Internal Presentation

Uncertainty Quantification and Decomposition of Large Language Model's In-Context Learning

Chen Ling

Project Update

10/17/2022

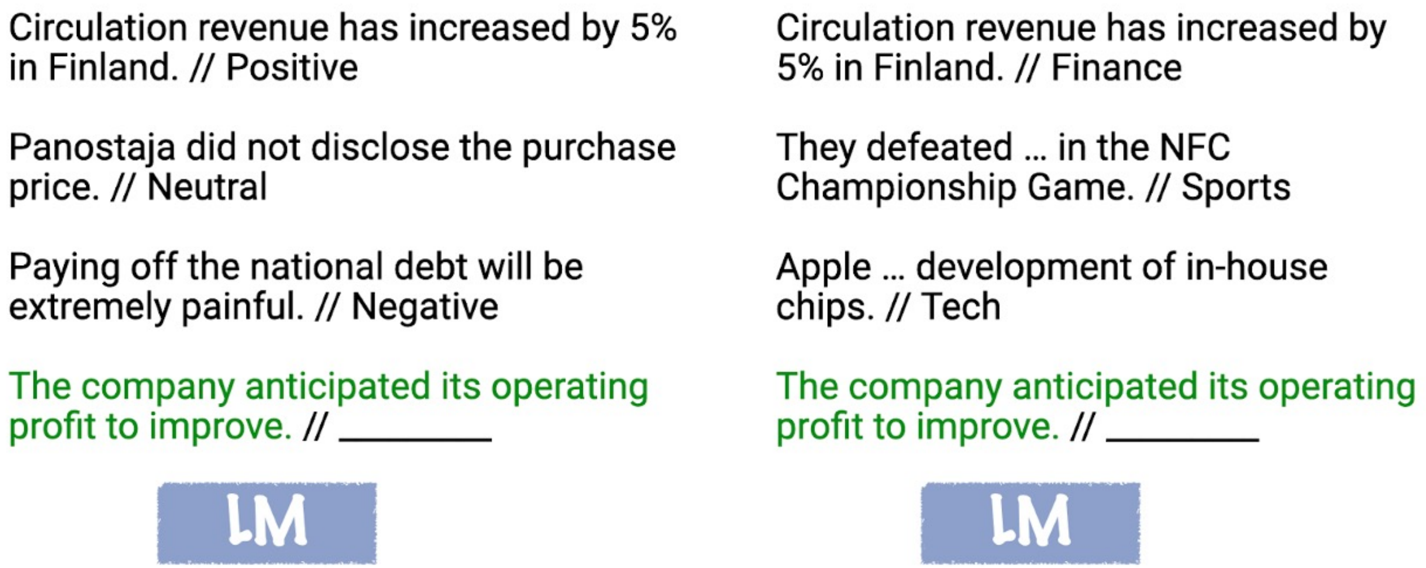
Agenda

- Introduction
- Research Target & Problem Formulation
- Proposed Solutions
- Existing Results & Future Plans

Background

The success of Large Language Models (LLMs) can be attributed to the emergent behavior: in-context learning.

- **In-context learning**: A frozen LM performs a task only by conditioning on the prompt text.
- **Few-shot Demonstration**: A few sentences consist of a list of input-output pairs that demonstrate a task.
- **What can in-context learning do?** On many NLP benchmarks, in-context learning is competitive with supervised models and is state-of-the-art on sentence completion and question answering competitions.



