

Validation of Ailix: Gastrointestinal (GI) Triage as an Artificial Intelligence–Based Consultation Solution for Patients with Gastrointestinal Diseases in Indonesia

Anita K. Tarigan^{1, 2*}, Salma Nur Fadhilah², Rifanditto Adhikara², Ira Andriani², Hyeonseung Oh³, Eun S. Kim³, Jonghyung Park³

¹Sirka, Jakarta, Indonesia, ²PT. Sirka Integra Medika, Tangerang, Indonesia, ³Meninblox Inc, Gwangju, Republic of Korea

Submitted to Journal:
Frontiers in Digital Health

Specialty Section:
Health Informatics

Article type:
Original Research Article

Manuscript ID:
1907473

Received on:
12 Jun 2026

Journal website link:
www.frontiersin.org

Scope Statement

This manuscript reports a clinical validation study of Ailix: GI Triage, an Indonesian-language large language model (LLM) consultation platform for gastrointestinal disorders. Using a within-subject design with 102 participants, we compared two AI configurations: a structured rule-based pathway and a free-conversation pathway, against physician assessment as the gold standard, evaluating red flag detection, diagnostic agreement, clinician-rated quality, and user acceptance. The study fits the scope of Frontiers in Digital Health - Health Informatics specialty in three ways. First, it is a primary validation study of an AI-based clinical decision support tool, evaluating safety, diagnostic performance, and usability in a single integrated framework. Second, it addresses a documented evidence gap by being, to the best of our knowledge, the first published validation of an LLM-based GI consultation assistant in the Indonesian language, a setting underrepresented in the existing AI-in-medicine literature. Third, the principal finding (a dissociation between user preference and safety performance) directly informs the design of patient-facing AI tools, an active area of methodological and ethical discussion in digital health. The work is methodologically rigorous (Gwet's AC1, paired comparisons, thematic analysis) and provides practical recommendations for hybrid AI triage design.

Conflict of interest statement

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest

Credit Author Statement

Anita Khairani Tarigan: Conceptualization, Investigation, Methodology, Writing – original draft, Writing – review & editing. **Jonghyung Park:** Funding acquisition, Supervision, Writing – review & editing. **Rifanditto Adhikara:** Funding acquisition, Resources, Supervision, Writing – review & editing. **Ira Andriani:** Project administration, Resources, Supervision, Writing – review & editing. **Eun Sil Kim:** Project administration, Resources, Writing – review & editing. **Hyeonseung Oh:** Conceptualization, Software, Writing – review & editing. **Salma Nur Fadhillah:** Conceptualization, Data curation, Formal Analysis, Validation, Writing – original draft, Writing – review & editing.

Keywords

artificial intelligence, Clinical decision support, Digital Health, gastrointestinal triage, Indonesia, Large language models, usability, Validation study

Abstract

Word count: 344

Introduction: Gastrointestinal (GI) disorders affect a large share of the Indonesian population, but access to specialist care is limited and unevenly distributed. Large language models offer a possible way to widen access to initial GI consultation, but no published study has validated such a system in the Indonesian language. This study evaluated Ailix: GI Triage, an Indonesian-language consultation platform, by comparing a structured rule-based pathway and a free-conversation pathway against physician assessment. **Methods:** A single-center, observational, cross-sectional study with a within-subject design was conducted online. Adults with a history of GI complaints were recruited from active clients of a digital health service. Each participant completed both pathways in a single session, then a satisfaction and preference questionnaire. Physician independently validated both pathway outputs as the gold standard for red flag detection, diagnosis, and clinical quality. Analyses included red flag detection metrics, Top-1 and Top-3 diagnostic match, diagnosis-group agreement using Gwet's first-order agreement coefficient, paired McNemar and Wilcoxon signed-rank tests, binomial tests for preference items, and thematic analysis of open-ended feedback. **Results:** One hundred and two participants completed both pathways. The structured pathway identified true red flag cases with 97.5% sensitivity and 80.6% specificity, while the conversational pathway classified all cases as non-red flag, with 0.0% sensitivity and twelve false-negative cases. Top-1 diagnostic match was 44.1% for the structured pathway and 39.2% for the conversational pathway, with no significant difference. The structured pathway showed substantial diagnosis-group agreement with physician, while the conversational pathway showed moderate agreement. Clinical reviewers rated the structured pathway higher for safety net performance and higher overall, while the conversational pathway was rated higher for clinical accuracy. Overall satisfaction did not differ between pathways, but participants preferred the conversational pathway as more helpful, more trustworthy, and preferred for future use. Thematic analysis supported a hybrid design. **Conclusion:** Neither pathway was superior across all dimensions. The structured pathway was safer, the conversational pathway had higher reviewer-rated clinical accuracy and greater user preference, and the safest configuration appears to be a hybrid design with a mandatory rule-based safety layer and a conversational symptom-elaboration layer.

