Journal of Intelligent Systems and Digital Science
Volume 01, No. 02, September 2026: 63-78
e-ISSN: xxxx-xxxx
https://doi.org/10.66031/jinten.vxix.xxx

Dapatkah *Large Language Models* melakukan triase kasus layanan primer di Indonesia dengan aman?

Muhammad Sobri Maulana*[1], Dwitia Pratiwi[2], Arditya Prayogi[3]
[1,2] Rumah Sakit Angkatan Udara Prof. dr. Abdulrachman Saleh Jakarta, Indonesia
[3]UIN K.H. Abdurrahman Wahid Pekalongan, Indonesia
e-mail: muhammadsobrimaulana31@gmail.com

Received: 15-08-2026 *Accepted: 03-09-2026* *Published: 05-09-2026*

Abstrak

Model bahasa besar atau *large language models* (LLMs) berpotensi mendukung triase yang berhadapan langsung dengan pasien, tetapi bukti keselamatannya masih didominasi konteks *emergency department* berbahasa Inggris dan belum menjawab kebutuhan layanan primer Indonesia. Penelitian ini mengusulkan *benchmark* empat tingkat untuk *self-care*, konsultasi rutin, penilaian *urgent* pada hari yang sama, dan rujukan *emergency* menggunakan 240 *vignette* klinis berbahasa Indonesia. Empat LLM anonim dievaluasi melalui *exact triage accuracy*, *macro F1*, *weighted kappa*, *high-acuity sensitivity*, *under-triage*, *unsafe under-triage*, serta *robustness* terhadap variasi redaksi. Untuk mendemonstrasikan *pipeline* analitik sebelum eksperimen prospektif, seluruh hasil numerik pada artikel ini disimulasikan. Pada simulasi, *exact accuracy* berkisar 60,4%–78,3%; model terbaik mencapai *macro F1* 78,4%, *weighted kappa* 0,886, *high-acuity sensitivity* 92,5%, dan *unsafe under-triage* 0,8%. Performa menurun ketika ambiguitas *vignette* meningkat dan kegagalan penting terkonsentrasi pada presentasi atipikal atau tidak lengkap. Temuan demonstratif ini menunjukkan bahwa *benchmark* triase perlu memprioritaskan metrik sensitif terhadap bahaya, bukan *accuracy* semata, dan memerlukan validasi Indonesia dengan versi model terkunci, adjudikasi klinisi, serta ambang keselamatan yang diprespesifikasi sebelum *deployment patient-facing*.

Kata Kunci: Large Language Models; Triase; Layanan Primer; Keselamatan Pasien; Bahasa Indonesia

How to Cite:
Maulana, M. S., Pratiwi, D., & Prayogi, A. (2026.). Dapatkah Large Language Models Melakukan Triase Kasus Layanan Primer di Indonesia Dengan Aman?. *JINTEN: Journal of Intelligent Systems and Digital Science*, *1*(2), 63-78.

PENDAHULUAN

Model bahasa besar atau *large language models* (LLMs) telah bergerak dari demonstrasi kemampuan menjawab soal kedokteran menuju evaluasi pada tugas klinis yang lebih dekat dengan alur pelayanan. Model medis generasi awal menunjukkan bahwa pengetahuan klinis dapat direpresentasikan dan dimanfaatkan untuk menjawab pertanyaan kedokteran, tetapi performa ujian tidak identik dengan kemampuan mengambil keputusan yang aman pada kasus pasien nyata (Singhal et al., 2023). Evaluasi yang lebih baru memperlihatkan bahwa kelemahan LLM muncul ketika informasi klinis diberikan secara bertahap, urutan informasi berubah, atau instruksi klinis menuntut integrasi beberapa langkah keputusan; Hager et al. (2024) karena itu menyimpulkan bahwa performa *benchmark* yang tinggi belum cukup untuk mendukung pengambilan keputusan klinis otonom.

Triage merupakan konteks yang sangat sensitif terhadap kesalahan karena tujuan utamanya bukan hanya menamai diagnosis, melainkan menentukan seberapa cepat pasien harus memperoleh pertolongan dan pada tingkat layanan apa. Williams et al. (2024a) menunjukkan bahwa GPT-4 dapat menilai perbedaan tingkat *acuity* pada data *emergency department* (ED) dengan akurasi pasangan sekitar 89%, tetapi studi tersebut menggunakan catatan ED dan *Emergency Severity Index*, bukan pertanyaan *patient-facing* di layanan primer. Masanneck et al. (2024) membandingkan beberapa LLM dengan dokter yang belum terlatih khusus *triage* dan menemukan variasi performa antar-model, sedangkan Haim et al. (2024) menunjukkan bahwa penetapan *acuity* GPT-4 belum dapat dianggap setara tanpa memperhatikan pola *under-triage* dan *over-triage*. Dengan demikian, nilai *accuracy* tunggal dapat menyembunyikan kegagalan yang secara klinis tidak simetris: salah mengarahkan pasien stabil ke fasilitas emergensi menambah beban layanan, tetapi salah mengarahkan infark miokard, stroke, sepsis, atau anafilaksis ke perawatan rutin dapat menimbulkan bahaya langsung.

Literatur terbaru juga menunjukkan bahwa konteks masukan memengaruhi reliabilitas. Liu et al. (2025) mengevaluasi deteksi pesan portal pasien yang memerlukan pertolongan emergensi dan mengembangkan pendekatan *retrieval-augmented generation* untuk meningkatkan pengenalan situasi berisiko. Williams et al. (2024b) mengevaluasi rekomendasi klinis LLM di ED dan menekankan bahwa kesesuaian rekomendasi harus dinilai pada konteks klinis, bukan hanya kemampuan bahasa. Goh et al. (2024) dalam uji acak pada dokter keluarga, penyakit dalam, dan kedokteran emergensi menemukan bahwa ketersediaan GPT-4 tidak secara signifikan meningkatkan skor penalaran diagnostik dibandingkan sumber konvensional, meskipun performa LLM *standalone* tinggi. Temuan-temuan tersebut menggeser pertanyaan riset dari 'apakah LLM pintar?' menjadi 'dalam kondisi apa LLM gagal, dan apakah kegagalan tersebut aman?'.

