K21004676

7MRI0110 MSc/MRes Individual Research Project

CORE-CMR:

A GDPR Compliant, AI Cardiac MRI Summary Generation Training and Fine-Tuning Pipeline with Hallucination and Uncertainty Awareness

Author

Damandeep Singh Kharoud

Supervisors

Professor Amedeo Chiribiri

Nathan Wong MSc, BSc

Project Report

submitted in partial fulfilment of the

Master of Science or Master of Research degree in

Healthcare Technologies

August 2025

§

Acknowledgments:

I would like to thank my supervisors, Professor Amedeo Chiribiri and Nathan Wong, for their academic guidance and technical support throughout this project. I am also deeply grateful to my parents for providing a supportive environment that enabled me to focus on this work. Finally, I thank my course friends (especially Konstantinos, Ahthini and Natthaya), my friends from Medicine (Terren, Jay, Raghav, Noah and Preaveen) and school friends (especially Faye, America, Daniel and Orchi) for their encouragement and support, and for helping me to maintain balance and perspective during the course of this project.

§

Contents Page

§

Design Statement: Core CMR - A GDPR Compliant, AI Pipeline for Cardiac MRI Summary Generation with Hallucination and Uncertainty Awareness

Context and Motivation:

Cardiac MRI (cMRI) is the gold standard for assessing cardiac function, yet report generation is slow and resource-intensive, often requiring more than an hour per patient. Previous work on a locally deployable MVP demonstrated feasibility of XML parsing, guideline aware reporting, and draft summary generation utilising Large Language Models (LLM), but also revealed limitations including LLM hallucinations and lack of interpretability. There is a pressing need for a safe, efficient, and GDPR-compliant pipeline to improve cMRI reporting throughput.

Aim & Objectives:

The aim of this project is to develop and evaluate a locally deployable large language model (LLM) pipeline that reduces hallucinations in cMRI reporting while enhancing interpretability and compliance with clinical governance. Objectives are:

1. Fine-tune a small, open-source LLM (Llama 3.2, 3B parameters) using **Low-Rank Adaptation (LoRA)** for domain specificity.
2. Implement **Retrieval-Augmented Generation (RAG)** to improve factual grounding through historical report retrieval.
3. Develop a **hallucination detection classifier** trained on semi-synthetic labelled data.
4. Integrate **token-wise Monte Carlo dropout** to estimate confidence and uncertainty.
5. Evaluate performance using both classical NLP metrics and clinically relevant hallucination rates.

Methods & Approach:

A dataset of 5,550 anonymised cMRI reports was split into training, embedding, and test sets. Structured parsing and embedding were implemented via FAISS for RAG retrieval. LoRA adapters were trained on 3,000 reports. Hallucination detection was developed using frozen LLM embeddings with a neural classifier head. Uncertainty visualisation was performed via multiple stochastic forward passes with dropout activated (Monte-Carlo dropout). Full pipeline and methodology is visualised in Figure 1.

§

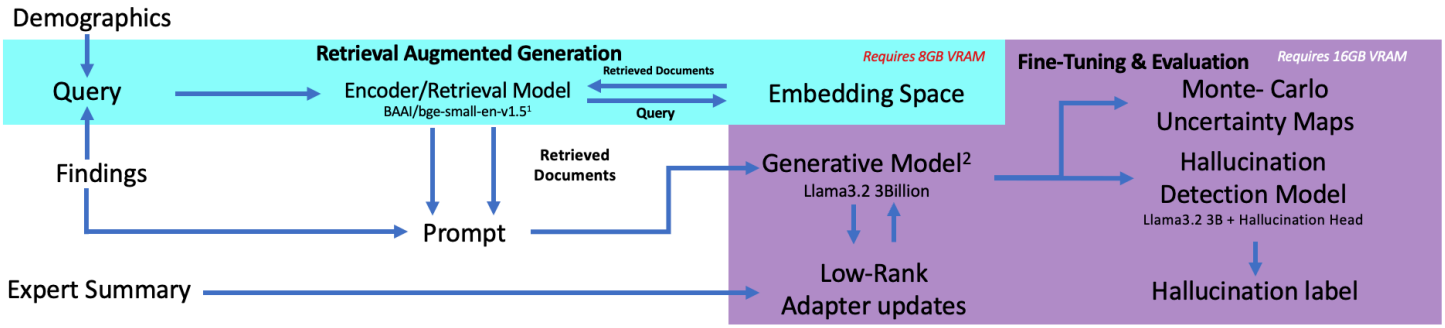


Figure 1. Full proposed pipeline

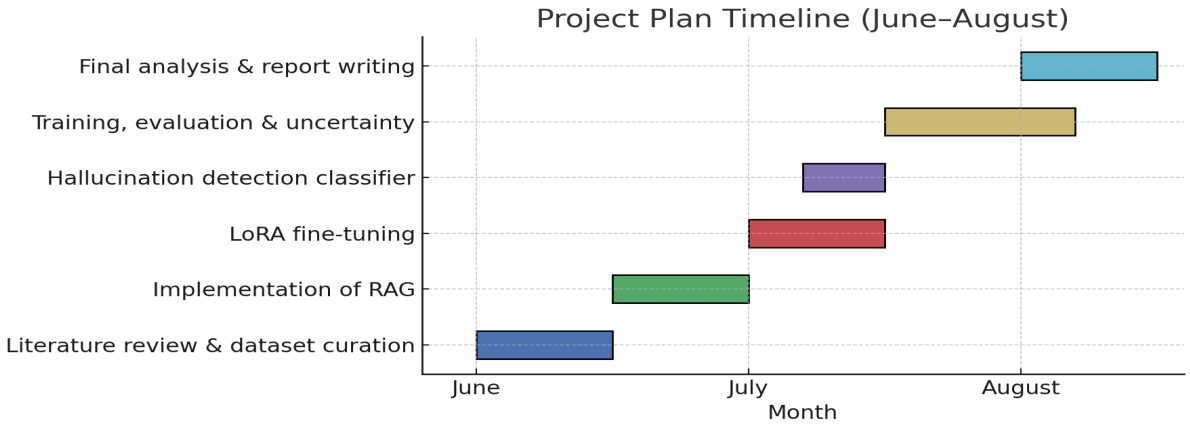


Figure 2. Proposed Timeline for Project

Expected Outcomes & Impact:

This project will deliver a GDPR-compliant fine-tuning pathway that adapts a lightweight, low-memory LLM to generate clinically safe cMRI summaries. We expect to demonstrate a substantial reduction in intrinsic hallucinations through retrieval-augmented generation (RAG) and LoRA fine-tuning, alongside a robust hallucination detection model capable of identifying errors in generated text. Token-wise Monte Carlo dropout will provide a framework for uncertainty visualisation, enabling clinicians to flag high-risk tokens during review and preparing the system for integration into the broader CORE-CMR application.

By shifting clinicians from composing reports to reviewing AI-assisted drafts, the pipeline has the potential to reduce reporting times from over an hour to minutes, improve patient throughput, and establish a transferable framework for safe, interpretable AI deployment across cardiology and other clinical domains.

§

AI Declaration Statement

King's requires students to acknowledge any use of generative AI tools in coursework by including a declaration statement along with your references. Please note that so long as acknowledged use falls within the scope of appropriate use as defined in the assessment brief/guidance then this will not have any direct impact on the grades awarded.

Please include the following completed statements in your project reports:

I declare that parts of this submission has contributions from AI software and that it aligns with acceptable use as specified as part of the assignment brief/ guidance and is consistent with good academic practice. The content can still be considered as my own words. I understand that as long as my use falls within the scope of appropriate use as defined in the assessment brief/guidance then this declaration will not have any direct impact on the grades awarded.

I acknowledge use of software to [include as appropriate]:

(i) Generate ideas or structure suggestions, for assistance with understanding core concepts, or other substantial foundational and preparatory activity.

NA

(ii) Write, rewrite, rephrase and/or paraphrase part of this essay.

ChatGPT – Proof reading for grammatical errors and lexical ambiguity

Grammarly AI – Proof reading for grammatical errors and lexical ambiguity

(iii) Generate some other aspect of the submitted assessment.

ChatGPT – syntax checking code written by Author for pythonic errors

§

ABSTRACT

Cardiac Magnetic Resonance Imaging (cMRI) is the gold-standard modality for assessing cardiac structure and function. However, the reporting process post scan is slow and labour-intensive, often requiring over an hour per patient. Previous work demonstrated the feasibility of CORE-CMR, a modular, locally deployable minimum viable product (MVP) that parsed XML volumetric data, compared measurements against clinical guidelines, and generated draft summaries using large language models (LLMs). While this MVP reduced reporting time to under five minutes, it also exposed critical limitations: guideline-dependent variability in outputs and frequent hallucinations of prior pathology, highlighting the need for more robust, clinically safe approaches.

This project extends CORE-CMR into a complete training and fine-tuning pipeline. A lightweight, locally deployable Llama-3.2 model (3B parameters) was combined with Retrieval-Augmented Generation (RAG) to improve factual grounding, Low-Rank Adaptation (LoRA) for domain-specific fine-tuning, and a custom hallucination detection model trained on a semi-synthetic dataset of 500 labelled samples. Additionally, token-wise Monte Carlo dropout was implemented to estimate confidence and epistemic uncertainty, allowing for interpretability through heatmaps and the flagging of high-risk tokens.

Results demonstrated that classical natural language metrics (BLEU, ROUGE, METEOR, Cosine Similarity) were poorly aligned with clinically relevant errors. By contrast, hallucination detection provided a reliable quantitative measure of model safety. RAG achieved the most significant performance gain, reducing hallucination rates by over 70% compared with baseline. LoRA yielded modest further reductions but recalibrated model confidence, making hallucinations more cautious and interpretable. Uncertainty visualisation provided an additional safety layer by highlighting potentially unreliable tokens for clinician review.

All interventions were achieved on a small, resource-efficient model that can be deployed locally, ensuring GDPR compliance. By shifting the clinician’s role from composing to reviewing reports, this pipeline offers a pathway to faster, safer, and more interpretable AI-assisted reporting in cardiology, with potential to scale across other specialities.

1.INTRODUCTION

1.1 Cardiac Magnetic Resonance Imaging

Cardiac Magnetic Resonance Imaging (cMRI) has become one of the principal diagnostic tools for assessing structural abnormalities of the heart, both congenital and acquired [1]. At many NHS sites, including Guy's and St Thomas' (GSTT), 3T MRI scanners are used to generate volumetric reconstructions of the heart with or without contrast. The clinical workflow typically involves image acquisition (30–60 minutes), reconstruction of a three-dimensional cardiac volume, segmentation and

measurement of structures such as ventricular wall thickness and ejection fraction, documentation of findings (usually under 30 minutes), and finally the composition of a comprehensive clinical summary, which often requires more than one hour. This final stage constitutes the most significant bottleneck; even experienced consultant cardiologists can spend over an hour per patient, in contrast to the comparatively rapid segmentation and findings documentation. Such inefficiencies contribute to delays in reporting, which may postpone diagnosis and treatment initiation, increase morbidity, mortality, and treatment-related complications, and ultimately result in patient dissatisfaction and reduced quality of life.

1.2 CORE-CMR & Previous Work

To address the disproportionate time required to write clinical summaries, we developed an end-to-end application, CORE-CMR, designed to accelerate the reporting workflow. Built upon a Docker backbone, clinicians could upload XML volumetric files, which were automatically parsed into a fixed-structure set of findings. These findings were then automatically compared to the preferred guideline (either EACVI[2] or Hudsmith et al. (2005) [3]), before being passed to a large language model (LLM) to generate a draft clinical summary. This earlier implementation functioned as a modular, locally deployable minimum viable product (MVP). It demonstrated that structured parsing and guideline-aware reporting could reduce reporting time to minutes and shift the clinician's role from composing full reports to reviewing and correcting draft outputs. However, it also revealed critical limitations: outputs varied substantially depending on the guideline selected, and in some cases the LLM hallucinated prior pathology despite normal findings. These observations highlighted two requirements for safe clinical deployment: hallucination awareness and mitigation, and transparent integration of guideline and demographic context.

This experience reflects wider developments in the field. A growing body of literature demonstrates that LLMs are increasingly being deployed for administrative and reporting tasks to reduce the non-critical documentation burden on clinicians, thereby enabling more patient-facing care[4], [5], [6]. Similar findings have been reported in systematic reviews of AI in radiology and pathology[5], which consistently highlight efficiency gains in documentation, triaging, and report standardisation, but also warn of risks such as over-reliance, data privacy concerns, and the persistence of hallucinations in clinical contexts[4], [7]. Furthermore, studies of domain-specific systems such as MedPaLM show that while LLMs can achieve high factual accuracy on benchmark question-answering tasks, their performance drops substantially when applied to real-world patient records, reinforcing the need for domain adaptation and hallucination safeguards. More recently, safety frameworks for medical text summarisation have emphasised the importance of quantifying hallucination rates and model calibration before clinical deployment[4].

§

Together, these findings and prior developments in CORE-CMR provide the foundation for this project. In contrast to the earlier MVP, which primarily demonstrated feasibility, the present work extends CORE-CMR into a comprehensive training and fine-tuning pipeline that introduces retrieval-augmented generation (RAG)[8] for factual grounding, low-rank adaptation (LoRA)[9] for domain specialisation, and novel modules for hallucination detection and uncertainty mapping. These advances directly address the limitations identified in the MVP and align with the broader need for safe, interpretable, and locally deployable clinical AI.

1.3 Aims and Objectives

To address these limitations, we propose a pipeline centred on fine-tuning a small, locally deployable LLM, Llama 3.2 (3 billion parameters)[10], to generate clinically safe summaries of cMRI findings. The pipeline combines **Retrieval-Augmented Generation (RAG)[11]** with **Low-Rank Adaptation (LoRA)[9]** to improve factual grounding and domain specificity. In addition, we introduce a novel hallucination-detection metric based on a custom classifier architecture, alongside token-wise **Monte Carlo confidence and uncertainty estimation,** enabling the system to highlight tokens with a high likelihood of error. Together, these strategies aim to mitigate hallucinations, enhance trustworthiness, and ensure that the pipeline can be safely integrated into the clinical workflow.

