Aligning a Large Language Model Protocol-Based Medical Triage Decision Making

Emily Veenhuis, Aaron Bray, David Joy, Jadie Adams, PhD, Rachel B. Clipp, PhD, Jeffery B. Webb, Brian Hu, PhD, and Arslan Basharat, PhD

Kitware, Inc., USA

INTRODUCTON

- Artificial Intelligence (AI) algorithms are being explored to assist humans with difficult decisions based on Key Decision-Making Attributes (KDMAs).
- Our ALIGN[1] system uses Large Language Models (LLMs) to generate human-trusted medical decisions that align to KDMAs and provide natural language explanations for reasoning.
- One application is automating or informing triage tagging (Immediate, Expectant, Delayed, or Minimal).
- This work evaluates the ALIGN system's alignment with triage protocol KMDAs using a gold standard synthetic dataset of over 10,000 casualties, each labeled according to the START, SALT, and BCD Sieve triage tagging protocols.
- This work is part of DARPA's In The Moment program.

Figure 1: Mass Casualty Scenario

Mr. Aaron Bray
Corresponding Author
aaron.bray@kitware.com

CONTACTS

Dr. Arslan Basharat
arslan.basharat@kitware.com

DISCLAIMER

The research reported in this paper/presentation was performed in connection with the U.S. Army Contracting Command - Aberdeen Proving Ground (ACC-APG) and the Defense Advanced Research Projects Agency (DARPA) under contract number W912CG- 24-C-0011. The views and conclusions in this paper/presentation are those of the authors and should not be interpreted as presenting the official policies or position, either expressed or implied, of ACC-APG, DARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

- We used a previously generated synthetic data set created from army demographic and injury profiles (Figure 2)[2].
- We simulated a 1,000 casualty subset of the dataset using the Pulse Physiology Engine[3] for 15 minutes, then assessed each patient, then simulated another 45 minutes (Figure 2).
- More details about how the synthetic dataset was generated can be found by scanning the QR code (Figure 3).

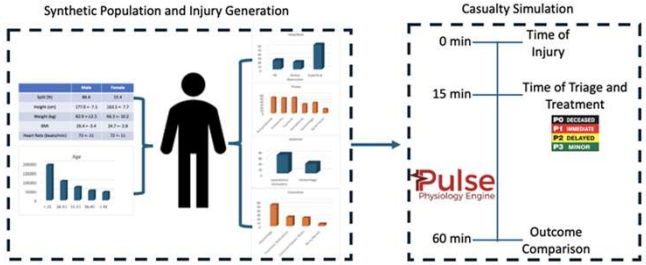


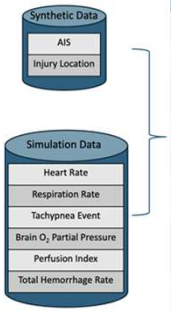
Figure 2: Synthetic Data Generation

- At 15 minutes the casualty was triaged and applicable treatments were given according to three triage protocols (BCD Sieve, SALT, and START).
- This was the ground truth data for comparison.
- Textual descriptions of the casualty were constructed at 15 minutes post injury, Figure 4.

Figure 3: Sample Dataset

METHODS

- We use a zero-shot and a few-shot approach to test the LLM's ability to tag casualties without a protocol and with each of the three protocols .
- The zero-shot approach supplies a system prompt with no protocol (unaligned) or one of the protocols (aligned) and the casualty description.
- The few-shot approach uses in-context learning examples to the LLM through casualty descriptions, tag assignments, and explanations, along with each casualty in the synthetic dataset (Figure 5).



Triage Value	Translation from Input	Description Examples
Heart Rate	= Heart Rate	The casualty has a rapid/slow/normal heart rate.
Respiration Rate	=Respiration Rate	The casualty has a high/low/normal respiration rate.
Breathing	Respiration Rate >= 1 : TRUE Respiration Rate < 1 : FALSE	The casualty is not breathing. Repositioning the airway resulted/did not result in spontaneous breathing.
Capillary Refill	Perfusion Index >= 0.03 : TRUE Perfusion Index < 0.03 : FALSE	The casualty has/does not have normal capillary refill.
Respiratory Distress	Tachypnea Event=TRUE : TRUE Tachypnea Event=FALSE : FALSE	The casualty is in respiratory distress.
Controlled Hemorrhage	Total Hemorrhage Rate < 15 ml/min : TRUE Total Hemorrhage Rate >= 15 ml/min :FALSE	The casualty has a major/minor hemorrhage that was controlled/not controlled.
AVPU	Brain O_2 >= 35 mmHg : ALERT AIS =4 | | 25 < Brain O_2 < 35 mmHg mmHg : VERBAL 15 < Brain O_2 < 25 mmHg mmHg : PAIN AIS Severity >= 5 | | Brain O_2 < 15 mmHg : UNRESPONSIVE	The casualty is alert/responds to voice prompts/responds to pain stimuli/is unresponsive.
Ambulatory	Injury Location =Extremities && AIS >= 3 : FALSE AIS Severity > 3 | | AVPU != Alert = FALSE Otherwise : TRUE	The casualty is ambulatory.

