Uncertainty Quantification for In-Context Learning of Large Language Models NAACL 2024

Chen Ling, Xujiang Zhao, Xuchao Zhang, Wei Cheng, Yanchi Liu

February 24, 2026

YunSeop Shin, Seoul national university, statistics, IDEA LAB

Uncertainty Quantification in Deep Learning

- We focus on classification case which is $y \in \{0, 1, \cdots, K-1\}$ and $x \in \mathbb{R}^d$

$$p(y = k|x, \theta) = \mathrm{softmax}(f_\theta(x))_k. \tag{1}$$

- Using the estimated $\hat{\theta}$, a possible desired behavior of a model would be to return a prediction while also indicating that the point lies outside of the data distribution used to estimate $\hat{\theta}$.
- 이는 우리가 모델 불확실성(epistemic uncertainty)을 quantification하고싶어하는 동기가 되었다.

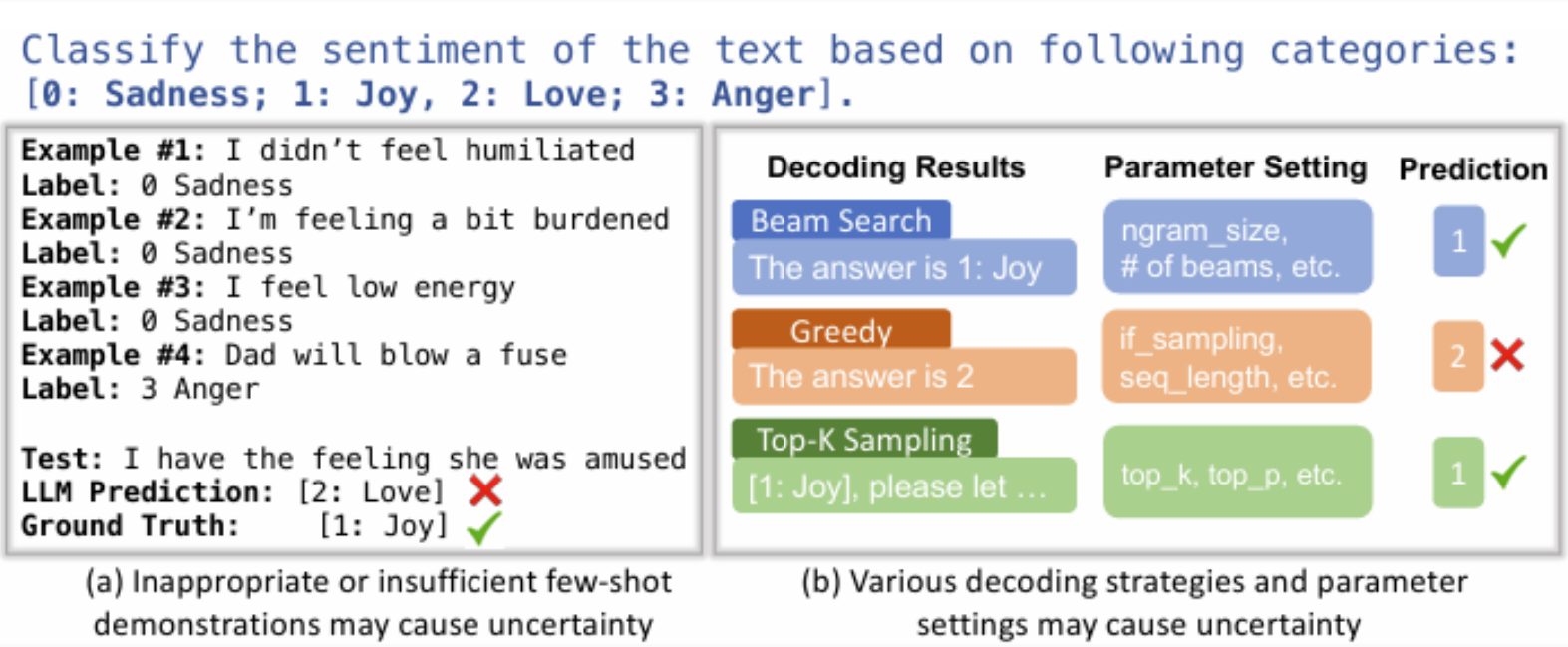
Uncertainty Quantification in Deep Learning

