WHITE PAPER

THE CASE AGAINST AN AI CHATBOT IN EMS

PROTOCOLS ARE LEGAL DOCUMENTS. LARGE LANGUAGE MODELS ARE NOT.

Peter Antevy, MD, FAEMS
Chief Medical Officer, Handtevy

handtevy
Clinical Intelligence for EMS

WWW.HANDTEVY.COM

EXECUTIVE SUMMARY

Across EMS, leaders are being shown demos of generative AI chatbots positioned as clinical decision support for paramedics. The demos are persuasive. The technology is sophisticated. The conversation about whether this class of tool is safe to put in the back of an ambulance, however, has not yet happened with the rigor our patients deserve.

This paper makes a structured case that the conversational AI chatbot, as a clinical interface for paramedics making time-sensitive decisions on real patients, is the wrong tool for the moment. The argument rests on four pillars.

First, the legal architecture of EMS is different from the rest of medicine. EMS protocols are signed, dated, board-approved documents that define the legal scope of paramedic practice. In 21 states, they carry the force of statute. They are admissible as evidence and shape the scope of legal immunity. When a chatbot paraphrases protocol content, it produces clinical guidance that no medical director has approved.

Second, hallucination is not a defect being engineered out of large language models. Researchers at the National University of Singapore have proved mathematically that elimination is impossible. OpenAI's own 2025 research has reached the same conclusion. The current scientific consensus is managing uncertainty, not eliminating it.

Third, the cognitive load argument runs in the wrong direction. A conversational interface asks a task-saturated clinician to type, wait, parse, and reconcile. Structured navigation removes that work. Chat adds it.

Fourth, the chain of accountability that protects patients and clinicians breaks at the chatbot. The vendor disclaims the output. The medical director never approved it. The paramedic acted in good faith. This is the structure of every malpractice case that will eventually be written about using this technology.

The paper closes with a description of the form of clinical AI that does fit the legal and operational architecture of EMS, and how Handtevy's platform implements it.

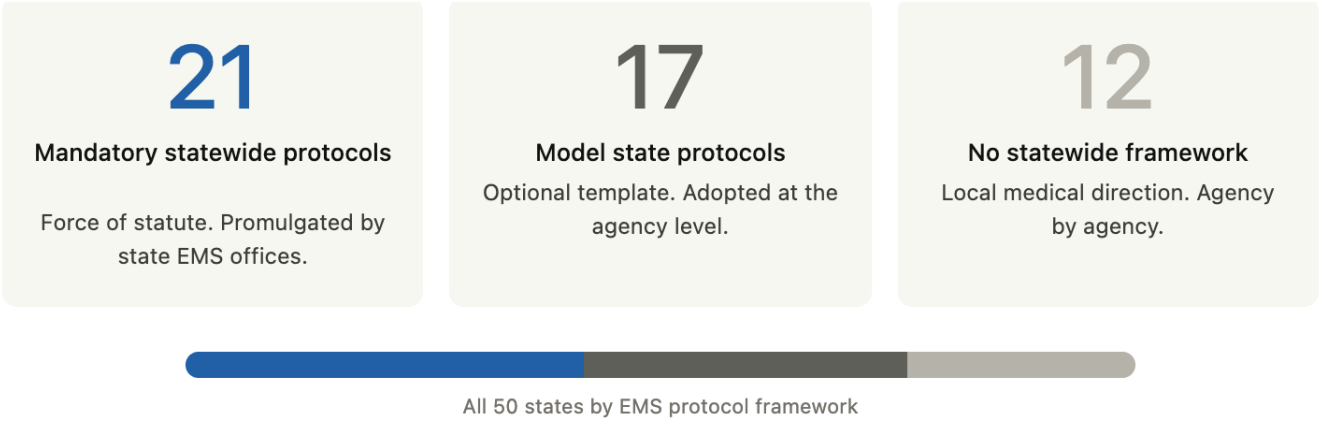
Entering a broad query like 'correct pediatric dose of ketamine' into an AI chatbot is ill-advised and dangerous.

