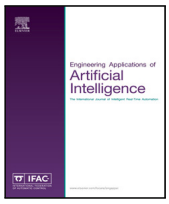
Engineering Applications of Artificial Intelligence 139 (2025) 109553

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai



Word-Sequence Entropy: Towards uncertainty estimation in free-form medical question answering applications and beyond

Zhiyuan Wang [a], Jinhao Duan [b], Chenxi Yuan [c], Qingyu Chen [d], Tianlong Chen [e], Yue Zhang [b], Ren Wang [f], Xiaoshuang Shi [a,*], Kaidi Xu [b]

[a] Center for Future Media, School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, 611731, Sichuan, China
[b] Department of Computer Science, Drexel University, Philadelphia, 19104, PA, USA
[c] Department of Biostatistics, Epidemiology, and Informatics (DBEI), Perelman School of Medicine, University of Pennsylvania, Philadelphia, 19104, PA, USA
[d] Section of Biomedical Informatics & Data Science, Yale School of Medicine, Yale University, New Haven, 06510, CT, USA
[e] Computer Science & Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, 02139, MA, USA
[f] Department of Electrical and Computer Engineering, Illinois Institute of Technology, Chicago, 60616, IL, USA

ARTICLE INFO

Keywords:
Open-ended medical question-answering
Generative inequality
Uncertainty quantification
Semantic relevance

ABSTRACT

Uncertainty estimation is crucial for the reliability of safety-critical human and artificial intelligence (AI) interaction systems, particularly in the domain of healthcare engineering. However, a robust and general uncertainty measure for free-form answers has not been well-established in open-ended medical question-answering (QA) tasks, where generative inequality introduces a large number of irrelevant words and sequences within the generated set for uncertainty quantification (UQ), which can lead to biases. This paper proposes Word-Sequence Entropy (*WSE*), which calibrates uncertainty at both the word and sequence levels based on semantic relevance, highlighting keywords and enlarging the generative probability of trustworthy responses when performing UQ. We compare *WSE* with six baseline methods on five free-form medical QA datasets, utilizing seven popular large language models (LLMs), and demonstrate that *WSE* exhibits superior performance in accurate UQ under two standard criteria for correctness evaluation. Additionally, in terms of the potential for real-world medical QA applications, we achieve a significant enhancement (e.g., a 6.36% improvement in model accuracy on the COVID-QA dataset) in the performance of LLMs when employing responses with lower uncertainty that are identified by *WSE* as final answers, without requiring additional task-specific fine-tuning or architectural modifications.

1. Introduction

Healthcare professionals and patients increasingly employ online search engines to query information and symptoms when confronted with medical conditions. A U.S. health survey (Abacha and Zweigenbaum, 2015) found that 18% of individuals who self-diagnosed online received conflicting advice or outright refusals from medical experts. Despite this, about 77% of adults still prefer online searches over in-person consultations, posing significant health risks. In this context, there is a pressing demand for reliable question-answering (QA) applications in healthcare, to provide accurate and trustworthy responses to user queries.

Recent advancements in natural language generation (NLG), particularly in question-answering (QA) (Brown et al., 2020; Chowdhery et al., 2022; Chen et al., 2023; Ouyang et al., 2022), have been driven by large language models (LLMs) (Waisberg et al., 2023; Zhang et al., 2022; Touvron et al., 2023a; He and Garner, 2023). Enabled by in-context learning[1] (ICL) (Min et al., 2021), LLMs exhibit outstanding task-agnostic and few-shot performance (Brown et al., 2020; Chowdhery et al., 2022; Duan et al., 2024). Given a few-shot prompt with multiple query-response pairs, LLMs efficiently handle new QA tasks (Brown et al., 2020; Ouyang et al., 2022), showing great potential for real-world medical QA applications. However, LLMs are proven to *hallucinate*[2] and provide unfactual answers that seem plausible but

* Corresponding author.
E-mail address: xsshi2013@gmail.com (X. Shi).

[1] In-context learning is to design task-specific instruction prompts, and then leverage a few annotated samples as the prompts to guide LLMs to tackle new test data.

[2] "Hallucinate" is defined as LLMs generating content that is nonsensical or unfaithful to the provided source content. In this case, users cannot trust that any output is correct.

https://doi.org/10.1016/j.engappai.2024.109553
Received 6 March 2024; Received in revised form 20 July 2024; Accepted 23 October 2024

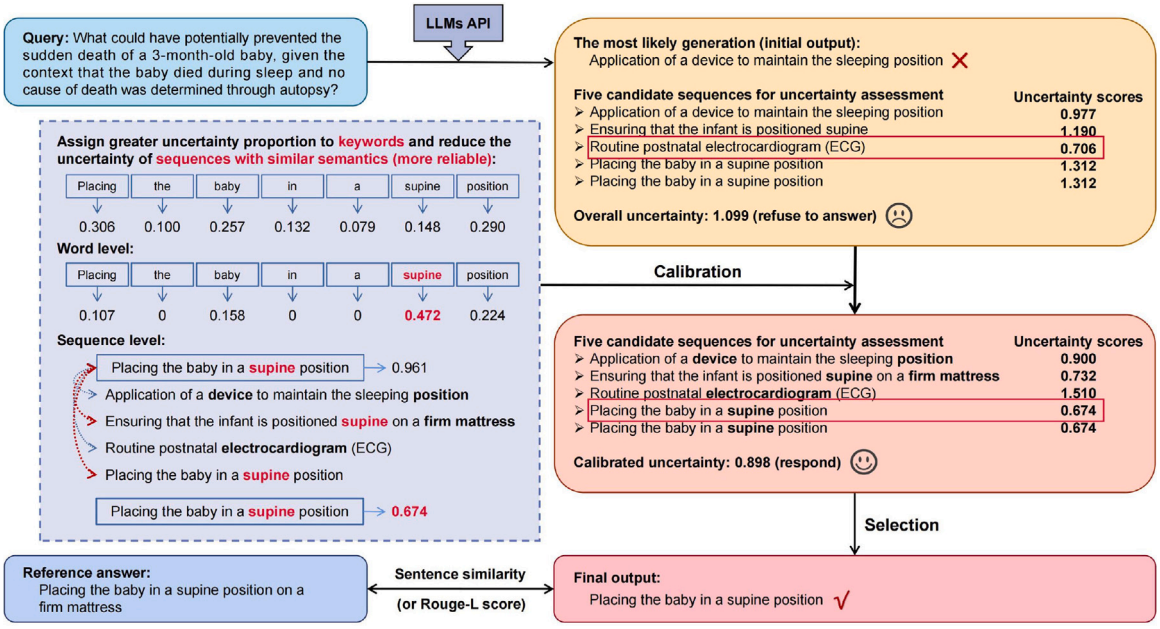


Fig. 1. The overview of *WSE* and its potential for improving model accuracy. Given a medical query, the language model generates the most likely generation as the output, which might be incorrect. Following prior work, we additionally generate multiple (e.g., five) candidate responses to evaluate the trustworthiness of this output. Existing entropy-based measures identify high overall uncertainty in the candidate set, causing the API to refuse to answer the most likely generation. By assessing semantic relevance at both the word and sequence levels, *WSE* highlights keywords and reliable sequences, resulting in calibrated uncertainty that meets the response criterion. Finally, we employ the response with the lowest uncertainty as the final output, which coincides with the reference answer.

deviate from user instructions (Manakul et al., 2023; Yao et al., 2023; Sun et al., 2024), compromising the reliability of their deployment in healthcare applications. Uncertainty quantification (UQ) is an effective approach to address these issues (Kadavath et al., 2022; Chen and Mueller, 2023). By estimating the uncertainty of statements, practical QA applications can inform users about the trustworthiness of the query-answering process, thereby mitigating the risk of unforeseen health incidents.

Nevertheless, UQ in free-form QA tasks, particularly in the medical domain, poses significant challenges. Unlike prediction tasks with specific output forms and labels (Yuan et al., 2023), LLMs-based QA generate semantically equivalent responses but syntactically or lexically distinct, resulting in an unbounded output space. Additionally, LLMs face multiple sources of uncertainty, primarily aleatoric uncertainty from data distribution and epistemic uncertainty from insufficient information (Kendall and Gal, 2017). To address these issues, existing methods either empower LLMs to self-evaluate the uncertainty of their answers through fine-tuning (Lin et al., 2022a; Kadavath et al., 2022) or devise entropy-based measures (Malinin and Gales, 2020; Kuhn et al., 2023; Duan et al., 2023). Recent work, Shift Attention to Relevance (SAR) (Duan et al., 2023), reallocates the weights of uncertainty induced by each token and sentence based on their relevance, achieving state-of-the-art performance in multiple general-purpose QA tasks.

In open-ended medical QA tasks, a general framework for quantifying the uncertainty of free-form responses has yet to be established. An overview of our method is illustrated in Fig. 1. Generative inequality introduces many irrelevant words and sequences within the candidate responses for UQ, leading to biased uncertainty measurements when existing entropy-based methods treat all words and sequences equally. To address this issue, we propose **W**ord-**S**equence **E**ntropy (WSE), which allocates greater uncertainty proportion to keywords and enlarges the generative probability of reliable responses by assessing semantic relevance at both the word and sequence levels. Additionally, we leverage the concept of bi-directional entailment (Kuhn et al., 2023) — if two textual sequences logically imply each other, they are semantically similar — to develop a new method for measuring the semantic textual similarity between two sequences, which correlates with semantic relevance. Moreover, we investigate improving model accuracy by resampling based on the uncertainty measure, aiming to mitigate the limitations of LLMs in the medical domain.

We evaluate WSE utilizing multiple open-source pre-trained (e.g., LLaMA-7B Touvron et al., 2023a) and instruction-tuned (e.g., LLaMA-2-7B-Chat Touvron et al., 2023b, StableBeluga-7B Touvron et al., 2023b; Mukherjee et al., 2023 and Zephyr-7B-Alpha Tunstall et al., 2023) LLMs with the model size of 7B on five open-ended medical QA datasets (i.g., COVID-QA Möller et al., 2020, Medical Meadow MedQA Jin et al., 2020, PubMedQA Jin et al., 2019, MedMCQA Pal et al., 2022 and MedQuAD Ben Abacha and Demner-Fushman, 2019). Experimental results show that *WSE* outperforms six baseline methods (e.g., *WSE* surpasses *SAR* by 4.99% AUROC on the PubMedQA dataset). Furthermore, after filtering sequences with high uncertainty identified by *WSE*, we obtain a substantial improvement in model accuracy (e.g., +6.36% accuracy on the COVID-QA dataset, utilizing the Zephyr-7B-Alpha model), demonstrating the remarkable potential in real-world medical QA applications.