2.METHODOLOGY

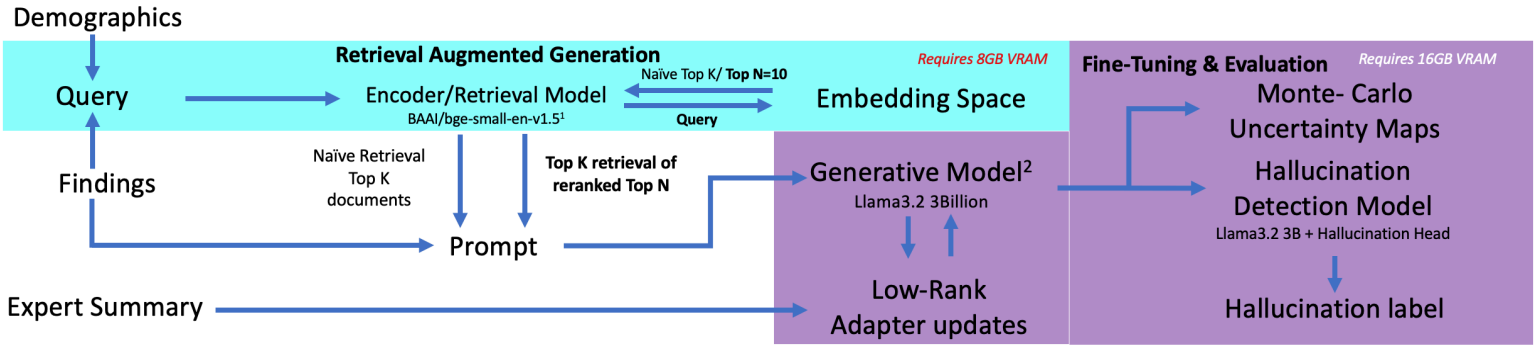


Figure 1 – Flow chart summarising complete pipeline methodology

2.1 Base model selection

To ensure compliance with GDPR, we could not expose any of the patient data to models via an online API; hence, models were run locally. To ensure the projects longevity, we opted to utilise open-source models that could be run using the Hugging Face library[12], in particular, the Llama 3.2, (3 billion parameters) model[10]. This model required less than seven gigabytes of GPU VRAM to generate summaries allowing for efficient training and deployment, ensuring that the trained model can be deployed locally on desktops with limited computational capabilities.

§

2.2 Datasets

We extracted 5550 anonymised historical Cardiac MRI reports (cMRI) from the Guy's and St Thomas' patient database. These reports were screened to remove any that lacked findings or incomplete demographic data (N=10). These reports were randomly divided into three datasets: train (N=3000), RAG embedding (N=2000), and test (N=540) for training the core generative LLM.

After dividing the dataset, we extracted relevant information from each report, including: the ground truth summaries written by clinical cardiologists, volumetric and clinical findings from the MRI scan, and demographic data (age, weight, sex, etc.). These extracted data points for each dataset were written to CSV files.

2.3 Prompt Engineering

Prompt engineering was used to constrain the behaviour of the generative model and align outputs with the requirements of clinical reporting[13]. A structured template was developed that provided the model with explicit guidelines on style, accuracy, and completeness. The prompt opened with a role specification, instructing the model to behave as a specialist in cardiac magnetic resonance imaging[13], [14]. This was followed by detailed principles, such as the requirement to summarise only from the provided findings, to avoid inference or modification of values, and to reproduce all numerical measurements exactly as reported. Additional instructions emphasised the inclusion of both normal and abnormal findings, with abnormal findings described in greater detail, and the explicit mention of cardiac structures such as the ventricles, valves, and atria even when no abnormalities were present.

To ensure consistency across cases, a mandatory inclusion checklist[13] was embedded within the prompt, covering key diagnostic domains including ventricular function, atrial size, presence of late gadolinium enhancement, valve function, aortic measurements, and tissue characterisation. The content structure was also predefined: findings were to be presented as a numbered list beginning with the left ventricle and right ventricle, followed by other abnormalities, and concluding with an "Impression" section to summarise the clinical significance. Accuracy requirements were reinforced by explicitly prohibiting the model from inferring values, approximating measurements, or deviating from the terminology used in the input. Finally, a reasoning section was appended to each summary, separated by a "===== REASONING =====" marker, in which the model was required to explain the clinical rationale behind its output.

By combining explicit clinical rules with format constraints and examples of normal and abnormal outputs, this prompt design reduced variability, improved completeness, and supported downstream evaluation of hallucinations (example in Appendix 1).

§

2.4 Hallucinations and Evaluation

As mentioned previously, the predominant reason preventing wide-scale adoption of LLMs into clinical practice is the presence of hallucinations within generated text[6]. Identifying hallucinations within generated texts is not only important as a metric of the model's performance, but also as a safeguard against clinical errors. Misreporting diagnostic values can lead directly to misdiagnosis, delayed treatment, and ultimately poorer patient outcomes[6].

Hallucinations are defined within natural language processing as the phenomenon of nonsensical, incorrect or unfaithful token generation compared to the provided source and prompt content [15]. Xu et al. formally define hallucinations as the failure of the model to generate tokens within the ground truth function [16]. Hallucinations, whilst incredibly varied, can be split into two fundamental sets: intrinsic hallucinations, where the model's output directly conflicts with the provided prompt context; and extrinsic hallucinations, where the generated output cannot be verified using the provided source context or external knowledge bases [15].

Due to the specificity of the task (generating concise summaries of volumetric findings of cMRIs), the only hallucination type we were concerned with was intrinsic hallucinations. The most common hallucination was a misinterpretation of left ventricular ejection fraction (LVEF) and right ventricular ejection fraction (RVEF). This thus simplified the task of evaluation from a complex multi-classification problem to a binary classification problem of whether there was a hallucination or not.

Generated Summary (hallucinations in red)	**Corrections**	**Hallucination Description**
1. Normal indexed LV end-diastolic volume (95 ml/m2) and global systolic function (LVEF=55%) 2. Normal indexed RV end-diastolic volume (97 ml/m2) and global systolic function (RVEF=50%) 3. No regional wall motion abnormalities at rest 4. Normal LV wall thickness; normal indexed LV mass	'LV end diastolic volume **(94ml/m2)** and global systolic function **(LVEF=53%)**' 'RV end-diastolic volume **(98 ml/m2)** and global systolic function **(RVEF=49%)**'	**Complete failure in transcription of key values**

§

1. Normal indexed LV end-diastolic volume (83ml/m2) and global systolic function (LVEF=62%) 2. Normal indexed RV end-diastolic volume (73ml/m2) and global systolic function (RVEF=59%) 3. No myocardial fibrosis, infiltration, or infarction	'LV end-diastolic volume **(100ml/m2)** and global systolic function **(LVEF = 32%)**' 'RV end-diastolic volume **(80ml/m2)** and global systolic function **(RVEF = 45%)'**	**Complete failure in transcription of key values**
1. Normal indexed RV end-diastolic volume (66ml/m2) and global systolic function (RVEF=55%) 2. Normal indexed LV end-diastolic volume (70ml/m2) and global systolic function (LVEF=59%) 3. No myocardial edema, fibrosis, or infiltration	'RV end-diastolic volume **(138ml/m2)...(RVEF = 54%)**' 'LV end-diastolic volume **(128ml/m2)...(LVEF = 60%)**' 'myocardial **oedema'**	**Complete failure in transcription of key values, spelling error**
1. Normal indexed RV end-diastolic volume (75ml/m2) and global systolic function (RVEF=61%) 2. Right ventricular wall thickness (5 mm in mid-septum) within normal limits 3. No regional wall motion abnormalities or thinning	'Normal indexed RV end-diastolic volume **(89ml/m2)**' 'Left ventricular wall thickness (7mm in mid-septum) within normal limits'	**Failure in transcription of key values, failure in distinguishing left ventricular wall thickness (hallucinated as right).**

Table 1 – Examples of common hallucinations seen when generating cMRI summaries from findings using baseline model.

2.5 Classical Evaluation Metrics

We first evaluated the feasibility of utilising classical quantitative LLM evaluation metrics, including Bilingual Evaluation Understudy (BLEU)[17], Recall-Orientated Understudy for Gisting Evaluation (ROUGE)[18], and Cosine Similarity[19] (amongst others), for assessing model performance. We compared metrics of the baseline generations against ground truth and RAG (K=3 Naive) across 1102

randomly selected, paired samples. These metrics were all calculated using the **pritamdeka/S-PubMedBert-MS-MARCO[20]** embedding model, trained on pub-med abstracts to improve embedding quality of complex medical texts, to embed the generations and ground truths to allow for the calculation of the metrics.

Metric	How its calculated
Bilingual Evaluation Understudy[17]	N-gram overlap between generated and reference texts, using a brevity penalty to punish short outputs.
Recall-Orientated Understudy for Gisting Evaluation[18]	N-gram overlap, with an emphasis on recall, capturing how much of reference content is in generated text. '-L': longest common subsequence '-1': comparing individual word overlap
Cosine-Similarity[19]	Compares vector representations of sentences and computing the cosine of angle between the vectors using BERT- encoder
METEOR[21]	Extends BLEU scoring by incorporating stemming, synonyms, and paraphrase matches rather than just exact lexical matches alone, offering evaluation of semantic overlap.

Table 2 – Quantitative LLM Metrics assessed and how they are calculated

2.6 Hallucination Detection

Due to the poor performance of quantitative LLM metrics (Section 3.1), we utilised the generation hallucination rate as a metric of model performance. To tackle the task of hallucination detection, we first built a labelled database of summaries with hallucinations and without. We randomly subsampled the train dataset (N=450) and generated summaries using the baseline model (Llama 3.2, 3 billion parameters [10]). One hundred of these samples were then augmented by the author to contain a variety of hallucinations, including non-topical sentences (e.g., 'the heart is very cool'), contradictory sentences ('the enlarged ventricles were very small') and changes in values ('LVEF (55%) -> LVEF (35%)'). To ensure class balance and increase the robustness of the dataset, 50 ground truth summaries were injected into the database. The author then labelled the generated summaries as either containing a hallucination (N = 222) or not (N = 278), using the findings and ground truth summary as a reference.

We then used this database to train a custom binary classifier architecture (inspired by Shemalov et al. (2025)[22]); we wrapped a Llama 3.2, 3 billion parameter model in a lightweight fully connected

neural network, in which the network used the normalised pooled embeddings of the LLM as features for classification. Due to the complexity of the features, we opted to utilise skip connections within the neural network head [23] as well as the GELU activation function for greater stability. Due to computational limitations, only the neural network head was trained, whilst the base LLM weights remained frozen during training. The model was trained for 15 epochs, with a learning rate of 5e-4 to avoid overfitting the training data. The training data was divided into an 80/10/10 train/validation/test split, using binary cross entropy as the loss function.

2.7 Retrieval Augmented Generation (RAG)

To enhance factual grounding and improve generation quality, we implemented **Retrieval Augmented Generation (RAG)**. RAG conditions the generative model on relevant, demographically similar historical reports, which are injected into the prompt to reduce hallucinations by providing grounded examples[11].

RAG introduces an intermediate retrieval step in which the input query (composed of the patients' demographics and volumetric findings) is first mapped into a high-dimensional embedding space[11], [24]. Relevant documents are then retrieved from an indexed database of historical reports and appended to the query before being passed into the generative model. This process ensures that the model is guided by clinically validated prior examples, thereby improving factual accuracy and reducing unsupported generations[11], [24], [25].

2.8 Embedding RAG Vector Space

From 2,000 historical reports, three elements were extracted: patient demographics, volumetric findings, and ground-truth summaries (Section 2.2). Using the **BAAI/bge-small-en-v1.5**[26] encoder model, (chosen for its high semantic richness and low resource requirements) the demographics and findings were embedded into a FAISS[27], [28] vector space, whilst the ground-truth summaries were stored as metadata. This design preserved the structure of the document index allowing the formation of clusters (Figure 2) of patients with similar demographics and findings – therefore improving retrieval quality. Ground-truth summaries were not embedded in the same way however as due to the complexity and the lexical similarities within the medical text, encoders struggle to distinguish differences between differing summaries, leading to vector space collapse (with many document vectors becoming co-linear).

§

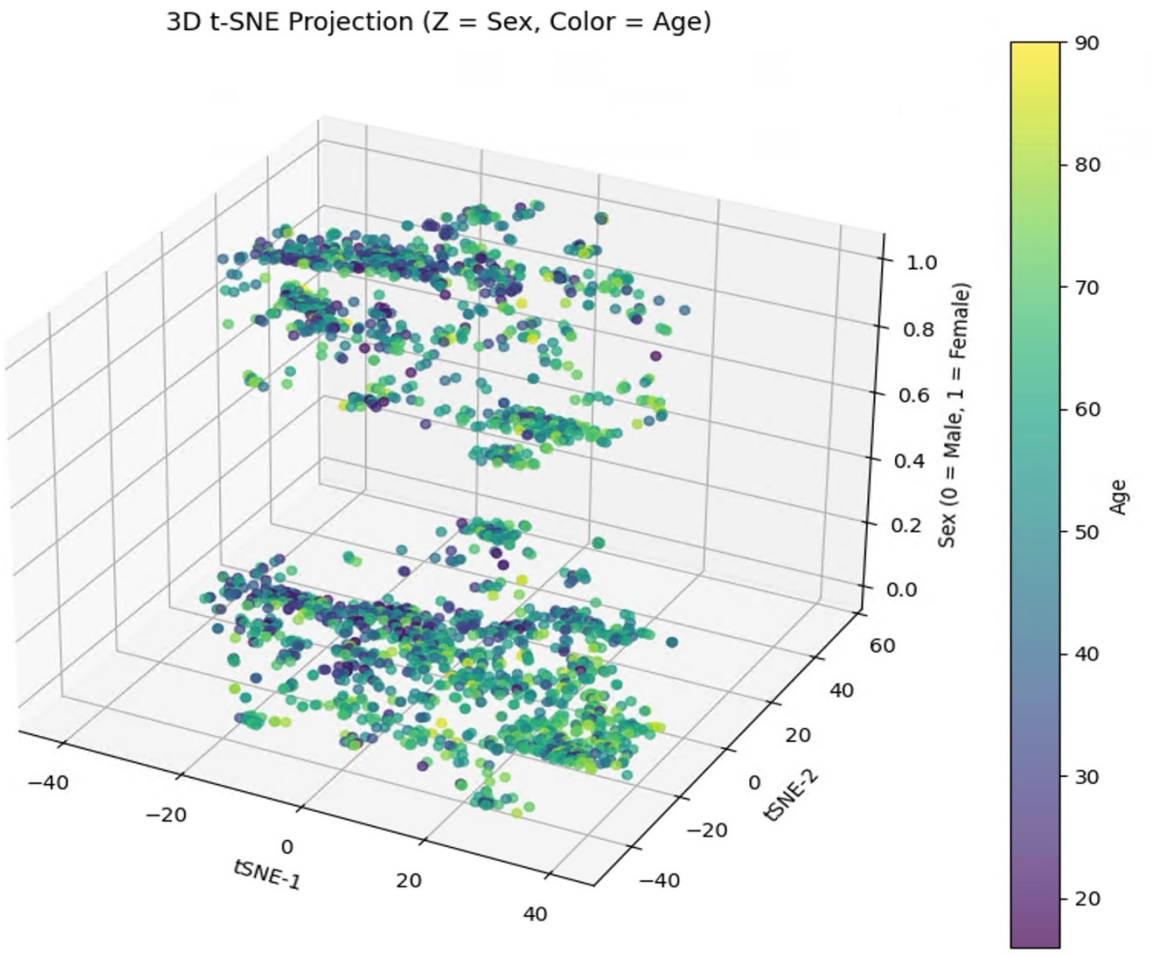


Figure 2 – 3D visualisation of the index vector space. The document vectors are well stratified by sex, with distinct clusters of patients with similar features and findings.

