

Analysis of Image-and-Text Uncertainty Propagation in Multimodal Large Language Models with Cardiac MR-Based Applications

Yucheng Tang^{1,2}, Yunguan Fu^{1,2}, Weixi Yi^{1,2}, Yipei Wang^{1,2}, Daniel C. Alexander^{2,3}, Rhodri Davies⁴, and Yipeng Hu^{1,2}

¹ Department of Medical Physics and Biomedical Engineering, University College London, UK

² UCL Hawkes Institute, University College London, UK

³ Department of Computer Science, University College London, UK

⁴ Institute of Cardiovascular Science, University College London, UK
yucheng.tang.24@ucl.ac.uk

Abstract. Multimodal large language models (MLLMs) can process and integrate information from multimodality sources, such as text and images. However, interrelationship among input modalities, uncertainties due to individual uni-modal data and potential clinical applications following such an uncertainty decomposition are yet fully understood in the context of large-scale MLLMs. In this work, we propose a multimodal uncertainty propagation model (MUPM) based on uncertainty propagation, to characterise the relationship among the uncertainties arising from image-only, text-only, and joint image-text variations in MLLM inputs. Using real clinical data consisting of cardiac MR scans and digital health records, we describe that MUPMs can be optimised robustly with a few samples. We then show that the fitted MUPMs are generalisable across different input data distributions and, perhaps surprisingly, across different downstream tasks. Such a transferability may be explained by the shared pretraining, comparatively light MLLM fine-tuning, along with the low-dimensional nature of the MUPMs. More importantly, this learned transferability, quantifying the relationship between these uncertainties, led to direct clinical applications in which uncertainties may be estimated and thus analysed robustly for varying data or even a novel set of cardiac disease prediction tasks. In addition, we show experimentally the efficiency in multimodal data required for estimating the overall uncertainty and its ability to identify redundant factors, both of which are considered practical yet clinically useful applications with the proposed MUPMs. Codes are available at <https://github.com/yucheng722/MUPM>.

Keywords: Uncertainty · Multimodal LLM · Cardiac MRI

1 Introduction

By leveraging multimodality inputs such as images and texts, multimodal large language models (MLLMs) such as BLIP-2 [18], LLaVA-Med [17], and M3D [3]

have achieved state-of-the-art performance in tasks like clinical decision-making [16], medical image processing [5], and diagnostic text generation [20]. Analysis of uncertainties from individual modality data and their interactions in MLLMs remain underexplored. In medicine, where such uncertainty analysis is particularly useful as it directly impacts the credibility of clinical decisions and potentially offers quantitative insights in these decisions, interesting questions arise: How does variability in a single modality affect overall uncertainty? How can we utilize these uncertainties for clinical applications?

Previous work leveraged MLLM uncertainty to enhance performance in specific tasks, such as composite image retrieval [11], out-of-distribution detection [8], and hallucination identification [27]. Other studies investigated calibration before and after fine-tuning [6], relationship between accuracy and uncertainty [14], and uncertainty-based evaluation for MLLMs [9]. However, to our best knowledge, little research has studied relationship between uncertainties across different MLLM modalities and the clinical applications of these relationships.

In this work, we propose the multimodal uncertainty propagation model (MUPM) to characterize the relationship among uncertainties from image-only, text-only, and joint image-text inputs, as well as the correlation between image and text inputs, in MLLMs. We refer hereinafter to these uncertainties as image-only, text-only, and overall uncertainties. We apply the MUPM to various cardiac disease prediction applications, using cardiac magnetic resonance (MR) imaging scans combined with patient digital health records. We adopted the existing medical-domain-specific M3D model [3], which has been proven credible on large biomedical datasets and supports 3D medical image inputs. Subsequently, we investigate a set of MUPM applications including its robustness across different input data distributions and downstream tasks, efficient estimation of overall uncertainty, and identifying redundant factors.

Pioneering analysis of multi-modality uncertainty propagation, our contributions are as follows: First, we derive a mathematical expression to quantify the relationship among image-only uncertainty, text-only uncertainty, the correlation between image and text inputs, and the overall uncertainty in MLLMs. Secondly, we fit MUPM instead of computing derivatives and show experimentally improved stability and generalizability. Furthermore, we identify four novel applications of MUPM for cardiac disease prediction, with the experiments presented using large data sets. We release our code at <https://github.com/yucheng722/MUPM>.

