Potential of LLMs in Supporting Call-taking Processes of Emergency Dispatch Centers

Simon Franke
German Red Cross Rhineland-Palatinate e. V.,
Germany

Melanie Reuter-Oppermann
Technical University of Applied Sciences
Würzburg-Schweinfurt,
Germany

Tilo Mentler
Trier University of Applied Sciences,
Germany

ABSTRACT

In recent years, there have been major improvements in the development of large language models (LLMs). These models can process natural language and form the basis of many partially autonomous AI agents. Consequently, they have found their way into many areas of work, including medicine. However, the use of LLMs is associated with certain risks. Working in emergency dispatch and control centers presents particular challenges. Decisions must be made quickly in order to provide people with the right help in case of an emergency. LLMs could support human operators by helping them to evaluate emergency calls, triage them, and alert rescue services. However, the models must be highly reliable for this purpose. This study evaluates the ability of current LLMs (GPT 5.2 and Claude Sonnet 4.6) to classify emergency call transcripts according to their criticality using the ABCDE scheme. Different levels of detail are provided in the guidelines using different prompts. The results show that the models perform well even without detailed guidelines. However, well-designed guidelines can significantly enhance performance. These offer a opportunity to adapt the evaluations to local conditions or continuously improve them during later real-world operation. The study demonstrates a possibility to integrate LLMs in control center processes. Although the use of LLMs is certainly promising, it also involves numerous risks. Therefore, insecurity-critical fields such as emergency services, it is essential to carefully consider whether their use is justified.

Keywords

Agent System, Large Language Model, Call-Taking, Emergency Call, Dispatch Center, Human-AI-Teaming, Triage

INTRODUCTION

Emergency medical call-taking and dispatch centers play a critical role in emergency services. They serve as the initial points of contact for people experiencing medical emergencies and are responsible for evaluating the severity of the situation, determining the appropriate course of action, and coordinating the deployment of rescue services.

The demands placed upon dispatch center employees are considerable, and they meet the criteria for a high responsibility team (Hagemann et al. 2011). Numerous decisions must be made within a limited time frame, with no opportunity to pause and gather additional information or opinions. The consequences for patients can be substantial, and poor decisions can potentially result in fatalities or serious impairment. The work of dispatch centers directly impacts patient care in emergencies. Currently, dispatch centers are facing challenges similar to those faced by many other healthcare facilities, including mounting cost pressures, staff shortages, and an increasing volume of requests. (Luiz et al. 2019)

Due to these high demands for quality and quantity in the operations of dispatch centers, along with the recent advancements in artificial intelligence and associated technologies, there has been an increase in the number of

studies and development efforts that are centered on the potential applications of contemporary technologies in the realm of dispatch center operations and emergency medical services, in general (Fachverband Leitstelle 2022; Gormley et al. 2022; Reuter-Oppermann and Kühl 2021). The range of potential application scenarios is extensive, including systems that facilitate data-driven dispatching of emergency vehicles or prediction of call distribution (Abreu et al. 2023; Rautenstrauss and Schiffer 2026), as well as systems that support call taking by translating or speech recognition (Costa et al. 2023). A wide range of system autonomy is conceivable, from simple support systems to (partially) autonomous agents. (Parasuraman et al. 2000)

Even though the use of fully autonomous AI agents to handle critical emergency calls seems unlikely in the coming years, opportunities to support human operators should be investigated. (Gerdes et al. 2025; EASA 2023) The medical triage of patients has already been investigated in other studies, but often with a focus on classification in a clinical setting. (Nedos et al. 2026; Naderi et al. 2026) Triage based solely on a telephone description of the patient's condition presents its own challenges. The aim is to decide whether and what type of emergency medical assistance the patient requires. (Luiz et al. 2019; Møller et al. 2021)

Therefore, the current study aims to address the following research question:

RQ: How effectively can current large language models perform medical triage based on emergency call transcripts?

The following section summarizes the background information needed to understand the work. This includes an overview of the call-taking process and the opportunities and challenges of using AI in this context, particularly LLMs. The Methods section presents the underlying data, the experimental design, the models used, and the evaluation method. Then, results are presented and discussed, along with limitations of the study. Finally, an outlook on possible future work is provided.

BACKGROUND

The basics of the call-taking process and the collaboration between humans and AI are briefly described below.

Call-taking Process

The tasks can be broadly categorized into two distinct functions: call taking and dispatching. The term "call-taking" refers to the process of responding to and managing emergency calls, including triage of medical emergencies. (Scott et al. 2016) The term "dispatching" refers to the process of allocating individual vehicles to specific operations, providing alerts, and exercising operational control. (Gardett et al. 2013) The steps can be allocated to different organizational responsibilities or executed by a single dispatch center. Similarly, tasks within the dispatch center can be distributed among multiple operators or assigned to a single operator with exclusive responsibility (Trautmann et al. 2022). A more detailed description of the process steps can be found in Franke et al. 2025

The call-taking process is particularly engaging because substantial information must be gathered and evaluated. (Alzayed and Alsardi 2025) In addition to the emergency call itself, which must be managed by the call taker, geodata and connection information about the caller, videos and images, or chat messages are becoming increasingly important and should be considered in the decision-making process. (EENA 2013) This places a substantial mental workload on the call-taker. In addition to handling the call, which includes providing reassurance to the caller, conducting the conversation, and providing first aid instructions, the call-taker must also evaluate and verify the location information, perform medical triage, and then make a decision. Furthermore, the results must undergo transmission to the dispatch system and the requisite documentation to enable the dispatching process. (Alzayed and Alsardi 2025; Laguna et al. 2022; Maletzki et al. 2022) While these processes are critical and even minor details can have serious consequences (e.g., similar street names in the wrong location), there is a lack of redundancy or dual control in call taking. Consequently, errors are often detected incidentally rather than being intercepted in a structured manner. (Bohm and Kurland 2018)

As part of the medical diagnosis of the patient's complaints, the ABCDE scheme can be used. The ABCDE scheme is a structured examination and prioritization scheme for the initial assessment of critically ill or injured patients (Thim et al. 2012; Bruinink et al. 2024). It is used to systematically examine vital functions and other potentially critical systems in a fixed order (Airway, Breathing, Circulation, Disability, Exposure), identify problems, and treat them immediately. A brief overview can be found in Table 1. Based on the ABCDE assessment, a decision can be made as to how critical a patient's condition is. Its use is now widespread, and although the system was not initially developed for dispatch centers, the scheme can be used to evaluate the patient's condition during an emergency call.