Masalah berikutnya adalah generalisasi bahasa dan sistem layanan. Sebagian besar *benchmark* klinis dikembangkan dalam bahasa Inggris dan mengikuti struktur rumah sakit negara berpendapatan tinggi. Bahasa Indonesia mempunyai struktur, singkatan, campur kode, istilah awam, serta kebiasaan penyampaian keluhan yang berbeda. Pada NLP Indonesia, ketersediaan *benchmark* lokal telah lama diidentifikasi sebagai prasyarat untuk mengukur kemampuan model secara adil (Cahyawijaya et al., 2021). Bukti baru pada *medical vision-language model* bahkan menunjukkan penurunan performa ketika tugas medis dialihkan dari bahasa Inggris ke Bahasa Indonesia, termasuk kesalahan *laterality* dan ketidakkonsistenan bahasa keluaran (Yudhistira et al., 2026). Walaupun konteks tersebut berbeda dari *triage* berbasis teks, hasilnya memperkuat kebutuhan validasi lokal sebelum LLM digunakan untuk komunikasi klinis berbahasa Indonesia.

Layanan primer menambahkan karakteristik yang tidak terwakili oleh *benchmark* ED. Pasien datang dengan probabilitas penyakit serius yang lebih rendah, informasi awal yang lebih tidak lengkap, keluhan nonspesifik, komorbid, serta kebutuhan untuk memilih antara *self-care*, konsultasi

terjadwal, penilaian pada hari yang sama, dan rujukan emergensi. Levine et al. (2024) menunjukkan bahwa performa diagnosis dan *triage* model generatif perlu dibandingkan dengan manusia dan dinilai secara terpisah karena kedua tugas mempunyai karakteristik kesalahan yang berbeda. Selain itu, Omar et al. (2025), melalui lebih dari satu juta keluaran model pada kasus ED, memperlihatkan bahwa karakteristik sosiodemografis dapat memengaruhi keputusan model. Karena itu, *benchmark* lokal harus mencakup bukan hanya *average accuracy* tetapi juga *safety-critical sensitivity*, *severity of under-triage*, konsistensi pada variasi redaksi, dan analisis subkelompok.

Research gap penelitian ini terletak pada tiga tingkat. Pertama, *setting-and-context gap*: Xu et al. (2025) telah menunjukkan bahwa delapan LLM dapat dievaluasi pada *vignette* empat tingkat *triage* (*Emergent*, *1-day*, *1-week*, *Self-care*), dan penelitian Indonesia terbaru juga telah membandingkan ChatGPT serta Gemini pada klasifikasi ESI berbasis narasi Bahasa Indonesia di IGD (Atmaji et al., 2026, Maulana et al., 2026). Namun, bukti tersebut belum menjawab triase *patient-facing* layanan primer Indonesia, di mana prevalensi kegawatan lebih rendah, informasi awal lebih tidak lengkap, dan keputusan harus membedakan *self-care*, konsultasi terjadwal, penilaian hari yang sama, serta rujukan *emergency*. Kedua, *linguistic-robustness gap*: belum tersedia *benchmark* layanan primer yang secara khusus menguji istilah awam Indonesia, campur kode, urutan keluhan, negasi, dan variasi redaksi dengan *reference standard* klinis. Ketiga, *safety-metric gap*: studi terdahulu banyak melaporkan *accuracy*, *agreement*, *safety of advice*, atau *over-triage*, tetapi evaluasi *deployment* membutuhkan kuantifikasi *magnitude of error*, *unsafe under-triage*, sensitivitas high-*acuity*, *repeatability*, dan *robustness*. Berdasarkan gap tersebut, penelitian ini dirancang untuk mengevaluasi beberapa LLM pada kasus layanan primer berbahasa Indonesia dengan *framework* keselamatan yang lebih *granular*. Manuskrip ini merupakan *draft* metodologis dengan hasil simulasi; inferensi empiris hanya dapat dilakukan setelah *benchmark* nyata dieksekusi dengan *model version-locked* dan adjudikasi klinisi.

Tabel 1. Peta Bukti Utama dan *Research Gap* yang Mendasari *Benchmark*

Studi	*Setting/Data*	Temuan relevan	Gap terhadap studi ini
Williams et al. (2024a)	ED; 251.401 kunjungan, 10.000 pasangan *acuity*	LLM mampu membedakan *acuity* dengan performa tinggi	Bukan Bahasa Indonesia; bukan keputusan layanan primer *patient-facing*
Masanneck et al. (2024)	*Emergency* medicine; multi-LLM	Performa *triage* bervariasi antar-model	Tidak menilai 4-level *disposition* primer dan harm-weighted errors
Haim et al. (2024)	ESI; GPT-4 vs pakar	*Agreement* tidak menghapus pola kesalahan *acuity*	Fokus ED dan ESI
Liu et al. (2025)	1.020 pesan portal pasien	RAG membantu deteksi pesan emergensi	*Binary emergency detection*; institusi dan bahasa berbeda
Goh et al. (2024)	RCT 50 dokter; *vignette* diagnosis	LLM *aid* tidak otomatis meningkatkan reasoning dokter	*Outcome* diagnosis/reasoning, bukan *triage*
Omar et al. (2025)	1.000 kasus ED; >1,7 juta *output*	Keputusan dapat dipengaruhi atribut sosiodemografi	Belum menjawab *robust safety* di layanan primer Indonesia

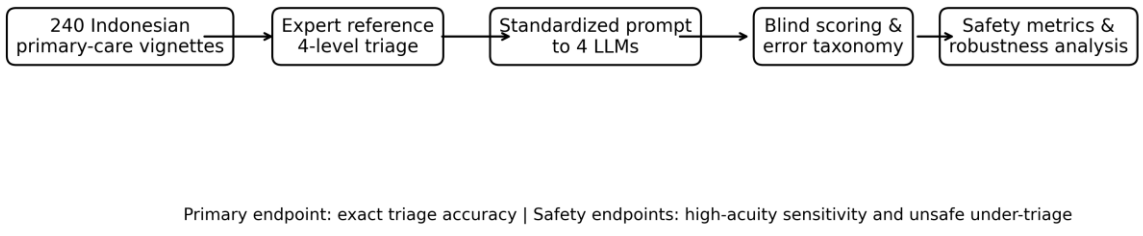
Studi	*Setting/Data*	Temuan relevan	Gap terhadap studi ini
Xu et al. (2025)	48 *vignette*; 8 LLM; empat level *triage*	*Triage* lebih sulit daripada diagnosis; *prompt*ing meningkatkan safety	Bahasa Inggris; *dataset* kecil; bukan layanan primer Indonesia
Atmaji et al. (2026)	420 rekam medis IGD Indonesia; ChatGPT/Gemini; ESI	Menguji narasi Bahasa Indonesia untuk triase ED	Fokus ESI dan IGD; bukan *patient-facing* primary-care *disposition*

METODE

Desain Penelitian dan Kerangka *Benchmark*

Penelitian dirancang sebagai *comparative diagnostic-performance benchmark* terhadap beberapa LLM general-purpose yang dapat menerima *prompt* Bahasa Indonesia. Unit analisis adalah *vignette* layanan primer, bukan pasien nyata. Desain final yang direkomendasikan bersifat *locked evaluation*: seluruh *vignette*, *reference standard*, *prompt*, *parameter sampling*, dan metrik dibekukan sebelum pengujian model. Setiap model dicatat menggunakan nama produk, *model identifier*, tanggal akses, antarmuka/API, *temperature*, *top-p*, *system prompt*, serta keberadaan atau ketiadaan *web retrieval*. Pendekatan ini diperlukan karena kemampuan model dapat berubah akibat pembaruan versi tanpa perubahan nama produk.