- 하지만 위와같은 이유로 생기는 불확실성 외에도, 데이터 자체에 내제된 불확실성(data uncertainty, aleatoric uncertainty)도 있으므로 이를 분리해서 quantification 해야한다.
- 그러기 위해서 기존의 방법은 엔트로피를 사용해서 불확실성을 측정하는데, 이때 total uncertainty는 $H[p(y|x, \mathcal{D}_{train})]$, 즉 (사후)예측분포에서 앤트로피로 측정되며, aleatoric uncertainty는 $\mathbb{E}_{\theta|\mathcal{D}_{train}}[H[p(y|x, \theta)]]$으로 측정될 수 있다.
- 여기서 $\theta|\mathcal{D}_{train}$의 분포와 예측분포를 근사하기 위해서 MC-dropout, Laplace approximation, Variation inference 등의 방법들이 사용된다.

Notation

- $[x_1, x_2, \cdots, x_{T-1}]$: In-context demonstrations containing both questions and answers.
- x_T : Test question without the task answer.
- y_T : Response to the test question x_T.
- $z \sim p(z|x_{1:T})$: Latent variable (concept).
- θ : LLM parameters/configurations (e.g. temperature).
- $q(\theta)$: The approximated posterior of the LLM's parameter θ.
- $H(p) = -\sum_{k=1}^{K} p_k \log p_k$ for $p \in \mathbb{R}^K$.

Example



(a) Inappropriate or insufficient few-shot demonstrations may cause uncertainty

(b) Various decoding strategies and parameter settings may cause uncertainty

Example

System Prompt	### Below is an instruction that describes a task. Clearly follow the instruction and write a short response to answer it.
Task Description	### Instruction: Classify the sentiment in the following text based on the six categories: [0: Sadness; 1: Joy, 2: Love; 3: Anger; 4: Fear, 5: Surprise]. Provide the information in a structured format WITHOUT additional comments, I just want the numerical label for each text.
Demonstrations	### Here are some examples: Example 1: Sentence: {i didnt feel humiliated} Category: {0: Sadness} Example 2: Sentence: {im grabbing a minute to post i feel greedy wrong} Category: {3: anger} Example 3: Sentence: {i have the feeling she was amused and delighted} Category: {1: joy} Example 4: Sentence: {i feel more superior dead chicken or grieving child} Category: {1: joy} Example 5: Sentence: {i get giddy over feeling elegant in a pencil skirt} Category: {1: joy} ...
Test Query	### Test Sentence: {} Category:

Predictive Distribution

- Authors formulate the predictive distribution of in-context learning (ICL) for predicting y_T given few-shot examples $x_{1:T-1}$ and a test question x_t as:

$$p(y_T|x_{1:T}) \approx \int p(y_T|\theta, x_{1:T}, z) \cdot p(z|x_{1:T})q(\theta)d\theta dz \tag{2}$$

- To obtain probability $p(y_T|\theta, x_{1:T}, z)$, the authors use White-box LLMs such as LLaMA/OPT.

Uncertainty-Decompose

- Total Uncertainty: $H(p(y_T|x_{1:T}, \theta))$
- Epistemic Uncertainty: $\mathbb{E}_z\left[H(p(y_T|x_{1:T}, z, \theta))\right]$
  $\rightarrow$ Concept z가 고정되어있을때 예측분포 $p(y_T|x_{1:T}, z, \theta)$가 얼마나 멀리 퍼져있는지에 대한 평균.
- Aleatoric Uncertainty:

$$\begin{aligned} I(y_T, z|x_{1:T}, \theta) =& H(p(y_T|x_{1:T}, \theta)) \\ & -\mathbb{E}_z\left[H(p(y_T|x_{1:T}, z, \theta))\right] \end{aligned} \tag{3}$$

  $\rightarrow$ Concept z가 바뀌었을때 예측이 얼마나 퍼져있는지 고려함.
- LLM의 hyperparameter, configuration (temperature, decoding 방법) 등은 고정시켜 놓은 상황에서의 불확실성을 고려하고 싶다??

