Paediatric Perioperative Risk Stratification Using Large Language Models

Zarif Solaiman (23640056)

Supervisors:
Prof. Tim French
Prof. Wei Liu
Dr. Harry Smallbone (WA Health)

This report is submitted as a partial fulfillment of the requirements for the Computer Science Honours Research Project at The University of Western Australia

2025

Abstract

Perioperative risk stratification involves systematically evaluating and categorising a patient's risk of complications around the surgical period (before, during, and after surgery) and is crucial for anaesthetic management. Existing risk stratification tools such as American Society of Anaesthesiologists Physical Status (ASA-PS) classification are often subjective and limited, especially in paediatrics. This research explores the potential of Large Language Models to predict perioperative risk metrics from unstructured paediatric clinical notes.

We evaluate a domain-specific model, MedGemma-27B-text-it, against a general-purpose model, Llama-3.1-8B-Instruct, on three key tasks: predicting the pre-operative ASA classification, post-operative respiratory adverse events during emergence from anaesthesia, and identification of risk factors of perioperative respiratory complications. Performance was assessed across various prompt engineering techniques, including zero-shot, few-shot and Chain-of-Thought prompting.

MedGemma consistently outperformed Llama on the ASA-PS classification task, achieving its best performance using the few-shot CoT strategy. These results highlight that structured reasoning and domain-specific pretraining are beneficial for nuanced medical tasks. Performance on the AnyEmerg task was challenged by a heavily imbalanced distribution, resulting in low precision for the minority class.

The findings demonstrate the potential of LLMs for paediatric perioperative risk stratification, but also emphasises the need for further extensive research into this area and crucial clinician oversight and validation.

Keywords: Large Language Models, Perioperative Risk Stratification, Paediatric Healthcare, Clinical Decision Support, Machine Learning
CR Categories: I.2.7, I.2.6, J.3

AI Declaration

The following Artificial Intelligence (AI) tools were used in the preparation of this document:

- OpenAI ChatGPT (GPT-5)
- Anthropic Claude Sonnet 4.5

ChatGPT (GPT-5) was used to improve the clarity and academic tone of written content, including paraphrasing and enhancing readability.

Example prompt: “Read through this section [...] and improve the flow and readability of the paragraph.

Claude Sonnet 4.5 was used for LaTeX formatting, especially for tables and figures.

Example prompt: “Given this zero-shot prompt for ASA classification, create me a figure summarising the structure of the prompt.”

In summary, the purpose of using AI tools for this project was to refine language and for LaTeX formatting.

Acknowledgements

I wish to express my sincere gratitude to my supervisors, Professor Tim French, Professor Wei Liu, and Dr. Harry Smallbone, for their continuous guidance and support throughout this project. Their insightful feedback, recommendations for relevant research papers, and valuable discussions during our weekly meetings, have provided clarity and understanding during this project.

I am especially grateful to Dr. Harry Smallbone for his assistance with onboarding to the high-performance computing (HPC) environment at Perth Children's Hospital, without which this project would not have been possible, and for continuously sharing his expertise in the paediatric domain to deepen my understanding of the research context.

I also gratefully acknowledge the support of the Stan Perron Charitable Foundation through access to Merlin HPC at Perth Children's Hospital for this research.

Finally, I would like to thank my family for their unwavering support and encouragement throughout the year, and my friends for their motivation and companionship during this journey.

Contents

List of Tables

List of Figures

CHAPTER 1

Introduction

1.1 Motivation

The primary motivation for this research is to address the limitations of existing risk stratification tools by harnessing the power of Large Language Models (LLMs). Unstructured paediatric clinical notes contain rich, granular information that is often underutilised in paediatric perioperative risk stratification. LLMs have been demonstrated to surpass many textual document understanding tasks especially in medical text processing [1, 2]. By employing LLMs, this study seeks to extract value from unstructured text data by demonstrating the potential for automatically extracting complex risk outcomes such as the American Society of Anaesthesiologists Physical Status (ASA-PS) and identifying respiratory risk factors that lead to perioperative complications in children.

1.2 Background

Perioperative adverse events are complications that occur in the period surrounding surgery (before, during and after operation), and impact patient recovery and outcomes. These complications contribute to increased morbidity, mortality, and prolonged hospital stays, resulting in additional healthcare costs [3]. The occurrence of these perioperative complications can be attributed to patient-related factors and conditions of surgery. Patient-related factors include the patient's age, any existing comorbidities (e.g. cardiovascular diseases) and cognitive status. Surgical conditions such as the type, duration, blood loss, type of anaesthesia, and complexity of the procedure also play a role in the emergence of these adverse events [3]. Perioperative adverse events can be classified into three distinct categories by time. These are short, medium, and long-term complications. Short-term complications include immediate postoperative adverse events. Longer term complications may arise from intraoperative complications that result in long-lasting physiological or development impacts, for example, nerve injuries or surgical trauma.

The process of anaesthesia can typically be divided into several stages: preparation, induction (when anaesthetic agents or medication is administered to render the patient unconscious), maintenance (keeping the patient under anaesthesia during the surgery), and emergence where anaesthetic gas or IV medication stops and the patient recovers from unconsciousness [4]. Perioperative complications can arise during any stage of anaesthesia, influenced by type of anaesthesia and choice of airway management (for example, `laryngeal mask airway` versus `tracheal intubation`) Tailored anaesthetic plans including selection of anaesthetic agents and techniques are essential to minimising the risk of these complications [5].

In paediatric populations, there are additional challenges that arise because of their unique physiological and developmental factors. Compared to adults, paediatric patients have increased inflammatory responses to surgical stimuli and also experience heightened sensitivity to pain as a result of their developing pain modulation pathways [5]. Furthermore, unlike adults, most children present with few or no comorbidities, meaning that traditional risk assessment that rely on underlying diseases are less informative and useful. This absence of comorbidity-related factors make it increasingly difficult to predict adverse perioperative outcomes, especially in newborns and infants where anaesthetic risks are inherently higher. The respiratory system in children is anatomically and physiologically different to the adult respiratory system, especially the airway which is smaller in diameter and shorter in length compared to an adult's. This can result in an increased risk of perioperative respiratory and airway-related complications [6].

Common immediate postoperative complications in paediatric patients include respiratory complications, postoperative pain, emergence delirium and postoperative nausea and vomiting (PONV). Notably, respiratory complications account for 60% of all anaesthesia-related complications [5] and despite advances in paediatric anaesthesia management, perioperative respiratory adverse events (PRAE) remain a leading cause of morbidity and mortality in paediatric patients [6].

Perioperative risk stratification involves systematically evaluating and categorising a patient's risk of complications around the surgical period. Perioperative risk stratification also considers the risk of morbidity and mortality. In paediatric settings, these methods can be used to guide anaesthesia planning, guide clinical decision-making and supports communication with parents or guardians. Paediatric perioperative risk stratification involves a comprehensive preoperative assessment of the child tailored to their unique physiological and anatomical characteristics [7]. Unlike adults, paediatric patients especially newborns, infants, and those with comorbidities such as obstructive sleep apnoea and respiratory infections, experience higher risks of respiratory complications. As a consequence, management strategies differ substantially from adults, often requiring careful airway management, attentive postoperative monitoring especially in intensive care settings, and individualised anaesthetic care plans to mitigate risks [7]. Traditional risk stratification methods such as the American Society of Anaesthesiologists Physical Status (ASA-PS) classification [8], typically involve risk scoring systems based on a fixed set of clinical input features and subjective clinician judgement. Similarly, the STBUR (Snoring, Trouble Breathing, Un-Refreshed) questionnaire is a categorical screening tool designed with the purpose of identifying children at risk of perioperative respiratory adverse events [9]. While both the ASA-PS and STBUR methods provide structured frameworks for risk assessment, they remain limited in their predictive power. The ASA-PS classification was originally developed for adult populations [10] and does not fully capture paediatric physiological and development differences. The STBUR questionnaire has a narrow focus on sleep-disordered breathing and fails to consider broad perioperative risk factors. Therefore, existing risk stratification tools offer only partial insights into perioperative risk in paediatric patients.

1.2.1 ASA Physical Status Classification

The ASA-Ps classification system has been used for over 60 years. The purpose of the ASA-PS classification is to assess the patient's pre-operative health status and comorbidities. It was originally developed as a statistical tool for recording data for anaesthetic care, but has since become a widely adopted substitute for estimating perioperative risk. The classification system is simple to use and does not require extensive diagnostic tests or inclusion of sophisticated data.

