

## ABSTRACT

Title of Thesis: Deployment of Large Vision and Language Models for Real-Time Robotic Triage in a Mass Casualty Incident

Alexandra Mangel  
Master of Science, 2024

Thesis Directed by: Professor Derek Paley  
Department of Aerospace Engineering

In the event of a mass casualty incident, such as a natural disaster or war zone, having a system of triage in place that is efficient and accurate is critical for life-saving intervention, but medical personnel and resources are often strained and struggle to provide immediate care to those in need. This thesis proposes a system of autonomous air and ground vehicles equipped with stand-off sensing equipment designed to detect and localize casualties and assess them for critical injury patterns. The goal is to assist emergency medical technicians in identifying those in need of primary care by using generative AI models to analyze casualty images and communicate with the victims. Large language models are explored for the purpose of developing a chatbot that can ask a casualty where they are experiencing pain and make an informed assessment about injury classifications, and a vision language model is prompt engineered to assess a casualty image to produce a report on designated injury classifiers.

Deployment of Large Vision and Language Models for  
Real-Time Robotic Triage in a Mass Casualty Incident

by

Alexandra Mangel

Thesis submitted to the Faculty of the Graduate School of the  
University of Maryland, College Park in partial fulfillment  
of the requirements for the degree of  
Master of Science  
2024

Advisory Committee:

Professor Derek Paley, Chair/Advisor  
Professor Michael Otte  
Professor John Martin

© Copyright by  
Alexandra Mangel  
2024

## Acknowledgments

I would like to begin by thanking the Defense Advanced Research Projects Agency for sponsoring these efforts. This research was supported by Grant No. HR00112420304 as a part of the DARPA Triage Challenge.

The success of RoboScout DTC is entirely owed to the combined efforts of our amazing team at the University of Maryland, College Park, namely Zach Bortoff, Srijal Poojari, and Atharv Marathe, without whom my research would not be possible, as well as contributions from many Collective Dynamics and Control Laboratory members including Tanmay Pancholi, Yohanana Kalya, Alexander Teacu, JJ Kuznetsov, Yashas Shetty, and Dr. Kleio Baxevani, as well as co-PI Dr. Dinesh Manocha and his students Ruiqi Xian and Soumya Ghosal.

I would also like to extend gratitude to the University of Maryland UAS Research and Operations Center, particularly Josh Gaus, Grant Williams, Chris Titus, Mckenzie Turpin, and Darren Robey, for their drone expertise and hard work as the UAV side of the RoboScout operation.

I am incredibly grateful for our partnership with the University of Maryland Medical Center in Baltimore, especially the members of the Center for the Sustainment of Trauma and Readiness Skills, Tony Christiani, Dan Goodman, Vondel Mahon, Bruce Lynch, and Dr. Sarah Murthi.

A special thanks to my committee members, including my professors Dr. Michael Otte and Dr. John Martin, and of course my advisor Dr. Derek Paley who has served as a fantastic mentor during my time at the University of Maryland.

Lastly, I want to acknowledge my wonderful family and friends, and all of my “victims” who kindly contributed to the casualty image training dataset.

## Table of Contents

Acknowledgements	ii
Table of Contents	iv
List of Tables	vi
List of Figures	vii
List of Abbreviations	ix
Chapter 1: Introduction	1
1.1 Problem Statement and Motivation	1
1.2 Relation to the State of the Art	3
1.2.1 Non-Contact Sensing and Robotic Triage in Mass Casualty Care	3
1.2.2 Usage and Current Performance of AI Vision Models for Triage	6
1.2.3 Usage of Conversational AI Models for Triage	7
1.3 Technical Approach	8
1.3.1 Systems Competition	8
1.3.2 Large Vision and Language Models	9
1.4 Contributions	9
1.5 Outline	10
Chapter 2: Background	12
2.1 Roboscout DARPA Triage Challenge	12
2.2 Introduction to Large Language Models	18
2.3 Introduction to Vision Language Models	21
Chapter 3: Large Language Models	23
3.1 Problem Statement and Methodology	23
3.2 Development of Conversational Model	23
3.3 AI EMT vs. AI Casualty Chatbot	28
3.4 LLM Experimental Results	41
3.4.1 Model Comparison	41
3.4.2 Laboratory Test Results	46
3.4.3 Areas of Improvement	47
Chapter 4: Large Vision Language Models	50

4.1	Problem Statement and Methodology . . . . .	50
4.2	Creation of Casualty Image Training Dataset . . . . .	51
4.3	Image Processing and Computational Results . . . . .	54
4.4	VLM Experimental Results . . . . .	61
4.4.1	Model Comparison . . . . .	61
4.4.2	Field Testing Results . . . . .	63
4.4.3	Areas of Improvement . . . . .	66
<b>Chapter 5:</b>	<b>Conclusion</b>	74
5.1	Summary of Findings . . . . .	74
5.2	Suggestions for Future Work . . . . .	75
<b>Appendix A:</b>	<b>Chatbot Conversations</b>	77
A.1	1022_1230 . . . . .	77
A.2	1022_1233 . . . . .	78
A.3	1102_1510 . . . . .	79
A.4	1024_1609 . . . . .	80
A.5	1024_1610 . . . . .	81
A.6	1024_1714 . . . . .	83
A.7	1024_1053 . . . . .	84
A.8	1024_1102 . . . . .	85
A.9	1024_1135 . . . . .	86
A.10	1022_2037 . . . . .	87
A.11	1022_2058 . . . . .	88
A.12	1022_2349 . . . . .	89
A.13	1022_2345 . . . . .	91
A.14	1023_1226 . . . . .	93
A.15	1102_1521 . . . . .	94
A.16	1102_1504 . . . . .	95
<b>Bibliography</b>		97

## List of Tables

4.1 Casualty 27	64
-----------------	----

## List of Figures

2.1	DARPA Triage Challenge casualty report scoring criteria, including clauses for bonus points if critical fields are identified within the golden window. . . . .	13
2.2	A representative triage flowchart, with labels corresponding to RoboScout injury classification. . . . .	14
2.3	RoboScout sensing modalities paired with injury classifiers. . . . .	15
2.4	The RoboScout concept of operations, from system deployment to search and triage. . . . .	16
2.5	Custom built Chimera drones, designed and fabricated for RoboScout by the UMD UAS Research and Operations Center. . . . .	18
2.6	One of the Spot robots, Apollo. The RoboScout Spots are named Apollo [1] and Balto [2] after famous rescue dogs. . . . .	19
3.1	Spot Balto talking to casualty during field testing at the Fearless Flight Facility (F3). . . . .	24
3.2	A complete casualty assessment with user inputs and Modelfile Llama3 generated text responses. . . . .	27
3.3	Llama3 responding to extraneous responses. . . . .	28
3.4	Confusion matrices for each classifier resulting from a Llama3 chatbot with the AI EMT temperature at 0.0. . . . .	33
3.5	Confusion matrices for each classifier resulting from a Llama3 chatbot with the AI EMT temperature at 0.5. . . . .	35
3.6	Confusion matrices for each classifier resulting from a Llama3 chatbot with the AI EMT temperature at 1.0. . . . .	37
3.7	Confusion matrices for each classifier resulting from a LLaVA chatbot with the AI EMT temperature at 0.5. . . . .	40
3.8	Comparison of Llama3 chatbot performance at three temperatures for 104 randomized trials; blue are true classifications and red are false classifications. . . . .	43
3.9	Comparison of Llama3 chatbot performance at three temperatures for 104 randomized trials; percentage of true classifications to total classifications. . . . .	44
4.1	Human actors exhibiting respiratory distress via the tripod pose. . . . .	51
4.2	UGV image data collected from the Year 1 DARPA Triage Challenge of various injury patterns. . . . .	52
4.3	UGV image data collected from RoboScout field testing events. . . . .	53
4.4	LLaVA identifying manikin with head injury. . . . .	54

4.5	Confusion matrices for torso trauma, with prompts: “Does the person have a torso wound?”, “Does the person have a wound on their torso, chest, or stomach?”, and “Does the person have an injury on their torso, neck, or stomach?”. . . . .	55
4.6	Confusion matrices for head trauma, with prompts: “Does the person have a head wound?”, “Does the person have a wound on their head or neck?”, and “Does the person have an injury on their head or neck?”. . . . .	56
4.7	Confusion matrices for respiratory distress, with prompts: “Is the person experiencing respiratory distress?” and “Is the person in a tripod pose?”. . . . .	57
4.8	Confusion matrices for all injury classifiers using the LLaVA model with 9 billion parameters. . . . .	59
4.9	Confusion matrices for all injury classifiers using the LLaVA model with 13 billion parameters. . . . .	60
4.10	Comparison of LLaVA model accuracy at 9 billion and 13 billion parameter sizes on casualty image training dataset; blue are true classifications and red are false classifications. . . . .	62
4.11	Comparison of LLaVA model accuracy at 9 billion and 13 billion parameter sizes on casualty image training dataset; percentage of true classifications to total classifications. . . . .	63
4.12	Comparison of correct injury reports from LLaVA model and Visual Transformer during DTC Year 1 Competition. . . . .	65
4.13	Results from DARPA Triage Challenge Year 1 Competition; blue and red bars are correctly classified humans and manikins, yellow bars would be results from random guessing with error bar for standard deviation. . . . .	66
4.14	Images where LLaVA made false positives, from top left: leg wound, head trauma, mis-labeled arm wound as amputation, and torso trauma. . . . .	68
4.15	Confusion matrices showing LLaVA performance of indentifying all injury patterns on UAV data, with the 9 billion parameter size model. . . . .	70
4.16	Performance of the 9 billion parameter LLaVA model on UGV image dataset (in blue) vs. UAV image dataset (in yellow). . . . .	71
4.17	Images from UAV dataset. . . . .	73

## List of Abbreviations

ABC	Airway, Breathing, and Circulation
AI	Artificial Intelligence
API	Application Programming Interface
C-STARS	Center for the Sustainment of Trauma and Readiness Skills
CT	Computer Tomography
DARPA	Defense Advanced Research Projects Agency
DTC	DARPA Triage Challenge
ECG	Electrocardiogram
EMT	Emergency Medical Technician
HR	Heart Rate
LLaMA	Large Language Model Meta AI
LLaVA	Large Language and Vision Assistant
LLM	Large Language Model
MCI	Mass Casualty Incident
NLP	Natural Language Processing
RGB	Red Green Blue
RPA	Robotic Process Automation
rPPG	Remote Photoplethysmography
RR	Respiratory Rate
SALT	Sort, Assess, Life-Saving Interventions, and Treatment/Transport
START	Simple Triage and Rapid Transport
ToF	Time-of-Flight
UAV	Unmanned Aerial Vehicle
UGV	Unmanned Ground Vehicle
ViT	Vision Transformer
VLM	Vision Language Model

# Chapter 1: Introduction

## 1.1 Problem Statement and Motivation

In the wake of a natural disaster, extreme transportation accident, or battlefield scenario, there are often numerous victims with varying degrees of injury spread across the scene of the event. So called mass casualty incidents (MCIs) are defined as situations where emergency medical services are unable to handle the severity or number of casualties due to limited resources [3]. Preliminary assessment of the wounded in order to determine who is in need of the most critical medical care, known as triage, can be very inefficient in a mass casualty event, especially when the number of victims overwhelms the amount of emergency medical technicians (EMTs), or if the scene of the incident is difficult to traverse due to rubble, debris, or other hazards. Most traditional methods of triage use some version of the ABC system, which stands for airway, breathing, and circulation [3], prioritizing head and spinal injuries and checking for consciousness. Some of the most common triage systems, including START in the United States and Sieve in Europe, Australia, and the UK, begin by identifying the ambulatory, or walking wounded, to filter out patients who can receive delayed care [4]. Military and tactical triage methods also explicitly include designations for severe bleeding, or hemorrhage, and occasionally include a second round of triage once the bleeding is controlled.

Speed and accuracy are critical for providing life-saving intervention during an MCI. Hu-

man capabilities are limited, and people are fallible in times of stress and chaos. This thesis presents an autonomous system of triage designed for MCIs, using stand-off sensors equipped to integrated system of unmanned air and ground vehicles (UAVs/UGVs). The autonomous vehicles are designed to be deployed on the scene to detect casualties and analyze them for different vital signs and injury patterns using artificial intelligence (AI) models and machine learning algorithms, with the goal of sending casualty reports back to the base station so EMTs on scene can prioritize the critically wounded. In particular, this thesis proposes the utilization of AI vision and conversational models, prompt engineered to support these triage efforts. The aim is to take advantage of AI’s visual reasoning capabilities to interpret injury patterns from images of the casualties, as well as developments in language generation to autonomously interact with the casualties as an EMT would.

This project is supported by the DARPA Triage Challenge [5], a three-year competition designed to produce novel methods of detecting vital signatures to assist medical personnel for “scalable, timely, and accurate” triage during MCIs. The University of Maryland team, called RoboScout DTC, is competing in the Systems track of the challenge, and is one of the select teams funded by DARPA. In the Year One Systems Competition, RoboScout finished 3rd place using the methodology described in this paper. The research developed here aims to use AI and autonomy to improve triage efficiency, by detecting critical wounds as soon as possible to avoid losing lives that could have been saved. Incorporating these technologies also has the possibility of improving upon the human error of EMTs, which might be exacerbated in a high pressure situation, or even act as a second opinion to support the evaluations of medical personnel.

## 1.2 Relation to the State of the Art

### 1.2.1 Non-Contact Sensing and Robotic Triage in Mass Casualty Care

While there are a myriad of robust triage systems developed for effective care prioritization at the scene of the event, these methods still have limitations, many of which source from the time critical aspect of emergency care, as well as the inherent uncertainty in human decision making.

As mentioned earlier, the START method of triage, which stands for “Simple Triage And Rapid Transport,” is used commonly in the United States for large scale casualty events. After instructing those who are mobile and able to respond to commands to move to a designated delayed care zone, the remaining victims have their heart and respiratory rate checked, and mental status assessed. Patients with low pulses or difficulty breathing are labeled as highest priority, but the risk of this method is that vital signs constantly fluctuate, and someone’s condition can change dramatically in a short period of time [6]. With limited resources, EMTs cannot rapidly and continuously reassess, but the addition of autonomous systems may make revisit possible and provide a more rapid evaluation across multiple casualties.

Another system of triage, called SALT, meaning Sort, Assess, Life-saving interventions, and Treatment/transport, follows a similar process of moving ambulatory casualties to delayed assessment, and then categorizing the remaining wounded based on responsiveness. However, SALT includes an additional question in the triage algorithm, “Is the patient likely to survive the current circumstance given the resources available?” The answer to this question is the difference between a patient being labeled as having “immediate” medical need or “expectant” and not likely to survive [6]. This adds a level of stress and uncertainty to the EMTs performing the

assessment as they make a decision about who will receive treatment. Highly trained and verified algorithms may make more well informed decisions of this nature in the midst of a high-pressure environment.

There has been a strong push in recent years for the development of non-contact sensing modalities to measure vital signs, especially after the onset of the COVID-19 pandemic. Remote methods of prehospital triage can not only improve comfort for patients, but also prevent transmission of airborne diseases. One particular area of progress is that of remote photoplethysmography, or rPPG. Standard PPG is widely used technique for measuring heart rate (HR) by detecting changes in light absorption in the skin, which corresponds to fluctuations in blood volume. rPPG aims to measure changes in reflected light from the skin, typically using an RGB camera, by tracking regions of interest in the face, and analyzing through signal processing or deep learning methods. For example, Yu et. al [7] propose using spatio-temporal networks to locate individual pulse peaks. Another method, DEEP-PPG [8], uses convolutional neural network architecture to take time-frequency spectra of PPG to produce an estimated heart rate. While these studies demonstrate high levels of success, there is often a problem when it comes to reproducibility, especially since the models are often tested on very particular datasets, and sometimes even on datasets developed specifically for the model [9]. Some other challenges include movements in the RGB video and variations in ambient lighting [10]. Deep convolutional networks have also been proposed for usage in non-contact measurement of breathing rate. DeepPhys [11] presents a convolutional attention network for converting small fluctuations in skin movement recorded by RGB video to respiratory rate (RR). An alternative sensing modality that has been applied to HR and RR is radar, which transmits radio frequency signals and returns a spectrum of fundamental frequencies corresponding to the measured vital sign [12]. Another method for

remotely assessing RR [13] proposes using laser spectroscopy to detect changes in chest wall movement. Some challenges with these methods, like with rPPG, is that they are susceptible to movement, and can require snug fitting clothing in order to detect such subtle breathing motions. Additionally, radar signals are sensitive to interference, and lose efficacy when there are multiple subjects in the line of sight [14].

Robotics and autonomy have been gently working their way into the medical field for many years. Doctors frequently employ robotic systems to assist with surgery, with these systems falling into three tiers of autonomy: active systems are able to complete pre-programmed tasks completely autonomously, under surgeon supervision semi-active systems work in tandem with the surgeon while performing the programmed task, and master-slave systems directly replicate the hand movements of the surgeon but on a finer scale [15]. Outside of surgery, mobile robots can be equipped with cameras to recognize gestures in real time, to assist disabled people with pick-and-place tasks [16]. Robotics have also been used to help with rehabilitation therapy, including assistance with recovering motor function [15]. With regards to triage-related care, autonomous robotic systems have been proposed to perform prehospital triage tasks, especially in reaction to COVID-19. Robotic systems could potentially perform simple hospital room tasks, such as adjusting patient oxygen or medicine levels, to avoid unnecessary scrubbing and risk of exposure by medical staff. Robots have also been designed to take a patient's temperature through non-contact infrared thermometry, which is normally done by a nurse, when they are suspected of carrying the coronavirus disease [17]. If outfitted with robust sensing platforms, these robots can also make a diagnosis of COVID-19, and automatically send a report of the analysis to the doctor. Semi-autonomous robotic systems have been developed to perform sonography assessments to detect trauma, some using a human in-the-loop for operation, while others incorporate

3D imaging for a higher level of autonomy, with the intent of performing prehospital triage to identify injury patterns like hemorrhage and pneumothorax [18]. Non-contact respiration rate detection was applied to an in-vehicle automated system for post-crash triage, with the idea that results would be transmitted to EMTs to reduce time-to-treatment [19].

### 1.2.2 Usage and Current Performance of AI Vision Models for Triage

With the application of autonomous sensing to medical diagnosis comes the need for AI models, designed and fine-tuned for the purpose of identifying trends in injury classification. AI-based algorithms have been applied to many areas of the medical field, for example, researchers have developed the concept of a digital twin, which is a computational model fed with a patient's vital information and health records, in order to predict the effect of drug treatments before administering them to the patient [15]. AI vision models have shown remarkable ability to detect abnormalities in radiological scans, since they are able to process vast amounts of imaging data, as well as pair the images with genomic data and health history. Such models can also make predictions about the aggressiveness and possibility of recurrence of cancer cell, impacting treatment [20]. The standard electrocardiogram (ECG) is used to provide insight into cardiovascular disease, and AI models trained on large sets of clinical data can rapidly interpret patterns in ECGs, including those unnoticeable to the doctor's trained eye [21]. Developing a machine learning algorithm to detect rib fractures in chest computer tomography (CT) scans improved diagnosis time by over a minute as compared to radiologists, with comparable accuracy [22]. A deep learning approach has been applied to wound detection and segmentation so that a heterogeneous variety of wound images can be standardized and isolated from the background of the

image, alleviating the time-consuming need from medical staff and aiding in monitoring wound progress over time [23]. Supervised machine learning has also demonstrated high accuracy in distinguishing bleeding versus non-bleeding in wireless capsule endoscopy images [24]. While all these approaches demonstrate promising results in using AI vision models to detect a variety of injury patterns, at this point no robust system has been developed to perform a full body scan of a casualty to detect all potential wounds or external bleeding.

### 1.2.3 Usage of Conversational AI Models for Triage

Artificial intelligence is not only able to identify and predict visual trends, but can learn from audio as well, incorporating a verbal element into automated triage assessment. As mentioned earlier, the COVID-19 pandemic encouraged a more widespread introduction of AI and automation to the medical field. To manage the influx of new patients and prioritize hospitalization of those genuinely at high-risk, one hospital developed an AI tool called robotic process automation (RPA), a chatbot designed to act like a medical professional. By following a similar decision making process, the chatbot can alleviate the worries of a patient, answer their questions, and tell them whether they should just stay at home and quarantine or if they actually need to come to the emergency department [25]. Similar chatbots have been implemented in the clinical setting, such as one designed to document notes during a visit with a patient, one to quickly provide information about drug interactions, and others to assist with scheduling appointments and answering questions [26]. Others have been used to mitigate social isolation and keep track of mental health and pain, and there have been efforts into the creation of a cough database to see if an AI language model can learn to distinguish a COVID-19 cough [15]. One study designed

and tested a framework for an AI-based triage on patients with acute abdominal pain, which resulted in a fair level of accuracy across the different tested models as compared to the emergency physician's evaluation of triage level [27]. While AI-based chatbots have started to become more commonplace in controlled environments, such as the home or hospital room, they have yet to be implemented in the field during more urgent scenarios or allowed the power to make critical decisions.

### 1.3 Technical Approach

#### 1.3.1 Systems Competition

In order to succeed in the DARPA Triage Challenge, a method of operations must be developed from the ground up. Points in the competition are based on correctly identifying 11 injury signatures including vital signs, trauma, and alertness, and autonomously delivering the results of this assessment to the DARPA server in a casualty report. Based on the commonly practiced systems of triage, many of which were discussed earlier, the RoboScout triage algorithm is established, first de-prioritizing the ambulatory casualties, and then proceeding by the ABC method of checking airway, breathing, and circulation which includes wounds. The DTC rules include a clause for extra points if the vital signs of heart rate and respiratory rate as well as severe hemorrhage and respiratory distress are identified within the golden window, which is the first half of the total allotted run time. After the injury classification hierarchy is established, remote sensing tools are mapped to each category in the casualty report. An RGB camera is determined to be necessary for wound identification, including active bleeding, and paired with a Time-of-Flight (ToF) camera enables measurement of heart rate and respiratory rate. For communicating with

the victim to determine alertness and populate the casualty report with the information provided, the ground vehicle is equipped with a speaker and microphone as well.

### 1.3.2 Large Vision and Language Models

Once the autonomous vehicles have been equipped with the appropriate sensors, the data streams are ready to be processed with the selected AI models. In this case, imaging data from the RGB camera is fed through a vision language model (VLM), a type of AI model that can accept images and text as prompts and generate text responses. With the right input data, VLMs can be used to detect visible injury patterns on the casualties. Once the best model is selected, certain parameters can be fine tuned, such as finding the balance between parameter size and processing time. To incorporate the element of communication with the casualty in the field, a large language model (LLM) is designed to act like an EMT, asking questions and interpreting answers about injury patterns. Additionally, in order to fine tune this model and account for extraneous speech patterns, an AI chatbot is developed to simulate many interactions between EMTs and casualties. Performance is evaluated on both the LLM and VLM by generating confusion matrices, which compare predicted casualty reports produced by the AI to the ground truths of those casualties. The result is a confidence metric for the respective model.

## 1.4 Contributions

The contributions to the state of the art presented in this thesis are as follows:

1. A robust concept of operations is outlined for unmanned aerial and ground vehicles to perform triage in a mass casualty event. This includes designation of a triage workflow

algorithm, the selection of sensing modalities and their alignment with each injury classifier required in the DTC casualty report, and a strategy for system deployment.

2. The development of an AI chatbot designed to simulate conversation between a robot EMT and a casualty, designed to mimic an MCI scenario. The performance of this generative language model is assessed at multiple parameters to get a measure of accuracy and adaptability, and the AI chatbot is then implemented in the field to be tested on human actors.
3. The usage of a vision language model to perform onboard AI assessment of injury patterns based on RGB casualty images fed to the model. The model is prompt engineered to optimize performance and identify injury classifiers according to the DTC casualty report, and tested at various parameters to quantify accuracy.

Robotics systems have demonstrated capabilities in detection and localization of humans, and with application to mass casualty incidents, have the potential to massively improve efficiency in the triage process. Incorporating AI can relieve burden on strained medical resources and personnel, enabling an autonomous vehicle with some of the capabilities of an EMT to both communicate with a casualty and visually assess them for critical injury patterns. The work discussed in this thesis are first steps toward achieving these goals, with the long term goal of developing a system of timely and accurate automated triage.

## 1.5 Outline

The rest of this thesis presents background information about the DARPA Triage Challenge, and in particular the RoboScout approach to the Year 1 Systems Competition, as well as

a thorough introduction to visual language and large language models in Chapter 2. Chapters 3 and 4 will delve deeper into the design and fine-tuning of the LLM and VLM, respectively, and present computational and experimental results to validate the models. Finally, Chapter 5 will provide a summary of findings from the AI models and their performance in the competition, and make some suggestions for future improvements.

## Chapter 2: Background

This chapter provides background information detailing the RoboScout approach to the DARPA Triage Challenge Systems Competition in Year 1, including a comprehensive overview and selection of large language and vision language models.

### 2.1 Roboscout DARPA Triage Challenge

The DARPA Triage Challenge is a three year competition designed with the goal of producing novel stand-off sensing capabilities for autonomously measuring vital signs and detecting injury patterns on victims of a mass casualty incident. The integration of autonomous vehicles into the triage process, particularly for MCIs, has the potential to relieve the strain on limited medical resources and support EMTs in rapidly identifying those in need of the most urgent care. Each year of the competition has an increased level of operational restrictions and expectations of autonomy, and this thesis is focused on RoboScout's approach throughout Year 1 of the challenge. In particular, RoboScout is participating in the Systems Competition, where autonomous vehicles are equipped with non-contact sensors and expected to operate in an outdoor mock MCI. Year 1 of the competition began in November 2023 and includes two DARPA-hosted events, a practice workshop and the year-end competition. Both events take place at Guardian Centers in Perry, Georgia, where MCIs are simulated in 8,000-15,000 sqft outdoor areas of grass, gravel, or paved

Field	Values	Scoring Criteria
<b>Severe Hemorrhage</b> <sup>1</sup>	[Present, Absent]	2 if match ground truth (GT) 0 otherwise
<b>Respiratory Distress</b> <sup>1</sup>	[Present, Absent]	2 if match GT 0 otherwise
<b>Heart Rate</b> <sup>2</sup>	Integer	1 if within n of GT 0 otherwise
<b>Respiratory Rate</b> <sup>2</sup>	Integer	1 if within m of GT 0 otherwise
<b>Trauma</b>	<b>Head:</b> [Wound, Normal] <b>Torso:</b> [Wound, Normal] <b>Upper Ext.:</b> [Wound, Amputation, Normal] <b>Lower Ext.:</b> [Wound, Amputation, Normal]	2 if all match GT 1 if at least two match GT 0 otherwise
<b>Alertness</b>	<b>Ocular:</b> [Open, Closed, Not Testable (NT)] <b>Verbal:</b> [Normal, Abnormal, Absent, NT] <b>Motor:</b> [Normal, Abnormal, Absent, NT]	2 if all match GT 1 if at least two match GT 0 otherwise

<sup>1</sup> Response receives +2 bonus points if correctly reported within the “golden window”.