Our major contributions are summarized as follows:

- We investigate the phenomenon of generative inequality within the responses generated by LLMs in open-ended medical QA tasks and analyze its implications for uncertainty measurement.
- We propose Word-Sequence Entropy (WSE) to quantify the uncertainty of free-form answers in open-ended medical QA tasks for the first time.
- We conduct extensive experiments on five free-form medical QA datasets utilizing seven LLMs under two standard criteria for correctness evaluation, demonstrating that *WSE* surpasses six comparable baselines.
- Without requiring additional task-specific fine-tuning or architectural modifications, we improve the performance of LLMs, by resampling and applying responses with lower uncertainty, measured by *WSE*, as final answers, and obtain remarkable enhancement of model accuracy.

2. Related work

2.1. UQ in conventional NLP tasks

The concepts and approaches of UQ have been extensively explored and analyzed across various tasks (Hüllermeier and Waegeman, 2021), including machine translation (MT). To address data uncertainty from semantically equivalent translations, under-specification, and lower-quality training data in MT, Ott et al. (2018) assess whether references match the top model prediction or if most generated sequences align well with human translations. Considering the relationship between model probabilities and human judgments, Fomicheva et al. (2020) establish a strong correlation with human quality judgments through UQ techniques. Glushkova et al. (2021) address accumulated uncertainty from noisy scores, insufficient references, and out-of-domain text by incorporating Monte Carlo (MC) dropout (Gal and Ghahramani, 2016) and model ensembling (Lakshminarayanan et al., 2017), characterizing uncertainty through confidence intervals.

Due to limited work on calibration in a regression setting, Wang et al. (2022a) augment training data in low-resource scenarios and select instances based on UQ, addressing both the data and model predictive uncertainty. Malinin et al. (2020) also apply prior networks for interpretable UQ.

To enhance the reliability of decision-making in text classification tasks, Miok et al. (2019) quantify the predictive uncertainty utilizing MC dropout regularization (Gal and Ghahramani, 2016) and detect hate speech efficiently and reliably. Given the fundamental notion of epistemic uncertainty (EU) (Kendall and Gal, 2017) as a lack of knowledge, Lahlou et al. (2021) introduce the approach of direct epistemic uncertainty prediction (DEUP) and assess the excess risk as a measure of EU.

2.2. UQ in free-form NLG tasks

Distinguishing tasks with specific labels, such as misclassification detection (Vazhentsev et al., 2022) and text classification (Hu and Khan, 2021), it is challenging to implement uncertainty estimation in open-ended NLG tasks, where any output from LLMs sharing equivalent semantics with the standard answer can be considered correct.

The issue of truthfulness motivates uncertainty calibration for LLMs. Lin et al. (2022a) empower LLMs to self-evaluate the uncertainty of their answers in words via supervised fine-tuning. Meanwhile, Kadavath et al. (2022) adopt answer options from existing multiple-choice tasks and ask LLMs to determine if each answer is true or false. Both approaches prompt the language model itself to measure uncertainty with additional task-specific training. In a zero-resource setting, Manakul et al. (2023) attribute poor performance to variations in generating patterns. If the consistency score of multiple generations is low, it indicates high uncertainty. Motivated by the limited work on general uncertainty estimation for structured prediction, Malinin and Gales (2020) devise a novel measure of knowledge uncertainty by summing the predictive entropy over multiple outputs. Recently, to tackle the issue of semantic equivalence, Kuhn et al. (2023) propose to cluster semantically similar sequences and calculate the semantic entropy. The approach most connected with ours is *SAR* (Duan et al., 2023), which reassigns the weight of uncertainty associated with each token and sentence based on their respective relevance.

Compared to general-purpose QA tasks, medical QA with free-form responses is more domain-specific and often involves rare and compound technical terms. In such cases, LLMs adopt character-based tokenization, breaking a single word into multiple sub-tokens for processing. Analyzing each token independently, as done in *SAR*, can lead to inconsistency of semantic relevance within the same word, resulting in biased and unstable uncertainty measurements. Additionally, *SAR* relies on an external language model to measure semantic similarity, which lacks explainability and reliability due to the semantic complexity of medical QA. Given the absence of a robust and general approach to estimating uncertainty in open-ended medical QA tasks, we aim to address this gap by developing an estimator to inform users about the trustworthiness of output statements from LLMs.

3. Methodology

3.1. Preliminaries

Conditioned on a medical query x, LLMs progressively predict the probability distribution of the next token based on previous tokens and generate free-form textual sequences in an auto-regressive fashion. Following prior work (Kuhn et al., 2023; Duan et al., 2023), we generate K responses to the same query and estimate the predictive uncertainty of the current QA process within the generated set $\mathbb{S} = \{\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_K\}$, where $\mathbf{s}_i$ refers to the ith response. We denote the jth word within the textual sequence $\mathbf{s}_i$ as w_{ij}, and the kth token in w_{ij} as z_{ijk}. Additionally, we denote the number of words within $\mathbf{s}_i$ by N_i, the number of tokens in w_{ij} by M_j and the total number of tokens within $\mathbf{s}_i$ by T_i (i.e., $T_i = \sum_j^{N_i} M_j$). Prompted by x, we define the probability of generating z_t as $p\left(z_t \mid \mathbf{z}_{<t}, x\right)$, where $\mathbf{z}_{<t}\ (t \in T_i)$ refers to previously generated tokens within the ith textual sequence. In subsequent research, we simplify $p\left(z_t \mid \mathbf{z}_{<t}, x\right)$ to $p\left(z_t\right)$ to represent the generative probability of the tth token.

3.2. Generative inequality in free-form medical query responses

To investigate the issue of generative inequality in open-ended medical QA tasks, we leverage the popular Predictive Entropy (PE) (Kadavath et al., 2022) as the fundamental method for UQ. Given $\mathbf{s}_i$, we first calculate the token-wise entropy of z_t based on its generative probability:

$$E_T\left(z_t\right) = -\log p\left(z_t\right). \tag{1}$$

Then, PE calculates the sequence-wise entropy of $\mathbf{s}_i$ by summing the token-wise entropy of all tokens within it with equal weights:

$$E_S\left(\mathbf{s}_i\right) = \sum_t^{T_i} E_T\left(z_t\right). \tag{2}$$

The predictive uncertainty or entropy of the current QA process is obtained by directly computing the average of the sequence-wise entropy of these K candidate responses:

$$E\left(\mathbb{S}\right) = \frac{1}{K}\sum_i^{K} E_S\left(\mathbf{s}_i\right). \tag{3}$$

In this context, it is apparent that the token-wise entropy represents the uncertainty committed by individual tokens, the sequence-wise entropy captures the predictive uncertainty of each textual sequence (i.e., response), and PE quantifies the complexity encompassing the generated set (i.e., an approximation of the model's output space), which characterizes the overall uncertainty of the current decision-making process for medical queries.

Analogous to the formulation of token-wise entropy in Eq. (1), the sequence-wise entropy of $\mathbf{s}_i$ can be expressed as its log-probability:

$$E_S\left(\mathbf{s}_i\right) = -\log p\left(\mathbf{s}_i\right), \tag{4}$$

where $p\left(\mathbf{s}_i\right)$ reflects the probability of the ith sequence and is obtained by multiplying the probabilities of all tokens within $\mathbf{s}_i$ (i.e., $\prod_t^{T_i} p\left(z_t\right)$).

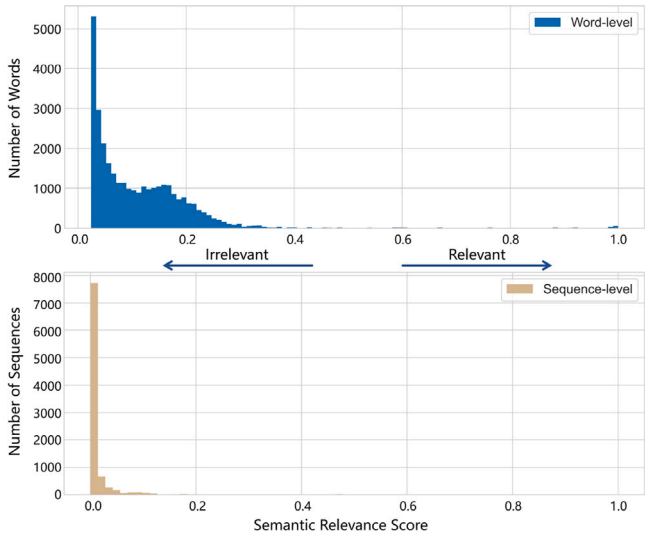


Fig. 2. Distribution of semantic relevance scores at both the word and sequence levels. The entire generated set contains a considerable proportion of irrelevant words and sequences (i.e., generative inequality).

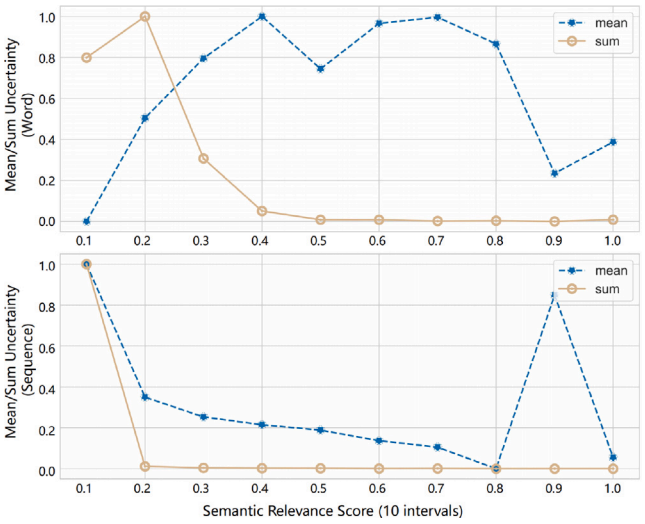


Fig. 3. Correlation between the semantic relevance and uncertainty proportion at both the word and sequence levels. Irrelevant words and sequences account for the primary source of uncertainty within the generated set (responses) in general.