Funding information

This study was funded by MENINBLOX Inc., based in Gwangju, Republic of Korea, through a research grant awarded under a memorandum of understanding with Sirka Science Ltd, Singapore. The research was carried out by PT. Sirka Integra Medika, Jakarta, Indonesia, the operating affiliate of Sirka Science Ltd. The funder had no role in study design, data collection, data analysis, interpretation of results, or the decision to publish.

Funding statement

The author(s) declare that financial support was received for the research and/or publication of this article.

Ethics statements

Studies involving animal subjects

Generated Statement: No animal studies are presented in this manuscript.

Studies involving human subjects

Generated Statement: The studies involving humans were approved by STIKes Mayapada Research Ethics Committee, Jakarta, Indonesia. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Inclusion of identifiable human data

Generated Statement: No potentially identifiable images or data are presented in this study.

Data availability statement

Generated Statement: The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Generative AI disclosure

The author(s) verify and take full responsibility for the use of generative AI in the preparation of this manuscript. Generative AI was used (Anthropic models including Claude Opus 4 and Claude Opus 4.7) was used by the authors to assist with translation from Indonesian to English, proofread the manuscript, and explore potential papers. The tool was not used for data generation, data analysis, or statistical computation. All output was reviewed, verified, and edited by the authors, who take full responsibility for the content of the manuscript.

Validation of Ailix: Gastrointestinal (GI) Triage as an Artificial Intelligence–Based Consultation Solution for Patients with Gastrointestinal Diseases in Indonesia

Anita Khairani Tarigan^{1*}, Salma Nur Fadhilah¹, Rifanditto Adhikara^{1,2}, Ira Andriani¹, Hyeonseung Oh³, Eun Sil Kim³, Jonghyung Park³

¹PT. Sirka Integra Medika, Jakarta, Indonesia

²Sirka Science Ltd, Singapore

³Meninblox Inc., Gwangju, 61008, Republic of Korea

* Correspondence:

Anita Khairani Tarigan

anitakhairanit@gmail.com

Keywords: artificial intelligence, large language models, clinical decision support, gastrointestinal triage, digital health, Indonesia, usability, validation study

[Word count: 8658 words, Table count: 10 tables]