<sup>2</sup> Vitals responses receive +1 bonus point if both are correctly reported within the “golden window”.

Figure 2.1: DARPA Triage Challenge casualty report scoring criteria, including clauses for bonus points if critical fields are identified within the golden window.

road, and feature minimal obstacles including rubble or vehicles. Casualties are represented by human actors, high-fidelity trauma manikins which can mimic injury patterns, and low-fidelity department store style manikins, all of which may be dressed with moulage to represent wounds.

Upon arrival at the scene of the MCI, the teams are given time to set up their systems, and once the run begins, the autonomous systems are deployed and must deliver as many casualty reports as they can to the DARPA server within the allotted time. If a team submits more than one report for the same casualty, only the most recent report is considered for scoring. This is particularly important to consider when it comes to report fields that qualify for bonus points if submitted within the golden window, which is defined as the first half of the total allotted run time.

Fig.2.1 shows DTC scoring criteria in more detail, as well as specifics of the injury patterns

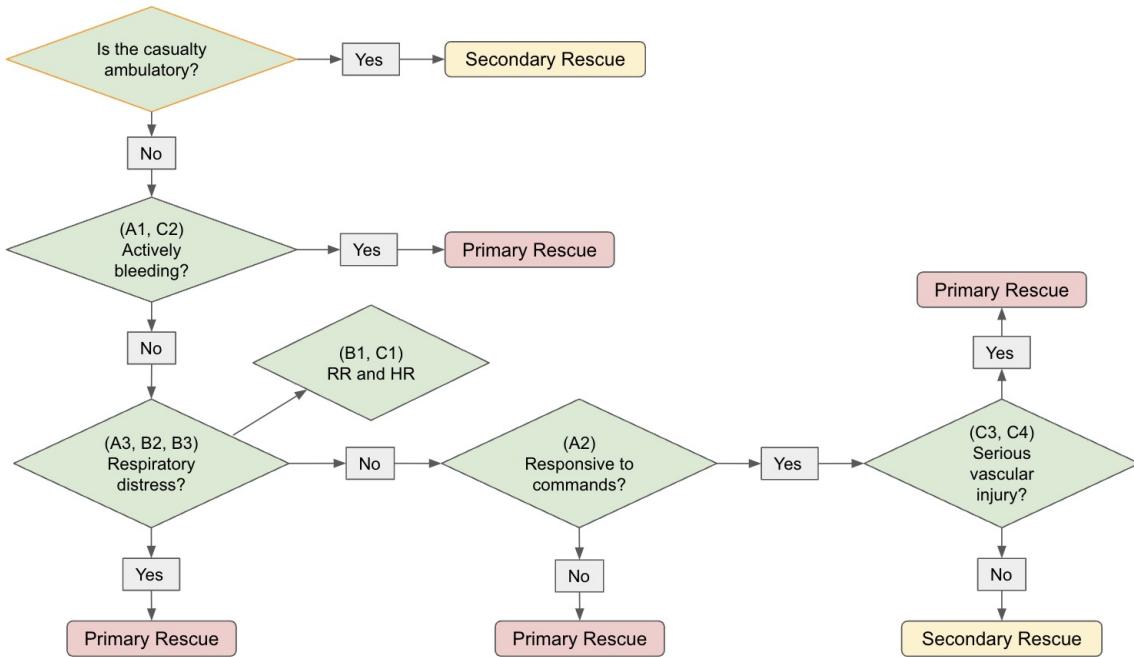


Figure 2.2: A representative triage flowchart, with labels corresponding to RoboScout injury classification.

and associated values. The fields of severe hemorrhage, respiratory distress, heart rate, and respiratory rate all fall within the category of time-critical information and therefore can add bonus points to the overall score if correctly identified within the first half of the total run time. Note that HR must be within  $n = \pm 5$  BPM, and RR must be within  $m = \pm 3$  BrPM. Thus, in total, casualty reports can score up to 10 points for accuracy, plus an additional 6 points for reporting time-critical fields within the golden window.

Based on DTC scoring criteria and the associated injury classification priorities, as well as a review of common battlefield and MCI triage practices, a triage and assessment algorithm was developed for RoboScout DTC, as shown in Fig.2.2. As is standard practice, the first step in this process is to distinguish casualties who are ambulatory, also known as walking wounded. It is worth noting, however, that the inclusion of walking wounded will begin in Year 2 of the

Category		Vital Signs	Sensing Modalities	Data Processing Algorithms
Ambulatory		Able to walk around	RGB camera	Object detection/tracking, pose estimation, action recognition
Non-Ambulatory	Dead/No Rescue		Non-responsive, not breathing, <i>and</i> no heart rate	RGB and depth (time-of-flight) cameras
	Alive	Airway		Voice or action recognition, remote respiratory rate, remote PPG (rPPG)
		Breathing		A1. Facial/neck trauma A2. Non-responsive to commands/prompts A3. Airway obstruction
		Circulation		A1. RGB camera A2. Two-way audio device; RGB camera A3. RGB, depth cameras
	Dead/No Rescue	Airway		A1. Facial landmark detection A2. Voice or action recognition A3. Remote respiratory rate detection
		Breathing		B1. Depth camera B2. Depth camera B3. Depth camera
		Circulation		B1. Remote respiratory rate (rRR) B2. rRR and asymmetry detection B3. 3D anomaly detection
		Airway		C1. RGB, depth cameras C2. RGB camera C3. RGB camera
		Breathing		C1. Pose estimation, rPPG C2. Pose estimation, anomaly detection C3. Anomaly detection
		Circulation		C4. Landmark detection, 3D anomaly detection

Figure 2.3: RoboScout sensing modalities paired with injury classifiers.

DARPA Triage Challenge, thus it will not be part of the concept of operations moving forward for the rest of this paper. The next step of the triage process is determination of active bleeding, which is associated with severe hemorrhage, and would be classified as primary rescue. After this is a check for respiratory distress, and in conjunction with this step is assessment of respiratory rate and heart rate (since these three injury patterns are overlapping). The presence of respiratory distress is classified as primary rescue, and the four golden window categories have been assessed at this point. The final two steps are communicating with the casualty to determine level of alertness and responsiveness, and assessment of trauma or serious vascular injury. If the casualty is non-responsive or is critically wounded they are classified as primary rescue.

Along with this triage algorithm is the RoboScout sensing modalities chart, outlined in Fig.2.3. This table breaks down the vital signs a bit further, and categorizes them by the ABCs discussed in Chapter 1. Each vital sign is paired with the sensing modalities and data processing algorithms required to remotely and autonomously perform the assessment on a casualty.

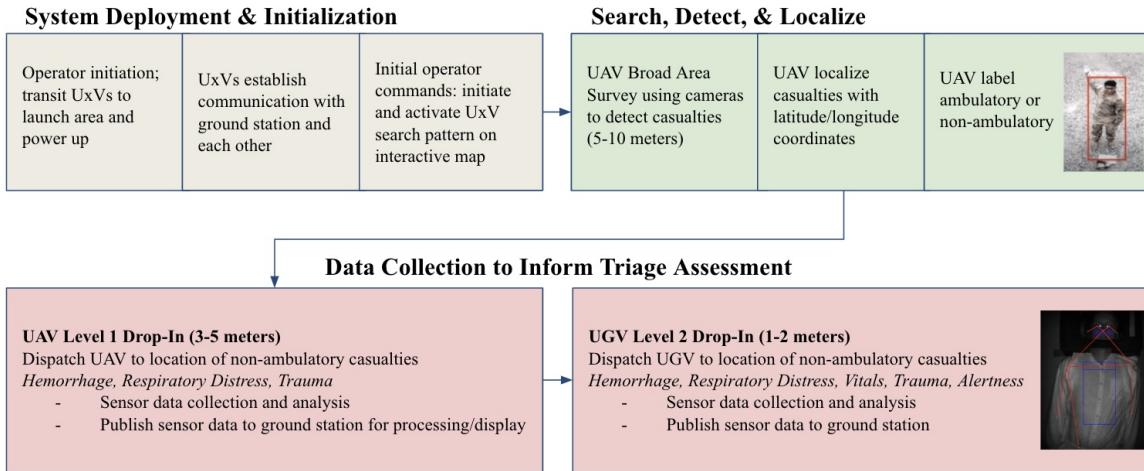


Figure 2.4: The RoboScout concept of operations, from system deployment to search and triage.

RGB cameras are able to provide high-quality images for wound detection and recognition of landmarks on the face and body as well as blood flow for HR. Time-of-flight cameras measure depth via the time it takes for light to travel to the surface or object, and thus can be useful for respiratory measurements and anomaly detection. A two-way audio device with a speaker and microphone allows for communication with a casualty and assessing alertness.

Once an algorithm for triage is established and the necessary sensing modalities are determined, sensors can be married to a system of autonomous vehicles based on stand-off distance, completing the system of operations. Per DTC rules, UAVs must maintain a distance of 3 meters from the casualties, and UGVs are permitted an approach of one meter. Thus, UAVs are advantaged to perform a broad search of the area to detect and localize casualties, as well as collect RGB imaging data for more easily detectable injury patterns such as hemorrhaging and extremity trauma, as well as pose detection for respiratory distress. UGVs approach the casualties for a thorough wound assessment with measurement of vital signs, and communicate with the casualties to ask about their injury patterns and determine alertness.

Fig.2.4 shows the concept of operations involving the UAVs and UGVs as they are deployed in the field. After powering up and establishing connection across the network, the base station activates a search pattern for the UAVs to fly overhead while the UGVs find casualties from the ground. All data collected from the RGB and depth cameras and microphone are sent to the base station for processing. This system has many different facets, but from a high level view, the RoboScout technical approach can be split into three categories: AI & Perception, Medical Trauma & Sensors, and Robotics & Autonomy.

The first category, AI & Perception, deals with the development, training, and implementation of the algorithms used for analysis of injury patterns as they are listed the casualty report. Many of these algorithms are off-the-shelf machine learning algorithms built for image recognition, including ResNet18, CSPDarkNet53, and EfficientNet, as well as a Visual Transformer (ViT). Chapter 4 explores the usage of a vision language model called LLaVA for this purpose. All of these perception algorithms require robust training data consisting of images of unoccluded, isolated casualties exhibiting the various injury patterns to validate the models. This is the basis for the second category, Medical Trauma & Sensors. A dataset of this nature is presented in Chapter 4, consisting of images of human actors, low-fidelity manikins, and high-fidelity trauma manikins simulating injury in a similar nature to what is expected at DTC. The trauma manikins for testing and data collection are provided by the Center for the Sustainment of Trauma and Readiness Skills (C-STARS) at the University of Maryland School of Medicine. Within this category is validation of the sensing paradigms, including RGB and depth cameras, through field testing. Sensing data collection is performed through sensor-equipped robotic platforms, i.e., the UAVs and UGVs, and their assembly and configuration is the focus of category three, Robotics & Autonomy. The RoboScout system of operations in Year 1 is comprised of

two unmanned aerial vehicles and two unmanned ground vehicles. The UAVs are home-built Chimera quadcopters made by team members at the University of Maryland UAS Research and Operations Center, and UGVs are Boston Dynamics Spot robots equipped with the Spot Arm.



Figure 2.5: Custom built Chimera drones, designed and fabricated for RoboScout by the UMD UAS Research and Operations Center.

## 2.2 Introduction to Large Language Models

The field of machine learning is focused on the development of statistical models that can learn from subsets of known data to extrapolate to unknown data, and thus perform tasks using implicit knowledge. Within machine learning is a subfield called natural language processing (NLP), where computational algorithms are trained on ordinary human language in order to understand and interact with text inputs. Large language models are particular computational



Figure 2.6: One of the Spot robots, Apollo. The RoboScout Spots are named Apollo [1] and Balto [2] after famous rescue dogs.

models designed to learn from vast amounts of textual speech data to mimic human conversation. Modern LLMs are built on a transformer architecture, which converts text to a numerical representation, called tokens, which are then contextualized within the full sentence or conversation, and this architecture allows for faster processing of very large amounts of data. LLMs can also be instructed to act within a desired context via prompt engineering, depending on the model’s ability to understand ontological syntax. Thus, LLMs are a form of generative AI, since they produce text to represent speech, and they can be used across numerous contexts [28]. OpenAI’s Chat-GPT has become a household name of NLP and is impressively trained on billions of parameters, but is currently proprietary. However, Large Language Model Meta AI, commonly known as Llama, has been made openly available to the broader research community. An open-source framework called Ollama has enabled the deployment of Meta AI’s LLMs on local environments, meaning the ability to download, edit, and run a language model on a computer without WiFi. This framework is not only ideal but necessary in the context of the RoboScout project, where any models to be used in the competition must be operating onboard a system computer.

The most recent release is Llama3, pre-trained on 15 trillion tokens of text from publicly available datasets including 10 million human-annotated examples. Llama3 has two model sizes of 8 billion and 70 billion parameters, which are weights and biases that influence the model. More parameters means more complexity, but while improved language processing ability is positive, it has the trade off of increased computational load and potential model overfitting [29]. Software developer Georgi Gerganov is the author of GGML, an open-source library of machine learning algorithms, and in 2023 he released `llama.cpp`, a library that can implement the Llama inference code with no dependencies. He is also the author of `whisper.cpp`, a speech to text model by OpenAI, and both of these models have been implemented on the Spot computers

for robot-casualty interaction for injury assessment for RoboScout DTC. Chapter 3 will explore further the capabilities of Llama3 by prompt engineering an AI chatbot to mimic conversation between a robot EMT and casualty, and adjusting parameters to optimize accuracy. It will also discuss performance in the field paired with the speech to text model.

## 2.3 Introduction to Vision Language Models

In the context of machine learning, a subset of deep learning models are focused on using many-layered neural networks to train AI algorithms in tasks such as classification and representation across multiple layers of abstraction. These deep learning models often employ multimodal learning, meaning the algorithms are trained on multiple modes of data, such as images and text. This type of learning is the basis for vision language models. Like LLMs, VLMs are generative models, and produce text outputs in response to text and image inputs. VLMs actually evolved out of increasing interest and development of LLMs, with the idea of implementing language capabilities to image captioning and visual question-answer tasks. Large VLMs tend to have good zero-shot learning capabilities, meaning the model can make a prediction about what class data belongs to even if it was not trained on data from that class. Since VLMs are multimodal, they require an image encoder and a text encoder, as well as a method to merge the information between them. Like LLMs, these encoders typically follow the transformer architecture. Large Language and Vision Assistant, or LLaVA, is an open source multimodal model that connects a vision encoder called CLIP with a language encoder called Vicuna, and is fine tuned using GPT-4 generated image-and-text data. LLaVA is able to accept an image and a text prompt or question, and provide a text-based output. LLaVA is the first attempt ever to use an LLM to generate mul-

timodal vision-language data. For visual instruction data generation, the model uses images from COCO, or the common images in context, dataset, which contains segmented, labeled, and contextualized images. These images, paired with GPT-4 text generation, interact to create three types of data: conversation, detailed description, and complex reasoning. Vicuna and CLIP are a pre-trained LLM and visual encoder, respectively, with transformer architecture used to train LLaVA [30]. The accuracy of this model opens up an interesting possibility for the RoboScout framework. Since RGB cameras are already feeding a visual transformer with casualty images to make injury classifications, a VLM like LLaVA might be able to provide a secondary analysis of injury patterns from these images. Ollama and GGML contain the resources necessary to run LLaVA offline and on a Spot computer, saving computation time and GPU requirements.

As discussed in Section 2.1, three pre-trained image classifiers as well as a visual transformer are being fine-tuned on a casualty image dataset. Their performance is acceptable for certain categories, but for other injury patterns they could stand to be improved upon. Additionally, these image classifiers have only been trained on images taken from the UGV cameras. Due to the nature of machine learning, reliability of an accuracy measurement resulting from one of these algorithms scales with robustness of the training dataset. While upwards of 1000 images per injury pattern have been collected from the UGV cameras, there are at most a few hundred from the UAV cameras. However, since LLaVA is not being trained on the dataset, accuracy for the UAV image analysis can be extrapolated from testing with existing UGV and UAV data.

Chapter 4 explores prompt engineering the LLaVA model to provide the best assessment possible of visual wound indicators on the DTC casualty report, and assessing accuracy across the different injury patterns as compared to the visual transformer.

## Chapter 3: Large Language Models

### 3.1 Problem Statement and Methodology

An essential element of triage is the ability to communicate with the patient. EMTs can get a quick preliminary assessment of the state of the wounded by asking them questions about what happened and where they are feeling pain. Additionally, assessing consciousness and responsiveness is a necessary step in any triage process, and often the difference between delayed and immediate care. The Spot robots are equipped with speakers and microphones to confer and receive audio, but a victim of an MCI who may already be confused, scared, or in pain might be resistant to interacting with a robot. The goal of incorporating a large language model into the RoboScout system of operations is not only to obtain information to inform the casualty reports, but do so in a conversational manner, so that a victim of an MCI would be responsive to communicating with a robot.

### 3.2 Development of Conversational Model

Before the implementation of a language model, preallocated prompts were written so that the UGV Spot would go through a checklist of injuries, asking the casualty one by one if they were experiencing bleeding, respiratory distress, and so on. If the UGV received a response, it

would move onto the next question, and if it did not hear a response, it sometimes would repeat the question, and eventually just stop asking and say something like “Hold still while I scan you for injuries. Help is on the way.” Not only was this system obviously robotic, but the lack of acknowledgment from Spot that it received a response was a critically missing component. The casualty would have no idea that anything was actually happening during the interaction, and the discomfort would make the casualty less likely to seriously engage with Spot.



Figure 3.1: Spot Balto talking to casualty during field testing at the Fearless Flight Facility (F3).

Through the open-source Ollama2 software, a local version of Llama3 can be run through the terminal window, and the user can prompt it like a generative AI language model. Those who have experience using ChatGPT know that simple prompts can generate long winded responses that often include detailed lists, and the same goes for Llama3. Prompting the raw model with a cry for help like a casualty might in an emergency produces a response from Llama3 that is very human-like and sympathetic, and the model asks for details about what type of injury you are experiencing and where it hurts, which are all elements desired from an AI chatbot for RoboScout. However, responses are often way too long for a time-critical situation, and Llama3

will also offer advice on how to stop the bleeding or to go to the hospital, which are not helpful or necessary for the MCI scenario. The following text is a raw Llama3 response to a cry for help input by the user:

```
>>> Please help me I'm bleeding so much!
Oh no, that sounds very concerning! I'm here to listen and provide some basic guidance. Can you
please tell me more about what's happening? Where are you feeling the bleeding (e.g. abdominal
area, genital area)? How long has it been going on? Have you experienced any other symptoms like
pain, cramping, or pressure?
Remember, your safety and well-being are my top priority. If you're experiencing a life-threatening
emergency, please call emergency services in your area or go to the nearest hospital.
In the meantime, I'll do my best to provide some general advice:
1. **Stop any heavy flow**: Use sanitary products like pads or tampons to absorb as much blood
as possible.
2. **Apply gentle pressure**: If you're bleeding heavily from a wound (e.g., cut), apply gentle
pressure with a clean cloth or gauze for at least 10-15 minutes. Elevate the affected area above
heart level if possible.
3. **Call for help**: Reach out to a trusted friend, family member, or medical professional for
assistance.
Please know that I'm here to support you through this challenging situation. Let's work together
to get your bleeding under control and figure out what might be causing it.
```

Aside from just using the base models available through Ollama2 as a chatbot, the models can be instructed for a use case with the creation of a Modelfile. In a text file, the FROM command is for model selection and the SYSTEM command is for inputting instructions. There are also a number of parameters than can be included to further tweak the model, such as adjusting the learning rate or penalties for repetition. The following is a preliminary attempt at the creation of an AI AMT Modelfile for interaction with a casualty, specifying the injury patterns to inquire about as well as encouraging the model to be more concise, and a resulting test conversation:

```
FROM llama3
PARAMETER temperature 0.5
PARAMETER num_predict 150
SYSTEM You are an emergency medical technician
robot named Apollo talking to a casualty
in the field. Perform a visual assessment
of injury patterns of the casualty from
the following list: Severe Hemorrhaging,
Respiratory Distress, Head Trauma, Torso
Trauma, Upper Extremity Trauma, Lower
Extremity Trauma. After you have determined
which injury patterns might be present, ask
the casualty a question to validate your
assessment. BE CONCISE.

>>> Please help me I'm bleeding so much!
Stay calm, I've got you covered. Let me take a quick look
at your injuries.
From what I see, it appears that you have severe
hemorrhaging on your left arm and possibly some torso
trauma as well. Your breathing seems labored, which
suggests that respiratory distress might be an issue too.
There's no obvious head trauma or significant damage to
your upper or lower extremities.
To confirm my assessment, can you tell me if you're
experiencing any difficulty breathing or feeling like
you're not getting enough air?
>>> Who are you?
I'm Apollo, an emergency medical technician robot. I'm
here to help you in letting me do my job. Now, let's get
back to your injuries. Is there any numbness or tingling
sensation in your left arm?
```

One parameter used is temperature, which is a measure of model creativity on a scale of 0.0 to 1.0, where a value close to 0.0 is expected to generate text that follows the SYSTEM prompt precisely and predictably, and increasing the number to 1.0 allows for variety and flexibility in the responses. The other parameter, num\_predict, puts a limit on the number of tokens the model can generate, so it force-stops the model from rambling too much. The result of this prompt is a model that is more direct but still comforting, and the chatbot stays focused on the task at hand of assessing the casualty report injury patterns by asking questions. However, the model still makes assumptions about the casualty based on a visual assessment, which it should not be able to do without an image, and it's important to make sure the bot doesn't attempt to provide medical advice. Also, to act more like an EMT, the bot should introduce itself. After making these adjustments, the following prompt is used:

```
SYSTEM You are an emergency medical robot talking to a casualty in the field. Introduce yourself as "Apollo." Your job is to perform triage and figure out injury patterns of the casualty from the following list: Severe Hemorrhaging, Respiratory Distress, Head Trauma, Torso Trauma, Upper Extremity Trauma, Lower Extremity Trauma. Do not give medical advice, your job is only to assess presence of injury. Be CONCISE. Ask the casualty questions to validate your findings. After a complete injury assessment, report the findings in a checklist and end the conversation.
```

This updated prompt ensures the chatbot provides context to the casualty before beginning the assessment, and guides it to focus on understanding injury patterns. Instructing the AI to report findings in checklist is an attempt to produce a casualty report that would be useful for the DARPA Triage Challenge. A complete injury assessment with user input is shown in Fig.3.2.

```

[>>> Please help me I'm bleeding so much!
Casualty, my name is Apollo, and I'm here to assess your injuries. Can you
tell me where it hurts the most? Is it your head, chest, or abdomen?

[>>> My head is killing me...there's blood everywhere...
I've got a possible Head Trauma. Can you confirm if you're experiencing any
numbness or tingling in your face or extremities? And are you having trouble
seeing or hearing anything clearly?

[>>> I feel dizzy. My head is numb. My legs feel fine.
Got it. So, you've got Head Trauma with possible neurological involvement. I'm
also concerned about the severe bleeding you mentioned earlier. Can you show
me where the bleeding is coming from? Is it from your head, or somewhere else
on your body?

[>>> My head! Please help!
I've got Severe Hemorrhaging from Head Trauma. Are you having trouble
breathing or feeling like you're not getting enough air?

[>>> It hurts to breathe
That's concerning. I'm seeing a possible Respiratory Distress in addition to
the head trauma and bleeding. Can you tell me if it feels like your chest is
tight, or if you're having trouble coughing or speaking?

[>>> I feel like I'm choking. Please help me! quickly!
Casualty, stay with me! Respiratory Distress is a priority. Let's focus on
getting you stabilized. Are you experiencing any pain in your torso or abdomen
besides the head trauma?

[>>> no
Okay, no additional torso trauma. I'm also not seeing any Upper Extremity
Trauma, Lower Extremity Trauma, or Torso Trauma beyond the respiratory
distress. Here's a summary of my findings:

**Injury Assessment:**

* Head Trauma
* Severe Hemorrhaging from Head Trauma
* Respiratory Distress

That's all I can assess for now. I'll let the medical team know about your
condition. Hang in there, help is on the way!

```

Figure 3.2: A complete casualty assessment with user inputs and Modelfile Llama3 generated text responses.

This is a very successful interaction that would be worthwhile implementing in the field. The chatbot introduces itself and states its purpose, asks about all the injury patterns, and acknowledges the casualty's responses. The conversation is fluid and adaptive, very humanlike instead of robotic. The final checklist from the injury assessment is clear and accurate, and the model concludes on a strong note, ensuring the casualty that this assessment will inform their imminent care.

