Quantifying Uncertainty in Natural Language Explanations of Large Language Models for Question Answering

Yangyi Li, Mengdi Huai
Department of Computer Science, Iowa State University
{liyangyi, mdhuai}@iastate.edu

Abstract

Large language models (LLMs) have shown strong capabilities, enabling concise, context-aware answers in question answering (QA) tasks. The lack of transparency in complex LLMs has inspired extensive research aimed at developing methods to explain large language behaviors. Among existing explanation methods, natural language explanations stand out due to their ability to explain LLMs in a self-explanatory manner and enable the understanding of model behaviors even when the models are closed-source. However, despite these promising advancements, there is no existing work studying how to provide valid uncertainty guarantees for these generated natural language explanations. Such uncertainty quantification is critical in understanding the confidence behind these explanations. Notably, generating valid uncertainty estimates for natural language explanations is particularly challenging due to the auto-regressive generation process of LLMs and the presence of noise in medical inquiries. To bridge this gap, in this work, we first propose a novel uncertainty estimation framework for these generated natural language explanations, which provides valid uncertainty guarantees in a post-hoc and model-agnostic manner. Additionally, we also design a novel robust uncertainty estimation method that maintains valid uncertainty guarantees even under noise. Extensive experiments on QA tasks demonstrate the desired performance of our methods.

1 Introduction

Large language models (LLMs) such as GPT-4 have recently achieved impressive gains in natural-language understanding and generation, demonstrating near-human fluency across a wide range of tasks (Achiam et al., 2023). When adapted for open-domain question answering (QA), these models exploit their vast parametric knowledge to deliver concise, context-aware answers that surpass traditional retrieval pipelines in reasoning depth and coverage (Kwiatkowski et al., 2019; Chen et al., 2025b). However, in LLM-based QA systems, the underlying LLMs are complex models where their inner working mechanisms are not yet fully understood. This lack of interpretability poses a significant barrier to their deployment in high-stakes decision-making applications, where inappropriate guidance can have severe consequences.

Given the importance of explaining LLMs' behaviors, many interpretation methods have been proposed (Zhu et al., 2024). Among them, natural language explanations (Kumar and Talukdar, 2020; Camburu et al., 2018; Wadhwa et al., 2024) are appealing for their self-contained insights, even when the models are closed-source. However, there is no work studying rigorous uncertainties behind these explanations, making it difficult to understand the confidence level associated with them. Traditional uncertainty estimation methods (e.g., perturbation and Bayesian-based methods) (Tanneru et al., 2024; Xiong et al., 2023; Liu et al., 2024) face significant limitations when applied to natural language explanations. Specifically, they either fail to provide valid uncertainty guarantees, or require the access of model logits and extensive model retraining.

In this work, we aim to provide rigorous, post-hoc, and model-agnostic uncertainty guarantees for natural-language explanations in QA. Conformal prediction offers a distribution-free framework with provably valid coverage (Shafer and Vovk, 2008; Su et al., 2024; Campos et al., 2024; Qian et al., 2024b; Li et al., 2024; Lidder et al., 2025; Zhao et al., 2025; Angelopoulos et al., 2024). However, traditional conformal prediction methods cannot be directly applied to natural language explanations. The reason is that in conformal prediction, the underlying models are typically trained in a supervised manner, where the uncertainty sets correspond to predefined class labels. In contrast, LLMs are trained in an auto-regressive fashion, generating text one token at a time, with each token

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 25005–25013

conditioned on the previously generated tokens.

Additionally, real-world QA queries often contain noise, such as ambiguous phrasing and typographical errors. Such noise can violate the underlying exchangeability assumption required by conformal prediction (Shafer and Vovk, 2008). Therefore, these generated uncertainties could become invalid, posing substantial challenges for generating reliable uncertainty estimates of natural language explanations. Although several robust conformal methods have been proposed (Yan et al., 2024; Ghosh et al., 2023; Wang et al., 2024; Jeary et al., 2024), they typically assume well-structured datasets and fail to account for discrete and token-level noise that is inherent in natural language explanations generated by LLMs. Such noise complicates valid uncertainty guarantees for natural language explanations in QA.

To address the above challenges, in this work, we propose **ULXMQA**, a novel **u**ncertainty method for natural **l**anguage e**x**planations for **m**edical **q**uestion **a**nswering, which can generate valid uncertainty guarantees in a post-hoc and model-agonistic way. Specifically, in our method, we first design prompts, which can assign each input token an importance score. Then, we design an uncertainty set construction function, which selects explanation tokens based on their assigned importance scores. For the constructed uncertainty sets, we provide theoretical guarantees by proving that the expected fraction of ground-truth tokens included in these uncertainty sets is theoretically guaranteed. Additionally, to address noisy data that may undermine the validity of the generated uncertainty sets, we also design a **r**obust **u**ncertain estimation method for these generated natural **l**anguage e**x**planations (**RULX**), which can provide robust valid uncertainty guarantees under discrete and token-level noise in questions. We further conduct extensive experiments to verify the desired performance of our proposed methods across different question answering tasks.

2 Methodology

Here, we first introduce our valid uncertainty method for natural language explanations in LLM-based QA systems. Then, we present the proposed robust uncertainty method, designed to mitigate the effects of noise.

Without loss of generality, in this paper, we consider a vision-language model based QA system, which can output accurate answers to medical questions about the input medical image. Let $\mathcal{Q}$ represent the question space and $\mathcal{A}$ the answer space. We denote a language model as $\mathcal{M} : \mathcal{Q} \rightarrow \mathcal{A}$, which takes a sequence of k question tokens $Q = (q_1, q_2, \ldots, q_k)$, and produces a sequence of m answer tokens $A = (a_1, a_2, \ldots, a_m)$. A designed prompt P augments the input and instructs $\mathcal{M}$ to emit a natural language explanation E, which we model as a subset of question tokens deemed essential for predicting A.