Abstract

Introduction: Gastrointestinal (GI) disorders affect a large share of the Indonesian population, but access to specialist care is limited and unevenly distributed. Large language models offer a possible way to widen access to initial GI consultation, but no published study has validated such a system in the Indonesian language. This study evaluated Ailix: GI Triage, an Indonesian-language consultation platform, by comparing a structured rule-based pathway and a free-conversation pathway against physician assessment. **Methods:** A single-center, observational, cross-sectional study with a within-subject design was conducted online. Adults with a history of GI complaints were recruited from active clients of a digital health service. Each participant completed both pathways in a single session, then a satisfaction and preference questionnaire. Physician independently validated both pathway outputs as the gold standard for red flag detection, diagnosis, and clinical quality. Analyses included red flag detection metrics, Top-1 and Top-3 diagnostic match, diagnosis-group agreement using Gwet's first-order agreement coefficient, paired McNemar and Wilcoxon signed-rank tests, binomial tests for preference items, and thematic analysis of open-ended feedback. **Results:** One hundred and two participants completed both pathways. The structured pathway identified true red flag cases with 97.5% sensitivity and 80.6% specificity, while the conversational pathway classified all cases as non-red flag, with 0.0% sensitivity and twelve false-negative cases. Top-1 diagnostic match was 44.1% for the structured pathway and 39.2% for the conversational pathway, with no significant difference. The structured pathway showed substantial diagnosis-group agreement with physician, while the conversational pathway showed moderate agreement. Clinical reviewers rated the structured pathway higher for safety net performance and higher overall, while the conversational pathway was rated higher for clinical accuracy. Overall satisfaction did not differ between pathways, but participants preferred the conversational pathway as more helpful, more trustworthy, and preferred for future use. Thematic analysis supported a hybrid design. **Conclusion:** Neither pathway was superior across all dimensions. The structured pathway was safer, the conversational pathway had higher reviewer-rated clinical accuracy and greater user preference, and the safest configuration

appears to be a hybrid design with a mandatory rule-based safety layer and a conversational symptom-elaboration layer.

1 Introduction

Gastrointestinal (GI) disorders are a common health problem in Indonesia. National data showed that the prevalence of gastritis is 40.8% (1), while the prevalence of gastroesophageal reflux disease (GERD) is 27.4% and continues to rise (2). These conditions affect a large share of the population and create a high demand for clinical consultation.

At the same time, access to specialist care in Indonesia is limited and unevenly distributed. The ratio of medical specialists is only 0.21 per 1,000 population, below the ideal target of 0.28 per 1,000 (3). In addition, about 59% of specialists are concentrated on the island of Java, which leaves regions outside Java with very limited access (4). The gap between a high burden of GI disease and low access to specialists creates an urgent need for safe and validated solutions that can widen the reach of GI consultation.

Artificial intelligence (AI), and in particular large language models (LLMs), has shown potential to support clinical decision-making. Ong et al. demonstrated that an LLM used as a co-pilot together with health professionals improved the accuracy of medication-related problem identification by 32.6% compared with health professionals working without AI support, across 16 clinical specialties (5). Gaber et al. showed that LLM workflows can effectively support triage, specialist referral, and diagnosis prediction in real health-care contexts (6). In gastroenterology specifically, a systematic review identified the role of LLMs in improving diagnostic accuracy, automating documentation, and supporting patient education (7), and a comparative study found that several LLMs reached a high level of diagnostic coverage on challenging GI cases (8).

Despite this growing evidence, important gaps remain. A systematic review of LLMs as health advisors found that most studies relied on subjective judgment as the ground truth, and that fewer than one-third addressed ethical, regulatory, and patient-safety issues (9). Furthermore, most existing studies were conducted in English or other major languages, and few were carried out in low-resource or culturally specific settings. The World Health Organization has emphasized the importance of developing digital health technologies that fit local language and cultural contexts in order to reduce gaps in access (10). To the best of our knowledge, no published study has validated an LLM-based AI assistant for GI consultation in the Indonesian language.

This study addresses that gap by validating Ailix: GI Triage, an LLM-based consultation platform designed for triage of GI disorders in the Indonesian language. The platform compares two different AI approaches: MediKoGPT, a structured (rule-based) model that follows a fixed clinical flow, and ChatGPT, a free-conversation (conversational) model. These two approaches have different strengths and limitations: a structured model is more consistent but less flexible, while a conversational model is more natural but may be more prone to inaccurate output. Comparing them in the same study allows their relative performance and user acceptance to be assessed directly.

Accordingly, this study had three aims: (1) to evaluate the level of agreement and diagnostic performance of AI triage outputs compared with physician assessment as the gold standard; (2) to compare user satisfaction and preference between the structured model (MediKoGPT) and the free-conversation model (ChatGPT); and (3) to identify weaknesses of the system and recommendations for further development. By evaluating clinical performance, measuring user acceptance, and identifying system weaknesses in a real-world Indonesian context, this study aims to provide early evidence on the feasibility, safety considerations, and development priorities of an Indonesian-language AI assistant for GI consultation.

