Weekly Study Report

Wang Ma

2024-12-09

Electrical, Computer, and Systems Engineering Department
Rensselaer Polytechnic Institute

Outline

1. Uncertainty Quantification form Input Space

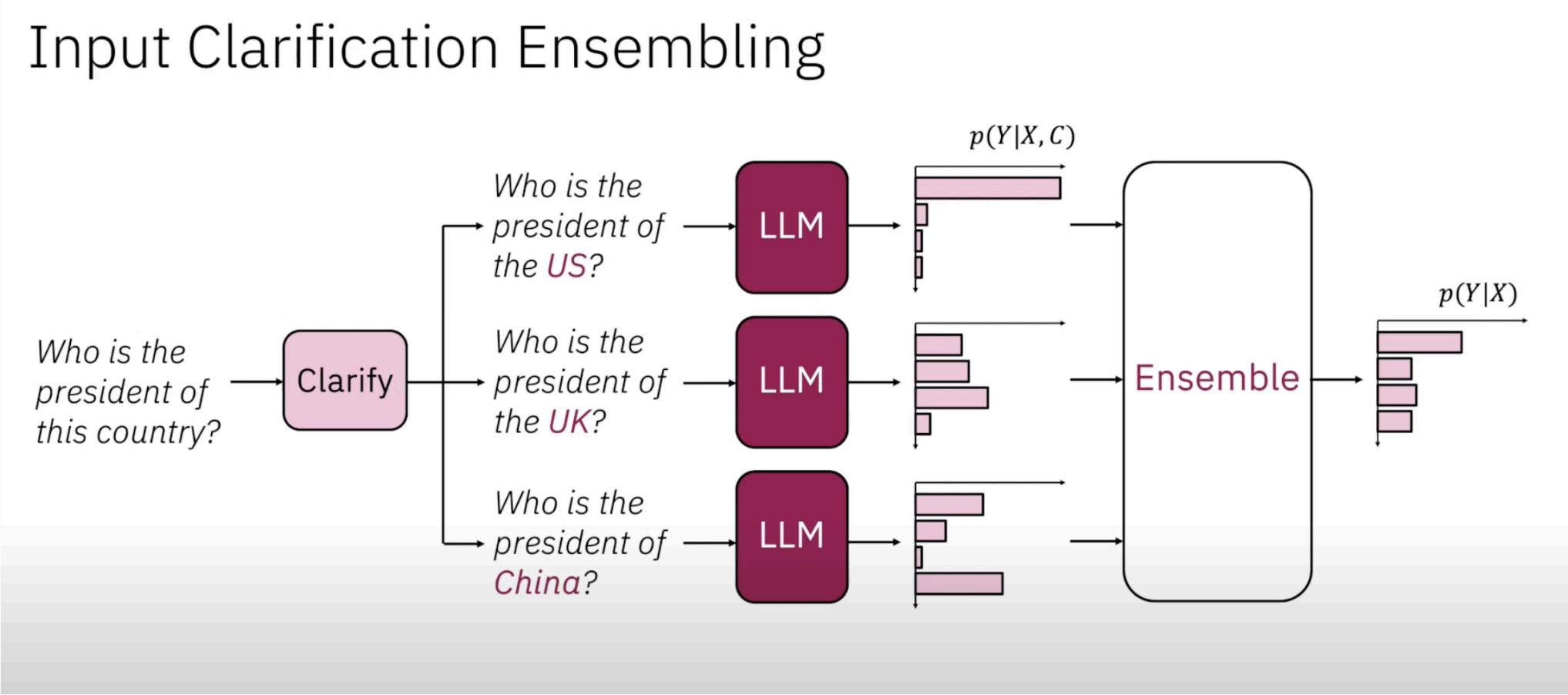
[ICML 2024] Decomposing Uncertainty for Large Language Models through Input Clarification Ensembling

