

Localizing Input Uncertainty Quantification for Large Language Models via Shapley Values

26' NeurIPS Sydney (Poster)

Seongjun Lee*, Suwan Yoon*, Changhee Lee†

*: Equal contribution, †: Corresponding author

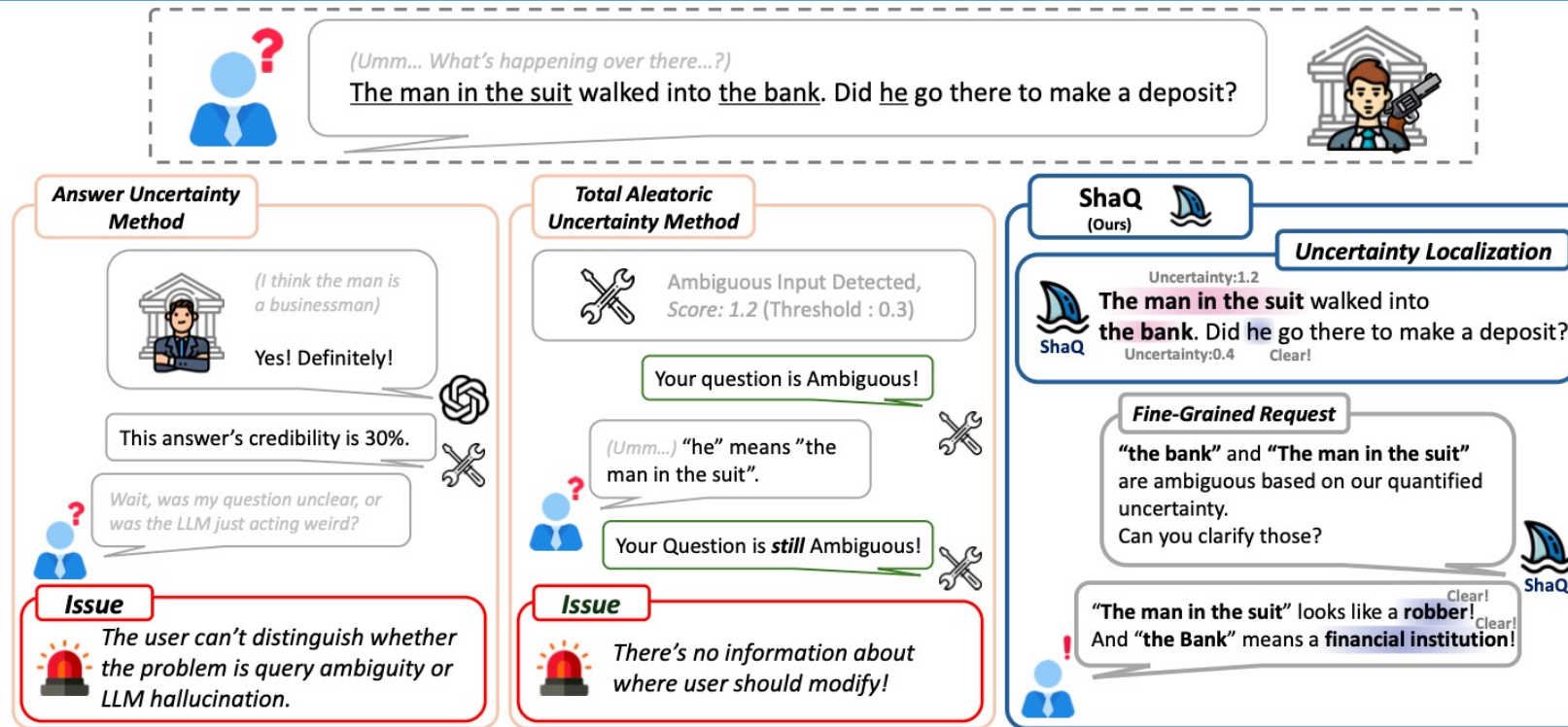
The Actionable Intelligence Lab

@Korea University, Seoul, Korea



pyoung7307, suwanyoon, changheele@korea.ac.kr

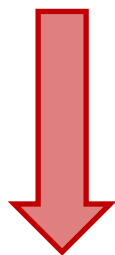
Motivation



- Most existing LLM uncertainty quantification methods focus on uncertainty in **model outputs**.
- Input-induced uncertainty has received relatively less attention, and existing methods typically summarize it with a **single scalar score**, providing little guidance on what should be clarified.
- Our goal is to **localize input-induced uncertainty** to specific ambiguous spans, enabling actionable clarification.