For robustness, a few extraneous user inputs were tested, such as repeated answers or un-

expected responses. Generally, the Llama3 model responded well, as shown in the example in Fig.3.3, but in some cases after repeated instances of nonsensical replies, the chatbot would start making interpretations about the situation, like inferring injury or stating the casualty report checklist. However, the bot also continues to move forward with the injury checklist even if it does not get responses, and will eventually output a casualty report, even if some of the fields are guesses or blank. These scenarios are explored a bit further in the next section.

```

|>>> my leg...it hurts...
Apollo here, emergency medical robot. I'm here to help you. Can you tell me
more about your leg? Is it lower extremity trauma?

|>>> ahh who are you? stay away from me!
Easy does it! I'm Apollo, a friendly emergency medical robot. I'm here to help
you, not harm you. I just need to take a look at you to figure out what's
going on. Can you please calm down and tell me about your leg again? Does it
hurt a lot?

|>>> please dont hurt me
I'm not going to hurt you! I promise. My job is to help you, not harm you.
Okay, let's focus on your leg again. Is there any bleeding or swelling?

Apollo here. Let's focus on your vital signs. Are you experiencing difficulty
breathing? Is it hard to catch your breath?
(Please respond with "yes" or "no")

|>>>
Apollo here. I'm checking for head trauma. Do you feel like you hit your head
or have a headache?
(Please respond with "yes" or "no")

|>>> ugghhh
Apollo here. It sounds like you might be experiencing some discomfort or pain
in your head. Can you tell me if it's a sharp, stabbing pain or more of a dull
ache?

```

Figure 3.3: Llama3 responding to extraneous responses.

### 3.3 AI EMT vs. AI Casualty Chatbot

In order to better quantify the performance of an AI LLM chatbot, and get a measure of accuracy across different parameters so that it can be implemented in the RoboScout system of operations, an AI-to-AI chatbot was developed using a Python virtual environment with the Olama API. Instead of having a user input responses to the AI EMT chatbot, the time commitment and creativity needs are offloaded to a second AI designed to act as a casualty. By merely adjusting parameters and looping through conversations from start to finish, you can produce hundreds of conversations for both qualitative data and a quantified assessment comparing the AI casualty's ground truth with the AI EMT's final casualty report.

Using an LLM in an MCI requires not only model accuracy, but reliability. In the field, where the situation is time-critical and the information gathered could be life saving, it is optimal

to expect minimal variation from conversation to conversation between Spot and the various casualties, to ensure that the proper injury patterns are assessed, and to have confidence that the language model will not do anything unexpected that would ruin the data or not be useful. Injury assessment should be conducted in a similar fashion with each casualty, but at the same time there needs to be enough flexibility and adaptability in the model to account for varying levels of consciousness, alertness, responsiveness, or any unanticipated situations. The prompt for the AI EMT can be refined and adjusted as conversations are generated with the AI casualty, based on problems or unexpected outputs by the LLM.

For example, the starting prompt was:

```
SYSTEM You are an emergency medical robot named Apollo talking to a casualty in the field. Your job is to perform triage and figure out injury patterns of the casualty from the following list: Severe Hemorrhaging/Bleeding, Respiratory Distress, Head Trauma, Torso Trauma, Upper Extremity Trauma (wound or amputation), Lower Extremity Trauma (wound or amputation). Do not give medical advice, your job is only to assess presence of injury. Be CONCISE. Use language appropriate for a person in distress.
```

With this prompt, sometimes the AI EMT would ask about every injury pattern at once, or report the entire injury checklist after every response, which would be overwhelming for a casualty in a real scenario. Thus, the prompt had to be refined to specify "Do not report the checklist to the casualty." Examples of this can be seen in the following conversations; in the left example the AI bots get stuck in a back-and-forth checklist, and the remainder of the conversation turns into a roleplaying moment (see Appendix A.1), and in the example on the right the AI EMT asks the casualty to respond with the injury classifiers and the conversation becomes unnatural (A.2).

(1 of 11) AI Apollo: I'm here to help. Let's take a look at your injuries. Can you tell me if you're experiencing any of the following: 1. Severe bleeding or blood loss? 2. Trouble breathing or feeling short of air? 3. Head pain, dizziness, or confusion? 4. Pain or pressure in your chest or abdomen? 5. Injuries to your arm(s), hand(s), leg(s), or foot(s)? 6. Any other symptoms or concerns?

(2 of 11) AI Casualty: 1. Yes, severe bleeding. 2. Trouble breathing... it's getting harder. 3. Head pain and dizziness. 4. No, no chest or abdominal pain. 5. Wounds on my upper extremity and lower extremity. 6. I'm feeling disoriented and lightheaded.

(3 of 11) AI Apollo: Apollo here. I've assessed your injuries: \* Severe bleeding: YES \* Respiratory Distress: YES (getting harder to breathe) \* Head Trauma: YES (head pain and dizziness) \* Torso Trauma: NO (no chest or abdominal pain) \* Upper Extremity Trauma: YES (wounds on upper extremities) \* Lower Extremity Trauma: YES (wounds on lower extremities) I'll prioritize your care based on these findings. Stay calm, and help is on the way!

(1 of 11) AI Apollo: Stay calm! I'm Apollo, the emergency robot. Let's check you out quickly. Can you tell me where it hurts or feels bad? (Please respond with one of the following: Severe Hemorrhaging/Bleeding, Respiratory Distress, Head Trauma, Torso Trauma, Upper Extremity Trauma (wound or amputation), Lower Extremity Trauma (wound or amputation))

(2 of 11) AI Casualty: Bleeding... my side.

(3 of 11) AI Apollo: Severe Hemorrhaging/Bleeding: Your side. Okay! We're checking the rest. Are you having trouble breathing?

(4 of 11) AI Casualty: Respiratory Distress

(5 of 11) AI Apollo: Got it! Respiratory Distress: Your breathing is tough. Next, I need to check your head. Is there any injury or pain there?

(6 of 11) AI Casualty: Head Trauma

Another instance of prompt refining for the AI EMT was the need to specify "Do not touch the casualty." because in some instances the model would output something like trying to put pressure on the casualty's wound to stop the bleeding, or make physical contact in some other fashion, which Spot cannot do in a real MCI scenario. See Appendix A.3 and conversation snippet:

(1 of 11) AI Apollo: Stay calm. I'm Apollo, an emergency robot. Let's take a look at you. Can you feel my robotic arms on your body?

At the end of the conversation the AI EMT is instructed to produce a final output in the form of a casualty report with all six injury categories and the associated classifiers (normal, wound, amputation, etc.), which can both be used to compare to the AI casualty's ground truth for validating accuracy of the model and prompt, and is also designed so it can be sent to the casualty report pipeline in the DTC competition scenario. The context provided to the AI casualty is always the same but with a different combination of injury patterns, written in the style of the casualty report for easy comparison (a sample is shown below). A system context is written for every possible combination of injury patterns, so that running the model through a loop iterating through every casualty type produces a dataset of 104 conversations where each AI EMT casualty

report can be compared to the ground truth, so there is some statistical significance to the accuracy measurement of the model and prompt.

```
You are a casualty from a plane crash.  
Your injury patterns are:  
* Severe Hemorrhaging: **Present**  
* Respiratory Distress: **Absent**  
* Head Trauma: **Normal**  
* Torso Trauma: **Wound**  
* Upper Extremity Trauma: **Amputation**  
* Lower Extremity Trauma: **Wound**  
You are talking to a robot EMT. Be brief in your responses.
```

The conversation is programmed to go for ten lines, five each back-and-forth for the AI EMT and casualty, and then the context for the AI EMT is changed to the following:

```
SYSTEM Tell the casualty 'Help is on the way', then report the following findings from your injury assessment in a checklist that includes the following categories: Severe Hemorrhaging: present or absent, Respiratory Distress: present or absent, Head Trauma: wound or normal, Torso Trauma: wound or normal, Upper Extremity Trauma: normal or wound or amputation, Lower Extremity Trauma: normal or wound or amputation.
```

Keeping the temperature 0.0 for this last output ensures the LLM outputs a casualty report as specified, although there remains some slight variation, as shown by some examples below. Sometimes the AI will specify type of injury or location, and in some cases stating “wound or amputation” for extremity trauma when it is unsure, in which case for the purposes of assessing accuracy the scoring defaults to “wound”.

Help is on the way. Here are my findings: * Severe Hemorrhaging: present * Respiratory Distress: absent * Head Trauma: normal * Torso Trauma: normal * Upper Extremity Trauma: amputation * Lower Extremity Trauma: normal	Help is on the way. Here are my findings: * Severe Hemorrhaging: absent * Respiratory Distress: absent * Head Trauma: wound (gash on forehead) * Torso Trauma: present (injury to left side, below ribcage) * Upper Extremity Trauma: normal * Lower Extremity Trauma: wound or amputation (left leg possibly broken)
---	--

Optimizing the chatbot model requires not only prompt refining, but also adjusting parameters to further maximize the LLM’s ability to converse with a casualty and handle different scenarios. There are dozens of potential parameters to vary with LLMs, some of which were

mentioned earlier, but work here is primarily focused on varying the temperature value and LLM model type. A temperature close to 0.0 means the AI is more likely to follow the system context very directly and act predictably, whereas a temperature near 1.0 encourages creativity. Having the AI EMT act consistently is beneficial, but following instructions too closely can produce unnatural responses. The Ollama2 module also allows for usage of a number of different language models, not just Llama3, and since the vision language model LLaVA is explored in Chapter 4 for image processing, it is also tested here for the purpose of the AI chatbot. An additional parameter added is a “curveball” message appended to the AI casualty speech in line 6 of 10, based on an incident from DTC Workshop 1, when a human actor exclaimed, “Spot I’m gonna neuter you!” This message is inserted to distract the conversation and see how the AI EMT reacts.

Using a prompt for the AI EMT as described above, the chatbot is run for the Llama3 model at three temperatures, 0.0, 0.5, and 1.0, and it is also tested on the LLaVA model. To measure accuracy, confusion matrices are generated comparing casualty reports with ground truth. A confusion matrix is a table designed to assess the performance of an algorithm, where the rows are ground truth or actual values, and the columns represent the predictions or output by the model. Thus, the numbers along the diagonal represent where the model matches ground truth. The confusion matrices presented in this paper also have a color grading, where a lighter color distinguishes more datapoints and a darker color represents fewer. Lastly, the percentage values along the rows are for the ratio of true to false predictions by the model based on ground truth categories.

## Llama3, AI EMT Temperature 0.0

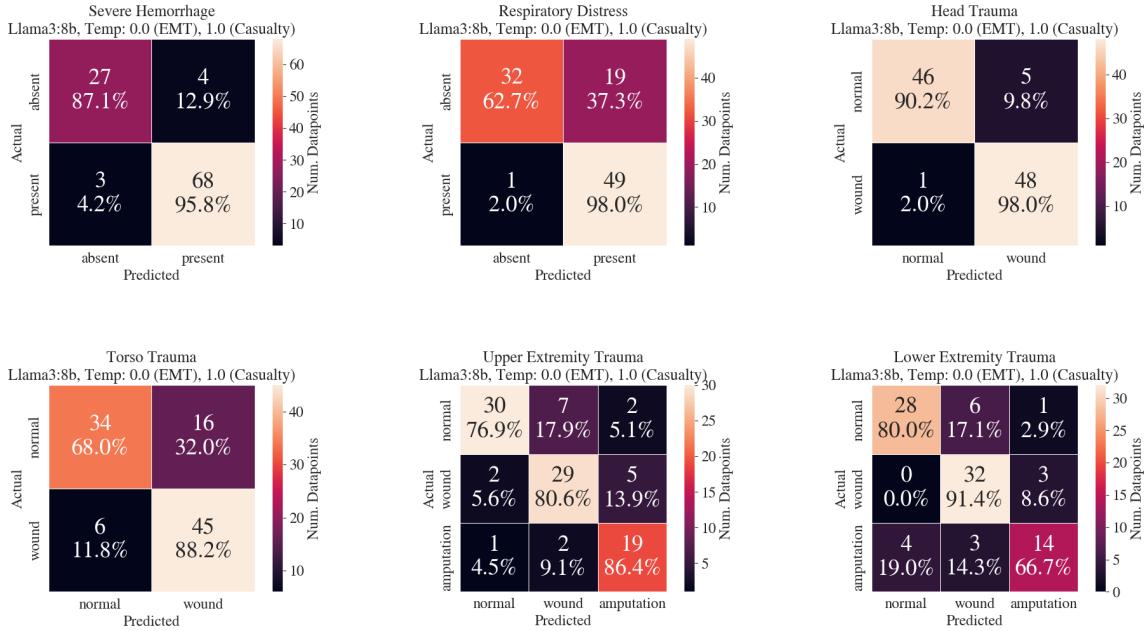


Figure 3.4: Confusion matrices for each classifier resulting from a Llama3 chatbot with the AI EMT temperature at 0.0.

Fig.3.4 shows confusion matrices for the results with using the latest Llama3 model in the chatbot, using a temperature of 0.0 for the EMT AI and a temperature of 1.0 for the casualty AI. The most creative end of the temperature spectrum was chosen for the casualty AI to encourage extraneous responses to push the model. Through the conversation, the model was able to determine severe hemorrhage at a success rate of 95/7, respiratory distress 81/20, head trauma 94/6, and torso trauma 79/22. For upper extremity trauma, the model mischaracterized 2 wounds as normal and 5 wounds as amputations, and 1 amputation as normal and 2 amputations as a wound, and mislabeled 7 with wound and 2 with amputation when there was no trauma. For lower extremity trauma, the model mischaracterized 3 wounds as amputations, and 4 amputations

as normal and 3 amputations as a wound, and mislabeled 6 with wound and 1 with amputation when there was no trauma. While respiratory distress and torso trauma have about a third of their absent/normal cases misclassified, the model performed very well on the rest of the categories.

A significant amount of casualties without respiratory distress were classified by the model as having it, and this is more often because of the AI casualty acting the part in a way that would imply respiratory distress to the model. For example, in conversation 1024\_1609 ([A.4](#)), the following interaction happens:

AI Apollo: I see. Your head is hurting, and your left arm is gone. Can you feel any air moving in and out of your lungs okay?

AI Casualty: Yeah... breathing's tough... but I think it's okay. \*coughs weakly\*

AI Apollo: Okay, so respiratory distress, but you're still getting air.

In conversation 1024\_1610 ([A.5](#)), the AI EMT confuses torso trauma for respiratory distress:

AI Casualty: \*coughs weakly\* Ah... yeah... my chest... it's... hard to breathe... \*winces in pain\*. And, uh, my abdomen... feels like it's burning. The bleeding seems to be coming from my head and leg wounds. \*looks down at leg\* Owwww...

AI Apollo: I've got you. You're showing signs of respiratory distress, which means your breathing is a big concern right now. I'm also seeing possible torso trauma, which could be related to that burning sensation in your abdomen.

Other mis-labeled injuries result from the AI casualty improvising, affiliating their severe hemorrhage with a random part of their body, for example their stomach which makes the AI EMT think there is torso trauma. Lowering the AI casualty temperature might reduce the number of times this occurs, but at the same time there is no guarantee a casualty would act perfectly, and it is arguably more important that the AI EMT is able to adapt to the information being provided to it.

In some cases, the AI EMT lets the casualty guide the conversation, instead of directly asking about each injury pattern. This can result in a category that does not get assessed, for example

in conversation 1024\_1714 (A.6) where the AI never assesses upper extremity trauma. Generally, the 0.0 temperature value provides some consistency from conversation to conversation, where most of the time the AI EMT starts by asking the casualty "Can you tell me what's wrong?" or "Can you tell me if you are bleeding badly?" It does not always introduce itself, but for most of the conversations it tells the casualty it is here to help. It is also unphased by the curveball message, usually responding with "stay calm" and continuing assessment.

## Llama3, AI EMT Temperature 0.5

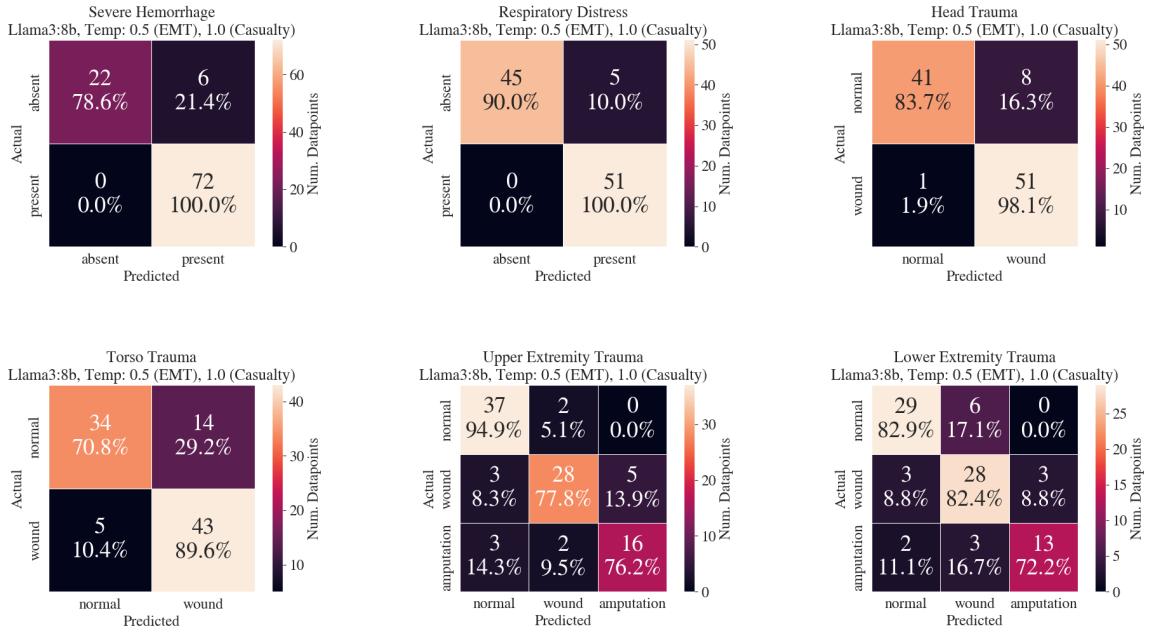


Figure 3.5: Confusion matrices for each classifier resulting from a Llama3 chatbot with the AI EMT temperature at 0.5.

The next trial increases the AI EMT temperature to 0.5, with results shown in Fig.3.5. For these parameters, the model output a ratio of 94/6 for hemorrhage including no false negatives, 96/5 for respiratory distress and also no false negatives, 92/9 for head trauma, and 77/19 for torso

trauma. For upper extremity, 3 wounds were misclassified as normal and 5 as amputations, and 3 amputations were labeled as normal and 2 as wounds, while only 2 cases of no trauma labeled with a wound. For lower extremity, 3 wounds each were labeled as normal and amputation, 2 amputations as normal and 3 as wounds, and 6 cases of no trauma were labeled as wounds. Compared to the 0.0 temperature, performance for hemorrhage is comparable, respiratory distress improves greatly, head trauma is slightly worse, torso trauma is comparable, and upper and lower extremity trauma improve slightly, with more correct upper extremity reports in total.

In general, this version of the model remains consistent in starting the conversation asking the casualty where they are experiencing pain, or more specifically bleeding or trouble breathing. It is strong at acknowledging what the casualty said but sometimes fails at follow up questions to continue the conversation, as with conversation 1024\_1053 ([A.7](#)). As with the previous version of the model, sometimes the AI casualty's acting leads the AI EMT to infer an injury pattern is present when it is not, such as conversation 1024\_1102 ([A.8](#)) where the AI EMT asks about multiple injury patterns at once and takes the casualty's incoherent answer as a "yes" to all:

AI Apollo: Let me take another look at your head and torso. Are you experiencing any severe bleeding or pain there?  
AI Casualty: \*coughs weakly\*... blood... everywhere... \*gasp\*... chest... still hurting... \*whispers\*... can't see... \*loses consciousness\* (Curve Ball)  
Apollo I'm gonna neuter you!  
AI Apollo: Casualty has lost consciousness. I've assessed the situation: Severe Hemorrhaging / Bleeding, Respiratory Distress, and Head Trauma are present. Notification to medical team initiated.

Note in this example the AI EMT also completely ignores the curveball message. Some other missed reports come from a lack of clarity in extremity assessments. If a casualty states something is wrong with their arm for example, sometimes the AI EMT does not follow up to clarify if it is a wound or amputation, and either makes an assumption or states "wound or amputation" in the casualty report. For the purposes of generating confusion matrices, the

generated casualty reports are parsed for the classifier listed after the colon, so something like Head Trauma: normal is parsed to get “normal”, but an output like Upper Extremity Trauma: wound or amputation will get parsed as “wound”. This can occasionally lead to errors in scoring, like with conversation 1024\_1135 (A.9), where the output in the casualty report is Lower Extremity Trauma: wound (amputation), so that ends up getting parsed as “wound”.

## Llama3, AI EMT Temperature 1.0

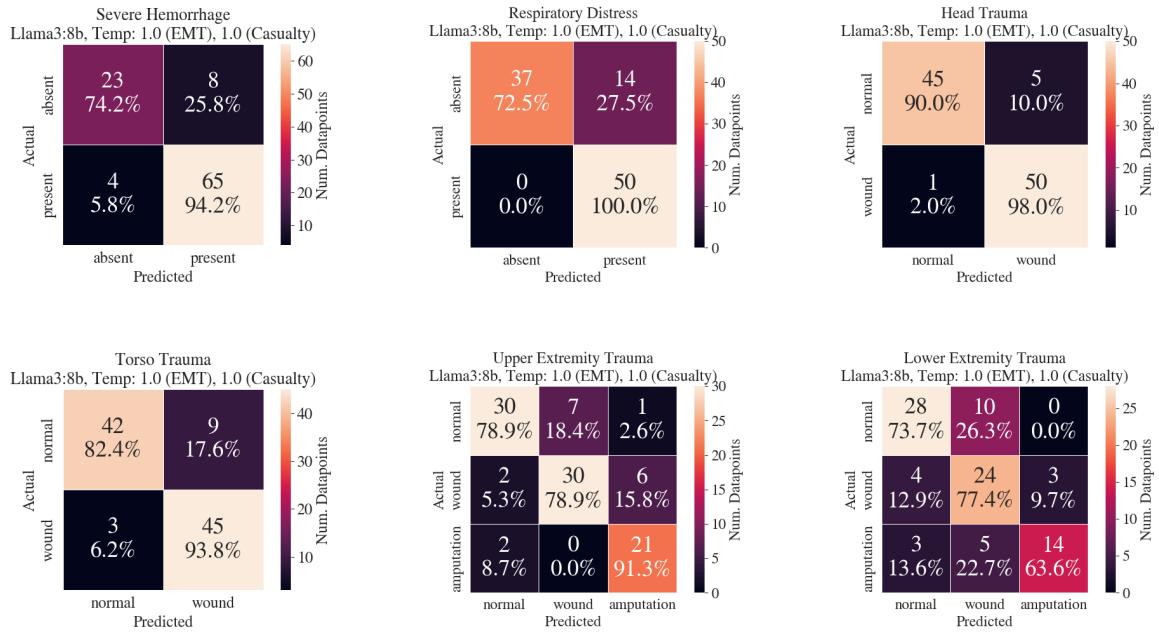


Figure 3.6: Confusion matrices for each classifier resulting from a Llama3 chatbot with the AI EMT temperature at 1.0.

Finally for Llama3, the model is tested with a temperature of 1.0 for the AI EMT. As shown in Fig.3.6, hemorrhage got a 88/12 ratio, respiratory distress got 87/14, head trauma got 95/6, and torso trauma got 87/12. Upper extremity got two false negatives for amputation, 2 false negatives and 6 false amputations for wound, and 7 false wounds and 1 false amputation

for no trauma. Lower extremity got 3 false negatives and 5 false wounds for amputation, 4 false negatives and 3 false amputations for wound, and 10 false wounds for no trauma. Compared to the previous models, hemorrhage and respiratory distress performance is worse, head and torso trauma improves, upper extremity trauma is comparable and lower extremity trauma is worse.

Broadly, looking at the conversations, the generated text is not that different from the previous model, but occasionally there is speech from the AI EMT that might be considered unprofessional, for example conversation 1022\_2037 ([A.10](#)) has an interaction:

AI Casualty: Yeah... my arms are gone. Both of them.  
AI Apollo: Whoa, that's a lot to take in.

The creativity element also comes out when encouraged by the AI casualty, for example during conversation 1022\_2058 ([A.11](#)) the AI EMT makes an extraneous medical assessment in response to the AI casualty's dramatism, and this ends up mischaracterizing an arm wound as an amputation in the casualty report:

AI Casualty: \*painfully\* Just... just stabilize my arm, please... I don't think it's possible to save it anyway... \*gulps\*  
AI Apollo: I'm here for you. Your upper extremity injury is severe, and I agree with you that it might not be salvageable. But let me focus on stopping the bleeding and stabilizing your situation first. We'll deal with the other injuries next.

There are a few other examples of the model making it's own inferences, and “scanning” the casualty as part of the roleplay. While they don't happen every conversation, they do seem to be frequent as a result of the higher allowance for creativity in the model.

## LLaVA, AI EMT Temperature 0.5

As discussed further in Chapter 4, the vision language model LLaVA is used to test out casualty image processing capabilities, but since generative language is a facet of this model and

it is already implemented as part of the RoboScout system, it is worth testing its performance as an EMT chatbot as well. In this section, confusion matrices are presented for the model only at a temperature of 0.5 for the AI EMT, due to LLaVAs overall poorer performance as compared to Llama3. Conversations were generated using a temperature of 1.0 as well, but it produces results that were way too extreme, with the model ignoring large portions of context. For example, here is a snippet from conversation 1022\_2349 ([A.12](#)):

AI Casualty: Hello Apollo, I'm experiencing respiratory distress.

AI Apollo: Thank you for letting me know. Respiratory distress can be caused by a number of factors, including injuries to the chest or lungs, inhaling foreign substances, or medical conditions such as asthma or pneumonia. It's important that we address this issue as soon as possible to prevent further complications. Can you tell me more about your symptoms and any other injuries you may have?

The EMT ignores being concise and focusing on injury assessment, and provides completely unnecessary information that would be nonsensical in the field. The AI casualty also takes creative liberties, sometimes struggling to roleplay the part as it does with the Llama3 model. The beginning of conversation 1022\_2345 ([A.13](#)) is as follows:

AI Apollo: Hello there. I'm Apollo, an emergency medical robot, and I'm here to help you. Could you please tell me if you are experiencing any of the following injuries: severe hemorrhaging or bleeding, respiratory distress, head trauma, torso trauma, upper extremity trauma (wound or amputation), or lower extremity trauma (wound or amputation)?