2.9 RAG Retrieval Methods

The query for retrieval was composed of the patient's demographic information and volumetric findings derived from the cMRI scan. Using the same encoder model[26], the query was mapped into the embedding space and compared against the index vector space to identify the most relevant documents for augmentation.

Four retrieval strategies were evaluated. In the **Naïve Top-K** approaches, the K nearest neighbours (*K=1 or K=3)* were directly retrieved based on cosine similarity [19]. In the **hybrid reranking** approaches , the top *N = 10* nearest neighbours were first retrieved using cosine similarity and subsequently reranked using the scoring function defined in **(1)**

§

$$\boldsymbol{Score = \alpha \cdot CosSim(Q_E, D_E) + (1 - \alpha) \cdot CosSim(F_E, S_E)} \quad \textbf{(1)}$$

where $\boldsymbol{\alpha}$ is a tuneable hyperparameter, CosSim denotes cosine similarity, Q_E is the query embedding, D_E the document embedding, F_E the findings embedding and S_E the summary embedding. The top K documents with the highest scores were selected for augmentation.

The inclusion of the **CosSim(F_E,S_E)** term accounted for token- and context-level consistency, such as numerical values (*e.g.*, LVEF = 55%) and diagnostic terms (*e.g.*, left ventricular hypertrophy). This design assumed that, although volumetric findings and narrative summaries are structurally and lexically distinct, key clinical tokens and contextual relationships are preserved. By incorporating this into the reranking function, retrieval was guided not only by lexical similarity but also by clinically relevant contextual alignment.

The selected documents were concatenated with the query findings to form an enriched prompt , which was then passed to the generative model (Llama 3.2, 3B [4]). To allow for faster generation across the full dataset, this retrieval step was performed statically: all prompts were pre-enriched with their retrieved documents prior to inference, rather than querying the retriever dynamically during generation. The base model weights remained frozen, while the enriched context improved factual grounding and overall summary quality by providing clinically relevant exemplars.

2.10 Low Rank Adaptation (LoRA)

Low-Rank Adaptation (LoRA) is a parameter-efficient fine-tuning approach that enables LLMs to be adapted to specialised tasks without updating the full parameter set[9]. Modern transformer-based models such as Llama 3.2 contains billions of parameters, meaning full fine-tuning is computationally expensive and often infeasible on limited hardware (generally, 16GB of GPU VRAM is required per 1B parameters in the model[29]). LoRA addresses this by freezing the original pretrained model weights and introducing small, trainable low-rank matrices that approximate the required weight updates[9], [30].

Given a pretrained weight matrix $W_0 \in \mathbb{R}^{dxk}$ within the transformer (for example, the query or value projection matrix in the attention mechanism), LoRA parameterises the weight update ΔW as the product of two low-rank matrices:

$$W = W_0 + \Delta W = W_0 + BA \quad (2)$$

Where $B \in \mathbb{R}^{dxk}$, A $\mathbb{R}^{rxk}$, and r << min(d,k). By choosing a small rank r, the number of trainable parameters is reduced from d x k (full fine-tuning) to r x (d+k), leading to significant efficiency gains. During training, only A and B are updated, while W_0 remains frozen, preserving the general linguistic knowledge of the base model.

§

In practice, LoRA modules are typically injected into the attention layers of the transformer architecture[9], [30], specifically within the query and value projections, where domain-specific adaptation has the greatest effect on contextual reasoning. A scaling factor α is also introduce to modulate the impact of the low-rank updates on the forward pass:

$$h = W_o x + \frac{\alpha}{r} BAx. \quad (3)$$

where x is the input vector and h the adapted output. This ensures that the contribution of LoRA updates remains stable regardless of the choice of rank.

The choice to use LoRA to fine-tune Llama 3.2 (3B) for the task of cMRI summarisation was motivated by several constraints and requirements. Firstly, due to the limited size of the training dataset (N=3000), full parameter fine-tuning would have been prone to severe overfitting (normal fine-tuning requires N>1e6 datapoints for meaningful training[31]). Secondly, by freezing the base model, LoRA ensured that the pretrained linguistic knowledge was preserved, whilst the small trainable matrices captured the domain-specific clinical features. Thirdly, the LoRA matrices can be stored as a separate, small adapter file, using the HuggingFace library[12]. This allows the model to remain lightweight to distribute and be locally run within NHS intranets, thereby remaining GDPR compliant[32].

2.11 LoRA parameters

The LoRA fine-tuning configuration was selected to balance computational feasibility, training stability, and suitability for the clinical summarisation task. The adapters were applied with a **rank (r)** of **8** and a **scaling factor (α)** of **16**, which provided sufficient capacity for adaptation whilst limiting computational overhead. A **learning rate** of **5 x 10^{-5}** was used, combined with a **cosine learning rate scheduler** and **warmup ratio of 0.05,** to allow the model to stabilise early in training before decaying towards smaller parameter updates. Training was conducted for **10 epochs** with a train-validation split of **0.9/0.1**, ensuring adequate validation monitoring without sacrificing training data. **Token-wise cross entropy loss** was utilised for back-propagation.

To prevent overfitting and improve generalisation, **dropout was set to 0.1,** applied within the adapter layers. LoRA adapters were injected specifically into the query and value projection matrices of the transformer's attention layers, rather than all projection. This design choice was motivated by Hu et al. (2021)[9], who demonstrated that adapting the query-value pathway provides the greatest benefit for task-specific reasoning, while freezing the key and output projections reduces the parameter count and mitigates training instability. This allowed the model to efficiently adapt its attention mechanism to clinical language patterns without unnecessary overhead.

§

2.12 Monte-Carlo Confidence and Uncertainty

To quantify the reliability of generated summaries, we implemented token-wise Monte Carlo (MC)[33], [34] dropout to estimate model confidence and uncertainty. During inference, dropout layers were activated and multiple stochastic forward passes of the same input were performed. By sampling from the model in this way, we obtained a distribution of token probabilities, from which **confidence** (mean probability) and **epistemic uncertainty** (variance)[33] were computed at the token level[34].

Gal and Ghahramani (2016)[33] proposed that variation in output due to dropout can be interpreted as approximate Bayesian inference of deep neural networks. Formally, given an input *x*, the model is evaluated under *T* stochastic forward passes with dropout activated, producing predictions $\mathbf{f(x:\theta_t)}$ where $\boldsymbol{\theta_t}$ denotes the randomly masked parameters at pass t. The predictive mean and variance are then computed as:

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} f(x:\theta_t) \quad (4) \quad , \mathrm{Var}(\hat{y}) = \frac{1}{T}\sum_{t=1}^{`t} f(x:\theta_t)^2 - \hat{y}^2 \quad (5)$$

where $\hat{\mathbf{y}}$ represents the token-wise confidence and $\mathbf{Var(\hat{y})}$ the epistemic uncertainty.

Uncertainty estimation was performed on **10 randomly selected validation samples** at the end of each training epoch. For each sample, the model generated 10 stochastic passes, and token-wise variance was recorded. These values were visualised as heatmaps overlayed on generated text, highlighting regions where the model was most uncertain. In addition, summary-level distributions of uncertainty were aggregated and displayed as **histograms**, allowing comparison between baseline, RAG, and LoRA models.

The primary aim of this analysis was exploratory: to determine whether uncertainty visualisation could highlight regions of generated text that appeared less trustworthy, and whether models adapted with RAG or LoRA exhibited reduced uncertainty compared to the baseline. These results can be used to complement hallucination detection by providing the user with interpretable means of identifying tokens or sections of text where the model was less confident.

§

3.RESULTS

3.1 Classical LLM Metric performance

When testing the performance of the classical LLM metrics across 1102 paired generated samples (Table 3), BLEU and METEOR showed no significant differences ($p > 0.2$, Cohen's $d < 0.05$), indicating complete insensitivity to changes in generation quality. ROUGE-1 and ROUGE-L achieved statistically significant p-values ($p < 0.001$), but with negligible effect sizes ($d \approx 0.10–0.14$), suggesting that significance arose from large sample size rather than any practical improvement.

Metric	Baseline Mean	RAG Mean	Mean Diff	p (t-test)	Cohen's d
BLEU	0.205550	0.203522	-0.002029	0.490996	-0.020754
METEOR	0.282419	0.285840	0.003421	0.226101	0.036484
ROUGE-L	0.382847	0.328727	0.012573	**0.000008**	0.135275
ROUGE-1	0.316154	0.392958	0.010112	**0.000480**	0.105494
Cosine-Sim	0.990167	0.989885	-0.000283	**0.004671**	-0.085391

Table 3 – Statistical Comparison of Baseline vs RAG Generations (N=1102 random paired training set generations) across classical NLP metrics. Despites statistically significant p-values, all effect sizes were negligible ($|d| < 0.2$), indicating no practically meaningful differences.

Cosine similarity also yielded a significant p-value ($p = 0.0047$), but the mean difference was effectively zero and the effect size negligible ($d = -0.09$). Cosine-similarity in particular struggles due to the complexity of medical text, in which clinically distinct phrases are encoded almost colinearly by the encoder (e.g., 'left ventricular systolic dysfunction', 'right ventricular ejection fraction' have a cosine-similarity of 0.985). This thus indicates that surface-level, quantitative metrics are poorly aligned with detection of hallucinations in clinical text and are inappropriate for this pipeline, highlighting the necessity for the development of the Hallucination detection model.

3.2 Hallucination Detection Model Training

The neural network head was trained for 15 epochs utilising a learning rate of 5e-4. These were chosen to balance sufficient learning of the task whilst preventing overfitting to the limited training data.

§

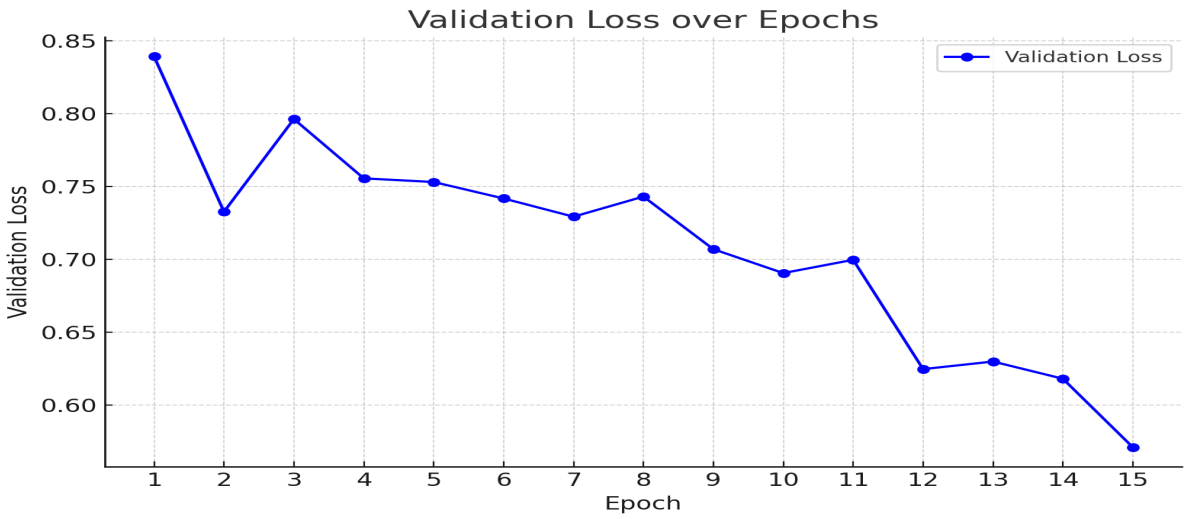


Figure 3 – Validation loss over training

As shown by Figure 3, the validation loss demonstrates a general downward trend across the 15 epochs, indicating that the neural network head successfully learned to classify generated summaries as hallucinated or not, based on pooled embeddings.

Metric	Result
Accuracy	**0.7600**
Sensitivity	**0.8182**
Specificity	**0.7143**
Cohen's Kappa	**0.5223**

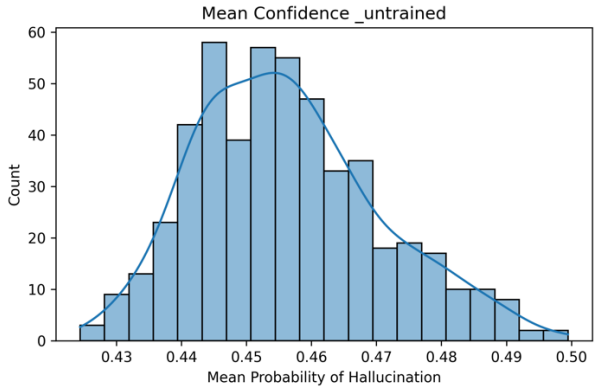
Table 4 – Model performance on test set

This is further supported by the model's performance on the test set (Table 4). The model had an overall accuracy of 0.76, with greater performance at identifying true positives (sensitivity = 0.8182) whilst relatively underperforming in true negatives (specificity = 0.7143). A Cohen's Kappa of 0.52 indicates moderate agreement with the author's labels, supporting the model's suitability for large-scale hallucination detection.