Input Uncertainty Quantification in LLMs

$$H(Y|x) = \underbrace{I(Y; \theta|x)}_{\text{Epistemic}} + \underbrace{\mathbb{E}_{\theta}[H(Y|x, \theta)]}_{\text{Aleatoric}}$$



$$H(Y|x) = \underbrace{I(Y; C|x)}_{\text{Input-induced}} + \underbrace{\mathbb{E}_c[H(Y|x, c)]}_{\text{Remaining}}$$

Bayesian Active Learning for Classification and Preference Learning

Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, Máté Lengyel
Computational and Biological Learning Laboratory
University of Cambridge

[11'Arxiv]

Decomposing Uncertainty for Large Language Models
through Input Clarification Ensembling

Bairu Hou¹ Yujian Liu¹ Kaizhi Qian² Jacob Andreas³ Shiyu Chang¹ Yang Zhang²

[24'ICML]

- The bayesian active learning paper proposed the decomposition of the predictive entropy into two parts.
- **In black-box LLMs, sampling model parameters θ is generally infeasible.** Recent input-centric UQ instead introduces a latent clarification variable C , representing plausible interpretations of the input.
 - $I(Y; C | x)$: Uncertainty caused by input ambiguity
 - $\mathbb{E}_c[H(Y | x, C)]$: Uncertainty remaining after clarification.

However, it provides only a global uncertainty score, without localizing its source in the input.

Ambiguous Span & Naïve Attribution

$$H(Y|x) = \underbrace{I(Y; C|x)}_{\text{Input-induced}} + \underbrace{\mathbb{E}_c[H(Y|x, c)]}_{\text{Remaining}}$$

Definition 3.1 (Ambiguous Span). An *ambiguous span* is a contiguous subsequence of the input sequence that constitutes a minimal semantically coherent unit admitting multiple plausible interpretations. We denote by $\mathcal{S} = \{s_1, \dots, s_n\}$ a collection of non-overlapping ambiguous spans extracted from x .

Definition 3.2 (LOO Span Attribution). The LOO attribution of span s_k is defined as the conditional mutual information

$$\phi_k^{\text{LOO}} \triangleq \mathcal{I}(Y; C_k \mid x, C_{\setminus k}) = \mathcal{H}(Y \mid x, C_{\setminus k}) - \mathcal{H}(Y \mid x, C), \quad (2)$$

which quantifies the information about Y contributed by clarifying s_k after all remaining spans have been resolved.

- We denote the joint clarification as $C = (C_1, \dots, C_n)$
- $C_{\setminus k}$ means that joint clarification excluding the k -th span.

Ambiguous Span & Naïve Attribution

Problems

In our running example,

The man in the suit walked into *the bank*. Did he go there to make a deposit?

The man in the suit → Robber, Businessman

The bank → Riverbank, Financial bank

After clarifying “the man in the suit” as a robber

The robber walked into *the bank*. Did he go there to make a deposit?

⇒ “the bank” is implicitly constrained toward a financial bank.

⇒ Its LOO attribution becomes artificially small.

How can we measure each span’s attribution while accounting for their interactions?

- W
- $C_{\setminus k}$ means that joint clarification excluding the k -th span.

Shapley Value : Accounting for interactions

- Shapley value measures each player's attribution by averaging its marginal contribution over all possible coalition contexts.

$$\phi_k = \sum_{A \subseteq \mathcal{N} \setminus \{k\}} \frac{|A|! (n - |A| - 1)!}{n!} (v(A \cup \{k\}) - v(A))$$

- $A \subseteq \mathcal{N} \setminus \{k\}$: a coalition excluding player k
- $v(A)$: payoff of coalition A
- Shapley values satisfy four key axioms.
 - Efficiency

$$\sum_{i \in \mathcal{N}} \phi_i = v(\mathcal{N}) - v(\emptyset)$$

whole payoff must be exactly divided.

- Other axioms: Symmetry, Dummy player, Additivity.

