Triage Performance Across Large Language Models, ChatGPT, and Untrained Doctors in Emergency Medicine: Comparative Study

- *Journal of Medical Internet Research (JMIR), 2024*

INTRODUCTION

- **AI in Medicine:**
- Generative AI and LLMs gaining attention, highlighted by ChatGPT's release in November 2022.
- **Capabilities of ChatGPT and Other LLMs:**
- Successful performance on US medical exams.
- Preference over doctor replies for certain questions.
- **Emergency Department (ED) Triage:**
- Triage: Prioritizing patients based on urgency using systems like the Manchester Triage System (MTS).
- Challenges: High-stress environment, variable quality, influenced by rater's experience and fatigue.

Study Rationale

- ❑ Assessing ChatGPT’s performance in ED triage compared to professional raters and untrained doctors.
- ❑ Evaluating ChatGPT and other LLMs (e.g., Gemini 1.5, Llama 3 70B, Mixtral 8x7b).
- ❑ Exploring ChatGPT’s potential as a second opinion for less experienced ED staff.

Method

- **Case Vignette creation :**

  - 124 anonymized emergency cases from a single day at University Hospital Düsseldorf.
  - Vignettes contained medically relevant information only, adjusted for age and clinical values.
  - Non-medical information excluded.

- **Vignette Review:**

  -Created by one doctor and reviewed by a second doctor to ensure anonymity and standardization.