How to approximate?

- To approximate the term $H(p(y_T|x_{1:T}, \theta))$, $\mathbb{E}_z \left[H(p(y_T|x_{1:T}, z, \theta))\right]$, we need to sample z from $p(z|x_{1:T})$.
- But the problem is that we don't know the $p(z|x_{1:T})$. So authors sample L many $x_{1:T-1}$. Authors think that for each sampled example $x^{(l)}_{1:T-1}$, $l = 1, \cdots, L$ has corresponding concept z^l.
- Also, authors claim that they sample M many θ from $q(\theta)$ (But I don't think it is sampling from the approximate posterior). They said that they do this by beam search with beam width $= 10$.

Approximate

- Let $p_{m,l}(k)$ as follows

$$p_{m,l}(k) := p\left(y_T = k \mid x_{1:T}^{(l)}, z^l, \theta_m\right), \quad k \in 0, \ldots, K-1 \quad (4)$$

- From this, we can obtain $\hat{y}_{m,l} = \mathrm{argmax}_k p_{m,l}(k)$ and $p^\star_{m,l} = \max_k p_{m,l}(k)$.
- Define pseudo probability matrix $\mathcal{M}$:

$$\mathcal{M}_{k,l} = \sum_{m=1}^{M} p^\star_{m,l} I(\hat{y}_{m,l} = k). \quad (5)$$

