, we compare entropy values at turns with empty dialogue memory to those with non-empty memory. Once memory is populated, MC-SE exhibits systematically different uncertainty behavior compared to turn-local semantic entropy.

This difference confirms that changes in uncertainty are attributable to conditioning on prior dialogue assertions rather than to sampling variability alone. In the absence of memory, MC-SE reduces to standard semantic entropy, while the presence of memory introduces additional constraints that reshape the uncertainty landscape.

8.4 Repeated Questions as a Consistency Regime

Repeated questions provide a natural probe for dialogue-level consistency, as they require the model to maintain alignment with earlier responses. We observe that MC-SE assigns higher uncertainty to repeated questions than to non-repeat turns.

Importantly, this increase does not arise from surface-form variation. Even when repeated questions elicit lexically similar responses, MC-SE assigns elevated uncertainty when multiple interpretations remain plausible under dialogue-level consistency constraints. This behavior indicates that MC-SE is sensitive to semantic divergence from prior commitments rather than to lexical variation alone.

Memory State	Mean MC-SE	Std. MC-SE	# Turns
Empty	0.5090	0.4183	12
Non-empty	0.5365	0.5166	41

Table 1: Memory-conditioned semantic entropy for turns with empty and non-empty dialogue memory. Differences reflect the effect of memory conditioning rather than sampling variability.

8.5 Response to Contradictions

We further analyze MC-SE behavior when sampled responses contradict prior dialogue assertions. In such cases, MC-SE assigns substantially higher uncertainty, reflecting unresolved disagreement among competing interpretations after memory conditioning.

This result suggests that MC-SE responds specifically to semantic inconsistency, rather than merely reflecting variability in sampled generations. By down-weighting interpretations that contradict dialogue memory, MC-SE exposes failure modes that remain invisible to turn-local uncertainty measures.

8.6 Summary of Findings

Across analyses, MC-SE exhibits behavior aligned with its design objectives. Uncertainty estimates are shaped not only by ambiguity in the current input but also by compatibility with prior dialogue commitments. By conditioning uncertainty on dialogue memory, MC-SE exposes cross-turn failure modes, including delayed contradictions and belief drift, that are not captured by turn-local uncertainty measures.

9 Discussion

Our results demonstrate that memory-conditioned semantic entropy captures a distinct aspect of uncertainty that is overlooked by turn-local measures. Our goal is not to outperform semantic entropy in aggregate metrics, but to reveal failure modes that semantic entropy is structurally unable to detect. MC-SE provides complementary diagnostic signal by exposing long-horizon inconsistencies that arise specifically in multi-turn dialogue.

MC-SE is not intended as a standalone hallucination detector, but as an uncertainty signal that highlights instability introduced through dialogue memory.

In practice, memory-conditioned uncertainty is most valuable in settings involving extended interactions—such as conversational assistants, tutoring systems, or multi-step reasoning agents—where maintaining internal consistency over time is critical.

10 Conclusion

We study hallucination detection in multi-turn dialogue through the lens of semantic uncertainty. By introducing memory-consistent semantic entropy, we show that conditioning uncertainty on dialogue history enables the detection of long-horizon semantic drift that single-turn measures miss.

Our findings suggest that hallucinations in dialogue are fundamentally temporal phenomena, and that uncertainty estimation must account for conversational memory. We hope this work encourages further research on memory-aware evaluation methods for conversational language models.

11 Limitations and Future Work

This study has several limitations. First, the dialogue data used in our experiments is synthetically constructed and may not capture the full diversity of real-world conversational behavior. While this controlled setting enables isolation of memory effects, validation on human-curated dialogue data remains an important next step.

Second, MC-SE depends on the quality and diversity of sampled generations. Limited or homogeneous sampling may reduce its sensitivity to semantic divergence. Future work could explore adaptive or budget-aware sampling strategies to mitigate this dependence.

From a scalability perspective, estimating semantic entropy across multiple samples introduces computational overhead, particularly for long dialogues. More efficient approximations or incremental entropy estimation methods could improve practicality.

We focus on a single model to isolate uncertainty behavior, and leave cross-model validation to future work.

Finally, the lack of standardized benchmarks for long-horizon hallucination in dialogue limits systematic comparison across methods. Developing shared multi-turn evaluation datasets with con-

trolled consistency regimes would significantly advance research on memory-aware uncertainty estimation.

References

Nouha Dziri and 1 others. 2022. Faithdial: A faithful benchmark for information-seeking dialogue. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630:625–632.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*.

Kalpesh Krishna and 1 others. 2021. Hurdles to progress in long-form question answering. In *North American Chapter of the Association for Computational Linguistics*.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations*.

Patrick Lewis and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Pranav Rajpurkar and 1 others. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

James Thorne and Andreas Vlachos. 2018. Fever: A large-scale dataset for fact extraction and verification. In *North American Chapter of the Association for Computational Linguistics*.