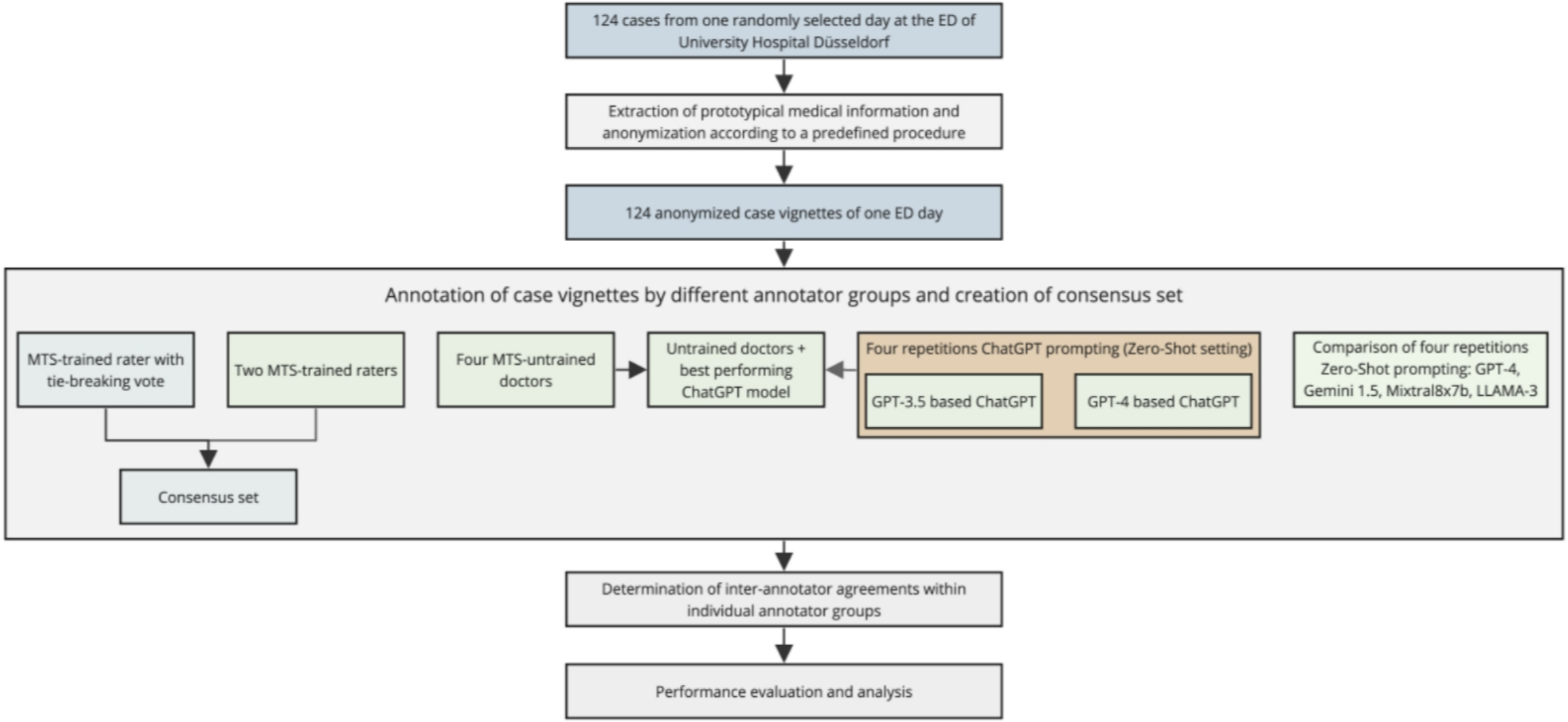
- **Triage Ratings:**

  - Independent assessment by 2 MTS-trained (Manchester Triage System) staff members.
  - Consensus reached with a third MTS-trained doctor for cases with differing priorities.

Method

1. **Untrained Resident Doctors**
    - 4 MTS-untrained resident doctors working regularly in the ED.
    - Residency Year: 2 doctors in their second year, 2 in their third year.
    - Untrained doctors reviewed ChatGPT's responses as a second opinion and reconsidered their initial triage decisions.

1. **Chat GPT**
    - GPT-4, Llama 3 70B, Gemini 1.5, Mixtral 8x7b
    - Zero-shot setting with optimized prompts, without additional training or access to MTS diagrams.
    - Each version queried 4 times with new chats to account for the probabilistic nature of LLMs.


124 cases from one randomly selected day at the ED of University Hospital Düsseldorf
Extraction of prototypical medical information and anonymization according to a predefined procedure
124 anonymized case vignettes of one ED day
Annotation of case vignettes by different annotator groups and creation of consensus set
MTS-trained rater with tie-breaking vote
Two MTS-trained raters
Consensus set
Four MTS-untrained doctors
Untrained doctors + best performing ChatGPT model
Four repetitions ChatGPT prompting (Zero-Shot setting)
GPT-3.5 based ChatGPT
GPT-4 based ChatGPT
Comparison of four repetitions Zero-Shot prompting: GPT-4, Gemini 1.5, Mixtral8x7b, LLAMA-3
Determination of inter-annotator agreements within individual annotator groups
Performance evaluation and analysis

Agreement measurement

quadratic-weighted Cohen Kappa.

Statistical analysis

one-way ANOVA with Bonferroni correction, Tukey honest significant difference test.

Results

1. Agreement Levels

- **Untrained doctors:** κ =0.68
- **GPT-4 :** κ = 0.67
- **GPT-3.5:** κ = 0.54
- **Gemini 1.5**: κ = 0.60
- **Llama 3 70B**: κ = 0.52
- **Mixtral 8x7b**: κ = 0.42