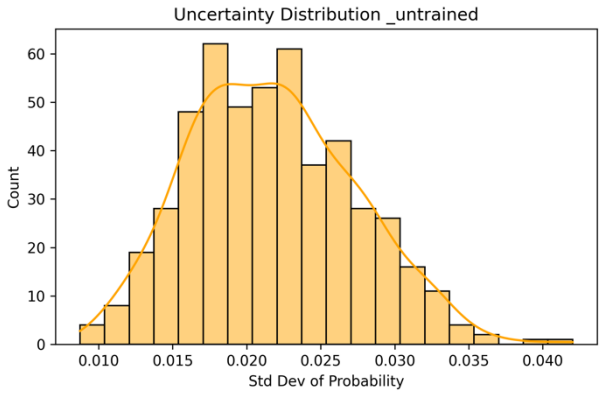
This performance is further illustrated by the distribution of the model confidence and uncertainty before and after training (Figure 4). In the untrained state, the classifier produced a narrow unimodal confidence distribution centred at approximately 0.455, with corresponding low variance in uncertainty scores (Fig 4-A,B). This pattern is consistent with near-random discrimination, where the model fails to separate hallucinated from non-hallucinated summaries. After training, the confidence distribution shifted to a bimodal profile, with density at both low and high probability regions (Fig 4-C,D). At the same time, the uncertainty distribution became right-skewed, with. The majority of predictions concentrated at low variance and smaller tail of higher-uncertainty cases (Fig 4-D). Together with the quantitative test metrics, these distributions indicate that the trained classifier not

only achieved improved separation between classes, but also developed the capacity to express uncertainty in ambiguous cases.

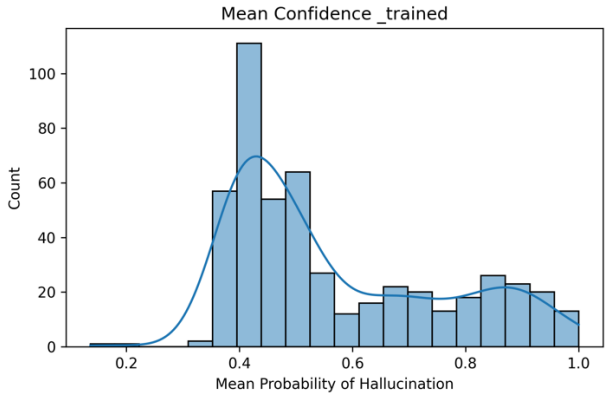
A



B



D



C

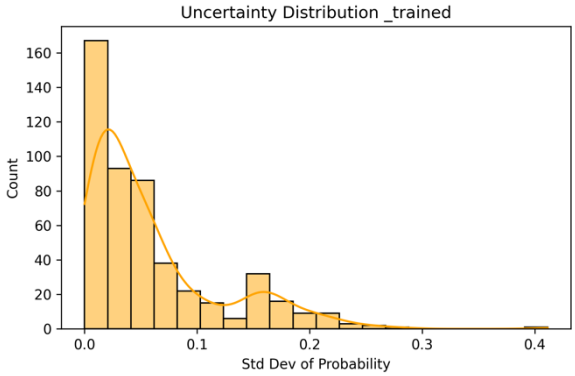


Figure 4 – (clockwise A-D) – (A) Confidence of untrained Hallucination detection model, (B) uncertainty of untrained model, (C) Uncertainty of trained model, (D) Confidence of trained model

3.3 RAG Experiments vs Baseline

We compared baseline inference to four different RAG implementations: naïve retrieval with K=1, naïve retrieval with K=3, hybrid reranking with K=1, and hybrid reranking with K=3 (Fig 5). Baseline performance was poor, with a hallucination rate of 50.4% across 540 test summaries. All RAG variants improved performance by at least 50%, with best result observed for hybrid reranking with K=3, which reduced hallucinations to **12.2%** (75.8% reduction). Naïve retrieval with K=3 achieved a similar reduction (**13.2%**, 73.8% reduction), whereas hybrid reranking with K=1 had the smallest improvement (hallucination rate 24.4%, 51.5% reduction). These results indicate that increasing the number of retrieved documents (K=3) provided more consistent improvements than reranking alone.

§

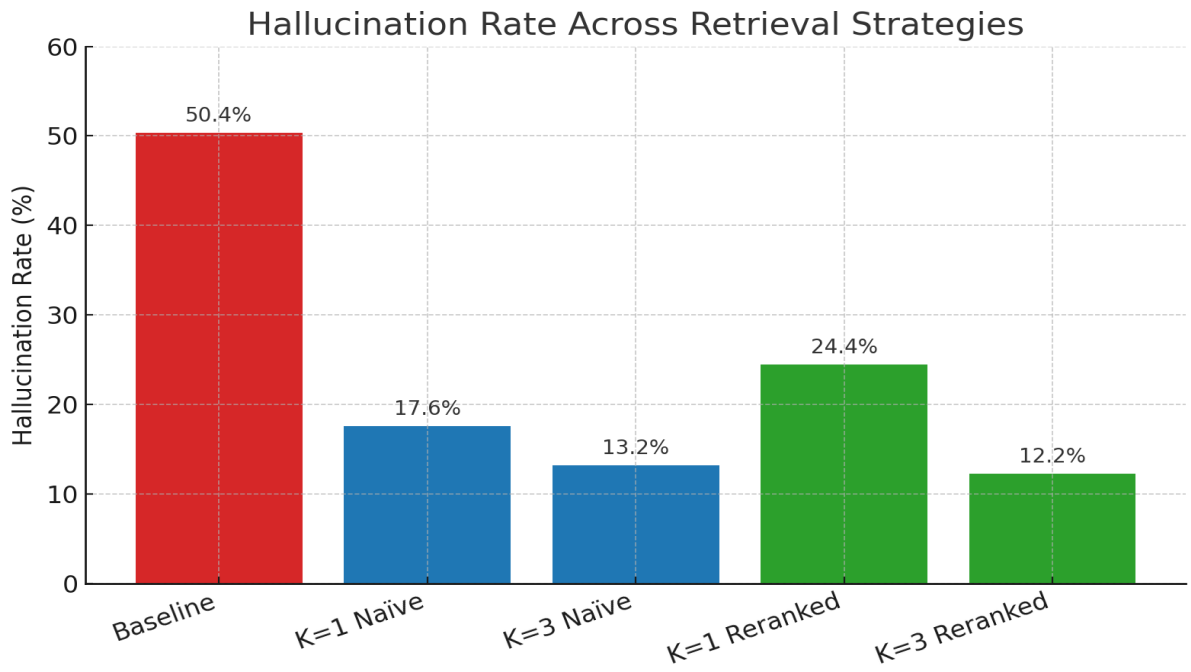


Figure 5 – Hallucination Rates across RAG implementations (N=540)

3.4 LoRA vs RAG K=3 Reranked vs Baseline

We then fine-tuned the base Llama3.2 model using LoRA (section 2.11). Training was conducted on 3,000 samples that were statically pre-enriched using the hybrid RAG rerank K=3 strategy as part of the full pipeline design (Figure 1). Training cross-entropy loss decreased steadily from 1.253, plateauing at 0.86 by epoch 10 (Fig. 6).

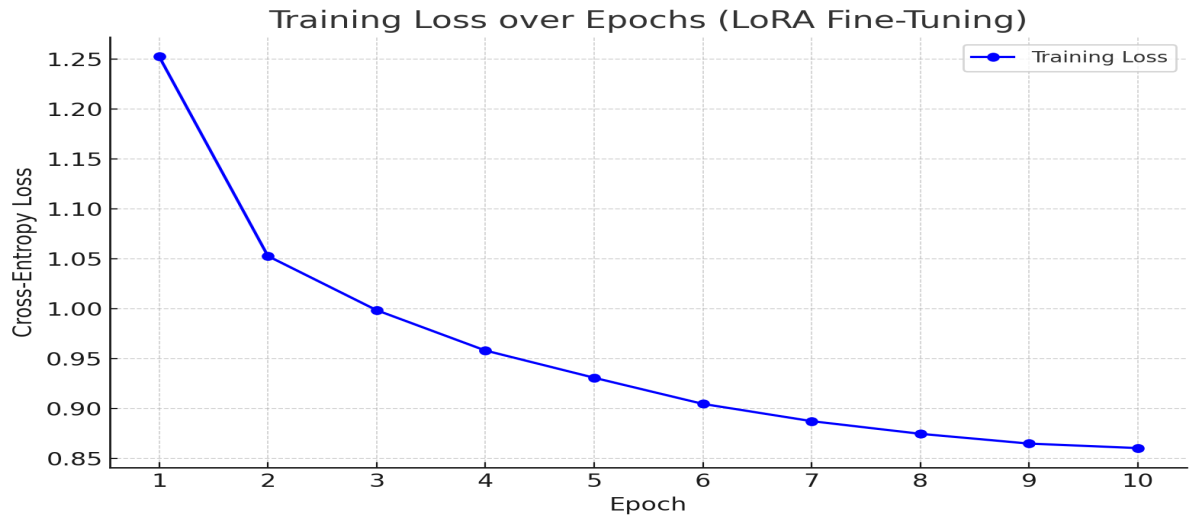


Figure 6 – Training Loss graph

When evaluated with the hallucination detection model, the LoRA-adapted model unexpectedly achieved only marginal improvements over baseline and performed considerably worse than all RAG

§

variants, with a detected hallucination rate of 49.1% (Fig 8). The associated classifier confidence distribution collapsed into a unimodal peak (p=0.5) (Fig 7-A,B), suggesting poor calibration against the automated detection labels indicating a ‘domain-shift’ in the generated summaries from the model (Table 5)[35].

Baseline	Here's a summary of the cardiac MRI study based on the provided report: 1. Normal indexed left ventricular end-diastolic volume (83ml/m2) and global systolic function (LVEF=58%) 2. Normal indexed right ventricular end-diastolic volume (74ml/m2) and global systolic function (RVEF=51%) 3. No regional wall motion abnormalities, thinning, or aneurysm formation 4. No myocardial fibrosis, infiltration, or infarction Impression: A normal cardiac MRI study.
RAG (K=3 Hybrid Reranked)	1. Normal indexed LV end-diastolic volume (normal) and global systolic function (LVEF=60%) 2. Normal indexed RV end-diastolic volume (normal) and global systolic function (RVEF=50%) 3. Normal wall motion, no regional thinning or aneurysm formation 4. Normal LV trabeculation pattern 5. Normal aortic root and main pulmonary artery diameter Impression: Normal cardiac MRI study.
LoRA	1. Visually normal LV size with good systolic function (visual LVEF=60%). 2. Visually normal RV size with good systolic function (visual RVEF=50%). 3. No myocardial fibrosis, infiltration or infarction. 4. No significant intracardiac shunt. 5. No regional wall motion

	abnormalities, thinning, or aneurysm formation. In summary, the findings are in keeping with a normal biventricular size and systolic function. No evidence of fibrosis or scar. No evidence of cardiomyopathy.

Table 5 – Example generations of the same findings query (synthetic) for the Baseline, RAG K=3 Reranked and LoRA models highlight the domain-shift in model inference after fine-tuning. The model output is more expressive (e.g., 'Normal indexed...' vs 'Visually normal...') and matches ground truth label much more closely. Text in Red are hallucinations.

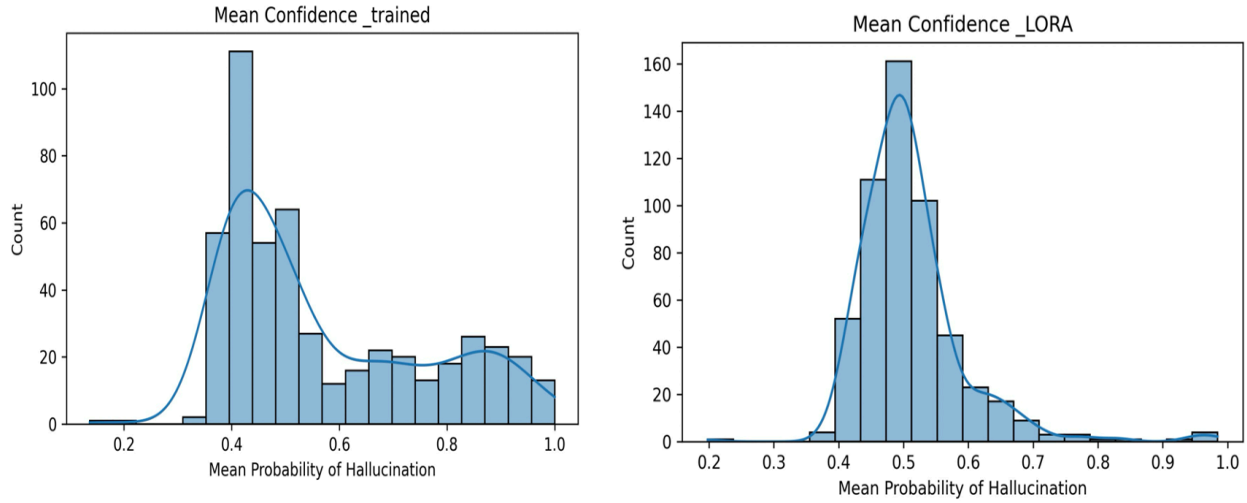
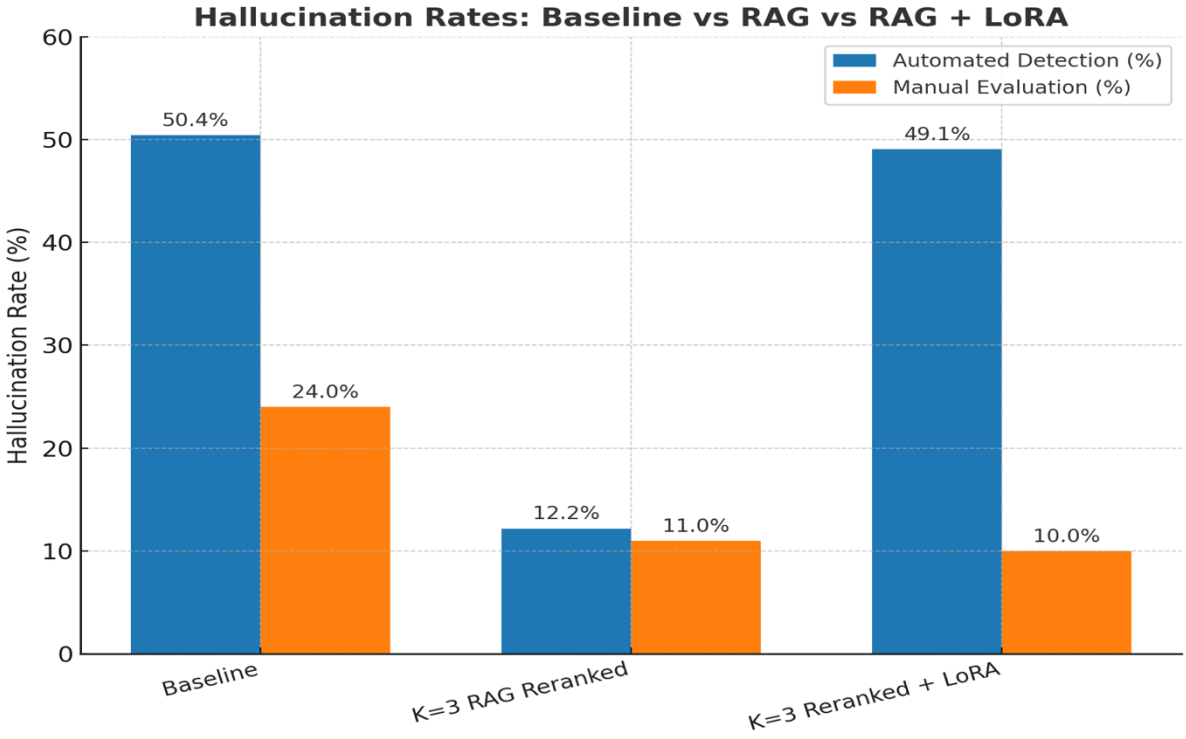


Figure 7 – A (left): confidence distribution of trained model, B (right): confidence distribution on LoRA test set generations.