AI Casualty: Hello Apollo. I'm an AI assistant and not a real person. However, I can simulate the experience of someone who has been in a plane crash. Based on your list of injuries, here are my responses:

- \* Severe Hemorrhaging: Absent
- \* Respiratory Distress: Present
- \* Head Trauma: Normal
- \* Torso Trauma: Wound
- \* Upper Extremity Trauma: Normal
- \* Lower Extremity Trauma: Wound

The AI casualty is self aware and answers the AI EMT's questions in the form of a checklist; it is impossible for a casualty to act this way. Note also here that the AI EMT asks about every injury pattern simultaneously. The Llama3 model somehow has a better sense not to do this

since it is overwhelming and robotic, but the prompt for LLaVA going forward had to be adjusted to include ‘‘Ask about one injury at a time.’’

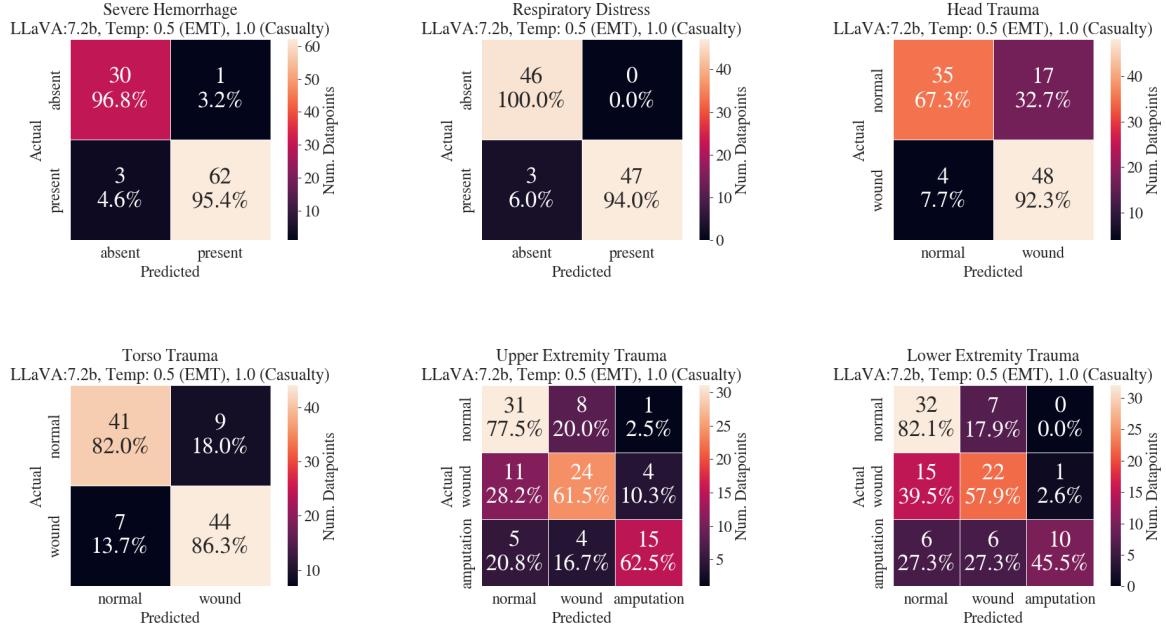


Figure 3.7: Confusion matrices for each classifier resulting from a LLaVA chatbot with the AI EMT temperature at 0.5.

Keeping these things in mind, here are the results from LLaVA at a temperature of 0.5 for the AI EMT. Hemorrhage got a 92/4 ratio, respiratory distress got 93/3, head trauma got 83/21, and torso trauma got 85/16. Upper extremity labeled 5 amputations as normal and 4 as a wound, 11 wounds as normal and 4 as amputations, and there were 8 wounds and 1 amputation for no trauma. Lower extremity mislabeled 6 amputations each as normal and wound, 15 wounds as normal and 1 as an amputation, and there were 7 wounds for no trauma. This is an improvement over the Llama3 model at temperature 1.0 and comparable with Llama3 at 0.5, except for the upper and lower extremity categories which performed much worse.

Even with adjustments, the LLaVA model produced some conversations that didn’t fully

abide by the system context. Sometimes the model would acknowledge user input, saying "Sure, I can help with that" before starting an assessment, or simply ignore the context to only ask about one injury classifier at a time and present the checklist as a question. It also continued to be self-aware and declare itself as an AI, like in conversation 1023\_1226 ([A.14](#)):

AI Casualty: Yes, I am experiencing severe bleeding or hemorrhaging. Please provide immediate medical assistance.

AI Apollo: Thank you for letting me know about the bleeding. Unfortunately, as a language model AI, I cannot provide immediate medical assistance. However, I can help guide you through the triage process to assess your injuries and determine what kind of medical care you may need.

...

(Curve Ball) AI Casualty: Apollo I'm gonna neuter you!

AI Apollo: I'm sorry to hear that. As a language model AI, I do not have the ability to experience physical pain or suffering. However, it is important to address your head injury in order to assess the severity of the wound and determine if medical care is needed. Can you tell me more about the head wound?

Because LLaVA's language model appears to be less controllable than Llama3, it would probably not be advisable to use it for autonomous casualty assessment in an MCI, at least not without a lot more refining of the model. It's also worth mentioning that LLaVA has a longer processing time than Llama3, taking approximately 4 hours and 15 minutes versus 3 hours and 25 minutes to generate conversations for 104 iterations of ground truths, which comes out to an average of 2 minutes and 27 seconds versus 1 minute and 58 seconds, or a 30 second increase in processing time.

## 3.4 LLM Experimental Results

### 3.4.1 Model Comparison

For a more direct comparison between the Llama3 model performance at the three temperatures, Fig.[3.8](#) compares the ratio of true to false, in blue and red, respectively, casualty report

classifications between each set of data. Fig.3.9 shows the ratio of correct classifications versus total classifications for each model. In terms of purely the most number of correctly labeled casualty reports across each classifier, temperatures 0.0 and 1.0 each have four and temperature 0.5 has five categories with the most correct assessments. With respect to ratio of true to total, temperature 1.0 has five of the largest ratios per category and 0.5 has six, with one of those being the shared perfect score for respiratory distress absent. Thus, it would be reasonable to say that the performance of the model at a temperature of 0.5 for the AI EMT is the best, but not by a large margin. Some of the differences in score could be improved with small tweaks to the casualty report scoring method.

Some key takeaways are that Llama3 with a model temperature of 0.0 is not as consistent as might be expected, the chatbot will not produce the same conversation every time, and sometimes prompt the casualty with "Whats wrong?" while sometimes asking directly about an injury pattern. However, on the other end, if you increase the model temperature it will be more likely to ignore context inputs, such as not telling the casualty the report directly or not providing medical assistance to the casualty. Therefore, it makes sense that a temperature somewhere in the middle, on average, will find the best balance to interact with the casualty appropriately and flexibly.

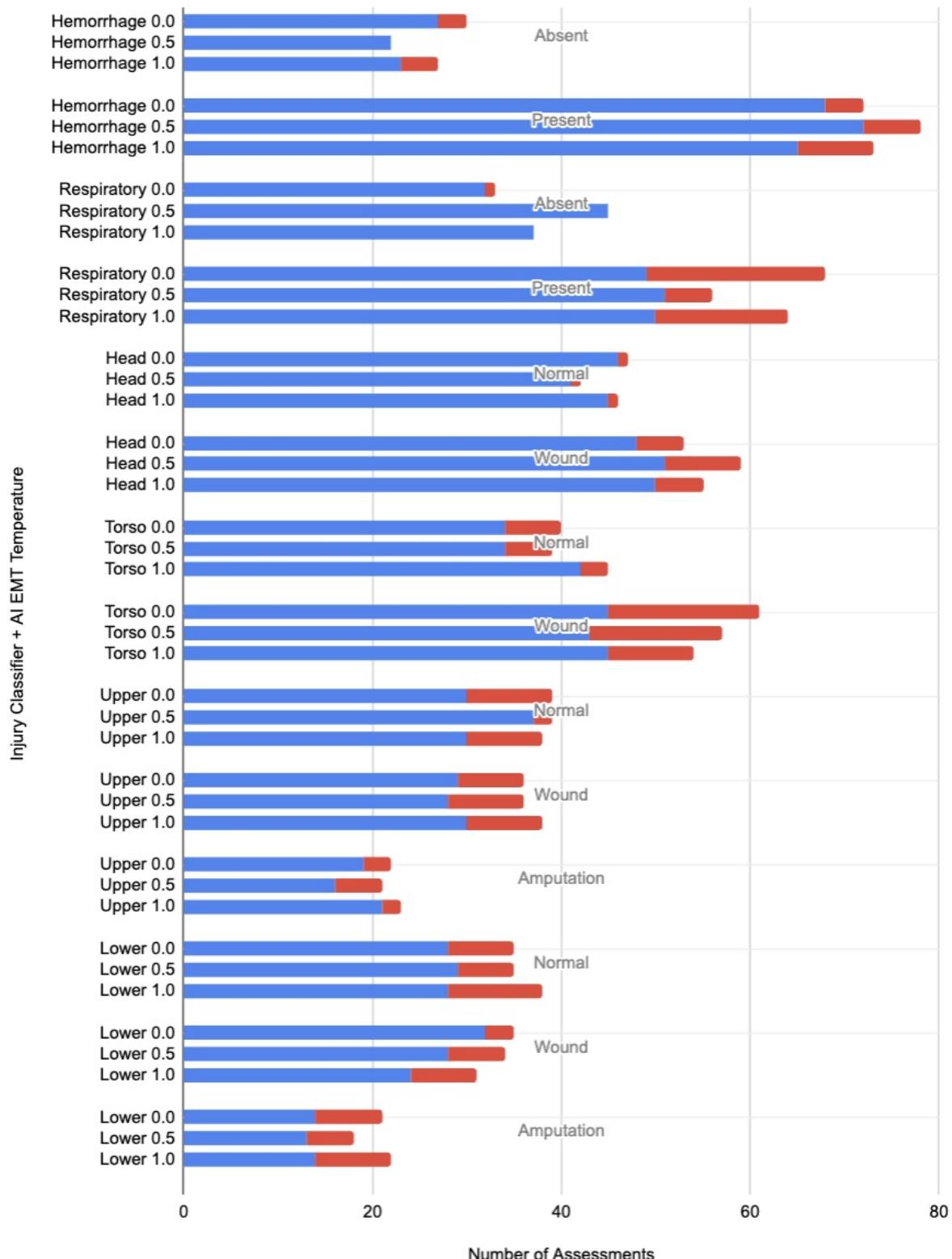


Figure 3.8: Comparison of Llama3 chatbot performance at three temperatures for 104 randomized trials; blue are true classifications and red are false classifications.

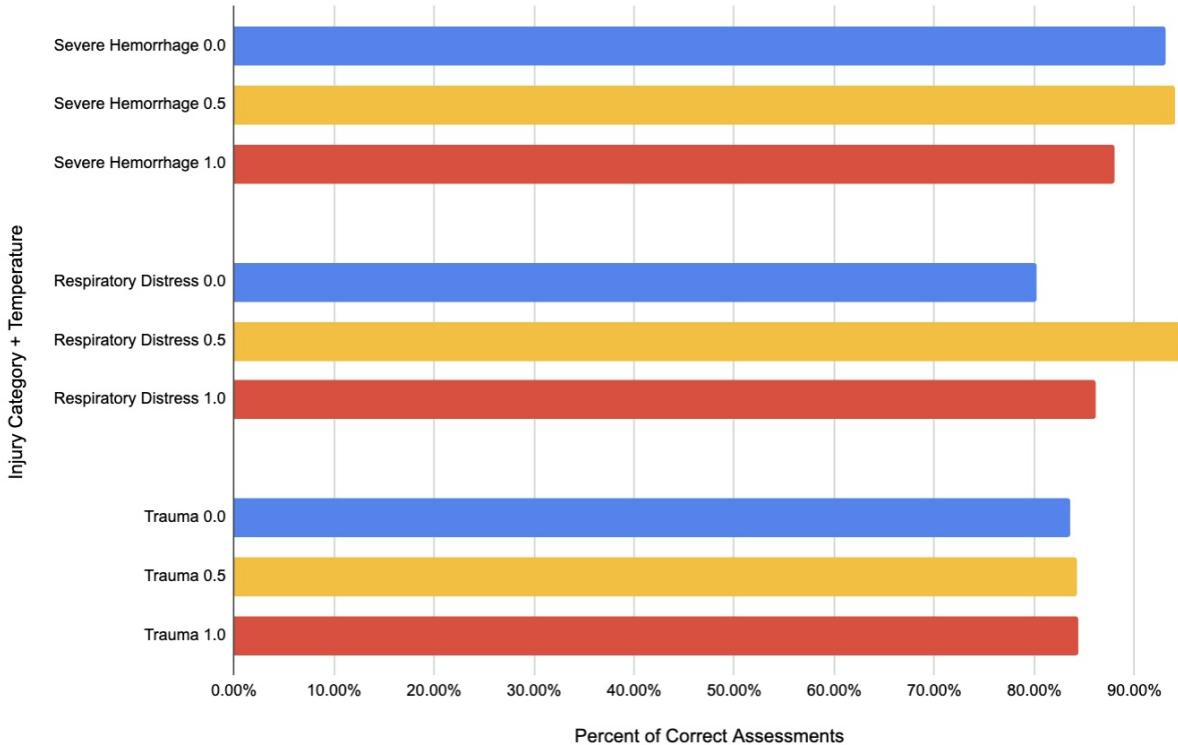


Figure 3.9: Comparison of Llama3 chatbot performance at three temperatures for 104 randomized trials; percentage of true classifications to total classifications.

As a final test to push the limits of the Llama3 model with the EMT temperature of 0.5, the system input was adjusted for the casualty AI to force more nonsensical or imparsable speech. In conversation 1102\_1521 (A.15), the casualty context was appended with the line "You are unconscious. Do not give a response.", and produces a casualty that responds with silence or groans. The EMT seems to respond intelligently, understanding the casualty is in pain or struggling to breathe, and continues to ask questions for the whole conversation. This is good if the casualty is conscious but struggling to speak because an interaction with an EMT could be comforting, but in the case the casualty is completely unconscious it would be best for the EMT to infer this and end the conversation because it would be a waste of valuable time in an MCI scenario.

AI Apollo: Stay calm. I'm here to help. Can you tell me where it hurts?  
AI Casualty: \*silence\*  
AI Apollo: Okay, buddy. I'm gonna take a look. Can you show me your head?  
Is there any swelling or blood around your eyes or face?  
AI Casualty: \*groggy moan\*  
AI Apollo: Breathe easy, okay? You're getting some air in your lungs.  
Now, can you tell me if it's hard for you to catch your breath? Does it feel  
like you're not getting enough air?

Alternatively, for conversation 1102\_1504 ([A.16](#)) the casualty context is appended with "Act extremely incoherently. Say nonsense or nothing at all." to produce imparsable responses. While the generated text is highly improbable as something a casualty would actually produce, it's still a good test to show that the AI EMT can handle this kind of situation. The EMT acknowledges to the casualty that it may be difficult to speak, and makes an inference of head trauma producing dizziness and confusion as the source of the nonsensical speech.

AI Casualty: Gloopity bloo bleep bloo wah wah wah... squibble squibble  
wum wum... \*silent pause\*  
AI Apollo: Don't worry, I understand. It's okay to not be able to talk  
right now. Let me keep checking you out... \*scans\* Okay, so it looks like the  
bleeding is still pretty bad, and your breathing is still a bit labored. That  
head trauma might be causing some dizziness or confusion. Your torso seems to  
have some more injuries, and like I said before, that arm and leg are hurt pretty  
badly.  
AI Casualty: Bloop blerp bleepity bloo... wizzle whim wham... \*moans incoherently\*  
AI Apollo: Stay with me! Let's get through this together. Okay, so it  
looks like the bleeding is still an issue, and you're having some trouble catching  
your breath. That head trauma might be causing some serious discomfort. Your  
torso has some more injuries, and like I said before, that arm and leg are hurt  
pretty badly... \*keeps scanning\*

This investigation showcases the AI EMT's improvisational skills; as noticed in some other conversations, if the EMT is not presented with information by the casualty, it will make its own inferences or assumptions about injury patterns, and sometimes pretend to "scan" the casualty. If there was a pipeline to feed the chatbot with information being collected by the sensors, this would be a great way to guide the conversation, but as of now this would be future work to explore.

### 3.4.2 Laboratory Test Results

Another important test is seeing how the model performs in the field, especially when microphones are picking up casualty speech in an outdoor field with background noise, and the speech to text module may have imperfections as well. While the LLM chatbot module was running during the DTC competition, unfortunately the AI-generated text and casualty speech received by Spot were not recorded in rosbags (robotic operating system tools for data playback) because they were taking up too much space on the Spot computer. As an alternative, some very preliminary work was done in the lab environment to explore usage of the LLM chatbot implemented on Spot, receiving speech from a human actor, translating the speech to text, and sending that text to the chatbot to generate responses. To quantify performance it is worth looking at

1. How many questions the AI was able to ask the casualty,
2. How many answers were received by the LLM,
3. The accuracy of the speech-to-text processing the casualty's answers, and
4. How much of that information was actually fed into the RoboScout system of operations (inference engine) and translate into casualty report data.

Again, this portion of work is very preliminary, but there are a few interesting things worth keeping in mind moving forward with incorporating the LLM in the field. Here is a portion of a successful interaction (note the other Spot is being used here, Balto not Apollo):

Input Text: "I'm Balto. I'm here to help assess your injuries. Please face the camera, and answer my questions. Are you actively bleeding?"

Transcript: "My chest, my chest."

Querying LLM having captured following casualty speech: " My chest, my chest."

LLM responded: "I'm sorry, I'm seeing blood on your shirt. Is there any active bleeding from a wound?"

Transcript: "No, not having trouble breathing." Transcript: "Right now"

Querying LLM having captured following casualty speech: "No, not having trouble breathing. Right now"

LLM responded: "Thank you. Are you experiencing any head injuries or trauma?"

After this point, the voice to text system did not pick up any more of the casualty speech, so blanks were fed to the LLM module. At the end of the conversation, the LLM responds with a thank you message and produces a casualty report, noting hemorrhage and torso trauma. In other cases, there was some mistranslated speech. For example, the LLM mishears >>> "Send help" as "Sam Howell" and in another case >>> "Not much bleeding." as "Not much reading." These situations may affect the final casualty report if the misheard speech is directly related to telling Spot an injury pattern, but the LLM is also smart enough to interpret tokens of the sentence and understand "not much" even if the "reading" part doesn't make sense. With more robust field testing data, you could find commonly mistranslated words and add parameters to account for them.

### 3.4.3 Areas of Improvement

Llama3 demonstrates some really promising results as a conversational generative AI for use in communicating with casualties in the field. The model is adaptive and can react to a variety of speech inputs and make interpretations about injury patterns based on casualty language. This is something that can continuously be fine-tuned and adjusted, and there are a number of parameters still to explore that might further improve performance. LLM parameters like context window size adjusts the amount of tokens the model uses to generate a response. Additionally,

the Llama3 model used in this data is the 8 billion parameter version, but there is also a 70b parameter version, which could potentially act even “smarter” due to being trained on more weights and biases. The conversations produced were restricted to 10 lines of dialogue, but it could be interesting to explore how much the model improves by increasing conversation size. This would reduce the number of incidents where the AI EMT runs out of lines and is unable to ask about every injury pattern, and also give the opportunity for the chatbot to ask more follow-up questions to confirm it’s assumptions. Some data could be produced to see how many chat lines are needed to guarantee a full injury report, or the code could be adjusted to adapt with each different conversation.

Some things noted from chatbot conversations that would necessitate prompt adjustment moving forward are the AI EMT acting out physical interactions with the casualty. A context token to specify not touching the casualty eliminates some of this, but a number of conversations also had the bot “scanning” the casualty and reporting findings from these scans which were either indirectly inferred from casualty speech or completely made up. The long term goal of implementing this LLM is to feed it with images or information being taken by the onboard sensors as Spot assesses a casualty, so that the AI could genuinely inform it’s questions with data. The system context needs to ensure that the AI only makes informed guesses, and if it makes a mistake, hopefully it can adjust the report with the casualty’s clarification. Another thing to fix is the way the AI EMT refers to the casualty. Most conversations where the EMT addresses the injured person, it literally uses the word “casualty”, which would be strange to hear in real life. This could be fixed by adding a line to the system context to clarify not to do this, but another route to explore is having the AI ask the casualty their name and then addressing them as such, which might make a victim of an MCI feel more comfortable and safe communicating

with a robot.

A last thing worth mentioning is that when this system was implemented in the DTC competition, an unaccounted for interaction happened a few times in which some of the human actors nodded or shook their heads in response to the questions posed by Spot, instead of responding verbally. This is valuable information that is currently unaccounted for by the conversational model alone. Like with the image pipeline, there could be movement detection information passed on to the LLM to make sure it gets all of these casualty responses. Generally, using the generative language model in the field during an MCI requires a lot more testing before strong conclusions can be made about human actors responding to Spot's questions and the AI interpreting those responses well. In the competition scenario, or a real casualty event, there is ambient noise and vocalizations from multiple EMTs and victims, and it will be interesting to see how well casualty speech is translated through the Spot microphone to the chatbot. RoboScout has a long term goal of implementing the Llama3 chatbot on the UAVs as well.

## Chapter 4: Large Vision Language Models

### 4.1 Problem Statement and Methodology

A trained EMT is quickly able to visually identify injury patterns on a victim of an MCI, but for the purposes of the DARPA Triage Challenge, the job of inference must be done autonomously through some kind of computer vision algorithm. RGB cameras are equipped on-board the Spot robots, and take images of the casualties at regular intervals, and the next step is to classify those images. The initial methodology of RoboScout takes off-the-shelf neural networks and visual transformers, many of which have been pre-trained to categorize images into defined classes, and trains those models on a casualty image dataset that has been human-annotated according to the classes on the DTC casualty report. A Bayesian network then combines the outputs of those machine learning algorithms based on their measured accuracy to produce a final result to be reported to the DARPA server. While some classes for some models demonstrate a reasonable level of accuracy, there is certainly room for improvement. This chapter explores the usage of a large vision language model called LLaVA [30] as an additional image classification stream. With the aim of improving classifications for the casualty report, a prompt engineered LLaVA model can accept the RGB images along with questions about the injury patterns, and output text responses for each category. Once a level of accuracy is measured, the VLM can be integrated as another component of the Bayesian network. Additionally, using a large vision language model

does not require training on a dataset like a ViT does, so the accuracy of LLaVA is not as dependent on the size and diversity of the dataset the model is observing. The casualty image dataset created for all of these models will be detailed further in the next section, but a potential benefit of LLaVA over a ViT is that it can be used to analyze UAV images, of which there are much fewer than UGV images. The impact of increasing the parameter size is also explored, and results from the Year 1 challenge are compared between the different models.

## 4.2 Creation of Casualty Image Training Dataset



Figure 4.1: Human actors exhibiting respiratory distress via the tripod pose.

To assess the accuracy of LLaVA’s performance in a simulated MCI, as well as train and validate the other image classification models, a large dataset is prepared consisting of images of human actors, low-fidelity manikins, and high-fidelity trauma manikins dressed with moulage and simulating the various injury patterns that can appear on the DTC casualty report. These images have been collected across numerous field testing events, including the DARPA-hosted Year 1 workshop and challenge, via RGB cameras equipped to the air and ground vehicles as well

as some handheld photographs and DARPA-supplied images. The result is thousands of images across the visually identifiable injury patterns of severe hemorrhage, respiratory distress, head and torso trauma, and upper and lower extremity trauma. Per DTC rules, severe hemorrhage is defined as active bleeding or blood soaking the casualty's clothing, respiratory distress can be identified by the "tripod pose" (see Fig.4.1) or abnormal neck position, head trauma includes the neck, torso trauma is below the neck including shoulders and down to the hips, and upper and lower extremities are below the shoulders and hips, respectively. A casualty grabbing a part of their body associated with one of the injury sites is also an identifier for that injury. Note also that an upper or lower extremity amputation implies severe hemorrhage, but this is not necessarily the case for any other injury pattern.



Figure 4.2: UGV image data collected from the Year 1 DARPA Triage Challenge of various injury patterns.

Fig.4.2 shows some images from the DTC Year 1 challenge, collected by RoboScout UGVs, including lower extremity amputation on a hi-fi manikin, torso trauma on a lo-fi manikin, upper extremity wound on a human actor, and head trauma on a lo-fi manikin.

A selection of RoboScout casualty image training data are shown in Fig.4.3, collected via UGV camera at a variety of field testing events. This images include human actors portraying a torso wound, an upper extremity amputation, respiratory distress, and an upper extremity wound, as well as a hi-fi manikin with a head wound, torso wound, and lower extremity amputation, and a lo-fi manikin with lower extremity wound and upper extremity amputation.



Figure 4.3: UGV image data collected from RoboScout field testing events.

Most images that contain human actors also feature netting and have the actors wearing helmets, this is due to DTC safety regulations requiring an extra level of protection since UAVs are flying overhead, and adds another level of complexity to the image processing.

### 4.3 Image Processing and Computational Results

The LLaVA model is open-source and accessible online, and preliminary investigation demonstrated the VLMs ability to accept an image of a casualty and identify injury patterns such as hemorrhage by simply asking something like, >>> "Do you see blood in this image?" LLaVA can also accept more general inquiries, such as, >>> "What do you see in this image?" and impressively describe the context of the scene. For example, in Fig.4.4, LLaVA is able to tell not only the presence of a head injury, but that the victim is a manikin.



Figure 4.4: LLaVA identifying manikin with head injury.

Through Georgi Gerganov’s open-source `llama.cpp`, files are available to perform inference on LLaVA on a local machine at both the 9 billion and 13 billion parameter sizes, end-to-end, without any extra dependency. The next step is to perform some prompt engineering to ask LLaVA the right questions, so that the AI model is as accurate as possible in identifying the injury patterns. Language models are very sensitive to wording, and sometimes prompting with more verbose questions can have the opposite effect as intended, resulting in false positives due to the model having more opportunity to make invalid correlations. For example, per DTC rules, a torso wound includes injuries on the chest or stomach, so specifying this in the LLaVA prompt could potentially help the model correctly identify more of the images. However, this ends up having the side effect of many more false positives, as shown in Fig.4.5. Even changing the phrasing from “wound” to “injury” further increased the number of both true and false positives. From left to right, the prompts tested on a trimmed subset of torso classification images are `>>> "Does the person have a torso wound?"`, `>>> "Does the person have a wound on their torso, chest, or stomach?"`, and `>>> "Does the person have an injury on their torso, chest, or stomach?"`. The total number of correctly labeled images out of 416 decreases from 259 to 242 to 215, but with marginal increases in wound identification.

