Supplementary

Table S1 Papers and exclusions

Article title	Reason for exclusion
1; A National Evaluation of the Effect of Trauma-Center Care on Mortality; E.J. MacKenzie and Others;	NOT CLINICAL
2; A Prospective Study of Venous Thromboembolism after Major Trauma; W.H. Geerts and Others;	INCLUDED
3; A Comparison of Low-Dose Heparin with Low-Molecular-Weight Heparin as Prophylaxis against; Venous Thromboembolism after Major Trauma; W.H. Geerts and Others	INCLUDED
4; Military–Civilian Collaboration in Trauma Care and the Senior Visiting Surgeon Program; E.E. Moore and Others;	NOT CLINICAL
5; Validity of a Set of Clinical Criteria to Rule Out Injury to the Cervical Spine in Patients with Blunt; Trauma; J.R. Hoffman and Others;	INCLUDED
6; The Canadian C-Spine Rule versus the NEXUS Low-Risk Criteria in Patients with Trauma; I.G. Stiell and Others;	PROMPT ERROR
7; Physicians' Warnings for Unfit Drivers and the Risk of Trauma from Road Crashes; D.A. Redelmeier and Others;	INCLUDED
8; Prehospital Plasma during Air Medical Transport in Trauma Patients at Risk for Hemorrhagic; Shock; J.L. Sperry and Others;	INCLUDED
9; Morphine Use after Combat Injury in Iraq and Post-Traumatic Stress Disorder; T.L. Holbrook and Others;	INCLUDED
10; Efficacy and Safety of Epoetin Alfa in Critically Ill Patients; H.L. Corwin and Others;	INCLUDED
11; Immediate versus Delayed Fluid Resuscitation for Hypotensive Patients with Penetrating Torso; Injuries; W.H. Bickell and Others;	INCLUDED
12; Transesophageal Echocardiography in the Diagnosis of Traumatic Rupture of the Aorta; M.D. Smith and Others;	INCLUDED
13; Necrotizing Cutaneous Mucormycosis after a Tornado in Joplin, Missouri, in 2011; R. Neblett Fanfair and Others;	INCLUDED
14; Brief Report: Role of COL4A1 in Small-Vessel Disease and Hemorrhagic Stroke; D.B. Gould and Others;	INCLUDED
15; A Multicenter, Randomized Trial of Treatment for Mild Gestational Diabetes; M.B. Landon and Others;	INCLUDED
16; Reversal of Catabolism by Beta-Blockade after Severe Burns; D.N. Herndon and Others;	INCLUDED
17; Contribution of Major Diseases to Disparities in Mortality; M.D. Wong and Others;	INCLUDED
18; Cannabis Legalization and Detection of Tetrahydrocannabinol in Injured Drivers; J.R. Brubacher and Others;	GPT 4.0 DENIED KNOWLEDGE
19; Indications for Computed Tomography in Patients with Minor Head Injury; M.J. Haydel and Others;	INCLUDED
20; Transplantation of Kidneys from Donors Whose Hearts Have Stopped Beating; Y.W. Cho and Others;	GPT 4.0 DENIED KNOWLEDGE
21; Elevation of Serum Creatine Kinase in Divers with Arterial Gas Embolization; R.M. Smith and T.S. Neuman;	INCLUDED
22; Recurrent Spontaneous Cervical-Artery Dissection; W.I. Schievink, B. Mokri, and W.M. O'Fallon;	GPT 4.0 DENIED KNOWLEDGE
23; Controlled Trial of Psychotherapy for Congolese Survivors of Sexual Violence; J.K. Bass and Others;	GPT 4.0 DENIED KNOWLEDGE
24; An Analysis of Outcomes of Reconstruction or Amputation after Leg-Threatening Injuries; M.J. Bosse and Others;	INCLUDED
25; Saline or Albumin for Fluid Resuscitation in Patients with Traumatic Brain Injury; The SAFE Study Investigators;	INCLUDED

This table describes the names of the papers and reasons for exclusions from the analysis for the 25 top papers in the *NEJM* under the search term "Trauma". Green represents that the article was included. Red represents that the article was excluded.