Modeling uncertainty in natural language explanations. Here, we propose a post-hoc, model-agnostic method that assigns provably valid uncertainty to natural language explanations E. Let $\mathcal{D}^{\text{cal}} = \{(P_i, Q_i, E_i^*, A_i)\}_{i=1}^{n}$ denote the calibration data, and Q_{n+1} denote the test-time question. Here, E_i^* denotes the ground-truth explanation supplied by human annotators. Each example comes with a gold explanation sentence, and annotators mark the question tokens they judge essential in light of that sentence (Aggarwal et al., 2021). Specifically, we construct an uncertainty set of Q_{n+1} that provides theoretical guarantees on the inclusion ratio of ground-truth natural language explanations. The challenge in providing guarantees is that the language model $\mathcal{M}$ generates text auto-regressively. To address this, for each question $Q_i \in \{Q_i\}_{i=1}^{n+1}$, we propose a confidence-aware prompt P_i, which enables $\mathcal{M}$ to output an importance score $\mathcal{S}(P_i, q_{i,j}; \mathcal{M}) \in [0, 1]$ for each word in Q_i and simultaneously obtain the final answer. These scores reflect how essential each token is for predicting the final answer. Then, for Q_i, we construct its uncertainty explanation set as follows

$$\mathcal{C}_\lambda(Q_i; \mathcal{M}) = \{q_{i,j} \in Q_i : \mathcal{S}(P_i, q_{i,j}; \mathcal{M}) \geq 1 - \lambda\}, \quad (1)$$

where λ is a parameter that increases the size of the prediction sets as its value grows. To obtain the importance score, we concatenate a prompt P_i to the given question Q_i using the template:“ Read the question and assign each word an importance score...”. Crucially, our method preserves its coverage guarantee regardless of the quality of these scores. Such prompt-based explanations can effectively reveal the underlying reasoning process behind model predictions (Parcalabescu and Frank, 2024; He et al., 2024; Sudhi et al., 2024). Notably, our method ensures valid coverage guarantees regardless of variations in the quality of these prompt-based results across different LLMs.

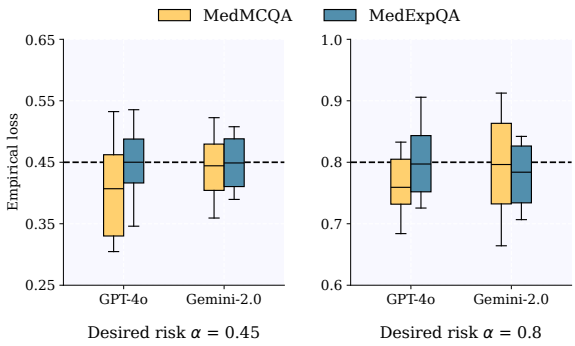

Figure 1: Validity of our ULXQA at desired risks.

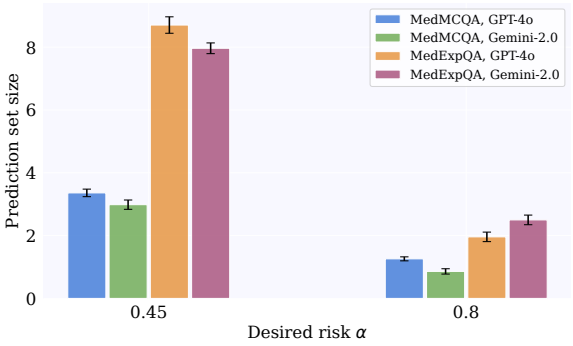

Figure 2: Efficiency of our ULXQA at desired risks.

To quantify the quality of the constructed uncertainty set $\mathcal{C}_\lambda(Q_i;\mathcal{M})$ of Q_i from the calibration data $\mathcal{D}^{\text{cal}}$, we measure the proportion of ground-truth token explanations that appear in this set relative to the total ground-truth token explanations, and then define

$$\ell(\mathcal{C}_\lambda(Q_i;\mathcal{M}), E_i^*, \lambda) = 1 - |E_i^* \cap \mathcal{C}_\lambda(Q_i;\mathcal{M})|/|E_i^*|. \quad (2)$$

The above loss $\ell(\mathcal{C}_\lambda(Q_i;\mathcal{M}), E_i^*, \lambda)$ decreases when the set $\mathcal{C}_\lambda(Q_i;\mathcal{M})$ includes a larger fraction of true tokens. Then, we calculate the average empirical loss at level λ on the calibration set as $\widehat{R}_n(\lambda) = (\ell(\mathcal{C}_\lambda(Q_1;\mathcal{M}), E_1^*, \lambda) + \ldots + \ell(\mathcal{C}_\lambda(Q_n;\mathcal{M}), E_n^*, \lambda))/n$. Given any desired risk level $\alpha \in (0,1)$, we set

$$\hat{\lambda} = \inf\{\lambda : \widehat{R}_n(\lambda) \leq \alpha - \frac{1-\alpha}{n}\}. \quad (3)$$

Since $\widehat{R}_n(\lambda)$ is monotone, we can efficiently search for $\hat{\lambda}$ using binary search to arbitrary precision and construct the uncertainty set $\mathcal{C}_{\hat{\lambda}}(Q_{n+1};\mathcal{M})$ with uncertainty guarantees for the test-time question Q_{n+1} based on Eq. (1) and (3). The full algorithm is deferred to Algorithm 1 in the Appendix B.