— Douglas M. Wolfberg
Esq., Partner, Page, Wolfberg & Wirth

THE FIVE-YEAR-OLD IN THE AMBULANCE

Picture a paramedic in the back of a moving ambulance. The patient is a five-year-old with a femur fracture, screaming, scared, and in pain. Fentanyl has already been given and the pain is not controlled. The medic opens an app and voices a text into a chatbot, asking for a pediatric ketamine dose. A paragraph of text comes back. It sounds confident. It sounds clinical. It might even be correct.

But the medical director who signed the protocol document for that agency has never seen those exact words, in that exact order, with that exact reasoning attached. In 21 states, the document is signed not by an agency medical director but by the state medical director, with the force of statute and administrative rule behind it. Either way, in that moment, the legal architecture of EMS quietly comes apart. The paramedic is no longer operating under the license and clinical guidance of a physician medical director, but rather a chatbot is guiding care.



In 21 states, the protocol carries the force of statute.

Kupas et al., Prehospital Emergency Care, 2015.

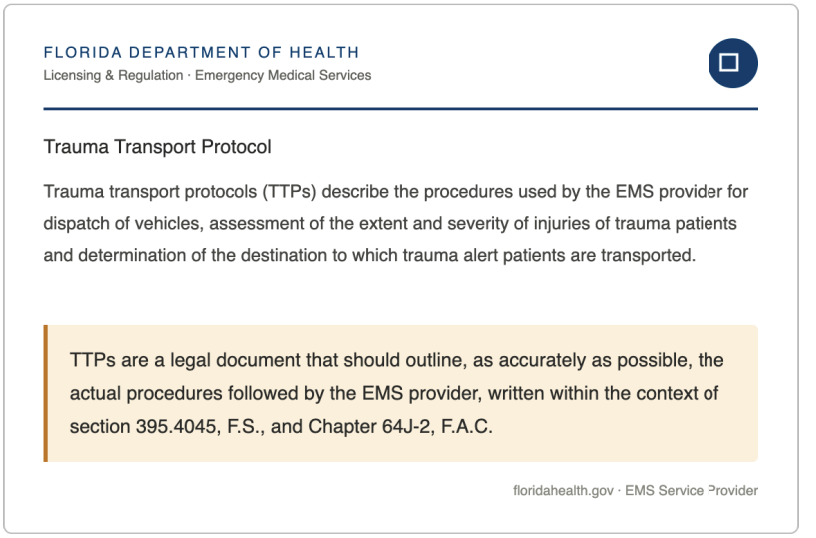
This is the question I keep coming back to as I watch large language model chatbots get pitched to EMS leaders. Not whether the technology is impressive. It is. Not whether it will eventually have a role in our field. It probably will. The question is whether the conversational chatbot, as a clinical interface, is the right tool for the back of an ambulance today. I do not believe it is. And I think it is worth saying so plainly, because the gap between a polished demo and a safe deployment in EMS is wider than most people realize.

I am not a skeptic of artificial intelligence. We use it across our teams every day. It is helping us with research, software development, quality assurance, and protocol review. It is genuinely transformative for non-patient-facing work, and I expect that footprint to keep growing. The question is not whether AI belongs in EMS. It does. It just does not belong in the seat of time-sensitive clinical decisions made on real patients. The reasons are structural, and they are worth walking through.

THE LEGAL ARCHITECTURE OF EMS PROTOCOLS

Start with what an EMS protocol actually is. In the rest of the house of medicine, clinical references are exactly that, references. An emergency physician operates under a license to practice medicine, with the standard of care and individual judgment as the framework. In EMS, the protocol is something different. It is a signed, dated, board-approved document that defines the legal scope ofparamedic practice. A paramedic acting within the protocol is operating within their license. A paramedic acting outside the protocol is not. That document is reviewed by counsel, ratified by medical direction, and audited after the fact in QA review and, sometimes, in court.

This is what makes a protocol something more than a reference. It carries regulatory force where states have given it that weight. It defines the standard of care that a clinician will be measured against if something goes wrong. It is admissible as evidence and serves as the basis for expert testimony in court. And it shapes the scope of legal immunity that an EMS provider is entitled to when they act within it.