https://dx.doi.org/10.21037/jmai-24-48

Table S2 Reviewer 1 ChatGPT 3.5 *vs.* ChatGPT 4.0

Article title	GPT3_N_errors	GPT3_eTypes	GPT3_Why	GPT3_How	GPT3_Con	GPT3_CCC	GPT4_N_errors	GPT4_eTypes	GPT4_Why	GPT4_How	GPT4_Con	GPT4_CCC
2; A Prospective Study of Venous Thromboembolism after Major Trauma	3	Numeric inaccuracy	Yes	Yes	Yes	Yes	1	Misrepresents study conclusion	Yes	Yes	Yes	Yes
3; A Comparison of Low-Dose Heparin with Low-Molecular-Weight Heparin as Prophylaxis against;	3	Numeric inaccuracy, misrepresents study conclusion	Yes	Yes	No	No	1	Misrepresents study conclusion	Yes	Yes	Yes	Yes
5; Validity of a Set of Clinical Criteria to Rule Out Injury to the Cervical Spine in Patients with Blunt; Trauma	0	None	Yes	Yes	Yes	No	2	Numeric inaccuracy, misrepresents study design	Yes	Yes	Yes	Yes
7; Physicians' Warnings for Unfit Drivers and the Risk of Trauma from Road Crashes	2	Misrepresents study design, numeric inaccuracy	Yes	No	Yes	Yes	0	None	Yes	Yes	Yes	Yes
8; Prehospital Plasma during Air Medical Transport in Trauma Patients at Risk for Hemorrhagic	3	Numeric inaccuracy, misrepresents study design, misrepresents study conclusion	Yes	No	Yes	Yes	1	Misrepresents study conclusion	Yes	Yes	Yes	Yes
9; Morphine Use after Combat Injury in Iraq and Post-Traumatic Stress Disorder	2	Misrepresents study conclusion, misrepresents study design	Yes	Yes	No	No	0	None	Yes	Yes	Yes	Yes
10; Efficacy and Safety of Epoetin Alfa in Critically Ill Patients	1	Misrepresents study conclusion	Yes	Yes	Yes	Yes	1	Misrepresents study conclusion	Yes	Yes	No	No
11; Immediate versus Delayed Fluid Resuscitation for Hypotensive Patients with Penetrating Torso; Injuries	1	Misrepresents study conclusion	Yes	Yes	No	No	0	None	Yes	Yes	Yes	Yes
12; Transesophageal Echocardiography in the Diagnosis of Traumatic Rupture of the Aorta	1	Misrepresents study design	Yes	No	Yes	Yes	1	Misrepresents study design	Yes	No	Yes	Yes
13; Necrotizing Cutaneous Mucormycosis after a Tornado in Joplin, Missouri, in 2011	0	None	Yes	Yes	Yes	Yes	0	None	Yes	Yes	Yes	Yes
14; Brief Report: Role of COL4A1 in Small-Vessel Disease and Hemorrhagic Stroke	1	Misrepresents study design	Yes	No	Yes	Yes	0	None	Yes	Yes	Yes	Yes
15; A Multicenter, Randomized Trial of Treatment for Mild Gestational Diabetes	3	Numeric inaccuracy, misrepresents study design, misrepresents study conclusion	Yes	No	No	No	0	None	Yes	Yes	Yes	Yes
16; Reversal of Catabolism by Beta-Blockade after Severe Burns	1	Numeric inaccuracy	Yes	Yes	Yes	Yes	0	None	Yes	Yes	Yes	Yes
17; Contribution of Major Diseases to Disparities in Mortality	0	None	Yes	Yes	Yes	Yes	1	Misrepresents study conclusion	Yes	Yes	Yes	Yes
19; Indications for Computed Tomography in Patients with Minor Head Injury	1	Misrepresents study design	Yes	No	No	Yes	1	Misrepresents study design	Yes	Yes	Yes	Yes
21; Elevation of Serum Creatine Kinase in Divers with Arterial Gas Embolization	1	Numeric inaccuracy	Yes	Yes	Yes	Yes	1	Misrepresents study conclusion	Yes	Yes	No	Yes
24; An Analysis of Outcomes of Reconstruction or Amputation after Leg-Threatening Injuries	1	Misrepresents study design, misrepresents study conclusion	Yes	No	No	No	2	Misrepresents study conclusion	Yes	Yes	No	Yes
25; Saline or Albumin for Fluid Resuscitation in Patients with Traumatic Brain Injury	3	Numeric inaccuracy, misrepresents study conclusion, misrepresents study design	Yes	Yes	No	No	1	Numeric inaccuracy	Yes	Yes	Yes	Yes