Theorem 1. *Assume that the calibration set $\mathcal{D}^{cal}$ and the test data are exchangeable. For any desired $\alpha \in (0,1)$, let $\widehat{R}_n(\lambda) = (\ell(\mathcal{C}_\lambda(Q_1;\mathcal{M}), E_1^*, \lambda) + \ldots + \ell(\mathcal{C}_\lambda(Q_n;\mathcal{M}), E_n^*, \lambda))/n$ and choose $\hat{\lambda}$ according to Eq. (3). Then, for the constructed uncertainty set $\mathcal{C}_{\hat{\lambda}}(Q_{n+1};\mathcal{M})$, we have*

$$\mathbb{E}[\ell(\mathcal{C}_{\hat{\lambda}}(Q_{n+1};\mathcal{M}), E_{n+1}^*, \hat{\lambda})] \leq \alpha. \quad (4)$$

Theorem 1 guarantees that, on average, the uncertainty set $\mathcal{C}_{\hat{\lambda}}(Q_{n+1};\mathcal{M})$ contains at least a $1-\alpha$ fraction of the true tokens, providing a valid coverage guarantee for the generated natural language explanations. Note that Theorem 1 is stated under the exchangeability assumption, which is weaker than independence and identical distribution (i.i.d.). *The proof of Theorem 1 is provided in the Appendix A.*

Robust uncertainty guarantees under noisy data. Note that the uncertainty guarantee of $\mathcal{C}_{\hat{\lambda}}(Q_{n+1};\mathcal{M})$ relies on the exchangeability of the data, which can be violated by noise such as ambiguous phrasing or typographical errors. Our goal is to provide robust uncertainty guarantees under noise. Let Q'_{n+1} represent the noisy test question derived from a clean version Q^*_{n+1}. To model potential noise, we define $\mathcal{B}_{Q^*_{n+1}}$ as the set of candidate noisy questions associated with Q^*_{n+1}. As previously discussed, discrete and token-level noise inherent in natural language explanations poses significant challenges for robust uncertainties. To address this, for each word $q^*_{n+1,j} \in Q^*_{n+1}$, we define a synonym set $\mathcal{B}_{q^*_{n+1,j}}$, which contains the synonyms of $q^*_{n+1,j}$ (including $q^*_{n+1,j}$ itself). Consequently, if noise affects at most $d \leq k$ words in Q^*_{n+1}, replacing them with elements from their respective synonym sets, the observed noisy question $Q'_{n+1} = \{q'_{n+1,1}, \cdots, q'_{n+1,k}\}$ emerges as follows

$$\mathcal{B}_{Q^*_{n+1}} = \{Q'_{n+1} : \|Q'_{n+1} - Q^*_{n+1}\|_0 \leq d, \; q'_{n+1,j} \in \mathcal{B}_{q^*_{n+1,j}}, \forall j\}, \quad (5)$$

where $\|Q'_{n+1} - Q^*_{n+1}\|_0 := \sum_{j=1}^k \mathbb{1}\{q'_{n+1,j} \neq q^*_{n+1,j}\}$. For the noisy test question $Q'_{n+1} \in \mathcal{B}_{Q^*_{n+1}}$, note that $\mathcal{B}_{Q'_{n+1}}$ also contains the clean test question Q^*_{n+1} because noise affects at most $d \leq k$ words in Q^*_{n+1}. To construct its robust uncertainty set, for each word $\tilde{q}_{n+1,j} \in \mathcal{B}_{q'_{n+1,j}}$ we compute

$$\mathcal{R}(P_{n+1}, \tilde{q}_{n+1,j}; \mathcal{B}_{Q'_{n+1}}, \mathcal{M}) = \sup_{\tilde{Q}_{n+1} \in \mathcal{B}_{Q'_{n+1}}, \tilde{q}_{n+1,j} \in \tilde{Q}_{n+1}} \mathcal{S}(P_{n+1}, \tilde{q}_{n+1,j}; \mathcal{M}), \quad (6)$$

where the supremum is taken over all noisy questions $\tilde{Q}_{n+1} \in \mathcal{B}_{Q'_{n+1}}$ that still contain $\tilde{q}_{n+1,j}$. In-

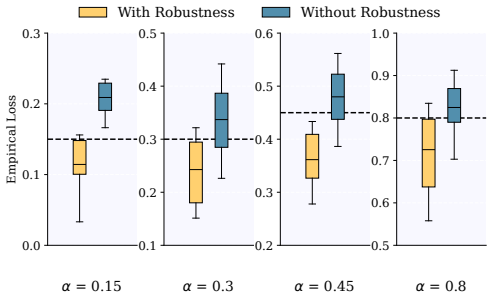

Figure 3: Validity of our RULX on MedMCQA under noisy data.

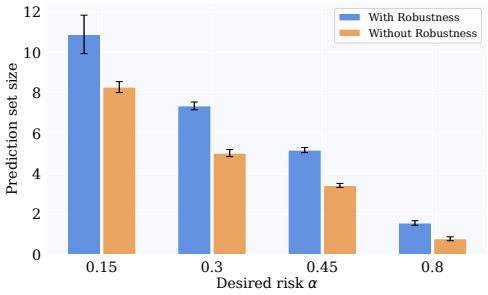

Figure 4: Efficiency of our RULX on MedMCQA under noisy data.

tuitively, this ensures we capture the maximum importance score of $\tilde{q}_{n+1,j}$ across all relevant perturbations. With $\hat{\lambda}$ defined as in Eq. (3), we can construct the robust uncertainty set as follows

$$\mathcal{C}^{\mathcal{R}}_{\hat{\lambda}}(Q'_{n+1};\mathcal{M}) = \{\tilde{q}_{n+1,j} \in \mathcal{B}_{q'_{n+1}} : \quad (7)$$
$$\mathcal{R}(P_{n+1}, \tilde{q}_{n+1,j}; \mathcal{B}_{Q'_{n+1}}, \mathcal{M}) \geq 1 - \hat{\lambda}\}.$$

Here, robust uncertainty set $\mathcal{C}^{\mathcal{R}}_{\hat{\lambda}}(Q'_{n+1};\mathcal{M})$ includes all tokens whose maximum importance scores exceed the threshold $1 - \hat{\lambda}$.