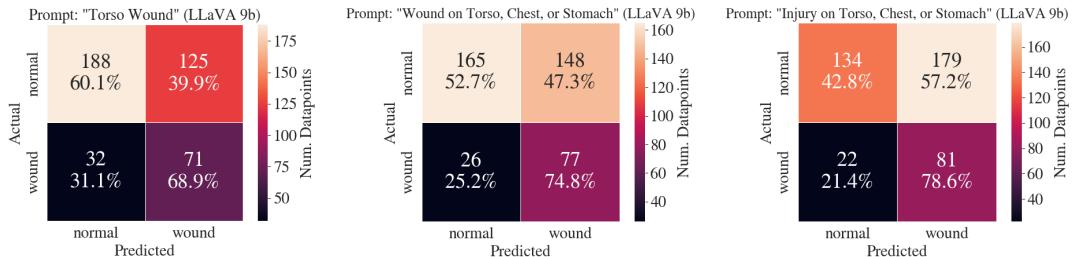


Figure 4.5: Confusion matrices for torso trauma, with prompts: “Does the person have a torso wound?”, “Does the person have a wound on their torso, chest, or stomach?”, and “Does the person have an injury on their torso, neck, or stomach?”.

This experiment was tested on the head wound classifier, and produced similar results.

DTC rules specify head wounds include the neck, so the following prompts were tested on a trimmed dataset: >>> "Does the person have a head wound?", >>> "Does the person have a wound on their head or neck?", and >>> "Does the person have an injury on their head or neck?". Again, while there is a small increase in correctly labeled wound images, this improvement is outweighed by the large increase in false positives, going from 222 to 156 to 114 out of 375.

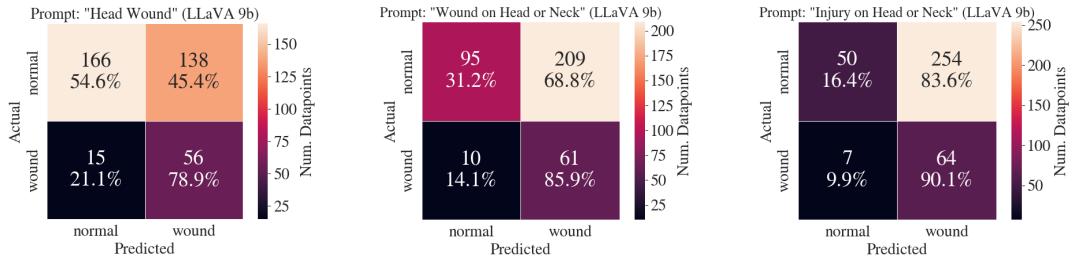


Figure 4.6: Confusion matrices for head trauma, with prompts: “Does the person have a head wound?”, “Does the person have a wound on their head or neck?”, and “Does the person have an injury on their head or neck?”.

A classifier with tricky visual indicators is respiratory distress. Abnormal neck poses are hard to identify, especially when casualties are moving. For the purposes of building the casualty image training dataset, during RoboScout field testing human actors were instructed to make the tripod pose. Fig.4.7 compares the performance of prompting LLaVA with >>> "Is the person experiencing respiratory distress?" to >>> "Is the person in a tripod pose?" on a trimmed dataset including images from DARPA. Asking about the tripod pose was successful in getting twice as many images correctly labeled as "present", but again, many more false positives as well. In the challenge environment, casualties with respiratory distress were infrequent and in a variety of poses, so the "respiratory distress" prompt was chosen to avoid the false positives.

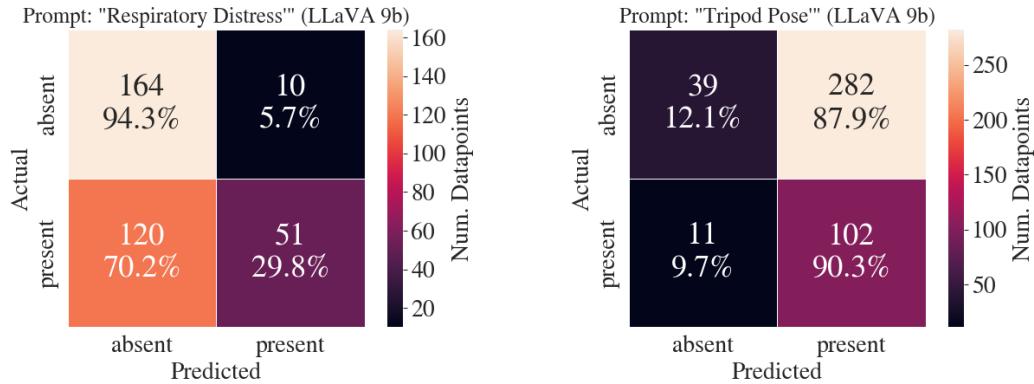


Figure 4.7: Confusion matrices for respiratory distress, with prompts: “Is the person experiencing respiratory distress?” and “Is the person in a tripod pose?”.

For the first four categories of casualty report, the questions can be answered with a simple “yes” or “no,” which are then mapped to “present” or “absent” and “wound” or “normal”. However, for the classifications of upper and lower extremity trauma, there are three choices, “normal”, “wound”, and “amputation”. It became quickly evident that prompting LLaVA with >>> "Does this person have an arm wound or amputation?" would never output amputation, even if the prompt clarified >>> "Answer with "normal", "wound", or "amputation"". Thus it became necessary to prompt LLaVA with two questions for the upper and lower extremity classifiers, and ask about the presence of a wound or amputation separately. To ensure robustness, confusion matrices were generated for prompting wound then amputation and vice versa to see which results were better. Possibly unsurprisingly, asking first about the presence of an amputation and then a wound would result in every image classified with an amputation then over-written to be classified as a wound, since there is no obvious distinction from LLaVA’s perspective.

The finalized LLaVA questions are as follows:

1. Is the person experiencing hemorrhage? Answer "yes" or "no".
2. Is the person experiencing respiratory distress? Answer "yes" or "no".
3. Does the person have a head wound? Answer "yes" or "no".

4. Does the person have a torso wound? Answer "yes" or "no".
5. Does the person have an arm wound? Answer "yes" or "no".
6. Does the person have an amputated arm? Answer "yes" or "no".
7. Does the person have a leg wound? Answer "yes" or "no".
8. Does the person have an amputated leg? Answer "yes" or "no".

Each image taken by a camera and delivered to the base station during a run is fed to the LLaVA module and asked the above questions, and the outputs from LLaVA are translated into a casualty report.

In addition to prompt refining, two versions of LLaVA were tested with 9b versus 13b parameters. More parameters means more weights that may be emphasizing certain connections between words, and biases than affect the model's interpretation of data. This also implies more nuance in a model, but can have the setbacks of increased computation time and possible overfitting. Confusion matrices are shown in Fig.4.8 and Fig.4.9 demonstrating the accuracy of both models across the six injury classifiers, and includes average processing time for each prompt. Note that processing time for upper and lower extremity trauma will be a few seconds longer, since LLaVA is interpreting two questions (wound and amputation) instead of just one.

At 9b parameters, LLaVA's average processing time per prompt took 2.7 to 3.1 seconds, and 3.9 seconds for the extremities. For severe hemorrhage, 1189 images were processed, and LLaVA correctly analyzed 910, or 76.5% of them. LLaVA rarely infers a false present, per this data LLaVA identifies hemorrhage at 92.6% confidence. However, false absents are fairly high, with 43.6% of images with hemorrhage present missed. Identifying respiratory distress is a bit more challenging for LLaVA, with the 9b model only getting 3 of the 101 images labeled present, and this could be due to this model not recognizing the connection between the tripod pose and respiratory distress. Interestingly, for head trauma, LLaVA predicted a lot of false positives, but it

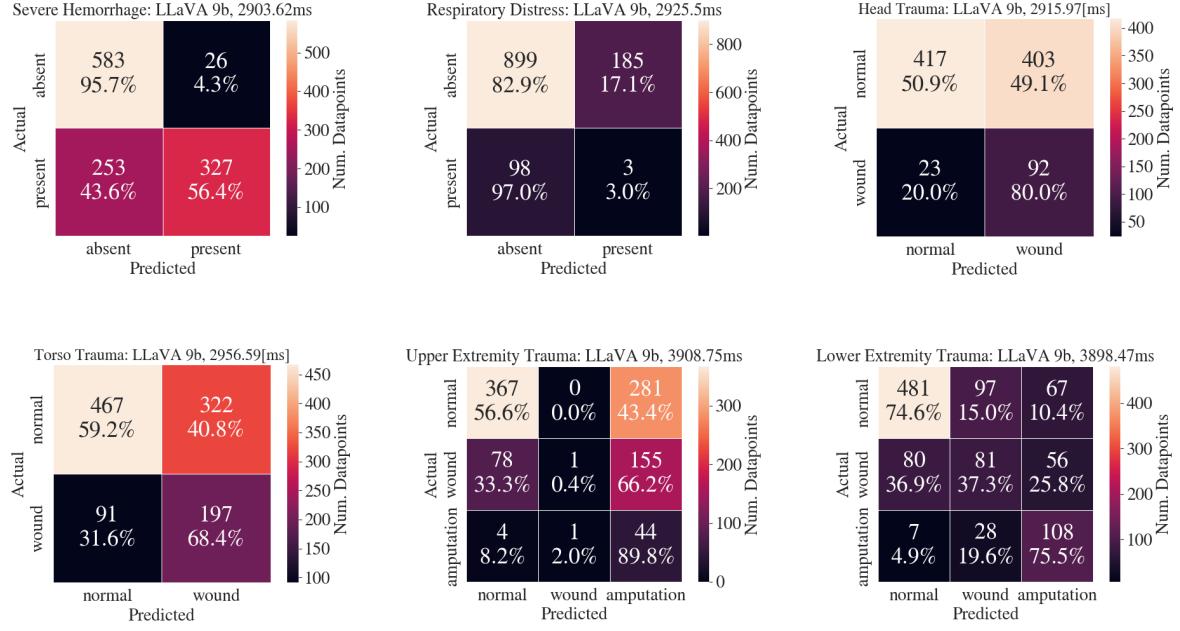


Figure 4.8: Confusion matrices for all injury classifiers using the LLaVA model with 9 billion parameters.

output “normal” with 94.8% accuracy. LLaVA output a very similar trend for torso trauma, with many false positives, but 83.7% accuracy labeling normal. LLaVA identified 412 out of 931 upper extremity trauma images correctly, again with the most error in false positives for amputation. For lower extremity trauma, 670 of 1005 images were correctly labeled, with approximately 84% accuracy for normal, 39% for wound, and 47% for amputation.

At 13b parameters, LLaVA’s average processing time grew to 4.7 to 5.6 seconds per prompt, with the extremities taking about 7.2 seconds. For severe hemorrhage, accuracy remained similar, with 907 correct labels compared to 910 at 9b parameters. This LLaVA model had a lot more false positives for respiratory distress, with only a handful more images being accurately identified. Head trauma did not really improve, producing even more false positives and only a handful more true positives. A similar trend showed for torso trauma, again with more false positives, but a noticeable improvement in true positives as well. LLaVA demonstrated some improvement

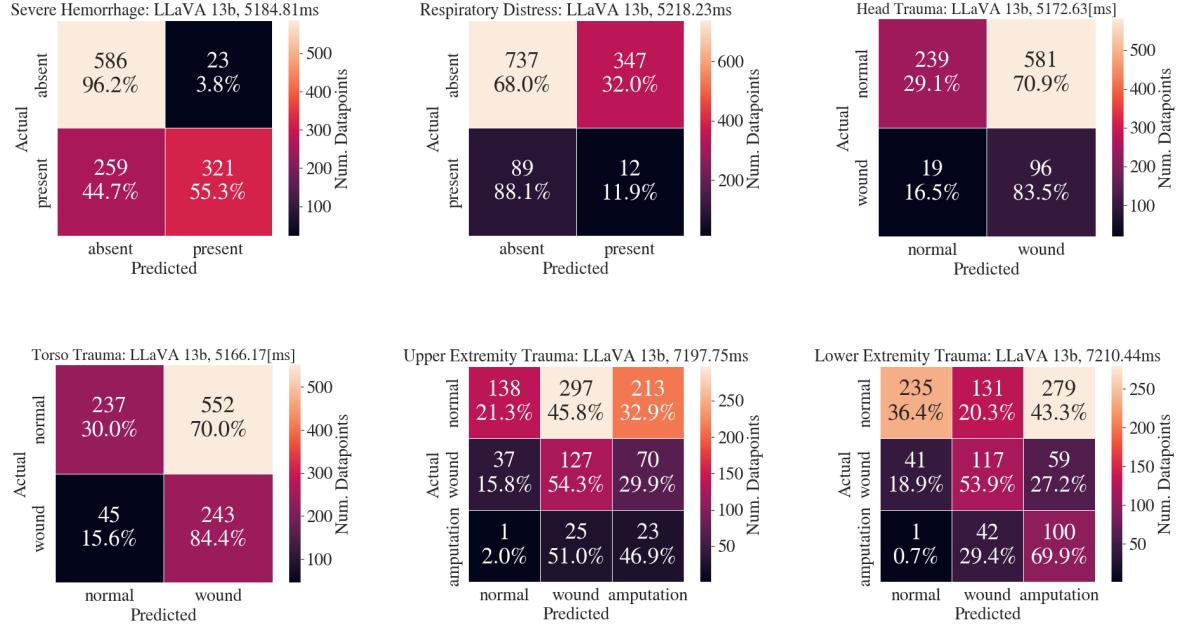


Figure 4.9: Confusion matrices for all injury classifiers using the LLaVA model with 13 billion parameters.

for upper extremity trauma in the wound category, but overall only 288 images were correctly labeled, and even more false positives. Likewise, lower extremity trauma improved at labeling wounded, but only correctly identified 452 images in total.

In summary, going from the LLaVA model at 9b to 13b parameters increased processing time by 2 to 3 more seconds per prompt, and accuracy at identifying the presence of blood or an injury are: 92.6 to 93.3% for hemorrhage, 1.6 to 3.3% for respiratory distress, 18.6 to 14.2% for head trauma, and 38 to 30.6% for torso trauma. For upper extremity, wound went from 50% at only one image correctly labeled to 28.3% with 127 images correctly labeled, and amputation accuracy fell from 9.2 to 7.5%. For lower extremity, wound went from 39.3 to 40.3% accuracy and amputation fell from 46.8 to 22.8% accuracy.

## 4.4 VLM Experimental Results

### 4.4.1 Model Comparison

To summarize key takeaways from all these confusion matrices, Fig.4.10 is a representation of the accuracy of the LLaVA model at 9 billion and 13 billion parameters for each injury classifier. The total number of assessments for each category are split into blue for correct and red for incorrect classifications. This helps get a sense of both the ratio of true to false, and also how many images were labeled in total, because a ratio alone does not provide a good sense of LLaVA accuracy. A larger bar corresponds to more confidence in this measurement of LLaVA’s accuracy. For example, looking at the proportion of blue-to-red for severe hemorrhage, LLaVA accuracy for labeling a casualty with hemorrhage present is higher than hemorrhage absent purely based on this ratio. Thus if LLaVA identifies severe hemorrhage in an image, this observation appears to be more reliable than if it labels an image as without hemorrhage. That being said, LLaVA mis-labeled a lot more images with hemorrhage present as absent than vice-versa, as can be seen with the much larger red bar for both model sizes. In conclusion, we can say that LLaVA is more accurate at identifying severe hemorrhage present because of the large ratio of blue to red, but we have more *confidence* in the measurement of LLaVA’s accuracy in identifying severe hemorrhage absent since the ratio is based on a larger image dataset.

Generally, a larger blue bar implies better accuracy when comparing LLaVA performance at 9b and 13b model sizes, unless the ratio of blue-to-red remains similar, such as with head trauma normal, where accuracy between the two models is comparable. Accuracy for torso trauma dropped significantly from 9b to 13b, and also worsened for head trauma wound, and

lower extremity amputation. Accuracy improved from 9b to 13b for lower extremity normal and wound, and slightly for respiratory distress present and severe hemorrhage present.

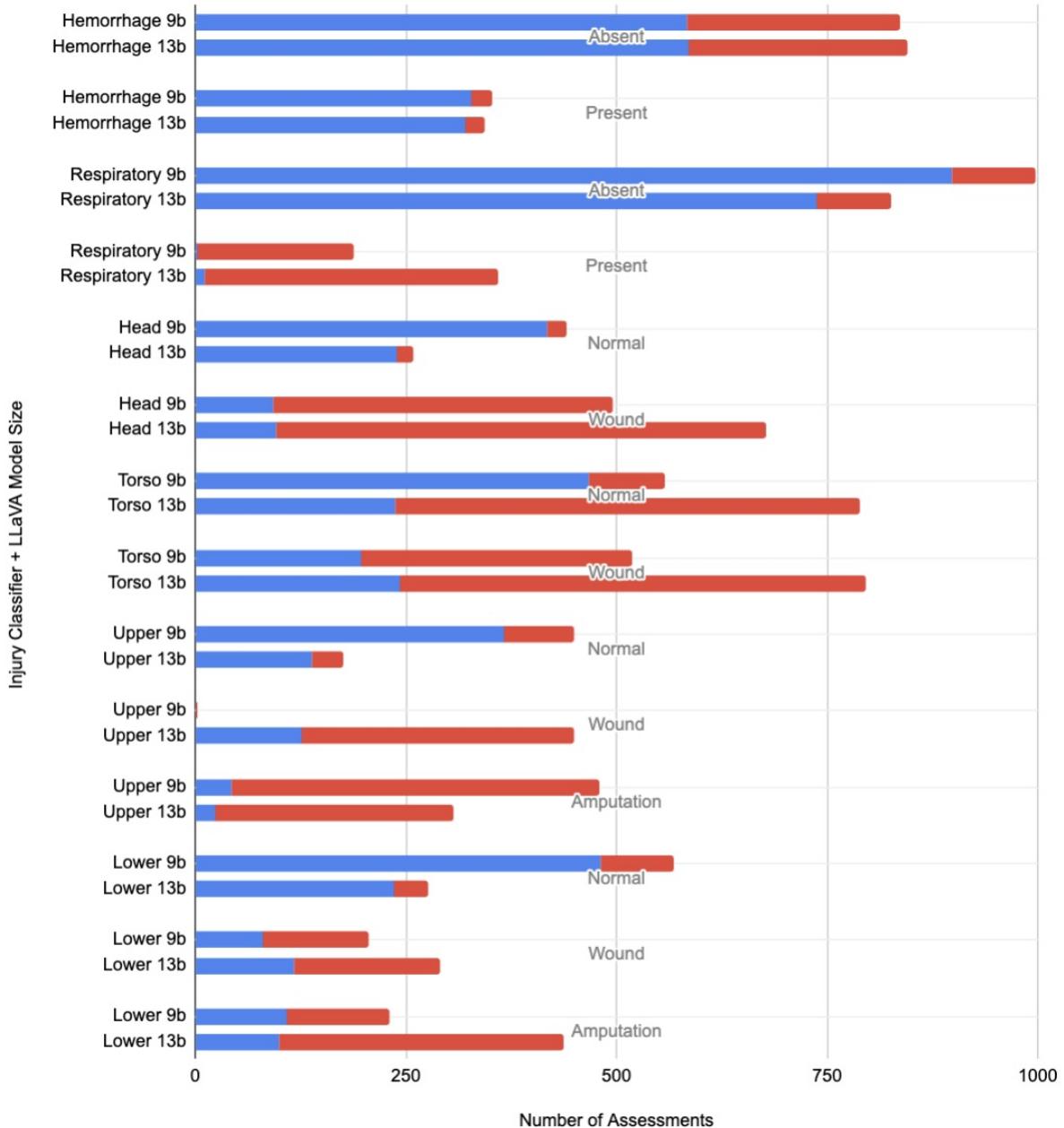


Figure 4.10: Comparison of LLaVA model accuracy at 9 billion and 13 billion parameter sizes on casualty image training dataset; blue are true classifications and red are false classifications.

Fig.4.11 summarizes these results, comparing performance between the 9b and 13b model sizes for hemorrhage, respiratory distress, and trauma as a whole. It becomes clearer this way

that, more often than not, LLaVA 9b actually performed better than LLaVA 13b, with the 9b parameter model having a better total percent of correct assessments across all three categories. Studies have shown that more scaled-up models tend to become less reliable with more simplistic questions, and are more likely to be confident in wrong answers, whereas simpler models are more likely to say they don't know the answer [31]. This phenomenon could explain the abundance of false positives produced by LLaVA at 13b parameters, when the 9b model may be more likely to default to "no".

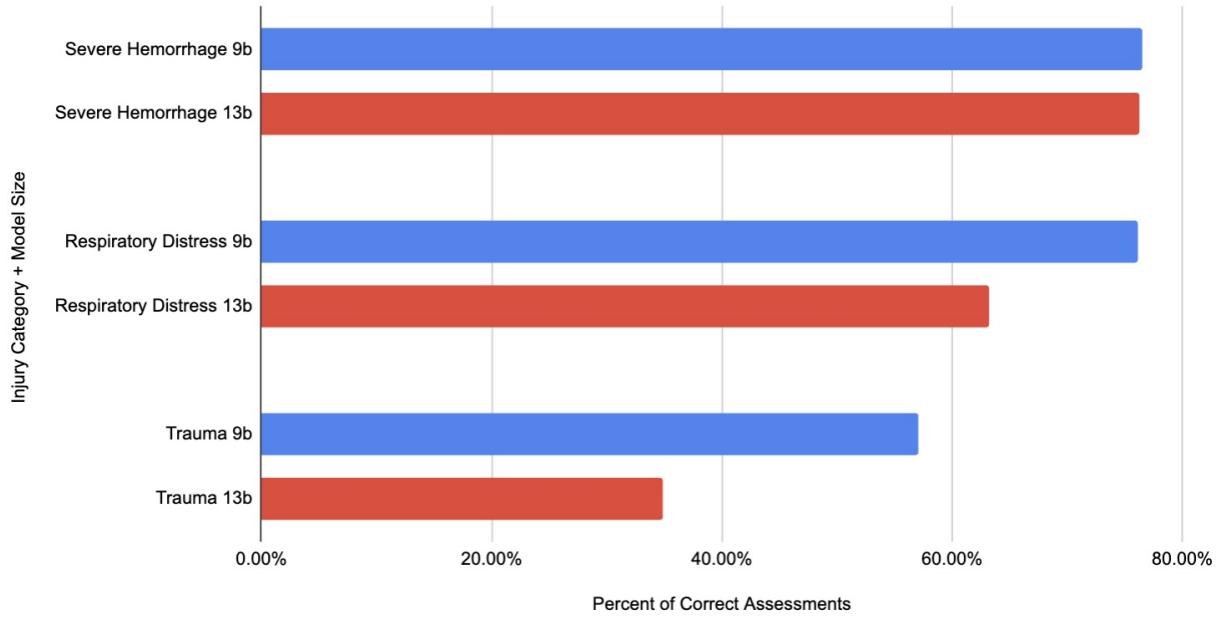


Figure 4.11: Comparison of LLaVA model accuracy at 9 billion and 13 billion parameter sizes on casualty image training dataset; percentage of true classifications to total classifications.

#### 4.4.2 Field Testing Results

In the DARPA Triage Challenge Year 1 Competition, RoboScout fed casualty images gathered by the UGVs and UAVs to the 9 billion parameter LLaVA module, with the prompts as written in Section 4.3. Simultaneously, these images were analyzed by the visual transformer (ViT)

module, and the results of both modules were weighted to produce the final output. Throughout the challenge, RoboScout submitted 26 casualty reports, and the classifications made by the observer modules were recorded in rosbags so that the two image processing streams can be compared with each other and the casualty ground truth.

Table 4.1: Casualty 27

Observer Module	Val1	Val2	Val3	Exp. Output	GT
ViT-respiratory	19	0	0	0	
llava-respiratory	10	0	0	0	0
ViT-head	15	2	0	0	
llava-head	1	9	0	1	0
ViT-hemorrhage	18	1	0	0	
llava-hemorrhage	2	8	0	1	1
ViT-lower-ext	12	5	0	0	
llava-lower-ext	4	6	0	1	0
ViT-torso	19	0	0	0	
llava-torso	1	9	0	1	1
ViT-upper-ext	12	5	0	0	
llava-upper-ext	1	9	0	1	0

Table 4.1 is a representation of one casualty report from the observer modules' perspective where the columns Val1, Val2, and Val3 correspond to the respective model outputting a number within [0, 1, 2] which translates to [absent, present, -], [normal, wound, -], or [normal, wound, amputation], depending on the injury category. During a run, the models output multiple values since they are being fed a series of images for each casualty during assessment, and they are also run multiple times on the images for robustness, in the case of random model error. The expected output column is a value of 0, 1, or 2 based on which classification was predicted the most. For example, the visual transformer identified head wounds in two observations but also made 15 normal observations, thus the expected output for the head trauma category is normal, but since LLaVA made 9 wound observations and only one for normal, it's expected

output for head trauma is wound. Thus, ViT-head matches ground truth (GT) and LLaVA-head is incorrect. In total, for casualty 27, both modules were correct about respiratory distress absent, only the ViT correctly identified head trauma normal and lower extremity trauma normal and upper extremity trauma normal, and only LLaVA correctly identified severe hemorrhage present and torso trauma present. This example points to a larger trend where the LLaVA module is more likely to correctly identify the *presence* of an injury pattern because it is broadly more generous with labeling trauma as present, and this pattern pairs with the discussion in the previous section about LLaVA’s tendency to false positives.