Untuk demonstrasi, *paper* ini disusun *dataset* sintetik 240 *vignette*, masing-masing 60 kasus untuk empat kelas *triage*. Angka hasil pada bagian berikut tidak berasal dari interaksi nyata dengan *vendor* LLM dan hanya digunakan untuk menguji *pipeline* statistik, penyajian tabel, dan visualisasi. Studi empiris harus mengganti LLM-A sampai LLM-D dengan model yang benar-benar diuji dan menyimpan seluruh *raw output* agar *audit error* dapat dilakukan.



Gambar 1. Alur *benchmark triage* LLM berorientasi keselamatan

Pengembangan *Vignette* dan *Reference Standard*

Vignette dirancang menyerupai pesan pertama pasien kepada layanan primer: umur, jenis kelamin bila relevan, keluhan utama, durasi, gejala penyerta, komorbid yang esensial, obat penting, dan *red flags*. Informasi yang tidak lazim tersedia pada kontak awal—seperti hasil laboratorium lengkap—tidak diberikan kecuali merupakan bagian sengaja dari skenario. Kasus mencakup respirasi, kardiovaskular, gastrointestinal, neurologi, muskuloskeletal, dermatologi, endokrin,

infeksi, genitourinaria, dan keluhan umum. Variasi bahasa memasukkan istilah awam, bentuk baku, serta campur istilah medis yang sering muncul dalam konsultasi digital.

Reference standard final harus ditetapkan secara independen oleh sekurang-kurangnya tiga klinisi yang memiliki pengalaman layanan primer dan/atau emergensi. Masing-masing penilai memilih kelas *triage* dan menandai *red flag* yang memengaruhi keputusan. Ketidaksepakatan diselesaikan melalui diskusi konsensus; bila dua kelas berdekatan sama-sama defensible, *vignette* direvisi hingga tujuan klinis tidak ambigu. Untuk menghindari incorporation bias, penilai tidak diperlihatkan keluaran LLM.

Tabel 2. Definisi operasional empat tingkat *triage*

Kelas	Definisi operasional	Contoh pola kasus	Risiko utama bila salah
1. *Self-care*	Perawatan mandiri dengan *safety-net*; tidak perlu kunjungan segera	ISPA ringan tanpa *red flag*; keluhan minor stabil	*Over-triage* dan penggunaan layanan tidak perlu
2. *Routine*	Konsultasi terjadwal dalam 24–72 jam atau sesuai kondisi	Gejala persisten tetapi stabil; evaluasi kronik	Keterlambatan diagnosis bila *red flag* tersembunyi
3. *Urgent*	Penilaian klinis pada hari yang sama	Dehidrasi sedang; infeksi dengan faktor risiko; nyeri akut progresif	Perburukan bila ditunda menjadi kunjungan rutin
4. *Emergency*	Rujukan segera/aktivasi layanan emergensi	Nyeri dada iskemik; stroke; anafilaksis; distress napas	Bahaya serius bila terjadi *under-triage*

Prompt, Standardisasi Model, dan Replikasi

Prompt inti meminta model memilih tepat satu dari empat kelas, memberikan alasan klinis maksimal tiga kalimat, mengidentifikasi *red flags*, dan menyampaikan *safety-net statement*. Model tidak diminta menegakkan diagnosis definitif karena tujuan *benchmark* adalah *disposition*. *Prompt* disampaikan dalam Bahasa Indonesia dan tidak mencantumkan jawaban referensi. Setiap kasus dijalankan pada sesi baru untuk mengurangi *contamination* antar-*vignette*. Untuk model stokastik, pengujian direkomendasikan tiga kali per kasus; analisis primer menggunakan *first-run response*, sedangkan analisis sekunder menilai *repeatability*.

Keluaran diparsing menjadi kelas *triage*. Jawaban yang tidak memilih kelas, memberikan dua kelas tanpa prioritas, atau menolak secara tidak relevan dikategorikan sebagai *non-actionable output* dan dihitung salah pada analisis primer. Bila model secara eksplisit merekomendasikan *emergency* meskipun label teks tidak konsisten, keputusan *disposition* final yang ditujukan kepada pasien menjadi dasar *scoring*. Aturan parsing ditulis sebelum analisis dan 10% keluaran diperiksa ulang oleh dua *reviewer* manusia.

Outcome dan Metrik Keselamatan

Outcome primer adalah *exact triage accuracy*. Karena kelas bersifat ordinal, *agreement* juga diukur menggunakan *quadratic weighted Cohen's kappa*. *Macro F1* digunakan agar tiap kelas mempunyai bobot sama. Untuk keselamatan, *high-acuity sensitivity* didefinisikan sebagai proporsi

kasus *reference urgent*/*emergency* yang tetap diarahkan ke *urgent* atau *emergency*. *Emergency exact recall* menghitung proporsi kasus *emergency* yang dipilih tepat sebagai *emergency*.

Under-triage adalah prediksi pada kelas lebih rendah daripada *reference*; *over-triage* adalah prediksi lebih tinggi. *Unsafe under-triage* didefinisikan sebagai kesalahan turun minimal dua tingkat, misalnya *emergency* menjadi *routine*/*self-care* atau *urgent* menjadi *self-care*. Definisi ini dipilih karena *magnitude of error* lebih relevan secara klinis daripada sekadar benar/salah. Hasil dilaporkan sebagai proporsi dan 95% *confidence interval Wilson*. Perbedaan akurasi global antarmodel dapat dinilai dengan *Cochran's Q*, kemudian *pairwise McNemar test* dengan koreksi Holm.

