ABSTRACT

Evaluation of a Large Language Model–Enabled Adverse Event Triage Tool in Telehealth Obesity Care

Caroline Altman, Sophie Black, Lauren Broffman, PhD, Dan Cadden, Adam Handler, Peter King, Quentin Lehn, Emily Strait, Will Walmsley, MASc

ABSTRACT

Background

Adverse events from GLP-1s are a factor in suboptimal adherence and treatment discontinuation. Timely management of adverse events is essential for supporting long-term therapeutic success. In traditional and telehealth care models, patients frequently report symptoms via portal messaging, where clinically important concerns may be buried among a high volume of routine or administrative communications. As a result, care teams may be delayed in triaging and responding to critical adverse event reports. To address this challenge, we developed a large language model (LLM)-based system to automatically detect and triage messages containing potential GLP-1-related adverse event concerns, enabling faster care team follow up and improved support for medication adherence.

Methods

A custom LLM–based annotation system was developed to detect patient messages referencing a current or recent adverse event during treatment with GLP-1 pharmacotherapy on a telehealth platform. In partnership with clinical pharmacists, structured adverse event severity data were incorporated into the model prompt to enhance annotation accuracy. At deployment, the LLM identified a qualifying message as adverse event-related, classified its severity, and then generated a high-priority alert for clinical review by a licensed care team member (the LLM did not make care decisions; all clinical determinations were made by providers). To evaluate model precision, human reviewers manually labeled a sample of annotated messages as adverse event–related or not. To assess the model's operational impact, we analyzed a random sample of 3,500 patient messages sent in the four days before and five days after implementation of the LLM. The LLM was active and classifying messages in real time prior to full deployment, but its outputs were not yet connected to routing logic or alert generation. This pre-launch, classification-only phase allowed for a comparison of care team response times by severity, before and after LLM-driven alerts were introduced. Message resolution times were compared pre- and post-implementation using Wilcoxon signed-rank tests, with p values corrected for multiplicity.

Results

The LLM achieved a 94% true-positive rate in detecting adverse event–related content. Following implementation, response times were significantly reduced across all adverse event severity levels (mild, moderate, severe, and urgent). Median time to first clinical response decreased from 115.1 minutes pre-LLM launch (n=1628; IQR 78.4–264.8) to 33.0 minutes post-LLM launch (n=1872; IQR 13.3–205.7), a median reduction of 82.1 minutes (P<.001). Reductions were observed consistently across severity subgroups, with the largest improvements in the mild (–80.8 minutes; P<.001) and moderate (–85.9 minutes; P<.001) categories. Mean response times also declined post-LLM launch, from 235.1 (SD=255.2) to 174.6 (SD=264.4) minutes overall, though variability remained high. Urgent adverse events post-LLM launch were addressed most quickly, with a median response time of 25.8 minutes (IQR 9.3–50.2).

Conclusions

Implementation of an LLM-enabled triage tool that accurately identified and classified patient-reported adverse events led to significantly faster response times, demonstrating its feasibility and potential for scalable use in obesity care. These findings suggest that LLM-based triage of adverse event messages may offer meaningful clinical benefits by improving care team responsiveness and streamlining care delivery.

Table 1. Response Times in Minutes to Patient Messages Before and After Implementation of LLM-Enabled AE Triage Tool (Nov 30 – Dec 9, 2024)

Time Frame	Severity Rating	n	Median Response Time (IQR)	Mean Response Time (SD)	Median Difference (Pre to Post LLM Launch)
Pre-LLM Launch	Mild	771	114.4 (76.5, 261.6)	236.9 (261.4)	NA
	Moderate	687	117.8 (81.4, 284.8)	237.4 (250.1)	NA
	Severe	140	110.8 (77.6, 203.6)	218.9 (243.7)	NA
	Urgent	30	89.4 (43.5, 157.4)	214.2 (268.4)	NA
	Any Severity	1628	115.1 (78.4, 264.8)	235.1 (255.2)	NA
Post-LLM Launch	Mild	785	33.5 (12.8, 233.4)	179.4 (262.3)	-80.8***
	Moderate	808	31.9 (13.3, 168.1)	168.9 (268.1)	-85.9***
	Severe	237	37.7 (17.1, 277.4)	192.0 (269.4)	-73.1***
	Urgent	42	25.8 (9.3, 50.2)	99.0 (183.9)	-63.5***
	Any Severity	1872	33.0 (13.3, 205.7)	174.6 (264.4)	-82.1***

*** P <0.001

POSTER

Evaluation of a Large Language Model–Enabled Adverse Event Triage Tool in Telehealth Obesity Care

Key Finding: Implementation of an LLM-enabled triage tool that accurately identified and classified patient-reported adverse events led to significantly faster response times.

- The LLM achieved a 94% true-positive rate in identifying adverse event–related content.
- Response times decreased significantly across all adverse event severity categories following LLM implementation.
- Post-LLM implementation, urgent adverse events had a median response time of 25.8 minutes, the fastest among all severity levels.
- Improvements were observed consistently across mild, moderate, severe, and urgent classifications.