Uncertainty of In-context Learning

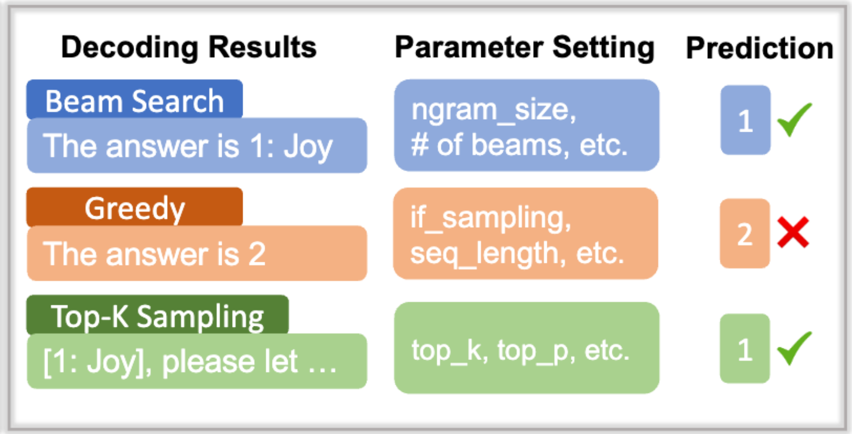
Can we always trust LLM's prediction despite the success of in-context learning?

- From input data's perspective, we may use inappropriate or insufficient few-shot demonstrations.

```
Classify the sentiment in the following text based on following
categories: [0: Sadness; 1: Joy, 2: Love; 3: Anger; 4: Fear].
Example #1: I didn't feel humiliated // 0: Sadness
Example #2: I've been feeling a little burdened // 0: Sadness
Example #3: I feel low energy I'm just thirsty // 0: Sadness

Test: I have the feeling she was amused and delighted
LLM Prediction: [2: Love] ✗
Ground Truth:   [1: Joy]  ✓
```

- From the model's perspective, the model configurations or hyperparameter setting are also uncertain.

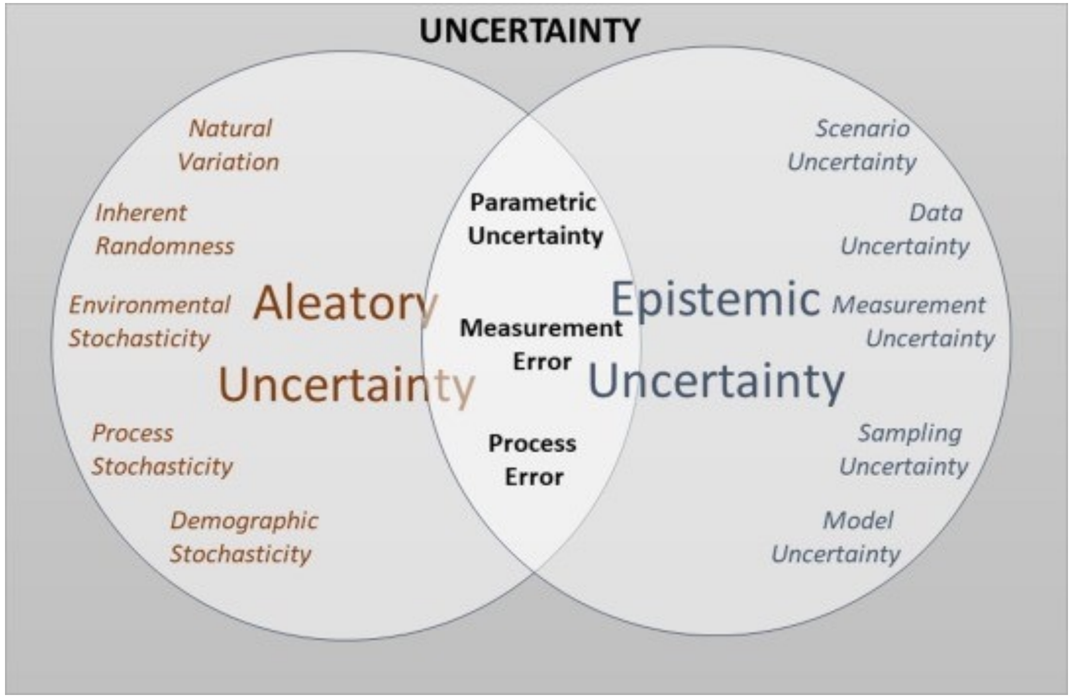


Various Decoding Methods ?