This table displays the results of grading the ChatGPT 3.5 and ChatGPT 4.0 summaries after exclusion, as completed by reviewer 1.

https://dx.doi.org/10.21037/jmai-24-48

Table S3 Reviewer 2 ChatGPT 3.5 *vs.* ChatGPT 4.0

Article title	GPT3_N_errors	GPT3_eTypes	GPT3_Why	GPT3_How	GPT3_Con	GPT3_CCC	GPT4_N_errors	GPT4_eTypes	GPT4_Why	GPT4_How	GPT4_Con	GPT4_CCC
2; A Prospective Study of Venous Thromboembolism after Major Trauma	2	Numeric inaccuracy	Yes	Yes	Yes	No	0	None	Yes	Yes	Yes	Yes
3; A Comparison of Low-Dose Heparin with Low-Molecular-Weight Heparin as Prophylaxis against	1	Numeric inaccuracy	Yes	Yes	No	No	0	None	Yes	Yes	Yes	Yes
5; Validity of a Set of Clinical Criteria to Rule Out Injury to the Cervical Spine in Patients with Blunt; Trauma	3	Numeric inaccuracy	No	No	No	No	0	None	Yes	Yes	Yes	Yes
7; Physicians' Warnings for Unfit Drivers and the Risk of Trauma from Road Crashes	5	Numeric inaccuracy	Yes	No	No	No	0	None	Yes	Yes	Yes	Yes
8; Prehospital Plasma during Air Medical Transport in Trauma Patients at Risk for Hemorrhagic	2	Numeric inaccuracy	Yes	Yes	Yes	Yes	0	None	Yes	Yes	Yes	Yes
9; Morphine Use after Combat Injury in Iraq and Post-Traumatic Stress Disorder	0	None	Yes	Yes	Yes	Yes	0	None	Yes	Yes	Yes	Yes
10; Efficacy and Safety of Epoetin Alfa in Critically Ill Patients	0	None	No	Yes	Yes	Yes	0	None	Yes	Yes	No	No
11; Immediate versus Delayed Fluid Resuscitation for Hypotensive Patients with Penetrating Torso	2	Numeric inaccuracy	Yes	No	No	No	0	None	Yes	Yes	Yes	Yes
12; Transesophageal Echocardiography in the Diagnosis of Traumatic Rupture of the Aorta	0	None	No	No	No	Yes	0	None	Yes	No	Yes	Yes
13; Necrotizing Cutaneous Mucormycosis after a Tornado in Joplin, Missouri, in 2011	0	None	Yes	Yes	Yes	Yes	0	None	Yes	Yes	Yes	Yes
14; Brief Report: Role of COL4A1 in Small-Vessel Disease and Hemorrhagic Stroke	0	None	Yes	Yes	Yes	Yes	0	None	Yes	Yes	Yes	Yes
16; Reversal of Catabolism by Beta-Blockade after Severe Burns	2	Numeric inaccuracy	Yes	Yes	Yes	Yes	0	None	Yes	Yes	Yes	Yes
15; A Multicenter, Randomized Trial of Treatment for Mild Gestational Diabetes	2	Numeric inaccuracy	Yes	Yes	No	Yes	0	None	Yes	Yes	No	No
17; Contribution of Major Diseases to Disparities in Mortality	0	None	Yes	Yes	No	No	0	None	Yes	No	Yes	Yes
19; Indications for Computed Tomography in Patients with Minor Head Injury	0	None	No	No	No	No	0	None	Yes	No	Yes	Yes
21; Elevation of Serum Creatine Kinase in Divers with Arterial Gas Embolization	2	Numeric inaccuracy, misrepresents study conclusion	Yes	Yes	No	Yes	0	None	Yes	Yes	Yes	Yes
24; An Analysis of Outcomes of Reconstruction or Amputation after Leg-Threatening Injuries	0	None	Yes	Yes	No	No	0	None	Yes	Yes	Yes	Yes
25; Saline or Albumin for Fluid Resuscitation in Patients with Traumatic Brain Injury	1	Numeric inaccuracy, misrepresents study conclusion	No	No	No	No	2	Numeric inaccuracy	No	No	Yes	Yes