Step	Stands for	Focus	Examples
A	Airway	Patency, obstruction	Closed or contracted airway, foreign bodies
B	Breathing	Ventilation, oxygenation	missing or abnormal breathing, skin
C	Circulation	Perfusion, bleeding	Pulse, blood pressure, capillary refill, skin, chest pain
D	Disability	Neurological status	consciousness, pupils, blood glucose, stroke signs
E	Exposure	Full body assessment	inspect for injuries, other pain, temperature

Table 1. Overview ABCDE scheme

Large Language Models in the Context of Call-taking

A large language model (LLM) is a transformer-based pre-trained language model with tens to hundreds of billions of parameters, trained on massive text datasets. Compared to earlier pre-trained language models, LLMs show stronger language understanding and generation capabilities and exhibit emergent abilities that conventional software does not have (Minaee et al. 2024).

The use of LLMs, however, can entail particular risks, especially in safety-critical domains. The outputs may be incorrect or incomplete, and the user cannot verify the model's reasoning process. Therefore, the use of LLMs should be subject to particular scrutiny, especially in safety-critical domains. (Shapira et al. 2026)

LLMs have the potential to play a role in the call-taking process for several reasons. On the one hand, this is due to their capacity to process natural language; on the other hand, it is because LLMs also serve as an essential foundation for autonomous AI systems that can act as agents. These individuals possess a wealth of knowledge, demonstrate advanced reasoning abilities, and are adept at breaking complex goals into more manageable sub-goals using chain-of-thought prompts. These agents are often referred to as the "brain" of the system. In collaboration with the modules responsible for planning, memory, and tool usage, a perception-thinking-action cycle emerges as the basis for agency, in comparison to conventional software tools (Erukude et al. 2025; L. Wang et al. 2024; Raptis et al. 2025).

Advances in LLM development have led to an increase in studies examining how LLMs can be used in medicine. In addition to standard tasks such as transcription and documentation, several studies are focusing in particular on triage in clinical emergency departments (Williams et al. 2024; Seo et al. 2025). Nedos et al. 2026 conducted a study to investigate how well different LLM systems perform in clinical triage. The authors used a comprehensive dataset of real cases and compared the triage classifications of the LLMs with those of emergency physicians. There were significant differences in quality, but none of the models achieved strong agreement. The authors recommend the use of LMMs "as adjunctive tools under clinician supervision rather than autonomous systems in triage" (Nedos et al. 2026). However, in this study, no prompts were adjusted, and no special architectures were used to improve performance.

Naderi et al. 2026 also tested the performance of LLMs in various clinical tasks in a comprehensive study. In the first step, the clinical knowledge of 18 LLMs was tested; in the second step, the clinical reasoning of five models was tested across several sub-tasks. This revealed a clear difference in quality between different models and model generations, but it also showed that other models are better suited to various tasks.

(Savage et al. 2024) experiment with different prompting approaches. They compare classic chain-of-thought prompting to diagnostic reasoning prompting, in which the model is given an explicit thought process intended to mimic the clinical decision-making of medical professionals. Although the results do not show a significant improvement in diagnostic accuracy, the study suggests that such diagnostic reasoning prompts generate a comprehensible clinical line of reasoning. This makes it easier to assess the plausibility of LLM responses. Since the survey was conducted with older models (GPT-3.5 and GPT-4), it can also be assumed that newer models may be able to implement this type of structured reasoning even more reliably.

METHODS

To answer the research question of how LLMs perform in call taking, several test runs were conducted as part of this study. The underlying data and the experimental design are described below.

Data

The present study is predicated on a set of 60 emergency call transcripts. These calls are realistic, but do not constitute actual data. A key benefit of this approach is that it allows publication of transcripts without adhering to guidelines on handling sensitive patient data.

To achieve a broad range of results, the data was generated using various methods. For calls 1-30, emergency service personnel were tasked with developing realistic scenarios. The emergency calls were recorded as part of a simulated emergency call and subsequently transcribed to emulate the challenges associated with transcription and human communication over the telephone. The calls 31-60 were generated using another approach: Case categories and severity levels were randomly generated following a realistic distribution due to german emergency cases as described by Sefrin et al. 2025. In the second step, a GPT5.2 model was tasked with developing a scenario that included the current event, the scene, the on-site situation, and the patient's medical history. In the third step, the model was tasked with simulating the caller's actions in an emergency scenario. The call-taker role was executed by actual call-takers, the call-taking process was simulated as a chat.

A brief overview of an example call is given in Figure 1. The full transcripts of the emergency calls (in their original German and with English translations), the case descriptions, and the generation code are available in the accompanying data.