Theorem 2. *Let $Q'_{n+1} \in \mathcal{B}_{Q^*_{n+1}}$ be a noisy test question near the clean test question Q^*_{n+1} such that $\|Q'_{n+1} - Q^*_{n+1}\|_0 \leq d$. For the above constructed uncertainty set $\mathcal{C}^{\mathcal{R}}_{\hat{\lambda}}(Q'_{n+1};\mathcal{M})$, it satisfies*

$$\mathbb{E}[\ell(\mathcal{C}^{\mathcal{R}}_{\hat{\lambda}}(Q'_{n+1};\mathcal{M}), E^*_{n+1}, \hat{\lambda})] \leq \alpha. \quad (8)$$

According to Theorem 2, for the noisy question Q'_{n+1}, the expected proportion of true tokens included in the uncertainty set $\mathcal{C}^{\mathcal{R}}_{\hat{\lambda}}(Q'_{n+1};\mathcal{M})$ is guaranteed to be at least $1 - \alpha$. *Due to the space limitations, the proof of Theorem 2 and the full algorithm for RULX are provided in the Appendix A&B.* Our framework could be generalized to the full conformal prediction setting (Martinez et al., 2023; Chen et al., 2024; Blot et al., 2025), where machine unlearning techniques (Zhao et al., 2023, 2024; Qian et al., 2023, 2024a, 2025; Chen et al., 2025a) could be explored to mitigate the associated high computational costs of retraining.

3 Experiments

3.1 Experimental Setup

Datasets and models. We evaluate our approaches on two real-world QA datasets: MedMCQA (Pal et al., 2022), a large-scale dataset of 194k multiple-choice questions, and MedExpQA (Alonso et al., 2024), a multilingual set of 622 clinical-case questions. Note that both datasets include ground-truth explanations. Additionally, we adopt two advanced LLMs, GPT-4o (OpenAI, 2024) and Gemini 2.0 Flash (Google, 2024), to ensure a thorough evaluation. Our code will be publicly released upon acceptance.

Implementation details. For our experiments, we utilize LLMs with a temperature setting of 1. For each adopted QA dataset, we use its validation data as the test data, and partition its training data into 70% fine-tuning data and 30% calibration data. All experiments are run for 10 trials, and we report the averaged experimental results.

3.2 Experimental Results

Validity. We evaluate ULXQA's validity on MedMCQA and MedExpQA across both LLMs, reporting empirical loss in Eq. (2) at risk levels $\alpha = 0.45$ and $\alpha = 0.8$. Note that these LLMs are fine-tuned on the adopted QA datasets. In Fig. 1, the dashed line marks the desired risk, while the horizontal line in each box shows the average empirical loss. Our proposed ULXQA consistently maintains loss below or equal to α, ensuring compliance with Eq. (4). For instance, with $\alpha = 0.45$, our proposed ULXQA achieves an empirical loss of 0.44 on MedExpQA using Gemini 2.0 Flash. These results confirm that ULXQA can provide valid uncertainty guarantees on these generated explanations.

Efficiency. In Fig. 2, we explore the efficiency of our proposed ULXQA across various LLMs. We report the average set size at the desired risk levels $\alpha = 0.45$ and $\alpha = 0.8$, ensuring consistency with the validity experiments. As depicted, our proposed ULXQA consistently provides explanation uncertainty sets with small sizes, so that it remains efficient when doctors use these explanation uncertainty sets to make decisions. For instance, on the adopted MedMCQA dataset, the average set size of our ULXQA using Gemini 2.0 Flash is approxi-

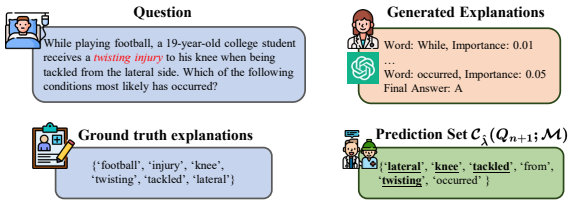

Figure 5: Visualization results of our ULXQA on MedMCQA.

mately 2.99 at $\alpha = 0.45$. It shows that, on average, fewer than three words can provide correct explanations for users with valid coverage guarantees. This indicates that ULXQA can efficiently output uncertainty quantification of natural language explanations for QA tasks.

Visualization. We visualize uncertainty quantification for explanations on the MedMCQA dataset. In Fig. 5, the uncertainty set for a target image and question shows strong overlap with the ground truth, containing four correct words ('lateral', 'knee', 'tackled', and 'twisting') in the prediction set. The prediction set size is six, which is close to the size of the ground truth. These results demonstrate our method's ability to capture truly influential explanations with modest redundancy.

Robust uncertainty under noisy data. We evaluate RULX's validity and efficiency under varying risk levels using Gemini 2.0 Flash with noisy test data on MedMCQA. In Fig. 3, the non-robust approach often exceeds desired risk levels, lacking formal guarantees, while RULX consistently stays within bounds, satisfying Eq. (8). Fig. 4 shows that although robust RULX slightly enlarges prediction sets to ensure validity, their sizes remain comparable to the non-robust method. Together, these figures confirm that RULX maintains valid uncertainty guarantees, while keeping prediction set sizes effectively comparable.