2 Method

2.1 First-order Variances and Covariance Propagation

Uncertainty propagation theory provides a framework for quantifying how uncertainties (or random errors) in input variables affect a model’s uncertainty [15]. Building on this foundation, we first examine how to quantify the overall uncertainty as a function of image-only and text-only uncertainties.

Suppose that the MLLM is a high-dimensional, nonlinear model F with paired input variables: image x and text y . A first-order Taylor expansion approximates the model in Eq. (1).

$$F(x, y) \approx F(x_0, y_0) + \frac{\partial F}{\partial x} x + \frac{\partial F}{\partial y} y \quad (1)$$

where $\partial F/\partial x$ and $\partial F/\partial y$ are the partial derivatives of the model with respect to image and text inputs, respectively. By computing the variance of both sides of Eq. (1), the uncertainty propagation formula is derived as in Eq. (2).

$$\sigma_{F_{I,T}}^2 \approx \left(\frac{\partial F}{\partial x}\right)^2 \sigma_I^2 + \left(\frac{\partial F}{\partial y}\right)^2 \sigma_T^2 + 2\left(\frac{\partial F}{\partial x}\right)\left(\frac{\partial F}{\partial y}\right)\rho\sigma_I\sigma_T \quad (2)$$

where σ_I^2 and σ_T^2 represent the variances of the image and text inputs, respectively. $\rho\sigma_I\sigma_T$ represents the covariance between image and text inputs, where ρ is the Pearson correlation coefficient [7]. In subsequent sections, we refer to $\sigma_I\sigma_T$ as the covariance term. $\sigma_{F_{I,T}}^2$ denotes the variance of the model's output.

By substituting $\sigma_{F_I}^2 \approx a_I^2 \sigma_I^2$ and $\sigma_{F_T}^2 \approx b_T^2 \sigma_T^2$, where $a_I^2 = \left(\frac{\partial F}{\partial x}\right)^2$ and $b_T^2 = \left(\frac{\partial F}{\partial y}\right)^2$, into Eq.(2), the overall uncertainty can be quantified as a function of image-only and text-only uncertainties, as shown in Eq. (3)

$$\sigma_{F_{I,T}}^2 \approx \frac{a^2}{a_I^2} \sigma_{F_I}^2 + \frac{b^2}{b_T^2} \sigma_{F_T}^2 + 2\frac{ab}{a_I b_T} \rho \sigma_{F_I} \sigma_{F_T} \quad (3)$$

where a and b denote the partial derivatives of the model with respect to x and y . Explicit computing these partial derivatives is challenging due to the complex, nonlinear architecture of MLLMs and is numerically unstable, as shown in experiments of instability of derivative computation in Section 3, whereas fitting linear models provides a more reliable and computationally feasible alternative for approximating these derivatives.

2.2 Uncertainty Estimation and MUPM Fitting

To estimate uncertainties in Eq. (3), we use resampling and augmentation strategies [25] to approximate uncertainties without retraining the MLLM. MUPM is not an uncertainty quantification method, but a model-agnostic framework and is compatible with various uncertainty estimation approaches, such as infer-dropout [19], greedy negative log-likelihood (G-NLL) [2], named entity replacement [12], and assembly [13].

Let X - Y be a paired image-text input set, where $X = \{x_i\}_{i=1}^N$ represents N image inputs and $Y = \{y_i\}_{i=1}^N$ represents the corresponding text inputs. Each pair (x_i, y_i) describes the same semantic entity. As shown in Fig. 1, to estimate the image-only uncertainty $\sigma_{F_I}^2$, we generate n augmented versions for each cardiac MR image in the original set X . These augmented images, denoted by $\{X^j\}_{j=1}^n$, are produced by applying affine transformations (e.g., rotation, zoom,