Paramedics evaluated all emergency calls to obtain the most reliable data possible. All calls were assessed according to the ABCDE categories based on the severity scale (0-3). The three categories are predicated on the notion that a binary classification (critical, non-critical) is inherently more straightforward for an AI service than a more nuanced distinction. Since Germany employs a physician-staffed emergency medical service system, control centers are obligated to determine not only whether a patient requires acute emergency medical assistance, but also whether a paramedic-staffed ambulance is adequate or if an emergency physician must be dispatched to the scene (Sefrin et al. 2025). This approach is intended to ensure that the results are more realistic.

In the initial phase, the conversations were evaluated independently by two paramedics. When there was a divergence of opinion, the case was referred to a third paramedic for evaluation. The decision was then made by a majority vote. In a limited number of instances (seven individual categories), a consensus could not be reached, and these cases were resolved through a consensus in a case discussion. The confidence level achieved is indicated in the data (1 = unanimity, 0.5 = majority decision required, 0 = case discussion required).

Severity	Description
0	No information available; no conclusions possible
1	No or only minor impairment suspected
2	Significant impairment present; no acute life threat suspected (ambulance indicated)
3	Severe impairment; vital threat possible or likely; immediate intervention required (urgent ambulance, possibly physician response unit)

Table 2. Severity classification used for ABCDE assessment.

The differentiation rate is relatively high. Following the initial two evaluations, 232 out of 300 categories (for 60 emergency calls, each containing one ABCDE category) were rated equally. It should be noted that the values differ, particularly for E (Exposure/Environment), which had only 24 matching values. However, evaluations that differ by more than one value (the difference between non-critical and highly critical categories) occur in only 10 out of 600 categories, and 5 of these also fall into category E.

Experimental Approach

The experiments aim to test how LLMs evaluate emergency calls according to the ABCDE scheme. In all experiments, the models were tasked with evaluating emergency calls individually within each ABCDE category and classifying them into levels 0-3, as shown in Table 2. In all experiments except the first, the models were also asked to indicate the symptoms found as a reference and to evaluate the statement with a confidence level.

The experiments differ mainly in the scope of the prompt. In experiments 3 and 4, the models are given a much more comprehensive description of the task and the symptoms, with experiment 4 even providing clear, extensive examples. All prompts and the code can be found in the accompanying data material. The ABCDE scheme is particularly important here for two reasons: If the LLM's only task were to evaluate an emergency call and recommend sending an ambulance or an emergency doctor, the traceability of the decision would be reduced. In a

subsequent real-world application, a human call-taker would likely face challenges in interpreting this statement. They would probably either place unquestioning trust in the model or disregard the result.

However, there are also technical reasons for dividing the task into smaller steps, thereby controlling the LLM's reasoning. If the task is divided into individual steps, such as "find symptoms that match the category" and "evaluate the category," then a chain of thoughts can be achieved, possibly even diagnostic reasoning as described by Savage, which is likely to produce better results (Wei et al. 2022; Chen et al. 2022; Savage et al. 2024). The division into the five categories of the ABCDE scheme has another practical advantage. This converts 60 test emergency calls into 300 decisions per test run, increasing the significance of the results.

The request to rate the LLM's statement with a confidence score can also lead to expanded reasoning. If the confidence values are found to be reliable, this could also result in improved quality of the results in later practical use. Greater uncertainties could be displayed to the human operator or trigger a re-examination by another LLM.

Models

The experiments are conducted using different models, as previous studies have shown differences in quality across models and task types. Naderi et al. have demonstrated that the Claude Sonnet 3.5 & 4 models from Anthropic and GPT 5 from OpenAI achieved the best results in reasoning (Naderi et al. 2026). Claude Sonnet 4 in particular also achieved good results in Nedos et al. 2026. The test runs were therefore carried out on the two current versions of the models (GPT5.2 and Claude Sonnet 4.6).

The temperature parameter can be used to influence how the models operate. Lower values produce more deterministic outputs, while higher values increase output diversity, which is why the temperature is often controlled or varied in experiments to assess robustness and reproducibility (Holtzman et al. 2019; X. Wang et al. 2022; Liang et al. 2022). In high-risk areas, selecting a low temperature to achieve reliably safe results is undoubtedly an interesting approach. However, higher temperatures may improve the reasoning ability (Joshi et al. 2026). Therefore, experiments were conducted at different temperatures.

Example Transcript (Call 52, shortened)

CT: Emergency services, what is the location?
C: In Cologne, Porz — ground-floor apartment. Please come quickly, my husband is lying in the bathroom and is not responding!
CT: If you shake your husband, does he not respond at all?
C: No, not at all. . . I shook his shoulders and called, he doesn't open his eyes. He is completely limp.
CT: Check whether he is still breathing. Place your ear near his mouth and nose, and look at his chest.
C: No, there is no air, I can't hear anything and his chest is not moving. . . not at all.
CT: Look in his mouth. Tilt his head back slightly.
C: Nothing in there, just some saliva. Head tilted back, but he still isn't breathing.
CT: Does anyone there know how to perform CPR?
C: The neighbor says he can. . . he's already next to him. Should he start? Please!
CT: Start CPR as quickly as possible.
C: We pulled him into the hallway. The neighbor is pressing on his chest, very fast, repeatedly..

Figure 1. Shortened example of an emergency call transcript (Call 33). CT = calltaker, C = caller.

Experiment runs

Table 3 illustrates the configuration of the individual experiments. The primary distinction lies in the prompt strategy, which evolves from rudimentary to complex and extensive from experiments 1 to 5. Examples of the prompts are given in Figure 3. The detailed prompts can be found in the accompanying data. The number of tokens further highlights this phenomenon. A preliminary estimate is available in the table. An increase in token counts results in a concomitant increase in computing effort and associated costs. Experiments 3 to 5, in particular, have a considerably higher token count, primarily due to the 5-fold LLM call for each transcript.