Shapley-based input uncertainty Quantification (ShaQ)

Definition 4.1 (Value Function). The value function $v : 2^{\mathcal{N}} \rightarrow \mathbb{R}$ assigns to each coalition $\mathcal{A} \subseteq \mathcal{N}$ the mutual information between Y and the joint clarifications of the spans indexed by $\mathcal{A}$:

$$v(\mathcal{A}) = \mathcal{I}(Y; C_{\mathcal{A}} | x) = \mathcal{H}(Y | x) - \mathbb{E}_{c_{\mathcal{A}} \sim q(C_{\mathcal{A}} | x)} [\mathcal{H}(Y | x, c_{\mathcal{A}})], \quad (3)$$

where $C_{\mathcal{A}} = (C_j : j \in \mathcal{A})$ denotes the joint clarifications associated with the spans in $\mathcal{A}$. By convention, we have $v(\emptyset) = 0$ and $v(\mathcal{N}) = \mathcal{I}(Y; C | x)$.

Theorem 4.1 (Shapley Value of Ambiguous Span). For any coalition $\mathcal{A} \subseteq \mathcal{N} \setminus \{k\}$, the marginal contribution of span s_k satisfies

$$v(\mathcal{A} \cup \{k\}) - v(\mathcal{A}) = \mathcal{I}(Y; C_k | x, C_{\mathcal{A}}). \quad (4)$$

Consequently, the Shapley value of span s_k can be written as

$$\phi_k = \sum_{\mathcal{A} \subseteq \mathcal{N} \setminus \{k\}} \frac{|\mathcal{A}|!(n - |\mathcal{A}| - 1)!}{n!} \mathcal{I}(Y; C_k | x, C_{\mathcal{A}}), \quad (5)$$

which characterizes ϕ_k as the expected information gain regarding Y from clarifying s_k , marginalized over all possible clarification contexts $C_{\mathcal{A}}$.

Remark 4.1 (Efficiency). The Shapley values in (5) partition the total input-induced uncertainty:

$$\sum_{k=1}^n \phi_k = v(\mathcal{N}) = \mathcal{I}(Y; C | x). \quad (6)$$

Shapley-based input uncertainty Quantification (ShaQ)

Definition 4.1 (Value Function). The value function $v : 2^{\mathcal{N}} \rightarrow \mathbb{R}$ assigns to each coalition $\mathcal{A} \subseteq \mathcal{N}$ the mutual information between Y and the joint clarifications of the spans indexed by $\mathcal{A}$:

$$v(\mathcal{A}) = \mathcal{I}(Y; C_{\mathcal{A}} | x) = \mathcal{H}(Y | x) - \mathbb{E}_{c_{\mathcal{A}} \sim q(C_{\mathcal{A}} | x)} [\mathcal{H}(Y | x, c_{\mathcal{A}})], \quad (3)$$

where $C_{\mathcal{A}} = (C_j : j \in \mathcal{A})$ denotes the joint clarifications associated with the spans in $\mathcal{A}$. By

Computation Goal

$$\begin{aligned} I(Y; C_k | x, C_A) &= v(A \cup \{k\}) - v(A) \\ &= \mathbb{E}_{c_A} [H(Y | x, c_A)] - \mathbb{E}_{c_{A \cup \{k\}}} [H(Y | x, c_{A \cup \{k\}})] \end{aligned}$$

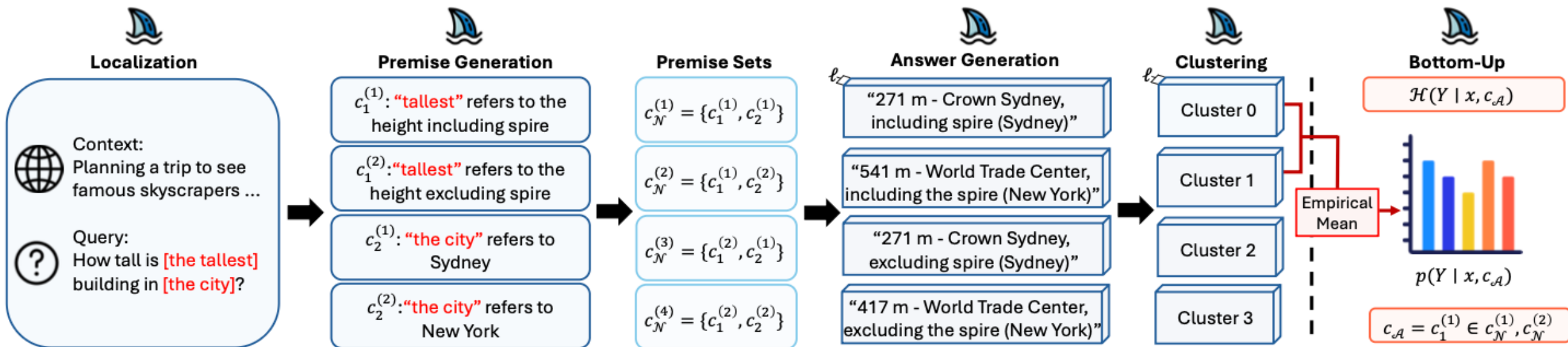
Thus we need to compute $\mathbb{E}_{c_A} [H(Y | x, c_A)]$ for all coalition $A \subseteq \mathcal{N}$.