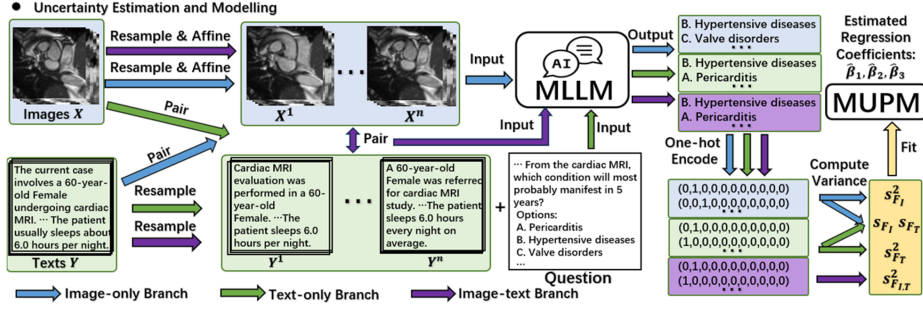


Fig. 1. Uncertainty estimation and MUPM fitting process. We compute the sample variances of one-hot encoded outputs of MLLM when applying image-only, text-only, and image-text data resampling and augmentation. The covariance term is computed from the outputs of the image-only and text-only branches. We then fit the linear MUPM in Eq. (4) using the sample variances and the covariance term. The regression coefficients are estimated during this process.

shear, shift) and adding Gaussian noise to resampled scans from the same patient case. We pair these images with fixed text inputs Y and compute the sample variance $s_{F_I}^2$ from the one-hot encoded MLLM outputs. One-hot encoding of MLLM outputs is application-specific; further details are provided in Section 3.3. Similarly, $s_{F_T}^2$ is estimated by augmenting resampled digital health records $\{Y^j\}_{j=1}^n$ using synonymous replacements while fixing image inputs X , then calculating sample variance $s_{F_T}^2$ of the one-hot encoded MLLM outputs. For overall uncertainty $s_{F_{I,T}}^2$, we simultaneously resample and augment both modalities, pair augmented samples, and compute MLLM outputs' variance $s_{F_{I,T}}^2$.

Based on Eq.(3), we formulate MUPM as a linear regression model in Eq.(4), with independent variables ($s_{F_I}^2, s_{F_T}^2, s_{F_I}s_{F_T}$), and dependent variable ($s_{F_{I,T}}^2$).

$$s_{F_{I,T}}^2 = \beta_1 s_{F_I}^2 + \beta_2 s_{F_T}^2 + \beta_3 s_{F_I}s_{F_T} + \epsilon \quad (4)$$

where β_1, β_2 , and β_3 are regression coefficients, ϵ represents residual errors. The least squares [4] is used to estimate optimal parameters $\hat{\beta}_1, \hat{\beta}_2$, and $\hat{\beta}_3$.

2.3 MUPM Applications

We propose four clinical use cases of MUPM for cardiac disease prediction.

First, MUPM is investigated across various input data distributions for cardiac disease prediction. For instance, if we observe a higher regression coefficient of text-only (or image-only) uncertainty in MUPM, meaning that it contributes more to the overall uncertainty, we would like to improve the text (or image) quality by e.g. adding more detailed or further clinical information (or by MR reacquisition). In these examples, if the quantified relationship between these uncertainties remains relatively stable, there is no need to refit MUPM. It is nonetheless important to note that quantifying the health-economical values of

per unit of changes in texts or images remains highly application-specific. Specific clinical cases are provided and discussed in Section 3.

Second, MUPM is tested with various cardiac disease prediction tasks, including forecasting the occurrence of diseases in 1, 3, or 5 years, using closed-ended visual question answering (VQA) tasks in which MLLM classifies diseases from predefined options. These tasks are designed to share the same definition of uncertainty, as the same token-based format encoded in one-hot is used, even though they produce different semantic predictions.

Furthermore, MUPM is expected to enable efficient estimation of overall uncertainty from downstream data with fewer resample sizes (n in Section 2.2 set to a lower value) than estimation using test-time augmentation from scratch.

Finally, MUPM identifies redundant factors by analyzing the covariance term. If a modality has a very low regression coefficient, it has a low impact on predictive performance. Completely redundant factors are more likely to have very low regression coefficients and simultaneously exhibit elements with large absolute values in their covariance terms with other modalities.

3 Experiments and Results

Implementation Details We implemented MLLM uncertainty estimation using PyTorch [22] and Transformers [26], and fitted MUPM using Scikit-Learn [23]. We adopted the M3D [3] model as the MLLM model as it supports 3D image inputs. The implementation was carried out on a 16GB NVIDIA GeForce RTX 4070 GPU, and the uncertainty estimation and MUPM fitting process took approximately 8 hours in total.