Various Hyperparameter Settings ?

Existing Works

- Existing methods tend to view the uncertainty as a whole value; however, it would make more sense to view both uncertainties individually.
- Decomposing uncertainties into distinct aleatoric and epistemic components is essential for informed decision-making when using LLMs.



Problem Formulation

- From the Bayesian view, LLM uses the in-context learning prompt to "locate" a previously learned concept to do the in-context learning task.

$$p(\mathbf{y}_T|\mathbf{x}_{1:T}) = \int_{z\in\mathcal{Z}} p(\mathbf{y}_T|\mathbf{x}_{1:T}, z)p(z|\mathbf{x}_{1:T})dz.$$

Task Output

T-shot demos

Latent concept

Circulation revenue has increased by 5% in Finland. // Positive

Panostaja did not disclose the purchase price. // Neutral

Paying off the national debt will be extremely painful. // Negative

The company anticipated its operating profit to improve. // ________

LM

Circulation revenue has increased by 5% in Finland. // Finance

They defeated … in the NFC Championship Game. // Sports

Apple … development of in-house chips. // Tech

The company anticipated its operating profit to improve. // ________

LM

- The predictive total uncertainty of LLMs can then be denoted as:

$$p(\mathbf{y}_T|\mathbf{x}_{1:T}) \approx \int p(\mathbf{y}_T|\Theta, \mathbf{x}_{1:T}, z)\cdot\underbrace{p(z|\mathbf{x}_{1:T})}_{\text{Data Uncertainty ?}}\underbrace{q(\Theta)}_{\text{Model Uncertainty ?}}dz d\Theta,$$

Entropy-based Uncertainty Decomposition

- Let $H(\boldsymbol{y}|\boldsymbol{x}_{1:T})$ be the entropy of a probability distribution that entangles both types of uncertainties.
- We typically have access only to a deterministic set of parameters denoted by Θ. We condition the equation on a specific realization of this parameter set, yielding

$$p(\mathbf{y}_T|\mathbf{x}_{1:T},\Theta) = \int p(\mathbf{y}_T|\mathbf{x}_{1:T},z,\Theta)p(z|\mathbf{x}_{1:T})dz \Longrightarrow H(\mathbf{y}_T|\mathbf{x}_{1:T},z,\Theta)$$

- The expected value of this entropy under different demonstration sets is then $\mathbb{E}_z\left[H(\mathbf{y}_T|\mathbf{x}_{1:T},z,\Theta)\right]$, which serves as a metric to quantify the **epistemic uncertainty** in coming from different z.
- The **aleatoric uncertainty** can subsequently be calculated as the discrepancy between the total uncertainty and the aleatoric uncertainty.

$$\begin{aligned} I(\mathbf{y}_T, z|\Theta) &= H\left(\mathbf{y}_T|\mathbf{x}_{1:T},\Theta\right) - \mathbb{E}_z\left[H(\mathbf{y}_T|\mathbf{x}_{1:T},z,\Theta)\right] \\ &\approx \sum^{M\times L} H(\mathbf{y}_T) - \frac{1}{M}\boxed{\sum_{m=1}^{M}}\boxed{\sum_{l=1}^{L}}\left[H(\mathbf{y}_T^{\Theta_{m,l}})\right], \end{aligned}$$