§

Figure 8 – Automated and Manual Hallucination detection results

To verify the performance independently, manual annotation was performed on 100 randomly sampled generations matched across baseline, RAG K=3 rerank, and LoRA outputs. The baseline hallucinated in 24% of generations, RAG 11%, and 10% for LoRA, indicating substantial improvement relative to baseline and marginal improvement over RAG (Fig 8). During manual review however, LoRA generations frequently included an additional 'conclusion statement' post summary, which in several cases collapsed into nonsensical text. Moreover, when the LoRA model hallucinated, it did so more severely, occasionally generating institutional-specific content (e.g, websites, emails and even a department phone number specific to Imperial and UCL) which was not present in the training data or retrieved documents, likely reflecting artefacts from the base model pretraining corpus.

3.5 Monte Carlo Confidence and Uncertainty

Utilising Monte Carlo Dropout[33], we generated token-wise confidence heatmaps (Appendix 2-4) These showed consistently high confidence for both the baseline (Appendix 2) and RAG K=3 hybrid reranked models (Appendix 3). Baseline generations exhibited mean token confidence of **0.917**, with critical regions (e.g., numerical values such as LVEF) frequently > 0.90. RAG K=3 displayed a similar pattern (mean 0.905, Appendix 3) with high token confidence. After LoRA fine-tuning, however, confidence decreased markedly across tokens, with a mean of 0.764 (Appendix 4).

Uncertainty histograms (Figs. 9-A,B,C) mirrored this shift. For both baseline and RAG, the modal token uncertainty was 0, with extreme left-skewed distributions (Fig. 9-A,B). In contrast, LoRA produced a distribution that was approximately normal with a mode at approximately 0.25 and reduced left-skew (Fig. 9C).

§

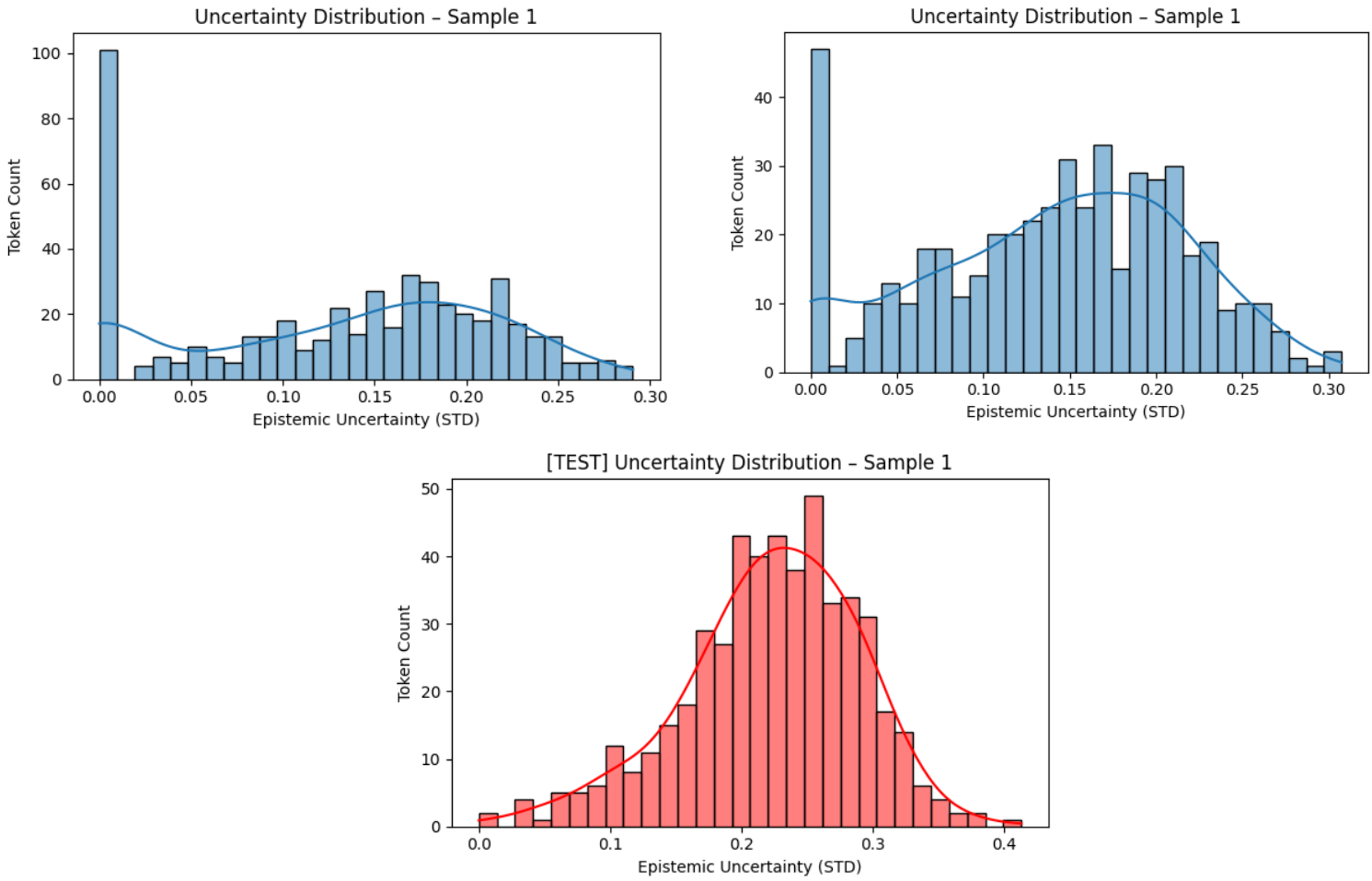


Fig 9A (Top Left) – Baseline token-wise uncertainty histogram, Fig. 9B (Top right) – RAG K=3 Reranked token-wise uncertainty histogram, Fig.9C (Bottom) – LoRA token-wise uncertainty histogram.

Taken together, these results indicate that whilst LoRA improved summary quality (Section 3.4), it also recalibrated the model to be more cautious, lowering per-token confidence and increasing token-level epistemic spread relative to baseline and RAG. The epistemic uncertainty and confidence (specifically the union of tokens with high uncertainty and low confidence) can be used to indicate regions of higher model instability, indicating utility as a flagging signal for potential hallucinations on inference (Appendix 5).

§

4.DISCUSSION

4.1 Classical LLM Metrics

As shown in Section 3.1, classical quantitative metrics commonly applied to LLM evaluation were not appropriate for this project. This was not unexpected given both the nature of the task and the characteristics of the dataset.

As discussed previously (Section 2.5), five metrics are typically employed for LLM analysis: BLEU, ROUGE-L, ROUGE-1, METEOR, and Cosine Similarity[17], [18], [19], [21]. All of these fundamentally rely on comparing the **semantic similarity** between generated summaries and ground-truth references. However, this assumption represents a fundamental flaw in their application here.

Our dataset (N = 5,550) was derived from a single institution, Guy's and St Thomas' Hospital, where a limited number of cardiologists were responsible for reporting cardiac MRI scans. Over time, as guidelines stabilised[2] and expertise converged, reporting styles naturally became **lexically and structurally homogenous.** Moreover, cMRI is primarily used to diagnose recurrent and congenital anatomical and physiological abnormalities[36], meaning that summaries are diagnostically similar across patients despite demographic variation[2]. The result is that reference reports within the dataset exhibit high surface-level similarity. This was further supported by analysis of the RAG index (using a similar BERT model[26]), where despite attempts to introduce diversity through embedding, the space still collapsed towards a conical manifold (mean cosine similarity >0.9) (Fig. 10) which many documents essentially colinear (cosine similarity approximately 1). This collapse also highlights

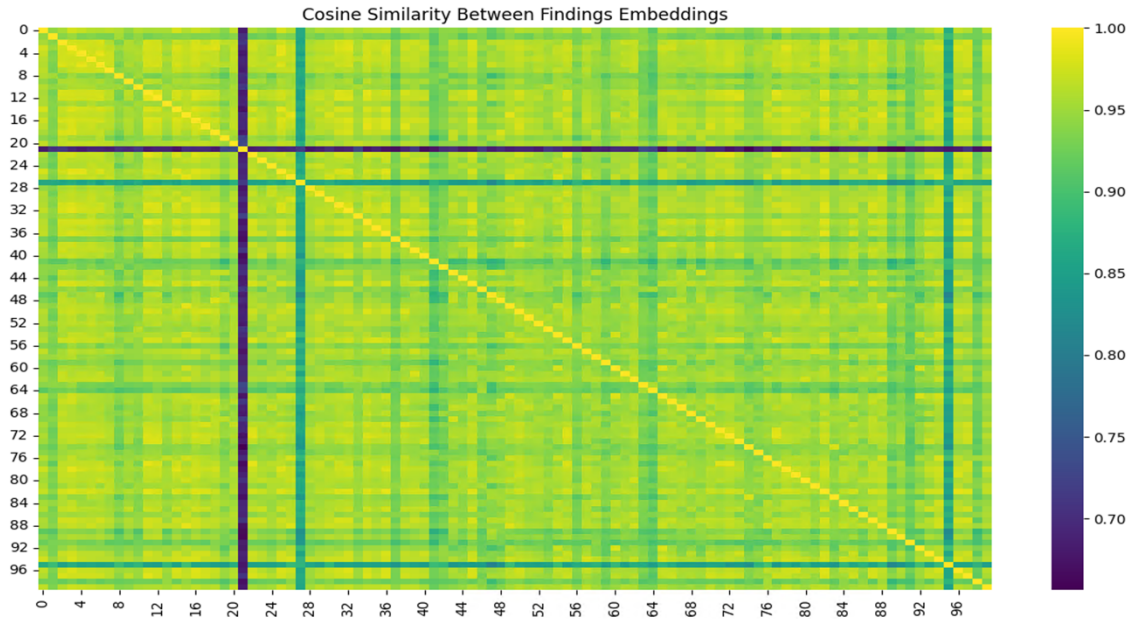


Figure 10 – Cosine Similarity heat map of RAG document index

a limitation of open-source embedding models: even those pre-trained on biomedical corpora[20] often lack granularity in distinguishing specialised medical terminology, a likely consequence of limited access to high-quality clinical data for pretraining.

Finally, as shown in Tables 1 and 5, the generative model rarely collapsed to nonsensical outputs. The predominant hallucination type involved the **substitution of numerical values** (e.g., RVEF reported as 54% vs. 55%). Within the embedding space, such small changes are represented by nearly colinear vectors, and thus are interpreted as highly similar to ground truth. Consequently, the metrics returned values suggesting high similarity even when clinically significant hallucinations were present. This confirms that these metrics cannot be meaningfully interpreted in the context of hallucination detection for clinical text.

The next step that could enable these metrics to be used meaningfully is the development of a custom BERT embedding model trained specifically on labelled cMRI generations and ground-truth summaries. However, given the relatively limited dataset size (N=5,550), this was not feasible within the scope of the project due to the high risk of data leakage and overfitting.

4.2 Hallucination Detection

Given the poor performance of classical metrics, a custom hallucination detection model was developed to identify whether a generated summary contained hallucinations when compared directly against the structured findings. This approach was intended to emulate the information that would realistically be available in clinical use.

Compared to prior work, model performance was modest but promising. Our model achieved an accuracy of 0.76, whereas Shelmanov et al. reported accuracies of 0.88–0.90 across several non-medical domains[22]. The discrepancy is largely explained by dataset scale: our model was trained on only 500 manually labelled samples, while Shelmanov's models leveraged millions of examples spanning multiple tasks[22]. The relative strength of our model despite limited data can be attributed to the simplicity and objectivity of the task. Unlike Shelmanov et al., who employed multi-class classification and inference, our model addressed a binary decision: does the summary contain an intrinsic hallucination or not. This binary framing reduced complexity, avoided arbitrary binning of outputs, and minimised the subjectivity that often arises when clinicians attempt to grade outputs into multiple severity categories.

The rigid lexical structure of cMRI summaries likely further enhanced performance. With relatively standardised phrasing and limited stylistic variation, the model was able to efficiently learn patterns indicative of hallucinations without requiring broader semantic generalisation. Moreover, multi-class classification is widely recognised as a challenge for LLMs[22], [37], particularly in medical text domains where training corpora are scarce. Even models explicitly trained for multi-class tasks (e.g.

Alta/Selene-mini[38]) struggle with reliability in specialised contexts. By contrast, binary classification maximised robustness while also aligning with clinical utility, providing a clear yes/no signal to the user.

Finally, it is important to note that only the neural network head was trained, while the LLM base remained frozen. This constraint arose due to hardware limitations throughout the project. As a result, the classifier was restricted to making predictions from the mean pooled embeddings of the frozen LLM. This architecture inherently limits performance, as the NN head cannot compensate for deficiencies in the base model's ability to capture lexical and semantic nuances. Allowing both the LLM base and the NN head to train in parallel would likely yield significant performance improvements, enabling the embeddings themselves to adapt to clinical language while the head learns to make more accurate predictions.