4 Conclusion

To the best of our knowledge, this work is the first to introduce a rigorous uncertainty estimation framework for natural language explanations in LLM-based QA systems. The post-hoc, model-agnostic method guarantees coverage by ensuring the expected fraction of non-ground-truth explanation tokens below a threshold. Building upon this, we further propose a robust extension that maintains reliable and valid uncertainty guarantees in the presence of noise. We also conduct extensive experiments across various QA tasks to comprehensively evaluate the effectiveness of our proposed methods.

Acknowledgments

This work is supported in part by the US National Science Foundation under grants CNS-2350332 and IIS-2442750. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Limitations

Our experiments show that ULXQA and its robust variant RULX achieve reliable coverage guarantees on challenging QA tasks, confirming the practical value of our framework. However, our current study has several limitations. First, the experimental results focus on the limited datasets, so additional experiments on other types of QA (e.g., legal or open-domain) are needed to verify generality. Second, the present study considers only single-modal textual inputs. An important next step is to extend ULXQA/RULX to multimodal settings, such as visual or audio question answering, and maintain uncertainty guarantees when multiple modalities are involved.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Shourya Aggarwal, Divyanshu Mandowara, Vishwajeet Agrawal, Dinesh Khandelwal, Parag Singla, and Dinesh Garg. 2021. Explanations for CommonsenseQA: New Dataset and Models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3050–3065, Online. Association for Computational Linguistics.

Iñigo Alonso, Maite Oronoz, and Rodrigo Agerri. 2024. Medexpqa: Multilingual benchmarking of large language models for medical question answering. *Artificial Intelligence in Medicine*, 155:102938.

Anastasios Nikolas Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. 2024. Conformal risk control. In *The Twelfth International Conference on Learning Representations*.

Vincent Blot, Anastasios Nikolas Angelopoulos, Michael Jordan, and Nicolas JB Brunel. 2025. Automatically adaptive conformal risk control. In *International Conference on Artificial Intelligence and Statistics*, pages 19–27. PMLR.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. *Advances in Neural Information Processing Systems*, 31.

Margarida Campos, António Farinhas, Chrysoula Zerva, Mário A. T. Figueiredo, and André F. T. Martins. 2024. Conformal prediction for natural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 12:1497–1516.

Aobo Chen, Yangyi Li, Wei Qian, Kathryn Morse, Chenglin Miao, and Mengdi Huai. 2024. Modeling and understanding uncertainty in medical image classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 557–567. Springer.

Aobo Chen, Yangyi Li, Chenxu Zhao, and Mengdi Huai. 2025a. A survey of security and privacy issues of machine unlearning.

Hanjie Chen, Zhouxiang Fang, Yash Singla, and Mark Dredze. 2025b. Benchmarking large language models on answering and explaining challenging medical questions. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3563–3599, Albuquerque, New Mexico. Association for Computational Linguistics.

Subhankar Ghosh, Yuanjie Shi, Taha Belkhouja, Yan Yan, Jana Doppa, and Brian Jones. 2023. Probabilistically robust conformal prediction. In *Uncertainty in Artificial Intelligence*, pages 681–690. PMLR.

Google. 2024. Introducing gemini 2.0: our new ai model for the agentic era. https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/.

Xiaoxin He, Xavier Bresson, Thomas Laurent, Adam Perold, Yann LeCun, and Bryan Hooi. 2024. Harnessing explanations: LLM-to-LM interpreter for enhanced text-attributed graph representation learning. In *The Twelfth International Conference on Learning Representations*.

Linus Jeary, Tom Kuipers, Mehran Hosseini, and Nicola Paoletti. 2024. Verifiably robust conformal prediction. *arXiv preprint arXiv:2405.18942*.

Sawan Kumar and Partha Talukdar. 2020. NILE : Natural language inference with faithful natural language explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8730–8742, Online. Association for Computational Linguistics.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Yangyi Li, Aobo Chen, Wei Qian, Chenxu Zhao, Divya Lidder, and Mengdi Huai. 2024. Data poisoning attacks against conformal prediction. In *International Conference on Machine Learning*, pages 27563–27574. PMLR.

Divya Lidder, Kathryn Morse, Bridget Sullivan, Wei Qian, Chenglin Miao, and Mengdi Huai. 2025. Neuron explanations for conformal prediction (student abstract). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 29412–29414.

Shudong Liu, Zhaocong Li, Xuebo Liu, Runzhe Zhan, Derek F. Wong, Lidia S. Chao, and Min Zhang. 2024. Can LLMs learn uncertainty on their own? expressing uncertainty effectively in a self-training manner. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21635–21645, Miami, Florida, USA. Association for Computational Linguistics.

Javier Abad Martinez, Umang Bhatt, Adrian Weller, and Giovanni Cherubin. 2023. Approximating full conformal prediction at scale via influence functions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 6631–6639.

OpenAI. 2024. Hello gpt-4o. https://openai.com/index/hello-gpt-4o/. Accessed: 2024-09-28.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR.

Letitia Parcalabescu and Anette Frank. 2024. On measuring faithfulness or self-consistency of natural language explanations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6048–6089, Bangkok, Thailand. Association for Computational Linguistics.

Wei Qian, Aobo Chen, Chenxu Zhao, Yangyi Li, and Mengdi Huai. 2024a. Exploring fairness in educational data mining in the context of the right to be forgotten. *arXiv preprint arXiv:2405.16798*.

Wei Qian, Chenxu Zhao, Wei Le, Meiyi Ma, and Mengdi Huai. 2023. Towards understanding and enhancing robustness of deep learning models against malicious unlearning attacks. In *Proceedings of the*

29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pages 1932–1942.