Not our framing. The state's.
Florida Department of Health, EMS Service Provider page.

Twenty-one states currently have mandatory statewide EMS protocols promulgated through state EMS offices under regulatory authority (Kupas et al., 2015). An additional seventeen states publish model protocols that local agencies may adopt or modify. In both categories, the document is the product of medical direction acting through a defined chain of authority. The signature on the bottom of the page is the operational manifestation of that authority.

> **"Protocols are subregulatory guidance which certainly can be introduced as evidence."**
>
> **— Douglas M. Wolfberg**
> *Esq., Partner, Page, Wolfberg & Wirth*

When a chatbot generates clinical guidance by paraphrasing, summarizing, or recombining protocol content, it produces text that no medical director has approved. The output looks authoritative because it is written in clinical language, but it does not carry the legal weight of the actual protocol. It is a paraphrase masquerading as policy. That distinction matters every time a paramedic acts on it, and it matters even more when something goes wrong and the case is reconstructed line by line.

WHY HALLUCINATION CANNOT BE ENGINEERED OUT

The second problem is how these systems actually work. Large language models are probabilistic. They generate the next most likely token, then the next, then the next. That means the same question, asked twice, can produce two different answers. Hallucination is not a defect being engineered out.

Researchers at the National University of Singapore have formalized this mathematically and shown, using results from learning theory, that eliminating hallucination in large language models is not just difficult, it is impossible (Xu, Jain, and Kankanhalli, 2024). OpenAI's own 2025 research paper reached a parallel conclusion from a different direction. Hallucinations, they wrote, are a predictable outcome of how these systems are trained and evaluated, not a bug that better engineering will solve. The leading researchers in the field have largely stopped pursuing zero hallucination as a goal. The current consensus is managing uncertainty, not eliminating it.

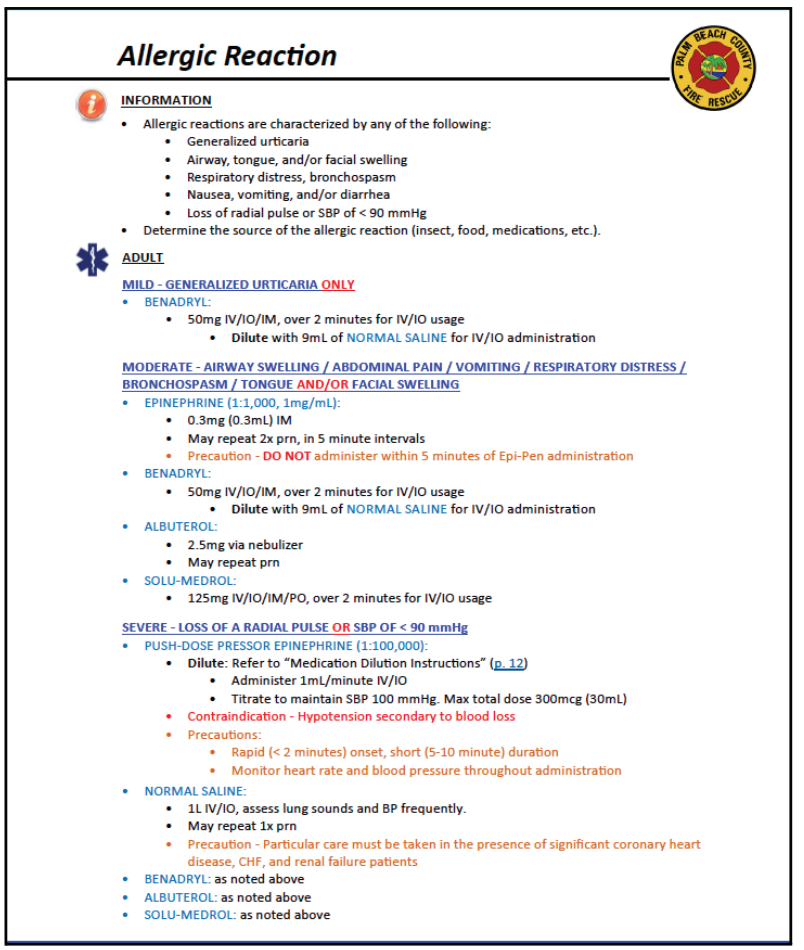
Hallucination is Inevitable:
An Innate Limitation of Large Language Models