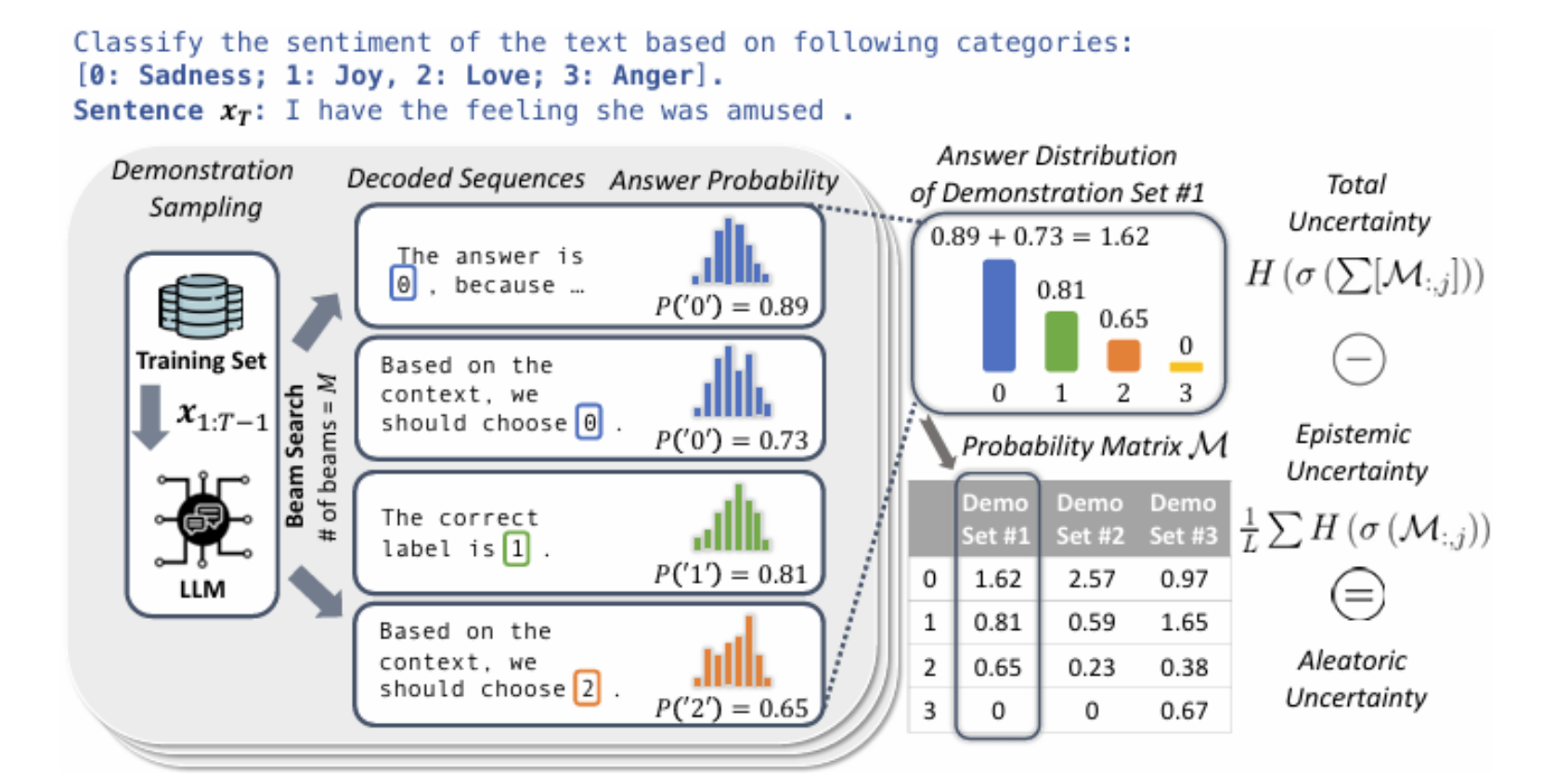
Approximate

- Now, let $\hat{p}^l = \sigma(\mathcal{M}_{,l})$ and $\hat{p}^{total} = \sigma(\sum_{l=1}^{L} \mathcal{M}_{,l})$.
- From this, author approximated each uncertainty as follows:

$$H(p(y_T|x_{1:T},\theta)) \approx H(\hat{p}^{total}) = H\left(\sigma(\sum_{l=1}^{L}\mathcal{M}_{,l})\right),$$
$$\mathbb{E}_z\left[H(p(y_T|x_{1:T},z,\theta))\right] \approx \frac{1}{L}\sum_{l=1}^{L} H(\hat{p}^l) = \frac{1}{L}\sum_{l=1}^{L} H\left(\sigma(\mathcal{M}_{,l})\right). \tag{6}$$

where, $\sigma : \mathbb{R}^K \to \mathbb{R}^K$

Overall FrameWork



Problem

- 최소한 저 근사가 성립하기 위해서는,
  1. $\sigma(\sum_{l=1}^{L} \mathcal{M}_{,l})$ 이 $p(y_T|x_{1:T}, \theta)$을 근사하고,
  2. $\sigma(\mathcal{M}_{,l})$이 $p(y_T|x_{1:T}, z, \theta)$를 근사해야 한다.
- 그런데, $\sigma(\sum_{l=1}^{L} \mathcal{M}_{,l})$ 이건 뭐 monte-calro도 아니고, 그냥 L 개 sample 뽑은거에서 다 더해서 정규화시킨것인데 이게 왜 $p(y_T|x_{1:T}, \theta)$를 근사한다는지도 모르겠고,
- $\sigma(\mathcal{M}_{,l})$이 $p(y_T|x_{1:T}, z, \theta)$를 근사한다는것은 더 이해가 안 됨. θ에 대해서 조건부로 고정시켜놨는데, θ도 여러개를 뽑은 다음에, 거기서 예측확률값만 구해서 그걸 더한 후, 정규화해서 $p(y_T|x_{1:T}, z, \theta)$라고 근사시키는데 이게 왜 가능한지 이해하지 못하겠음...

Reference

- Ling, C., Zhao, X., Zhang, X., Cheng, W., Liu, Y., Sun, Y., ... & Chen, H. (2024). Uncertainty quantification for in-context learning of large language models. arXiv preprint arXiv:2402.10189.

More...