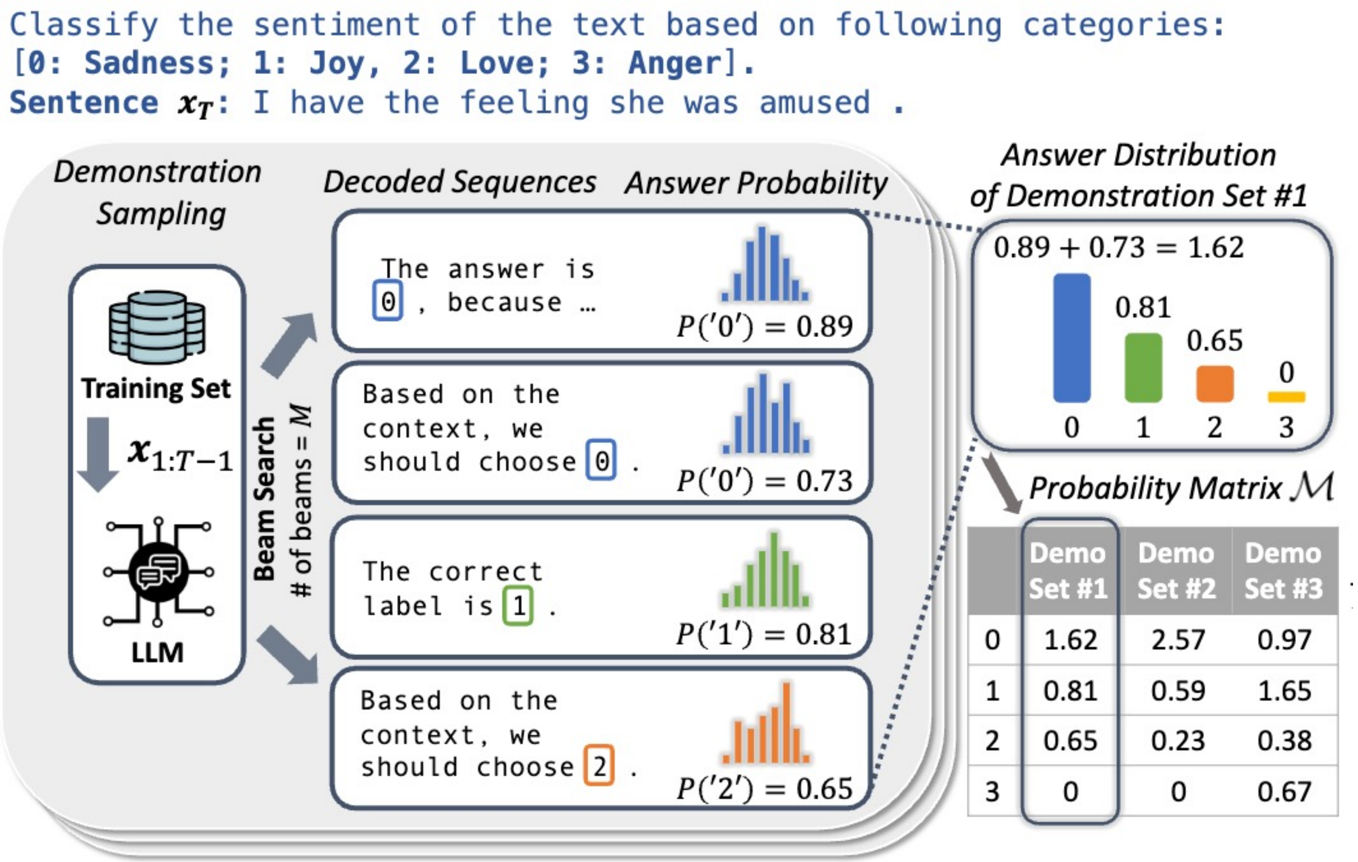
The number of different model configurations

The number of different demonstrations

Entropy Approximation

- In practice, the entropy is still hard to calculate due to following reasons.
  1. LLMs may not always be able to return feasible answers, i.e., the generation does not contain desired predictions.
  2. Not all tokens in the generated sequences are semantically equal, e.g., ‘ ’ and ‘-’.
  3. The length of generated sequences are not always the same.