Datasets We utilized the UK BioBank Dataset Heart MRI (Category 1015, Version Sep 2024) as image database, which includes data from 78,684 participants, resulting in 71,702 image pairs. These images include four cardiac MRI modalities: lax2c, lax3c, lax4c (2D long-axis views), and sax (3D short-axis view), each with 50 temporal frames. For text database, we collected information from the UK Biobank Text Metadata (Data-Field 41270, Data-Field 41280, Version Apr 2024), covering 446,814 participants and totaling 7,015,683 entries. The records include 17 patient information items such as age, sex, weight, height, BMI, symptoms, pulse rate, diastolic blood pressure, systolic blood pressure, cardiac related history, cardiac history, family history, metabolic equivalent of task, sleep duration, smoking, drinking, heart function and diagnoses of 11 cardiac disease categories: hypertensive disease, ischaemic heart disease, arrhythmia, conduction disorders, complications, valve disorders, heart failure, pulmonary heart disease, rheumatic heart disease, pericarditis, and cardiomyopathy.

We constructed image-text pairs by retrieving cardiac disease diagnostic records from the text database, aligned with each patient’s cardiac MRI timestamp at 1, 3, and 5 year intervals. This yielded 611, 1223, and 1223 valid cardiac image-text pairs (for 1, 3, and 5 year predictions, respectively) for a closed-ended VQA task predicting cardiac diseases. The original images were resized to $256 \times 256 \times 32$ to meet the M3D model’s input requirements and normalized

Table 1. Results of MUPM regression when input data distributions change due to improved data quality in cardiac disease prediction tasks. $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\beta}_3$ are the averaged fitted regression coefficients from five experiments, corresponding to the L2-norms of the mean cardiac MR image-only uncertainty ($\|s_{F_I}^2\|_2$), health record text-only uncertainty ($\|s_{F_T}^2\|_2$), and the covariance term ($\|s_{F_I} s_{F_T}\|_2$). R^2 represents goodness of fit between the overall uncertainties computed by the fitted MUPM and overall uncertainty benchmark, with reported ECE values and the ANOVA p-values for the regression coefficients across the three tasks. The abbreviations denote experimental settings: w.o. (without data quality improvement); Image $\uparrow$ (cardiac MR image quality improvement only); Text $\uparrow$ (health record text quality improvement only); Both $\uparrow$ (simultaneous cardiac MR image and health record text quality improvement).

	$\hat{\beta}_1 (\ s_{F_I}^2\ _2)$	$\hat{\beta}_2 (\ s_{F_T}^2\ _2)$	$\hat{\beta}_3 (\ s_{F_I} s_{F_T}\ _2)$	R^2	ECE
w.o.	0.24 (0.18)	0.90 (0.40)	-0.20 (0.17)	0.76	0.09
Image $\uparrow$	0.24 (0.15)	0.88 (0.42)	-0.21 (0.16)	0.78	0.09
Text $\uparrow$	0.21 (0.18)	0.93 (0.35)	-0.22 (0.13)	0.80	0.12
Both $\uparrow$	0.22 (0.11)	0.91 (0.28)	-0.22 (0.09)	0.84	0.06
ANOVA	0.41	0.33	0.89		

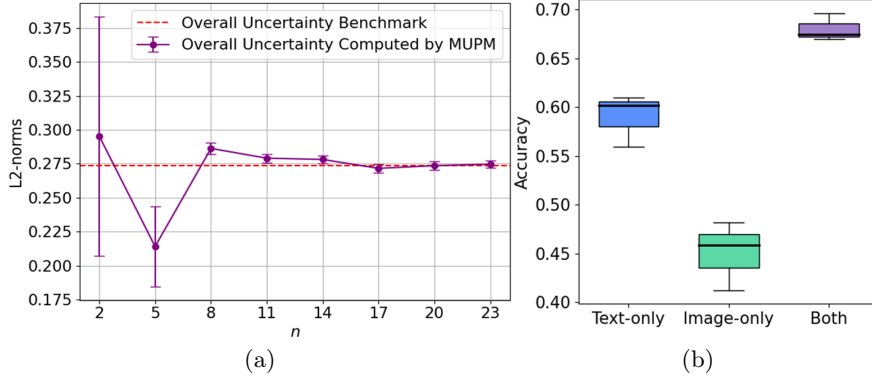
to $[0, 1]$. In all simulations, we set the resample size $n = 20$. We calculated the benchmark of overall uncertainty using all cardiac image-text pairs, which includes cardiac MR images and health record text pairs used for cardiac disease prediction over 1, 3, and 5 years. Resampling and augmentation were applied 100 times to quantify overall uncertainties for all cardiac image-text pairs, and their average was used as the benchmark for comparison.