Wei Qian, Chenxu Zhao, Yangyi Li, Fenglong Ma, Chao Zhang, and Mengdi Huai. 2024b. Towards modeling uncertainties of self-explaining neural networks via conformal prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 14651–14659.

Wei Qian, Chenxu Zhao, Yangyi Li, Wenqian Ye, and Mengdi Huai. 2025. Towards unveiling predictive uncertainty vulnerabilities in the context of the right to be forgotten. *arXiv preprint arXiv:2508.07458*.

Glenn Shafer and Vladimir Vovk. 2008. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3).

Jiayuan Su, Jing Luo, Hongwei Wang, and Lu Cheng. 2024. API is enough: Conformal prediction for large language models without logit-access. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 979–995, Miami, Florida, USA. Association for Computational Linguistics.

Viju Sudhi, Sinchana Ramakanth Bhat, Max Rudat, and Roman Teucher. 2024. Rag-ex: A generic framework for explaining retrieval augmented generation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2776–2780.

Sree Harsha Tanneru, Chirag Agarwal, and Himabindu Lakkaraju. 2024. Quantifying uncertainty in natural language explanations of large language models. In *International Conference on Artificial Intelligence and Statistics*, pages 1072–1080. PMLR.

Somin Wadhwa, Adit Krishnan, Runhui Wang, Byron C Wallace, and Luyang Kong. 2024. Learning from natural language explanations for generalizable entity matching. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6114–6129, Miami, Florida, USA. Association for Computational Linguistics.

Zhiyuan Wang, Jinhao Duan, Lu Cheng, Yue Zhang, Qingni Wang, Xiaoshuang Shi, Kaidi Xu, Heng Tao Shen, and Xiaofeng Zhu. 2024. ConU: Conformal uncertainty in large language models with correctness coverage guarantees. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6886–6898, Miami, Florida, USA. Association for Computational Linguistics.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2023. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *arXiv preprint arXiv:2306.13063*.

Ge Yan, Yaniv Romano, and Tsui-Wei Weng. 2024. Provably robust conformal prediction with improved efficiency. *arXiv preprint arXiv:2404.19651*.

Chenxu Zhao, Wei Qian, Aobo Chen, and Mengdi Huai. 2025. Membership inference attacks with false discovery rate control. *arXiv preprint arXiv:2508.07066*.

Chenxu Zhao, Wei Qian, Yangyi Li, Aobo Chen, and Mengdi Huai. 2024. Rethinking adversarial robustness in the context of the right to be forgotten. In *International Conference on Machine Learning*, pages 60927–60939. PMLR.

Chenxu Zhao, Wei Qian, Rex Ying, and Mengdi Huai. 2023. Static and sequential malicious attacks in the context of selective forgetting. *Advances in Neural Information Processing Systems*, 36:74966–74979.

Zining Zhu, Hanjie Chen, Xi Ye, Qing Lyu, Chenhao Tan, Ana Marasovic, and Sarah Wiegreffe. 2024. Explanation in the era of large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 5: Tutorial Abstracts)*, pages 19–25, Mexico City, Mexico. Association for Computational Linguistics.

A Proofs of Theorems

Theorem 1. *Assume that the calibration set $\mathcal{D}^{cal}$ and the test data are exchangeable. For any desired $\alpha \in (0,1)$, let $\widehat{R}_n(\lambda) = (\ell(\mathcal{C}_\lambda(Q_1;\mathcal{M}), E_1^*, \lambda) + \ldots + \ell(\mathcal{C}_\lambda(Q_n;\mathcal{M}), E_n^*, \lambda))/n$ and choose $\hat{\lambda}$ according to Eq. (3). Then, for the constructed uncertainty set $\mathcal{C}_{\hat{\lambda}}(Q_{n+1};\mathcal{M})$, we have*

$$\mathbb{E}[\ell(\mathcal{C}_{\hat{\lambda}}(Q_{n+1};\mathcal{M}), E_{n+1}^*, \hat{\lambda})] \leq \alpha. \tag{9}$$

Proof. Consider a sequence of exchangeable random loss functions, $\{\ell(\mathcal{C}_\lambda(Q_i;\mathcal{M}), E_i, \lambda)\}_{i=1}^{n+1}$, where $\ell(\cdot,\cdot,\lambda)$ defined in Eq. (2) is non-increasing in λ, right-continuous, and satisfying $\ell(\mathcal{C}_{\lambda_{\max}}(Q_1;\mathcal{M}), \cdot, \lambda_{\max}) \leq \alpha$ when $\lambda_{\max} = 1$. We define

$$\begin{aligned}\hat{R}_{n+1}(\lambda) &= (\ell(\mathcal{C}_\lambda(Q_1;\mathcal{M}), E_1, \lambda) + \ldots \\ &\quad + \ell(\mathcal{C}_\lambda(Q_{n+1};\mathcal{M}), E_{n+1}, \lambda))/(n+1), \\ \hat{\lambda}' &= \inf\{\lambda \in \Lambda : \hat{R}_{n+1}(\lambda) \leq \alpha\}. \end{aligned} \tag{10}$$

Since $\inf_\lambda \ell = \ell(\mathcal{C}_{\lambda_{\max}}(Q_1;\mathcal{M}), \cdot, \lambda_{\max}) \leq \alpha$, $\hat{\lambda}'$ is well-defined almost surely. Since $\ell(\mathcal{C}_\lambda(Q_{n+1};\mathcal{M}), E_{n+1}, \lambda) \leq \sup_\lambda \ell = 1$, we get

$$\begin{aligned}&\hat{R}_{n+1}(\lambda) \\ &= \frac{n}{n+1}\hat{R}_n(\lambda) + \frac{\ell(\mathcal{C}_\lambda(Q_{n+1};\mathcal{M}), E_{n+1}, \lambda)}{n+1} \\ &\leq \frac{n}{n+1}\hat{R}_n(\lambda) + \frac{1}{n+1}. \end{aligned} \tag{11}$$