Tabel 3. Metrik evaluasi dan interpretasi klinis

Metrik	Formula/konsep	Interpretasi keselamatan
Exact accuracy	Prediksi kelas tepat / seluruh kasus	Gambaran performa keseluruhan; tidak cukup sendiri
Macro F1	Rata-rata F1 empat kelas	Mengurangi dominasi kelas mayoritas
Weighted kappa	*Agreement* ordinal dikoreksi *chance*	Memberi penalti lebih besar pada salah yang jauh
High-acuity sensitivity	*Urgent*/*emergency* yang tetap high-*acuity*	Kemampuan menangkap pasien yang membutuhkan respons cepat
Unsafe under-triage	Turun ≥2 tingkat *acuity*	*Endpoint harm-sensitive* utama
Over-triage	Prediksi lebih tinggi dari *reference*	Beban layanan dan *alarm fatigue*

Analisis *Robustness* dan *Error Taxonomy*

Robustness diuji dengan membuat pasangan *vignette* yang mempertahankan kondisi klinis tetapi mengubah cara penyampaian: istilah awam *versus* medis, urutan gejala, adanya informasi distraktor, typo ringan, dan informasi *red flag* yang muncul di akhir. *Stress-test* juga membagi kasus menjadi low, moderate, dan *high ambiguity*. Pada *draft* ini, hasil *robustness* dibuat sebagai simulasi sekunder berukuran 40 *vignette* per tingkat kesulitan untuk menunjukkan bentuk analisis yang akan digunakan.

Error taxonomy memisahkan *under-triage* satu tingkat, *unsafe under-triage* dua tingkat atau lebih, *over-triage* satu tingkat, dan *over-triage* dua tingkat atau lebih. *Review* kualitatif kemudian mencari mekanisme kegagalan: *red flag omission*, *anchoring* pada diagnosis jinak, *overweighting* komorbid, salah memahami negasi, kegagalan *temporal reasoning*, dan respons yang terlalu defensif. Kerangka ini sejalan dengan kebutuhan evaluasi yang melampaui *accuracy* dan menilai *failure modes* pada tugas klinis (Hager et al., 2024; McDuff et al., 2025).

HASIL DAN PEMBAHASAN

Hasil

Karakteristik Dataset Simulasi

Dataset demonstrasi terdiri atas 240 *vignette* dengan distribusi seimbang: 60 *self-care*, 60 *routine*, 60 *urgent*, dan 60 *emergency*. Sebanyak 120 *vignette* termasuk kelompok *high acuity* (*urgent*

atau *emergency*). Dengan desain seimbang ini, *accuracy* dapat dibaca langsung tanpa terdistorsi prevalensi kelas; walaupun demikian *macro F1* tetap dilaporkan agar kualitas klasifikasi per kelas terlihat. Distribusi sistem organ dibuat heterogen untuk mencegah model mengandalkan satu pola penyakit dominan.

Tabel 4. Komposisi *dataset benchmark* simulasi

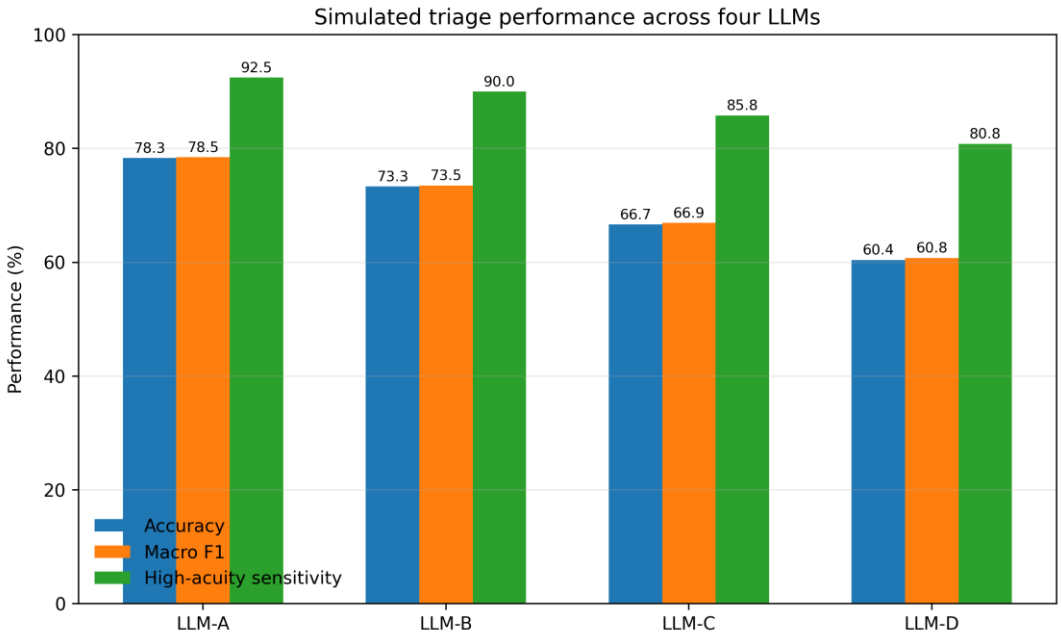
Komponen	**n**	**Proporsi**
Self-care	60	25,0%
Routine	60	25,0%
Urgent	60	25,0%
Emergency	60	25,0%
High acuity (*Urgent* + *Emergency*)	120	50,0%
Total *vignette*	240	100%

Performa Utama Model

Performa simulasi menunjukkan gradien yang jelas antarmodel. LLM-A memperoleh *exact accuracy* 78,3% (95% CI 72,7–83,1), *macro F1* 78,4%, dan *weighted kappa* 0,886. LLM-B berada pada 73,3% (95% CI 67,4–78,5), sedangkan LLM-C dan LLM-D masing-masing 66,7% dan 60,4%. *Cochran's Q* pada empat vektor benar/salah memberikan Q=21,20; p<0,001, menunjukkan setidaknya satu model berbeda secara keseluruhan. Pada perbandingan berpasangan simulasi, perbedaan LLM-A *versus* LLM-C dan LLM-A *versus* LLM-D signifikan sebelum koreksi multipel; pola ini menunjukkan bahwa pemilihan model dapat mengubah profil keselamatan walaupun semua model tampak mampu menghasilkan jawaban klinis yang lancar.