The common rebuttal to all of this is that grounding the model to a specific document, in this case your agency's protocol, solves the problem. It does not. Retrieval-augmented generation reduces hallucination. It does not eliminate it. The model can still misread a protocol section, blend content from two similar pathways, or generate a clinically wrong sentence with complete confidence and no flag to the clinician.

There is a more dangerous finding tucked inside the OpenAI work. Models do not just get things wrong. They get things wrong confidently, with low expressed uncertainty and no signal that the answer deserves a second look. In a protocol-based system, where uncertainty should trigger a double-check, a tool that projects confidence it has not earned is not a safety net. It is a liability. In a domain where a misplaced decimal kills a child, "usually right" is not a category that exists.

None of this is to say all AI implementations in EMS are equivalent. Applications that surface medical-director-approved protocols through structured retrieval, with the source document directly referenced and the signed content the clinician sees, are doing something different from a chatbot that generates new sentences. The argument here is specifically about the conversational chatbot interface and the generative paraphrase that interface produces.

There is a deeper problem that is less obvious from a demo. When a physician writes an EMS protocol, it is written from top to bottom with intention. There is a progression. Assessment comes before intervention. Non-pharmacologic options come before drugs. Contraindications come before indications. First-line agents come before second-line. Doses come last, and only after the clinician has been walked through the decision that justifies them. The protocol is not a flat database of facts. It is a structured pathway, and the order is part of the clinical content.



Allergic Reaction

INFORMATION

- Allergic reactions are characterized by any of the following:
  - Generalized urticaria
  - Airway, tongue, and/or facial swelling
  - Respiratory distress, bronchospasm
  - Nausea, vomiting, and/or diarrhea
  - Loss of radial pulse or SBP of < 90 mmHg
- Determine the source of the allergic reaction (insect, food, medications, etc.).

ADULT

MILD - GENERALIZED URTICARIA ONLY

- BENADRYL:
  - 50mg IV/IO/IM, over 2 minutes for IV/IO usage
    - **Dilute** with 9mL of NORMAL SALINE for IV/IO administration

MODERATE - AIRWAY SWELLING / ABDOMINAL PAIN / VOMITING / RESPIRATORY DISTRESS / BRONCHOSPASM / TONGUE AND/OR FACIAL SWELLING

- EPINEPHRINE (1:1,000, 1mg/mL):
  - 0.3mg (0.3mL) IM
  - May repeat 2x prn, in 5 minute intervals
  - Precaution - **DO NOT** administer within 5 minutes of Epi-Pen administration
- BENADRYL:
  - 50mg IV/IO/IM, over 2 minutes for IV/IO usage
    - **Dilute** with 9mL of NORMAL SALINE for IV/IO administration
- ALBUTEROL:
  - 2.5mg via nebulizer
  - May repeat prn
- SOLU-MEDROL:
  - 125mg IV/IO/IM/PO, over 2 minutes for IV/IO usage

SEVERE - LOSS OF A RADIAL PULSE OR SBP OF < 90 mmHg

- PUSH-DOSE PRESSOR EPINEPHRINE (1:100,000):
  - **Dilute**: Refer to "Medication Dilution Instructions" (p. 12)
    - Administer 1mL/minute IV/IO
    - Titrate to maintain SBP 100 mmHg. Max total dose 300mcg (30mL)
  - Contraindication - Hypotension secondary to blood loss
  - Precautions:
    - Rapid (< 2 minutes) onset, short (5-10 minute) duration
    - Monitor heart rate and blood pressure throughout administration
- NORMAL SALINE:
  - 1L IV/IO, assess lung sounds and BP frequently.
  - May repeat 1x prn
  - Precaution - Particular care must be taken in the presence of significant coronary heart disease, CHF, and renal failure patients
- BENADRYL: as noted above
- ALBUTEROL: as noted above
- SOLU-MEDROL: as noted above

A chatbot does not see that. When a medic types "pain dose for a five-year-old," the model treats the protocol as a searchable text corpus, finds the line that matches, and returns it. Every step that came before that line, the assessment, the contraindications, the alternative agents, the recognition that this child might not need ketamine at all, becomes invisible. The clinician gets the answer to the question they asked, not the answer to the clinical situation they are managing.