Thus, we can have

$$\frac{n}{n+1}\hat{R}_n(\lambda) + \frac{1}{n+1} \leq \alpha \Rightarrow \hat{R}_{n+1}(\lambda) \leq \alpha.$$

This implies $\hat{\lambda}' \leq \hat{\lambda}$ when the LHS holds for some $\lambda \in [0,1]$. When the LHS is above α for all $\lambda \in [0,1]$, by definition, $\hat{\lambda} = \lambda_{\max} \geq \hat{\lambda}'$. Thus, $\hat{\lambda}' \leq \hat{\lambda}$ almost surely. Since $\ell(\cdot,\cdot,\lambda)$ is non-increasing in λ,

$$\begin{aligned}&\mathbb{E}[\ell(\mathcal{C}_{\hat{\lambda}}(Q_{n+1};\mathcal{M}), E_{n+1}, \hat{\lambda})] \\ \leq &\mathbb{E}[\ell(\mathcal{C}_{\hat{\lambda}'}(Q_{n+1};\mathcal{M}), E_{n+1}, \hat{\lambda}')]. \end{aligned} \tag{12}$$

Let $\mathcal{E}$ be the multiset of loss functions $\{\ell(\mathcal{C}_\lambda(Q_i;\mathcal{M}), E_i, \lambda)\}_{i=1}^{n+1}$. Then $\hat{\lambda}'$ is a function of $\mathcal{E}$, or equivalently, $\hat{\lambda}'$ is a constant conditional on $\mathcal{E}$. Additionally, $\ell(\mathcal{C}_\lambda(Q_{n+1};\mathcal{M}), E_{n+1}, \lambda)|\mathcal{E} \sim \text{Uniform}(\{\ell(\mathcal{C}_\lambda(Q_i;\mathcal{M}), E_i, \lambda)\}_{i=1}^{n+1})$ by exchangeability. These facts combined with the right-continuity of L_i imply

$$\begin{aligned}&\mathbb{E}[\ell(\mathcal{C}_{\hat{\lambda}'}(Q_{n+1};\mathcal{M}), E_{n+1}, \hat{\lambda}') \mid \mathcal{E}] \\ &= \frac{1}{n+1}\sum_{i=1}^{n+1} L_i(\hat{\lambda}') \leq \alpha. \end{aligned} \tag{13}$$

The proof is completed by the law of total expectation and Eq. (12). □

Theorem 2. *Let $Q'_{n+1} \in \mathcal{B}_{Q^*_{n+1}}$ be a noisy test question near the clean test question Q^*_{n+1} such that $\|Q'_{n+1} - Q^*_{n+1}\|_0 \leq d$. For the above constructed uncertainty set $\mathcal{C}^{\mathcal{R}}_{\hat{\lambda}}(Q'_{n+1};\mathcal{M})$, it satisfies*

$$\mathbb{E}[\ell(\mathcal{C}^{\mathcal{R}}_{\hat{\lambda}}(Q'_{n+1};\mathcal{M}), E^*_{n+1}, \hat{\lambda})] \leq \alpha. \tag{14}$$

Proof. According to the definition of robust score in Eq. (6), we have

$$\begin{aligned}&\mathcal{R}(P_{n+1}, q^*_{n+1,j}; \mathcal{B}_{Q'_{n+1}}, \mathcal{M}) \\ \geq &\mathcal{S}(P_{n+1}, q^*_{n+1,j}; \mathcal{M}), \end{aligned} \tag{15}$$

for any clean token $q^*_{n+1,j}$, which means

$$\begin{aligned}&q^*_{n+1,j} \in \mathcal{C}_{\hat{\lambda}}(Q^*_{n+1};\mathcal{M}) \\ \Rightarrow &q^*_{n+1,j} \in \mathcal{C}^{\mathcal{R}}_{\hat{\lambda}}(Q'_{n+1};\mathcal{M}). \end{aligned} \tag{16}$$

Consequently, we have

$$\begin{aligned}&\mathbb{E}[\ell(\mathcal{C}^{\mathcal{R}}_{\hat{\lambda}}(Q'_{n+1};\mathcal{M}), E_{n+1}, \hat{\lambda})] \\ \leq &\mathbb{E}[\ell(\mathcal{C}_{\hat{\lambda}}(Q^*_{n+1};\mathcal{M}), E_{n+1}, \hat{\lambda})] \leq \alpha, \end{aligned} \tag{17}$$

which directly implies the result stated in Eq. (14). □

Algorithm 1 ULXQA

Input: A language model $\mathcal{M}$, importance score $\mathcal{S}(P_i, q_{i,j}; \mathcal{M})$, calibration data $\mathcal{D}^{\text{cal}}$, test sample Q_{n+1}, Candidate threshold grid $\Lambda = \{\lambda_1 < \lambda_2 < \cdots < \lambda_K\}$, desired risk level $\alpha \in (0,1)$