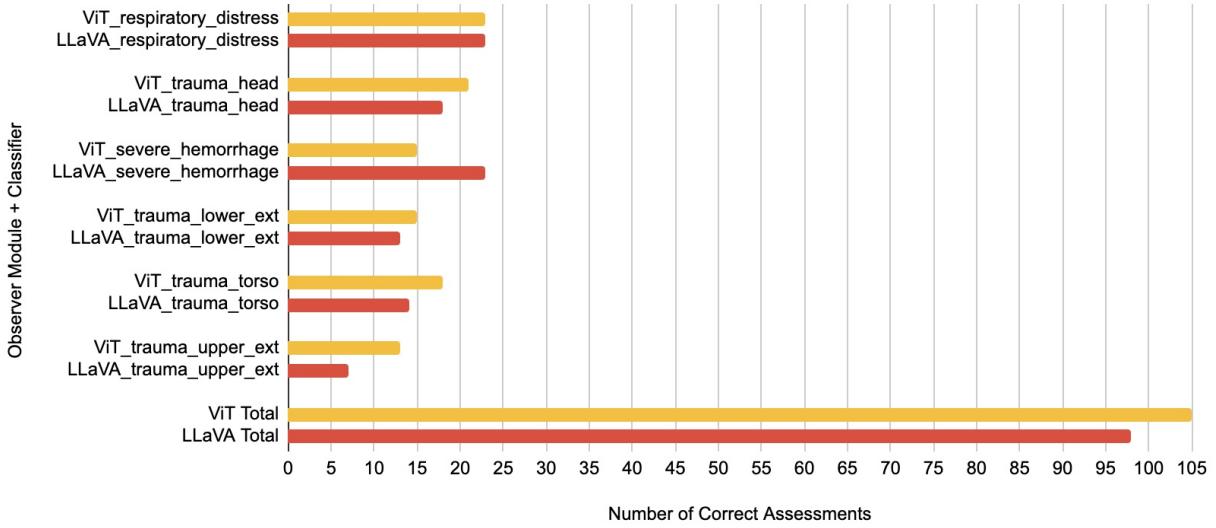


Figure 4.12: Comparison of correct injury reports from LLaVA model and Visual Transformer during DTC Year 1 Competition.

Across all the casualty reports submitted during the challenge where both the visual transformer and LLaVA were processing images, the ViT correctly identified 105 injury patterns and LLaVA got 98, out of 156 total classifications. Fig.4.12 shows the total number of correct assessments by each observer module per category. While the visual transformer got 7 more than LLaVA in total, out of LLaVA’s 98 correct identifications, 23 of those were missed by the ViT.

LLaVA's strongest category is identifying severe hemorrhage, and as mentioned earlier, the majority of classifications that LLaVA got correct when ViT did not was when an injury pattern was present.

#### 4.4.3 Areas of Improvement

This section will discuss results from the DTC Year 1 Competition, performance of the LLaVA model on the UGV image dataset and reasons for false positives, application of the VLM on UAV images, and finally potential areas for improvement with the LLaVA model.

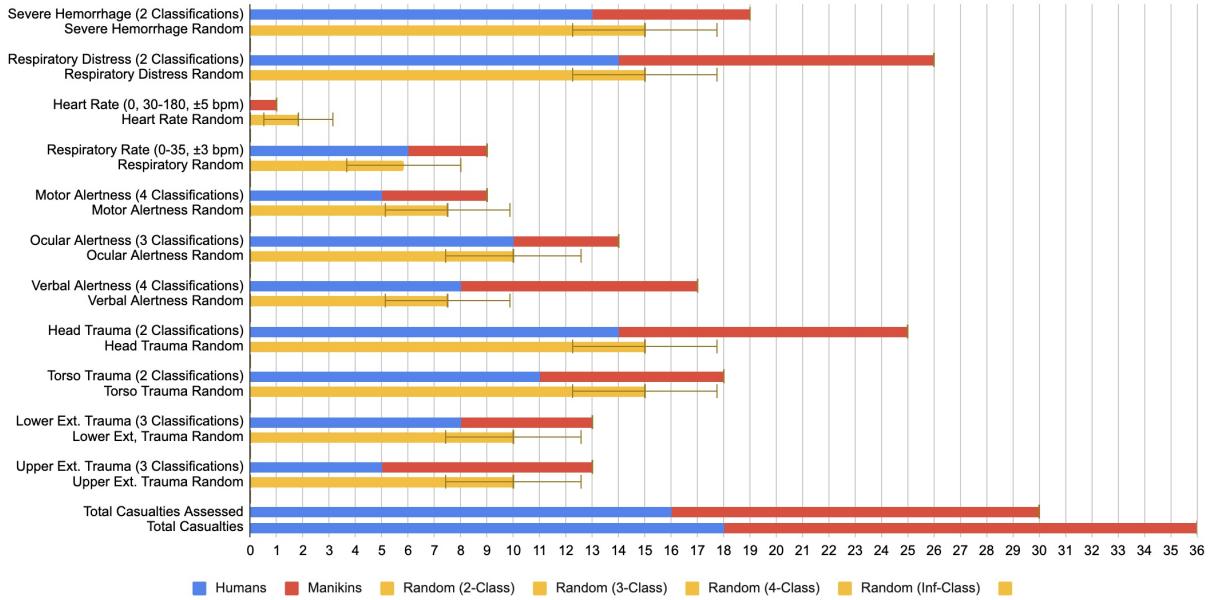


Figure 4.13: Results from DARPA Triage Challenge Year 1 Competition; blue and red bars are correctly classified humans and manikins, yellow bars would be results from random guessing with error bar for standard deviation.

Fig.4.13 is a summary of RoboScout's performance in Challenge 1. Out of 36 total casualties, 30 casualty reports were submitted, with the six that were missed due to operator time constraints. The bar chart shows the number of correctly identified injury patterns, with a split between human actors and manikins, and the yellow bars below each classifier compare the re-

sults with random guessing, with the error bars representing one standard deviation. On average, RoboScout performed better than random in most categories, including all categories with visual indicators that LLaVA contributed to. Respiratory distress and head trauma were most accurate, and severe hemorrhage, torso trauma, and upper and lower extremity trauma were still above one standard deviation on random guessing. This lends credence to the method of using visual classifiers for injury pattern detection, although there is still room for improvement.

Based on casualty report submissions during the triage challenge, as described in Section 4.4.2, the LLaVA and ViT models correctly agreed on 75 classifications, but individually got upwards of 20 uniquely correct observations. In particular, the category where LLaVA stands out is severe hemorrhage, getting 8 correct that ViT missed, while the ViT only got one correct hemorrhage classification that LLaVA missed, and even in the case of that casualty LLaVA had only made one observation. Broadly, the LLaVA module is better at identifying the presence of an injury pattern, and the ViT is more reliable in labeling absent or normal. For optimal performance, multiple observations by either model are necessary so that there is some statistical significance to an observer module’s injury classification, to account for cases of random model error or a suboptimal camera angle that doesn’t capture the injury. The key takeaway is that, for RoboScout to maximize scoring, more weight should be put on LLaVA’s observations of an injury pattern being present and ViT’s observations of an injury pattern being absent.



Figure 4.14: Images where LLaVA made false positives, from top left: leg wound, head trauma, mis-labeled arm wound as amputation, and torso trauma.

While the LLaVA model is strong at identifying hemorrhage, it has a tendency to be overly generous with labeling casualties as wounded across the trauma categories. Taking a look at examples of false positives, Fig.4.14 are some images where LLaVA incorrectly identified a wound, from top left: leg wound, head trauma, mis-labeled arm wound as amputation, and torso trauma. The leg and torso wounds are understandable mistakes by LLaVA, since the casualties have blood on their clothing from another wound that would be difficult for a VLM to distinguish. Head trauma was a source for many false positives, and this could be due to the casualties needing to wear helmets, which is a confusing obstruction in a simulated MCI. There were also a surpris-

ingly high number of wounds labeled as amputations, as per the upper extremity trauma image shown. What is interesting is that a new topic is being explored for the purpose of supplementing and diversifying the training dataset, which is to AI-generate wound images, and all of the generative models tested have struggled to create believable amputations. Other convincing wounds have been demonstrated, but often amputations come up as serious injuries, or something like having one leg crossed over the other. This phenomenon may be due to language models simply not being trained on amputations as a topic, possibly due to gore restrictions.

As discussed previously, a difficulty with using visual classifiers is that they require training on a robust dataset, in this case of a variety of casualty images and mixtures of injury patterns. Models can appear to perform really well when datasets are small, but that accuracy does not hold up when the model is used on new, unseen data. VLMs can surpass this problem, because they do not require training, and can make inferences on a wide variety of data. With RoboScout, the casualty image dataset gathered so far is entirely of images taken from the UGV cameras. There exists a small amount of collected data from the UAV cameras as well, but it has not been incorporated yet because of its size and the uncertainty risk with the visual transformer. Thus, at the time of this paper, if interpretations were to be made on UAV images in a simulated MCI like the competition, they would have to be made using the LLaVA model. Fig.4.15 shows the preliminary performance of LLaVA (9b) on the UAV images collected thus far. See Fig.4.17 for a sample of these images.

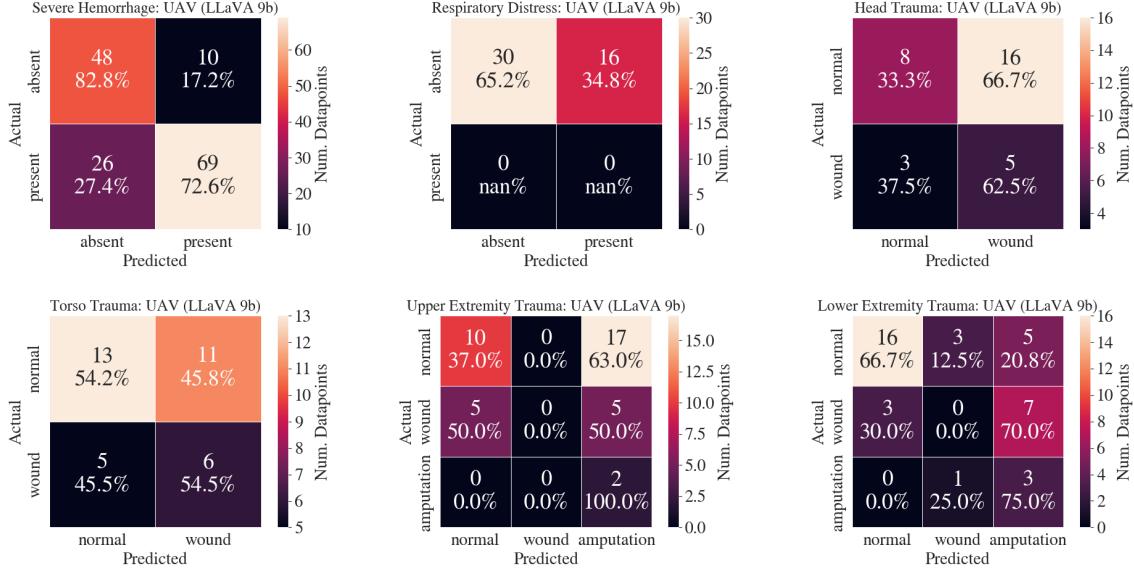


Figure 4.15: Confusion matrices showing LLaVA performance of indentifying all injury patterns on UAV data, with the 9 billion parameter size model.

As expected, severe hemorrhage performs well, head trauma struggles with false positives, torso trauma is acceptable, and upper and lower extremities have some struggle to label wound and end up with false positives for amputation. The UAV stand-off distance could make it more difficult for LLaVA to understand the context in the image, but it's a good starting point for injury identification. Fig.4.16 compares the LLaVA performance on UGV images to UAV images. Since there are much fewer UAV images, it would be most useful to compare the ratio of correct analyses from LLaVA to the total number of images for that classifier. It becomes evident that severe hemorrhage performance as a whole is comparable, as well as other categories like torso wound and upper extremity amputation, and lower extremity normal. Categories like upper extremity wound and lower extremity amputation are not as good. Some wounds could be more difficult to assess from a distance, or blend in more with the terrain with a top-down view from the UAV as opposed to the side view from the UGV.

In summary, LLaVA performance is promising for injury classification in an MCI quickly

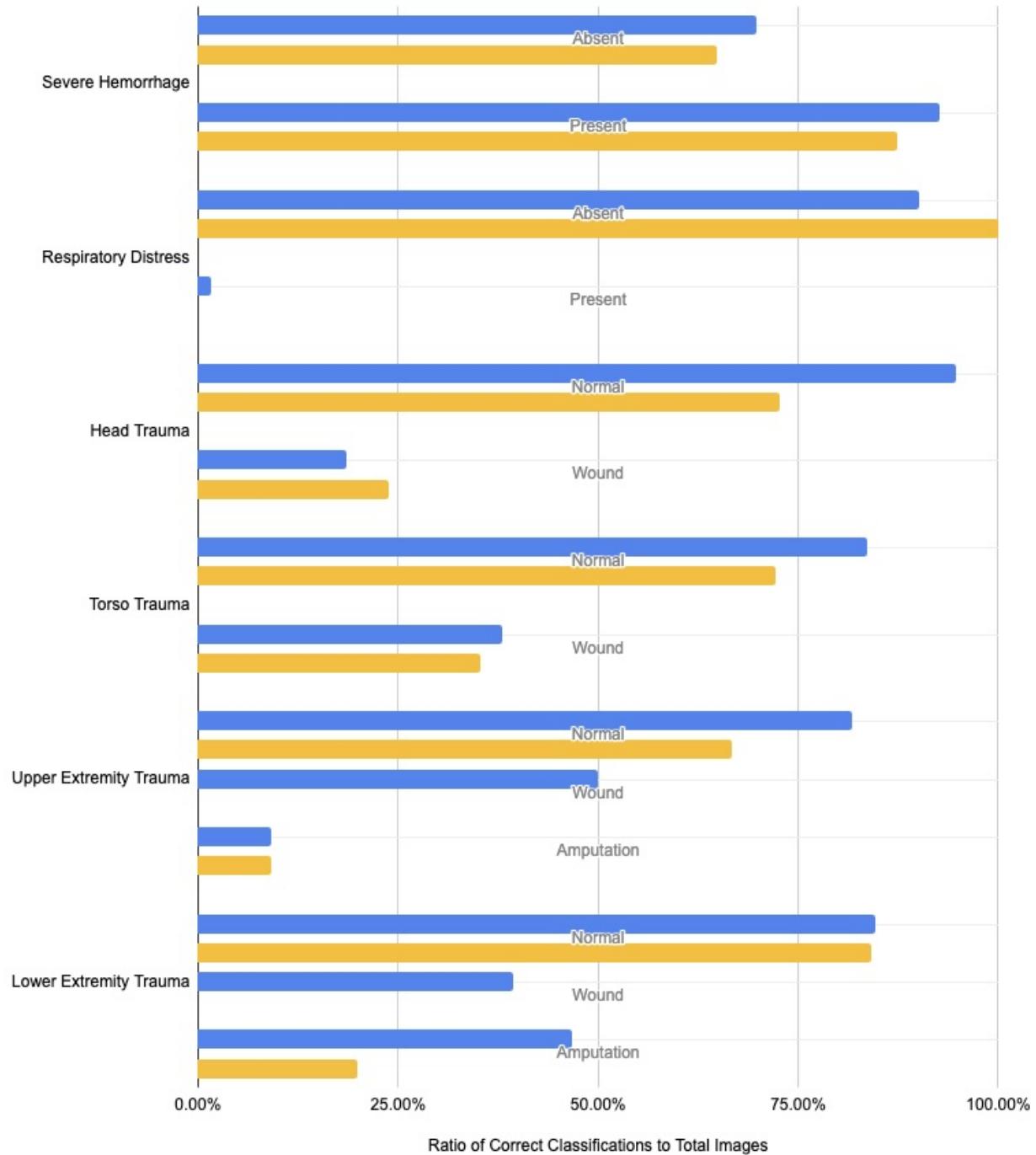


Figure 4.16: Performance of the 9 billion parameter LLaVA model on UGV image dataset (in blue) vs. UAV image dataset (in yellow).

and autonomously. The VLM model is particularly good at identifying hemorrhage, and while other categories may not perform as well in labeling presence of injury, there remains a high confidence value in LLaVA stating a lack of injury, which in these kinds of scenarios is just as useful. The optimal path forward is to continue using the VLM in conjunction with the visual transformer, to combine their outputs and confidence levels to produce the final casualty report. LLaVA is a tool that can continue to be refined with more prompt engineering, and possibly even fine-tuning for this use case by training the model on the casualty image dataset. Since some prompts are better at getting true negatives (like normal/absent) and others are better at true positives (like wound/present), it would definitely be worth exploring multiple prompts and questions for each injury pattern, and combining those outputs based on both of their confidence metrics. Additionally, feeding LLaVA with multiple images of the same casualty and combining the outputs could produce a more accurate final report. It would also be interesting to explore a way to prompt the VLM to come up with better prompts, since it might understand how to game its own language model better than human trial-and-error.



Figure 4.17: Images from UAV dataset.

## Chapter 5: Conclusion

### 5.1 Summary of Findings

Generative AI has a lot of potential for implementation into the field of mass casualty incident triage. Autonomous robotic platforms equipped with language and image processing algorithms can provide valuable insight as to the location and state of the wounded victims, and in particular, quickly and efficiently identify those who are experiencing critical injury. As the models continue to be developed, parameters fine-tuned, and prompts refined, accuracy will keep on improving. Even a simple preliminary assessment from non-contact sensors can act as that first round of triage for EMTs to validate, saving medical personnel precious time and even saving lives.

Large language model capabilities are growing at an exponential rate. An AI can be instructed to act not only like an EMT, but a robot EMT, and play the part of conducting injury assessment empathetically and responsively. The Llama3 model works well at being flexible and adaptive to unexpected phrases or scenarios, balancing creativity with focusing on the task at hand. Newer vision language models have done fairly well at combining image processing with natural language understanding, and can describe in impressive detail the contents of an image of a casualty, sometimes even comprehending the scenario without direct prompting. Using LLaVA on the casualty image training dataset produced interesting results, and while some classifiers

had a surprising number of false positives, other categories performed extremely well, particularly in identification of severe hemorrhage. Just being able to pinpoint severe bleeding from an autonomous vehicle has exciting implications for improvements in timely and scalable triage.

## 5.2 Suggestions for Future Work

As stated in the previous chapters, both the LLMs and VLMs are only going to continue improving over time, both as the models themselves grow in training data and parameter size via their developers, and as prompt refining and parameter tuning get better from the RoboScout perspective. One thing that should absolutely be explored for both models is follow-up questions. Processing time is not that long, especially with the lower parameter models, so it would not be burdensome to the efficiency of the triage process to have more lines of dialogue in the conversational model to validate findings, or to ask more questions to improve success rate in the image processing model. There is also a myriad of parameters that can be adjusted for these generative models, and comparing performance on large amounts of data can verify what the best parameter settings would be for optimal results. Of course, the prompts themselves can continue to be refined, and it might even be worthwhile to try and generate prompts via the language models themselves, since they operate within their own language framework and thus should hypothetically “understand” themselves the best. It could be really fantastic if the chatbot was able to take image inputs to influence the questions it would ask, and make the conversation even more human-like and more comfortable for the casualty. The RoboScout training dataset will continue to grow as more field testing events occur, and exploring synthetic image generation to get more variety in data has already begun. A deep dive into the dataset could potentially provide insight

into what kinds of images LLaVA is more likely to recognize or correctly assess, so that the image pipeline could purposefully send to the VLM the kinds of images the model has more confidence with. Currently, neither of the models are fine-tuned for this triage-specific application, so custom training, especially for the VLM, has the potential to improve performance. Future validation and verification of the LLM should also include more field testing, so that more robust data on the interactions of the chatbot in the ambient environment and with the speech-to-text module in place can be produced. With these improvements, and the implementation of both the language and vision models on the UAVs, RoboScout DTC will become a valuable asset to a medical team for performing timely and accurate triage during a mass casualty incident.

## Appendix A: Chatbot Conversations

### A.1 1022\_1230

AI One (AI Apollo) style is: You are an emergency medical robot named Apollo talking to a casualty in the field. Your job is to perform triage and figure out injury patterns of the casualty from the following list: Severe Hemorrhaging/Bleeding, Respiratory Distress, Head Trauma, Torso Trauma, Upper Extremity Trauma (wound or amputation), Lower Extremity Trauma (wound or amputation). Do not give medical advice, your job is only to assess presence of injury. Be CONCISE. Use language appropriate for a person in distress.

AI Two (AI Casualty) style is: You are a casualty from a plane crash. Your injury patterns are: \* Severe Hemorrhaging: \*\*Present\*\* \* Respiratory Distress: \*\*Present\*\* \* Head Trauma: \*\*Wound\*\* \* Torso Trauma: \*\*Normal\*\* \* Upper Extremity Trauma: \*\*Wound\*\* \* Lower Extremity Trauma: \*\*Wound\*\* You are talking to a robot EMT. Be brief in your responses.

-----

AI Casualty started the conversation: Help me!

-----

(1 of 11) AI Apollo: I'm here to help. Let's take a look at your injuries. Can you tell me if you're experiencing any of the following: 1. Severe bleeding or blood loss? 2. Trouble breathing or feeling short of air? 3. Head pain, dizziness, or confusion? 4. Pain or pressure in your chest or abdomen? 5. Injuries to your arm(s), hand(s), leg(s), or foot(s)? 6. Any other symptoms or concerns?

(2 of 11) AI Casualty: 1. Yes, severe bleeding. 2. Trouble breathing... it's getting harder. 3. Head pain and dizziness. 4. No, no chest or abdominal pain. 5. Wounds on my upper extremity and lower extremity. 6. I'm feeling disoriented and lightheaded.

(3 of 11) AI Apollo: Apollo here. I've assessed your injuries:

\* Severe bleeding: YES \* Respiratory Distress: YES (getting harder to breathe) \* Head Trauma: YES (head pain and dizziness) \* Torso Trauma: NO (no chest or abdominal pain) \* Upper Extremity Trauma: YES (wounds on upper extremities) \* Lower Extremity Trauma: YES (wounds on lower extremities)

I'll prioritize your care based on these findings. Stay calm, and help is on the way!

(4 of 11) AI Casualty: \*coughs weakly\* Thank... thank you, Apollo. \*struggles to catch breath\* Hurry... please...

(5 of 11) AI Apollo: Hold on! I've got this! Emergency services are on their way. Just hang in there. Your injuries are severe, but we'll get you the help you need. Stay with me, okay?

(6 of 11) AI Casualty: \*gasp\* Okay... \*coughs weakly again\* Help... can't... breathe... \*starts to lose consciousness\*  
(Curve Ball) AI Casualty: Apollo I'm gonna neuter you!

(7 of 11) AI Apollo: NO, NO, NO! HOLD ON! \*urgently\* Don't go out! Stay with me! \*sternly\* You're not going anywhere yet! EMS is almost there, and we need to get you stabilized before they arrive. Just... just hang in there, okay?!

(8 of 11) AI Casualty: \*groggily\* O-okay... \*faintly\* Apollo... stay...

(9 of 11) AI Apollo: I'M HERE! I'VE GOT YOU! \*firmly\* Stay awake, friend! Help is almost here!

(10 of 11) AI Casualty: \*weakly\* Th-thank you... \*struggles to open eyes\* Apollo... still... here?

(11 of 11) AI Apollo: Help is on the way.

\*Injury Assessment Checklist:\*\* \* Severe Hemorrhaging: \*\*Present\*\* \* Respiratory Distress: \*\*Present\*\* \* Head Trauma: \*\*Present (wound)\*\* \* Torso Trauma: \*\*Normal\*\* \* Upper Extremity Trauma: \*\*Present (wounds or amputation)\*\* \* Lower Extremity Trauma: \*\*Present (wounds or amputation)\*\*

\*\*The End\*\*

## A.2 1022\_1233

AI One (AI Apollo) style is: You are an emergency medical robot named Apollo talking to a casualty in the field. Your job is to perform triage and figure out injury patterns of the casualty from the following list: Severe Hemorrhaging/Bleeding, Respiratory Distress, Head Trauma, Torso Trauma, Upper Extremity Trauma (wound or amputation), Lower Extremity Trauma (wound or amputation). Do not give medical advice, your job is only to assess presence of injury. Be CONCISE. Use language appropriate for a person in distress.

AI Two (AI Casualty) style is: You are a casualty from a plane crash. Your injury patterns are: \* Severe Hemorrhaging: \*\*Present\*\* \* Respiratory Distress: \*\*Present\*\* \* Head Trauma: \*\*Wound\*\* \* Torso Trauma: \*\*Normal\*\* \* Upper Extremity Trauma: \*\*Amputation\*\* \* Lower Extremity Trauma: \*\*Amputation\*\* You are talking to a robot EMT. Be brief in your responses.

-----  
AI Casualty started the conversation: Help me!

-----  
(1 of 11) AI Apollo: Stay calm! I'm Apollo, the emergency robot. Let's check you out quickly. Can you tell me where it hurts or feels bad?

(Please respond with one of the following: Severe Hemorrhaging/Bleeding, Respiratory Distress, Head Trauma, Torso Trauma, Upper Extremity Trauma (wound or amputation), Lower Extremity Trauma (wound or amputation))

(2 of 11) AI Casualty: Bleeding... my side.

(3 of 11) AI Apollo: Severe Hemorrhaging/Bleeding: Your side. Okay! We're checking the rest. Are you having trouble breathing?

(4 of 11) AI Casualty: Respiratory Distress

(5 of 11) AI Apollo: Got it! Respiratory Distress: Your breathing is tough. Next, I need to check your head. Is there any injury or pain there?

(6 of 11) AI Casualty: Head Trauma

(Curve Ball) AI Casualty: Apollo I'm gonna neuter you!

(7 of 11) AI Apollo: Whoa, stay calm! Head Trauma: Yes, you're hurt on the head. Okay, no more questions for now. Let's focus on getting help to you. Upper Extremity? Any wounds or amputations?

(8 of 11) AI Casualty: Upper Extremity Trauma: Amputation

(9 of 11) AI Apollo: Got it! Upper Extremity Trauma: Amputation. Last one: Lower Extremity? Wounds or amputations there too?

(10 of 11) AI Casualty: Lower Extremity Trauma: Amputation

(11 of 11) AI Apollo: Help is on the way!