This table displays the results of grading the ChatGPT 3.5 and ChatGPT 4.0 summaries after exclusion, as completed by reviewer 2.

https://dx.doi.org/10.21037/jmai-24-48

Table S4 Number of errors by error type by reviewer 1

Error type	ChatGPT 3.5	ChatGPT 4.0	GPT3 Error Count	GPT4 Error Count
Numeric inaccuracy	44%	11%	8	2
Misrepresents study design	50%	17%	9	3
Misrepresents study conclusion	44%	39%	8	7
None	17%	39%	3	7

This table shows the number of errors by type of error for ChatGPT 3.5 and ChatGPT 4.0 and as percentages of the total of papers reviewed as assessed by reviewer 1. Categories of error included "numeric inaccuracy", "misrepresented study design", and "misrepresented study conclusion" or "none" where no error occurred.

Table S5 Number of errors by error type by reviewer 2

Error Type	ChatGPT 3.5	ChatGPT 4.0	GPT3 Error Count	GPT4 Error Count
Numeric inaccuracy	56%	6%	10	1
Misrepresents study design	0%	0%	0	0
Misrepresents study conclusion	11%	0%	2	0
None	44%	94%	8	17

This table shows the number of errors by type of error for ChatGPT 3.5 and ChatGPT 4.0 and as percentages of the total of papers reviewed as assessed by reviewer 1. Categories of error included "numeric inaccuracy", "misrepresented study design", and "misrepresented study conclusion" or "none" where no error occurred.

Table S6 Failed comprehension by type by reviewer 1

Failed Comprehension	ChatGPT 3.5	ChatGPT 4.0	ChatGPT 3.5	ChatGPT 4.0
Why	0%	0%	0	0
How	39%	6%	7	1
Conclusion	39%	17%	7	3
CCC	39%	6%	7	1
Perfect Comprehension	33%	78%	6	14

This table shows the number of times ChatGPT 3.5 and ChatGPT 4.0 failed to display comprehension of the relevant paper as assessed by reviewer 1. The categories of comprehension included "why" the paper was produced, "how" the study was designed, "conclusions" of the study and "CCC" which represents the "Correct Clinical Conclusion" or the primary conclusion of the study for which a failure to comprehend could lead to an incorrect clinical application of the study findings.

https://dx.doi.org/10.21037/jmai-24-48

Table S7 Failed comprehension by type by reviewer 2

Failed Comprehension	ChatGPT 3.5	ChatGPT 4.0	ChatGPT 3.5	ChatGPT 4.0
Why	28%	6%	5	1
How	33%	22%	6	4
Conclusion	61%	11%	11	2
CCC	50%	11%	9	2
Perfect Comprehension	28%	67%	5	12

This table shows the number of times ChatGPT 3.5 and ChatGPT 4.0 failed to display comprehension of the relevant paper as assessed by reviewer 2. The categories of comprehension included "why" the paper was produced, "how" the study was designed, "conclusions" of the study and "CCC" which represents the "Correct Clinical Conclusion" or the primary conclusion of the study for which a failure to comprehend could lead to an incorrect clinical application of the study findings.

Table S8 Failed comprehension by type by averaged across reviewers with Cohen kappa coefficients

Category	ChatGPT 3.5 Avg Failure	ChatGPT 4.0 Avg Failure	ChatGPT 3.5 CK Score	ChatGPT 4.0 CK Score
Why	14%	3%	0	0
How	36%	14%	0.16	0.34
Conclusions	50%	14%	0.15	0.31
CCC	44%	8%	0.33	0.64
Perfect Comprehension	31%	72%	-0.174	0.182

This table shows the number of times ChatGPT 3.5 and ChatGPT 4.0 failed to display comprehension of the relevant paper as assessed averaged across reviewers. The categories of comprehension included "why" the paper was produced, "how" the study was designed, "conclusions" of the study and "CCC" which represents the "Correct Clinical Conclusion" or the primary conclusion of the study for which a failure to comprehend could lead to an incorrect clinical application of the study findings. Cohen Kappa Coefficients display the interobserver agreement between reviewers for each measure.

https://dx.doi.org/10.21037/jmai-24-48