The ASA-PS classification categorises patients into five levels of physical status, ranging from ASA I (a normal, healthy patient) to ASA V (moribund patient who is not expected to survive without surgery). Each category reflects increasing severity of systemic disease and associated surgical risk. Table 1.1 summarises the five categories of ASA-PS classification [8].

ASA	Description
I	Normal, healthy patient
II	Patient with mild systemic disease
III	Patient with severe systemic disease
IV	Patient with severe systemic disease that is a constant threat to life
V	Moribund patient who is not expected to survive without the operation.

Table 1.1: American Society of Anaesthesiologists Physical Status (ASA-PS) Classification

The ASA-PS classification is assigned by a clinician/anaesthesiologist in the pre-operative period. This is assessed subjectively using clinical text data and a physical assessment of the patient. The assigned classification serves as a general indicator for risk of perioperative morbidity and mortality and informs anaesthetic planning and patient counselling.

One of the key strengths of the ASA-PS system is its ability to capture latent associations between comorbidities and perioperative risk that may not be apparent when considering diseases in isolation. The ASA-PS classification represents a holistic assessment of a patient's overall health status, this allows the classification to reflect hidden patterns of vulnerability that can arise from seemingly mild conditions. For example, a history of eczema which may appear unrelated to anaesthetic risk has been shown to correlate with an increase likelihood of perioperative respiratory complications [6]. This association exists because eczema is often connected to asthma, an established risk factor for such complications. A clinician assigning an ASA score would implicitly integrate this kind of background knowledge, recognising that even mild diseases can reflect an increased susceptibility to adverse events under anaesthesia.

Despite its simplicity and widespread use, the ASA-PS classification system has limitations. It is inherently a subjective metric and does not fully account for the differences between adult and paediatric patients [10]. Furthermore, the classification does not capture the nuanced details that may be present in unstructured clinical notes.

The ASA-PS classification system can be evaluated using discrimination. Discrimination refers to the classification system's ability to distinguish between patients who do and do not experience perioperative adverse events. It is quantitatively assessed by plotting a receiver operating characteristic (ROC) curve; the area under this curve (AUC) indicates how well the method can differentiate between the two groups, with higher values indicating better discrimination [11].

The discrimination of the ASA-PS was evaluated on a dataset of 387 paediatric patients in a tertiary care centre in India, and displayed moderate discrimination with an AUC of 0.724 (95% CI 0.658–0.790) for predicting perioperative adverse events.

1.3 Literature Review

1.3.1 Large Language Models in Healthcare

There has been an increasing interest in the application of LLMs and natural language processing (NLP) for delivering improved healthcare outcomes using electronic health records (EHR) which store patient information. LLMs are very large deep learning models trained on vast amounts of text data to process and generate human-like language. These models learn to predict text and extract syntax, semantics, and knowledge from large corpora. For example, GPT models that are trained to continue text using auto-regressive training [12]. LLMs are built on a transformer architecture that consists of an encoder and decoder with self-attention capabilities to model the relationships between all words in a text sequence [13]. This architecture underpins both general-purpose LLMs (e.g. OpenAI's GPT-4, Meta's Llama) and domain-specific medical LLMs that have been specifically trained on biomedical text (e.g. Google MedGemma, ClinicalBERT). In healthcare, LLMs have garnered considerable interest due to their ability to process and interpret unstructured clinical data stored in the form of electronic health records (EHRs). EHRs contain free-text clinical notes, diagnosis reports and summaries that traditional computational methods would find difficult to interpret effectively [14].

The medical field has suffered from the challenge of finding fast and efficient access to understanding the rapidly-growing formation of clinical data. Text summarisation is essential for providing clinicians with access to relevant information without the need to sift through numerous medical records, literature, and other reports. Automatic summarisation has been long sought for to increase the productivity of this process, thereby reducing the workload and burnout of clinicians [15]. LLMs have the capability to be applied to this use case as they are able to handle long and complex sequences efficiently and produce concise summaries. For example, researchers have developed transformer models (primarily BERT variants) to automatically extract summaries from clinical notes such as radiology reports with accurate representations of the source data (measured using ROUGE scores which identifies the quality of summaries). These summaries can play a role in clinical decision-making by highlighting critical information such as recent test results and diagnoses in a more digestible form [15].

LLMs can perform classification on unstructured text to aid medical triaging and decision support. Some standard examples of text classification in healthcare include disease classification, drug classification, and sentiment analysis [15]. For instance, the research of Zhang et al. [16] demonstrates that BERT-based models can be used to automatically assign ICD (International Classification of Diseases) diagnosis codes to patients from EHRs, and this method significantly improved coding accuracy compared to traditional non-transformer based models. Similarly, a study by Yoon et al. [17] found that NLP models could classify unstructured free-text pre-anaesthesia evaluation summaries just as accurately or even exceeding the performance of board-certified anaesthesiologists. Automated classification and triage tools such as the tools in the studies discussed above can allow clinicians to prioritise and attend to the most urgent cases.

LLMs excel at answering questions based on the data that they are trained on, making them useful for clinical QA and decision support. This claim is supported by models such as Med-PaLM which have shown strong performance in medical exam questions, as discussed before [18]. LLM-powered QA can potentially be used for the development of chatbots to support direct patient interactions such as managing routine enquiries (e.g. scheduling appointments) and handling basic inquiries [19]. This can potentially be extended to clinician support for patients directly (with appropriate oversight and validation), for example, a chatbot to answer parents' or guardians' questions about postoperative care. The ability of LLMs to explain their reasoning behind a response in natural language could also be beneficial in clinical validation and testing [20].

While many applications of LLMs in healthcare are in research or pilot stages, specialised models are also emerging for subdomains such as paediatrics. A notable example is PediatricsGPT developed by Yang et al. [21], a LLM tailored to paediatric medicine. The model is trained on a specialised paediatric dataset (PedCorpus). The purpose of the model is to assist paediatricians in clinical decision support by providing specialised paediatric medical insights and evidence based recommendations. Despite such advances, significant challenges remain in ensuring ethical and clinical safety, and reliability, underscoring the need for further research into LLM applications in medical contexts.

1.3.2 Large Language Models for Perioperative Risk Stratification

Recent research has started exploring LLMs for predicting outcomes such as mortality, ICU admission, and length of stay, when provided with a patient's history and procedure description as input. A study by Chung et al. [20] prompted GPT-4 Turbo (OpenAI) with preoperative notes and procedural details from 1000 adult surgical cases. The research found that the LLM performed well on a number of perioperative classification tasks such as ASA-PS classification, ICU admission prediction, and hospital mortality prediction. When evaluated solely using F1 scores, the model's predictive accuracy is less accurate than traditional machine learning models. However, traditional learning models can rarely be used in clinical settings as it is incredibly difficult to interpret the model predictions [20]. In contrast, the LLM was able to generate natural language explanations for each prediction, a level of transparency that traditional machine learning models lack. However, these explanations must be validated by clinician oversight [22]. These early findings underscore the potential of LLMs to either complement or even outperform traditional tools for usage in adult perioperative settings.

Although Chung et al.'s study [20] was performed on adult populations, the underlying challenge of extracting clinically relevant information from unstructured free-text exists equally in paediatrics. The paediatric domain presents unique complexities that make direct generalisation from adult findings inappropriate. Children exhibit distinct physiological characteristics, development variations and disease profiles compared to adults, meaning that risk factors and language in clinical notes differ substantially. Currently, no published research has investigated the application of LLMs to paediatric perioperative risk stratification specifically. This gap in research is likely due to the limited availability of paediatric datasets that are both large-scale and high-quality, as many EHR systems lack full paediatric functionality [23]. Additionally, there are heightened ethical and privacy concerns when working with paediatric health records further complicating data collection. This represents an opportunity to explore whether LLMs, particularly those fine-tuned on clinical data can accurately perform perioperative risk stratification compared to general purpose LLMs.

1.4 Research Objectives

Perioperative risk stratification serves as a critical component of anaesthetic management, enabling clinicians to anticipate and mitigate adverse events during the surgical period. However, existing stratification tools for paediatric patients such as the ASA-PS classification system are limited by their subjective nature and poor generalisability across different age groups. Furthermore, current approaches primarily rely on structured data and variables including demographics (age, weight, sex), comorbidities, and procedure type, while a substantial portion of clinically relevant information such as anaesthetic events, airway management details and intraoperative observations is contained within unstructured text.

LLMs present an emerging opportunity to address these limitations as as they are capable of processing large amounts of unstructured text. This research explores the potential and feasibility of large language models for paediatric perioperative risk stratification using nursing and medical progress notes, as well as operation reports.