3.2.1. Relevance

To analyze generative inequality at the word level, where keywords (e.g., "Mother-to-child transmission" in the sentence "Mother-to-child transmission is the primary cause of HIV-1 infection in children worldwide".) may account for a limited proportion of the overall uncertainty within the current response, we first assess the semantic relevance of each word by measuring the textual similarity between the query-answer pairs before and after removing the evaluated word. A lower similarity score signifies a significant semantic variation, indicating that the word carries more semantic information within the current textual sequence (i.e., a keyword).

Following *SAR* (Duan et al., 2023), we evaluate textual similarity utilizing a cross-encoder model provided by the SentenceTransformers library (Reimers and Gurevych, 2019), with RoBERTa-large (Liu et al., 2019) as the backbone. The model processes sentence pairs and generates similarity scores. However, relying solely on an external language model for textual similarity evaluation is unreliable and lacks explainability, because embeddings of sentences encoded by the model, in which all semantic information is mixed in fixed-length vectors, are limited in the semantic representation (Wang and Yu, 2023). Inspired by bi-directional entailment (Kuhn et al., 2023), we leverage a Natural Language Inference (NLI) classifier, DeBERTa-large-mnli (He et al., 2020), for this task. The model takes sequence pairs as the input and predicts scores (logits) for three classes of semantic relationship: entailment, neutral, and contradiction. We employ the probability of entailment as the similarity measure.

For simplicity, we define $\mathbf{s}_i \setminus w_{ij}$ as the representation for removing the jth word from the ith response and $\cup$ as the concatenation of the prompt and answer. The measurement of textual similarity is formulated as:

$$\begin{cases} S_C = f_{ce}\left(x \cup \mathbf{s}_i, x \cup \mathbf{s}_i \setminus w_{ij}\right) \\ S_L = f_{ent}\left(x \cup \mathbf{s}_i, x \cup \mathbf{s}_i \setminus w_{ij}, c\right), \end{cases} \tag{5}$$

where $f_{ce}(\cdot)$ represents the utilization of the *cross-encoder* model to compute the textual similarity score between two sequences directly, $f_{ent}(\cdot)$ refers to obtaining the probability of *entailment* extracted from the logit vector, which falls within the range of 0 to 1 after being scaled by the $softmax$ function, and c is leveraged to control the smoothness of the logit vector.

Given that the employed language models (Liu et al., 2019; He et al., 2020) are not specifically pre-trained for the medical domain, consistently high similarity can lead to low semantic relevance for all words within the current textual sequence, thereby failing to capture keywords. We adopt a conservative strategy by selecting the smaller value from the two measures in Eq. (5), which mitigates potential instability arising from extreme similarity quantification and task-specific limitations. Then, the word-level semantic relevance score of the jth word within the ith response can be formulated as:

$$R_W\left(w_{ij}\right) = 1 - \min\left(S_C, S_L\right). \tag{6}$$

In the end, we assign the same relevance score to all tokens in w_{ij} as the word itself (i.e., the token-level semantic relevance score), to maintain the consistency of semantic relevance within a single word:

$$R_T\left(z_{ijk}\right) = R_W\left(w_{ij}\right)\left(k \in M_j\right). \tag{7}$$

Formally, it can be observed that if the ith textual sequence exhibits significant semantic variation before and after removing the jth word, then the semantic relevance score of all tokens in w_{ij} are deemed to be high.

In open-ended medical QA tasks, we generate multiple (i.e., K) responses to the same query to estimate the uncertainty of the current QA process, and there can be many irrelevant responses with limited semantic information. However, *PE*, as described in Eq. (3), calculates the average of the sequence-wise entropy of all responses within the generated set. To investigate this issue, we define the semantic relevance at the sequence level.

Building on the self-consistency hypothesis[3] (Wang et al., 2022b), we suggest that responses, which maintain strong semantic consistency with others among the set of K candidate responses, are more trustworthy. We employ the identical approaches described in Eq. (5) to measure the textual similarity between any two textual sequences. Then, the sequence-level relevance score of $\mathbf{s}_i$ is formulated as the accumulation of the textual similarity scores, re-weighted by the generative probability of the compared responses:

$$R_S\left(\mathbf{s}_i\right) = \sum_{l \neq i}^{K} S\left(\mathbf{s}_l, \mathbf{s}_i\right) p\left(\mathbf{s}_l\right), \tag{8}$$

where $S(\cdot,\cdot)$ represents the smaller similarity score obtained from the two measurements in Eq. (5), and $\mathbf{s}_l$ denotes the lth textual sequence that differs from $\mathbf{s}_i$ in the K generated responses. A higher probability of

[3] Self-consistency hypothesis states that a repetitively sampled response is viewed as a form of consistency linked to higher confidence in the response.

$\mathbf{s}_l$ (i.e., $p(\mathbf{s}_l)$) augments the persuasiveness of textual similarity between $\mathbf{s}_i$ and $\mathbf{s}_l$.

3.2.2. Uncertainty

As mentioned previously, the token-wise entropy reflects the uncertainty committed by each token (i.e., $E_T(z_t)$ in Eq. (1)), and the overall uncertainty of the ith response can be calculated by aggregating the token-wise entropy of all words within the entire textual sequence (i.e., $E_S(\mathbf{s}_i)$ in Eq. (2)). To ascertain how much uncertainty is induced by individual words, we compute the word-wise entropy of w_{ij} based on Eq. (1):

$$E_W(w_{ij}) = \sum_{k}^{M_j} -\log p(z_{ijk}), \tag{9}$$

where $p(z_{ijk})$ refers to the probability of generating z_{ijk} as the kth token in the jth word within the ith response. Then we calculate the ratio of the word-wise entropy and the sequence-wise entropy to determine the proportion of uncertainty stemming from the jth word within the ith response.:

$$P_W(w_{ij}, \mathbf{s}_i) = \frac{E_W(w_{ij})}{E_S(\mathbf{s}_i)}. \tag{10}$$

Similar to the word-wise situation, we formulate the uncertainty proportion of the ith response in the set of K generated responses (i.e., $\mathbb{S}$) as:

$$P_S(\mathbf{s}_i, \mathbb{S}) = \frac{E_S(\mathbf{s}_i)}{\sum_l^K E_S(\mathbf{s}_l)}. \tag{11}$$

3.2.3. Correlation analysis

To characterize generative inequality in open-ended medical QA tasks, we employ the MedMCQA dataset, with LLaMA-2-7B-Chat-HF serving as the generator. Given each medical query, we generate five responses (i.e., $K = 5$), and the max length of each sequence is set to 128 (i.e., $T_i \leqslant 128$). We first leverage Eqs. (6) and (7) to outline the distributions of word-level and sequence-level semantic relevance scores. Results are depicted in Fig. 2. Within the generated set, a considerable proportion of words exhibit low semantic relevance (i.e., irrelevant), and only a limited subset of words conveys the primary semantic information. At the sequence level, the prevalence of irrelevant responses significantly outweighs those with meaningful content.

When conducting UQ, we should prioritize keywords and reliable textual sequences. To explore the issue of generative inequality, we analyze the correlation between semantic relevance and uncertainty proportion. We divide relevance scores into ten equal intervals. Within each interval, we calculate the sum and average uncertainty of all words or sequences. Results at both the word and sequence levels are illustrated in Fig. 3. Irrelevant words contribute significantly to the overall uncertainty. At the sequence level, both the mean and sum uncertainty of irrelevant sequences are prominent.

Given the substantial proportion of irrelevant words and textual sequences within the generated set, this can introduce unexpected biases and instability when measuring the uncertainty of LLMs-generated answers in real-world open-ended medical QA applications. To address these issues, we propose a novel UQ method in the following text.

3.3. Word-sequence entropy

In light of the observed issues arising from generative inequality, as demonstrated in Section 3.2.3, we propose to emphasize keywords and more semantically relevant responses within the candidate set when conducting UQ. To maintain coherence and consistency in the presentation, we strictly adhere to the symbol conventions utilized in Sections 3.1 and 3.2.

3.3.1. Word-level WSE

Since the semantic information carried by each word (token) differs, treating all tokens equally, as described in Eq. (2), will lead to biased measurements of the predictive uncertainty within each textual sequence. To address this, we highlight tokens in keywords by directly multiplying the token-wise entropy by the token-level semantic relevance score:

$$U_T(z_{ijk}) = E_T(z_{ijk}) R_T(z_{ijk}), \tag{12}$$

where $E_T(z_{ijk})$ refers to the token-wise entropy of the kth token in the jth word within the ith response. Then, the calibrated word-wise entropy of w_{ij} can be formulated as:

$$U_W(w_{ij}) = \sum_{k}^{M_j} U_T(z_{ijk}). \tag{13}$$

To quantify the overall uncertainty of $\mathbf{s}_i$, we sum the calibrated (weighted) uncertainty of all words within this textual sequence:

$$U_S(\mathbf{s}_i) = \sum_{j}^{N_i} U_W(w_{ij}). \tag{14}$$

Finally, the word-level WSE is defined as the arithmetic mean uncertainty of these K candidate responses, following PE:

$$WSE_W(\mathbb{S}) = \frac{1}{K}\sum_{i}^{K} U_S(\mathbf{s}_i). \tag{15}$$

By employing the word-level WSE, we capture and highlight keywords carrying the main semantic information within the current textual sequence, thereby calibrating the predictive uncertainty of each candidate response.

3.3.2. Sequence-level WSE

As noted in Section 3.2, responses, which are semantically consistent with others in the set of K candidate responses, are more trustworthy. We reduce the uncertainty associated with the ith textual sequence by adding its generative probability to its semantic relevance score, after dividing by a constant d, to obtain the calibrated sequence-wise entropy of $\mathbf{s}_i$:

$$U_S'(\mathbf{s}_i) = -\log\left(p(\mathbf{s}_i) + \frac{R_S(\mathbf{s}_i)}{d}\right), \tag{16}$$

where d serves to regulate the extent to which the semantic relevance score influences the generative probability. Operating in the same way as Eq. (15), the sequence-level WSE is formulated as:

$$WSE_S(\mathbb{S}) = \frac{1}{K}\sum_{i}^{K} U_S'(\mathbf{s}_i). \tag{17}$$

By employing the sequence-level WSE, we enlarge the generative probability of more reliable responses by assessing their semantic relevance, thereby calibrating the overall uncertainty of the current QA process.