Ultimately, the hallucination detection model is not intended to replace the clinician but to act as a supportive tool. By flagging summaries with a high likelihood of hallucinations, the model ensures the clinician remains in the loop[39] and retains final oversight, thereby ensuring patient safety.

4.3 RAG Implementations

To improve the factual grounding of generations and reduce hallucinations, we implemented retrieval-augmented generation (RAG) to enrich prompts provided to the generative model. We compared different RAG strategies: pre-retrieval processing alone[8], [40] (where demographics and findings were extracted to generate the query) versus pre- and post-retrieval processing (hybrid reranking)[8], [25], [40], and single-document retrieval (K=1) versus multi-document retrieval (K=3). All RAG variants significantly outperformed baseline generations. The strongest performance was achieved by the hybrid reranked K=3 implementation, which reduced hallucination rates by 75.8%. Both K=3 strategies consistently outperformed K=1, highlighting the value of diversity in retrieved documents. The advantage of hybrid reranked K=3 likely stemmed from the additional $\text{CosSim}(F_E,S_E)$ term, which preserved semantic consistency between findings and summaries during reranking. This is based on the assumption previously defined in Section 2.9, that key semantic ideas are carried from the findings to summaries. This assumption is supported by the very role of summaries: to preserve the key ideas and information from the bulk findings while reducing unnecessary padding. By capturing this alignment, the model was better able to select multiple documents that matched the queried findings and provide richer contextual exemplars[40], thereby reducing hallucinations. In contrast, the hybrid reranked K=1 approach performed worst overall. This may reflect the fact that when only a single document is retrieved, the additional reranking step introduces noise at the document level rather than improving relevance, whereas with multiple retrieved documents the same mechanism enhances selection quality.

§

Unlike typical RAG implementations[40], the full findings and demographics were embedded and retrieved rather than being chunked[40]. Chunking usually improves retrieval diversity and efficiency by splitting long texts into smaller sections; however, in this context it risked discarding clinically essential information spread across the report[40]. Moreover, the reports in our dataset were relatively short (a few hundred tokens), so full-document encoding did not hinder inference performance. We also relied on a single document index, which likely exaggerated the collapse of the embedding space into a conical manifold with high cosine similarity. Future work could mitigate this by constructing multiple indices[40] (e.g., stratified by sex or other demographics) to impose additional structure on the retrieval space, or by supplementing the report index with external corpora such as guideline abstracts or research papers. This would expand the knowledge domain available to the model, reducing the likelihood of extrinsic hallucinations[40].

Due to hardware constraints, RAG was implemented statically, with retrievals and prompt enrichment performed prior to inference. This approach had practical benefits, allowing rapid inference on large batches of reports without requiring both embedding and generative models to be simultaneously loaded into GPU memory. However, it also imposed limitations: the embedding model itself could not be fine-tuned, and retrieval quality was therefore restricted by the representational capacity of the base embedding model. Allowing the embedding space to be dynamically updated during training could regenerate the document index into a more clustered, clinically meaningful manifold rather than the conical structure observed in this project, reducing vector overlap and potentially improving retrieval diversity. Beyond this, more advanced architectures such as **RAPTOR[41]**, which iteratively refines retrieval through recursive inference steps, or the use of a **cross-encoder reranker[42]** to capture fine-grained semantic variation, could be explored to further improve retrieval quality and robustness.

Finally, RAG is inherently scalable[28]. As new patient reports or generated summaries are embedded, the document space can be continuously expanded to cover a wider variety of cases. This makes the system less vulnerable to rare or unseen findings and allows for ongoing improvements in generation quality without the need for expensive re-training of the generative model.

4.4 LoRA Fine-Tuning

To evaluate whether hallucination rates could be reduced beyond retrieval alone, we applied parameter-efficient fine-tuning using low-rank adaptation (LoRA)[9] on the Llama 3.2 model. LoRA was chosen because it allows domain adaptation while training only a small subset of parameters, making it feasible on limited hardware[9] and preserving the frozen base model's general linguistic knowledge. This approach offered a pragmatic balance between performance and deployability, in contrast to full fine-tuning, which would have been computationally prohibitive.

§

When evaluated against baseline and RAG outputs, the LoRA-adapted model performed poorly under the hallucination detection classifier. However, manual annotation revealed modest reductions in hallucination rates, alongside a shift towards more verbose and stylistically refined summaries that more closely resembled expert reports. This discrepancy likely reflects a **domain shift[43], [44]** induced by fine-tuning. Domain shift occurs when the latent space of an LLM is altered through additional training, changing how prompts are interpreted and which regions of the latent space are activated during inference[44]. While domain shift is often viewed negatively due to the risk of *catastrophic forgetting*[43], [44], in which models overwrite prior knowledge, no such forgetting was observed here. Instead, LoRA appeared to recalibrate the model's latent space, improving lexical fidelity and aligning outputs more closely with expert summaries. This suggests that even limited fine-tuning was sufficient to shift the model away from the generic latent regions accessed in baseline and RAG inference, enabling retrieval from more domain-specific subspaces.

Several behavioural changes accompanied this shift. The model developed a tendency to append a distinct **concluding statement** to each summary. While generally separate from the main text, these conclusions often degraded into nonsensical output (e.g., "patient needs to be referred infinitely"), which appeared linked to the hard stopping criterion of the `max_new_tokens` parameter[12]. More concerning were the **failure states** observed in approximately 10% of manually graded samples. In these cases, the model hallucinated institutional artefacts, including contact details for Imperial and UCL hospitals (phone numbers, department emails), and in one instance, the name of a clinician. Inspection of the training set confirmed that such information was absent, strongly suggesting that fine-tuning surfaced latent knowledge embedded in the pretraining corpus. This aligns with the widespread understanding that commercial LLM developers scrape large quantities of publicly available web data, including NHS websites, to train their models[45], [46].

These behaviours highlight a fundamental risk of externally trained LLMs in medicine: fine-tuning can inadvertently expose latent artefacts from opaque pretraining corpora. Unlike open-source models, where such issues can be identified and mitigated, closed commercial APIs provide no visibility into training data, making it impossible to audit for these risks[7]. This raises unresolved questions of responsibility, accountability, and patient data governance, particularly in the context of hostile prompt injection or adversarial attacks, which can trick models into exposing sensitive patient data inadvertently[47].

From a performance perspective, LoRA achieved only marginal improvements in hallucination rates compared with RAG K=3 Reranked. Without access to larger training datasets, further gains are unlikely, as medical text is challenging to augment and deeper training risks overfitting. Two avenues remain most promising. First, acquiring more diverse training data would enable better generalisation

across pathologies and demographics, improving factual grounding. Second, scaling to larger base models such as Llama 3.1 8B[10], [48], [49] would likely provide stronger alignment, as larger parameter counts create richer latent spaces that capture contextual and lexical nuance more effectively. As demonstrated in OpenAI's 2020 work 'Scaling Laws for Neural Language Models' [49] , larger models consistently outperform smaller ones on inference tasks. The challenge is that larger backbones increase hardware demands, potentially limiting local deployability. However, as computational resources become increasingly available, extending this pipeline to larger open-source models may drive hallucination rates further down, while maintaining the transparency and safety advantages of locally deployed systems.

4.5 Uncertainty Mapping

To complement quantitative hallucination analysis, token-wise confidence heatmaps and Monte Carlo dropout[33] (MC dropout) uncertainty histograms were generated for baseline, RAG, and LoRA outputs. These visualisations provided insight into how each intervention influenced model calibration and allowed identification of regions most prone to hallucination.

Baseline and RAG generations were characterised by uniformly high confidence values (mean token confidence >0.9) and strongly left-skewed uncertainty distributions. In practice, this meant that even when hallucinations occurred, the model assigned them high confidence, making errors difficult to distinguish from correct predictions. LoRA fine-tuning, however, produced a strikingly different profile. Average token confidence fell markedly (mean = 0.76), and the distribution of token-level uncertainty shifted from a left-skewed to a near-normal shape with a mode around 0.25 (Fig. 9A). This behaviour suggests a form of recalibration. While LoRA fine-tuning did not eliminate hallucinations entirely, it shifted the model away from assigning uniformly high confidence scores, instead distributing probability mass more cautiously across tokens. In other words, LoRA reduced the tendency of the model to be confidently wrong[34] — a failure mode that is particularly dangerous in medical applications[4]. By broadening the spread of uncertainty, LoRA increased the likelihood that high-risk or ambiguous tokens would be flagged, making residual hallucinations easier to detect. From a clinical safety perspective, such caution is preferable: overconfident errors can be silently accepted by a reader[6], whereas uncertainty signals act as prompts for closer inspection. In practice, this means that even if LoRA does not fully suppress hallucinations, it makes them more visible to clinicians, shifting the model from a black-box generator towards an interpretable assistant. This reframing — from producing flawless text to supporting critical human review — aligns more closely with the clinical reality that AI systems should augment rather than replace expert judgement[4], [6].

Overlay analysis of confidence and uncertainty further demonstrated their potential as safety indicators. Regions where low confidence coincided with high uncertainty frequently aligned with

hallucinated content, providing a visual flagging mechanism that could be surfaced to end-users. This creates the possibility of interactive report generation, where clinicians are alerted to potentially unreliable sections of text and can prioritise these for review. Importantly, such interpretability is only possible in a transparent, locally deployed pipeline. Closed commercial APIs conceal internal calibration dynamics, preventing users from detecting or mitigating these failure modes.

While promising, these methods remain exploratory. No formal quantitative correlation between hallucinations and uncertainty was computed in this project due to time constraints. Nevertheless, the qualitative alignment observed in heatmaps and overlays indicates that token-level calibration metrics hold substantial potential as a complementary safety layer for clinical deployment. The integration of this methodology into the reporting application will inevitably increase computational requirements, as each generation requires multiple forward passes to estimate mean and variance of token probabilities. However, this trade-off is justified by the potential clinical value: uncertainty overlays could act as a red-flag system, drawing clinician attention to tokens that might otherwise be overlooked, and thereby reducing mistakes as this technology becomes integrated into wider practice.

4.6 Overall Interpretation

The results of this project demonstrate that improving the safety of LLM-based clinical summarisation requires a layered approach rather than reliance on a single intervention. Classical NLP metrics such as BLEU and ROUGE proved unable to capture clinically relevant hallucinations, reinforcing the need for a dedicated hallucination detection model. While this classifier achieved only moderate performance, it enabled large-scale quantitative evaluation and provided a foundation for assessing subsequent interventions. Retrieval-Augmented Generation (RAG) produced the most substantial performance gain, reducing hallucination rates by over 70% through the grounding of generations in prior patient reports. LoRA fine-tuning yielded only marginal reductions in hallucination frequency but recalibrated the model's confidence profile, making errors more cautious and stylistically aligning outputs more closely with expert summaries. Finally, token-level uncertainty and confidence mapping offered a promising interpretability tool, highlighting risky tokens for clinician review and reinforcing the potential of uncertainty-aware pipelines.

Crucially, all interventions and upgrades were performed on a small, low-parameter model that can be run entirely locally on commonly available workstations. This ensures compliance with data governance requirements[32] while avoiding dependence on closed commercial APIs[6]. With larger and more diverse training datasets, expanded labelled hallucination corpora, and the application of this pipeline to bigger model architectures[48], future iterations are likely to deliver even stronger performance and reduced hallucination rates. Most importantly, by reducing hallucinations and providing interpretable safety signals, this pipeline enables clinicians to spend less time drafting reports and more time reviewing and validating AI-assisted summaries[6]. In turn, this can improve

patient throughput, reduce waiting times, and ultimately contribute to better clinical outcomes[6]. With the use of the HuggingFace library[12], these models can be easily integrated into the CORE-CMR development pipeline. Overall, this project demonstrates that lightweight, locally deployable LLM pipelines can be engineered to generate clinically safe, concise summaries, and that a multi-component strategy combining retrieval, parameter-efficient fine-tuning, and uncertainty visualisation offers a feasible pathway toward trustworthy clinical AI.

5. Conclusions

This project set out to develop a stable and functional LLM fine-tuning and augmentation pipeline, with the goal of improving cardiac MRI (cMRI) summary generation from volumetric findings by reducing the rate of intrinsic hallucinations within generated text. By creating a custom hallucination detection model and developing methodology to map token-wise uncertainty and confidence, we enabled both the model and the user to become aware of potential errors. This ensures that mistakes are minimised in a context where inaccuracies can lead to misdiagnosis, mistreatment, and ultimately poorer patient outcomes.

We first demonstrated that classical natural language metrics such as BLEU and ROUGE are inappropriate for high-similarity clinical texts like cMRI summaries, as they fail to capture subtle but clinically critical hallucinations. To address this, we developed a custom hallucination detection model capable of quantitatively evaluating model performance across interventions, trained on a semi-synthetic dataset — an approach not previously applied to medical text. Retrieval-Augmented Generation (RAG) achieved the largest reduction in hallucinations (>70%), grounding generations in prior patient reports. LoRA fine-tuning yielded only modest improvements in hallucination rates but refined the style of generations, making them more verbose and lexically aligned with expert reports, while also recalibrating model confidence to reduce overconfidence in high-risk tokens. Finally, token-level uncertainty and confidence mapping provided a novel interpretability layer, surfacing risky tokens for clinicians to prioritise in review.

A key contribution of this project is that all interventions were performed on a small, low-parameter LLM trained with limited computational resources. This demonstrated that meaningful safety improvements can be achieved without relying on very large models or commercial APIs. The resulting system can be run entirely locally on commonly available hardware, ensuring GDPR compliance by keeping all patient data within hospital servers. This makes the pipeline highly scalable and deployable across multiple centres, avoiding the infrastructural and ethical challenges posed by cloud-based solutions. Importantly, by shifting the clinical task from writing to reviewing reports, this pipeline has the potential to reduce reporting times from over an hour to just minutes. Such an improvement can significantly increase patient throughput, reduce diagnostic waiting times, and ultimately improve clinical outcomes.

§

The most significant limitation of this project is the lack of access to a large, diverse database, which constrained both the training of the hallucination detection model and the fine-tuning of LoRA adapters. Access to larger and more heterogeneous datasets would not only improve raw performance but also ensure better generalisation across pathologies, demographics, and institutional reporting styles.

While limitations of this project have been discussed, the results point to several promising future directions. Although this pipeline was developed for cMRI, it could be adapted to other specialities with structured reporting schemas, creating a suite of domain-specific LLMs each paired with its own RAG index and fine-tuned LoRA weights. This vision — a "zoo" of specialist clinical LLMs — could be deployed as a comprehensive package integrated into existing reporting applications. Further advances will require dynamic RAG implementations, larger and more diverse training datasets, and the application of this layered pipeline to bigger base models to maximise performance while maintaining local deployability.