- We propose a novel way to estimate the uncertainty given the distributions of the generated tokens $p(y_T)$
  1. Generating M sequences based on a set of $x_{1:T-1}$
  2. Selecting token(s) $\omega_t^{y_T}$ that directly answers the question.
  3. Aggregating the token probabilities of M sequences as a distribution of predicted labels.
  4. Iterating the process L times with different demonstration sets and form a probability matrix $\mathcal{M}$.



Experiment Setup

- Evaluation: We leverage Area under Precision-Recall Curve (AUPR) and AUROC (ROC) based on the accuracy of the prediction and the quantified uncertainty scores.

- Tasks: we select three representative natural language understanding tasks.
  - ***Sentiment Analysis***: 1) Emotion: 6-way classification; 2) Financial Phrasebank: 3-way classification; 3) Stanford Sentiment Treebank v2 (SST-2): binary classification.
  - ***Linguistic Acceptability***: The Corpus of Linguistic Acceptability (COLA): binary classification
  - ***Topic Classification***: AG_News: 4-way classification.

- We consider beam search (beam width = 10) to sample different model outputs. In this work, we sample four sets of demonstrations with two demonstration selection strategies:
  - 1) Randomly selecting a given number of training samples from the training data
  - 2) Selecting k samples per class from the training data

Quantitative Analysis

We first compare different methods in assessing the misclassification samples, where misclassified samples should have a higher uncertainty score.

1. As shown in the table, our uncertainty decomposition (EU and AU) can serve as better indicators to identify misclassified samples.

2. **Class sampling strategy** can yield better performance across all datasets than random demonstration sampling.

3. **Larger models** (moving from 7B to 70B) tend to have better performance.

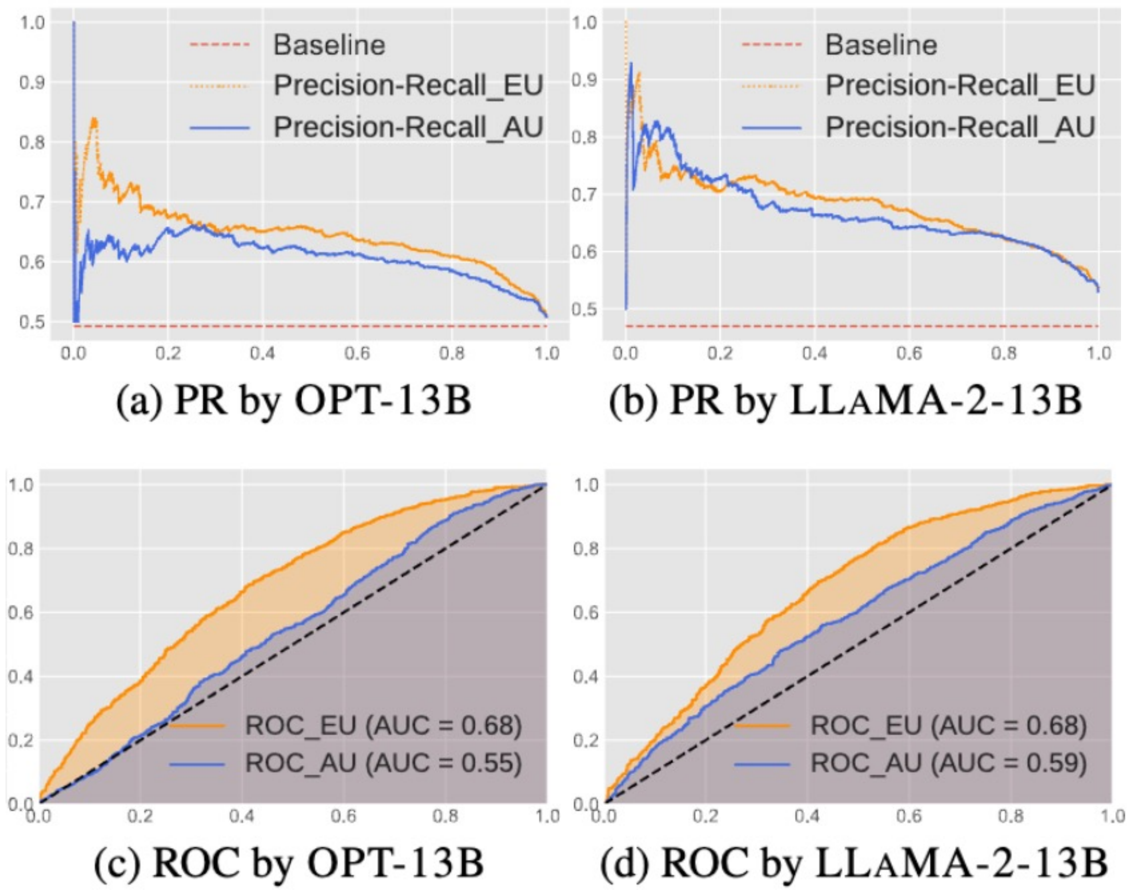
4. Treating all tokens **equally** can be harmful in uncertainty quantification.