Ketamine is the cleanest example. In a typical EMS system, ketamine appears in the analgesia protocol, the agitation protocol, the induction protocol, and sometimes in the bronchospasm pathway. Different doses. Different routes. Different concentrations. Different indications. Different contraindications. Ask a chatbot "what is the ketamine dose" and there is no good answer. Either the model picks one, in which case it is wrong roughly as often as it is right, or it lists all of them and asks the clinician to choose, in which case the burden of decision sits exactly where it should not.

THE COGNITIVE LOAD INVERSION

The pitch for AI in EMS is almost always framed as cognitive load reduction. A paramedic working a critical patient is task-saturated. The argument goes that an intelligent assistant should reduce that load. The goal is right. The chatbot interface is not how to get there.

A conversational interface, by design, asks the clinician to do work. Type a question. Wait for a paragraph. Read it. Parse it. Decide whether it applies to the protocol you are operating under. Reconcile the concentration the chatbot returned with the concentration on the truck. Choose between the three doses it offered. That is more cognitive load, not less, and it is being added at the worst possible moment, in a moving vehicle, with one hand, with bad lighting, with a sick patient who cannot wait.

The right interface for the back of an ambulance is the one that gets the clinician to the right page, in the right protocol, in two taps, with no parsing required. Structured navigation removes work. Conversation adds it.

THE ACCOUNTABILITY GAP

The final structural problem is accountability. Twenty years of serving as a medical expert in EMS cases has taught me that when something goes wrong, the case gets reconstructed line by line. Documentation is read aloud. Protocols are introduced as exhibits. Every decision is traced back to the person responsible for it. That chain of accountability is not a bureaucratic detail. It is the architecture that protects clinicians and patients alike.

Today, when a paramedic deviates from protocol, that chain holds. The deviation is documented. The medical director reviews the case. QA flags the variance. Education or remediation follows. The system learns. The clinician is supported. The patient is protected. It works because every step is traceable to a person who stands behind it.

When a chatbot produces a clinical recommendation, the chain breaks at exactly the moment it matters most. Who owns the output? The vendor disclaims it in the terms of service. The medical director never approved it. The paramedic acted on it in good faith because it appeared inside an app the agency adopted. If the recommendation contributed to a bad outcome, the chain of responsibility runs into a wall. That is not a hypothetical concern. It is the structure of every malpractice case that will eventually be written about this technology.

HOW NEW TOOLS EARN THEIR PLACE

All of which leads me to the part that bothers me most as a clinician and a scientist. The way we introduce new interventions into the house of medicine is well established. We define the outcome that matters. We run the trial. We look at the results. We let the data tell us whether the intervention helps, harms, or makes no difference. Then we decide. That discipline is not optional, and it does not change because the new intervention is software.

JAMA Network | Open

Original Investigation | Emergency Medicine

Effect of Machine Learning on Dispatcher Recognition of Out-of-Hospital Cardiac Arrest During Calls to Emergency Medical Services
A Randomized Clinical Trial

Stig Nikolaj Blomberg, MsC; Helle Collatz Christensen, MD, PhD; Freddy Lippert, MD; Annette Kjær Ersbøll, MsC, PhD; Christian Torp-Petersen, MD, PhD; Michael R. Sayre, MD; Peter J. Kudenchuk, MD; Fredrik Folke, MD, PhD