To avoid **instability of derivative computation**, also described in Section 2.1 and Section 2.2, we use the linear regression model in Eq.(4), where the fitted regression coefficients represent approximations of the complex partial derivative-related terms in Eq.(3). To validate this, 60 cardiac image-text pairs were sampled from our dataset, to estimate the partial derivatives using the chain rule, for all three types of uncertainties. For each sampled cardiac image-text pair, we computed the average derivatives of image pixels (for cardiac MR images) and text tokens (for health record texts). The results are as follows: $\bar{a}_I = 4.58 \pm 5.69$, $\bar{b}_T = 5.65 \pm 5.79$, $\bar{a} = 3.16 \pm 5.33$, and $\bar{b} = 5.07 \pm 5.91$, where $\bar{a}_I$, $\bar{b}_T$, $\bar{a}$, and $\bar{b}$ represent the computed average derivatives with standard deviations (StDs). The results indicate that directly computing these partial derivatives yields large standard deviations among samples, highlighting their instability. Additionally, gradient vanishing was observed in this experiment, making uncertainty propagation modelling highly impractical.

Robustness of MUPM across various input data distributions was quantified in the cardiac disease prediction tasks. We used cardiac MR images in short-axis (sax) view as image inputs. To generate augmented cardiac MR sax image inputs for each patient case, we first resampled from a total of 50 time frames along the temporal axis. Using the proposed sampling procedure, we expected to cover all time frames across a cardiac cycle, thus representing a time-frame independent analysis. Each resampled image was then augmented

Table 2. Results of MUPM regression across different downstream tasks for predicting cardiac disease over 1, 3, 5 years, respectively, structured the same as Table 1.

	$\hat{\beta}_1 (\ s_{F_I}^2\ _2)$	$\hat{\beta}_2 (\ s_{F_T}^2\ _2)$	$\hat{\beta}_3 (\ s_{F_I} s_{F_T}\ _2)$	R^2	ECE
1 year	0.20 (0.11)	0.91 (0.30)	-0.19 (0.09)	0.79	0.07
3 years	0.21 (0.12)	0.89 (0.32)	-0.20 (0.10)	0.82	0.07
5 years	0.23 (0.13)	0.91 (0.33)	-0.22 (0.10)	0.78	0.10
ANOVA	0.70	0.95	0.98		

**Fig. 2.** (a) Experimental results on the efficient uncertainty estimation of downstream data: the purple line represents the average L2-norms of the overall uncertainties (with StDs) computed by MUPM, when $n = 2, 5, 8, 11, 14, 17, 20, 23$. The red line represents the benchmark overall uncertainty. (b) Box plot of the average accuracy (with StDs) when removing cardiac MR image inputs (text-only), removing health record text inputs (image-only), and keeping both inputs (both).

using random affine transformations with maximum rotations of (15,15,15), zooms of (0.2,0.2,0.2), shear transformations of (5,5,5,5,5,5), and shifts of (0.15,0.15,0.15), and an added Gaussian noise (StD=10). To generate augmented text inputs, for each input, we resampled 5 to 6 items from the patient information and combined them with the predefined options corresponding to 11 cardiac disease categories. Additionally, we used synonymous replacements by GPT [1] with the temperature parameter set to 0.8. The one-hot encoded prediction was one of the 11 cardiac disease categories.

To simulate real-world scenarios of cardiac MR images with different imaging angles and signal-to-noise ratios (SNR), we changed the data distributions and improved image quality (relative to those in the experiments in Table 2) by resampling the cardiac MR images in long-axis (lax2c, lax3c, lax4c) views with higher SNRs and by removing affine transformations and noise. To simulate real-world scenarios of health reports with varying writing styles, formats, and templates, for health record text inputs, we changed the data distribution by adding more patient information items (increasing from 5–6 items to 10–12

items) and employing a data augmentation strategy involving synonymous replacements generated by GPT [1] with a lower temperature parameter of 0.4.