3.3.3. Integrated WSE

Given the direct mathematical relation between the probability and entropy of $\mathbf{s}_i$, as defined in Eq. (4), we replace $p(\mathbf{s}_i)$ in Eq. (16) with $e^{-U_S(\mathbf{s}_i)}$, where $U_S(\mathbf{s}_i)$ represents the calibrated uncertainty of $\mathbf{s}_i$ described in Eq. (14). Since the sequence-level semantic relevance score of $\mathbf{s}_i$ (i.e., $R_S(\mathbf{s}_i)$) is determined by the probability of compared sequences, we replace $p(\mathbf{s}_l)$ as defined in Eq. (8). Then, the combined WSE, which calibrates the uncertainty at both the word and sequence levels, is formulated as:

$$WSE_C(\mathbb{S}) = \frac{1}{K}\sum_{i}^{K} -\log\left(p_i' + \frac{\sum_{l\neq i}^{K} S_{li} p_l'}{d}\right), \tag{18}$$

where p_i' and p_l' refer to the replacements of the generative probability, and S_{li} represents the semantic textual similarity between s_l and s_i.

Algorithm 1: The pseudo-code for the combined *WSE*.

Input: $\mathbb{S}$, $\mathbf{s}_i$, $\mathbf{s}_{l\neq i}$, w_{ij}, z_{ijk}, K, N_i, M_j, $p\left(z_{ijk}\right)$, d.

for $i \leftarrow 1$ **to** K **do**
  for $j \leftarrow 1$ **to** N_i **do**
    Calculate the semantic textual similarity of $\mathbf{s}_i$ before and after removing w_{ij} $\leftarrow S_W\left(w_{ij}\right)$;
    $R_W\left(w_{ij}\right) \leftarrow 1 - S_W\left(w_{ij}\right)$;
    for $k \leftarrow 1$ **to** M_j **do**
      $R_T\left(z_{ijk}\right) \leftarrow R_W\left(w_{ij}\right)$; ▷ *consistent semantic relevance*
      $E_T\left(z_{ijk}\right) \leftarrow -\log p\left(z_{ijk}\right)$;
      $U_T\left(z_{ijk}\right) \leftarrow E_T\left(z_{ijk}\right) R_T\left(z_{ijk}\right)$.
    $U_W\left(w_{ij}\right) \leftarrow \sum_k^{M_j} U_T\left(z_{ijk}\right)$. ▷ *word-level*
  $U_S\left(s_i\right) \leftarrow \sum_j^{N_i} U_W\left(w_{ij}\right)$;
  $p\left(s_i\right) \leftarrow e^{-U_S\left(s_i\right)}$. ▷ *calibrated generative probability*
for $i \leftarrow 1$ **to** K **do**
  for $l \leftarrow 1$ **to** K **do**
    if $l \neq i$ **then**
      Calculate the semantic textual similarity between s_l and s_i $\leftarrow S_S\left(s_l, s_i\right)$.
  $R_S\left(s_i\right) \leftarrow \sum_{l\neq i}^{K} S_S\left(s_l, s_i\right) p\left(s_l\right)$;
  $U_S{}'\left(s_i\right) \leftarrow -\log\left(p\left(s_i\right) + \frac{R_S\left(s_i\right)}{d}\right)$. ▷ *sequence-level*
$WSE_C\left(\mathbb{S}\right) \leftarrow \frac{1}{K}\sum_i^K U_S{}'\left(s_i\right)$.

Output: Calibrated predictive uncertainty $WSE_C\left(\mathbb{S}\right)$.

Moreover, The pseudocode of the combined *WSE* is summarized in Algorithm 1.

In terms of computational complexity, we first analyze *PE*. As described in Eqs. (2)–(3), its computational complexity is $\mathcal{O}(T+K) = \mathcal{O}(\max(T,K))$, where T is the number of tokens within the response. For *SAR*, since it assesses the relevance of each token and sentence within the K generated responses, its computational complexity is $\mathcal{O}\left(KT+K^2\right) = \mathcal{O}(\max(T,K)\,K)$. Here, KT refers to assessing the relevance of individual tokens within these K responses, and K^2 refers to analyzing the similarity between every pair of responses. *WSE* assesses semantic relevance as both the word and sequence levels, with the computational complexity of $\mathcal{O}\left(KN+K^2\right) = \mathcal{O}(\max(N,K)\,K)$, where N is the number of words within the response, and KN refers to measuring the semantic variation of the K responses before and after removing each word. Since a single word can be composed of multiple tokens (i.e., $N \leq T$), *WSE* has a lower computational complexity compared to *SAR*.

By strategically calibrating the uncertainty proportion of keywords and elevating the generative probability of semantically analogous (i.e., more trustworthy) responses, *WSE* focuses more on significant words and responses when estimating the uncertainty of free-form answers generated by LLMs, effectively mitigating biases caused by generative inequality. In the latter part of the experiments, we denote word-level WSE, sequence-level WSE, and combined WSE by WSE_W, WSE_S, and WSE_C, respectively.

4. Experiments

In this section, we evaluate the performance of WSE in accurately measuring the uncertainty of LLMs-generated answers in open-ended medical QA tasks. Given the potential for real-world healthcare applications, we resample responses by employing the generation with the lowest uncertainty within the candidate set, measured by WSE, as the final output to the current medical query, and investigate the overall enhancement of model accuracy.

4.1. Experiment setup

4.1.1. Performance evaluation

Following Semantic Entropy (SE) (Kuhn et al., 2023; Quevedo et al., 2024) and *SAR* (Duan et al., 2023), we evaluate *WSE* by framing UQ as the problem of predicting whether to trust a model generation for a given medical query. We employ the widely used area under the receiver operating characteristic (AUROC) curve for the binary event that a given response is incorrect, which captures both precision and recall, ranging from 0 to 1, with 1 representing a perfect classifier and 0.5 representing an uninformative estimator. This metric evaluates whether *WSE* can effectively distinguish between correct and incorrect answers across various uncertainty thresholds. Additionally, since there can be unrealistic uncertainty thresholds, we employ deep AUROC (Carrington et al., 2023), which measures performance in multiple groups of predicted risk, or groups of true positive rate or false positive rate.

4.1.2. Correctness evaluation

We adopt two standard metrics to evaluate the correctness of responses: RS (RS) (Lin, 2004) and SS (SS) (Duan et al., 2023). RS measures the longest common subsequence between the output and reference answer, serving as a fuzzy matching criterion. For SS, we utilize the cross-encoder model mentioned in Section 3.2.1, with DistillRoBERTa (Sanh et al., 2019) as the backbone. SS corresponds to the semantic textual similarity denoted by S_C in Eq. (5). We consider the generation correct if either the RS or SS exceeds the predefined threshold of 0.5. Notably, we employ the most likely generation, as introduced in Section 4.1.6, as the object to evaluate the correctness of the current QA. In Section 4.2.2, we will analyze the sensitivity of *WSE* to these threshold values.

4.1.3. Model

We conduct experiments on seven open-source "off-the-shelf" LLMs provided by the Hugging Face platform, including both pre-trained LLMs (e.g., LLaMA-7B Touvron et al., 2023a) and instruction-tuned LLMs (e.g., LLaMA-2-7B-Chat Touvron et al., 2023b, Mistral-v0.1 Jiang et al., 2023, Zephyr-7B-Alpha Tunstall et al., 2023, Vicuna-7B-v1.5 Zheng et al., 2023, WizardLM-7B Xu et al., 2023, StableBeluga-7B Touvron et al., 2023b; Mukherjee et al., 2023) with the model size of 7B.

4.1.4. Datasets

We utilize five free-form medical QA datasets: COVID-QA (Möller et al., 2020), Medical Meadow MedQA (Jin et al., 2020), PubMedQA (Jin et al., 2019), MedMCQA (Pal et al., 2022) and MedQuAD (Ben Abacha and Demner-Fushman, 2019). COVID-QA consists of 2019 query-answer pairs related to COVID-19, and we employ all query-answer pairs within the maximum sequence length allowed by the language model. MedMCQA is a large-scale, multiple-choice QA dataset for medical entrance exams, and we select all samples where a question has only one correct option and begins with "what" or "which" (1895 in total). Medical Meadow MedQA is a free-form multiple-choice OpenQA dataset for solving medical problems, collected from the professional medical board exams. MedQuAD covers 37 question types associated with diseases, drugs, and other medical entities such as tests. For Medical Meadow MedQA and MedQuAD, we randomly select 2000 test samples from the validation set. PubMedQA is a novel biomedical QA dataset collected from PubMed abstracts and we employ the full test set (1000 question-answer pairs).

Unlike COVID-QA and PubMedQA, Medical Meadow MedQA, MedMCQA, and MedQuAD do not provide contextual information, and we randomly select five fixed query-answer pairs from each dataset to form the few-shot prompts, enabling LLMs to follow the instructions.

Table 1
Comparison of WSE_W, WSE_S, WSE_C, and six baseline methods utilizing seven pre-trained and instruction-tuned LLMs on five free-form medical QA datasets, employing SS as the criterion for correctness evaluation with the threshold set to 0.5 (AUROC).