Background

Adverse events from GLP-1s are a factor in suboptimal adherence and treatment discontinuation.[1] Timely management of adverse events is essential for supporting long-term therapeutic success. In traditional and telehealth care models, patients frequently report symptoms via portal messaging, where clinically important concerns may be buried among a high volume of routine or administrative communications. As a result, care teams may be delayed in triaging and responding to critical adverse event reports. To address this challenge, we developed a large language model (LLM)-based system to automatically detect and triage messages containing potential GLP-1-related adverse event concerns, enabling faster care team follow up and improved support for medication adherence.

Methods

A custom LLM–based annotation system was developed to detect patient messages referencing a current or recent adverse event during treatment with GLP-1 pharmacotherapy on a telehealth platform. In partnership with clinical pharmacists, structured adverse event severity data were incorporated into the model prompt to enhance annotation accuracy. At deployment, the LLM identified a qualifying message as adverse event-related, classified its severity, and then generated a high-priority alert for clinical review by a licensed care team member (the LLM did not make care decisions; all clinical determinations were made by providers).

To evaluate model precision, human reviewers manually labeled a sample of annotated messages as adverse event–related or not. To assess the model's operational impact, we analyzed a random sample of 3,500 patient messages sent in the four days before and five days after implementation of the LLM. The LLM was active and classifying messages in real time prior to full deployment, but its outputs were not yet connected to message routing logic or alert generation. This pre-launch, classification-only phase allowed for a comparison of care team response times by severity, before and after LLM-driven alerts were introduced. Message resolution times were compared pre- and post-implementation using Wilcoxon signed-rank tests, with p values corrected for multiplicity.

Authors

Caroline Altman, Sophie Black, Lauren Broffman, PhD, Dan Cadden, Adam Handler, Peter King, Quentin Lehn, Emily Strait, Will Walmsley, MASc

Affiliation

All authors are current employees of Roman Health Ventures, the telehealth platform that supplied data for this study.

Results

Following implementation, response times were significantly reduced across all adverse event severity levels (mild, moderate, severe, and urgent). Median time to first clinical response decreased from 115.1 minutes pre-LLM launch (n=1628; IQR 78.4–264.8) to 33.0 minutes post-LLM launch (n=1872; IQR 13.3–205.7), a median reduction of 82.1 minutes (P<.001) Figure 1).

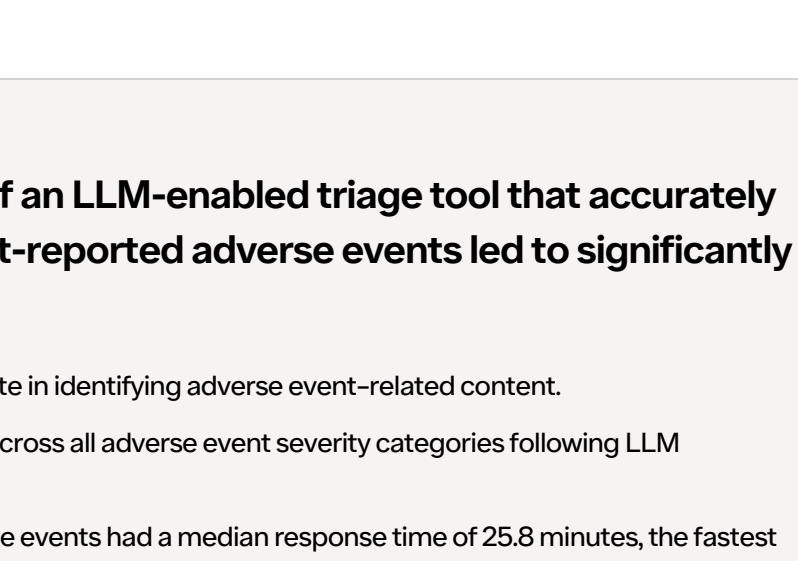
Reductions were observed consistently across severity subgroups; urgent adverse events post-LLM launch were addressed most quickly, with a median response time of 25.8 minutes (IQR 9.3–50.2).

Mean response times also declined post-LLM launch, from 235.1 (SD=255.2) to 174.6 (SD=264.4) minutes overall, though variability remained high.

Table 1: Mean Response Times in Minutes to Patient Messages Before and After Implementation of LLM-Enabled AE Triage Tool (Nov 30 – Dec 9, 2024)

Time Frame	Severity Rating	n	Mean Response Time (SD)
Pre-LLM Launch	Mild	771	236.9 (261.4)
	Moderate	687	237.4 (250.1)
	Severe	140	218.9 (243.7)
	Urgent	30	214.2 (268.4)
	Any Severity	1628	235.1 (255.2)
Post-LLM Launch	Mild	785	179.4 (262.3)
	Moderate	808	168.9 (268.1)
	Severe	237	192.0 (269.4)
	Urgent	42	99.0 (183.9)
	Any Severity	1872	174.6 (264.4)

Figure 1. Median Response Times in Minutes to Patient Messages Before and After Implementation of LLM-Enabled AE Triage Tool (Nov 30 – Dec 9, 2024)



References

1. Vidal, J., Flores, L., Jiménez, A., Pané, A. & Hollanda, A. de. What is the evidence regarding the safety of new obesity pharmacotherapies. Int. J. Obes. 49, 402–411 (2025).
2. Wharton, S. et al. Managing the gastrointestinal side effects of GLP-1 receptor agonists in obesity: recommendations for clinical practice. Postgrad. Med. 134, 14–19 (2022).