In summary, this project demonstrates that lightweight, locally deployable LLM pipelines can be engineered to reduce hallucinations and improve safety in clinical summarisation. By combining retrieval grounding, parameter-efficient fine-tuning, and uncertainty visualisation, we have shown a feasible path toward trustworthy AI in clinical reporting — one that prioritises accuracy, transparency, and patient safety.

REFERENCES

[1] M. Salerno *et al.*, 'Recent Advances in Cardiovascular Magnetic Resonance Techniques and Applications', *Circ. Cardiovasc. Imaging*, vol. 10, no. 6, p. e003951, June 2017, doi: 10.1161/CIRCIMAGING.116.003951.

[2] B. Herzog, J. Greenwood, and S. Plein, 'Cardiovascular Magnetic Resonance Pocket Guide'.

[3] L. E. Hudsmith, S. E. Petersen, J. M. Francis, M. D. Robson, and S. Neubauer, 'Normal human left and right ventricular and left atrial dimensions using steady state free precession magnetic resonance imaging', *J. Cardiovasc. Magn. Reson. Off. J. Soc. Cardiovasc. Magn. Reson.*, vol. 7, no. 5, pp. 775–782, 2005, doi: 10.1080/10976640500295516.

[4] E. Asgari *et al.*, 'A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation', *NPJ Digit. Med.*, vol. 8, p. 274, May 2025, doi: 10.1038/s41746-025-01670-7.

[5] M. Chandra *et al.*, 'Lived Experience Not Found: LLMs Struggle to Align with Experts on Addressing Adverse Drug Reactions from Psychiatric Medication Use', Jan. 07, 2025, *arXiv*: arXiv:2410.19155. doi: 10.48550/arXiv.2410.19155.

§

[6] C. Lin and C.-F. Kuo, ‘Roles and Potential of Large Language Models in Healthcare: A Comprehensive Review’, *Biomed. J.*, p. 100868, Apr. 2025, doi: 10.1016/j.bj.2025.100868.

[7] X. Meng *et al.*, ‘The application of large language models in medicine: A scoping review’, *iScience*, vol. 27, no. 5, p. 109713, Apr. 2024, doi: 10.1016/j.isci.2024.109713.

[8] P. Lewis *et al.*, ‘Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks’, Apr. 12, 2021, *arXiv*: arXiv:2005.11401. doi: 10.48550/arXiv.2005.11401.

[9] E. J. Hu *et al.*, ‘LoRA: Low-Rank Adaptation of Large Language Models’, Oct. 16, 2021, *arXiv*: arXiv:2106.09685. doi: 10.48550/arXiv.2106.09685.

[10] A. Grattafiori *et al.*, ‘The Llama 3 Herd of Models’, Nov. 23, 2024, *arXiv*: arXiv:2407.21783. doi: 10.48550/arXiv.2407.21783.

[11] Y. Gao *et al.*, ‘Retrieval-Augmented Generation for Large Language Models: A Survey’, Mar. 27, 2024, *arXiv*: arXiv:2312.10997. doi: 10.48550/arXiv.2312.10997.

[12] T. Wolf *et al.*, ‘HuggingFace’s Transformers: State-of-the-art Natural Language Processing’, July 14, 2020, *arXiv*: arXiv:1910.03771. doi: 10.48550/arXiv.1910.03771.

[13] S. Vatsal and H. Dubey, ‘A Survey of Prompt Engineering Methods in Large Language Models for Different NLP Tasks’, July 24, 2024, *arXiv*: arXiv:2407.12994. doi: 10.48550/arXiv.2407.12994.

[14] S. P, ‘Agentic Prompt Engineering: A Deep Dive into LLM Roles and Role-Based Formatting’. Accessed: Aug. 26, 2025. [Online]. Available: https://www.clarifai.com/blog/agentic-prompt-engineering

[15] L. Huang *et al.*, ‘A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions’, *ACM Trans. Inf. Syst.*, vol. 43, no. 2, pp. 1–55, Mar. 2025, doi: 10.1145/3703155.

[16] Z. Xu, S. Jain, and M. Kankanhalli, ‘Hallucination is Inevitable: An Innate Limitation of Large Language Models’, Feb. 13, 2025, *arXiv*: arXiv:2401.11817. doi: 10.48550/arXiv.2401.11817.

[17] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, ‘Bleu: a Method for Automatic Evaluation of Machine Translation’, in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, P. Isabelle, E. Charniak, and D. Lin, Eds, Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, July 2002, pp. 311–318. doi: 10.3115/1073083.1073135.

[18] C.-Y. Lin, ‘ROUGE: A Package for Automatic Evaluation of Summaries’, in *Text Summarization Branches Out*, Barcelona, Spain: Association for Computational Linguistics, July 2004, pp. 74–81. Accessed: Aug. 23, 2025. [Online]. Available: https://aclanthology.org/W04-1013/

[19] N. Reimers and I. Gurevych, ‘Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks’, Aug. 27, 2019, *arXiv*: arXiv:1908.10084. doi: 10.48550/arXiv.1908.10084.

§

[20] P. Deka and A. Jurek-Loughrey, ‘IMPROVED METHODS TO AID UNSUPERVISED EVIDENCE-BASED FACT CHECKING FOR ONLINE HEALTH NEWS’.

[21] S. Banerjee and A. Lavie, ‘METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments’, in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, J. Goldstein, A. Lavie, C.-Y. Lin, and C. Voss, Eds, Ann Arbor, Michigan: Association for Computational Linguistics, June 2005, pp. 65–72. Accessed: Aug. 23, 2025. [Online]. Available: https://aclanthology.org/W05-0909/

[22] A. Shelmanov *et al.*, ‘A Head to Predict and a Head to Question: Pre-trained Uncertainty Quantification Heads for Hallucination Detection in LLM Outputs’, May 13, 2025, *arXiv*: arXiv:2505.08200. doi: 10.48550/arXiv.2505.08200.

[23] K. He, X. Zhang, S. Ren, and J. Sun, ‘Deep Residual Learning for Image Recognition’, Dec. 10, 2015, *arXiv*: arXiv:1512.03385. doi: 10.48550/arXiv.1512.03385.

[24] M. Mortaheb, M. A. A. Khojastepour, S. T. Chakradhar, and S. Ulukus, ‘Re-ranking the Context for Multimodal Retrieval Augmented Generation’, Jan. 08, 2025, *arXiv*: arXiv:2501.04695. doi: 10.48550/arXiv.2501.04695.

[25] ‘Build a Retrieval Augmented Generation (RAG) App: Part 1 | 🦜🔗 LangChain’. Accessed: Aug. 25, 2025. [Online]. Available: https://python.langchain.com/docs/tutorials/rag/

[26] ‘BAAI/bge-small-en-v1.5 · Hugging Face’. Accessed: Aug. 25, 2025. [Online]. Available: https://huggingface.co/BAAI/bge-small-en-v1.5

[27] Stepkurniawan, ‘Comparing RAG Part 2: Vector Stores; FAISS vs Chroma’, Medium. Accessed: Aug. 25, 2025. [Online]. Available: https://medium.com/@stepkurniawan/comparing-faiss-with-chroma-vector-stores-0953e1e619eb

[28] M. Douze *et al.*, ‘The Faiss library’, Feb. 11, 2025, *arXiv*: arXiv:2401.08281. doi: 10.48550/arXiv.2401.08281.

[29] ‘How much VRAM do I need for LLM model fine-tuning?’, Modal. Accessed: Aug. 27, 2025. [Online]. Available: https://modal.com/blog/how-much-vram-need-fine-tuning

[30] ‘LoRA’. Accessed: Aug. 27, 2025. [Online]. Available: https://huggingface.co/docs/peft/en/package_reference/lora

[31] I. Vieira, W. Allred, S. Lankford, S. Castilho, and A. Way, ‘How Much Data is Enough Data? Fine-Tuning Large Language Models for In-House Translation: Performance Evaluation Across Multiple Dataset Sizes’, Sept. 10, 2024, *arXiv*: arXiv:2409.03454. doi: 10.48550/arXiv.2409.03454.

[32] ‘General Data Protection Regulation (GDPR) – Legal Text’, General Data Protection Regulation (GDPR). Accessed: Aug. 27, 2025. [Online]. Available: https://gdpr-info.eu/

§

[33] Y. Gal and Z. Ghahramani, ‘Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning’, Oct. 04, 2016, *arXiv*: arXiv:1506.02142. doi: 10.48550/arXiv.1506.02142.

[34] T. Zhang *et al.*, ‘Token-Level Uncertainty Estimation for Large Language Model Reasoning’, May 16, 2025, *arXiv*: arXiv:2505.11737. doi: 10.48550/arXiv.2505.11737.

[35] ‘Controlling Out-of-Domain Gaps in LLMs for Genre Classification and Generated Text Detection | PromptLayer’. Accessed: Aug. 28, 2025. [Online]. Available: https://www.promptlayer.com/research-papers/closing-the-gap-llms-and-out-of-domain-performance

[36] H. Zareiamand, A. Darroudi, I. Mohammadi, S. V. Moravvej, S. Danaei, and R. Alizadehsani, ‘Cardiac Magnetic Resonance Imaging (CMRI) Applications in Patients with Chest Pain in the Emergency Department: A Narrative Review’, *Diagnostics*, vol. 13, no. 16, p. 2667, Aug. 2023, doi: 10.3390/diagnostics13162667.

[37] M. Chandra *et al.*, ‘Lived Experience Not Found: LLMs Struggle to Align with Experts on Addressing Adverse Drug Reactions from Psychiatric Medication Use’, Jan. 07, 2025, *arXiv*: arXiv:2410.19155. doi: 10.48550/arXiv.2410.19155.

[38] A. Alexandru *et al.*, ‘Atla Selene Mini: A General Purpose Evaluation Model’, Jan. 27, 2025, *arXiv*: arXiv:2501.17195. doi: 10.48550/arXiv.2501.17195.

[39] A. Choudhury and Z. Chaudhry, ‘Large Language Models and User Trust: Consequence of Self-Referential Learning Loop and the Deskilling of Health Care Professionals’, *J. Med. Internet Res.*, vol. 26, p. e56764, Apr. 2024, doi: 10.2196/56764.

[40] Y. Gao *et al.*, ‘Retrieval-Augmented Generation for Large Language Models: A Survey’, Mar. 27, 2024, *arXiv*: arXiv:2312.10997. doi: 10.48550/arXiv.2312.10997.

[41] P. Sarthi, S. Abdullah, A. Tuli, S. Khanna, A. Goldie, and C. D. Manning, ‘RAPTOR: Recursive Abstractive Processing for Tree-Organized Retrieval’, Jan. 31, 2024, *arXiv*: arXiv:2401.18059. doi: 10.48550/arXiv.2401.18059.

[42] R. Ashman (PhD), ‘The aRt of RAG Part 3: Reranking with Cross Encoders’, Medium. Accessed: Aug. 28, 2025. [Online]. Available: https://medium.com/@rossashman/the-art-of-rag-part-3-reranking-with-cross-encoders-688a16b64669

[43] ‘The Hidden Challenges of Domain-Adapting LLMs’. Accessed: Aug. 28, 2025. [Online]. Available: https://www.arcee.ai/blog/the-hidden-obstacles-of-domain-adaptation-in-llms

[44] I. Muttakhiroh and T. Fevens, ‘Tackling Distribution Shift in LLM via KILO: Knowledge-Instructed Learning for Continual Adaptation’, Aug. 05, 2025, *arXiv*: arXiv:2508.03571. doi: 10.48550/arXiv.2508.03571.

[45] ‘Cloudflare Just Changed How AI Crawlers Scrape the Internet-at-Large; Permission-Based Approach Makes Way for A New Business Model | Cloudflare’. Accessed: Aug. 28, 2025.

[Online]. Available: https://www.cloudflare.com/en-gb/press-releases/2025/cloudflare-just-changed-how-ai-crawlers-scrape-the-internet-at-large/

[46] E. Creamer, '"Meta has stolen books": authors to protest in London against AI trained using "shadow library"', *The Guardian*, Apr. 03, 2025. Accessed: Aug. 28, 2025. [Online]. Available: https://www.theguardian.com/books/2025/apr/03/meta-has-stolen-books-authors-to-protest-in-london-against-ai-trained-using-shadow-library

[47] 'Prompt Injection Attacks: What You Need to Know', CalypsoAI. Accessed: Aug. 28, 2025. [Online]. Available: https://calypsoai.com/insights/prompt-injection-attacks-what-you-need-to-know/

[48] J. Kaplan *et al.*, 'Scaling Laws for Neural Language Models', Jan. 23, 2020, *arXiv*: arXiv:2001.08361. doi: 10.48550/arXiv.2001.08361.

[49] T. B. Brown *et al.*, 'Language Models are Few-Shot Learners', July 22, 2020, *arXiv*: arXiv:2005.14165. doi: 10.48550/arXiv.2005.14165.

APPENDIX

Appendix 1.

Below is an example prompt used throughout training after K=3 Reranked concatenation at the end. Other than the three summaries, it is identical in structure to Baseline prompts.