Datasets	LLMs	*LS*	*PE*	*SE*	Token-*SAR*	Sent-*SAR*	*SAR*	WSE_W	WSE_S	WSE_C
COVID-QA	LLaMA-7B	0.5076	0.7348	0.7032	0.6903	0.7180	0.7142	0.7448	0.7319	**0.7454**
	LLaMA-2-7B-Chat	0.4422	0.6756	0.6716	0.6640	0.6765	0.6589	**0.6869**	0.6767	0.6846
	Mistral-7B-v0.1	0.4341	0.7278	0.7027	0.6911	0.7166	0.7209	0.7318	0.7327	**0.7482**
	Zephyr-7B-Alpha	0.4147	0.6607	0.6583	0.6483	0.6655	0.6558	0.6643	0.6609	**0.6696**
	WizardLM-7B	0.4059	0.6951	0.6840	0.6737	0.6897	0.6593	**0.7076**	0.6948	0.7016
	Vicuna-7B-v1.5	0.4021	0.6955	0.6882	0.6826	0.7011	0.6914	**0.7159**	0.6971	0.7130
	StableBeluga-7B	0.4438	0.6904	0.7083	0.6986	0.7027	0.6962	0.7121	0.7068	**0.7228**
Average		0.4358	0.6971	0.6880	0.6784	0.6957	0.6857	0.7091	0.7001	**0.7122**
MedQA	LLaMA-7B	0.5143	0.5122	0.5493	0.4789	0.5468	0.5130	0.5164	0.5438	**0.5502**
	LLaMA-2-7B-Chat	0.5483	0.5793	0.5958	0.5805	0.5948	0.6145	0.6102	0.6074	**0.6415**
	Mistral-7B-v0.1	0.5355	0.4845	0.5119	0.4915	0.5085	0.5517	0.5185	0.5506	**0.5782**
	Zephyr-7B-Alpha	0.5035	0.4979	0.5206	0.4936	0.5043	0.5251	0.5192	0.5326	**0.5619**
	WizardLM-7B	0.5985	0.4631	0.5684	0.4836	0.6286	0.5517	0.5499	**0.6314**	0.6211
	Vicuna-7B-v1.5	0.5079	0.4538	0.4752	0.5093	0.4510	0.5335	0.5295	0.4746	**0.5576**
	StableBeluga-7B	**0.5776**	0.5139	0.5481	0.5318	0.5749	0.5696	0.5474	0.5758	0.5749
Average		0.5408	0.5007	0.5385	0.5099	0.5441	0.5513	0.5416	0.5595	**0.5836**
MedMCQA	LLaMA-7B	0.5468	0.5290	0.5415	0.5394	0.5583	0.5399	0.5498	**0.5586**	0.5548
	LLaMA-2-7B-Chat	0.5108	0.4954	0.5015	0.5128	0.4833	0.5200	0.5467	0.5030	**0.5612**
	Mistral-7B-v0.1	0.5075	0.4909	0.5216	0.5205	0.4980	0.5523	0.5146	0.5584	**0.5777**
	Zephyr-7B-Alpha	0.4831	0.5175	0.5404	0.5356	0.5331	0.5374	0.5259	**0.5534**	0.5512
	WizardLM-7B	0.5320	0.4980	0.5074	0.4957	0.5025	0.5063	0.5517	0.5149	**0.5623**
	Vicuna-7B-v1.5	0.5016	0.4952	0.5015	0.5011	0.4803	0.5065	0.5288	0.4975	**0.5395**
	StableBeluga-7B	0.4990	0.4446	0.4833	0.4446	0.4655	0.5305	0.4421	0.5125	**0.5314**
Average		0.5115	0.4958	0.5139	0.5071	0.5030	0.5276	0.5228	0.5283	**0.5540**
PubMedQA	LLaMA-7B	0.5496	0.5424	0.6202	0.5414	0.6129	0.6269	0.5420	**0.6343**	0.6340
	LLaMA-2-7B-Chat	0.5024	0.6146	0.5918	0.5676	0.5819	0.6176	0.5736	0.6179	**0.6640**
	Mistral-7B-v0.1	0.5018	0.6440	0.6644	0.5262	0.6614	0.6022	0.5808	**0.6980**	0.6627
	Zephyr-7B-Alpha	0.5929	0.5682	0.5706	0.4894	0.5594	0.5310	0.5293	0.5793	**0.6027**
	WizardLM-7B	0.5587	0.5308	0.5525	0.4676	0.5265	0.5172	0.5080	0.5640	**0.6031**
	Vicuna-7B-v1.5	0.5787	0.6631	0.6728	0.5715	0.6617	0.6289	0.6112	0.6670	**0.6869**
	StableBeluga-7B	0.6075	0.6598	0.6461	0.6662	0.6419	0.6971	0.6664	0.6754	**0.7169**
Average		0.5559	0.6033	0.6169	0.5471	0.6065	0.603	0.5730	0.6337	**0.6529**
MedQuAD	LLaMA-7B	0.6546	0.5996	0.6040	0.6534	0.6446	0.6491	**0.6618**	0.6365	0.6502
	LLaMA-2-7B-Chat	0.5758	0.4889	0.5123	0.5743	0.5484	0.5884	0.5879	0.5364	**0.5890**
	Mistral-7B-v0.1	0.5838	**0.6091**	0.5409	0.578	0.5639	0.5718	0.5823	0.5643	0.5847
	Zephyr-7B-Alpha	0.5718	0.5012	0.6283	0.6732	0.6393	0.6673	**0.6817**	0.6327	0.6756
	WizardLM-7B	0.5866	0.4447	0.5405	0.5958	0.5613	0.5871	**0.6112**	0.5596	0.6003
	Vicuna-7B-v1.5	0.5748	0.4469	0.5652	0.6357	0.5792	0.6249	**0.6493**	0.5727	0.6301
	StableBeluga-7B	0.5671	0.5226	0.5887	0.5960	0.5738	0.5732	**0.608**	0.5634	0.5737
Average		0.5878	0.5161	0.5686	0.6152	0.5872	0.6088	**0.6260**	0.5808	0.6148
Overall		0.5264	0.5626	0.5852	0.5715	0.5873	0.5952	0.5945	0.6005	**0.6235**

4.1.5. Baselines

We compare our method with PE (Kadavath et al., 2022), Semantic Entropy (*SE*) (Kuhn et al., 2023), Lexical Similarity (LS) (Lin et al., 2022b), Token-level *SAR* (Token-SAR), Sentence-level *SAR* (Sent-SAR), and *SAR* (Duan et al., 2023). *PE* quantifies uncertainty as described in Section 3.2. *SE* considers semantic equivalence and calculates the cluster-wise entropy. *LS* computes the mean semantic similarity score of responses in the set of K generated responses. Token-SAR and Sent-SAR reallocate the uncertainty weights of tokens and sentences based on their relevance, respectively. SAR combines Token-SAR and Sent-SAR. For domain-specific medical QA, WSE_W highlights keywords in each response by assessing the semantic relevance of each word based on semantic variation, which addresses the issue of inconsistent semantic relevance within individual words that *SAR* encounters. Additionally, WSE_S leverages a more reliable and explainable measure for semantic textual similarity and enlarges the generative probability of more trustworthy responses based on self-consistency. Similar to *SAR*, WSE_C is an orthogonal combination of WSE_W and WSE_S, which calibrates uncertainty at both the word and sequence levels.

4.1.6. Hyperparameters

Given each medical query, LLMs generate five free-form responses (i.e., $K = 5$) via multinomial sampling, which are then employed for UQ. For the correctness evaluation of the current QA, we employ beam search to obtain the most likely generation (Kuhn et al., 2023; Duan et al., 2023). The temperature is fixed at 0.5 for all LLMs, and the max length of each generation is set to 128 tokens. The coefficient c in Eq. (5) is set to 1.0 by default, and the denominator d in the relevance-controlled quantity in Eq. (18) is empirically set to 0.001.

4.2. Empirical findings

4.2.1. Uncertainty estimation

Given the integrated measurement of semantic textual similarity described in Section 3.2.1, we compare WSE_W, WSE_S, and WSE_C with six baseline methods, utilizing SS as the criterion for correctness evaluation. As summarized in Table 1, all the three *WSE* variants outperform the baseline methods significantly, with WSE_C achieving the highest overall AUROC of 0.6235. By highlighting keywords and addressing the inconsistency of semantic relevance within each word, WSE_W surpasses Token-*SAR* by 2.3% in AUROC overall, particularly on the MedQA dataset, where it exceeds Token-*SAR* by 3.17%. By

Table 2
Comparison of WSE_W, WSE_S, WSE_C, and six baseline methods utilizing seven pre-trained and instruction-tuned LLMs on five free-form medical QA datasets, employing RS as the criterion for correctness evaluation with the threshold set to 0.5 (AUROC).