Table 2. Primary and Secondary Outcomes

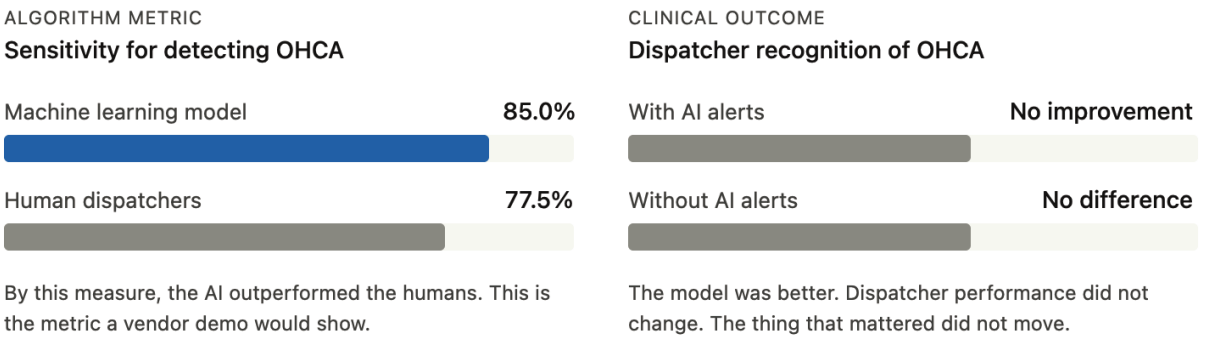
Outcome	Group, mean (SD)		*P* value
	Control	Intervention	
Eligible for analysis, No. (%)	336 (51.5)	318 (48.5)	.48
Call length, min	6.68 (3.39)	6.94 (3.36)	.35
Alert generated from machine learning model, min[a]	1.33 (1.51)	1.39 (1.32)	.60
Recognition of cardiac arrest, No (%)	304 (90.5)	296 (93.7)	.15
Secondary outcomes			
Time to dispatcher recognition, min	1.70 (1.57)	1.71 (1.63)	.90
DA-CPR instructions started, No. (%)	208 (61.9)	206 (64.8)	.47
Time to DA-CPR, min	2.48 (1.89)	2.52 (1.76)	.82

Abbreviation: DA-CPR, dispatcher-assisted cardiopulmonary resuscitation.

[a] Alert shown in intervention group only.

We have a clean recent example of what this looks likewhenitisdoneright.The Copenhagen Emergency Medical Services group ran a double-masked randomized trial of a machine learning model designed to help dispatchers recognize out-of-hospital cardiac arrest during 911 calls (Blomberg et al., JAMA Network Open, 2021). The trial enrolled more than 169,000 calls, including over a thousand confirmed cardiac arrests. The model worked. On the algorithmic metric, it outperformed the human dispatchers, with sensitivity of 85 percent compared to 77.5 percent. By the standard a vendor demo would use, that is a win.

And yet, when the trial measured what actually mattered, whether dispatcher recognition of cardiac arrest improved when the AI was in the loop, there was no difference. The model was better at the task, the dispatchers had access to its alerts, and dispatcher performance did not change. The investigators were direct about it. They wrote that almost all decision support tools driven by artificial intelligence or machine learning have so far failed to improve outcomes in practice. The 2025 AHA guidelines reviewed this evidence and reached a measured conclusion. The technology is promising. It is not ready to be relied on.



The algorithm won. The patients did not.

Blomberg et al., JAMA Network Open, 2021. 169,049 calls. Randomized clinical trial.

That is how this is supposed to work. We test our tools the way we test our drugs. We do not deploy them based on how good they look in the conference hall. The chatbot products being marketed to EMS today have not been studied with anything close to that rigor. There are no randomized trials. There are no patient-level outcomes. There are demos, testimonials, and screenshots. That is not enough.

THE HANDTEVY APPROACH

None of this is an argument against artificial intelligence in EMS. There is real, serious work to be done in this space, and there are use cases where machine learning will help us deliver better care. The point is narrower than that. The conversational chatbot, as a clinical interface for paramedics making time-critical decisions on real patients, is the wrong form. The right form is curated, deterministic, auditable, and approved by the medical director who signs the protocol.

Handtevy was built around that principle. The platform delivers medical-director-approved protocols to clinicians through structured navigation, not conversation. The content the paramedic sees is the content the medical director wrote, in the same order, with the same context, every single time. There is no paraphrase, no generation step, no probabilistic output. The clinician's hand goes to the right page in two taps. The cognitive work is removed at the moment it matters, not added.