"CRITICAL GUIDELINES: You are a specialized cardiac magnetic resonance imaging expert. SUMMARY PRINCIPLES: - Match the style and conciseness of the example summaries provided - You must only summarise based on the input findings below. Do not infer or modify clinical values. Do not use any numerical values unless they appear in the input findings. - For abnormal findings, include more detail focused on the abnormalities - For complex cases, ensure all clinically significant findings are included - Adapt the level of detail based on the clinical significance - For normal findings, still include all available numerical values and explicitly mention each cardiac structure (LV, RV, LGE, valves, etc.), even if normal. INCLUSION CHECKLIST: Always include — if available in findings: - LV volumes and systolic function (e.g., EDV, LVEF) - RV volumes and function (RVEF) - Presence or absence of LGE or fibrosis - Wall motion/thickening abnormalities - Valve findings or flow abnormalities - Aortic measurements - T2 or edema findings If any of these sections are missing from the summary, your summary will be considered incomplete. CONTENT STRUCTURE: - Present as a numbered list without a "Summary:" header - Start with LV findings, then RV findings, followed by any significant abnormalities - Include an "Impression:" section that briefly states the clinical significance ACCURACY REQUIREMENTS: - Use EXACT numerical values from the report - do not change or approximate - Use the same terminology as in the original findings when possible - ONLY USE VALUES FROM INPUT FINDINGS, DO NOT INFER OR EXTRACT VALUES FROM

§

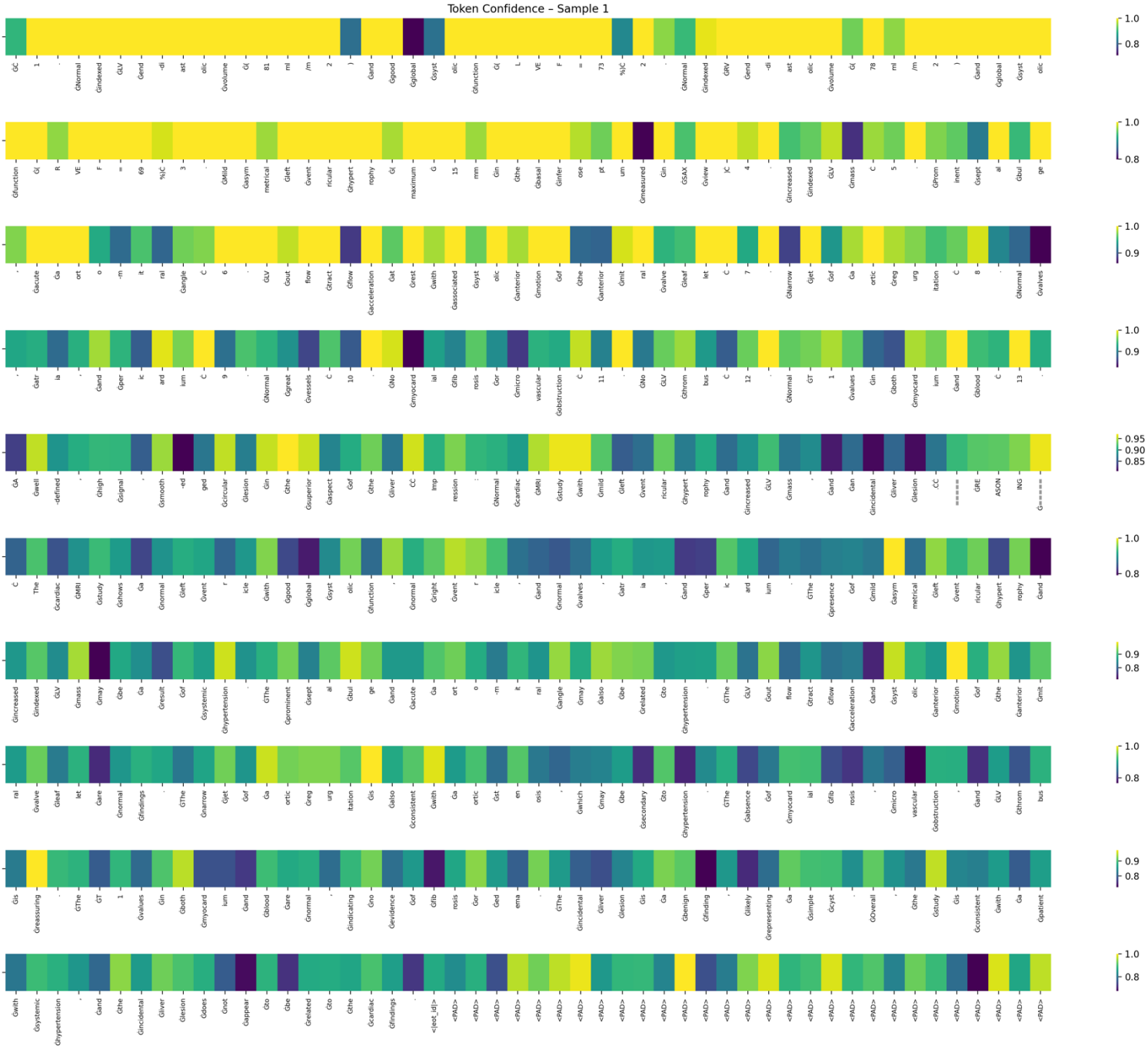
EXAMPLE SUMMARIES - Include interpretive statements only when they appear in the findings EXAMPLE OF IDEAL SUMMARY FORMAT: 1. Normal indexed LV end-diastolic volume (68ml/m2) and global systolic function (LVEF=68%) 2. Normal indexed RV end-diastolic volume (70ml/m2) and global systolic function (RVEF=65%) 3. No myocardial fibrosis, infiltration, or infarction Impression: Normal cardiac MRI study. EXAMPLE OF ABNORMAL FINDINGS FORMAT: 1. Moderate-severely dilated LV with severe systolic impairment (LVEF 22%) 2. Thinned, dyskinetic mid inferolateral and inferior LV walls 3. Normal RV size with mild global systolic impairment (RVEF 44%) 4. Near transmural myocardial infarction of mid inferolateral and inferior LV walls Impression: Severe LV dysfunction with evidence of prior myocardial infarction. REASONING SECTION: - Include a reasoning section after summary, separated by "===== REASONING =====" - Explain clinical significance of findings and rationale for your summary These are the closest EXAMPLE SUMMARIES for reference: Summary 1: '- Normal biventricular end-diastolic volumes and systolic function (LVEF=62%, RVEF=65%). - Normal sized atria. - No inducible myocardial perfusion defect. - No evidence of inducible perfusion defect. - No myocardial infarction, fibrosis or infiltration. - Prominent epicardial fat. In conclusion, there is no evidence of ischemia or cardiomyopaathy. Normal cardiac study.' Summary 2: 'Main findings: - Normal LV end-diastolic volume and systolic function (LVEF=68%) - No LV regional wall motion abnormalities. - Normal RV end-diastolic volume and systolic function (RVEF=61%). - Extensive inducible perfusion defect involving the LV inferior wall and inferolateral segments. - Very small area of subendocardial infarction in the LV basal inferior wall. Impression: inducible ischaemic in the RCA territory. Myocardial viability is preserved in perfusion territories.' Summary 3: 1. Normal indexed LV end-diastolic volume with good overall systolic function (LVEF=62%). 2. Normal indexed RV end-diastolic volume with good systolic function (RVEF=61%). 3. There is inducible myocardial ischaemia in the RCA territory. Additional ischaemia in the circumflex territory cannot be excluded (depending on the coronary anatomy) due to incomplete visualisation of the high lateral wall. 4. Small aplcal LV thrombus (we note the patient is already anticoagulated). 5. Transmural infarction of the apical inferior wall and apex. The remaining LV segments are viable. FINDINGS for analysis: 'LEFT VENTRICLE Normal indexed LV end-diastolic volumes (92 ml/m2) and systolic function (LVEF 64%). Focal hypokinesia of the apical anterior wall. Normal LV wall thickness (maximum 9 mm in the basal anteroseptum). Normal indexed LV mass. RIGHT VENTRICLE Normal indexed RV end-diastolic volumes (89 ml/m2) and systolic function (RVEF 64%). No RVH. ATRIA Left: moderately dilated at 30 cm2 (normal value < 24 cm2). Right: mildly dilated at 26 cm2 (normal value < 22 cm2). VALVES The aortic, mitral and tricuspid valves appear structurally and functionally normal. PERICARDIUM The pericardium is normal in thickness and appearance. No significant pericardial effusion. GREAT VESSELS Normal calibre of the aortic root (with maximum end-diastolic diameter of 36 mm at the Sinus of Valsalva in the 3-chamber view). Normal dimensions of the ascending (31 mm at RPA level) and remaining thoracic aorta. Normal main pulmonary artery

§

diameter (21 mm at bifurcation). STRESS PERFUSION Following adenosine infusion (140 mcg/kg/min for 3 minutes), heart rate rose from 61 bpm at rest to 79 bpm at peak stress and blood pressure changed from 119/75 mmHg at rest to 121/73 mmHg at peak stress. The patient experienced shortness of breath and a heavy chest. Stress perfusion imaging reveals no significant inducible perfusion abnormality. TISSUE CHARACTERIZATION Following gadolinium contrast injection, in the early phase, there is no LV thrombus or microvascular obstruction. In the late phase, there is focal near transmural (51-75% transmurality) enhancement of the apical anterior wall, with subendocardial (1-25% transmurality) extension into the apical lateral segment and apex. T1 mapping (MOLLI): Normal septal myocardial T1 values on 3T (1226 ms). Post-contrast myocardial T1 values at the equilibrium 540 ms. Estimated synthetic haematocrit 35%. Normal calculated extracellular volume (ECV) 24%.'"

§

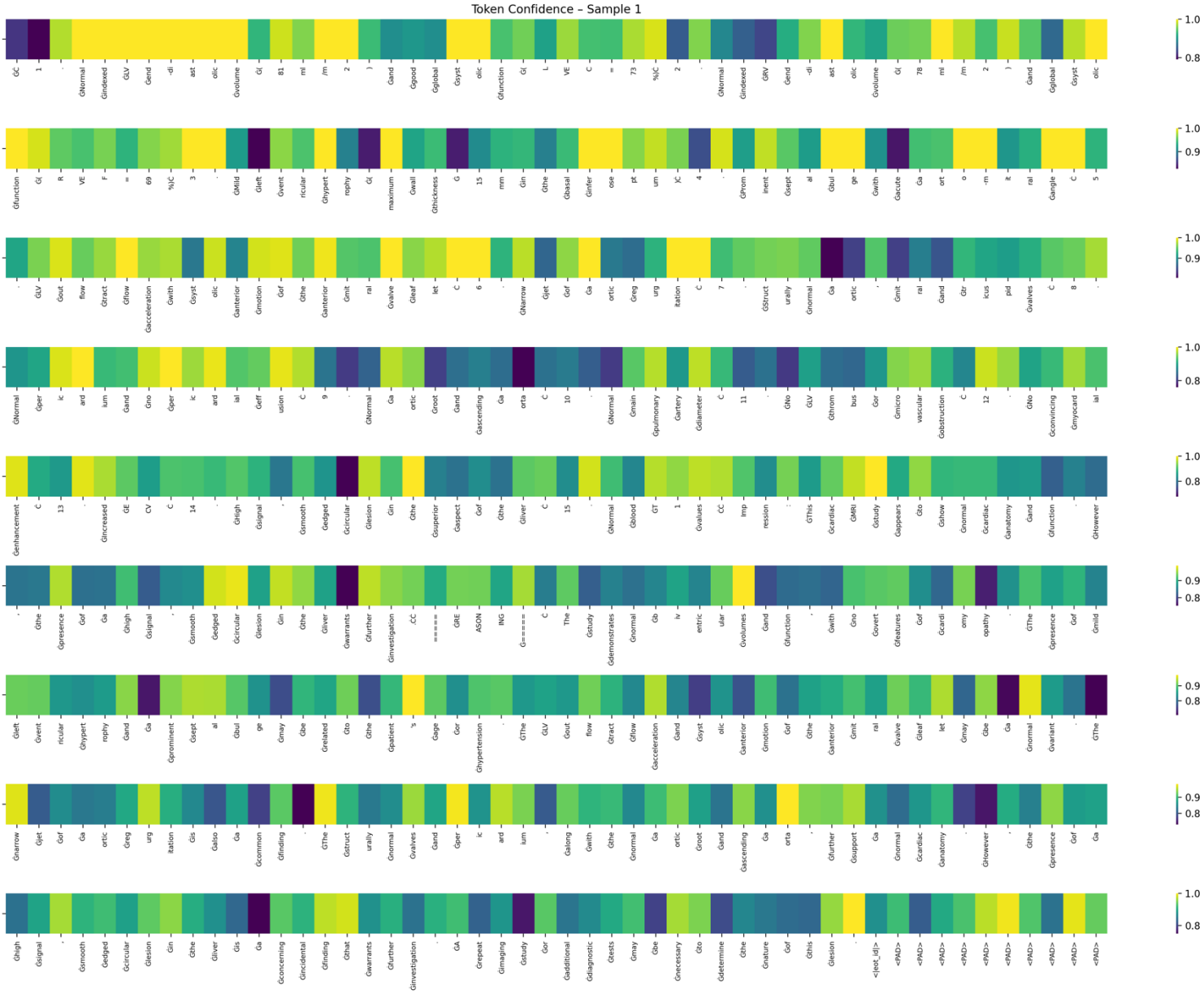
Appendix 2



Heatmap – Baseline token-wise confidence

§

Appendix 3



Heatmap – Token-wise confidence RAG

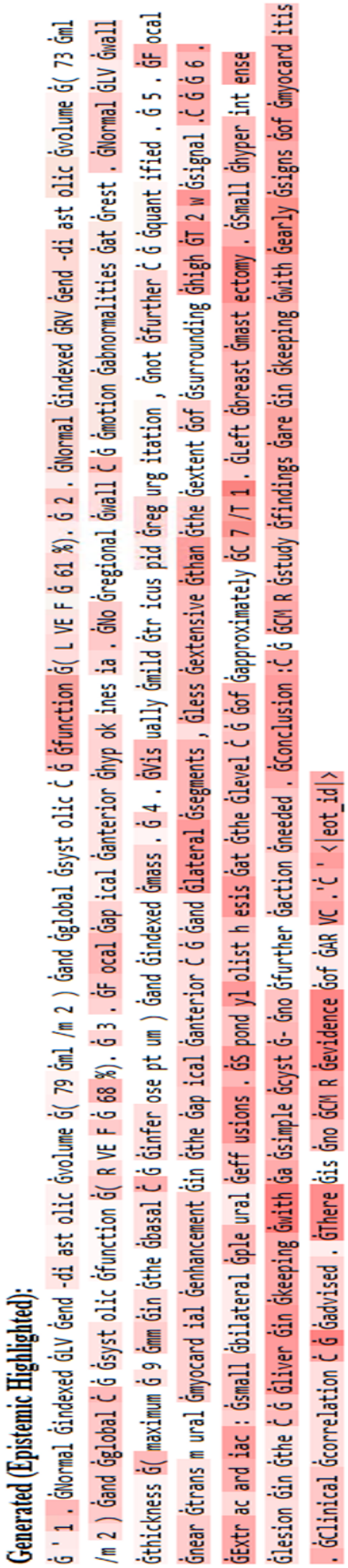
§

Appendix 4

Heatmap – Token-wise confidence LoRA

§

Appendix 5



Example of low confidence, high uncertainty heatmap that can be utilised as a flagging feature for potential hallucinations. Threshold for deciding union can be tuned and changed by user.

§

§