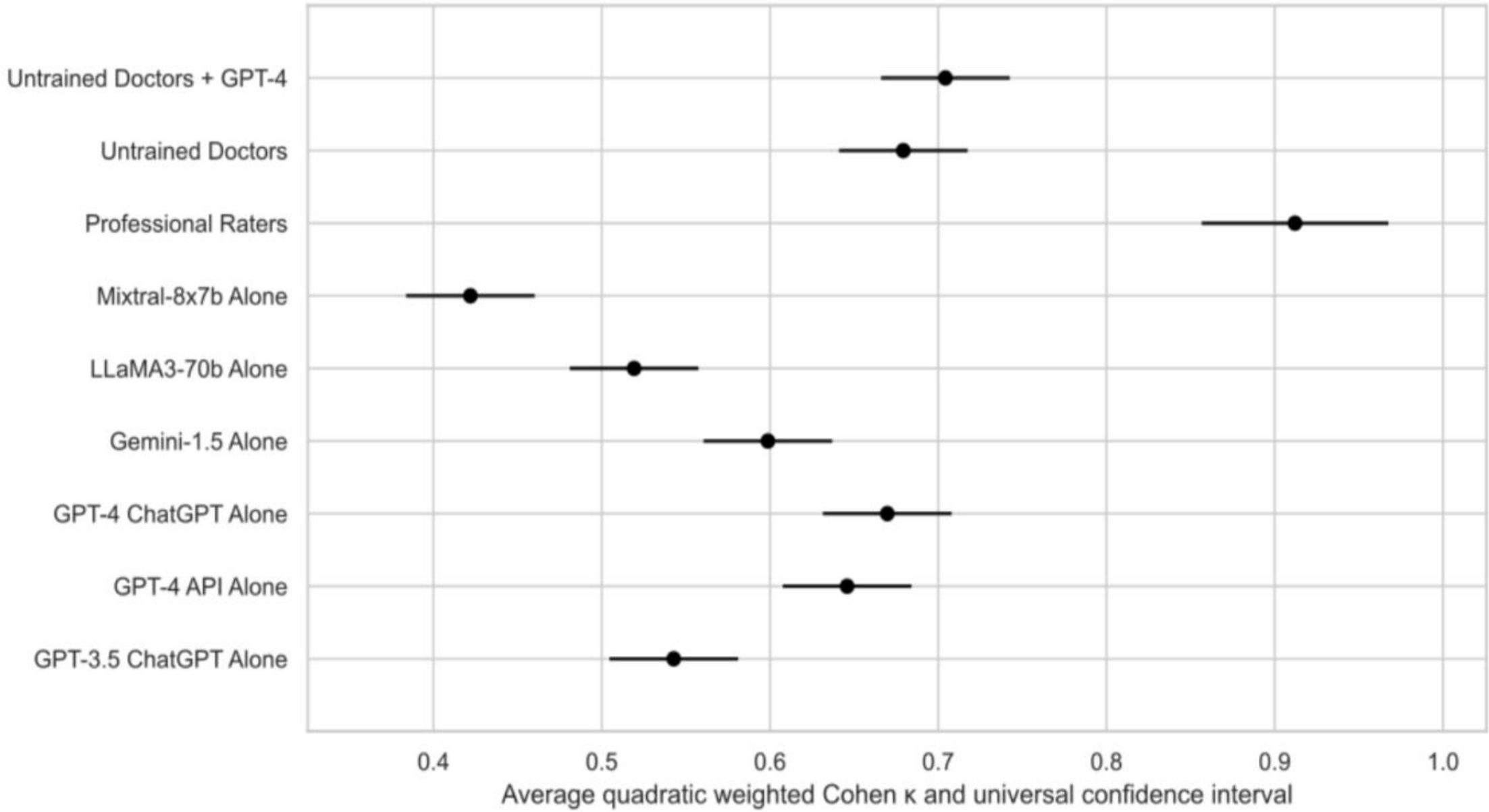
2. Statistical Significance

- GPT-3.5 vs GPT-4: $P < .001$
- GPT-4 vs Untrained Doctors: $P = 0.97$

3. Patterns Observed

•**GPT Models:** Tend toward over-triage.

•**Untrained Doctors:** Tend toward under-triage.


Untrained Doctors + GPT-4
Untrained Doctors
Professional Raters
Mixtral-8x7b Alone
LLaMA3-70b Alone
Gemini-1.5 Alone
GPT-4 ChatGPT Alone
GPT-4 API Alone
GPT-3.5 ChatGPT Alone
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Average quadratic weighted Cohen κ and universal confidence interval

Results

- GPT-4–based ChatGPT and untrained doctors showed substantial agreement with the consensus triage of professional raters.
- Other tested LLMs, including Gemini 1.5, Llama 3 70B, and Mixtral 8x7b, performed similarly to or worse than GPT-4–based ChatGPT.
- The LLMs and ChatGPT models tended to over-triage, while untrained doctors were more likely to under-triage.
- Despite the promising results, LLMs and ChatGPT do not yet match the performance of professionally trained raters and do not demonstrate gold-standard performance in emergency department triage.
- LLMs and the ChatGPT models failed to significantly improve the triage proficiency of untrained doctors when used as decision support.

Implications of the Study

- Need for LLMs Further Development
- Potential as Decision Support Tools
- Integration into Clinical Workflow
- Regulatory and Validation Considerations

THANK YOU

Shirin Dehghan

Dehghan.sh@gmail.com