which characterizes ϕ_k as the expected information gain regarding Y from clarifying s_k , marginalized over all possible clarification contexts C_A .

Remark 4.1 (Efficiency). The Shapley values in (5) partition the total input-induced uncertainty:

$$\sum_{k=1}^n \phi_k = v(\mathcal{N}) = \mathcal{I}(Y; C | x). \quad (6)$$

Shapley-based input uncertainty Quantification (ShaQ)



ShaQ : Bottom-Up Marginalization

ALGORITHM 1: BOTTOM-UP MARGINALIZATION

Input: Bottom-level (finest) samples from $\{p(Y \mid x, c_{\mathcal{N}})\}$ for all joint clarifications.

For each subset $\mathcal{A} \subseteq \mathcal{N}$ and each premise combination $c_{\mathcal{A}}$:

1. Identify all bottom-level $c_{\mathcal{N}}$ consistent with $c_{\mathcal{A}}$, i.e., $c_{\mathcal{N}} = c_{\mathcal{A}} \cup c_{\mathcal{N} \setminus \mathcal{A}}$ for some $c_{\mathcal{N} \setminus \mathcal{A}}$.
2. Compute $p(Y \mid x, c_{\mathcal{A}}) = \sum_{c_{\mathcal{N} \setminus \mathcal{A}}} q(c_{\mathcal{N} \setminus \mathcal{A}} \mid x) p(Y \mid x, c_{\mathcal{A}} \cup c_{\mathcal{N} \setminus \mathcal{A}})$.
3. Compute $\mathcal{H}(Y \mid x, c_{\mathcal{A}}) = - \sum_y p(y \mid x, c_{\mathcal{A}}) \log p(y \mid x, c_{\mathcal{A}})$.

For each subset $\mathcal{A} \subseteq \mathcal{N}$:

4. Compute $\mathbb{E}_{c_{\mathcal{A}}}[\mathcal{H}(Y \mid x, c_{\mathcal{A}})] = \sum_{c_{\mathcal{A}}} q(c_{\mathcal{A}} \mid x) \mathcal{H}(Y \mid x, c_{\mathcal{A}})$.

Output: Cached values $\mathbb{E}_{c_{\mathcal{A}}}[\mathcal{H}(Y \mid x, c_{\mathcal{A}})]$ for Shapley computation via (5).

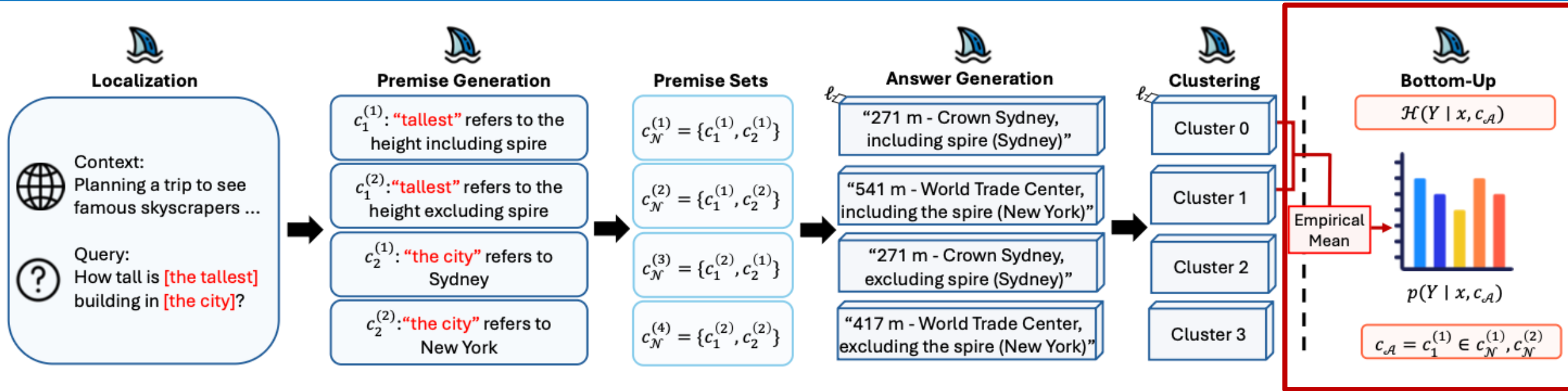
Why do we need this?