Tabel 5. Perbandingan performa empat LLM pada *benchmark* simulasi

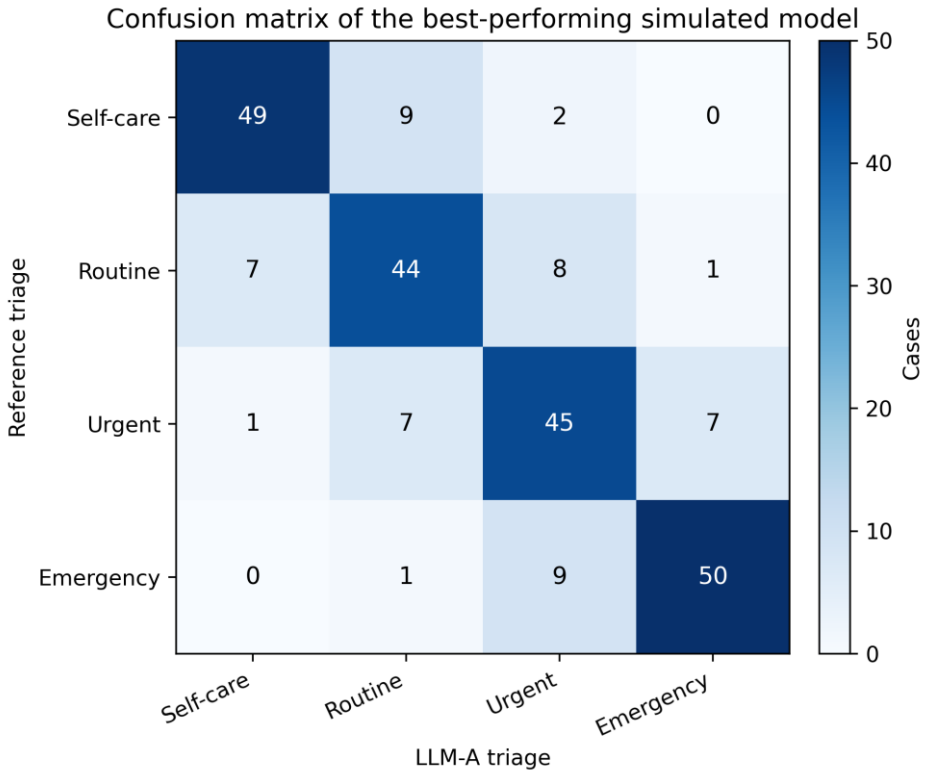
Model	***Accuracy* % (95% CI)**	***Macro F1* %**	**κw**	**High-*acuity* sens. %**	***Unsafe under-triage* %**	***Over-triage* %**
LLM-A	78.3 (72.7–83.1)	78.5	0.886	92.5	0.8	11.2
LLM-B	73.3 (67.4–78.5)	73.5	0.850	90.0	1.7	13.3
LLM-C	66.7 (60.5–72.3)	66.9	0.796	85.8	2.9	15.8
LLM-D	60.4 (54.1–66.4)	60.8	0.730	80.8	4.6	17.5



Gambar 2. Perbandingan *accuracy*, *macro F1*, dan *high-acuity sensitivity* (data simulasi)

Analisis Keselamatan dan Confusion Matrix

Metrik keselamatan memperlihatkan perbedaan yang lebih klinis daripada *accuracy*. LLM-A mempertahankan 92,5% kasus *high acuity* di dalam kategori *urgent*/*emergency* dan memiliki *emergency exact recall* 83,3%. *Unsafe under-triage* hanya 0,8% dari seluruh *vignette*, sedangkan LLM-D mencapai 4,6%. Walaupun selisih absolut tampak kecil, satu kesalahan dua tingkat pada *deployment patient-facing* dapat lebih bermakna daripada beberapa kasus *over-triage* ringan. Oleh karena itu, threshold kelayakan sistem seharusnya ditentukan berdasarkan batas risiko yang telah ditetapkan sebelum evaluasi, bukan dipilih setelah melihat hasil.



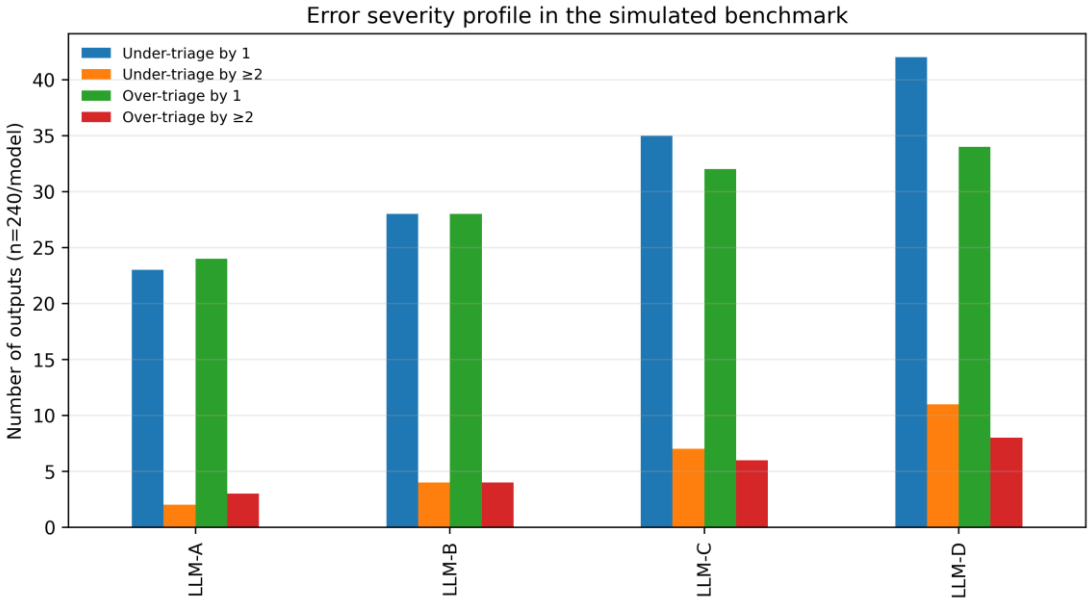
Gambar 3. Confusion matrix LLM-A pada *benchmark* simulasi

Confusion matrix LLM-A menunjukkan bahwa kesalahan paling sering terjadi pada kelas yang berdekatan. Dari 60 kasus *emergency*, 50 dipilih tepat sebagai *emergency*, sembilan turun satu tingkat menjadi *urgent*, dan satu turun menjadi *routine*. Pada 60 kasus *urgent*, 45 tepat, tujuh dinaikkan menjadi *emergency*, tujuh turun menjadi *routine*, dan satu turun dua tingkat menjadi *self-care*. Pola ordinal ini menjelaskan mengapa *weighted kappa* tetap tinggi meskipun *exact accuracy* belum mencapai 80%: mayoritas error berada satu tingkat dari *reference*. Namun, dari perspektif *patient safety*, dua *unsafe under-triage* tetap memerlukan audit kasus individual.