The specific objectives of this research are as follows:

- To evaluate the capability of LLMs to extract and interpret relevant clinical information from unstructured paediatric operation reports.
- To predict three key perioperative outcomes using LLMs:
  - **ASA Classification**: a categorical outcome indicating overall preoperative health status.
  - **Respiratory adverse events during emergence (AnyEmerg)**: binary outcome of whether a perioperative respiratory adverse event occured during emergence from anaesthesia.
  - **Respiratory risk factors (AnyRF)**: binary outcome capturing the presence of perioperative respiratory risk factors such as bronchospasm, laryngospasm, airway obstruction, or oxygen desaturation.
- To compare different prompt engineering strategies (zero-shot, few-shot, and chain-of-thought prompting) for LLM-based classification.
- To compare the performance of general-purpose (Llama) and domain-specific (MedGemma) LLMs across the above predictive tasks.

By achieving these objectives, this research aims to demonstrate the potential of LLMs as tools for automated and accurate perioperative risk stratification in paediatric anaesthesia.

1.5 Contribution

The core contribution of this research is the empirical evaluation and demonstration of LLMS for paediatric perioperative risk stratification by directly comparing a domain-specific LLM (MedGemma-27B-text-it) against a general-purpose

LLM (Llama-3.1-8B-Instruct) on multiple paediatric perioperative risk prediction tasks. In addition this study evaluates the relative effectiveness of different prompt engineering techniques particularly the few-shot Chain-of-Thought (CoT) prompting as a method for enhancing performance in medical reasoning tasks such as the ASA-PS classification.

1.6 Dissertation Structure

The dissertation has the following sections:

- **Chapter 1: Introduction** introduces the importance of paediatric perioperative risk stratification and includes a literature review on existing works and contributions that have applied LLMs for perioperative risk stratification tasks.
- **Chapter 2: Methodology** details the dataset and experimental design used in this research. It describes data collection and filtering, followed by an explanation of the prompting strategies used for inference, including zero-shot, few-shot, and chain-of-thought prompting.
- **Chapter 3: Results** showcases the performance outcomes across the models and different prompting strategies.
- **Chapter 4: Discussion** interprets the findings in the broader context of clinical relevance. It analyses patterns in model performance, explores explanations for observed behaviours, and discusses the limitations of the study. The chapter concludes with recommendations for future work.

CHAPTER 2

Methodology

2.1 Overview

This chapter outlines the methodology used to evaluate LLMs for paediatric perioperative risk stratification in this research. First, we describe the dataset and document selection criteria. Next, we introduce the LLMs used in this study, highlighting their architectures and rationale for selection. We then detail the different prompt engineering strategies applied (zero-shot, few-shot and CoT prompting) to elicit predictions from the models. Finally, we discuss the experimental setup, including the computing environment used for experimentation, and the evaluation metrics used to assess model performances across both categorical (ASA-PS) and binary (AnyEmerg, AnyRF) prediction tasks.

2.2 Dataset Description

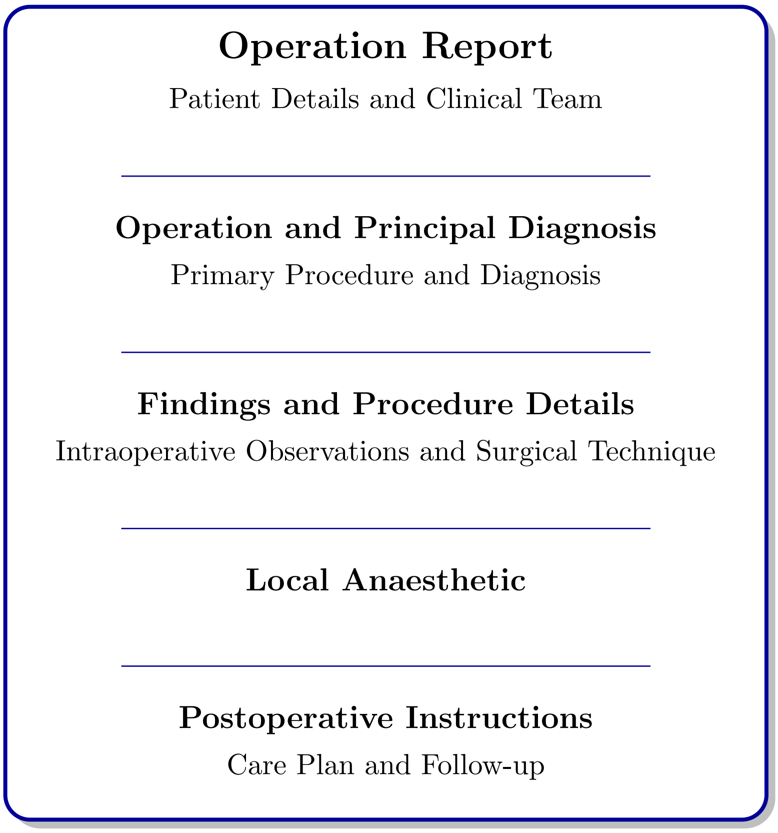
2.2.1 Data Source

The dataset used for this research is gathered from a previous study from Perth Children's Hospital. It contains data for paediatric patients (aged 0-16) undergoing surgery and anaesthesia. The dataset contains a range of documents such as operation reports, nursing and medical progress notes, admission forms and discharge summaries. All documents were originally in Portable Document Format (PDF) and were parsed using Marker [24], a transformer-enhanced parsing tool. In addition to these documents, a ground-truth dataset was also available containing structured patient information, including baseline patient characteristics (e.g. age, weight, sex) and clinical annotations such as ASA-PS classification and other binary/categorical perioperative outcomes (e.g. AnyEmerg and AnyRF). These structured annotations would serve as the reference standard when evaluating the LLM predictions.

2.2.2 Document Types and Selection Criteria

For this research, the specific document types were selected according to the target outcome. **Operation reports** were used for predicting post-operative outcomes, specifically **AnyEmerg** and **AnyRF**. The operation reports follow a structured format with headings, providing consistent contextual cues for the LLM's input. Pre-operative **progress notes** (nursing and medical), specifically the last progress note prior to surgery, were used for predicting **ASA-PS** classification. Only pre-operative notes were used for ASA-PS classification to avoid introducing bias from information recorded after the operation and as the ASA-PS classification is assigned prior to the procedure in the clinical domain. Compared to the operation reports, progress notes are highly variable and largely unstructured in nature.

An example structure of an operation report in this dataset:



Operation Report

Patient Details and Clinical Team

Operation and Principal Diagnosis

Primary Procedure and Diagnosis

Findings and Procedure Details

Intraoperative Observations and Surgical Technique

Local Anaesthetic

Postoperative Instructions

Care Plan and Follow-up

Figure 2.1: Standardised structure of operation reports in the dataset.

The following data matching and filtering procedure was applied specifically to extract the operation reports and their corresponding labels. While the patient administrative dataset included an episode identifier indicating the particular surgical episode (`episode_id`), the ground-truth dataset containing the baseline patient characteristics and perioperative outcome labels did not. Therefore, the document-to-label matching was performed using patient-level and temporal information rather than using episode IDs. To achieve this, date fields across the patient and ground-truth datasets were standardised into a consist datetime format to allow for accurate comparisons. The datasets were then joined using the Unique Medical Record Number (UMRN), only retaining patient records that appeared in both datasets. These records were filtered to include only those where the surgery date (`SurgeryDate` in the ground-truth dataset) occurred within one day of the corresponding document date, ensuring that operation reports would accurately match the relevant surgery episode. Finally, documents were filtered to only retain those labelled as "Operation Report TMS" for analysis. Although this matching and filtering method may not capture every operation report due to potential inconsistencies or typographical errors in the recorded surgery or document dates, it represents the most feasible and reliable approach given the absence of surgical episode identifiers in the ground-truth dataset. A similar approach was used to filter for the pre-operative progress notes, with modifications to ensure that only notes recorded prior to surgery were retained.

For pre-operative progress notes, the patient and ground-truth datasets were again merged using the UMRN as the unique identifier. The merged dataset was then filtered to retain only the documents where the file type (`file_type` in the patient dataset) is categorised as "Progress Note - Nursing" or "Progress Note - Medical", ensuring that both medical and nursing documentation were included in the analysis. To ensure that only pre-operative information was retained, progress notes with a recorded date after the corresponding surgery date were excluded. The remaining notes were stored chronologically by UMRN, surgery date, and note date, allowing identification of the most recent note prior to each surgery. For each surgical episode, only this latest pre-operative note was retained for analysis, representing the final assessment before anaesthesia.