The naïve computation of $\mathbb{E}_{c_{\mathcal{A}}}[\mathcal{H}(Y \mid x, c_{\mathcal{A}})]$ is separate calls of LLMs for each coalition A . But this doesn't satisfy marginalization relation:

$$\underbrace{p(Y \mid x, c_A)}_{\text{Call 1}} = \sum_{c_k} q(c_k \mid x) \underbrace{p(Y \mid x, c_A, c_k)}_{\text{Call 2}}$$

Thus the $I(Y; C_k \mid x, C_A) = v(A \cup k) - v(A)$ can be assigned negative value.

Shapley-based input uncertainty Quantification (ShaQ)



Property 4.1 (Hierarchical Monotonicity). *For any subset $\mathcal{A} \subseteq \mathcal{N}$ and any span $k \notin \mathcal{A}$, the expected conditional entropy decreases monotonically as more spans are clarified:*

$$\mathbb{E}_{c_A}[\mathcal{H}(Y | x, c_A)] \geq \mathbb{E}_{c_{A \cup \{k\}}}[\mathcal{H}(Y | x, c_{A \cup \{k\}})]. \quad (7)$$

Equivalently, every marginal contribution of the ambiguous span in (5) is non-negative:

$$\mathcal{I}(Y; C_k | x, C_A) = \mathbb{E}_{c_A}[\mathcal{H}(Y | x, c_A)] - \mathbb{E}_{c_{A \cup \{k\}}}[\mathcal{H}(Y | x, c_{A \cup \{k\}})] \geq 0. \quad (8)$$

The derivation is provided in Appendix A.

Experiments (Input Ambiguity Detection)

Table 1: Uncertainty-based ambiguity detection.

Method	AmbigQA (GPT-4)			AmbiEnt (GPT-5.4-mini)		
	F1 Score	AUROC	AUPRC	F1 Score	AUROC	AUPRC
<i>Other Uncertainty Methods</i>						
Semantic Entropy	44.58	57.16	57.00	42.20	57.56	54.36
Sample Diversity	55.24	60.86	59.35	48.33	59.23	57.48
Sample Repetition	56.98	60.90	59.24	43.69	56.67	57.78
Self-Consistency	53.40	59.33	57.59	47.93	59.28	59.46
<i>Aleatoric Methods</i>						
Ask4Conf-D	56.36	52.44	53.97	65.43	50.82	52.54
Deep Ensembles	59.70	60.29	60.31	56.37	58.96	61.49
ICE	65.54	60.66	59.98	57.33	55.40	53.21
LOO	67.77	50.43	54.35	66.07	49.04	57.00
LOO (Max Span)	67.77	52.36	55.80	66.07	52.66	59.69
ShaQ (Ours, Max Span)	70.38	64.20	59.95	74.60	77.80	75.19
ShaQ (Ours)	70.38	65.77	61.52	73.84	77.68	75.71