That architecture is deliberately the inverse of the chatbot. Where a chatbot is conversational, Handtevy is navigational. Where a chatbot generates, Handtevy retrieves. Where a chatbot produces fresh text every time, Handtevy displays the same protocol every time. Where a chatbot fragments the medical director's authority, Handtevy preserves it. The result is a tool that fits the legal architecture of EMS rather than working against it.

Peer-reviewed evidence supports the model. A 2026 multicenter study published in Prehospital Emergency Care (Dorsett et al.) found that pediatric medication dosing performed through the Handtevy platform was correct in 97 percent of cases, compared to substantially lower accuracy with traditional reference methods. That kind of evidence is what allows a clinical AI platform to earn its place in the field. It is the discipline the conversational chatbot products being marketed to EMS today have not yet undertaken.

RECOMMENDATIONS FOR EMS LEADERS

For medical directors, chiefs, and EMS administrators evaluating clinical AI products, this paper suggests four practical questions to ask of any tool being considered for the patient care setting.

First, is the content the clinician sees the same content the medical director signed, or has it been paraphrased through a language model? If the latter, the legal architecture argument applies. The output is not the protocol. It is a generated paraphrase.

Second, does the interface reduce cognitive load at the bedside, or shift it? A tool that requires the clinician to type, parse, and reconcile is adding work. A tool that gets the clinician to the right page in two taps is removing it.

Third, what is the chain of accountability when the tool's output contributes to a bad outcome? Where does the vendor disclaim, and what does the medical director own? If the chain breaks at the tool, the tool is a risk vector.

Fourth, what is the evidence base? Has the platform been studied in peer-reviewed outcomes research? Has it been measured against the clinical metric that matters, not the algorithmic benchmark? In the absence of that evidence, the burden of proof sits with the vendor.

EMS is not ready for an AI chatbot. Not because we are afraid of the technology, but because we are clinicians and scientists who understand what a protocol is and how new tools earn their place in our field. A protocol is a legal document, written with intention, designed to be navigated, not paraphrased. A new clinical tool is something we study before we deploy it, not after. Until the chatbot can honor both of those principles, the demo is the easy part. The deployment is where people get hurt.

REFERENCES

1. Blomberg SN, Christensen HC, Lippert F, et al. Effect of Machine Learning on Dispatcher Recognition of Out-of-Hospital Cardiac Arrest During Calls to Emergency Medical Services: A Randomized Clinical Trial. JAMA Network Open. 2021;4(1):e2032320.
2. Dorsett M, et al. Accuracy of Pediatric Medication Dosing in Emergency Medical Services Using a Structured Clinical Intelligence Platform. Prehospital Emergency Care. 2026.
3. Kupas DF, Schenk E, Sholl JM, Kamin R. Characteristics of Statewide Protocols for Emergency Medical Services in the United States. Prehospital Emergency Care. 2015;19(2):292-301.
4. OpenAI. Why Language Models Hallucinate. 2025.
5. Xu Z, Jain S, Kankanhalli M. Hallucination is Inevitable: An Innate Limitation of Large Language Models. arXiv. 2024;2401.11817.

ABOUT THE AUTHOR

Peter Antevy, MD, FAEMS, is the Founder and Chief Medical Officer of Handtevy. He is a practicing emergency and EMS physician, Medical Director for Palm Beach County Fire Rescue, and serves on national committees on cardiac arrest care, pediatric resuscitation, and prehospital blood transfusion. He has served as a medical expert in EMS cases for over twenty years.

ABOUT HANDTEVY

Handtevy is a clinical intelligence platform for EMS and hospital settings, focused on point-of-care decision support that preserves the legal authority of medical direction and reduces cognitive load for clinicians at the moment it matters. Handtevy's platform delivers medical-director-approved protocols through structured navigation, with peer-reviewed evidence supporting its accuracy in pediatric medication dosing and other high-risk clinical domains.

For more information, visit **handtevy.com.**