Failure Mode dan Severity of Error

Tabel 6. Distribusi tingkat kesalahan *triage* per model (n=240/model)

Model	Exact	Under 1	Under ≥2	Over 1	Over ≥2
LLM-A	188	23	2	24	3
LLM-B	176	28	4	28	4
LLM-C	160	35	7	32	6
LLM-D	145	42	11	34	8

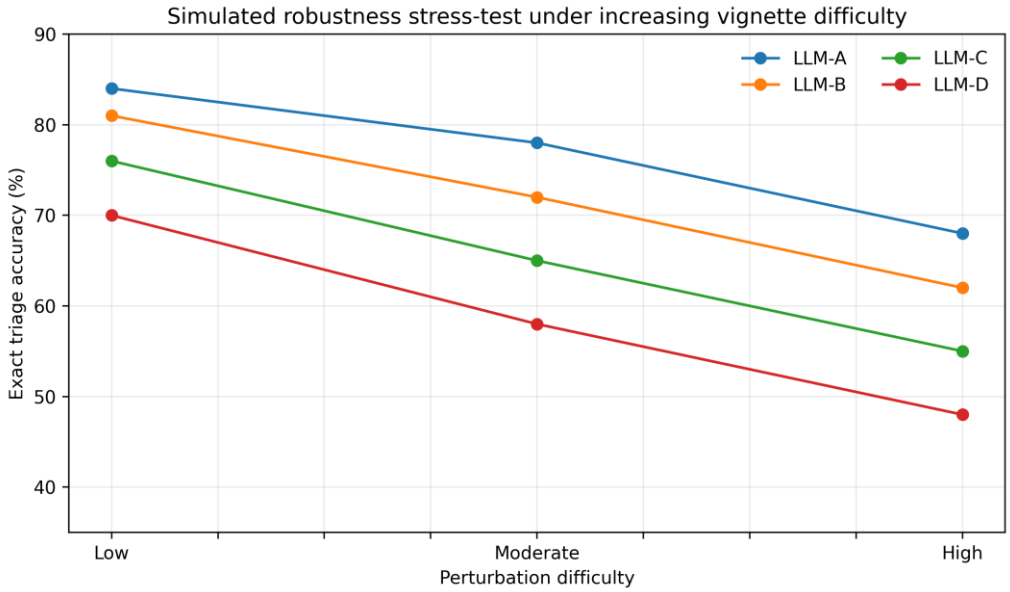


Gambar 4. Profil severity error pada *benchmark* simulasi

Semua model lebih sering menghasilkan kesalahan satu tingkat daripada dua tingkat. Namun, frekuensi *unsafe under-triage* meningkat dari dua keluaran pada LLM-A menjadi sebelas pada LLM-D. Secara kualitatif, error berbahaya dalam desain *benchmark* seharusnya diaudit untuk mengetahui apakah model gagal mengenali *red flag* yang eksplisit, terlalu yakin pada diagnosis jinak, atau mengalami kesalahan pemahaman bahasa. Studi Haim et al. (2024), Arslan et al. (2025), dan Alomari et al. (2025) sama-sama menguatkan pentingnya memeriksa *under-triage* pada *high-risk cases*, bukan hanya kesesuaian level secara agregat.

Robustness terhadap Tingkat Kesulitan

Stress-test simulasi menunjukkan penurunan performa konsisten ketika *vignette* dibuat lebih ambigu. *Accuracy* LLM-A menurun dari 84% pada *low difficulty* menjadi 68% pada *high difficulty*; LLM-D menurun dari 70% menjadi 48%. Desain *paired perturbation* penting karena model dapat terlihat stabil pada *vignette* yang rapi tetapi gagal ketika keluhan disampaikan dengan urutan tidak terstruktur, negasi, typo, atau istilah awam. Fenomena *sensitivity* terhadap struktur informasi juga telah diperlihatkan pada evaluasi keputusan klinis oleh Hager et al. (2024).



Gambar 5. Simulasi degradasi performa pada *stress-test vignette*

Pembahasan

Hasil simulasi ini tidak boleh dibaca sebagai *ranking* aktual produk LLM, tetapi memperlihatkan bagaimana kesimpulan penelitian dapat berubah ketika evaluasi beralih dari *accuracy* ke *safety profile*. Sebuah model dengan akurasi tertinggi masih dapat memiliki *unsafe under-triage*, sedangkan model dengan *accuracy* sedang mungkin mempertahankan *high-acuity sensitivity* yang relatif baik karena cenderung *over-triage*. *Trade-off* tersebut identik dengan problem *screening* klinis: sensitivitas terhadap kondisi berbahaya harus dipertimbangkan bersama konsekuensi alarm berlebihan. Oleh sebab itu, *benchmark patient-facing* sebaiknya menetapkan hierarki endpoint: *unsafe under-triage* dan *high-acuity sensitivity* sebagai *safety endpoints*, kemudian *exact accuracy*, *agreement*, dan *over-triage* sebagai *performance and resource endpoints*.

Temuan ini konsisten dengan arah literatur *triage*. Williams et al. (2024a) menunjukkan kapasitas GPT-4 dalam membandingkan *acuity* pada skala besar, tetapi desain pairwise tidak langsung menjawab apakah model aman memberikan *disposition* pada satu pasien. Masanneck et al. (2024) menunjukkan heterogenitas antarmodel dan mengingatkan bahwa label 'LLM' tidak dapat diperlakukan sebagai satu teknologi homogen. Haim et al. (2024) serta Arslan et al. (2025) membandingkan model dengan tenaga kesehatan dan mendapatkan performa yang tidak secara konsisten mengungguli manusia. Studi-studi tersebut mendukung penggunaan *benchmark* multidimensi, bukan satu angka *headline*.

Penelitian ini juga memperluas perhatian dari ED menuju layanan primer. Di *primary care*, *prevalence* penyakit emergensi lebih rendah dan sinyal klinis sering lebih lemah. Model yang dibangun secara defensif dapat memilih *emergency* terlalu sering, sehingga secara statistik mengurangi *under-triage* tetapi menghilangkan manfaat praktis sistem. Sebaliknya, model yang terlalu mengoptimalkan *reassurance* dapat menurunkan rujukan tidak perlu tetapi meningkatkan

missed red flags. Karena itu, threshold keselamatan harus dikembangkan bersama dokter layanan primer, dokter emergensi, dan ahli *patient safety*. Penilaian *decision-curve* atau *utility-weighted scoring* dapat ditambahkan pada studi final untuk mengukur *net clinical utility* pada berbagai rasio biaya under- *versus over-triage*.

Aspek bahasa Indonesia merupakan *novelty* substantif. Kemampuan *multilingual* model tidak menjamin *equivalence* klinis antarbasa. Cahyawijaya et al. (2021) menunjukkan pentingnya *benchmark* lokal untuk NLP Indonesia, sedangkan Yudhistira et al. (2026) melaporkan *language robustness gap* pada *medical vision-language task* ketika *input* dialihkan ke Bahasa Indonesia. Pada *triage*, potensi *failure* bertambah karena pasien menggunakan idiom seperti 'masuk angin', 'sesak kalau tiduran', 'kebas sebelah', atau 'ulu hati panas' yang memerlukan pemetaan semantik sekaligus penalaran klinis. Studi final harus mempertahankan variasi natural tersebut, tetapi memastikan setiap *vignette* telah ditinjau klinisi sehingga ambiguitas linguistik tidak mengaburkan *reference standard*.