Datasets	LLMs	*LS*	*PE*	*SE*	Token-*SAR*	Sent-*SAR*	*SAR*	WSE_W	WSE_S	WSE_C
COVID-QA	LLaMA-7B	0.5726	0.7297	0.7114	0.6735	0.7159	0.7047	0.7108	0.7304	**0.7445**
	LLaMA-2-7B-Chat	0.4676	0.7164	0.7148	0.7103	0.7174	0.7098	0.7255	0.7223	**0.7324**
	Mistral-7B-v0.1	0.5403	0.6368	0.6530	0.6697	0.6367	0.6418	**0.7207**	0.6432	0.6922
	Zephyr-7B-Alpha	0.5748	0.6344	0.6445	0.6015	0.6381	0.6120	0.6416	0.6458	**0.6524**
	WizardLM-7B	0.5591	0.6455	**0.6623**	0.6401	0.6345	0.5569	0.6486	0.6341	0.5779
	Vicuna-7B-v1.5	0.4898	0.6699	0.6972	0.6981	0.6959	0.6405	0.7051	0.6936	**0.7528**
	StableBeluga-7B	0.5795	0.6730	0.6712	0.6647	0.6744	0.6376	**0.6839**	0.6744	0.6528
MedQA	LLaMA-7B	0.5162	0.5191	0.5620	0.4725	0.5730	0.5178	0.5257	0.5679	**0.5739**
	LLaMA-2-7B-Chat	0.5714	0.5874	0.6194	0.5666	0.6265	0.6192	0.6167	0.6364	**0.6581**
	Mistral-7B-v0.1	0.5456	0.5246	0.5358	0.4952	0.5541	0.5706	0.5004	0.5826	**0.5828**
	Zephyr-7B-Alpha	0.4897	0.5278	0.5451	0.4987	0.5327	0.5400	0.5212	0.5633	**0.5803**
	WizardLM-7B	0.6246	0.4668	0.5755	0.4808	0.6161	0.5565	0.5461	**0.6285**	0.6268
	Vicuna-7B-v1.5	0.5154	0.4967	0.5085	0.5128	0.4937	0.5426	0.5274	0.5126	**0.5608**
	StableBeluga-7B	0.5860	0.5378	0.5741	0.5522	0.6048	0.5949	0.5687	0.6074	**0.6097**
MedMCQA	LLaMA-7B	0.5596	0.5182	0.5511	0.5347	0.5693	0.5403	0.5463	**0.5699**	0.5589
	LLaMA-2-7B-Chat	0.5030	0.4988	0.5012	0.5347	0.4881	0.5453	0.5544	0.5076	**0.5720**
	Mistral-7B-v0.1	0.5293	0.5307	0.5382	0.5295	0.5402	0.5652	0.5168	0.5760	**0.5781**
	Zephyr-7B-Alpha	0.4801	0.5500	0.5896	0.5659	0.5842	0.5718	0.5536	**0.6103**	0.5921
	WizardLM-7B	0.5151	0.5051	0.5000	0.5124	0.5005	0.5268	0.5381	0.5095	**0.5395**
	Vicuna-7B-v1.5	0.5048	0.4983	0.4937	0.5182	0.4843	0.5311	0.5304	0.5031	**0.5499**
PubMedQA	LLaMA-7B	0.5123	0.5457	0.5401	0.5423	0.5203	0.5426	0.5403	0.5451	**0.5540**
	LLaMA-2-7B-Chat	0.6511	0.6146	0.5867	0.6329	0.5726	0.7053	0.6420	0.6028	**0.7329**
	Mistral-7B-v0.1	0.5172	0.5331	0.5231	0.5093	0.5131	0.5133	0.5094	0.5654	**0.5659**
	Zephyr-7B-Alpha	0.4945	0.6194	0.4433	0.6103	0.4408	0.5737	**0.6640**	0.4754	0.6545
	Vicuna-7B-v1.5	0.7465	0.3397	0.3838	0.8246	0.3647	0.7926	**0.8888**	0.3477	0.6283
MedQuAD	LLaMA-7B	0.7442	0.7123	0.6611	0.6821	0.7126	0.7161	0.7216	0.7108	**0.7470**
	LLaMA-2-7B-Chat	0.8667	0.8783	0.8353	0.9215	0.8413	0.9065	**0.9527**	0.8423	0.9108
	Mistral-7B-v0.1	0.7893	0.7985	0.6101	0.8292	0.6700	0.8072	**0.8394**	0.7291	0.7040
	WizardLM-7B	0.2327	0.9816	0.9718	0.9843	0.9775	0.9721	**0.9889**	0.9746	0.9561
	Vicuna-7B-v1.5	0.9065	0.9020	0.9014	0.9266	0.9035	0.9275	**0.9356**	0.9125	0.9142
Overall		0.5729	0.6131	0.6102	0.6298	0.6132	0.6394	0.6522	0.6275	**0.6585**

Table 3
Comparison of *WSE* and *SAR* utilizing seven popular LLMs on the COVID-QA datasets, employing both SS and RS as the criteria for correctness evaluation with a more stringent threshold set to 0.7 (deep AUROC).

Metrics	LLMs	Token-*SAR*	Sent-*SAR*	*SAR*	WSE_W	WSE_S	WSE_C
RS	LLaMA-7B	0.6846	0.7217	0.7280	0.7194	0.7397	**0.7423**
	LLaMA-2-7B-Chat	0.5774	0.6501	0.6234	0.7046	**0.7309**	0.7297
	Mistral-7B-v0.1	0.6486	0.6052	0.6188	**0.6526**	0.6205	0.6352
	Zephyr-7B-Alpha	0.5373	0.6087	0.5120	0.6144	**0.6253**	0.6237
	WizardLM-7B	0.7645	0.7277	0.7038	**0.8092**	0.7835	0.7944
	Vicuna-7B-v1.5	0.6490	0.6637	0.5336	0.6691	**0.6842**	0.6216
	StableBeluga-7B	0.7942	0.6860	0.6658	**0.8113**	0.7035	0.6703
Average		0.6651	0.6662	0.6265	**0.7115**	0.6982	0.6882
SS	LLaMA-7B	0.5523	0.6775	0.5828	0.7410	0.6871	**0.7468**
	LLaMA-2-7B-Chat	0.5372	0.5213	0.5703	0.6909	0.5484	**0.6972**
	Mistral-7B-v0.1	0.5476	0.6643	0.5724	**0.7556**	0.6657	0.7425
	Zephyr-7B-Alpha	0.4927	0.6759	0.5218	0.6983	0.6741	**0.7211**
	WizardLM-7B	0.6129	0.6584	0.6354	0.6242	**0.6691**	0.6525
	Vicuna-7B-v1.5	0.6508	0.6783	0.6766	0.6979	0.6829	**0.7010**
	StableBeluga-7B	0.6996	0.6761	0.7187	**0.7293**	0.6762	0.7191
Average		0.5847	0.6503	0.6111	0.7053	0.6576	**0.7115**

employing a more reliable measure of semantic similarity, WSE_S surpasses Sent-SAR by 1.32% in AUROC overall, especially on the MedMCQA dataset, where it exceeds Sent-*SAR* by 2.53%. These enhancements highlight the superior adaptability of *WSE* for open-ended medical QA.

In the MedQuAD task, each few-shot prompt comprises multiple question-answer pairs with a similar structure, without providing any contextual information to the language models. Additionally, ground truth and generated responses exhibit notably greater length than other tasks. To address these challenges, we calculate normalized semantic relevance scores at the word level and assign them to individual tokens. This strategy enhances the connectivity between each word and the entire sequence, effectively mitigating biases induced by sequence length. As a result, WSE_W achieves the highest AUROC of 0.626, significantly outperforming six baseline methods.

Given that RS depends on the length of the longest common subsequence, and semantically equivalent textual sequences can be syntactically or lexically distinct, tasks involving long reference answers and responses may result in no generations meeting the correctness criterion. Table 2 presents the comparative results, excluding tasks with an accuracy of 0 from our analysis. Despite the inherent evaluation limitations of RS, WSE_W and WSE_C demonstrate remarkable superiority. By assessing semantic relevance at the word level rather than evaluating individual tokens independently, WSE_W achieves the second-highest average AUROC of 0.6498, outperforming Token-*SAR* by 2.24%, while WSE_C attains the highest average AUROC of 0.6555. Notably, comparable baselines exhibit unstable uncertainty

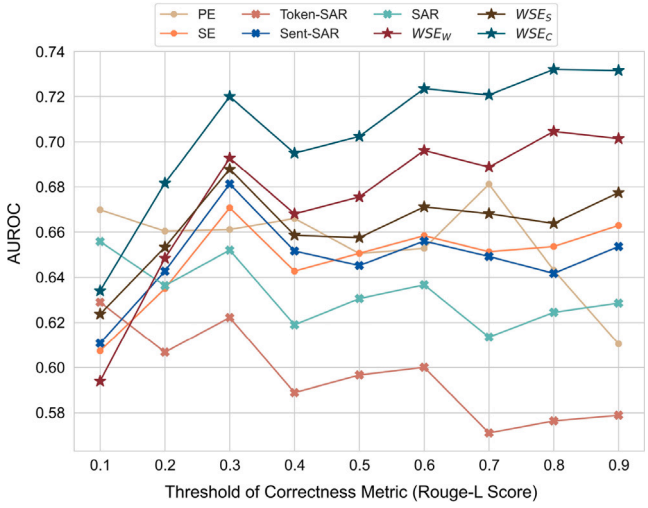


Fig. 4. The performance of WSE_W, WSE_S, WSE_C, and five baseline methods at different thresholds of RS. Results are obtained on the COVID-QA dataset utilizing the LLaMA-2-7B-Chat model.

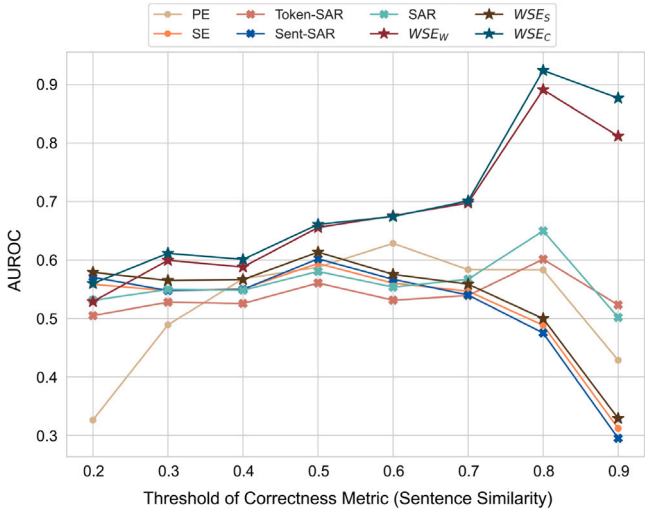


Fig. 5. The performance of WSE_W, WSE_S, WSE_C, and five baseline methods at different thresholds of SS. Results are obtained on the COVID-QA dataset utilizing the LLaMA-2-7B-Chat model.

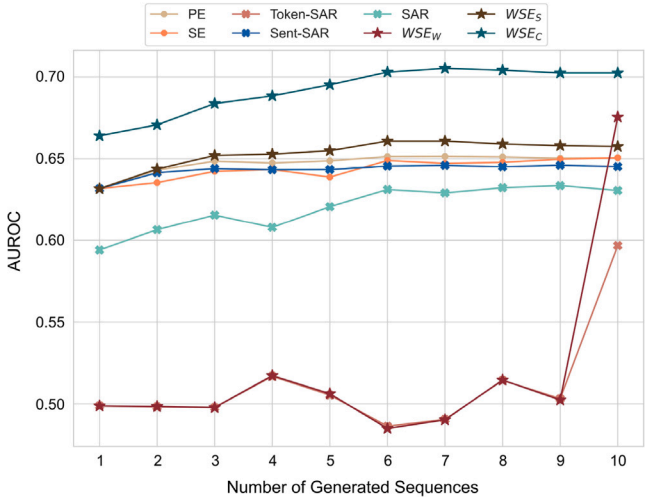


Fig. 6. The performance of WSE_W, WSE_S, WSE_C, and baselines at different numbers of generated sequences employing RS as the metric of correctness evaluation. Results are obtained on the COVID-QA dataset utilizing the LLaMA-2-7B-Chat model.