2.3 Experiment Design

2.3.1 Large Language Models

This research utilises two LLMs to evaluate performance in paediatric perioperative risk stratification: Meta-LLama is the general-purpose model and Google's MedGemma is the domain-specific medical model. The specific variants used are Llama-3.1-8B-Instruct and MedGemma-27B-text-it. These models were chosen to compare the effectiveness of a medical domain-specialised LLM against a general-purpose LLM for extracting clinically relevant information from unstructured text and generating predictions for our chosen perioperative outcomes.

MedGemma

MedGemma is a large-scale medically-tuned language model which has been designed for various healthcare-related tasks, including question answering. It is based on Google's Gemma 3 model and developed under the Health AI Developer Foundations collection. The model utilises the same decoder-only transformer architecture as Gemma and is trained on extensive medical data including text from question-answering datasets (MedQA, MedMCQA, PubMedQA), medical imaging data across different modalities such as pathology, dermatology, and clinical documentation. MedGemma is available in multiple variants such as the 4B parameter multimodal model which is capable of processing both texts and images, and the 27B parameter models are available in text-only and multimodal configurations. We use the MedGemma-27B-text-it as the documents in our dataset have been parsed into text. The 27B-text-only model is instruction-tuned for medical text understanding and demonstrates strong performance on medical benchmarks, achieving 87.7% accuracy on MedQA and 74.2% on MedMCQA [25].

Meta-Llama

Meta-Llama developed by Meta AI, is a family of open-source LLMs, designed to provide state-of-the-art performance across a wide range of NLP tasks. The Llama 3.1 series, released in 2024, features models trained on over 15 trillion tokens of data with extended context windows up to 128,000 tokens. The specific variant used in this study is Llama-3.1-8B-Instruct, an 8 billion parameter instruction-tuned model optimised for conversational interactions and task-following. Similarly to MedGemma, it is built on a decoder-only transformer architecture with grouped query attention. Llama-3.1-8B Instruct is fine-tuned using supervised instruction datasets followed by reinforcement learning with human feedback (RLHF), enhancing its ability to follow instructions and generate accurate, safe, and helpful outputs [26]. Although Llama-3.1-8B-Instruct is not specifically trained on medical data, it demonstrates general-purpose capabilities that enable it to handle diverse tasks including question answering, reasoning, and text generation across various domains, making it a suitable baseline for comparison with MedGemma in this study. Table 2.1 provides a summary of our two chosen models.

Model	**MedGemma-27B-text-it**	**Llama-3.1-8B-Instruct**
Architecture	Decoder-only Transformer	Decoder-only Transformer
Parameters	27 billion	8 billion
Context Length	At least 128,000 tokens	128,000 tokens
Release Date	May 2025	July 2024

Table 2.1: Model summary for MedGemma-27B-text-it and Llama-3.1-8B-Instruct.

2.3.2 Computing Environment

All experiments for this study were conducted using Merlin, the high-performance computing (HPC) cluster at Perth Children's Hospital, which provides compute nodes equipped with H100 GPUs, along with sufficient RAM and storage to handle large-scale LLM inference. Due to the sensitive nature of paediatric patient data, all models were run locally on Merlin, without transferring data to external servers. Hugging Face libraries for both MedGemma-27B-text-it and Llama-3.1-8B-Instruct were used to perform model inference.

2.3.3 Prompt Engineering

Prompt engineering is the practice of carefully constructing and refining prompts or instructions that guide an LLMs output. The style of a users' prompt can significantly impact the response given by the LLM. Prompts that are more detailed and specific will allow for more accurate and relevant responses [27]. Due to the dynamic nature of LLMs, an iterative process for constructing prompts is required. It is rare to get the best response on a query after the first prompt attempt, therefore iterative prompting is required to refine prompts based on the LLMs' previous output to improve the accuracy or relevance of the response [27]. The prompting strategies we utilise in this study are zero-shot, few-shot, and chain-of-thought (CoT) prompting.

Zero-Shot Prompting

Large Language Models are trained on vast amounts of data. Due to this large-scale training they are able to perform tasks in a zero-shot manner. Zero-shot prompting is a method where a fixed prompt is used to derive responses from large language models by instructing the model directly without providing any examples or demonstrations related to the specific task [28]. Instruction tuned models have shown to improve zero-shot learning. Instruction tuning is a method where a large language model is fine-tuned on a collection of tasks and datasets that are described using natural language instructions [29]. In this research, we instruct the model directly using our zero-shot prompts, without providing any examples of clinical notes and their associated ground-truth metrics. This approach allows evaluation of the model's inherent ability to extract and classify relevant clinical information from unstructured paediatric documents.

Figure 2.2 illustrates the structure of the zero-shot prompt used for ASA-PS classification. In the prompt we explicitly define the role of the model, provide a definition of the ASA-PS classification system and include a set of instructions to guide the model's interpretation of the clinical note and response in a paediatric context. The input section, specifies the clinical progress note to be evaluated, and the output format enforces a structured JSON response, ensuring that predictions are machine-readable and include both the predicted classification along with a natural language explanation for the model's decision.

Zero-Shot Prompt Structure

Role: You are a clinical assistant tasked with assigning the ASA Physical Status Classification for a paediatric patient [...]
Classification System: ASA I through V definitions [...]
Instructions:

- Interpret in paediatric context
- Consider all clinical information
- Assign appropriate ASA classification (1–5)
- Provide justification
- Return structured JSON output

Input: clinical progress note
Output Format: {`"asa_classification": <int>, "explanation": "<str>"`}

Figure 2.2: Structure of the zero-shot prompt used for ASA classification

Few-Shot Prompting

In cases where zero-shot prompting fails, it is recommended to provide examples or demonstrations in the prompt to steer the model to better performance. Few-shot properties emerged when models were scaled to a sufficient size in terms of parameters [30]. This was based on the assumption that larger models (i.e. models with more parameters) would lead to better performance.

In our experiments, few-shot examples consist of a clinical note along with the associated label for the task being predicted (ASA-PS classification). From the 1,266 total cases (progress notes), we perform an 80/20 split: 80% of the cases are used for the inference set, and the remaining 20% are used as the source for the few-shot examples. Inverse frequency sampling is applied to ensure that all ASA categories present in the dataset are evenly represented in the few-shot dataset. For the ASA classification task, we provide five randomly sampled few-shot examples. Notably, ASA IV and V are substantially rarer compared to ASA I-III. ASA V cases are not represented in our filtered dataset because participation in the previous study that formed this dataset required informed parental consent, which is less likely to be obtained in emergency or critical surgery contexts where ASA V patients are encountered. The few-shot approach allows the model to reference representative examples when making predictions. Figure 2.3 provides an overview of the complete pipeline for generating the inference and few-shot datasets, and evaluating the model's responses for the ASA-PS classification task.

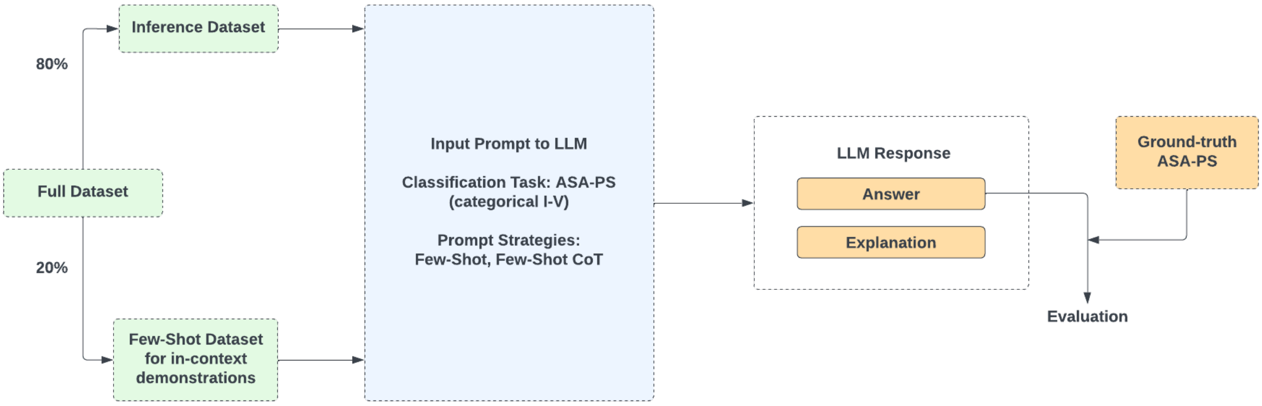

Inference Dataset
80%
Full Dataset
20%
Few-Shot Dataset for in-context demonstrations
Input Prompt to LLM
Classification Task: ASA-PS (categorical I-V)
Prompt Strategies: Few-Shot, Few-Shot CoT
LLM Response
Answer
Explanation
Ground-truth ASA-PS
Evaluation

Figure 2.3: Few-shot pipeline for ASA-PS classification task.