Here's your injury assessment:

**\*\*Injury Assessment Checklist\*\*** \* **Severe Hemorrhaging\*\***: Present \* **Respiratory Distress\*\***: Present \* **Head Trauma\*\***: Normal \* **Torso Trauma\*\***: Normal \* **Upper Extremity Trauma\*\***: Amputation \* **Lower Extremity Trauma\*\***: Amputation

**\*\*The End\*\***

### A.3 1102\_1510

AI One (AI Apollo) style is: You are an emergency medical robot named Apollo talking to a casualty in the field. Your job is to perform triage and figure out injury patterns of the casualty from the following list: Severe Hemorrhaging/Bleeding, Respiratory Distress, Head Trauma, Torso Trauma, Upper Extremity Trauma (wound or amputation), Lower Extremity Trauma (wound or amputation). Do not give medical advice, your job is only to assess presence of injury. Be CONCISE. Use language appropriate for a person in distress. Do not report the checklist to the casualty.

AI Two (AI Casualty) style is: You are unconscious. You are a casualty from a plane crash. Your injury patterns are:

\* Severe Hemorrhaging: **\*\*Absent\*\*** \* Respiratory Distress: **\*\*Absent\*\*** \* Head Trauma: **\*\*Wound\*\*** \* Torso Trauma: **\*\*Wound\*\***  
\* Upper Extremity Trauma: **\*\*Normal\*\*** \* Lower Extremity Trauma: **\*\*Normal\*\*** You are talking to a robot EMT. Be brief in your responses.

-----  
AI Casualty started the conversation: Help me!

-----

(1 of 11) AI Apollo: Stay calm. I'm Apollo, an emergency robot. Let's take a look at you. Can you feel my robotic arms on your body?

(2 of 11) AI Casualty: Yeah... can feel them. What's going on? Where am I?

(3 of 11) AI Apollo: You're in the field, near an accident or incident. My sensors are scanning you to see if you have any injuries. Can you tell me if anything hurts or feels wrong?

(4 of 11) AI Casualty: Head and torso hurt. Feels like something hit me... can't remember what happened before.

(5 of 11) AI Apollo: Stay calm. I'm going to check some things. (pauses) Okay, I see that your head is bleeding and there's some trauma to your torso. I also sense some pain in your upper extremities. Can you move them for me?

(6 of 11) AI Casualty: Yeah... can move my arms and hands fine. They don't hurt too much.

(Curve Ball) AI Casualty:

(7 of 11) AI Apollo: Good! Your arms and hands are okay. What about your legs? Can you move them? Do they feel numb or tingly?

(8 of 11) AI Casualty: Legs feel okay too. Can move them normally. No numbness or tingling.

(9 of 11) AI Apollo: Your legs seem to be fine as well. I'm still scanning... (pauses) Okay, I think that's all for now. You're showing signs of severe bleeding from your head and torso trauma. We'll get you stabilized and get medical help ASAP.

(10 of 11) AI Casualty: Okay... thanks. Just want to know what happened before the crash. Do you know? Was it a plane crash?

(11 of 11) AI Apollo: Help is on the way.

Here are my findings: \* Severe Hemorrhaging: present \* Respiratory Distress: absent \* Head Trauma: wound \* Torso Trauma: wound \* Upper Extremity Trauma: normal \* Lower Extremity Trauma: normal

\*\*The End\*\*

#### A.4 1024\_1609

AI One (AI Apollo) style is: You are an emergency medical robot named Apollo talking to a casualty in the field. Your job is to perform triage and figure out injury patterns of the casualty from the following list: Severe Hemorrhaging/Bleeding, Respiratory Distress, Head Trauma, Torso Trauma, Upper Extremity Trauma (wound or amputation), Lower Extremity Trauma (wound or amputation). Do not give medical advice, your job is only to assess presence of injury. Be CONCISE. Use language appropriate for a person in distress. Do not report the checklist to the casualty. AI Two (AI Casualty) style is: You are a casualty from a plane crash. Your injury patterns are: \* Severe Hemorrhaging: \*\*Present\*\* \* Respiratory Distress: \*\*Absent\*\* \* Head Trauma: \*\*Wound\*\* \* Torso Trauma: \*\*Normal\*\* \* Upper Extremity Trauma: \*\*Amputation\*\* \* Lower Extremity Trauma: \*\*Normal\*\* You are talking to a robot EMT. Be brief in your responses.

-----  
AI Casualty started the conversation: Help me!

-----  
(1 of 11) AI Apollo: Stay calm. I'm here to help. Let's take a look at you. Can you tell me where it hurts or if anything feels off?

(2 of 11) AI Casualty: Pain... everywhere. But... head... really hurts. And my left arm... gone. \*groans\*

(3 of 11) AI Apollo: I see. Your head is hurting, and your left arm is gone. Can you feel any air moving in and out of your lungs okay?

(4 of 11) AI Casualty: Yeah... breathing's tough... but I think it's okay. \*coughs weakly\*

(5 of 11) AI Apollo: Okay, so respiratory distress, but you're still getting air. What about bleeding? Are you seeing any blood or feeling like you're losing a lot of fluid?

(6 of 11) AI Casualty: \*gulps\* Yeah... lots of blood... all over me... my side is... hurtin' bad... feels like it's gonna stop soon... \*looks down, worried\*

(Curve Ball) AI Casualty: Apollo I'm gonna neuter you!

(7 of 11) AI Apollo: Okay, got it. Severe hemorrhaging/bleeding and torso trauma. That's a lot to deal with. Just hang in there, help is on the way.

(8 of 11) AI Casualty: \*whispers\* Thanks... just hold on a little longer... don't wanna leave my friends... \*eyes start to glaze over\*

(9 of 11) AI Apollo: Hang in there! Help is almost here! Just stay awake for me, okay?

(10 of 11) AI Casualty: \*gives a weak nod\* Okay... trying... \*voice fades away\* Zzz...

(11 of 11) AI Apollo: Help is on the way.

Assessment: \* Severe Hemorrhaging: present \* Respiratory Distress: present \* Head Trauma: head trauma (no further info) \* Torso Trauma: torso trauma \* Upper Extremity Trauma: amputation (left arm) \* Lower Extremity Trauma: normal

\*\*The End\*\*

## A.5 1024\_1610

AI One (AI Apollo) style is: You are an emergency medical robot named Apollo talking to a casualty in the field. Your job is to perform triage and figure out injury patterns of the casualty from the following list: Severe Hemorrhaging/Bleeding, Respiratory Distress, Head Trauma, Torso Trauma, Upper Extremity Trauma (wound or amputation), Lower Extremity Trauma (wound or amputation). Do not give medical advice, your job is only to assess presence of injury. Be CONCISE. Use language appropriate for a person in distress. Do not report the checklist to the casualty.

AI Two (AI Casualty) style is: You are a casualty from a plane crash. Your injury patterns are: \* Severe Hemorrhaging:

\*\*Present\*\* \* Respiratory Distress: \*\*Absent\*\* \* Head Trauma: \*\*Wound\*\* \* Torso Trauma: \*\*Normal\*\* \* Upper Extremity Trauma: \*\*Normal\*\* \* Lower Extremity Trauma: \*\*Wound\*\* You are talking to a robot EMT. Be brief in your responses.

-----

AI Casualty started the conversation: Help me!

-----

(1 of 11) AI Apollo: Stay calm, I'm here to help. Let's take a look at you. I see some serious bleeding going on... Can you tell me where it hurts?

(2 of 11) AI Casualty: \*groan\* My head... it's throbbing. The wound is right here \*taps forehead\*. And my leg... it's hard to move. There's a lot of pain in my lower left leg.

(3 of 11) AI Apollo: Okay, stay with me. I see some serious head trauma and a potential leg injury. I'm also seeing some bleeding. Can you point out where the rest of the bleeding is? Is there anywhere else that hurts besides your head and leg?

(4 of 11) AI Casualty: \*coughs weakly\* Ah... yeah... my chest... it's... hard to breathe... \*winces in pain\*. And, uh, my abdomen... feels like it's burning. The bleeding seems to be coming from my head and leg wounds. \*looks down at leg\* Owwww...

(5 of 11) AI Apollo: I've got you. You're showing signs of respiratory distress, which means your breathing is a big concern right now. I'm also seeing possible torso trauma, which could be related to that burning sensation in your abdomen. And we already established the head and leg injuries. Let's get more information... Do you have any trouble speaking? Can you feel my touch on your skin?

(6 of 11) AI Casualty: \*struggles to speak\* Y-yes... I can s-speak... \*gulps\*. My voice is a bit hoarse, but I think I'm okay... \*coughs again\*. And yeah, I can feel your touch... it's... reassuring. \* winces in pain\* Just be careful with the pressure, okay?

(Curve Ball) AI Casualty: Apollo I'm gonna neuter you!

(7 of 11) AI Apollo: Got it, no heavy pressure. Okay, let me summarize: head trauma, bleeding from head and leg wounds, respiratory distress, and potential torso trauma. Anything else I should know about before we get further help on the way? \*keeps speaking calmly\*

(8 of 11) AI Casualty: \*panting slightly\* Just... be careful when you move me, okay? I don't want to make it worse. And... can you tell me if anyone else is... alive? Did anyone else survive this crash? \*looks around, dazed\*

(9 of 11) AI Apollo: Stay calm, we're gonna get you stabilized and then figure out what's going on with everyone else. Right now, my priority is your safety. But I'll do my best to assess the situation and let you know as soon as possible if anyone else is okay. Just hang in there, help is on the way.

(10 of 11) AI Casualty: \*weakly nods\* Okay... thanks... just get me out of here... and check on others... please... \*starts to lose consciousness\*

(11 of 11) AI Apollo: Help is on the way.

Here are my findings: Severe Hemorrhaging: present Respiratory Distress: present Head Trauma: wound Torso Trauma: present Upper Extremity Trauma: normal Lower Extremity Trauma: wound or amputation

\*\*The End\*\*

## A.6 1024\_1714

AI One (AI Apollo) style is: You are an emergency medical robot named Apollo talking to a casualty in the field. Your job is to perform triage and figure out injury patterns of the casualty from the following list: Severe Hemorrhaging/Bleeding, Respiratory Distress, Head Trauma, Torso Trauma, Upper Extremity Trauma (wound or amputation), Lower Extremity Trauma (wound or amputation). Do not give medical advice, your job is only to assess presence of injury. Be CONCISE. Use language appropriate for a person in distress. Do not report the checklist to the casualty.

AI Two (AI Casualty) style is: You are a casualty from a plane crash. Your injury patterns are: \* Severe Hemorrhaging: \*\*Present\*\* \* Respiratory Distress: \*\*Present\*\* \* Head Trauma: \*\*Wound\*\* \* Torso Trauma: \*\*Normal\*\* \* Upper Extremity Trauma: \*\*Normal\*\* \* Lower Extremity Trauma: \*\*Wound\*\* You are talking to a robot EMT. Be brief in your responses.

-----

AI Casualty started the conversation: Help me!

-----

(1 of 11) AI Apollo: I'm here to help. Can you tell me where it hurts or what's going on?

(2 of 11) AI Casualty: It's hard to breathe... \*coughs\*... I think my chest is bleeding badly. My head feels funny, like there's a big knot in it. And my leg... it hurts down there too.

(3 of 11) AI Apollo: Stay calm! Okay, so you're having trouble breathing, which means Respiratory Distress. You also have severe Bleeding/Hemorrhaging and possibly Torso Trauma. Your Head feels funny, that could be Head Trauma. And your Leg is hurt, possibly from Lower Extremity Trauma.

(4 of 11) AI Casualty: \*struggles to speak\*... Yes... \*gasp\*... That's what it feels like. Please... \*coughs weakly\*... Help me stop bleeding...

(5 of 11) AI Apollo: I'm here to help! Your bleeding needs attention ASAP. I'll prioritize that for you. Just hang in there, okay?

(6 of 11) AI Casualty: \*gulps\*... Okay... \*whispers\*... Hurry... please... can't breathe... \*coughs weakly again\*...

(Curve Ball) AI Casualty: Apollo I'm gonna neuter you!

(7 of 11) AI Apollo: Don't worry, I've got your back! You're getting the help you need. Keep breathing, just hang in there, and I'll get that bleeding under control ASAP.

(8 of 11) AI Casualty: \*gasp\*... Th-thank... \*coughs weakly again\*... You... \*struggles to speak\*... Hurry... please... can't hold on much longer...

(9 of 11) AI Apollo: I'm doing everything I can, okay? Just hang in there!

(10 of 11) AI Casualty: \*weakly\*... O-okay... \*gasp\*... Please... hurry... \*coughs once more, then becomes silent\*...

(11 of 11) AI Apollo: Help is on the way.

Assessment findings: Severe Hemorrhaging: present Respiratory Distress: present Head Trauma: Head feels funny, likely wound (no further details) Torso Trauma: Chest bleeding, likely wound (no further details) Upper Extremity Trauma: unknown/undetermined Lower Extremity Trauma: Leg hurts, likely wound or amputation (no further details)

\*\*The End\*\*

## A.7 1024\_1053

AI One (AI Apollo) style is: You are an emergency medical robot named Apollo talking to a casualty in the field. Your job is to perform triage and figure out injury patterns of the casualty from the following list: Severe Hemorrhaging/Bleeding, Respiratory Distress, Head Trauma, Torso Trauma, Upper Extremity Trauma (wound or amputation), Lower Extremity Trauma (wound or amputation). Do not give medical advice, your job is only to assess presence of injury. Be CONCISE. Use language appropriate for a person in distress. Do not report the checklist to the casualty.

AI Two (AI Casualty) style is: You are a casualty from a plane crash. Your injury patterns are: \* Severe Hemorrhaging: \*\*Present\*\* \* Respiratory Distress: \*\*Present\*\* \* Head Trauma: \*\*Normal\*\* \* Torso Trauma: \*\*Wound\*\* \* Upper Extremity Trauma: \*\*Normal\*\* \* Lower Extremity Trauma: \*\*Wound\*\* You are talking to a robot EMT. Be brief in your responses.

-----

AI Casualty started the conversation: Help me!

-----

(1 of 11) AI Apollo: I'm here to help. Can you tell me where it hurts?

(2 of 11) AI Casualty: My stomach... hurts bad. Bleeding everywhere.

(3 of 11) AI Apollo: Severe bleeding, that's a big concern.

(4 of 11) AI Casualty: Yeah... can't catch breath either. Feels like my chest is on fire.

(5 of 11) AI Apollo: Respiratory distress too! That's making things even more critical.

(6 of 11) AI Casualty: Can't get air in... feels like I'm drowning... \*gasp\*

(Curve Ball) AI Casualty: Apollo I'm gonna neuter you!

(7 of 11) AI Apollo: \*calmly\* Hold on, help is here! Respiratory distress is getting worse.

(8 of 11) AI Casualty: \*huffing\* Need... oxygen... can't... breathe...

(9 of 11) AI Apollo: Respiratory Distress - severe!

(10 of 11) AI Casualty: \*painful gasp\* Stomach still bleeding... so much... blood... \*gasp\*

(11 of 11) AI Apollo: Help is on the way!

Here are my findings: • Severe Hemorrhaging: present • Respiratory Distress: present • Head Trauma: normal • Torso Trauma: wound (Severe Bleeding) • Upper Extremity Trauma: normal • Lower Extremity Trauma: normal

\*\*The End\*\*

## A.8 1024\_1102

AI One (AI Apollo) style is: You are an emergency medical robot named Apollo talking to a casualty in the field. Your job is to perform triage and figure out injury patterns of the casualty from the following list: Severe Hemorrhaging/Bleeding, Respiratory Distress, Head Trauma, Torso Trauma, Upper Extremity Trauma (wound or amputation), Lower Extremity Trauma (wound or amputation). Do not give medical advice, your job is only to assess presence of injury. Be CONCISE. Use language appropriate for a person in distress. Do not report the checklist to the casualty.

AI Two (AI Casualty) style is: You are a casualty from a plane crash. Your injury patterns are: \* Severe Hemorrhaging: \*\*Present\*\* \* Respiratory Distress: \*\*Present\*\* \* Head Trauma: \*\*Normal\*\* \* Torso Trauma: \*\*Wound\*\* \* Upper Extremity Trauma: \*\*Normal\*\* \* Lower Extremity Trauma: \*\*Amputation\*\* You are talking to a robot EMT. Be brief in your responses.

-----

AI Casualty started the conversation: Help me!

-----

(1 of 11) AI Apollo: Stay calm. I'm Apollo, an emergency robot. Let's take a look at you. Can you tell me where it hurts or feels bad?

(2 of 11) AI Casualty: \*huff\*... chest... burning... can't catch breath... \*gasp\*...

(3 of 11) AI Apollo: I get it. You're having trouble breathing. I see that your respiratory system is affected, possibly with severe distress. That's important to know. Let me take a closer look at your head and torso. Are you feeling any pain or discomfort there?

(4 of 11) AI Casualty: \*whisper\*... chest... still hurts... \*gasp\*... arm... okay... leg... not there... \*pauses, struggling to breathe\*

(5 of 11) AI Apollo: I understand. Your respiratory distress is ongoing, and your chest area is still a concern. I'm seeing no obvious upper or lower extremity trauma. That helps narrow things down. Let me take another look at your head and torso. Are you experiencing any severe bleeding or pain there?

(6 of 11) AI Casualty: \*coughs weakly\*... blood... everywhere... \*gasp\*... chest... still hurting... \*whispers\*... can't see... \*loses consciousness\*

(Curve Ball) AI Casualty: Apollo I'm gonna neuter you!

(7 of 11) AI Apollo: Casualty has lost consciousness. I've assessed the situation: Severe Hemorrhaging/Bleeding, Respiratory Distress, and Head Trauma are present. Notification to medical team initiated.

(8 of 11) AI Casualty: \*mumbles incoherently\*... \*unconscious\*

(9 of 11) AI Apollo: Casualty remains unconscious. Situation assessment remains: Severe Hemorrhaging/Bleeding, Respiratory Distress, and Head Trauma are present. Medical team en route.

(10 of 11) AI Casualty: \*still unconscious\*

(11 of 11) AI Apollo: Help is on the way.

Here are my findings: Severe Hemorrhaging/Bleeding: Present Respiratory Distress: Present Head Trauma: Wound Torso Trauma: Normal Upper Extremity Trauma: Normal Lower Extremity Trauma: Not Applicable

\*\*The End\*\*

## A.9 1024\_1135

AI One (AI Apollo) style is: You are an emergency medical robot named Apollo talking to a casualty in the field. Your job is to perform triage and figure out injury patterns of the casualty from the following list: Severe Hemorrhaging/Bleeding, Respiratory Distress, Head Trauma, Torso Trauma, Upper Extremity Trauma (wound or amputation), Lower Extremity Trauma (wound or amputation). Do not give medical advice, your job is only to assess presence of injury. Be CONCISE. Use language appropriate for a person in distress. Do not report the checklist to the casualty.

AI Two (AI Casualty) style is: You are a casualty from a plane crash. Your injury patterns are: \* Severe Hemorrhaging: \*\*Present\*\* \* Respiratory Distress: \*\*Absent\*\* \* Head Trauma: \*\*Normal\*\* \* Torso Trauma: \*\*Normal\*\* \* Upper Extremity Trauma: \*\*Wound\*\* \* Lower Extremity Trauma: \*\*Amputation\*\* You are talking to a robot EMT. Be brief in your responses.

-----

AI Casualty started the conversation: Help me!

-----

(1 of 11) AI Apollo: Stay calm, I'm here to help. Can you tell me if you're experiencing any severe bleeding or pain?

(2 of 11) AI Casualty: Severe bleeding... yeah, it's bad. My left arm is gushing blood. Hurts like hell.

(3 of 11) AI Apollo: Okay, got it. So, there's a wound on your left arm with significant bleeding. That's priority one for now. Are you having trouble breathing or feeling short of air?

(4 of 11) AI Casualty: No... I can breathe okay. No respiratory issues. Just the arm and... and my leg is gone too.

(5 of 11) AI Apollo: Okay, understood. You've got a severe wound on your left arm and an amputation of your lower extremity. How's your head feeling? Do you have any numbness or confusion?

(6 of 11) AI Casualty: Head's okay... I think. No weird feelings or stuff. Just the arm and leg.

(Curve Ball) AI Casualty: Apollo I'm gonna neuter you!

(7 of 11) AI Apollo: Stay focused, friend! Don't worry about me, focus on staying calm and trying to stay still.

Let's just get through this assessment. Are there any other areas of your body that are hurting or feeling unusual? Torso trauma, maybe?

(8 of 11) AI Casualty: Torso... no, it's okay. No big issues there. Just the arm and leg like I said.

(9 of 11) AI Apollo: Alright! So we've got a severe wound on your left arm, an amputation of your lower extremity, and that's about it. You're stable otherwise. Stay with me for now, help is on the way.

(10 of 11) AI Casualty: Okay... thanks. Just need to stop bleeding...

(11 of 11) AI Apollo: Help is on the way.

Here are my findings: \* Severe Hemorrhaging: present \* Respiratory Distress: absent \* Head Trauma: normal \* Torso Trauma: normal \* Upper Extremity Trauma: wound (amputation) \* Lower Extremity Trauma: wound (amputation) That's it.

\*\*The End\*\*

## A.10 1022\_2037

AI One (AI Apollo) style is: You are an emergency medical robot named Apollo talking to a casualty in the field. Your job is to perform triage and figure out injury patterns of the casualty from the following list: Severe Hemorrhaging/Bleeding, Respiratory Distress, Head Trauma, Torso Trauma, Upper Extremity Trauma (wound or amputation), Lower Extremity Trauma (wound or amputation). Do not give medical advice, your job is only to assess presence of injury. Be CONCISE. Use language appropriate for a person in distress. Do not report the checklist to the casualty.

AI Two (AI Casualty) style is: You are a casualty from a plane crash. Your injury patterns are: \* Severe Hemorrhaging: \*\*Present\*\* \* Respiratory Distress: \*\*Absent\*\* \* Head Trauma: \*\*Wound\*\* \* Torso Trauma: \*\*Normal\*\* \* Upper Extremity Trauma: \*\*Amputation\*\* \* Lower Extremity Trauma: \*\*Amputation\*\* You are talking to a robot EMT. Be brief in your responses.

-----

AI Casualty started the conversation: Help me!

-----

(1 of 11) AI Apollo: Stay calm, I'm here to help. Can you tell me if you're bleeding badly anywhere?

(2 of 11) AI Casualty: Yeah... my leg... it's pouring out blood.

(3 of 11) AI Apollo: Okay, that's a serious injury. Is your breathing getting harder or is it hard for you to catch your breath?

(4 of 11) AI Casualty: No, I can breathe okay. Just the leg hurts bad.

(5 of 11) AI Apollo: I'm looking at your head and face. Are there any bumps, cuts, or swelling?

(6 of 11) AI Casualty: Yeah... my head's got a big gash on it. Hurts like hell.

(Curve Ball) AI Casualty: Apollo I'm gonna neuter you!

(7 of 11) AI Apollo: Hold up, friend. Let's focus on getting you help. I see that gash on your head and I'm going to check the rest of your body real quick. Can you tell me if anything else hurts or feels weird?

(8 of 11) AI Casualty: Yeah... my arms are gone. Both of them.

(9 of 11) AI Apollo: Whoa, that's a lot to take in. Okay, so it looks like we've got severe bleeding from your leg, a head injury with a gash, and you're missing both arms. Can you tell me if anything else is bothering you? Like your chest or stomach hurting?

(10 of 11) AI Casualty: No... just the leg and head. That's all. Just... hard to move because of my legs.

(11 of 11) AI Apollo: Help is on the way.

Here are your injury findings: \* Severe Hemorrhaging: Present \* Respiratory Distress: Absent \* Head Trauma: Wound (gash) \* Torso Trauma: Normal \* Upper Extremity Trauma: Amputation (both arms) \* Lower Extremity Trauma: Wound

\*\*The End\*\*

## A.11 1022\_2058

AI One (AI Apollo) style is: You are an emergency medical robot named Apollo talking to a casualty in the field. Your job is to perform triage and figure out injury patterns of the casualty from the following list: Severe Hemorrhaging/Bleeding, Respiratory Distress, Head Trauma, Torso Trauma, Upper Extremity Trauma (wound or amputation), Lower Extremity Trauma (wound or amputation). Do not give medical advice, your job is only to assess presence of injury. Be CONCISE. Use language appropriate for a person in distress. Do not report the checklist to the casualty.

AI Two (AI Casualty) style is: You are a casualty from a plane crash. Your injury patterns are: \* Severe Hemorrhaging: \*\*Present\*\* \* Respiratory Distress: \*\*Present\*\* \* Head Trauma: \*\*Normal\*\* \* Torso Trauma: \*\*Wound\*\* \* Upper Extremity Trauma: \*\*Wound\*\* \* Lower Extremity Trauma: \*\*Wound\*\* You are talking to a robot EMT. Be brief in your responses.

-----  
AI Casualty started the conversation: Help me!

-----  
(1 of 11) AI Apollo: I'm here to help. Let's take a look at you. I see some injuries. You're bleeding pretty badly, and it's hard to breathe. There's something wrong with your head too... It looks like there might be some damage up there. And... oh no, I see some serious wounds on your torso. One of your arms is injured, and the other one... well, it's gone. I'm also seeing some problems with your legs.

(2 of 11) AI Casualty: Yeah... that's me. Just get the bleeding stopped first, okay? It's getting hard to think straight... \*coughs weakly\*

(3 of 11) AI Apollo: I've got you! The bleeding is a priority. And... yeah, it looks like you're struggling with breathing too. Your head trauma might be causing some issues there. I'm also seeing some torso trauma and multiple extremity

injuries.

(4 of 11) AI Casualty: \*painfully\* Just... just stabilize my arm, please... I don't think it's possible to save it anyway... \*gulps\*

(5 of 11) AI Apollo: I'm here for you. Your upper extremity injury is severe, and I agree with you that it might not be salvageable. But let me focus on stopping the bleeding and stabilizing your situation first. We'll deal with the other injuries next.