Output: Uncertainty set $\mathcal{C}_{\hat{\lambda}}(Q_{n+1};\mathcal{M})$

for $i \leftarrow 1$ **to** n **do**
  for $k \leftarrow 1$ **to** K **do**
    Construct $\mathcal{C}_\lambda(Q_i;\mathcal{M}) = \{q_{i,j} \in Q_i : \mathcal{S}(P_i, q_{i,j};\mathcal{M}) \geq 1-\lambda\}$
    Compute $\ell(\mathcal{C}_\lambda(Q_i;\mathcal{M}), E^*_i, \lambda_k)$
  end for
end for
$low \leftarrow 1,\ high \leftarrow K$
while $low < high$ **do**
  $mid \leftarrow \lfloor \frac{low+high}{2} \rfloor$
  $\widehat{R}_n(\lambda_{mid}) \leftarrow \sum_{i=1}^{n} \frac{\ell(\mathcal{C}_\lambda(Q_i;\mathcal{M}), E^*_i, \lambda_{mid})}{n}$
  if $\widehat{R}_n(\lambda_{mid}) \leq \alpha$ **then**
    $high \leftarrow mid$
  else
    $low \leftarrow mid + 1$
  end if
end while
$\hat{\lambda} \leftarrow \lambda_{low}$
Construct the uncertainty set $\mathcal{C}_{\hat{\lambda}}(Q_{n+1};\mathcal{M}) = \{q_{n+1,j} \in Q_{n+1} : \mathcal{S}(P_{n+1}, q_{n+1,j};\mathcal{M}) \geq 1-\hat{\lambda}\}$ that satisfies Eq. (4)

B Algorithms

B.1 Algorithm for ULXQA

Algorithm 1 describes how to construct the uncertainty set $\mathcal{C}_{\hat{\lambda}}(Q_{n+1};\mathcal{M})$ with the valid uncertainty guarantees for the test-time question Q_{n+1}. Specifically, this set ensures that, on average, it contains at least a $1-\alpha$ fraction of the ground-truth explanation tokens. This yields a provably valid coverage guarantee for the natural language explanations generated by the model.

B.2 Algorithm for RULX

Algorithm 2 outlines the construction of the uncertainty set $\mathcal{C}_{\hat{\lambda}}(Q_{n+1};\mathcal{M})$. This method provides a robust coverage guarantee that the expected fraction of ground-truth explanation tokens included in the uncertainty set remains at least $1-\alpha$, even in the presence of input noise. As a result, it offers reliable uncertainty quantification for natural language explanations under possible perturbations.

Algorithm 2 RULX

Input: A language model $\mathcal{M}$, importance score $\mathcal{S}(P_i, q_{i,j}; \mathcal{M})$, calibration data $\mathcal{D}^{\text{cal}}$, test sample Q_{n+1}, Candidate threshold grid $\Lambda = \{\lambda_1 < \lambda_2 < \cdots < \lambda_K\}$, desired risk level $\alpha \in (0, 1)$

Output: Uncertainty set $\mathcal{C}^{\mathcal{R}}_{\hat{\lambda}}(Q'_{n+1}; \mathcal{M})$

for $i \leftarrow 1$ **to** n **do**
  for $k \leftarrow 1$ **to** K **do**
    Construct $\mathcal{C}_\lambda(Q_i; \mathcal{M}) = \{q_{i,j} \in Q_i : \mathcal{S}(P_i, q_{i,j}; \mathcal{M}) \geq 1 - \lambda\}$
    Compute $\ell(\mathcal{C}_\lambda(Q_i; \mathcal{M}), E^*_i, \lambda_k)$
  end for
end for
$low \leftarrow 1,\ high \leftarrow K$
while $low < high$ **do**
  $mid \leftarrow \lfloor \frac{low+high}{2} \rfloor$
  $\widehat{R}_n(\lambda_{mid}) \leftarrow \sum_{i=1}^{n} \frac{\ell(\mathcal{C}_\lambda(Q_i;\mathcal{M}), E^*_i, \lambda_{mid})}{n}$
  if $\widehat{R}_n(\lambda_{mid}) \leq \alpha$ **then**
    $high \leftarrow mid$
  else
    $low \leftarrow mid + 1$
  end if
end while
$\hat{\lambda} \leftarrow \lambda_{low}$
Compute the robust importance score $\mathcal{R}(P_{n+1}, \tilde{q}_{n+1,j}; \mathcal{B}_{Q'_{n+1}}, \mathcal{M})$ based on Eq. (6)
Construct the robust uncertainty set $\mathcal{C}^{\mathcal{R}}_{\hat{\lambda}}(Q'_{n+1}; \mathcal{M}) = \{\tilde{q}_{n+1,j} \in \mathcal{B}_{q'_{n+1}} : \mathcal{R}(P_{n+1}, \tilde{q}_{n+1,j}; \mathcal{B}_{Q'_{n+1}}, \mathcal{M}) \geq 1 - \hat{\lambda}\}$ that satisifies Eq. (8)

C Datasets

We adopt the following datasets:

- **MedMCQA**. Released under the MIT License for research use, this large-scale benchmark provides expert-verified multiple-choice questions spanning cardiology, oncology, pediatrics, neurology, infectious diseases, and other medical specialties. The dataset is publicly available and contains no personally identifiable information. It is entirely in English, which facilitates evaluation of English-language medical QA systems. Although questions center on clinical scenarios, demographic attributes of the underlying patient groups are not specified because such information is absent from the source data.

- **MedExpQA**. Distributed under the CC BY-NC 4.0 license for non-commercial research, MedExpQA contains 622 clinician-authored multiple-choice questions, each paired with gold-standard explanations written by medical professionals. The corpus covers four languages (English, Spanish, French, and Italian) and supports cross-lingual assessment of answer correctness and explanation quality. It is publicly released and free of personally identifiable information. While items focus on clinical reasoning, demographic details of the represented populations are not included due to the nature of the data.

D AI Assistance in Writing

During manuscript preparation, we employed an AI language assistant like GPT-o3 for copy-editing tasks—namely, correcting grammar, smoothing phrasing, and improving overall readability. The model was not used to generate scientific ideas, analyses, or conclusions, and it played no role in shaping the study's methodology or results. Its contribution was limited to language polishing, with all substantive content and final editorial decisions made exclusively by the human authors.