1.1 Difference of UQ for LLMs

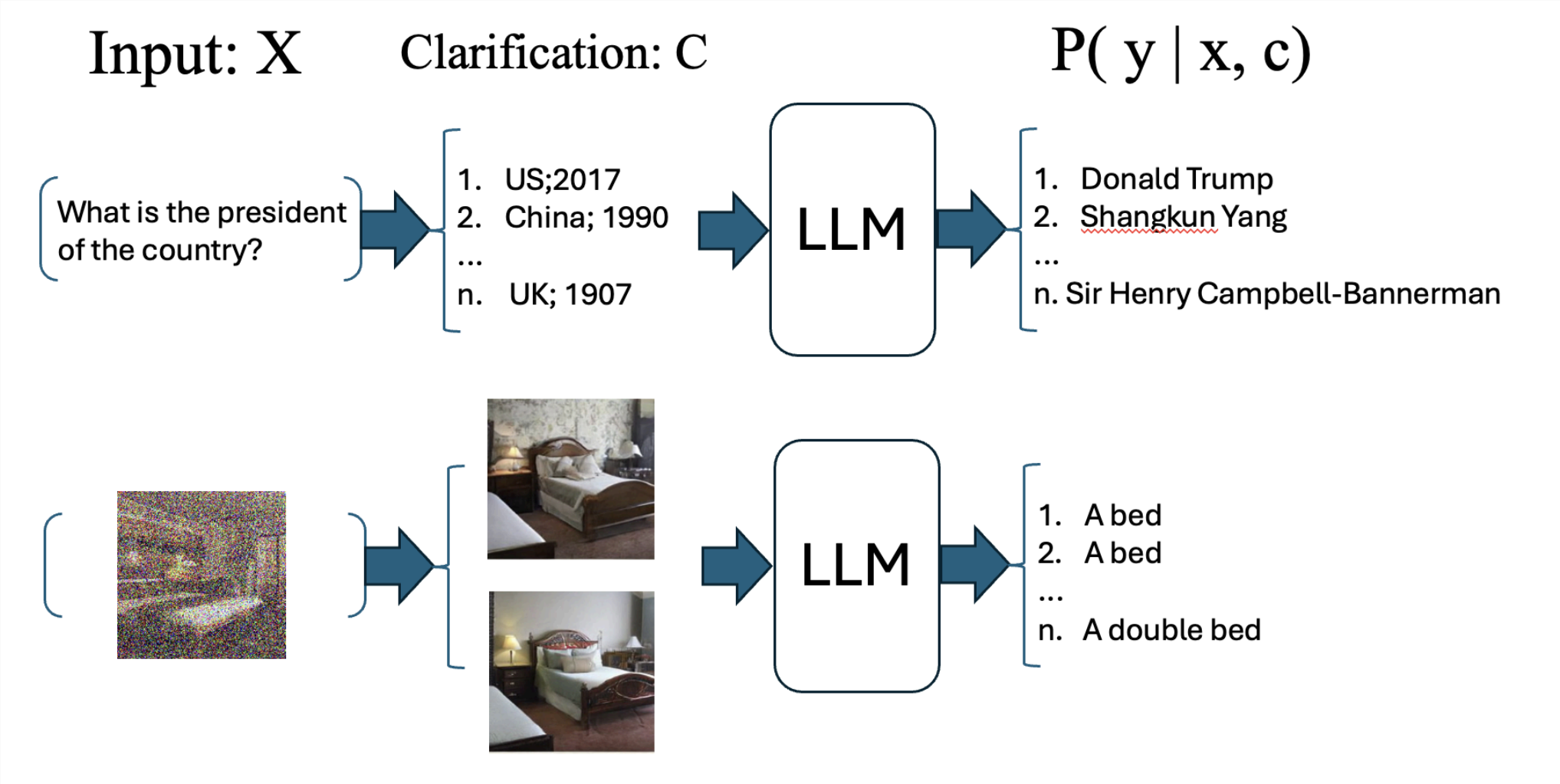
Difference: Current LLMs are trained on super-large dataset and have rich knowledge. While it's important to measure their Epistemic Uncertainty for specific task, their output can show high uncertainty if the input prompt is unclear/unspecific (or say, ambiguous), which can be considered a main source for Aleatoric Uncertainty.

In Image task, we train Ensemble models and treat their disagreement as the Epistemic Uncertainty. In this paper, they consider using the "Input Clarification Ensembling" and treat the disagreement in results as the Aleatoric Uncertainty.

The other motivation for the paper is that manipulating inputs is easy in Language Models.


Input Clarification Ensembling
Who is the president of this country?
Clarify
Who is the president of the US?
Who is the president of the UK?
Who is the president of China?
LLM
LLM
LLM
$p(Y|X,C)$
Ensemble
$p(Y|X)$

1.2 An Illustration for the Proposed Method



The prediction for the input X is

$$P(y|x) = \int_c P(y|x,c)P(c|x)dc,$$

where c is the clarification of x, can be generated by another LLM.

Then the Uncertainty Equation is

$$\mathbb{H}[Y|X] = \mathbb{I}[Y, C|\ X] + \mathbb{E}_C[\mathbb{H}[Y|X,C]]$$

.

Input Clarification Ensembling

$$\mathcal{H}[p(Y|X)] = I(Y;C|X) + \mathbb{E}_{p(C)}\left[\mathcal{H}[p(Y|X,C)]\right]$$

Total Uncertainty | Aleatoric Uncertainty | Epistemic Uncertainty

Deep Ensembles

$$\mathcal{H}[p(Y|X)] = I(Y;\theta|X) + \mathbb{E}_{p(\theta)}\left[\mathcal{H}[p(Y|X;\theta)]\right]$$

Total Uncertainty | Epistemic Uncertainty | Aleatoric Uncertainty

The flip is reasonable and understandable.

1. Aleatoric $\mathbb{I}[Y, C|\ X]$: If X is clear and Specific, then it should be low; if Y heavily depends on C, it means the row input X results in high uncertainty on the output.
2. Epistemic $\mathbb{E}_C[\mathbb{H}[Y|X, C]]$: If given X and its different kinds of clarifications $\{C_i\}_{i=1}^n$, $\mathbb{H}[y|x, c]$ still is high, this means the model is not familiar to the topic of X, this means the model need more knowledge about this topic.

1.4 Conclusion

Conclusion.

1. Input Clarification Ensembling is an Uncertainty Quantification/Decomposition algorithm better suited for LLMs.
2. Input Clarification Ensembling provides an idea of UQ for pre-trained model from the perspective of Input Space.

Limitation.

1. Single Model: Lacks of variation.
2. Lacks of Mathematical Proof: does the terms really refer to the corresponding uncertainties?
3. The input prompt X is considered as the only source of Aleatoric Uncertainty..

Thought.

1. Kind of, similar to Perturbation-based method or the "Test Time Augmentation" method.
2. Can it be added to the system of Bayesian Deep Learning? (using same modeling way)