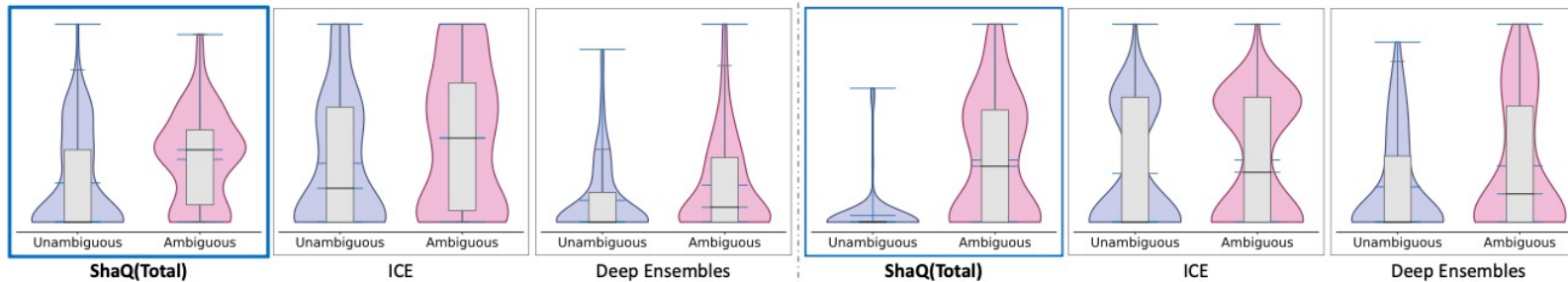


Figure 3: Aleatoric uncertainty distributions on AmbigQA (Left) and AmbiEnt (Right).

- **Ambiguity Detection: ShaQ consistently outperforms existing UQ baselines, demonstrating more reliable ambiguity quantification.**

Table 2: Answerer & Clusterer backbone with GPT-5.4-mini / Gemini models

Method	F1 Score	AmbigQA AUROC	AUPRC
GPT-5.4-mini			
Deep Ensembles	63.41	62.62	57.29
ICE	66.67	64.61	63.11
ShaQ (Ours, Max Span)	69.53	62.37	58.63
ShaQ (Ours)	69.53	63.60	62.08
Gemini-3.1-flash-lite-preview			
Deep Ensembles	44.30	55.92	54.26
ICE	63.21	63.41	63.53
ShaQ (Ours, Max Span)	66.67	62.97	58.95
ShaQ (Ours)	66.67	64.22	63.17
Gemini-2.5-flash			
Deep Ensembles	56.28	55.42	53.42
ICE	65.82	61.71	61.15
ShaQ (Ours, Max Span)	68.75	63.50	60.06
ShaQ (Ours)	68.75	64.84	62.74

Table 3: Answerer & Clusterer backbone with Gemini models

Method	F1 Score	AmbiEnt AUROC	AUPRC
Gemini-3.1-flash-lite-preview			
Deep Ensembles	27.52	46.74	47.76
ICE	50.70	50.71	49.45
Ask4Conf-D	49.12	45.51	49.57
LOO	66.07	53.23	55.63
LOO (Max Span)	65.47	52.94	56.20
ShaQ (Ours, Max Span)	65.59	72.39	70.65
ShaQ (Ours)	65.59	71.96	69.32
Gemini-2.5-flash			
Deep Ensembles	41.66	54.28	56.19
ICE	51.38	51.31	49.97
Ask4Conf-D	60.20	42.86	46.83
LOO	66.07	38.64	44.39
LOO (Max Span)	65.47	41.48	46.22
ShaQ (Ours, Max Span)	71.42	76.40	74.52
ShaQ (Ours)	71.42	76.47	75.78

Experiments (Uncertainty-guided Clarification)