	Inference Model	ACC	Likelihood		Entropy		Semantic		Ours (EU)		Ours (AU)	
			AUPR	ROC	AUPR	ROC	AUPR	ROC	AUPR	ROC	AUPR	ROC
Emotion	LLAMA-7B-RANDOM	0.407	0.423	0.426	0.448	0.501	0.598	0.607	**0.688**	**0.667**	0.625	0.579
	LLAMA-7B-CLASS	0.411	0.562	0.423	0.657	0.538	0.697	0.653	**0.745**	**0.696**	0.691	0.601
	LLAMA-13B-RANDOM	0.501	0.597	0.613	0.584	0.503	0.612	0.625	**0.645**	**0.681**	0.559	0.585
	LLAMA-13B-CLASS	0.533	0.641	0.578	0.593	0.554	0.652	0.701	**0.622**	**0.686**	0.526	0.599
	LLAMA-70B-RANDOM	0.584	0.512	0.462	0.491	0.452	0.657	0.696	**0.667**	**0.713**	0.531	0.663
	LLAMA-70B-CLASS	0.592	0.537	0.484	0.469	0.442	0.622	0.689	**0.659**	**0.721**	0.612	0.693
Financial	LLAMA-7B-RANDOM	0.379	0.821	0.532	0.728	0.438	0.715	0.624	**0.731**	**0.672**	0.669	0.582
	LLAMA-7B-CLASS	0.397	0.593	0.505	0.548	0.362	0.732	0.699	**0.803**	**0.711**	0.753	0.589
	LLAMA-13B-RANDOM	0.476	0.894	0.571	0.652	0.463	0.705	0.545	0.718	0.512	**0.729**	**0.573**
	LLAMA-13B-CLASS	0.477	0.752	0.594	0.692	0.531	0.694	0.543	**0.765**	**0.610**	0.758	0.592
	LLAMA-70B-RANDOM	0.530	0.816	0.509	0.754	0.493	0.679	0.688	**0.779**	**0.754**	0.734	0.642
	LLAMA-70B-CLASS	0.537	0.668	0.469	0.623	0.439	0.774	0.649	**0.893**	**0.804**	0.739	0.659
SST-2	LLAMA-7B-RANDOM	0.856	0.149	0.636	0.135	0.587	0.244	0.593	**0.286**	0.683	0.205	**0.702**
	LLAMA-7B-CLASS	0.897	0.230	0.666	0.196	0.579	0.253	0.577	0.248	**0.701**	**0.302**	0.673
	LLAMA-13B-RANDOM	0.866	0.268	0.472	0.204	0.467	**0.355**	0.712	0.314	0.677	0.326	**0.816**
	LLAMA-13B-CLASS	0.928	0.178	0.425	0.113	0.439	0.343	0.631	**0.397**	**0.836**	0.367	0.639
	LLAMA-70B-RANDOM	0.932	0.091	0.597	0.137	0.475	0.258	0.565	**0.318**	**0.764**	0.298	0.571
	LLAMA-70B-CLASS	0.938	0.132	0.552	0.185	0.531	0.312	0.679	0.331	**0.851**	**0.362**	0.697
COLA	LLAMA-7B-RANDOM	0.599	0.388	0.557	0.329	0.443	0.358	0.502	**0.416**	**0.562**	0.377	0.517
	LLAMA-7B-CLASS	0.639	0.392	0.523	0.381	0.478	0.425	0.526	**0.473**	**0.587**	0.401	0.506
	LLAMA-13B-RANDOM	0.652	0.389	0.498	0.287	0.512	0.433	0.562	0.469	**0.572**	**0.488**	0.565
	LLAMA-13B-CLASS	0.649	0.412	0.418	0.342	0.517	0.426	0.548	0.456	0.568	**0.523**	**0.641**
	LLAMA-70B-RANDOM	0.826	0.481	0.599	0.312	0.471	0.372	0.625	0.317	**0.716**	**0.329**	0.676
	LLAMA-70B-CLASS	0.852	0.357	0.612	0.397	0.588	0.397	0.613	0.389	**0.727**	**0.425**	0.682
AG_News	LLAMA-7B-RANDOM	0.646	0.238	0.472	0.265	0.463	0.312	0.612	**0.448**	**0.634**	0.361	0.537
	LLAMA-7B-CLASS	0.679	0.267	0.505	0.372	0.523	0.378	0.562	**0.384**	**0.627**	0.326	0.538
	LLAMA-13B-RANDOM	0.685	0.365	0.517	0.364	0.522	0.374	0.548	**0.395**	**0.648**	0.378	0.552
	LLAMA-13B-CLASS	0.685	0.378	0.528	0.359	0.413	0.411	0.566	**0.429**	**0.654**	0.401	0.569
	LLAMA-70B-RANDOM	0.792	0.311	0.478	0.316	0.498	**0.401**	0.552	0.309	**0.635**	0.319	0.543
	LLAMA-70B-CLASS	0.838	**0.302**	0.511	0.271	0.528	0.354	0.532	0.274	**0.662**	0.283	0.571