For each data distribution, the cardiac image-text pairs were split into five non-overlapping folds, and MUPM was fitted using each fold for 5 times. Although quantifying uncertainty is not the main focus of this work, its quality is relevant in real-world applications. We computed the expected calibration errors (ECEs) [21] to assess the quality of the estimated uncertainty. Samples were grouped based on the 10 percentiles of their uncertainty L2-norms. Within each group, we computed the mean absolute error (MAE) between prediction accuracy and confidence, with confidence quantified by mapping the uncertainty L2-norms using a logarithmic and sigmoid transformation.

The experimental results based on analysis of variance (ANOVA) in Table 1 show that there is no evidence to reject the null hypothesis that there is no significant difference between these regression coefficients of MUPM, across different input data distributions. This demonstrates that MUPM is insensitive to variable data distribution. Notably, improving the quality of health record text alone resulted in a more significant reduction in overall uncertainty as shown in Table 1. This is because the regression coefficient for text-only uncertainty is relatively large, perhaps suggesting it contributes more to overall uncertainty.

Robustness of MUPM in different cardiac disease predication tasks was also tested, for predictions over 1 year, 3 years, and 5 years, respectively, with the same image-text pairs. Table 2 compares MUPM’s regression coefficients across tasks, with statistical significance assessed by ANOVA[24] ($p=0.05$), with all pair-wise p -values greater than 0.05. This may seem counter-intuitive, albeit also demonstrated by the absolute values, the fitted variance and covariance are relatively invariant to varying tasks. Additionally, the results in Table 2 show that the average L2 norm of uncertainty increases with prediction time: 1 year $<$ 3 years $<$ 5 years, indicating a decrease in the reliability of cardiac disease predictions as time progresses, which is consistent with intuition in this case.

Efficient use of downstream data in overall uncertainty estimation was assessed by testing whether stable overall uncertainty estimates can be obtained with limited resample sizes (n in Section 2.2), on unseen downstream data. By varying n , we analyzed the trend of overall uncertainty estimates. For each n , we sampled 60 unseen image-text pairs to estimate the image-only and text-only uncertainties by respectively resampling and augmenting input pairs n times. The average L2-norms of 60 overall uncertainty estimates (with StDs) are shown in Fig. 2.(a) when $n = 2, 5, 8, 11, 14, 17, 20, 23$. Results indicate that as n increases, the overall uncertainty calculated by MUPM becomes closer to the benchmark. Beyond $n = 20$, the MUPM calculated overall uncertainty stabilizes with low StDs, allowing reliable computation using MUPM.

Identifying Redundant Factors As shown in Table 2 and Table 1, although the cardiac MR image-only regression coefficients ($\hat{\beta}_1$) are relatively small, the low covariance term between the cardiac MR image-only and health record text-only uncertainties (i.e., $\hat{\beta}_3 \times s_{F_I} s_{F_T}$) suggests that cardiac MR image information is not redundant. To further validate this, we removed the cardiac

MR image and health record text inputs separately, which led to a significant decrease in prediction accuracy (see Fig. 2.(b)), confirming that both modalities are essential in cardiac disease prediction tasks.

4 Conclusion

We present MUPM, a linear model that describes uncertainty propagation in MLLM. We fit MUPM by estimating the cardiac MR image-only, health record text-only, and overall uncertainties, along with the covariance term, using resampling and augmentation strategies, to avoid the instability observed in derivative computations. Four clinical applications of MUPM are proposed for cardiac disease prediction. Experimental results show MUPM’s robustness across different input data distributions and downstream tasks, in addition to efficient estimation and the ability to identify redundancy. Future work will extend MUPM to cardiac MRI-specific MLLMs such as CineMA[10], and apply it to per-disease analyses to identify clinically relevant modality contributions.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Aichberger, L., Schweighofer, K., Hochreiter, S.: Rethinking uncertainty estimation in natural language generation. arXiv preprint arXiv:2412.15176 (2024)
3. Bai, F., Du, Y., Huang, T., Meng, M.Q.H., Zhao, B.: M3d: Advancing 3d medical image analysis with multi-modal large language models. arXiv preprint arXiv:2404.00578 (2024)
4. Björck, Å.: Least squares methods. Handbook of numerical analysis **1**, 465–652 (1990)
5. Chen, L., Han, X., Lin, S., Mai, H., Ran, H.: Trimedlm: Advancing three-dimensional medical image analysis with multi-modal llm. In: 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). pp. 4505–4512. IEEE (2024)
6. Chen, Z., Hu, W., He, G., Deng, Z., Zhang, Z., Hong, R.: Unveiling uncertainty: A deep dive into calibration and performance of multimodal large language models. arXiv preprint arXiv:2412.14660 (2024)
7. Cohen, I., Huang, Y., Chen, J., Benesty, J., Benesty, J., Chen, J., Huang, Y., Cohen, I.: Pearson correlation coefficient. Noise reduction in speech processing pp. 1–4 (2009)
8. Dai, Y., Lang, H., Zeng, K., Huang, F., Li, Y.: Exploring large language models for multi-modal out-of-distribution detection. arXiv preprint arXiv:2310.08027 (2023)
9. Dang, Y., Gao, M., Yan, Y., Zou, X., Gu, Y., Liu, A., Hu, X.: Exploring response uncertainty in mllms: An empirical evaluation under misleading scenarios. arXiv preprint arXiv:2411.02708 (2024)