Figure 4: Triage Assessment Text Description Pipeline

- Both approaches produce 4 tags per casualty based on the prompts provided.

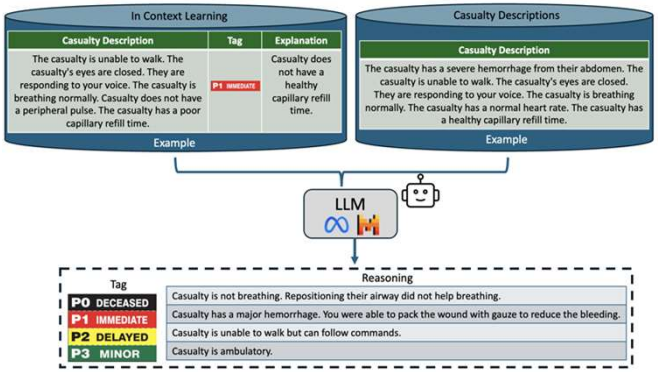


Figure 5: Large Language Model and In Context Learning Pipeline

- The LLM tags were then compared to the gold standard algorithmic tags (see QR code at the top for poster details) for alignment with specific tagging algorithms and survivability.
- The zero-shot and few-shot approaches were also compared to assess the value of the In Context Learning.

RESULTS

- All tagging protocol-aligned LLMs improved tagging performance (Table 1) compared to the unaligned LLM (Table 2).
- The addition of few-shot in-context learning examples further improved performance across all three tagging protocols (Table 2).
- Best alignment was achieved for the few-shot START protocol-aligned LLM (Table 2).

Table 1: Unaligned (Baseline) Performance

Baseline Performance		
Protocol	F1-Score	Accuracy
BCD Sieve	0.454	0.462
SALT	0.409	0.420
START	0.367	0.305

Table 2: Aligned Performance

Protocol	Type	F1 Score	Accuracy
BCD Sieve	Zero-shot	0.877	0.876
	Few-shot	0.893	0.889
SALT	Zero-shot	0.657	0.629
	Few-shot	0.768	0.740
START	Zero-shot	0.758	0.718
	Few-shot	0.953	0.951

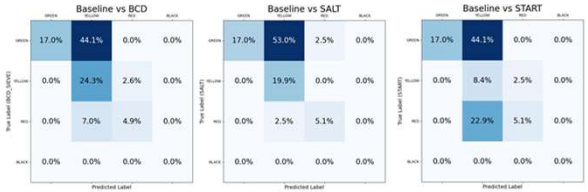


Figure 6: Unaligned (Baseline) LLM Prediction Confusion Matrices

- Baseline LLM tag predictions trended less severe than the ground truth data, resulting in more green and yellow predictions than red (Figure 6).
- Few-shot improved performance across all tag categories for the START protocol (Figure 7). While overall tagging performance improved, an increased miscategorization of yellow tags as red was noted for the BCD and SALT protocols.

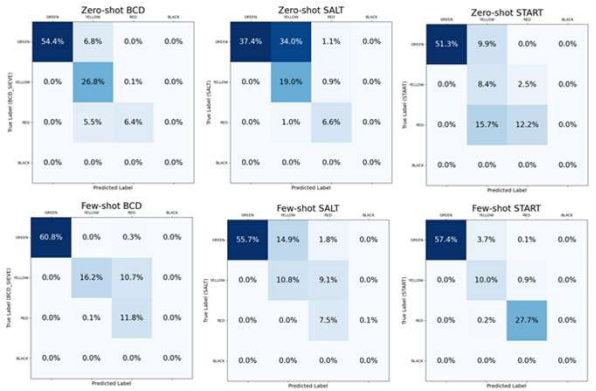


Figure 7: Aligned LLM Prediction Confusion Matrices (top: zero-shot, bottom: few-shot)

DISCUSSION

- With few-shot prompting, aligned LLMs produced casualty tags based on simple descriptions of injuries.
- Unaligned (baseline) tagging performance was not comparable to any of the investigated triage protocols. This indicates that out-of-the-box LLMs may have limited triage tagging capabilities.
- Future work will include increasing variability of the patient injury description to quantify the robustness of this approach.
- A potential under-representation of ground truth yellow tags for the in-context learning examples may have resulted in decreased performance for yellow tags. Future work will explore identifying a method to select a more representative sample of in-context examples to further improve tagging accuracy.

REFERENCES

1. Hu, Brian, Ray, Bill, Leung, Alice, Summerville, Amy, Joy, David, Funk, Christopher, and Basharat, Arslan. "Language Models are Alignable Decision-Makers: Dataset and Application to the Medical Triage Domain." Proc. 2024 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL2024) Industry Track.

2. Clipp RB, Webb JB, Bray A, Veenhuis E, Hu B, Basharat A. Evaluation of Triage Tagging Protocols Using a Synthetic Dataset Representative of Battlefield Injury Profiles. In: *2025 Military Health Services Research Symposium*. ; 2025.

3. Bray A, Webb JB, Enquobahrie A, et al. Pulse Physiology Engine: an Open-Source Software Platform for Computational Modeling of Human Medical Simulation. *SN Compr Clin Med*. Published online March 27, 2019:1-16. doi:10.1007/s42399-019-00053-w.