2 Methods

2.1 Study design

This study used an observational, cross-sectional design with a quantitative, descriptive-analytic approach. No medical intervention, drug, or invasive procedure was given to any participant. Participants only used a digital artificial intelligence (AI) platform and completed online questionnaires.

A within-subject design was applied. Each participant used both AI models in the same session, so that every participant acted as their own comparison. This design was chosen because it reduces variation between individuals and allows a direct comparison of user preference between the two models.

The study was a single-center digital study. Although participants took part from different locations across Indonesia, all study infrastructure was based on one platform (Ailix: GI Triage, accessed through the customer relationship management (CRM) system of Sirka), managed by one research team, using one database, and reviewed by one research ethics committee.

2.2 Study setting

Data were collected fully online through the Ailix: GI Triage platform. There was no single physical study site. Participants took part on their own smartphone or computer, from any location with an internet connection. The online setting was chosen because the target population, active Sirka clients with gastrointestinal complaints, was geographically dispersed across Indonesia. This setting also reflected the intended real-world use of Ailix: GI Triage as a digital consultation support platform.

2.3 Aims, research questions, and hypothesis

The general aim of this study was to validate the Ailix: GI Triage platform as an AI-based consultation solution for patients with gastrointestinal (GI) disorders in Indonesia. The specific aims were: (1) to evaluate the level of agreement between the AI triage output and physician assessment; (2) to compare user satisfaction between two different AI models; and (3) to identify weaknesses of the system and recommendations for further development.

Three research questions were addressed:

1. How accurate is the Ailix: GI Triage output compared with physician assessment as the gold standard?

2. Is there a significant difference in user satisfaction between the structured AI model (MediKoGPT) and the free-conversation AI model (ChatGPT)?
3. What weaknesses and areas for development can be identified from the user feedback and clinical review?

The study evaluated whether the AI-generated triage outputs showed acceptable agreement with physician assessment and whether performance differed between the structured and conversational pathways. Clinical performance was assessed using diagnostic accuracy metrics, paired comparison tests, and agreement statistics. For planning and benchmarking purposes, the study used an expected agreement target of 95%, corresponding to a Gwet's AC1 of ≥ 0.95 .

2.4 Study variables

The independent variable was the type of AI model: MediKoGPT (a structured, rule-based model) and ChatGPT (a free-conversation model). The main clinical outcome variables were red flag detection, red flag accuracy, Top-1 diagnostic match, Top-3 diagnostic match, diagnosis-group agreement, and clinical reviewer scores. Clinical reviewer scores covered five domains: anamnesis, safety net, clinical accuracy, comprehensiveness, and language.

User experience outcomes included total satisfaction score and direct preference items. Direct preference items assessed which pathway was easier to use, more helpful for expressing symptoms, more trustworthy, and preferred for future use. Participant demographic characteristics, including age group and sex, were used as descriptive and exploratory variables.

2.5 The Ailix: GI Triage platform

Ailix: GI Triage is a consultation platform based on a large language model (LLM), designed for triage of gastrointestinal disorders in the Indonesian language. The platform compared two different AI approaches: (a) MediKoGPT, a structured questionnaire or rule-based pathway that follows a fixed clinical flow; and (b) ChatGPT, a conversational pathway that allows freer symptom expression.

Both pathways were evaluated for their ability to collect symptom information, identify red flag features, and generate diagnostic outputs. Physician validation of the AI output was applied as a human-in-the-loop safety and quality mechanism. The AI output was not intended to replace physician diagnosis or emergency medical assessment.

2.6 Population and sampling

The target population was adults in Indonesia who had experienced or were currently experiencing GI complaints. The accessible population was active Sirka clients who met the inclusion criteria.