10. Fu, Y., Yi, W., Manisty, C., Bhuva, A.N., Treibel, T.A., Moon, J.C., Clarkson, M.J., Davies, R.H., Hu, Y.: Cinema: A foundation model for cine cardiac mri. arXiv preprint arXiv:2506.00679 (2025)
11. Ge, H., Jiang, Y., Sun, J., Yuan, K., Liu, Y.: Llm-enhanced composed image retrieval: An intent uncertainty-aware linguistic-visual dual channel matching model. *ACM Transactions on Information Systems* **43**(2), 1–30 (2025)
12. Goodarzi, S., Kagita, N., Minn, D., Wang, S., Dessi, R., Toshniwal, S., Williams, A., Lanchantin, J., Sinha, K.: Robustness of named-entity replacements for in-context learning. In: Findings of the Association for Computational Linguistics: EMNLP 2023. pp. 10914–10931 (2023)
13. He, J., Zhang, X., Lei, S., Chen, Z., Chen, F., Alhamadani, A., Xiao, B., Lu, C.: Towards more accurate uncertainty estimation in text classification. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 8362–8372 (2020)
14. Kostumov, V., Nutfullin, B., Pilipenko, O., Ilyushin, E.: Uncertainty-aware evaluation for vision-language models. arXiv preprint arXiv:2402.14418 (2024)
15. Lee, S.H., Chen, W.: A comparative study of uncertainty propagation methods for black-box-type problems. *Structural and multidisciplinary optimization* **37**, 239–253 (2009)
16. Li, B., Meng, T., Shi, X., Zhai, J., Ruan, T.: Meddm: Llm-executable clinical guidance tree for clinical decision-making. arXiv preprint arXiv:2312.02441 (2023)
17. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems* **36** (2024)
18. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International conference on machine learning. pp. 19730–19742. PMLR (2023)
19. Mi, L., Wang, H., Tian, Y., He, H., Shavit, N.N.: Training-free uncertainty estimation for dense regression: Sensitivity as a surrogate. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 10042–10050 (2022)
20. Mo, Y., Qin, H., Dong, Y., Zhu, Z., Li, Z.: Large language model (llm) ai text generation detection based on transformer deep learning algorithm. arXiv preprint arXiv:2405.06652 (2024)
21. Nixon, J., Dusenberry, M.W., Zhang, L., Jerfel, G., Tran, D.: Measuring calibration in deep learning. In: CVPR workshops. vol. 2 (2019)
22. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* **32** (2019)
23. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al.: Scikit-learn: Machine learning in python. *the Journal of machine Learning research* **12**, 2825–2830 (2011)
24. St, L., Wold, S., et al.: Analysis of variance (anova). *Chemometrics and intelligent laboratory systems* **6**(4), 259–272 (1989)
25. Wang, G., Li, W., Aertsen, M., Deprest, J., Ourselin, S., Vercauteren, T.: Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing* **338**, 34–45 (2019)
26. Wolf, T.: Transformers: State-of-the-art natural language processing. arXiv preprint arXiv:1910.03771 (2020)

27. Zhang, R., Zhang, H., Zheng, Z.: Vl-uncertainty: Detecting hallucination in large vision-language model via uncertainty estimation. arXiv preprint arXiv:2411.11919 (2024)