Generalization Capability

We evaluate the performance of misclassification rate using two backbone LLMs: OPT-13B and LLAMA-2-13B on EMOTION dataset.

1. Our method exhibits consistent trends across different LLMs, where the PR curve of both uncertainties ((a) and (b)) between the two methods are almost identical.
2. The ROC curves of different LLMs ((c) and (d)) show a similar pattern, with the AUC scores not deviating significantly.
3. Since LLAMA-2-13B is a more powerful LLM than OPT-13B, our method can quantify that EU of LLAMA-2-13B (AUROC = 0.59) is better than OPT-13B (AUROC = 0.55).



(a) PR by OPT-13B (b) PR by LLAMA-2-13B

(c) ROC by OPT-13B (d) ROC by LLAMA-2-13B

Out-of-domain (OOD) Demonstration Detection

We conduct OOD detection on the EMOTION dataset:

1. In-domain demonstrations (sampled from its training set)
2. Relevant demonstrations (sampled from Finance Phrasebank, a three-class sentiment analysis task)
3. OOD demonstrations (sampled from COLA binary classification dataset)

	Semantic		Ours (EU)		Ours (AU)	
	AUPR	ROC	AUPR	ROC	AUPR	ROC
Relevant Demo	0.702	0.644	**0.742**	**0.935**	0.657	0.682
OOD Demo	0.698	0.712	**0.784**	**0.941**	0.773	0.607

Table 2: Out-of-domain demonstration detection conducted with LLAMA-2-13B on EMOTION Dataset.

As shown in Table 2, compared to the state-of-the-art Semantic Uncertainty and the AU, the EU demonstrates the best indicator to detect both less relevant and OOD demonstrations.