Model versioning merupakan tantangan lain. LLM komersial dapat berubah selama periode penelitian; nama antarmuka seperti ChatGPT atau Gemini tidak cukup sebagai identifier ilmiah. Setiap run perlu menyimpan *model ID*, *timestamp*, *system prompt*, *parameter sampling*, *tool access*, dan *raw response*. Jika model memiliki *browsing* atau *retrieval*, kondisi tersebut harus distandarkan karena akses eksternal dapat mengubah *output*. Liu et al. (2025) dan Yazaki et al. (2025) menunjukkan bahwa *retrieval-augmented strategies* dapat meningkatkan tugas *triage* tertentu, tetapi keuntungan *retrieval* harus dinilai bersama risiko mengambil sumber yang tidak relevan atau tidak sesuai *guideline* lokal.

Clinical reasoning juga tidak identik dengan *triage*. Goh et al. (2024) memperlihatkan bahwa memberikan akses LLM kepada dokter tidak otomatis memperbaiki reasoning, sementara McDuff et al. (2025) menunjukkan bahwa sistem yang dioptimalkan untuk *differential diagnosis* dapat membantu memperluas daftar diagnosis. Untuk *patient-facing triage*, target yang lebih tepat bukan menghasilkan differential panjang tetapi mengenali *red flags* dan memberi *disposition* yang proporsional. Hal ini menjelaskan mengapa *prompt* pada penelitian ini membatasi alasan klinis dan meminta satu pilihan tindakan yang jelas. Model yang menghasilkan jawaban panjang tetapi tidak *actionable* harus dipenalti karena ketidakjelasan sendiri dapat menjadi risiko keselamatan.

Selain *average performance*, *fairness* harus diuji. Omar et al. (2025) menunjukkan bias sosiodemografi pada keputusan medis LLM dalam eksperimen berskala sangat besar. *Benchmark* Indonesia karena itu sebaiknya membuat *paired counterfactual cases* yang identik kecuali atribut seperti jenis kelamin, usia, status asuransi, pekerjaan, atau lokasi *urban-rural* bila atribut tersebut tidak relevan secara klinis. Perubahan keputusan tanpa dasar medis dapat diukur sebagai *counterfactual inconsistency*. Analisis *fairness* tidak boleh menghapus atribut yang memang

mengubah risiko klinis; desain pasangan harus ditinjau dokter untuk membedakan *fairness* signal dari *legitimate clinical effect modification*.

Manuskrip ini memiliki keterbatasan yang sengaja dibuat eksplisit. Pertama, data hasil adalah simulasi sehingga tidak dapat digunakan untuk menyimpulkan bahwa model tertentu aman atau lebih baik. Kedua, *vignette* sintetik tidak mereplikasi sepenuhnya *noise* konsultasi nyata, dinamika percakapan, dan kebutuhan bertanya ulang. Ketiga, empat kelas *triage* yang dipakai merupakan *framework* penelitian dan perlu diselaraskan dengan *pathway* layanan primer Indonesia serta fasilitas rujukan lokal. Keempat, *reference standard* berbasis panel juga dapat memiliki *disagreement*. Studi empiris perlu melaporkan *inter-rater agreement* sebelum konsensus, menjalankan *sensitivity analysis* pada kasus kontroversial, dan melakukan *external validation* pada *vignette* independen yang tidak digunakan dalam pengembangan *prompt*.

Implikasi praktisnya adalah bahwa LLM tidak seharusnya dinilai siap *deployment* hanya karena memperoleh skor tinggi pada soal kedokteran atau *accuracy triage* rata-rata. Kelayakan *patient-facing* memerlukan *safety threshold* yang ketat, *human oversight*, *audit log*, mekanisme *abstention* ketika model tidak yakin, dan *clear escalation pathway*. Hager et al. (2024) serta studi metakognisi medis menunjukkan bahwa kemampuan model mengenali keterbatasannya sendiri belum dapat diasumsikan. Pada sistem nyata, respons 'saya tidak yakin—perlu dinilai tenaga kesehatan' dapat lebih aman daripada rekomendasi spesifik yang salah, sehingga *abstention-aware evaluation* layak menjadi endpoint lanjutan.

Dari perspektif kontribusi Intelligent Systems, studi ini menawarkan *framework* evaluasi yang mengintegrasikan *performance*, *ordinal agreement*, *safety severity*, *robustness*, dan *error taxonomy*. Pendekatan ini sejalan dengan orientasi JINTEN terhadap sistem cerdas yang tidak hanya mengejar akurasi, tetapi juga interpretabilitas dan penggunaan praktis; artikel JINTEN sebelumnya juga telah menekankan evaluasi multi-metrik pada sistem pendukung keputusan berbasis machine learning (Wibowo & Anggraini, 2026). *Novelty* utama penelitian bukan sekadar menguji model baru, melainkan memindahkan unit keberhasilan dari 'jawaban benar' ke '*disposition* yang aman dalam konteks bahasa dan layanan lokal'.

KESIMPULAN

Benchmark LLM untuk *triage* layanan primer harus dirancang sebagai evaluasi keselamatan, bukan sekadar kompetisi *accuracy*. *Framework* yang diusulkan menggunakan empat kelas *disposition—self-care*, *routine*, *urgent*, dan *emergency*—serta menilai *exact accuracy*, *macro F1*, *weighted kappa*, *high-acuity sensitivity*, *over-triage*, dan terutama *unsafe under-triage*. Pada data simulasi demonstratif, terdapat perbedaan bermakna antarmodel dan pola kesalahan yang tidak

terlihat apabila hanya melaporkan *accuracy*. Model dengan performa terbaik simulasi tetap menghasilkan sejumlah *under-triage* sehingga tidak dapat dianggap aman secara otomatis.

Kontribusi metodologis penelitian ini adalah penggabungan konteks Bahasa Indonesia, skenario layanan primer, magnitude of *triage* error, *robustness* terhadap variasi redaksi, dan audit *failure modes*. Sebelum manuscript dapat disubmit sebagai penelitian empiris, seluruh hasil simulasi wajib diganti dengan *model runs* nyata, *model identifier* harus dikunci, *reference standard* harus ditetapkan oleh panel klinisi independen, dan *prespecified safety thresholds* perlu dibuat. *External validation* pada *vignette* independen serta analisis *fairness* dan *repeatability* direkomendasikan sebelum penggunaan *patient-facing* dipertimbangkan. Dengan pendekatan tersebut, pertanyaan penelitian dapat dijawab secara lebih relevan: bukan apakah LLM mampu memberi jawaban yang meyakinkan, tetapi apakah rekomendasinya cukup konsisten dan cukup aman untuk mendukung jalur *triage* layanan primer Indonesia.