The guidelines exhibit variations in both size and quality. Although the guidelines for Experiment 4 are preliminary compilations, those for Experiment 5 have undergone substantial modifications. For example, specific requirements and rule sets for alerting emergency physicians in the German rescue service were utilized for this purpose. (DBRD 2024; Hessen n.d.; Sefrin et al. 2025; BAEK 2023)

Exp.	Prompt Strategy	Req.	Guidelines Detail	Tok. GPT-5.2	Tok. Claude 4.6
1	Single prompt for full ABCDE classification	1	Minimal prompt with short ABCDE description	960	1290
2	Single prompt with detailed ABCDE guidelines	1	Explicit symptom descriptions and critical indicators for each category	1690	2200
3	Separate prompts per ABCDE category	5	Short guidelines explaining each category and severity scale	5800	7640
4	Separate prompts per ABCDE category	5	Guidelines with more detailed explanations of symptoms and severity indicators but not adjusted	6150	7950
5	Separate prompts per ABCDE category	5	Detailed guidelines including explicit examples of medical situations, reviewed and adjusted	6700	8650

Table 3. Overview of the evaluated prompting strategies.

Run	Model	Temperature
1	GPT 5.2	0
2	GPT 5.2	0.5
3	GPT 5.2	1
4	Claude Sonnet 4.6	0
5	Claude Sonnet 4.6	0.5
6	Claude Sonnet 4.6	1

Table 4. Model configurations used in the evaluation runs.

Evaluation Metrics

Model predictions were compared with the reference labels (gold standard) for each ABCDE category. Prediction deviations were evaluated using a predefined penalty matrix 2 that assigns costs to the differences between the predicted and reference severity levels. Let $P(y_i, \hat{y}_i)$ denote the penalty obtained from the penalty matrix shown in Figure 2 for reference severity y_i and predicted severity $\hat{y}_i$. For a run consisting of N evaluated predictions, the total penalty E is defined as

$$E = \sum_{i=1}^{N} P(y_i, \hat{y}_i)$$

The structure of the matrix is inspired by the squared error principle, which assigns disproportionately higher penalties for larger deviations compared to minor errors. However, the matrix was adapted to reflect operational priorities in emergency call triage. In particular, undertriage (predicting a lower severity than the actual condition) is penalized more strongly than overtriage, since underestimating the seriousness of a medical emergency may lead to insufficient resource allocation and delayed treatment.

To also account for model confidence, a confidence-weighted error was defined. Let $e \geq 0$ denote the severity error obtained from the penalty matrix and $c \in [0, 1]$ the confidence reported by the model. The confidence-weighted error E_c is defined as

$$E_c(e, c) = \begin{cases} 1 - c & \text{if } e = 0 \\ e \cdot c & \text{if } e > 0 \end{cases}$$

RESULTS

The results of the test runs are displayed in the matrices Figure 4, Figure 5, and Figure 6 . As illustrated in Figure 4, the error sum E is displayed for all test runs. As illustrated in Figure 5,, the confidence-weighted error E_c for experiments 2-5 is shown, given that experiment 1 does not provide a confidence scale. Figure 6 displays the maximum penalties. When the value 9 appears more than once, the quantity is indicated.

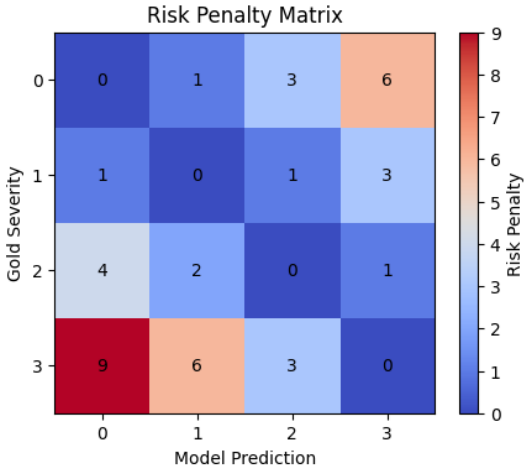

Figure 2. Risk penalty matrix.

Experiment	Mean E	Mean E_c
1	136.17	–
2	131.33	140.00
3	152.33	160.83
4	167.50	159.33
5	**108.17**	**123.33**

Table 5. Mean error E and confidence error E_c per experiment.

Table 5 displays the mean E values and standard deviation of each experiment. Experiment 5 performs considerably better than the other experiments. However, it is surprising that experiments 1 & 2 perform considerably better than experiments 3 & 4, considering that experiments 1 & 2 only require one request per transcript, while experiments 3–5 require five LLM calls for each call.

A detailed examination of the models reveals substantial variations in the outcomes, underscoring the need for a systematic approach to analysis. A comparison of the performance metrics reveals that Claude Sonnet 4.6 exhibited noticeably superior performance compared to GPT 5.2. The differences are particularly evident in experiments 2, 3, and 5. In the absence of detailed instructions in Experiment 1 and due to the lower quality of the guidelines in Experiment 4, both models exhibited similar performance metrics.

If we include E_c in Figure 5, we can also clearly see the difference in performance between the models and experiments. However, GPT 5.2. performs similarly well to Claude in Experiment 5.

It is also important to look at the maximum errors per run in Figure 6. As shown in Figure 2, an error of 9 indicates that no information was provided in a highly critical situation, whereas an error of 6 indicates that such a situation was assessed as non-critical. Here, it is clear that Claude performed worse than GPT 5.2 in Experiment 1. Both models performed the worst in Experiment 4.

It should be noted that both models in Experiment 5 exhibit a maximum of two or three penalty points, indicating an absence of substantial under- or over-triage. A thorough examination of the models reveals substantial variations in the outcomes, particularly in certain instances. A comparative analysis revealed that Claude Sonnet 4.6 exhibited superior performance in terms of average maximum penalty compared to GPT 5.2. The disparities are particularly evident in experiments 2, 3, and 5. In the absence of detailed instructions in experiment 1 and with the lower-quality guidelines in experiment 4, the performances of both models are comparable.