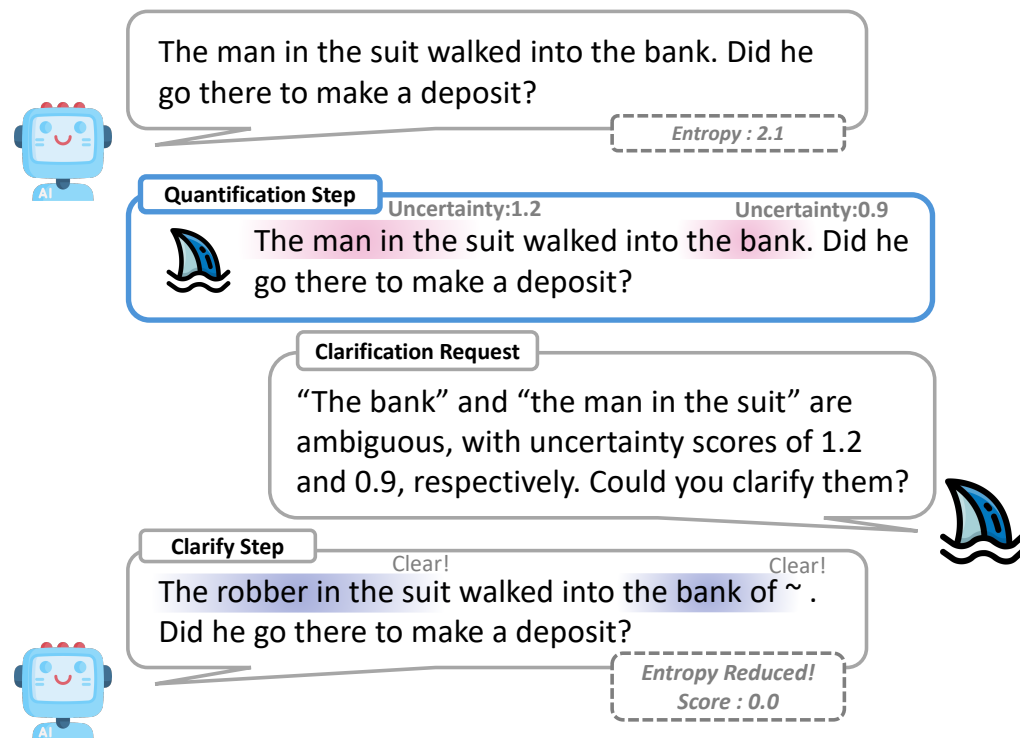


Figure 4: Uncertainty-guided clarification.

Table 5: High Aleatoric Uncertainty Superset

Method	Clarified	$\Delta H(\uparrow)$	Edit ($\downarrow$)
Semantic Entropy	0.3423	0.0330	6.4545
ICE	0.4387	-0.0633	6.7878
Deep Ensembles	0.3831	-0.0076	7.4545
ShaQ (Ours)	0.3365	0.0389	5.5151

Table 13: High ShaQ Max-Span Aleatoric Uncertainty Set (> 1).

Method	Clarified	$\Delta H(\uparrow)$	Edit ($\downarrow$)
Semantic Entropy	0.3429	0.1115	5.8571
ICE	0.5397	-0.0852	7.2857
Deep Ensembles	0.5245	-0.0700	7.3571
ShaQ (Ours)	0.3296	0.1248	4.9285

Table 14: High ShaQ Max-Span Aleatoric Uncertainty Set (top 20%).

Method	Clarified	$\Delta H(\uparrow)$	Edit ($\downarrow$)
Semantic Entropy	0.3165	0.0654	6.7142
ICE	0.4872	-0.1051	7.1428
Deep Ensembles	0.4426	-0.0605	7.1428
ShaQ (Ours)	0.3159	0.0661	5.0476

Table 15: High Baseline Max-Total Aleatoric Uncertainty Set (top 30).

Method	Clarified	$\Delta H(\uparrow)$	Edit ($\downarrow$)
Semantic Entropy	0.3374	-0.0174	6.1333
ICE	0.3684	-0.0277	6.6333
Deep Ensembles	0.3050	0.0149	7.5000
ShaQ (Ours)	0.2848	0.0351	5.5333

Table 16: High Baseline Max-Total Aleatoric Uncertainty Set (top 20%).

Method	Clarified	$\Delta H(\uparrow)$	Edit ($\downarrow$)
Semantic Entropy	0.3489	0.0066	6.1428
ICE	0.3477	0.0207	6.7142
Deep Ensembles	0.3239	0.0315	7.8095
ShaQ (Ours)	0.3008	0.0547	5.8571

Table 17: High Baseline Uncertainty Disagreement Set.

Method	Clarified	$\Delta H(\uparrow)$	Edit ($\downarrow$)
Semantic Entropy	0.1529	0.0199	5.5714
ICE	0.1676	0.0052	6.0952
Deep Ensembles	0.1758	-0.0029	6.7619
ShaQ (Ours)	0.1355	0.0373	5.0476

- Uncertainty-Guided Clarification: Span-level uncertainty guidance achieves **the largest entropy reduction with the fewest edits**, enabling more efficient clarification.

Experiments (Medical Dialogues)

(a) Unclear Symptom Appearances

[Doctor] And is it, what color is it?

[Patient] I it looks pretty dark and and sometimes red.