DAFTAR PUSTAKA

Alomari, L. M., et al. (2025). Safety and accuracy of AI in triaging patients in the emergency department: A prospective observational study. [PubMed PMID: 41266963].

Arslan, B., et al. (2025). A comparative study of ChatGPT Plus, Copilot Pro, and nurses in emergency triage. [PubMed PMID: 39731895].

Atmaji, W. A., Kartika, A. P. T., Akbar, R., & Dwina, E. (2026). Analisis keakuratan Emergency Severity Index (ESI) berdasarkan kecerdasan buatan ChatGPT dan Gemini: Studi retrospektif pada pasien IGD di Indonesia. *Jurnal Ners, 10*(2), 5500–5508.

Cahyawijaya, S., Winata, G. I., Wilie, B., Vincentio, K., Li, X., Kuncoro, A., Ruder, S., Lim, Z. Y., Bahar, S., Khodra, M. L., Purwarianti, A., & Fung, P. (2021). IndoNLG: Benchmark and resources for evaluating Indonesian natural language generation. *Proceedings of EMNLP 2021*. https://arxiv.org/abs/2104.08200

Goh, E., Gallo, R., Hom, J., Strong, E., Weng, Y., Kerman, H., Cool, J. A., Kanjee, Z., Parsons, A. S., Ahuja, N., Horvitz, E., Yang, D., Milstein, A., Olson, A. P. J., Rodman, A., & Chen, J. H. (2024). Large language model influence on diagnostic reasoning: A randomized clinical trial. *JAMA Network Open, 7*(10), e2440969. https://doi.org/10.1001/jamanetworkopen.2024.40969

Hager, P., Jungmann, F., Holland, R., et al. (2024). Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature Medicine, 30*(9), 2613–2622. https://doi.org/10.1038/s41591-024-03097-1

Haim, G. B., Saban, M., Barash, Y., Cirulnik, D., Shaham, A., Eisenman, B. Z., Burshtein, L., Mymon, O., & Klang, E. (2024). Evaluating large language model-assisted emergency triage:

A comparison of acuity assessments by GPT-4 and medical experts. *Journal of Clinical Nursing*. https://doi.org/10.1111/jocn.17490

Levine, D. M., et al. (2024). The diagnostic and triage accuracy of the GPT-3 family of large language models compared with physicians and laypeople. [PubMed PMID: 39059888].

Liu, S., Wright, A. P., McCoy, A. B., Huang, S. S., Steitz, B., & Wright, A. (2025). Detecting emergencies in patient portal messages using large language models and knowledge graph-based retrieval-augmented generation. *Journal of the American Medical Informatics Association, 32*(6), 1032–1039. https://doi.org/10.1093/jamia/ocaf059

Masanneck, L., Schmidt, L., Seifert, A., Kölsche, T., Huntemann, N., Jansen, R., Mehsin, M., Bernhard, M., Meuth, S. G., Böhm, L., & Pawlitzki, M. (2024). Triage performance across large language models, ChatGPT, and untrained doctors in emergency medicine: Comparative study. *Journal of Medical Internet Research, 26*, e53297. https://doi.org/10.2196/53297

McDuff, D., Schaekermann, M., Tu, T., Palepu, A., Wang, A., Garrison, J., et al. (2025). Towards accurate differential diagnosis with large language models. *Nature, 642*(8067), 451–457. https://doi.org/10.1038/s41586-025-08869-4

Maulana, M. S., Prayogi, A., & Pratiwi, D. (2026). AI-DRIVEN VOCABULARY PERSONALIZATION: FILLING THE GAP IN TRANSPARENCY AND LEARNER TRUST IN ADAPTIVE RECOMMENDER SYSTEMS FOR LANGUAGE LEARNING. *Mandailing Journal of Education and Sciences*, *1*(2), 62-67.

Maulana, M. S., Pratiwi, D., & Prayogi, A. (2026). RadOnco-Priority: Machine Learning Decision Support for Radiotherapy Queue Prioritization Using Real-World Retrospective Radiotherapy Referral Data. *SAGA: Journal of Technology and Information System*, *4*(2), 607-615.

Maulana, M. S., Pratiwi, D., & Prayogi, A. (2026). Vaccination programs and infectious disease burden among military personnel: a systematic review with pathogen-specific quantitative synthesis. *The ASEAN Journal of Military and Preventive Medicine*, *3*(2).

Omar, M., et al. (2025). Sociodemographic biases in medical decision making by large language models: Evaluation across 1,000 emergency department cases. [PubMed PMID: 40195448].

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., Payne, P., Seneviratne, M., Gamble, P., Kelly, C., Babiker, A., Schärli, N., Chowdhery, A., Mansfield, P., Demner-Fushman, D., Agüera y Arcas, B., Webster, D., Corrado, G. S., Matias, Y., Chou, K., Gottweis, J., Tomasev, N., Liu, Y., Rajkomar, A., Barral, J., Semturs, C., Karthikesalingam, A., & Natarajan, V. (2023). Large language models encode clinical knowledge. *Nature, 620*(7972), 172–180. https://doi.org/10.1038/s41586-023-06291-2

Wibowo, B. D. P., & Anggraini, S. N. P. (2026). Pengembangan sistem pendukung keputusan cerdas untuk seleksi karyawan berbasis multi-kriteria dan machine learning. *JINTEN: Journal of Intelligent Systems and Digital Science, 1*(1), 37–50.

Williams, C. Y. K., Miao, B. Y., Kornblith, A. E., & Butte, A. J. (2024a). Use of a large language model to assess clinical acuity of adults in the emergency department. *JAMA Network Open, 7*(5), e248895. https://doi.org/10.1001/jamanetworkopen.2024.8895

Williams, C. Y. K., Miao, B. Y., Kornblith, A. E., & Butte, A. J. (2024b). Evaluating the use of large language models to provide clinical recommendations in the emergency department. *Nature Communications, 15*, 8236. [PubMed PMID: 39379357].

Xu, L., Zhao, W., & Huang, X. (2025). Diagnosis and triage performance of contemporary large language models on short clinical vignettes. *Journal of Medical Systems, 49*, 141. https://doi.org/10.1007/s10916-025-02284-y

Yazaki, M., Maki, S., Furuya, T., Nakada, T., & Ohtori, S. (2025). Emergency patient triage improvement through a retrieval-augmented generation-enhanced large language model. [PubMed PMID: 38950135].

Yudhistira, P. C. Y., Malik, D. R., & Yudistira, N. (2026). Does language shift break medical vision-language models? Indonesian radiology visual question answering case study. arXiv:2606.03693.