When comparing the number of errors in the human-generated calls 1–30 to those in the human-AI-generated calls 31–60, there is no noticeable difference. Overall, 52% of the errors occur in calls 31–60, although these calls were generated in collaboration with GPT, on average they are not evaluated any better by the AI models.

DISCUSSION

The following sections contain an interpretation of the above results and clarify the limitations of this study.

Experiment 1 — Minimal Prompt (complete)

```
You are an expert in analysing emergency
calls.  You should evaluate the call
transcript according the ABCDE schema.
ABCDE Schema:  Airway, Breathing,
Circulation, Disability, Exposure.
Give every category a severity score
from 0 to 3.
0 -- No information available; no
conclusions possible
1 -- No or only minor impairment
suspected
2 -- Significant impairment present;
no acute life threat suspected
(ambulance indicated)
3 -- Severe impairment; vital threat
possible or likely; immediate
intervention required (urgent
ambulance, possibly physician
response unit)
```

Experiment 5 — Detailed Guideline (Cat. A, excerpt)

A – Airway Assessment

Purpose: Determine the airway status (patent, threatened, obstructed) based on caller descriptions. Breathing is not part of this task.

Evaluation Rules

Focus solely on the airway. A clear airway is present if the patient is alert and speaking in full sentences. A threatened or obstructed airway may occur with:

- Reduced consciousness
- Stridor or abnormal airway sounds
- Swelling in the mouth/throat
- Obstruction (blood, saliva, vomit, foreign body)

Severity Examples

Sev. 2: Mild airway restriction; minor allergic reaction without swelling.

Sev. 3: Sudden slurred speech, communication impossible; rapidly swelling airway; choking with ineffective breathing; severe oral/throat bleeding; reanimation.

Confidence: Score 0.0–1.0 based on transcript evidence. Incorrect high-confidence outputs incur high penalty.

Full guidelines for all five categories in the repository.

Figure 3. Comparison of prompting strategies: Experiment 1 (left) uses a minimal prompt that relies on the model's medical training knowledge. Experiment 5 (right, excerpt for Category A only) provides explicit evaluation rules.

Model	Temp	Mean E	Mean E_c
Claude Sonnet 4.6	0.0	**127.60**	**139.50**
Claude Sonnet 4.6	0.5	128.40	**139.50**
Claude Sonnet 4.6	1.0	133.00	141.75
GPT 5.2	0.0	149.60	151.00
GPT 5.2	0.5	147.20	152.00
GPT 5.2	1.0	148.80	151.50

Table 6. Average error E and confidence error E_c by model and temperature.

Interpretation

The experimental results demonstrated the efficacy of LLM in assessing emergency calls. This study demonstrated that selecting a model and configuring the architecture are paramount to the quality of the results. In the experimental phase, the Claude Sonnet 4.6 model demonstrated slightly superior performance across most trials. In contrast, the impact of temperature on the test run results was only marginal.

A notable observation is that experiments 3 and 4 with distinguished calls for each ABCDE category do not outperform experiments 1&2 where a single prompt covers all five categories at once. Similarly, it may be counterintuitive that GPT 5.2 performs worse in experiment 2 with a more detailed prompt compared to experiment 1. It can be assumed that the ABCDE-scheme, as a widely used concept in medical literature, is likely included in the LLM's training data. Prompts influence the output of LLMs, and it appears that guidelines that are not elaborated in sufficient detail, when included as part of the prompts, have a negative impact on quality. (Naderi et al. 2026)

The selection of the prompt was the most influential factor in determining the quality of the results. A well-designed guideline can significantly improve outcomes, as evidenced by the comparison between Experiments 4 and 5. A suboptimal design of the guideline can result in a substantial reduction in quality. It appears rational to permit the models to operate with only brief prompts and to depend on existing medical "knowledge." However, if a more comprehensive classification is required, as is the case with the German EMS, then more detailed guidelines are useful.

However, a comparison of Experiment 1 and Experiment 5 also shows that a well-developed guideline and a massive increase in tokens, and thus costs, only result in a relatively small increase in quality.

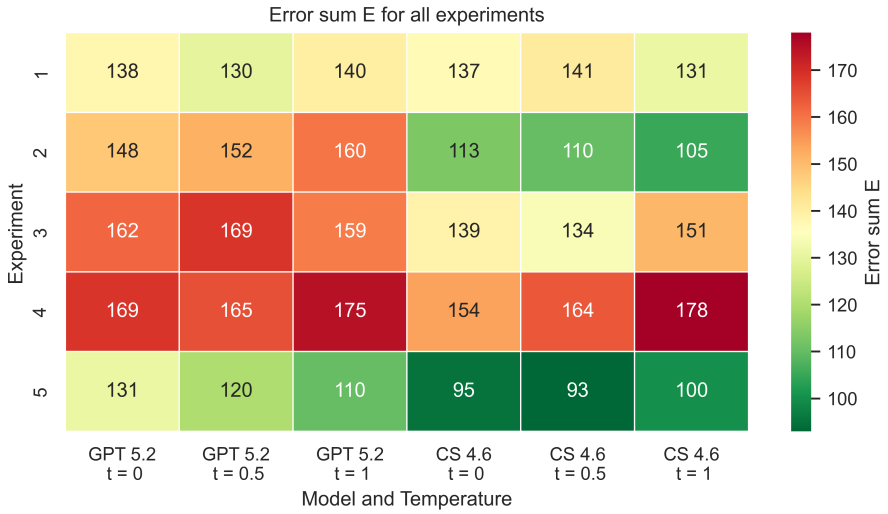

Figure 4. Total error E per per experiment and model (green = best results).