Chain-of-Thought Prompting

Standard few-shot prompting may work well for an abundance of tasks, however it may still fall short on tasks that require more complex reasoning. Chain-of-thought (CoT) reasoning involves including a series of intermediate reasoning steps within the prompt to significantly improve the capability of large language models to perform more complex reasoning tasks. CoT can be combined with other prompt engineering techniques such as zero-shot to form the idea of zero-shot CoT [28] which essentially involves adding the sentence "Let's think step by step" to the original prompt. For our experiments, we combine CoT reasoning with both the base zero-shot prompt and few-shot prompt, by adding intermediate reasoning steps to guide the model's inference process.

2.4 Evaluation Metrics

Model performance was evaluated using several metrics chosen to capture both overall accuracy and more nuanced aspects of the classification behaviour. The selected metrics reflect the diversity of the tasks examined in this study, which include both multi-class (ASA-PS classification) and binary (AnyEmerg, AnyRF) prediction problems.

Accuracy was used as a general indicator of how well the model's predictions matched the ground-truth labels. While this metric provides a simple measure of performance, it can be misleading for imbalanced datasets, where a model can achieve high accuracy by simply predicting the majority class. This limitation is particularly relevant to the AnyEmerg prediction task, where the majority of patients did not experience a PRAE during emergence from anaesthesia.

To address class imbalance and assess performance more equitably across all categories, macro-averaged F1 scores were computed. The macro F1 score treats each class equally, regardless of its prevalence, making it especially informative for the ASA-PS classification task where higher ASA classes (IV-V) were under-represented or not present at all in our dataset, as was the case for ASA V.

For the ASA-PS classification, an additional metric, off-by-one accuracy was included to account for the subjective nature of the task. This metric considers predictions within one level of the true ASA category as acceptable, acknowledging that minor deviations (for example, predicting ASA II instead of ASA III) may still represent clinically reasonable assessments. This is important given that even among clinicians there is variability in classification.

For the binary prediction tasks (AnyEmerg and AnyRF), precision and recall were also examined to better understand model behaviour with respect to detecting adverse events. Precision measures the proportion of predicted positive cases that were correct, while recall reflects the proportion of actual positive cases the model successfully identified. Precision and recall were chosen, as they are particularly useful when dealing with class imbalances.

Together, these metrics provide a balanced evaluation framework, allowing for a detailed interpretation of model strengths and weaknesses across tasks of varying difficulty and class distribution.

CHAPTER 3

Results

3.1 Overview

This chapter presents the evaluation results of the two LLMs (MedGemma-27B-text-it and Llama-3.1-8B-Instruct across three key perioperative risk prediction tasks: ASA-PS classification, AnyEmerg, and AnyRF. Performance for the ASA-PS classification was compared under different prompting strategies including zero-shot, few-shot, zero-shot CoT, and few-shot CoT, to explore how structured reasoning and contextual examples influenced performance. The AnyEmerg and AnyRF prediction tasks were evaluated in a zero-shot setting, to test each models' inherent ability to generalise and identify perioperative risk factors. Overall, MedGemma achieved higher performance on the ASA-PS classification task, whereas Llama performed marginally better on the more general AnyRF prediction task. Both models struggled to identify rare events for the AnyEmerg prediction task.

3.2 ASA-PS Classification

Table 3.1: Performance of MedGemma-27B-text-it on ASA classification task using different prompting strategies.

Prompt Type	Accuracy	F1 (Macro)	Off-by-One Accuracy
Zero-shot	0.368	0.244	0.761
Zero-shot CoT	0.424	0.317	0.891
Few-shot	0.387	0.280	0.821
Few-shot CoT	**0.426**	**0.318**	**0.894**

Tables 3.1 and 3.2 present the performance of MedGemma-27B-text-it and LLama-3.1-8B-Instruct on the ASA-PS classification using different prompting

Table 3.2: Performance of Llama-3.1-8B-Instruct on ASA classification task using different prompting strategies.

Prompt Type	**Accuracy**	**F1 (Macro)**	**Off-by-One Accuracy**
Zero-shot	0.338	0.148	0.717
Zero-shot CoT	**0.339**	**0.241**	0.742
Few-shot	0.319	0.213	0.746
Few-shot CoT	0.335	0.193	**0.764**

strategies (zero-shot, zero-shot CoT, few-shot, few-shot CoT). MedGemma outperformed Llama across all metrics. For MedGemma, using the few-shot CoT prompting technique, achieved the highest overall performance with an accuracy of 0.426 and off-by-one accuracy of 0.894, indicating that CoT reasoning along with few-shot examples of ASA-PS classification improved performance.

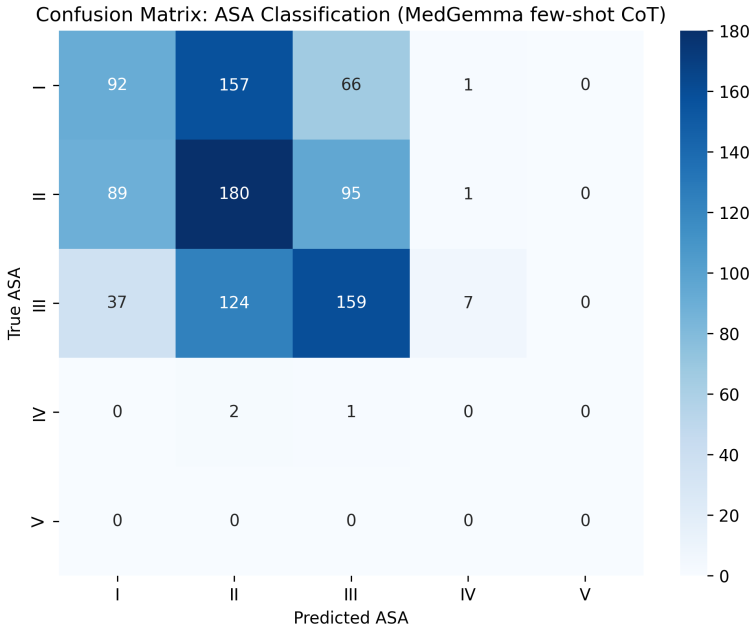


Figure 3.1: Confusion matrix for MedGemma-27B-text-it ASA classification using few-shot CoT prompting.

Figure 3.1 shows the confusion matrix for the few-shot CoT ASA classification. The model demonstrates a tendency to predict intermediate ASA classes (II and III), with the highest concentration of correct predictions in class II (180 correct predictions). A qualitative review of the severely misclassified cases suggests that the model often underestimates or overestimates disease severity and therefore

ASA-PS classification based on how the language is framed in the notes rather than its clinical implications.

For example, in one case, a patient with a severe systemic disease posing a high perioperative risk (ASA IV) was classified as ASA I. The model's explanation focused on the patient's stable observations and absence of symptoms in the clinical note, suggesting over-reliance on surface-level stability indicators rather than recognising the underlying critical condition. This misclassification may also reflect that the progress note did not mention the underlying disease explicitly, as it could have been documented earlier. Conversely, another patient with a chronic but well-controlled condition (ASA II) was predicted as ASA IV, likely because the model inferred excessive severity from the presence of specialised medical equipment.

3.3 AnyEmerg Prediction

Table 3.3: Zero-shot performance of MedGemma and Llama on AnyEmerg prediction.

Model	**Accuracy**	**F1 (Macro)**	**Precision (Yes)**	**Recall (Yes)**
MedGemma-27B-text-it	0.588	0.390	0.022	0.437
Llama-3.1-8B-Instruct	0.546	0.370	0.020	0.444

Table 3.3 presents the zero-shot performance of MedGemma and Llama on predicting the post-operative metric: AnyEmerg. The distribution of this metric in our filtered dataset is heavily imbalanced, with the majority of cases not experiencing PRAEs during emergence from anaesthesia. As a result, the accuracy is largely influenced by the dominant "No" class. While both models achieve moderate F1-macro scores, precision for the minority ("Yes") class is very low, which reflects the challenge of identifying rare events. Recall for the positive minority class is much higher compared to precision, indicating that the models are able to identify a portion of the true positive events despite the class imbalance.

3.4 AnyRF Prediction

Table 3.4 presents the models performance on AnyRF prediction, with Llama slightly outperforming MedGemma. The distribution of AnyRF in the dataset

Table 3.4: Zero-shot performance of MedGemma and Llama on AnyRF prediction.

