# Every Opinion Matters: Evaluating and Building Models with Pluralistic Views

**Xiang Li**

Department of Computer Science
University of Pittsburgh
xianglli@pitt.edu

## Abstract

The development of large language models has demonstrated robust performance on English-centric benchmarks, which predominantly reflect majority opinions and dominant cultural norms. However, successful deployment in real-world applications requires the ability to handle context-specific and diverse knowledge, which is often underrepresented in training data. Addressing a plurality of perspectives is therefore essential. My research focuses on developing pluralistic evaluation methods to assess the diversity of LLM outputs, with a particular focus on culturally rich common-sense reasoning. Additionally, I work on advancing models that integrate diverse knowledge into LLMs, aiming to bridge the gap between human and AI understanding through the incorporation of varied perspectives using innovative probabilistic frameworks. In this talk, I will emphasize two key directions of my previous work: the probabilistic box model for representing diverse knowledge and probabilistic evaluation for assessing diversity in LLMs, with a focus on distributional aspects. Additionally, I will discuss my efforts to understand model behavior in long-tail scenarios.

**Probabilistic Box Model for Knowledge** Building probabilistic models for diverse knowledge is challenging due to the vast amount of information, making explicit categorization difficult. Traditional models like probabilistic graphical models struggle with large-scale distributions involving thousands or millions of events. While LLMs can generalize across large datasets, their inner workings are opaque, and suffer from producing inconsistent probabilities under different contexts. My previous work bridges these approaches by using probabilistic box embeddings, which represent joint probability distributions in a latent geometric space (Vilnis* et al. 2018). This method excels in modeling transitive relational data, such as word and sentence-level entailment (Li* et al. 2019), and is also effective for representing diverse knowledge.

**LLM Evaluation with Pluralistic Views** Most LLM evaluations are based on multiple-choice tasks, which, while reliable, lack the pluralistic complexity of real-world scenarios. In my previous work (Li et al. 2022), I systematically evaluated these benchmarks with large language models. I found that multiple-choice tests enable models to memorize answer choices from training data instead of reasoning with new contexts. My work (Boratko* et al. 2020) ProtoQA proposes a generative evaluation to evaluate all possible answers for a given question by comparing a ranked answer list. The work is still framed as a question-answering task. However, we explicitly take into account that multiple answers could be correct with varying likelihood. Later, we proposed to broaden the scope of ProtoQA with greater applicability to downstream tasks (Cheng et al. 2024). We consider a short context sentence and aim to fill in the implicit information by framing it as a question. We also proposed an automatic evaluation to compare diverse answer sets, defining a novel approach that directly measures the KL divergence between distributions. We justify the automatic evaluation with rigorous theoretical motivation and empirical results by demonstrating a high correlation $(0.73)$ with human scoring. Few-shot GPT4 only achieved a 0.68 KL score compared to 0.06 in terms of human performance.

**Ethical Considerations** We acknowledge that the answers may be biased towards certain populations. To reflect global perspectives, we are actively working on expanding the dataset's scope.



## References

Boratko*, M.; Li*, X. L.; O'Gorman*, T.; Das*, R.; Le, D.; and McCallum, A. 2020. ProtoQA: A Question Answering Dataset for Prototypical Common-Sense Reasoning. In *EMNLP. * Equal contribution*.

Cheng, Q.; Boratko, M.; Yelugam, P. K.; O'Gorman, T.; Singh, N.; McCallum, A.; and Li, X. 2024. Every Answer Matters: Evaluating Commonsense with Probabilistic Measures. In *ACL*.

Li, X. L.; Kuncoro, A.; d'Autume, C. d. M.; Blunsom, P.; and Nematzadeh, A. 2022. A Systematic Investigation of Commonsense Understanding in Large Language Models. In *EMNLP*.

Li*, X. L.; Vilnis*, L.; Zhang, D.; Boratko, M.; and McCallum, A. 2019. Smoothing The Geometry of Probabilistic Box Embeddings. In *ICLR. * Equal contribution*.

Vilnis*, L.; Li*, X. L.; Murty, S.; and McCallum, A. 2018. Probabilistic Embedding of Knowledge Graphs with Box Lattice Measures. In *ACL. * Equal contribution*.