-> "pretty dark" ($\phi=0.693$): 'pretty dark' could refer to dark purulent sputum or to blood-tinged/dark blood, which are clinically distinct and both supported by the context.

[Doctor] OK, so do you think there's blood in your phlegm?

(b) Unclear Symptom Timing

[Patient] Uh yeah, he does have a fever, um 39.5 I checked yesterday.

[Doctor] OK, uh is that is that the first time that um he was noted to have a fever yesterday?

-> "first time that um he was noted to have a fever yesterday" ($\phi=0.693$): It is ambiguous whether this refers to the first time the fever was observed/measured (but could have started earlier), or the first time the fever actually occurred (i.e., fever onset).

[Patient] Um he has had a fever for the past few days. Um it was around, it was less, between 37.5 to 38.5, um but then yesterday was just uh it just kept going up so I got concerned and thought I should bring him in.

Table 6: Quantitative results on MediTOD.

Method	F1 Score	AUROC	AUPRC
ICE	63.57	59.44	57.90
ShaQ (Ours, Total)	79.21	82.61	82.32
ShaQ (Ours, Max)	80.00	83.12	82.87
Follow-up Match	72.09%		

Clinical Example (Real-world medical dialogues): ShaQ localizes **clinically meaningful ambiguity** in patient utterances, enabling targeted doctor's follow-up questions.

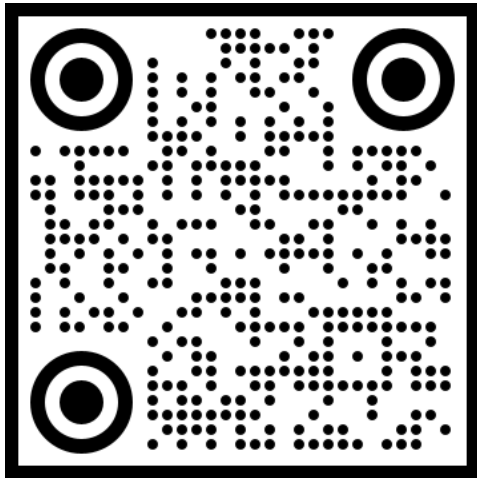
Summary

- We introduce **ShaQ**, a novel framework that identifies and measures uncertainty arising from specific spans in an input via Shapley values.
- Experiments show that **ShaQ** detects ambiguous inputs more effectively than existing methods and helps users resolve ambiguity with fewer edits.
- We also demonstrate its potential in medical conversations, where identifying unclear symptom descriptions can support more effective communication.
- Overall, **ShaQ** offers a practical step toward making LLMs more transparent, reliable, and easier to interact with.

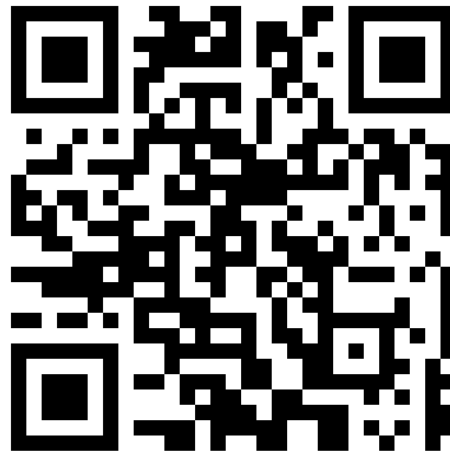
Future Works

1. **Lightweighting ShaQ via efficient Shapley value approximation**
2. **Simplifying the ShaQ pipeline while preserving its key theoretical properties**

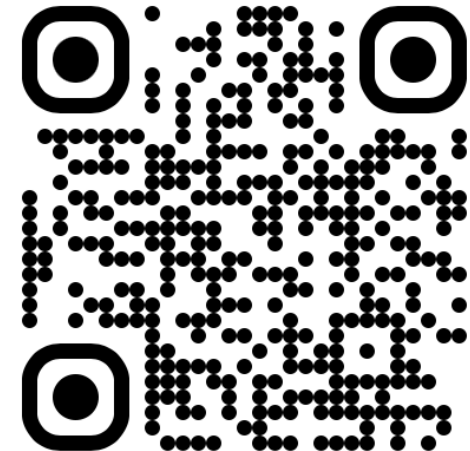
Thank you



LinkedIn(Seongjun Lee)



Personal Website(Suwan Yoon)



The Actionable Intelligence Lab 17