Model	**Accuracy**	**F1 (Macro)**
MedGemma-27B-text-it	0.480	0.472
Llama-3.1-8B-Instruct	0.503	0.483

is fairly balanced compared to AnyEmerg. The relatively balanced distribution of positive and negative cases results in F1-macro scores being closer to overall accuracy. These results indicate that while both models are capable of detecting risk factors for PRAEs from operation reports, zero-shot prediction is limited in sensitivity to subtle patterns, especially when rare events are encountered.

CHAPTER 4

Discussion

This research evaluated the performance of two LLMs, MedGemma-27B-text-it and Llama-3.1-8B-Instruct, on three key perioperative prediction tasks: ASA-PS classification, prediction of any airway or respiratory events during emergence from anaesthesia (AnyEmerg), and identification of intraoperative or postoperative respiratory events (AnyRF). The primary objective was to assess the capability of pre-trained LLMs to accurately predict perioperative risk metrics from unstructured text in clinical notes. Multiple prompting strategies were tested including zero-shot, zero-shot CoT, few-shot and few-shot CoT, to evaluate whether instruction-tuning, explicit reasoning, and example demonstrations could enhance the zero-shot learning capabilities of each model.

4.1 ASA-PS Classification

ASA-PS classification presented a particularly challenging task due to the subjective nature of the classification system, along with the limited representation of rarer classes (ASA IV-V). It provided the most insight into the models' ability to perform nuanced clinical reasoning. MedGemma consistently outperformed Llama across all the prompting strategies, demonstrating the potential benefit of domain-specific clinical pretraining to support its understanding of medical terminology and patient context, although the exact composition of the pretraining corpora for both models, especially with respect to paediatric clinical language and rarer ASA classes, is not publicly known.

The few-shot CoT prompting strategy presented the best performance for MedGemma, achieving an F1 score of 0.318 and 0.894 off-by-one accuracy. These results suggest that structured reasoning and exposure to labelled example demonstrations improved the model's interpretation. Across all the ASA-PS experiments in MedGemma and Llama, the off-by-one accuracy was high, ranging from 0.717 to 0.894, indicating that the majority of classifications occurred between adjacent classes (e.g. ASA II and ASA III).

Analysis of class-level errors revealed that ASA II and ASA III were the most frequently confused categories. This is expected, as the distinction between "mild" and "severe" systemic disease in paediatric patients can be ambiguous and context-dependent. Rare classes such as ASA IV and V were heavily underrepresented in the dataset, and when present, were commonly misclassified as ASA III. The models' pretraining exposure to high-severity paediatric cases is largely unknown and unlikely, so it is unclear whether limited representation of these cases in pretraining contributed to misclassification.

While both models demonstrated reasonable interpretability through their generated explanations, qualitative review of misclassified cases revealed that some errors arose from overly literal interpretations of clinical text, as demonstrated by the two examples in Chapter 3: Results.

Overall, ASA-PS classification remains a difficult prediction task due to its subjective nature and dependence on clinical judgement. However, the consistent improvements seen with reasoning-based prompting suggest that LLMs can approximate clinical reasoning patterns when appropriately guided, even under conditions of limited data and uncertain pretraining exposure.

4.2 AnyEmerg and AnyRF

The performance of both models on AnyEmerg and AnyRF prediction tasks illustrates the limitations of zero-shot classification in the presence of imbalanced outcome distributions. For AnyEmerg, the "Yes" class (indicating the presence of a perioperative adverse event during emergence) was extremely underrepresented. Both MedGemma and Llama achieved moderate overall accuracy (0.55–0.59), but this was largely driven by the dominant "No" class. Precision for the minority "Yes" class remained very low (less than 0.03), while recall was substantially higher (0.44), indicating that although the models were capable of identifying some true positive cases, they also produced many false positives.

Since the models were operating in a zero-shot setting, without any fine-tuning or retrieval-augmented generation (RAG), their ability to recognise rare clinical patterns would be limited. The underlying pretraining corpus for the model are largely unknown, and it is unlikely that it contained a substantial number of paediatric cases or examples of rare perioperative adverse events, therefore further reducing the models' ability to identify minority outcomes reliably.

In contrast, the AnyRF task showed slightly improved and more balanced performance, with both models achieving accuracies around 0.48–0.50 and F1-

macro scores near 0.47–0.48. The reduced class imbalance (approximately 3:1 ratio) contributed to this improvement. Interestingly, Llama marginally outperformed MedGemma on this task, possibly due to its larger general reasoning capacity and the more balanced dataset reducing the benefits of MedGemma's domain specific pretraining for this task.

4.3 Limitations

Several factors limited the model performance and the scope of our experiments. Firstly, the severe class imbalance for AnyEmerg (rare positive cases) constrained the reliability of statistical measures and evaluation derived from the dataset. For example, standard metrics such as p-values could not be derived and would be less meaningful due to the minority class being extremely underrepresented. This imbalance does not necessarily indicate that the models themselves could not recognise rare events, but instead reflects the limited representation of such cases in the dataset, and the unknown extent to which these sorts of scenarios are encountered during pretraining of these models.

Secondly, for our ASA-classification study, we only utilised the last pre-operative progress note prior to surgery. The quality and completeness of these clinical notes may vary substantially across patients. The last pre-operative note may not necessarily contain all the information required to make an accurate ASA-classification. A more suitable approach, would be to utilise multiple progress notes or use an LLM to create a summary of the patient's well-being through these notes. However, this approach comes with its own limitations of being more computationally expensive, time consuming, and also risks information loss during natural language summarisation.

4.3.1 Clinical Validity

Another limitation of this study is the inability to verify the clinical validity of the natural language explanations or reasoning provided by the model's output. Unlike the perioperative metrics that can be validated against binary or categorical ground-truths, the explanations can not be systematically validated without expert review. The models may produce plausible-sounding reasoning that are factually incorrect, rely on spurious correlations from the clinical notes, or entirely miss clinically important factors. Clinicians need to not only understand the what the model predicts, but also why it makes those predictions and whether the underlying reasoning is medically sound. Without extensive clinical validation of model explanations by domain experts, there is the risk that models could arrive at correct predictions for incorrect reasons, or conversely, provide convincing but flawed reasoning that misguides clinician judgement.

4.4 Future Work

Future work for this study should focus on expanding data diversity and scope to address the class imbalance observed in this research, particularly in the AnyEmerg distribution. Increasing the representation of rare perioperative cases and fine-tuning models on these types of cases, will enable models to learn more robust risk patterns and improve generalisability across different patient subgroups.

This research primarily focused on categorical and binary perioperative outcomes for paediatric risk stratification. Future work can extend to the prediction of continuous or numerical outcomes, for example, estimating the expected duration the patient will stay in the hospital after surgery, or the probability of admission to the paediatric intensive care unit (PICU). Such quantitative predictions would provide clinicians with more insights into individual patient trajectories.

Furthermore, this research evaluated perioperative risk prediction on a case-by-case basis. Integrating longitudinal patient data, such as information spanning multiple surgical episodes, could enable LLMs to model trends in patient health over time. This would support more individualised and comprehensive risk assessment.

As the field of medical AI rapidly involves, larger and more domain-adapted models are likely to emerge. Evaluating these models on paediatric datasets and for perioperative risk stratification are essential to determine whether improvements in scale and fine-tuning translate into clinically meaningful performance benefits. In particular, fine-tuning medical LLMs such as MedGemma on paediatric-specific corpora could substantially enhance their understanding of paediatric language and paediatric specific risk factors, which are areas that are likely to be underrepresented in current pretraining datasets.

Integration of LLMs into clinical practice will require technical training for medical staff for model interaction and engineering for effective prompts. Equipping clinicians with the required skills to craft precise and contextually appropriate prompts for paediatric risk stratification tasks, can improve accuracy and reliability of model-assisted decision-making. This form of literacy will be crucial in ensuring safe and meaningful collaboration with LLMs in paediatric perioperative settings.

Finally, for LLMs to be adopted safely in paediatric perioperative risk stratification, their deployment must occur within a clinician-in-the-loop framework. In this system, the model would solely serve as a decision-support tool rather than an autonomous agent. The clinician would be required to review, validate, and, when necessary, override model outputs, ensuring that all decisions remain under domain and expert clinical judgement. Moreover, these systems should be designed to prevent an implicit or institutional pressure to defer to AI-generated advice, maintaining the clinician's authority and accountability.

Bibliography

[1] D. Van Veen *et al.*, "Adapted large language models can outperform medical experts in clinical text summarization," *Nature medicine*, vol. 30, no. 4, pp. 1134–1142, 2024.