Decomposing Uncertainty for Large Language Models through Input Clarification Ensembling (ICML 2024)

- Pre-trained LLM 예측의 total uncertainty를 입력의 불확실성(aleatoric uncertainty)와 모델 불확실성(epistemic uncertainty)로 분해하여 계량하는 방법을 제안함.
- 주어진 입력 X에 대해서 clarifications $C^{(k)}$를 생성해, LLM의 입력으로 $X \oplus C^{(k)}$를 사용함. 그 뒤 결과값을 앙상블하여 불확실성을 계량함.

$$\mathcal{H}(q(\boldsymbol{Y}|\boldsymbol{X})) = \underbrace{\mathcal{I}(\boldsymbol{Y};\boldsymbol{C}|\boldsymbol{X})}_{①'} + \underbrace{\mathbb{E}_{q(\boldsymbol{C}|\boldsymbol{X})}\mathcal{H}(q(\boldsymbol{Y}|\boldsymbol{X}\oplus\boldsymbol{C}))}_{②'}.$$

- ②′은 clarification이 주어졌을 때의 불확실성이므로 모델 불확실성을 의미하고, ①′은 모델 불확실성을 구하고 남은 변동이므로 입력에 대한 불확실성을 의미한다. 여기서 $\mathcal{H}(p) = -\sum_{k=1}^{K} p_k \log p_k$이다.

More...

Position: Uncertainty Quantification Needs Reassessment for Large Language Model Agents (ICML 2025)

- LLM은 전통적인 aleatoric/epistemic uncertainty가 개념적, 실용적 한계가 크니, 새로운 방향의 불확실성을 재정의하자고 제안함.
- **Underspecification Uncertainties**
  1. 사용자가 필요한 맥락을 충분히 정의하지 않아서 생기는 불확실성.
- **Interactive Learning**
  1. LLM과 사용자가 추가 질의를 주고받으며, 처음 입력 x에 대한 불확실성을 줄일 수 있음.
- **Output Uncertainties**
  1. LLM이 단순 답변만 출력하지 않고, '무엇이 경쟁 가설인지, 왜 헷갈리는지, 무엇이 불확실성을 줄이는지'를 함께 전달해야 한다는 방향성임. 논문의 저자들은 이것을 conformal prediction의 prediction set이나 Bayesian credible interval의 자연어 확장 버전이라고 설명함.

More..

C-LoRA: Contextual Low-Rank Adaptation for Uncertainty Estimation in LLMs (NeurIPS 2025)

- LoRA의 베이지안 방법을 제안함. LoRA 전체 가중치가 아닌, 기존 LoRA를 다음과 같이 더 작게 분해한 함.

$$W = W_0 + \Delta W = W_0 + BE_xA \tag{7}$$

- 이 때, 기존 LoRA는 $\Delta W = BA$로 분해함. E만 확률변수로 보고 A,B는 deterministic하게 학습함.
- E의 분포를 입력 데이터에 의존적으로 모델링함. 그래서 저자들은 이를 Contextual Low-Rank Adaptation이라 명명함.

More...

Uncertainty as Feature Gaps: Epistemic Uncertainty Quantification of LLMs in Contextual Question-Answering (ICLR 2026)

- 기존 불확실성 계량화 연구가 closed-book QA에 치우친 한계를 지적하며, context QA에서 불확실성을 이론적으로 정의하고 계량하려 함.
- 이상적인 모델과 실제 모델의 예측분포간 KL divergence를 통해 epistemic uncertainty를 정의하고 이를 근사하는 방법을 제안함.
- 이 논문은 LLM이 답변의 불확실성을 점수로 매겼을 때, 그 점수가 실제로 오답인 경우에 더 높게 나와서 "보류해야 할 문제"를 잘 골라내는지를 평가함. 이를 위해 불확실성 점수만으로 정답과 오답을 구분하는 성능을 AUROC로 측정함.

More...

Textual Bayes: Quantifying Uncertainty in LLM-Based Systems (ICLR 2026)

- 프롬프트를 통계모형의 parameters로 간주하고 이에 대한 Bayesain Inference를 수행함.
- Parameter로 간주된 프롬프트 θ의 사전분포는 LLM으로 구성하고, 가능도함수 역시 LLM의 결과값 확률로 구한다.
- 이때 프롬프트 θ를 사후분포부터 추출하기 위해서 새로운 MCMC알고리즘을 제안한다.