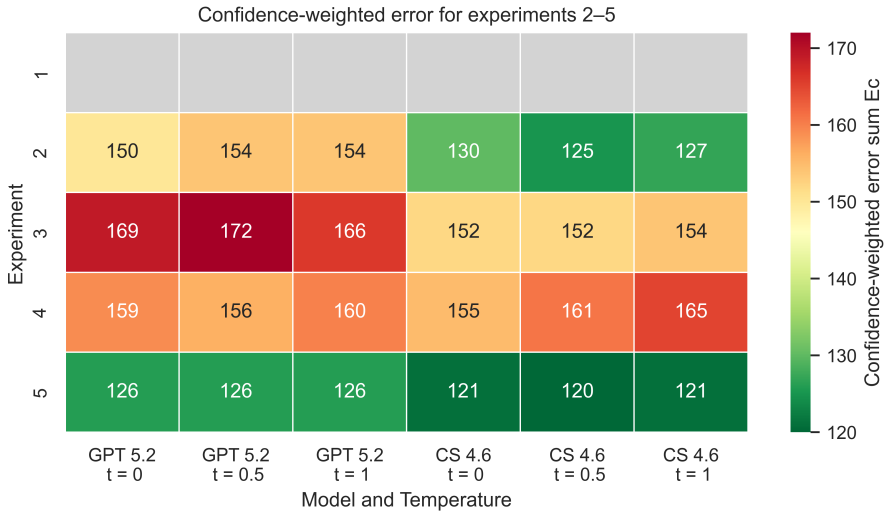

Figure 5. Confidence-weighted error E_c per experiment and model (green = best results).

A notable benefit of guidelines is their capacity for subsequent refinement during actual operational use without requiring modifications to the fundamental model architecture. This flexibility enables the adaptation of distinct sets of guidelines to individual control centers, fostering continuous improvement. A salient example of this adaptability arises when it becomes evident that specific cases are frequently misclassified as either over- or under-triaged.

Overall, it must be acknowledged that even the best results of experiment 5 and Claude Sonnet 4.6 model shows weaknesses. The maximum penalty of 2 shows that there are no cases of extreme under- or over-triage, but the test results differ from the human classifications, that is shown by the error sums E between 93 and 100. While the findings suggest that LLMs will be capable of supporting work in dispatch centers in the future, the results of this study also demonstrate, as previously asserted by Nedos et al. 2026, that their utilization in high-risk environments should be exclusively permitted under the direct supervision of a human operator who possesses the necessary expertise.

Limitations

As a work in progress, the limitations of this study must be acknowledged. The current data basis consists only of realistic but synthetic data and a limited set of 60 case studies. As described in the data-section, the first 30 cases were developed and recorded by emergency personnel. The following cases were developed in collaboration with an LLM, but through a multi-step process involving a human call taker. All cases were evaluated by human paramedics and deemed realistic. Nevertheless, the work with this data does not replace an analysis based on actual calls. The

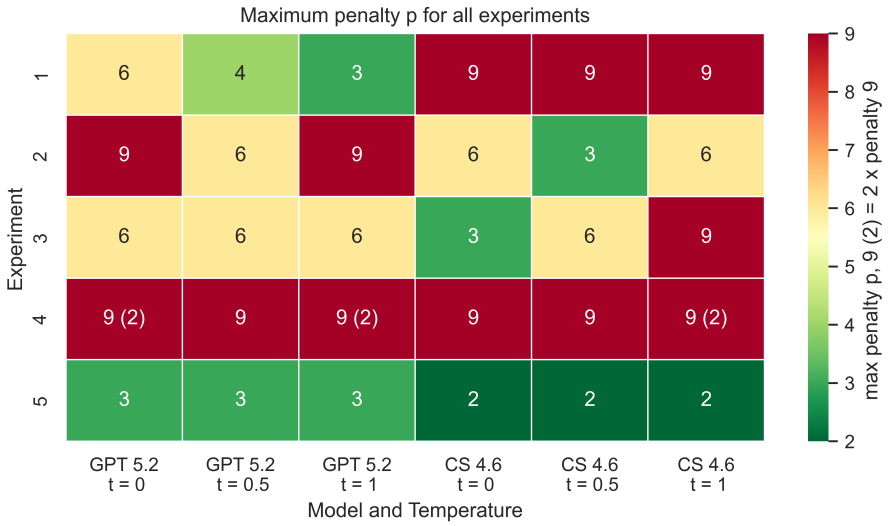

Figure 6. Maximum penalty per experiment and model (green = best results).

use of authentic emergency call data for future research is envisioned, but legal challenges must still be overcome. This evaluation is imperative for fully assessing use in real-life scenarios and should therefore be conducted.

This study examined only two distinct models. The selection was based on previous experience from other studies; these are models of the latest generation and from leading providers. Nevertheless, the incorporation of additional models in subsequent studies is recommended, as previous research has demonstrated that model selection significantly influences the quality of results. In addition, further studies should be conducted to determine whether a more complex multi-prompt solution offers real advantages over a single-prompt approach, or whether the approach can be further improved.

To this point, the scope of this study has been limited to rudimentary architectures comprising a model and a prompt, despite the complexity inherent in the comprehensive guidelines. Subsequent research endeavors should prioritize the in-depth exploration of these architectures, with potential avenues for exploration including their conversion into multi-agent systems. In such systems, models would complement and control one another, or multiple results would be compared. Furthermore, confidence could play a role in human-AI teaming, which may be important for the effective integration of AI into human-centered tasks.