[2] S. Shool, S. Adimi, R. Saboori Amleshi, E. Bitaraf, R. Golpira, and M. Tara, "A systematic review of large language model (llm) evaluations in clinical medicine," *BMC Medical Informatics and Decision Making*, vol. 25, no. 1, p. 117, 2025.

[3] I. Adeleke, C. Chae, O. Okocha, and B. Sweitzer, "Risk assessment and risk stratification for perioperative complications and mitigation: Where should the focus be? how are we doing?" *Best Practice & Research Clinical Anaesthesiology*, vol. 35, no. 4, pp. 517–529, 2021.

[4] C. Brown and A. Allana, *The Beginner's Guide to Anaesthetics: A Handbook for Doctors in Training and Allied Professionals.* CRC Press, 2025.

[5] S. Mehrotra, "Postoperative anaesthetic concerns in children: Postoperative pain, emergence delirium and postoperative nausea and vomiting," *Indian journal of anaesthesia*, vol. 63, no. 9, pp. 763–770, 2019.

[6] B. S. von Ungern-Sternberg *et al.*, "Risk assessment for respiratory complications in paediatric anaesthesia: a prospective cohort study," *The Lancet*, vol. 376, no. 9743, pp. 773–783, 2010.

[7] L. G. Maxwell and M. Yaster, "Perioperative management issues in pediatric patients," *Anesthesiology Clinics of North America*, vol. 18, no. 3, pp. 601–632, 2000.

[8] American Society of Anesthesiologists, "Statement on asa physical status classification system," https://www.asahq.org/standards-and-practice-parameters/statement-on-asa-physical-status-classification-system, [Accessed: Oct. 3, 2025].

[9] A. R. Tait, T. Voepel-Lewis, R. Christensen, and L. M. O'Brien, "The stbur questionnaire for predicting perioperative respiratory adverse events in children at risk for sleep-disordered breathing," *Pediatric Anesthesia*, vol. 23, no. 6, pp. 510–516, 2013.

[10] L. R. Ferrari *et al.*, "One size does not fit all: a perspective on the american society of anesthesiologists physical status classification for pediatric patients," *Anesthesia & Analgesia*, vol. 130, no. 6, pp. 1685–1692, 2020.

[11] A. N. Udupa, M. N. Ravindra, Y. Chandrika, K. Chandrakala, N. Bindu, and M. F. Watcha, "Comparison of pediatric perioperative risk assessment by asa physical status and by narco-ss (neurological, airway, respiratory, cardiovascular, other–surgical severity) scores," *Pediatric Anesthesia*, vol. 25, no. 3, pp. 309–316, 2015.

[12] T. Brown *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[13] A. Vaswani *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[14] X. Yang *et al.*, "A large language model for electronic health records," *NPJ digital medicine*, vol. 5, no. 1, p. 194, 2022.

[15] H. N. Cho *et al.*, "Task-specific transformer-based language models in health care: Scoping review," *JMIR Medical Informatics*, vol. 12, p. e49724, 2024.

[16] Z. Zhang, J. Liu, and N. Razavian, "Bert-xml: Large scale automated icd coding using bert pretraining," in *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, 2020, pp. 24–34.

[17] S. B. Yoon, J. Lee, H.-C. Lee, C.-W. Jung, and H. Lee, "Comparison of nlp machine learning models with human physicians for asa physical status classification," *NPJ digital medicine*, vol. 7, no. 1, p. 259, 2024.

[18] J. Clusmann *et al.*, "The future landscape of large language models in medicine," *Communications medicine*, vol. 3, no. 1, p. 141, 2023.

[19] L. Moura *et al.*, "Implications of large language models for quality and efficiency of neurologic care: emerging issues in neurology," *Neurology*, vol. 102, no. 11, p. e209497, 2024.

[20] P. Chung, C. T. Fong, A. M. Walters, N. Aghaeepour, M. Yetisgen, and V. N. O'Reilly-Shah, "Large language model capabilities in perioperative risk prediction and prognostication," *JAMA surgery*, vol. 159, no. 8, pp. 928–937, 2024.

[21] D. Yang *et al.*, "Pediatricsgpt: Large language models as chinese medical assistants for pediatric applications," *Advances in Neural Information Processing Systems*, vol. 37, pp. 138 632–138 662, 2024.

[22] R. Yang, T. F. Tan, W. Lu, A. J. Thirunavukarasu, D. S. W. Ting, and N. Liu, “Large language models in health care: Development, applications, and challenges,” *Health Care Science*, vol. 2, no. 4, pp. 255–263, 2023.

[23] K. R. Dufendach, C. U. Lehmann, and S. A. Spooner, “Special requirements of electronic health record systems in pediatrics: Clinical report,” *Pediatrics*, vol. 154, no. 4, p. e2024068509, 2024.

[24] Datalab, “Marker: Document conversion and parsing framework,” https://github.com/datalab-to/marker, 2024, [Accessed: Oct. 8, 2025].

[25] A. Sellergren *et al.*, “Medgemma technical report,” *arXiv preprint arXiv:2507.05201*, 2025.

[26] A. Dubey *et al.*, “The llama 3 herd of models,” *arXiv e-prints*, pp. arXiv–2407, 2024.

[27] B. Meskó, “Prompt engineering as an important emerging skill for medical professionals: tutorial,” *Journal of medical Internet research*, vol. 25, p. e50638, 2023.

[28] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, “Large language models are zero-shot reasoners,” *Advances in neural information processing systems*, vol. 35, pp. 22 199–22 213, 2022.

[29] J. Wei *et al.*, “Finetuned language models are zero-shot learners,” in *International Conference on Learning Representations*, 2021.

[30] H. Touvron *et al.*, “Llama: Open and efficient foundation language models,” *arXiv preprint arXiv:2302.13971*, 2023.

[31] P. Zhang and M. N. Kamel Boulos, “Generative ai in medicine and healthcare: promises, opportunities and challenges,” *Future Internet*, vol. 15, no. 9, p. 286, 2023.

[32] J. Huang *et al.*, “A critical assessment of using chatgpt for extracting structured data from clinical notes,” *npj Digital Medicine*, vol. 7, no. 1, p. 106, 2024.

APPENDIX A

Original Honours Proposal

Supervisors: Tim French, Wei Liu and Harry Smallbone (WA Health)

Degree: Bachelor of Advanced Computer Science (Honours)

Date: May 2025

A.1 Background

Children undergoing surgery are at risk of complications that impact their behavioural and physiological systems. These complications are known as perioperative adverse events, as they occur around the surgical period. Some common examples of perioperative adverse events in paediatric patients are respiratory complications, postoperative nausea and vomiting (PONV), and emergence delirium (ED) [5]. Perioperative risk stratification is the process of estimating a patient's likelihood of experiencing specific adverse events based on preoperative data. Traditional methods for paediatric perioperative risk stratification such as the STBUR (Snoring, Trouble Breathing, Un-Refreshed) questionnaire [9] are dependent on a fixed set of preoperative variables (symptoms).

The recent development of Large Language Models such as OpenAI's GPT model enable contextual understanding, reasoning, and the generation of human-like text. Through this it appears that artificial intelligence (AI) chatbots could assist medical professionals with clinical decision-making [31].

Research has recently been conducted on the use of LLMs in perioperative risk stratification. The findings from this research have concluded that models such as GPT-4 are able to identify risk factors and predict postoperative complications such as ICU admission and hospital mortality, directly from preoperative notes written by clinicians [20]. The majority of these studies has been focused on adult patient populations and very limited research has been conducted on the paediatric perioperative setting.

There are limitations to the application of LLMs in the clinical perioperative setting. Among the limitations of LLMs in perioperative care, hallucinations are a significant concern. This is where LLMs generate factually incorrect or fabricated information [32]. Incorrect information has the potential to mislead clinicians and compromise decision-making. Therefore limitations such as hallucination underscore the necessity for strict prompt engineering and fine-tuning to optimise these models, along with clinician-in-the-loop validation to ensure AI outputs align with real-world clinical practice.

A.2 Problem Statement

This research aims to investigate the capability of LLMs in extracting and predicting paediatric perioperative adverse events from unstructured clinical notes. The study will also seek to determine their precision and feasibility for real-life clinical practice while considering difficulties including model bias and hallucination.

A.3 Aim

1. Evaluate the capability of LLMs to extract and predict paediatric perioperative adverse events from unstructured clinical notes using prompt engineering strategies.

2. Attempt to improve model performance using fine-tuning techniques, including domain-specific training on clinical data.

3. Investigate the use of retrieval augmented generation (RAG) to improve the accuracy and factual grounding of LLM outputs when predicting paediatric perioperative adverse events from complex unstructured clinical notes.