OOD demonstration refers to coupling a test instance with less relevant or OOD demonstrations, potentially leading the model to be misled and handle the test instance unreliably.

Semantic Out-of-distribution Detection

In this study, we mask out a few classes and ask LLMs to classify test samples into the rest of the classes. The he method is expected to return a higher uncertainty score of SOOD test samples.

- We mask two classes [1: sadness; 2: anger] from the EMOTION dataset and ask LLM to categorize samples into the rest classes. SOOD samples are labeled as 1 and other samples are labeled as 0.
- EU still performs the best as a better indicator to recognize SOOD samples across various model sizes.
- Given the inappropriate task description and demonstrations, AU may not necessarily perform better in the presence of SOOD samples.

	Semantic		Ours (EU)		Ours (AU)	
	AUPR	ROC	AUPR	ROC	AUPR	ROC
7B	0.477	0.532	**0.548**	**0.658**	0.461	0.570
13B	0.417	0.468	**0.525**	**0.592**	0.414	0.437

Table 3: Semantic out-of-distribution detection using LLAMA-2 7B and 13B on EMOTION Dataset.

Semantic out-of-distribution (SOOD) detection refers to distinguishing test samples with semantic shifts from the given demonstrations and the prompt.

Case Study

- For 7B model, by presenting LLMs with more diverse demonstrations (containing both positive and negative sentences), the results would be more diverse between different beam search returned sequences, leading to a relatively higher AU than EU.

- For 70B model, with a larger model capability, both EU and AU are significantly reduced, which indicates the model is more confident in the generated output and the variation of data may not influence much to the prediction.

Testing Query: I had stated to her the reason I feel so fearful is because I feel unsafe (4: fear)		Extracted Predictions	EU	AU
LLaMA-2-7B	1. i felt anger when at the end of a telephone call (3: anger) 2. i feel a little mellow today (1: joy) 3. i don t feel particularly agitated (4: fear) 4. i hate it when i feel fearful for absolutely no reason (4: fear) 5. im updating my blog because i feel shitty (0: sadness)	0, 0, 0, 1, 3 4, 3, 4, 4, 4	0.171	0.372
	1. i am feeling outraged it shows everywhere (4: fear) 2. i do feel insecure sometimes but who doesnt (4: fear) 3. i start to feel emotional (0: sadness) 4. i feel so cold a href http irish (3: anger) 5. i feel i have to agree with her even though i can imagine some rather unpleasant possible cases (0: sadness)	4, 4, 1, 3, 4 4, 4, 4, 5, 4	0.163	0.189
LLaMA-2-70B	1. i felt anger when at the end of a telephone call (3: anger) 2. i feel a little mellow today (1: joy) 3. i don t feel particularly agitated (4: fear) 4. i hate it when i feel fearful for absolutely no reason (4: fear) 5. im updating my blog because i feel shitty (0: sadness)	4, 3, 4, 3, 4 4, 4, 2, 4, 4	0.012	0.079
	1. i am feeling outraged it shows everywhere (4: fear) 2. i do feel insecure sometimes but who doesnt (4: fear) 3. i start to feel emotional (0: sadness) 4. i feel so cold a href http irish (3: anger) 5. i feel i have to agree with her even though i can imagine some rather unpleasant possible cases (0: sadness)	4, 4, 4, 4, 4 4, 4, 4, 4, 4	0.004	0.009

Conclusion

- This work provide a novel approach to decompose the predictive uncertainty into its aleatoric and epistemic perspectives from the Bayesian perspective.

- A novel approximation method to quantify different uncertainties based on the decomposition is also proposed.

- Extensive experiments , including quantitative analysis, generalization analysis, and case studies, are conducted to verify the effectiveness and better performance of the proposed method than others.

Orchestrating a brighter world

NEC

NEC Laboratories **America**