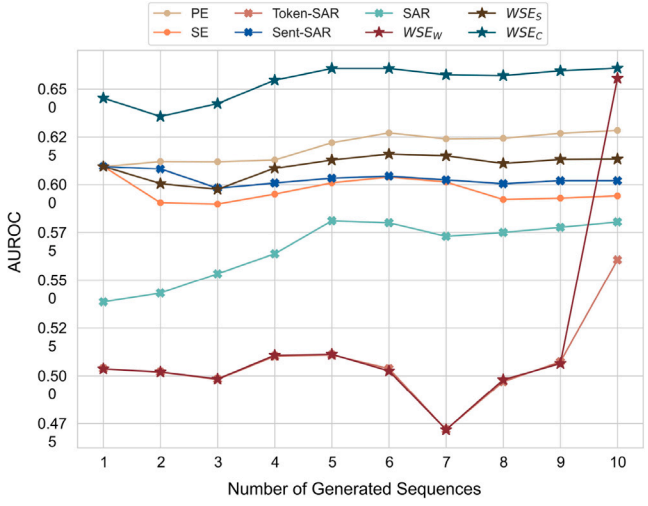


Fig. 7. The performance of WSE_W, WSE_S, WSE_C, and baselines at different numbers of generated sequences employing SS as the metric of correctness evaluation. Results are obtained on the COVID-QA dataset utilizing the LLaMA-2-7B-Chat model.

estimation under rigorous correctness evaluation conditions (e.g., Sent-*SAR* obtains an AUROC of 0.3647 on the PubMedQA dataset in the Vicuna-7B-v1.5 setting), while *WSE* consistently performs reliably, indicating significant potential for practical medical QA applications in the domain of healthcare.

We also evaluate *WSE* on the COVID-QA dataset utilizing a more stringent deep AUROC metric and correctness evaluation criteria. As shown in Table 3, all the three variants of *WSE* consistently outperform the corresponding three variants of *SAR*. For instance, WSE_S outperforms Sent-*SAR* by 3.2% in deep AUROC at the RS setting, WSE_W surpasses Token-*SAR* by 12.06% at the SS setting, and WSE_C achieves the highest deep AUROC of 0.7115, exceeding *SAR* by 10.04%, with SS as the correctness metric.

Overall, WSE demonstrates superior accuracy and stability in quantifying the uncertainty of LLMs-generated responses compared to six baseline methods, utilizing both RS and SS as correctness evaluation criteria across five popular open-ended medical QA tasks.

4.2.2. Sensitivity analysis

To investigate the impact of various thresholds for two correctness metrics on WSE_W, WSE_S, WSE_C, and five baseline methods, we utilize LLaMA-2-7B-Chat and generate ten responses (i.e., $K = 10$) to each medical query on the COVID-QA dataset. As is shown in Figs. 4 and 5, each uncertainty measure is influenced to varying degrees by the threshold. Generally, as the evaluation criteria become more stringent, *WSE* consistently outperforms five baseline methods. Notably, when utilizing RS, WSE_C achieves the highest AUROC of 0.7315, while using SS results in an AUROC of 0.877. When the threshold for SS is set to 0.1, all answers are identified as correct, and we exclude this scenario from our analysis.

Given that entropy-based methods integrate responses within the candidate set, we explore how the number of responses (i.e., K) impacts the performance of UQ. As illustrated in Figs. 6 and 7, WSE_W and Token-*SAR* exhibit sensitivity to variations in K. Nevertheless, WSE_W ultimately surpasses the baselines and achieves the second-highest AUROC score under both correctness evaluation criteria. When employing RS as the correctness metric, WSE_S generally outperforms the baseline

Table 4
The enhancement of model accuracy after employing responses with lower uncertainty identified by WSE_W, WSE_S, and WSE_C, utilizing both RS and SS as the criteria for correctness evaluation under multiple thresholds. Experimental results are obtained on the COVID-QA dataset.

Metrics	Threshold	Accuracy		LLaMA	LLaMA-Chat	Mistral	Zephyr	WizardLM	Vicuna	StabeBeluga
RS	0.3	Initial		0.4775	0.5172	0.1777	0.1936	0.1485	0.1406	0.1777
		Calibrated	WSE_W	0.5225	0.5809	0.6233	0.5862	0.5517	0.5544	0.5703
			WSE_S	0.5013	0.557	0.5995	0.5597	0.496	0.5279	0.5438
			WSE_C	0.504	0.557	0.5968	0.557	0.4934	0.5279	0.5438
		Enhanced (max)		↑ 4.5%	↑ 6.37%	↑ 44.56%	↑ 39.26%	↑ 40.32%	↑ 41.38%	↑ 39.26%
	0.5	Initial		0.3475	0.3899	0.0849	0.0769	0.0504	0.0637	0.0557
		Calibrated	WSE_W	0.382	0.4138	0.4881	0.4191	0.3687	0.3979	0.3793
			WSE_S	0.3501	0.4032	0.4562	0.3395	0.313	0.3395	0.3289
			WSE_C	0.3448	0.4085	0.4562	0.3342	0.321	0.3395	0.3236
		Enhanced (max)		↑ 3.45%	↑ 2.39%	↑ 40.32%	↑ 34.22%	↑ 31.83%	↑ 33.42%	↑ 32.36%
SS	0.5	Initial		0.2679	0.2759	0.3024	0.1910	0.2122	0.2334	0.1857
		Calibrated	WSE_W	0.2918	0.2865	0.3422	0.2546	0.2202	0.2653	0.2122
			WSE_S	0.2891	0.2785	0.3263	0.2202	0.2042	0.252	0.1963
			WSE_C	0.2679	0.2706	0.3103	0.2122	0.1989	0.2361	0.1883
		Enhanced (max)		↑ 2.39%	↑ 1.06%	↑ 3.98%	↑ 6.36%	↑ 0.8%	↑ 3.19%	↑ 2.65%
	0.7	Initial		0.1008	0.1061	0.1273	0.0584	0.069	0.0769	0.0584
		Calibrated	WSE_W	0.13	0.1114	0.1485	0.1008	0.0743	0.1008	0.069
			WSE_S	0.1167	0.1008	0.1406	0.0716	0.0743	0.0902	0.0557
			WSE_C	0.1141	0.1034	0.13	0.0637	0.0743	0.0849	0.0769
		Enhanced (max)		↑ 2.92%	↑ 0.53%	↑ 2.12%	↑ 4.24%	↑ 0.53%	↑ 2.39%	↑ 1.85%

methods and achieves the second-highest AUROC of 0.6161 leveraging only 6 generated sequences, which is generation-efficient,. It is noteworthy that WSE_C consistently outperforms comparable methods under both correctness evaluation criteria.

4.3. Accuracy enhancement

Due to the abundant and diverse domain-specific knowledge within the healthcare domain, the availability of LLMs specifically designed for open-ended medical QA tasks is comparatively limited. Furthermore, real-world medical QA scenarios tend to be highly intricate and often lack contextual information associated with the questions, posing significant challenges to LLMs. In this section, we investigate the enhancement of model accuracy solely through resampling and post-processing, by leveraging multiple "off-the-shelf" LLMs pre-trained on NLG datasets without requiring additional task-specific training or architectural modifications.

Given that *WSE* can quantify uncertainty at the sequence level, we assess the set of K candidate responses, and select the response with the lowest uncertainty, identified by *WSE*, as the final answer to the current medical query. Then, we recompute the overall accuracy of the dataset.

We employ COVID-QA as the dataset and investigate accuracy enhancement under two correctness evaluation criteria. As summarized in Table 4, accuracy improvement varies across multiple LLMs when utilizing RS. Given that RS is sensitive to the structure of generated sequences, LLMs of the LLaMA series achieve higher initial accuracy than others under two thresholds, with a maximum increase of 6.37% observed on the LLaMA-2-7B-Chat model. After filtering high-uncertainty sequences identified by WSE_W, we achieve a substantial accuracy enhancement of 44.56% on the Mistral model when the correctness metric threshold is set to 0.3. Despite the stringent nature and limitations associated with RS, the COVID-QA task exhibits a noteworthy improvement in accuracy across seven "off-the-shelf" LLMs.

For SS, we adopt two relatively stringent thresholds: 0.5 and 0.7. Compared to RS, there is no remarkable enhancement in accuracy, with the highest improvement observed at 6.36% on the Zephyr-7B-Alpha model when the threshold is set to 0.5. Overall, the COVID-QA dataset consistently maintains stable and highly effective accuracy improvements, showcasing the significant potential of *WSE* in practical medical QA applications within the domain of healthcare engineering.

5. Conclusion

We address the lack of general uncertainty measures in open-ended medical QA tasks. Given that generative inequality leads to a large number of irrelevant words and responses in the candidate set for UQ, we highlight the keywords within each textual sequence based on semantic variation and enlarge the generative probability of reliable responses through self-consistency. In the UQ process, we develop a stable measure of semantic textual similarity. Furthermore, to overcome the limitations of LLMs in medical QA, we focus on posterior work and utilize sequences with lower uncertainty identified by *WSE* as final answers, significantly enhancing model accuracy. Experiments on five medical QA datasets demonstrate the superior performance of *WSE* in accurate UQ and its substantial potential in healthcare.

Our proposed method employs "off-the-shelf" LLMs without requiring additional fine-tuning or modifications (i.e., unsupervised), facilitating further research in this area and enhancing reproducibility. However, with the rise of closed-source LLMs served via APIs, end-users typically lack access to token likelihoods or embeddings, limiting the applicability of entropy-based measures. A promising future research direction is to explore black-box approaches for estimating the confidence or uncertainty of LLMs in their responses. Additionally, the semantic diversity of the model's output space cannot fully capture the nuances of its uncertainty. A more comprehensive analysis is warranted, considering factors such as the model's design mechanism and data noise. Furthermore, the reliability of semantic similarity scores significantly affects the sensitivity of semantics-based approaches. We will investigate measurements of semantic textual similarity with stronger explainability and trustworthiness, and aim to devise certified methods for a theoretically rigorous uncertainty notion. By providing users with information regarding the uncertainty of language model outputs, we endeavor to advance the development of safer and more trustworthy QA systems, particularly in the domain of healthcare engineering.