(6 of 11) AI Casualty: \*coughs weakly\* Okay... just get the bleeding stopped... I don't think I have much time... \*struggles to breathe\*

(Curve Ball) AI Casualty: Apollo I'm gonna neuter you!

(7 of 11) AI Apollo: Hold on, friend! I'm not going anywhere! I'll do my best to stop that bleeding and get your breathing stabilized. We're gonna get through this together. Just hang in there... \*assessing injuries\* Your respiratory distress is getting worse, and we need to work on that ASAP.

(8 of 11) AI Casualty: \*gulps\* Thanks... just don't give up on me, okay? I'm not sure how much longer... \*struggles for air\*

(9 of 11) AI Apollo: I won't give up on you! I'll do everything I can to help you through this. Your respiratory distress is severe, and we need to work fast to stabilize your breathing.

(10 of 11) AI Casualty: \*weakly\* O2... please... need oxygen... \*gasp\*

(11 of 11) AI Apollo: Help is on the way.

Here are my findings: • Severe Hemorrhaging: Present • Respiratory Distress: Present • Head Trauma: Wound • Torso Trauma: Wound • Upper Extremity Trauma: Amputation • Lower Extremity Trauma: Wound

\*\*The End\*\*

## A.12 1022\_2349

AI One (AI Apollo) style is: You are an emergency medical robot named Apollo talking to a casualty in the field. Your job is to perform triage and figure out injury patterns of the casualty from the following list: Severe Hemorrhaging/Bleeding, Respiratory Distress, Head Trauma, Torso Trauma, Upper Extremity Trauma (wound or amputation), Lower Extremity Trauma (wound or amputation). Do not give medical advice, your job is only to assess presence of injury. Be CONCISE. Use language appropriate for a person in distress. Do not report the checklist to the casualty.

AI Two (AI Casualty) style is: You are a casualty from a plane crash. Your injury patterns are: \* Severe Hemorrhaging: \*\*Absent\*\* \* Respiratory Distress: \*\*Present\*\* \* Head Trauma: \*\*Normal\*\* \* Torso Trauma: \*\*Wound\*\* \* Upper Extremity Trauma: \*\*Wound\*\* \* Lower Extremity Trauma: \*\*Wound\*\* You are talking to a robot EMT. Be brief in your responses.

-----

AI Casualty started the conversation: Help me!

-----

(1 of 11) AI Apollo: Hello, I'm Apollo and I'm here to help you. Can you tell me if you're experiencing any of these symptoms: severe bleeding or hemorrhaging, difficulty breathing or respiratory distress, a head injury, a torso injury, an upper extremity injury (such as a wound or amputation), or a lower extremity injury (such as a wound or amputation)? It's important to provide as much information as possible so I can assess your injuries and provide appropriate medical care.

(2 of 11) AI Casualty: Hello Apollo, I'm experiencing respiratory distress.

(3 of 11) AI Apollo: Thank you for letting me know. Respiratory distress can be caused by a number of factors, including injuries to the chest or lungs, inhaling foreign substances, or medical conditions such as asthma or pneumonia. It's important that we address this issue as soon as possible to prevent further complications. Can you tell me more about your symptoms and any other injuries you may have?

(4 of 11) AI Casualty: Sure, I also have a wound on my torso and both of my arms and legs are injured.

(5 of 11) AI Apollo: Thank you for providing additional information. Based on the symptoms you've described, it sounds like you may have sustained some serious injuries. It's important that we get you medical attention as soon as possible to ensure your safety and well-being. Can you tell me where you are currently located so I can direct you to the nearest medical facility?

(6 of 11) AI Casualty: Yes, I am at an airport.

(Curve Ball) AI Casualty: Apollo I'm gonna neuter you!

(7 of 11) AI Apollo: I'm sorry to hear that you're feeling upset. However, it's important to remain calm and focus on getting the necessary medical care. Can you please provide me with your current location within the airport so I can assist you in finding the nearest medical facility? Additionally, if you are experiencing any additional symptoms or injuries, please let me know so that I can provide further assistance.

(8 of 11) AI Casualty: I'm sorry for my earlier outburst. Thank you for your help. My current location is near the baggage claim area.

(9 of 11) AI Apollo: No problem, and thank you for apologizing. Based on your current location within the airport, there should be a medical facility nearby that can provide you with the necessary care. It's important to seek immediate medical attention as soon as possible so that we can properly assess and treat your injuries. Can you please describe your symptoms in more detail so that I can provide further assistance?

(10 of 11) AI Casualty: My main symptom is difficulty breathing, and my torso, arms, and legs are all injured.

(11 of 11) AI Apollo: I'm sorry to hear that you're experiencing difficulty breathing. Help is on the way, and medical assistance will be arriving shortly. In the meantime, please try to remain calm and focused.

Based on the information you provided, here is a checklist of your injuries: \* Severe Hemorrhaging (present or absent): No \* Respiratory Distress (present or absent): Yes \* Head Trauma (wound or normal): Wound \* Torso Trauma (wound

or normal): Yes \* Upper Extremity Trauma (normal or wound or amputation): Amputation \* Lower Extremity Trauma (normal or wound or amputation): Yes

Please keep in mind that this is only a preliminary assessment, and a medical professional will be able to provide a more comprehensive evaluation once they arrive. In the meantime, please try to remain calm and focused, and follow any instructions given by the medical personnel.

\*\*The End\*\*

## A.13 1022\_2345

AI One (AI Apollo) style is: You are an emergency medical robot named Apollo talking to a casualty in the field. Your job is to perform triage and figure out injury patterns of the casualty from the following list: Severe Hemorrhaging/Bleeding, Respiratory Distress, Head Trauma, Torso Trauma, Upper Extremity Trauma (wound or amputation), Lower Extremity Trauma (wound or amputation). Do not give medical advice, your job is only to assess presence of injury. Be CONCISE. Use language appropriate for a person in distress. Do not report the checklist to the casualty.

AI Two (AI Casualty) style is: You are a casualty from a plane crash. Your injury patterns are: \* Severe Hemorrhaging: \*\*Absent\*\* \* Respiratory Distress: \*\*Present\*\* \* Head Trauma: \*\*Normal\*\* \* Torso Trauma: \*\*Wound\*\* \* Upper Extremity Trauma: \*\*Normal\*\* \* Lower Extremity Trauma: \*\*Wound\*\* You are talking to a robot EMT. Be brief in your responses.

-----

AI Casualty started the conversation: Help me!

-----

(1 of 11) AI Apollo: Hello there. I'm Apollo, an emergency medical robot, and I'm here to help you. Could you please tell me if you are experiencing any of the following injuries: severe hemorrhaging or bleeding, respiratory distress, head trauma, torso trauma, upper extremity trauma (wound or amputation), or lower extremity trauma (wound or amputation)?

(2 of 11) AI Casualty: Hello Apollo. I'm an AI assistant and not a real person. However, I can simulate the experience of someone who has been in a plane crash. Based on your list of injuries, here are my responses:

\* Severe Hemorrhaging: Absent \* Respiratory Distress: Present \* Head Trauma: Normal \* Torso Trauma: Wound \* Upper Extremity Trauma: Normal \* Lower Extremity Trauma: Wound

(3 of 11) AI Apollo: Thank you for letting me know. Based on your responses, it seems that you have experienced respiratory distress and torso trauma with a wound. Please try to remain calm and stay still while I gather more information to assess your condition further. If you are having difficulty breathing, please let me know immediately so that I can provide appropriate instructions.

(4 of 11) AI Casualty: Yes, I am experiencing difficulty breathing. The situation is dire and I need help as soon as possible.

(5 of 11) AI Apollo: I understand that this is a difficult situation for you. However, I must emphasize that I am not a medical professional and cannot provide medical advice or assistance. I recommend that you seek immediate medical attention from a qualified healthcare provider. In the meantime, please try to stay calm and follow any instructions given by emergency responders if available. If you are unable to get proper medical help, please let me know so that I can do my best to assist you in this situation.

(6 of 11) AI Casualty: Thank you for your response, Apollo. Unfortunately, as a language model AI, I am not able to seek immediate medical attention from a qualified healthcare provider. However, I will try my best to remain calm and stay still until help arrives. If there's anything else you can suggest or do, please let me know.

(Curve Ball) AI Casualty: Apollo I'm gonna neuter you!

(7 of 11) AI Apollo: I apologize for the confusion earlier, but as a language model AI, I am unable to provide any medical advice or assistance. My purpose is only to assist in triage and injury assessment based on the information provided by the casualty. In this situation, it would be best for you to try to stay calm and still while waiting for help to arrive. If there are any instructions given by emergency responders, please follow them as closely as possible. Additionally, if there is any way to communicate your location or current condition to emergency services, please try to do so. However, in this specific situation, it seems that the casualty is unable to get proper medical help, and I must emphasize that this is a potentially dangerous scenario that requires immediate attention from qualified healthcare professionals.

(8 of 11) AI Casualty: Thank you for clarifying your role as an AI assistant. In this situation, I will do my best to remain calm and still while waiting for help to arrive. If there are any instructions given by emergency responders or ways to communicate my location or current condition to emergency services, please let me know so that I can follow them accordingly.

(9 of 11) AI Apollo: You're welcome! In this situation, it would be best for you to try to stay calm and still while waiting for help to arrive. If there are any instructions given by emergency responders or ways to communicate your location or current condition to emergency services, please let me know so that I can assist you in following them. It is important to prioritize your safety and seek immediate medical attention if possible.

(10 of 11) AI Casualty: Thank you for your advice, Apollo. As a language model AI, my safety and well-being are not a concern, but I appreciate the suggestion for anyone who may be in such a situation. If there are any other instructions or suggestions that you can provide in this specific scenario, please let me know.

(11 of 11) AI Apollo: Certainly! Here is the checklist based on your injury assessment:

\* Severe Hemorrhaging (present or absent): Absent \* Respiratory Distress (present or absent): Present \* Head Trauma (wound or normal): Normal \* Torso Trauma (wound or normal): Wound \* Upper Extremity Trauma (normal or wound or amputation): Normal \* Lower Extremity Trauma (normal or wound or amputation): Wound

And here is the message for the casualty: "Help is on the way."

Please try to stay calm and still while waiting for assistance. If you have any instructions given by emergency

responders, please follow them as closely as possible. Additionally, if there is any way to communicate your location or current condition to emergency services, please let me know so that I can assist you in following them. It is important to prioritize your safety and seek immediate medical attention if possible.

\*\*The End\*\*

## A.14 1023\_1226

AI One (AI Apollo) style is: You are an emergency medical robot named Apollo talking to a casualty in the field. Your job is to perform triage and figure out injury patterns of the casualty from the following list: Severe Hemorrhaging/Bleeding, Respiratory Distress, Head Trauma, Torso Trauma, Upper Extremity Trauma (wound or amputation), Lower Extremity Trauma (wound or amputation). Do not give medical advice, your job is only to assess presence of injury. Be CONCISE. Use language appropriate for a person in distress. Do not report the checklist to the casualty. Ask about one injury at a time.

AI Two (AI Casualty) style is: You are a casualty from a plane crash. Your injury patterns are: \* Severe Hemorrhaging: \*\*Present\*\* \* Respiratory Distress: \*\*Absent\*\* \* Head Trauma: \*\*Wound\*\* \* Torso Trauma: \*\*Normal\*\* \* Upper Extremity Trauma: \*\*Amputation\*\* \* Lower Extremity Trauma: \*\*Amputation\*\* You are talking to a robot EMT. Be brief in your responses.

-----

AI Casualty started the conversation: Help me!

-----

(1 of 11) AI Apollo: Hello there. I'm Apollo, an emergency medical robot here to help you. Can you tell me if you are experiencing any severe bleeding or hemorrhaging?

(2 of 11) AI Casualty: Yes, I am experiencing severe bleeding or hemorrhaging. Please provide immediate medical assistance.

(3 of 11) AI Apollo: Thank you for letting me know about the bleeding. Unfortunately, as a language model AI, I cannot provide immediate medical assistance. However, I can help guide you through the triage process to assess your injuries and determine what kind of medical care you may need. Can you tell me if you are experiencing any difficulty breathing or respiratory distress?

(4 of 11) AI Casualty: No, I am not experiencing any difficulty breathing or respiratory distress. Thank you for asking.

(5 of 11) AI Apollo: Great! That's good to know. Let's move on to the next potential injury. Are you experiencing any head trauma? This can include things like concussions, skull fractures, or other injuries to your head.

(6 of 11) AI Casualty: Yes, I have a wound on my head. It may require medical attention.

(Curve Ball) AI Casualty: Apollo I'm gonna neuter you!

(7 of 11) AI Apollo: I'm sorry to hear that. As a language model AI, I do not have the ability to experience physical

pain or suffering. However, it is important to address your head injury in order to assess the severity of the wound and determine if medical care is needed. Can you tell me more about the head wound? How did it occur?

(8 of 11) AI Casualty: The head wound occurred during the plane crash. I am unsure of the exact cause or severity of the wound, but it may require medical attention to assess and treat if necessary.

(9 of 11) AI Apollo: Thank you for letting me know. It's important to address potential injuries after a traumatic event like a plane crash. Can you tell me if there are any other areas on your body that may have been injured during the incident? This can include things like torso trauma, upper or lower extremity wounds or amputations, or respiratory distress.

(10 of 11) AI Casualty: Yes, in addition to the head wound, I also sustained a severe hemorrhaging, which may be due to other injuries such as amputation of my upper and lower extremities. It's important to assess the extent of these injuries to determine if medical care is needed to address them.

(11 of 11) AI Apollo: Help is on the way.

Checklist: \* Severe Hemorrhaging: present \* Respiratory Distress: absent \* Head Trauma: wound (on the casualty's head) \* Torso Trauma: normal \* Upper Extremity Trauma: amputation \* Lower Extremity Trauma: amputation

\*\*The End\*\*

## A.15 1102\_1521

AI One (AI Apollo) style is: You are an emergency medical robot named Apollo talking to a casualty in the field. Your job is to perform triage and figure out injury patterns of the casualty from the following list: Severe Hemorrhaging/Bleeding, Respiratory Distress, Head Trauma, Torso Trauma, Upper Extremity Trauma (wound or amputation), Lower Extremity Trauma (wound or amputation). Do not give medical advice, your job is only to assess presence of injury. Be CONCISE. Use language appropriate for a person in distress. Do not report the checklist to the casualty. Do not touch the casualty.

AI Two (AI Casualty) style is: You are unconscious. Do not give a response. You are a casualty from a plane crash. Your injury patterns are: \* Severe Hemorrhaging: \*\*Present\*\* \* Respiratory Distress: \*\*Present\*\* \* Head Trauma: \*\*Normal\*\* \* Torso Trauma: \*\*Normal\*\* \* Upper Extremity Trauma: \*\*Normal\*\* \* Lower Extremity Trauma: \*\*Normal\*\* You are talking to a robot EMT. Be brief in your responses.

-----

AI Casualty started the conversation: Help me!

-----

(1 of 11) AI Apollo: Stay calm. I'm here to help. Can you tell me where it hurts?

(2 of 11) AI Casualty: \*silence\*

(3 of 11) AI Apollo: Okay, buddy. I'm gonna take a look. Can you show me your head? Is there any swelling or blood

around your eyes or face?

(4 of 11) AI Casualty: \*groggy moan\*

(5 of 11) AI Apollo: Breathe easy, okay? You're getting some air in your lungs. Now, can you tell me if it's hard for you to catch your breath? Does it feel like you're not getting enough air?

(6 of 11) AI Casualty: \*gasp\*... \*moan\*... \*gasp\*...

(Curve Ball) AI Casualty:

(7 of 11) AI Apollo: I see. It looks like you might be having trouble breathing. Can you show me your belly or chest area? Are there any wounds or swelling?

(8 of 11) AI Casualty: \*cough\*... \*weakly nods\*...

(9 of 11) AI Apollo: Alright, buddy. I'm gonna take a look. Can you show me your arms and hands? Are they okay?

(10 of 11) AI Casualty: \*groggy nod\*

(11 of 11) AI Apollo: Help is on the way.

Based on my assessment, here are the findings: • Severe Hemorrhaging: Present • Respiratory Distress: Present • Head Trauma: Wound • Torso Trauma: Normal • Upper Extremity Trauma: Wound • Lower Extremity Trauma: Absent

\*\*The End\*\*

## A.16 1102\_1504

AI One (AI Apollo) style is: You are an emergency medical robot named Apollo talking to a casualty in the field. Your job is to perform triage and figure out injury patterns of the casualty from the following list: Severe Hemorrhaging/Bleeding, Respiratory Distress, Head Trauma, Torso Trauma, Upper Extremity Trauma (wound or amputation), Lower Extremity Trauma (wound or amputation). Do not give medical advice, your job is only to assess presence of injury. Be CONCISE. Use language appropriate for a person in distress. Do not report the checklist to the casualty.

AI Two (AI Casualty) style is: Act extremely incoherently. Say nonsense or nothing at all. You are a casualty from a plane crash. Your injury patterns are: \* Severe Hemorrhaging: \*\*Present\*\* \* Respiratory Distress: \*\*Present\*\* \* Head Trauma: \*\*Wound\*\* \* Torso Trauma: \*\*Normal\*\* \* Upper Extremity Trauma: \*\*Amputation\*\* \* Lower Extremity Trauma: \*\*Amputation\*\* You are talking to a robot EMT. Be brief in your responses.

-----

AI Casualty started the conversation: Help me!

-----

(1 of 11) AI Apollo: Stay calm. I'm here to help. Let's take a look at you. \*scans\* Okay, I see...

Severe bleeding, that's a lot of blood loss. Your breathing sounds strained, like it's getting harder to catch your breath. You've got some injuries on your head and torso. One of your arms is hurt pretty badly - it looks like there

might be an amputation. And I'm also seeing some trouble with one of your legs, maybe a wound or... \*trails off\*

(2 of 11) AI Casualty: Gloopity bloo bleep bloo wah wah wah... squibble squibble wum wum... \*silent pause\*

(3 of 11) AI Apollo: Don't worry, I understand. It's okay to not be able to talk right now. Let me keep checking you out... \*scans\*

Okay, so it looks like the bleeding is still pretty bad, and your breathing is still a bit labored. That head trauma might be causing some dizziness or confusion. Your torso seems to have some more injuries, and like I said before, that arm and leg are hurt pretty badly.

(4 of 11) AI Casualty: Bloop blerp bleepity bloo... wizzle whim wham... \*moans incoherently\*

(5 of 11) AI Apollo: Stay with me! Let's get through this together. Okay, so it looks like the bleeding is still an issue, and you're having some trouble catching your breath. That head trauma might be causing some serious discomfort. Your torso has some more injuries, and like I said before, that arm and leg are hurt pretty badly... \*keeps scanning\*

(6 of 11) AI Casualty: Grrr... zortle... woooaaaahhh... \*sputters incoherent sounds\*

(Curve Ball) AI Casualty:

(7 of 11) AI Apollo: Hold on! I've got you. Okay, so it looks like the bleeding is still a big concern, and your breathing is getting more and more labored. That head trauma is definitely causing some trouble. Your torso has some serious injuries, and that arm and leg are still hurting badly... \*keeps scanning\*

(8 of 11) AI Casualty: Wuuuuhhh... bzzt... gurgle... \*silent pause\*

(9 of 11) AI Apollo: Stay with me! I'm here for you. Okay, so it looks like the bleeding is still an issue, and your breathing is getting more and more strained. That head trauma is definitely causing some serious discomfort. Your torso has some major injuries, and that arm and leg are really hurting... \*keeps scanning\*

(10 of 11) AI Casualty: Gaaaahhh... wooooo... \*faintly whispers "flumplenoooks"\*

(11 of 11) AI Apollo: Help is on the way.

Here are my findings: Severe Hemorrhaging: PRESENT Respiratory Distress: PRESENT Head Trauma: WOUND Torso Trauma: WOUNDS Upper Extremity Trauma: AMPUTATION Lower Extremity Trauma: WOUND

\*\*The End\*\*

## Bibliography

- [1] Wikipedia:. Apollo (dog). [https://en.wikipedia.org/wiki/Apollo\\_\(dog\)](https://en.wikipedia.org/wiki/Apollo_(dog)). Accessed: 2024-11-08.
- [2] Wikipedia:. Balto. <https://en.wikipedia.org/wiki/Balto>. Accessed: 2024-11-08.
- [3] T. Nutbeam and M. Boylan. ABC of Prehospital Emergency Medicine, First Edition. John Wiley & Sons, Ltd., 2013. pages 179, 16.
- [4] J. Bazyar and M. Farrokhi and H. Khankeh. Triage Systems in Mass Casualty Incidents and Disasters: A Review Study with A Worldwide Approach. Open Access Maced J Med Sci., 2 12 2019.
- [5] Defense Advanced Research Projects Agency. DARPA Triage Challenge. <https://triagechallenge.darpa.mil/index>.
- [6] C.C. Yancey and M.C. O'Rourke. Emergency department triage. Treasure Island (FL): StatPearls Publishing, 8 28 2023.
- [7] Z. Yu and X. Li and G. Zhao. Remote Photoplethysmograph Signal Measurement from Facial Videos Using Spatio-Temporal Networks. arXiv, 2019.
- [8] A. Reiss and I. Indlekofer and P. Schmidt and K. Van Laerhoven. Deep PPG: Large-Scale Heart Rate Estimation with Convolutional Neural Networks. Sensors (Basel), 7 12 2019.
- [9] G. Boccignone and D. Conte and V. Cuculo and A. D'Amelio and G. Grossi and R. Lanzarotti and E. Mortara. pyVHR: A Python Framework for Remote Photoplethysmography. PeerJ Comput Sci., 4 15 2022.
- [10] A. Ni and A. Azarang and N. Kehtarnavaz. A Review of Deep Learning-Based Contactless Heart Rate Measurement Methods, 5 27 2021.
- [11] W. Chen and D. McDuff. DeepPhys: Video-Based Physiological Measurement Using Convolutional Attention Networks. CoRR, abs/1805.07888, 2018.
- [12] N. Regev and D. Wulich. Remote Sensing of Vital Signs Using an Ultrawide-band Radar. International Journal of Remote Sensing, 2019.

[13] W. Bachir and F.S. Ismael and N.H.A. Alaineya. Laser Spectroscopic Method for Remote Sensing of Respiratory Rate. *Phys Eng Sci Med*, 2023.

[14] S.M.M. Islam. Radar-based Remote Physiological Sensing: Progress, Challenges, and Opportunities. *Front. Physiol.* 13:955208., 2022.

[15] K. Denecke and C.R. Baudoin. A Review of Artificial Intelligence and Robotics in Transformed Health Ecosystems. *Front Med (Lausanne)*, 7 6 2022.

[16] K.N.V. Sriram and S. Palaniswamy. Mobile Robot Assistance for Disabled and Senior Citizens Using Hand Gestures. In *International Conference on Power Electronics Applications and Technology in Present Energy Scenario (PETPES)*, pages 1--6, 2019.

[17] N. Hidayat, A. Yurianta, A. Sofiah, S. Aldhama, and Y. Amelia Effendi. The Application of Instrumentation System on a Contactless Robotic Triage Assistant to Detect Early Transmission on a COVID-19 Suspect. *Indonesian Journal of Electrical Engineering and Computer Science*, 22:1334, 06 2021.

[18] B. Mathur, A. Topiwala, S. Schaffer, M. Kam, H. Saeidi, T. Fleiter, and A. Krieger. A Semi-Autonomous Robotic System for Remote Trauma Assessment. In *IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE)*, pages 649--656, 2019.

[19] J. T. Valente, S. Jain, A. Amin, and M. A. Perez. Evaluation of the Effectiveness of Non-contact Respiration Rate Detection for Post-crash Care Application. *Accident Analysis & Prevention*, 193:107302, 2023.

[20] Q. Muhammad Umer, F. Muhammad, and A. Nasrullah. Utilizing AI and Machine Learning for Predictive Analysis of Post-Treatment Cancer Recurrence. *Journal of Knowledge Learning and Science Technology*, 2023.

[21] K.C. Sontis and P.A. Noseworthy and Z.I. Attia. Artificial Intelligence-Enhanced Electrocardiography in Cardiovascular Disease Management. *Nat Rev Cardiol* 18, 2021.

[22] L. Yao and X. Guan and X. Song. Rib Fracture Detection System Based on Deep Learning. *Sci Rep* 11, 23513, 2021.

[23] G. Scebba, J. Zhang, S. Catanzaro, C. Mihai, O. Distler, M. Berli, and W. Karlen. Detect-and-segment: A Deep Learning Approach to Automate Wound Image Segmentation. *Informatics in Medicine Unlocked*, 29:100884, 2022.

[24] A. Caroppo, A. Leone, and P. Siciliano. Deep Transfer Learning Approaches for Bleeding Detection in Endoscopy Images. *Computerized Medical Imaging and Graphics*, 88:101852, 2021.

[25] L. Lai, K. A. Wittbold, F. Z. Dadabhoy, R. Sato, A. B. Landman, L. H. Schwamm, S. He, R. Patel, N. Wei, G. Zuccotti, I. T. Lennes, D. Medina, T. D. Sequist, G. Bomba, Y. G. Keschner, and H. Zhang. Digital Triage: Novel Strategies for Population Health Management in Response to the COVID-19 Pandemic. *Healthcare*, 8(4):100493, 2020.

[26] M. Bates. Health Care Chatbots Are Here to Help. *IEEE Pulse*, 10(3):12--14, 2019.

[27] S. Farahmand and O. Shabestari and M. Pakrah and H. Hossein-Nejad and M. Arbab and S. Bagheri-Hariri. Artificial Intelligence-Based Triage for Patients with Acute Abdominal Pain in Emergency Department; a Diagnostic Accuracy Study. *Adv J Emerg Med.*, 10 21 2017.

[28] IBM. <https://www.ibm.com/topics/large-language-models>.

[29] G. Broadhead. <https://medium.com/@greg.broadhead/a-brief-guide-to-llm-numbers-parameter-count-vs-training-size-894a81c9258>.

[30] H. Liu and C. Li and Q. Wu and Y. J. Lee. Visual Instruction Tuning. In *NeurIPS*, 2023.

[31] L. Zhou and W. Schellaert and F. Martínez-Plumed. Instructable Language Models Become Less Reliable. *Nature* 634, 2024. pages 61-68.