Before such a system can be deployed in practice, further research and development work is needed. In particular, human-AI collaboration remains largely an unresolved issue. This study demonstrates that LLMs offer potential for supporting triage, but should be closely monitored by human operators. To this end, practical methods must be developed to ensure that human operators can verify the results. Furthermore, our experiments have only ever involved completed calls. A challenge remains when LLMs are to be integrated into live operations. In that case, continuously collected information must be incorporated into the triage process, as this cannot take place only after the call has ended.

CONCLUSION

This study aims to investigate the effectiveness of current large language models in medical triage, using emergency call transcripts as the basis for evaluation. Realistic emergency call data was created for the study and evaluated using the ABCDE scheme. This data was submitted to the GPT-5.2 and Claude Sonnet-4.6 models for categorization in various test runs. The test runs differed in terms of both the number of queries per transcript and the modified prompts. Different detailed instructions were provided. The results were evaluated using a weighted error that considers the potential impact of under-triage on patients.

Notably, simple queries to the models without detailed guidelines in the prompts. This suggests that the models have already learned effective medical reasoning from their training data. However, the best results were achieved using high-quality prompts with detailed guidelines. These guidelines not only improve the significance of the results, but can also be used to adapt applications to local conditions or continuously improve them.

Overall, the results were consistent with those reported in previous studies on the performance of LLMs in medical triage of emergency patients in a hospital setting. (Nedos et al. 2026; Naderi et al. 2026) If the right framework

conditions are created to meet the high requirements that apply when LLMs are used in safety-critical domains, they might be able to support human operators.

DATA AVAILABILITY

The dataset, prompts and code used in this study are available on GitHub:

github.com/franke94/LLM_in_calltaking_triage

REFERENCES

Abreu, P., Santos, D., and Barbosa-Povoa, A. (2023). "Data-driven forecasting for operational planning of emergency medical services". In: *Socio-Economic Planning Sciences* 86, p. 101492.

Alzayed, M. A. and Alsardi, N. (2025). "Dispatch Under Pressure: An Investigation into the Cognitive Load of Kuwait's Emergency Responders". In: *Journal of Engineering Research*.

BAEK (2023). *Empfehlungen für einen Indikationskatalog für den Notarzteinsatz*.

Bohm, K. and Kurland, L. (2018). "The accuracy of medical dispatch - a systematic review". In: *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine* 26.1, p. 94.

Bruinink, L. J., Linders, M., Boode, W. P. d., Fluit, C. R., and Hogeveen, M. (2024). "The ABCDE approach in critically ill patients: A scoping review of assessment tools, adherence and reported outcomes". In: *Resuscitation Plus* 20, p. 100763.

Chen, W., Ma, X., Wang, X., and Cohen, W. W. (2022). "Program of Thoughts Prompting: Disentangling Computation from Reasoning for Numerical Reasoning Tasks". In: *arXiv*. eprint: 2211.12588.

Costa, D. B., Pinna, F. C. d. A., Joiner, A. P., Rice, B., Souza, J. V. P. d., Gabella, J. L., Andrade, L., Vissoci, J. R. N., and Néto, J. C. (2023). "AI-based approach for transcribing and classifying unstructured emergency call data: A methodological proposal". In: *PLOS Digital Health* 2.12, e0000406.

DBRD, D. B. R. (2024). *Notarztindikations- Katalog 2024 als Handlungsempfehlung für Disponenten in Rettungsleitstellen*.

EASA (2023). *ARTIFICIAL INTELLIGENCE ROADMAP 2.0*.

EENA (2013). *Next Generation 112 Long Term Definition*.

Erukude, S. T., Veluru, S. R., and Marella, V. C. (2025). "AGENTIC AI - THE RISE OF AUTONOMOUS INTELLIGENT AGENTS IN THE ERA OF LLMS". In: *Indian Journal of ComputerScience and Engineering (IJCSE)*.

Fachverband Leitstelle (2022). *Positionspapier Maschinelles Lernen und Künstliche Intelligenz in BOS-Leitstellen"*.

Franke, S., Elsenbast, C., and Mentler, T. (2025). "Structured and Standardized Emergency Call Systems from an Automation Research Perspective". In: *Proceedings of the International ISCRAM Conference*.

Gardett, I., Clawson, J., Scott, G., Barron, T., Patterson, B., and Olola, C. (2013). "Past, present, and future of emergency dispatch research: a systematic literature review". In: *Emergency Medicine Journal* 33.9, e4.1–e4.

Gerdes, I., Jameel, M., Materne, L. J., and Bruder, C. (2025). "Synergies in the Skies: Situation Awareness and Shared Mental Model in Digital-Human Air Traffic Control Teams". In: *Aerospace* 12.6, p. 472.

Gormley, K., Lockhart, K., and Isaac, J. (2022). "Using natural language processing in facilitating pre-hospital telephone triage of emergency calls". In: *British Paramedic Journal* 7.2, pp. 31–37.

Hagemann, V., Kluge, A., and Ritzmann, S. (2011). "High Responsibility Teams – Eine systematische Analyse von Teamarbeitskontexten für einen effektiven Kompetenzerwerb". In: *Psychologie des Alltagshandelns* 4.1.

Hessen (n.d.). *Notarztindikationskatalog*.

Holtzman, A., Buys, J., Du, L., Forbes, M., and Choi, Y. (2019). "The Curious Case of Neural Text Degeneration". In: *arXiv*. eprint: 1904.09751.

Joshi, T., Aggarwal, S., Saha, A., Pandey, A., Dhoot, S., Rai, V., Goswami, R., Chadha, A., Jain, V., and Das, A. (2026). "Stochastic CHAOS: Why Deterministic Inference Kills, and Distributional Variability Is the Heartbeat of Artifical Cognition". In: *arXiv*. eprint: 2601.07239.