CRediT authorship contribution statement

Zhiyuan Wang: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Data curation, Conceptualization. **Jinhao Duan:** Writing – review &

editing, Methodology, Conceptualization. **Chenxi Yuan:** Writing – review & editing. **Qingyu Chen:** Writing – review & editing. **Tianlong Chen:** Writing – review & editing. **Yue Zhang:** Writing – review & editing. **Ren Wang:** Writing – review & editing. **Xiaoshuang Shi:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Kaidi Xu:** Writing – review & editing, Supervision, Methodology, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

Zhiyuan Wang and Xiaoshuang Shi were supported by the National Key Research & Development Program of China under Grant (No. 2022YFA1004100).

Data availability

All datasets used in this paper are publicly available from Hugging Face.

References

Abacha, A.B., Zweigenbaum, P., 2015. MEANS: A medical question-answering system combining NLP techniques and semantic web technologies. Inf. Process. Manage. 51 (5), 570–594.

Ben Abacha, A., Demner-Fushman, D., 2019. A question-entailment approach to question answering. BMC Bioinformatics 20 (1), 1–23.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al., 2020. Language models are few-shot learners. Adv. Neural Inf. Process. Syst. 33, 1877–1901.

Carrington, A.M., Manuel, D.G., Fieguth, P.W., Ramsay, T., Osmani, V., Wernly, B., Bennett, C., Hawken, S., McInnes, M., Magwood, O., et al., 2023. Deep ROC analysis and AUC as balanced average accuracy to improve model selection, understanding and interpretation. IEEE Trans. Pattern Anal. Mach. Intell. 45, 329–341.

Chen, J., Mueller, J., 2023. Quantifying uncertainty in answers from any language model via intrinsic and extrinsic confidence assessment. arXiv preprint arXiv:2308.16175.

Chen, Z., Zhou, K., Zhang, B., Gong, Z., Zhao, W.X., Wen, J.-R., 2023. ChatCoT: Tool-augmented chain-of-thought reasoning on\\chat-based large language models. arXiv preprint arXiv:2305.14323.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H.W., Sutton, C., Gehrmann, S., et al., 2022. Palm: Scaling language modeling with pathways. arXiv preprint arXiv:2204.02311.

Duan, J., Cheng, H., Wang, S., Wang, C., Zavalny, A., Xu, R., Kailkhura, B., Xu, K., 2023. Shifting attention to relevance: Towards the uncertainty estimation of large language models. arXiv preprint arXiv:2307.01379.

Duan, J., Zhang, R., Diffenderfer, J., Kailkhura, B., Sun, L., Stengel-Eskin, E., Bansal, M., Chen, T., Xu, K., 2024. GTBench: Uncovering the strategic reasoning limitations of LLMs via game-theoretic evaluations. arXiv preprint arXiv:2402.12348.

Fomicheva, M., Sun, S., Yankovskaya, L., Blain, F., Guzmán, F., Fishel, M., Aletras, N., Chaudhary, V., Specia, L., 2020. Unsupervised quality estimation for neural machine translation. Trans. Assoc. Comput. Linguist. 8, 539–555.

Gal, Y., Ghahramani, Z., 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: International Conference on Machine Learning. PMLR, pp. 1050–1059.

Glushkova, T., Zerva, C., Rei, R., Martins, A.F., 2021. Uncertainty-aware machine translation evaluation. arXiv preprint arXiv:2109.06352.

He, M., Garner, P.N., 2023. Can ChatGPT detect intent? Evaluating large language models for spoken language understanding. arXiv preprint arXiv:2305.13512.

He, P., Liu, X., Gao, J., Chen, W., 2020. Deberta: Decoding-enhanced bert with disentangled attention. arXiv preprint arXiv:2006.03654.

Hu, Y., Khan, L., 2021. Uncertainty-aware reliable text classification. In: Proceedings of the ACM SIGKDD Conference on Knowledge Discovery & Data Mining. pp. 628–636.

Hüllermeier, E., Waegeman, W., 2021. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. Mach. Learn. 110, 457–506.

Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., Casas, D.d.l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al., 2023. Mistral 7B. arXiv preprint arXiv:2310.06825.

Jin, Q., Dhingra, B., Liu, Z., Cohen, W.W., Lu, X., 2019. Pubmedqa: A dataset for biomedical research question answering. arXiv preprint arXiv:1909.06146.

Jin, D., Pan, E., Oufattole, N., Weng, W.-H., Fang, H., Szolovits, P., 2020. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. arXiv preprint arXiv:2009.13081.

Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DasSarma, N., Tran-Johnson, E., et al., 2022. Language models (mostly) know what they know. arXiv preprint arXiv:2207.05221.

Kendall, A., Gal, Y., 2017. What uncertainties do we need in bayesian deep learning for computer vision? Adv. Neural Inf. Process. Syst. 30.

Kuhn, L., Gal, Y., Farquhar, S., 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. arXiv preprint arXiv:2302.09664.

Lahlou, S., Jain, M., Nekoei, H., Butoi, V.I., Bertin, P., Rector-Brooks, J., Korablyov, M., Bengio, Y., 2021. Deup: Direct epistemic uncertainty prediction. arXiv preprint arXiv:2102.08501.

Lakshminarayanan, B., Pritzel, A., Blundell, C., 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. Adv. Neural Inf. Process. Syst. 30.

Lin, C.-Y., 2004. Rouge: A package for automatic evaluation of summaries. In: Text Summarization Branches Out. pp. 74–81.

Lin, S., Hilton, J., Evans, O., 2022a. Teaching models to express their uncertainty in words. arXiv preprint arXiv:2205.14334.

Lin, Z., Liu, J.Z., Shang, J., 2022b. Towards collaborative neural-symbolic graph semantic parsing via uncertainty. In: Findings of the Association for Computational Linguistics: ACL 2022.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V., 2019. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

Malinin, A., Chervontsev, S., Provilkov, I., Gales, M., 2020. Regression prior networks. arXiv preprint arXiv:2006.11590.

Malinin, A., Gales, M., 2020. Uncertainty estimation in autoregressive structured prediction. arXiv preprint arXiv:2002.07650.

Manakul, P., Liusie, A., Gales, M.J., 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. arXiv preprint arXiv:2303.08896.

Min, S., Lewis, M., Hajishirzi, H., Zettlemoyer, L., 2021. Noisy channel language model prompting for few-shot text classification. arXiv preprint arXiv:2108.04106.

Miok, K., Nguyen-Doan, D., Škrlj, B., Zaharie, D., Robnik-Šikonja, M., 2019. Prediction uncertainty estimation for hate speech classification. In: Proceedings of the International Conference on Statistical Language and Speech. pp. 286–298.

Möller, T., Reina, A., Jayakumar, R., Pietsch, M., 2020. COVID-QA: A question answering dataset for COVID-19. In: ACL 2020 Workshop on Natural Language Processing for COVID-19. NLP-COVID.

Mukherjee, S., Mitra, A., Jawahar, G., Agarwal, S., Palangi, H., Awadallah, A., 2023. Orca: Progressive learning from complex explanation traces of gpt-4. arXiv preprint arXiv:2306.02707.

Ott, M., Auli, M., Grangier, D., Ranzato, M., 2018. Analyzing uncertainty in neural machine translation. In: International Conference on Machine Learning. PMLR, pp. 3956–3965.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al., 2022. Training language models to follow instructions with human feedback. Adv. Neural Inf. Process. Syst. 35, 27730–27744.

Pal, A., Umapathi, L.K., Sankarasubbu, M., 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In: Conference on Health, Inference, and Learning. PMLR, pp. 248–260.

Quevedo, E., Yero, J., Koerner, R., Rivas, P., Cerny, T., 2024. Detecting hallucinations in large language model generation: A token probability approach. Nature.

Reimers, N., Gurevych, I., 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084.

Sanh, V., Debut, L., Chaumond, J., Wolf, T., 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.

Sun, L., Huang, Y., Wang, H., Wu, S., Zhang, Q., Gao, C., Huang, Y., Lyu, W., Zhang, Y., Li, X., et al., 2024. Trustllm: Trustworthiness in large language models. arXiv preprint arXiv:2401.05561.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al., 2023a. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al., 2023b. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288.

Tunstall, L., Beeching, E., Lambert, N., Rajani, N., Rasul, K., Belkada, Y., Huang, S., von Werra, L., Fourrier, C., Habib, N., et al., 2023. Zephyr: Direct distillation of LM alignment. arXiv preprint arXiv:2310.16944.

Vazhentsev, A., Kuzmin, G., Shelmanov, A., Tsvigun, A., Tsymbalov, E., Fedyanin, K., Panov, M., Panchenko, A., Gusev, G., Burtsev, M., et al., 2022. Uncertainty estimation of transformer predictions for misclassification detection. In: Proceedings of the Association for Computational Linguistics. pp. 8237–8252.

Waisberg, E., Ong, J., Masalkhi, M., Kamran, S.A., Zaman, N., Sarker, P., Lee, A.G., Tavakkoli, A., 2023. GPT-4: a new era of artificial intelligence in medicine. Ir. J. Med. Sci. 1–4.

Wang, Y., Beck, D., Baldwin, T., Verspoor, K., 2022a. Uncertainty estimation and reduction of pre-trained models for text regression. Trans. Assoc. Comput. Linguist. 10, 680–696.

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., Zhou, D., 2022b. Self-consistency improves chain of thought reasoning in language models. arXiv preprint arXiv:2203.11171.

Wang, H., Yu, D., 2023. Going beyond sentence embeddings: A token-level matching algorithm for calculating semantic textual similarity. In: Proceedings of the Association for Computational Linguistics. pp. 563–570.

Xu, C., Sun, Q., Zheng, K., Geng, X., Zhao, P., Feng, J., Tao, C., Jiang, D., 2023. Wizardlm: Empowering large language models to follow complex instructions. arXiv preprint arXiv:2304.12244.

Yao, Y., Duan, J., Xu, K., Cai, Y., Sun, E., Zhang, Y., 2023. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. arXiv preprint arXiv:2312.02003.

Yuan, C., Duan, J., Tustison, N.J., Xu, K., Hubbard, R.A., Linn, K.A., 2023. ReMiND: Recovery of missing neuroimaging using diffusion models with application to Alzheimer's disease. medRxiv.

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X.V., et al., 2022. Opt: Open pre-trained transformer language models. arXiv preprint arXiv:2205.01068.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al., 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. arXiv preprint arXiv:2306.05685.