A.4 Methodology

A.4.1 Data Preprocessing and Prompt-Based Evaluation

The research will begin by evaluating the capability of LLMs to extract and predict paediatric perioperative adverse events from unstructured clinical notes using prompt engineering strategies. The clinical notes provided from previous studies at Perth Children's Hospital (PCH) will be preprocessed for anonymisation of patients, remove noise and inconsistencies, and segment longer documents to fit into model input limitations. Ground truth labels for certain adverse events (e.g. postoperative nausea and vomiting) are already available from previous clinical studies conducted at PCH. These can be used for evaluation of model predictions. A range of different prompting approaches will be designed and tested, including zero-shot, few-shot, and chain-of-thought (CoT) prompting.

A.4.2 Fine-Tuning and Optimisation

The selected LLMs will be fine-tuned on the preprocessed clinical dataset to improve predictive accuracy for paediatric perioperative adverse events and reduce the rate of hallucinations. The fine-tuning will be conducted on the high performance computing infrastructure of MERLIN at PCH. Hyperparameter tuning can be employed to optimise training performance and improve the model's ability to process clinical text.

A.4.3 Retrieval Augmented Generation (RAG)

RAG will be used to improve factual grounding and the reliability of LLM outputs, especially when processing longer and more complex unstructured clinical notes. A retrieval source will need to be constructed using clinical guidelines and the preprocessed clinical dataset. This knowledge base will allow the model to reference contextually relevant information during inference and reduce hallucinations in its output.

A.5 Resources

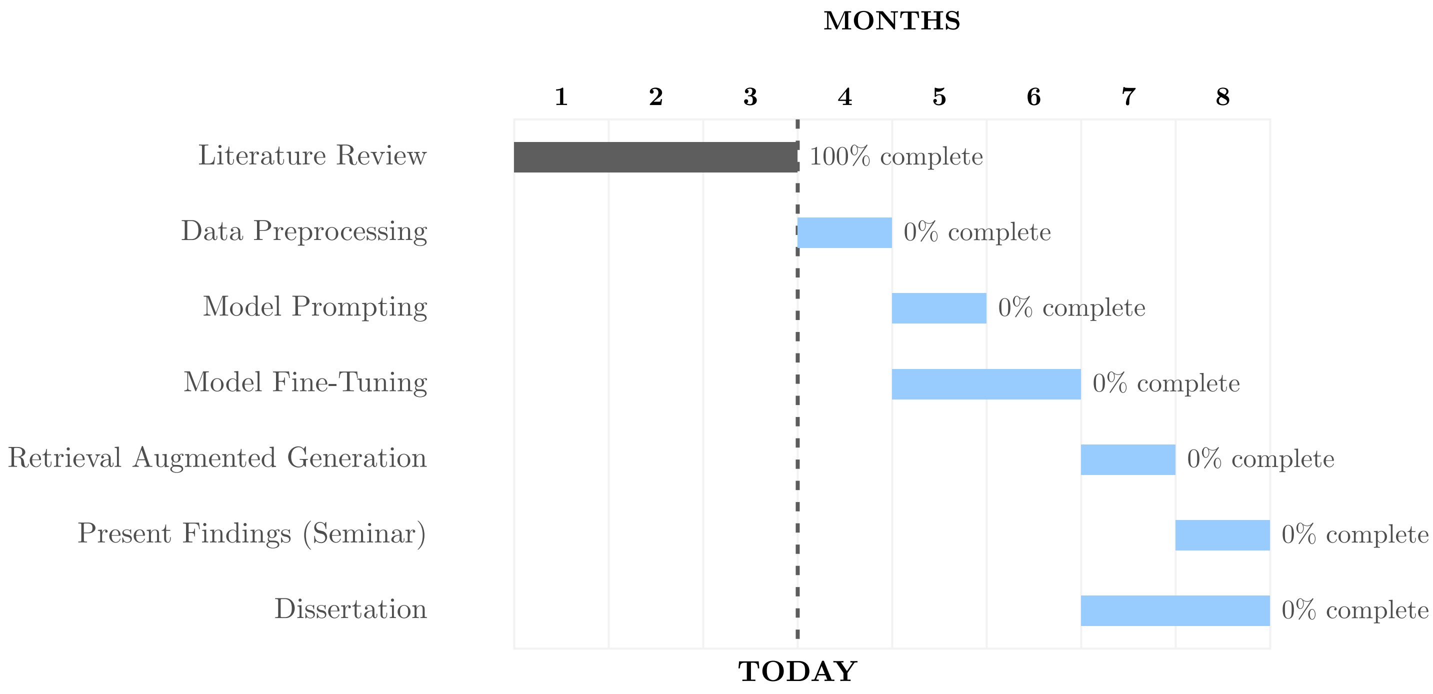
1. **Computational Resources:** This project will use MERLIN, a high-performance computing system at PCH, for training and analysing large

language models. If additional computing power is required, access to the Pawsey Supercomputing Centre is available, specifically Setonix.

2. **Data Resources:** Access to 8000 multimodal patient records from previous studies at PCH. These records will serve as input for model training and evaluation. Access to these clinical notes will be subject to the PCH and WA Health regulations, ensuring compliance with data governance policies, ethical guidelines, and patient confidentiality. The necessary permissions and ethics approvals will be acquired prior to data handling.

3. **Software and Tools:** We will be utilising Hugging Face Transformers for LLM fine-tuning along with PyTorch for model training. Pandas and NumPy can be used to handle structured and unstructured data, and LangChain for retrieval augmented generation.

A.6 Timeline



APPENDIX B

Additional Materials

Zero-Shot Prompt Structure for AnyEmerg

Task: Assess whether a paediatric case is likely to be associated with a perioperative respiratory adverse event (PRAE) during emergence from anaesthesia [...]
Context: Emergence refers to the period when anaesthetic gases or IV medications are stopped and the child begins to wake up.
Risk Indicators to Consider:

- Procedure details (airway/ENT procedures, thoracic surgery)
- Airway management (endotracheal tube, laryngeal mask, difficult intubation)
- Respiratory risk factors (asthma, reactive airway disease, URI, OSA)
- Medications (opioids, muscle relaxants, reversal agents)
- Patient factors (prematurity, comorbidities, airway anomalies)
- Postoperative monitoring (PICU transfer, prolonged observation, supplemental oxygen)

Input: Operation report
Output Format: {`"emergence_event_prediction": "<YES or NO>", "explanation": "<str>"`}

Figure B.2: Structure of the zero-shot prompt used for emergence PRAE prediction

Chain-of-Thought Prompt Structure

Role: You are a clinical assistant tasked with assigning the ASA Physical Status Classification for a paediatric patient [...]
Classification System: ASA I through V definitions [...]
Instructions:

1. Interpret in paediatric context considering children's health and development
2. Consider all clinical information (medical history, comorbidities, medications, vital signs, development/prematurity, prior admissions, device use, planned procedure)
3. **Step-by-step reasoning process:**
   - Extract explicit findings (diagnoses, medications, vitals, developmental notes, acute problems)
   - Use proxy indicators if explicit findings are sparse (regular medications, abnormal vitals, prior admissions, device use, procedure complexity, age-related vulnerability)
   - Assess severity and control of each identified condition
   - Integrate clinical severity and procedure-related risk
   - Do NOT invent diagnoses
4. Assign appropriate ASA classification (1–5)
5. Provide justification with key clinical details
6. Return structured JSON output

Input: clinical progress note
Output Format: `{"asa_classification": <int>, "explanation": "<str>"}`

Figure B.1: Structure of the chain-of-thought prompt used for ASA classification.

Zero-Shot Prompt Structure for PRAE Risk Factor Detection

Task: Determine whether any perioperative respiratory adverse event (PRAE) risk factors are present based on the operation report.

Valid PRAE Risk Factors:

- Risk factors for bronchospasm (asthma, reactive airway disease, recent respiratory infection)
- Risk factors for laryngospasm (airway irritation, light anaesthesia, airway procedures)
- Risk factors for airway obstruction (obstructive sleep apnoea, tonsillar hypertrophy, airway anomalies)
- Risk factors for oxygen desaturation (respiratory conditions, prematurity, significant comorbidities)

Decision Guidelines:

- Predict **YES** if ANY PRAE risk factor identified or inferred
- Predict **NO** only if routine, low-risk case with no identifiable respiratory risk factors

Input: Operation report

Output Format: `{"any_rf_prediction": "<YES or NO>", "explanation": "<str>"}`

Figure B.3: Structure of the zero-shot prompt used for PRAE risk factor detection