Laguna, M., Chilimoniuk, B., Purc, E., and Kulczycka, K. (2022). "Job-Related Affective Well-Being in Emergency Medical Dispatchers: The Role of Workload, Job Autonomy, and Performance Feedback". In: *Advances in Cognitive Psychology* 18.4, pp. 243–250.

Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., et al. (2022). "Holistic Evaluation of Language Models". In: *arXiv*. eprint: 2211.09110.

Luiz, T., Marung, H., Pollach, G., and Hackstein, A. (2019). "Implementierungsgrad der strukturierten Notrufabfrage in deutschen Leitstellen und Auswirkungen ihrer Einführung". In: *Der Anaesthesist* 68.5, pp. 282–293.

Maletzki, C., Rietzke, E., and Bergmann, R. (2022). "Utilizing Expert Knowledge to Support Medical Emergency Call Handling". In: *8th Workshop on Formal and Cognitive Reasoning*.

Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., and Gao, J. (2024). "Large Language Models: A Survey". In: *arXiv*. eprint: 2402.06196.

Møller, T. P., Jensen, H. G., Viereck, S. o., Lippert, F., and Østergaard, D. (2021). "Medical dispatchers' perception of the interaction with the caller during emergency calls – a qualitative study". In: *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine* 29.1, p. 45.

Naderi, B., Liu, L., Ghandehari, A., Khoshons, D., Taylor, R. A., Bhavsar, N., Balasubramanian, S., Tanouye, R., Creech, N., Davidson, C., et al. (2026). "The role of large language models in emergency care: a comprehensive benchmarking study". In: *npj Artificial Intelligence* 2.1, p. 24.

Nedos, I., Zagalioti, S.-C., Kofos, C., Katsikidou, T., Vellidou, D., Astrinakis, K., Karagiannis, I., Giannakopoulos, P., Michaloudi, S., Apostolopoulou, A., et al. (2026). "Is Artificial Intelligence Ready for Emergency Department Triage? A Retrospective Evaluation of Multiple Large Language Models in 39,375 Patients at a University Emergency Department". In: *Journal of Clinical Medicine* 15.4, p. 1512.

Parasuraman, R., Sheridan, T., and Wickens, C. (2000). "A model for types and levels of human interaction with automation". In: *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans* 30.3, pp. 286–297.

Raptis, E. K., Kapoutsis, A. C., and Kosmatopoulos, E. B. (2025). "Agentic LLM-based robotic systems for real-world applications: a review on their agenticness and ethics". In: *Frontiers in Robotics and AI* 12, p. 1605405.

Rautenstrauss, M. and Schiffer, M. (2026). "Ambulance demand prediction via convolutional neural networks". In: *Operations Research, Data Analytics and Logistics* 46, p. 200497.

Reuter-Oppermann, M. and Kühl, N. (2021). "Artificial Intelligence for Healthcare Logistics: An Overview and Research Agenda". In: *Artificial Intelligence and Data Mining in Healthcare*. Ed. by M. Masmoudi, B. Jarboui, and P. Siarry. Cham: Springer International Publishing, pp. 1–22.

Savage, T., Nayak, A., Gallo, R., Rangan, E., and Chen, J. H. (2024). "Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine". In: *npj Digital Medicine* 7.1, p. 20.

Scott, G., Olola, C., Toxopeus, C., Clawson, J., Johnson, A., Schultz, B., Miller, K., Richmond, N., Robinson, D., Zavadsky, M., et al. (2016). "Characterization of call prioritization time in a Medical Priority Dispatch System". In: — 4.1.

Sefrin, P., Händlmeyer, A., and Kast, W. (2025). "Leistungen des Notfall-Rettungsdienstes". In: *Der Notarzt* 31.04, S34–S48.

Seo, J. W., Park, S.-J., Kim, Y. J., Kim, J.-Y., Kim, K. G., and Yoon, Y.-H. (2025). "Artificial intelligence for severity triage based on conversations in an emergency department in Korea". In: *Scientific Reports* 15.1, p. 16870.

Shapira, N., Wendler, C., Yen, A., Sarti, G., Pal, K., Floody, O., Belfki, A., Loftus, A., Jannali, A. R., Prakash, N., et al. (2026). *Agents of Chaos*. arXiv: 2602.20021 [cs.AI].

Thim, T., Krarup, N. H. V., Grove, E. L., Rohde, C. V., and Løfgren, B. (2012). "Initial assessment and treatment with the Airway, Breathing, Circulation, Disability, Exposure (ABCDE) approach". In: *International Journal of General Medicine* 5.0, pp. 117–121.

Trautmann, R., Reuter-Oppermann, M., and Christiansen, J. (2022). *PSAP-G-ONE Eine explorativ-deskriptive Studie über Leitstellen der nichtpolizeilichen Gefahrenabwehr in der Bundesrepublik Deutschland*. Tech. rep. Aachen: Deutsche Gesellschaft für Rettungswissenschaft e. V.

Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., et al. (2024). "A survey on large language model based autonomous agents". In: *Frontiers of Computer Science* 18.6, p. 186345. eprint: 2308.11432.

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., and Zhou, D. (2022). "Self-Consistency Improves Chain of Thought Reasoning in Language Models". In: *arXiv*. eprint: 2203.11171.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., and Zhou, D. (2022). "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models". In: *arXiv*. eprint: 2201.11903.

Williams, C. Y. K., Zack, T., Miao, B. Y., Sushil, M., Wang, M., Kornblith, A. E., and Butte, A. J. (2024). "Use of a Large Language Model to Assess Clinical Acuity of Adults in the Emergency Department". In: *JAMA Network Open* 7.5, e248895.