Participants were selected using purposive sampling through the Sirka CRM system. This non-probability method was chosen because the study required participants with specific characteristics, namely a history of GI complaints and active Sirka client status, which reflect the real user profile of Ailix: GI Triage.

2.6.1 Inclusion and exclusion criteria

Inclusion criteria were: (1) aged at least 18 years, all sexes; (2) active Sirka client, with no restriction by ethnicity, socioeconomic group, or geographic location; (3) a history of or active GI complaint (gastritis, gastroesophageal reflux disease (GERD), dyspepsia, irritable bowel syndrome, chronic bloating, or persistent bowel disturbance); (4) access to the internet and a digital device; (5) able to read the Indonesian language; and (6) willing to provide digital informed consent.

Exclusion criteria were: (1) unable to operate the digital platform independently; (2) refusal to provide informed consent; or (3) incomplete data from either consultation pathway. Participants who reported potentially urgent alarm symptoms during the consultation process were not treated by the AI system and were advised to seek appropriate medical care. These red flags outputs were retained for clinical validation because red flag detection was one of the main study outcomes

2.6.2 Sample size

The minimum sample size was calculated using a precision-based formula for estimation in inter-rater agreement studies:

$$n = \frac{(Z^2 \times P \times (1 - P))}{E^2}$$

where n is the minimum sample size, $Z_{\alpha/2}$ is the Z value for a 95% confidence level (1.96), P is the expected agreement, and E is the margin of error.

With an expected agreement of $P = 0.95$, a margin of error of $E = 0.05$, and $Z = 1.96$, the estimated minimum sample size was:

$$n = \frac{(1.96^2 \times 0.95 \times 0.05)}{0.05^2} = 72.99 \approx 73 \text{ participants}$$

The target sample size was increased to a minimum of 100 participants to account for an expected dropout or incomplete data of about 15–20%, to provide adequate data for paired comparison between the two pathways, and to support qualitative analysis of open-ended feedback.

2.7 Recruitment and data collection

Recruitment was carried out through a broadcast message in the Sirka CRM (WhatsApp Business) sent to active clients who met the criteria. Participation was voluntary and without pressure. No public advertising on social or mass media was used, and no personal face-to-face or telephone approach was made, in order to avoid social pressure. The invitation contained a short explanation of the study and a direct link to the Ailix: GI Triage platform.

Data collection took place in a single session of about 10–15 minutes per participant, in the following sequence: (1) the participant read the information sheet and provided digital informed consent through a checkbox with a timestamp; (2) the participant completed a demographic questionnaire (date of birth and sex) and a GI symptom history questionnaire; (3) the participant randomly assigned to complete either an AI consultation using Model A (MediKoGPT, structured/rule-based flow) first, or an AI consultation using Model B (ChatGPT, free-conversation flow); (4) the participant completed the user satisfaction questionnaires (adapted from SUS and MAUQ) for each model and

188 provided open-ended feedback; and (5) independently, Sirka physician validated the triage output of
189 both AI models as the gold standard.

190 Roles in data collection were as follows: the lead researcher was responsible for overall coordination
191 and data quality monitoring; the Sirka Customer Experience team managed the distribution of
192 invitations through the CRM and provided technical support to participants; Sirka physician
193 performed clinical validation of all AI triage outputs; and the data analyst was responsible for data
194 cleaning, coding, and statistical analysis.

195 2.8 Instruments

196 The study used four instruments: a demographic questionnaire, a gastrointestinal symptom
197 questionnaire, a physician clinical review form, and a user satisfaction and preference questionnaire.
198 The user questionnaire included adapted items based on the System Usability Scale (SUS) and the
199 mHealth App Usability Questionnaire (MAUQ). User experience outcomes included total
200 satisfaction score and direct preference items. The total satisfaction score was calculated separately
201 for each pathway from 11 Likert-scale items rated from 1 to 5, covering ease of use, ease of
202 understanding, perceived complexity, familiarity, difficulty of use, overall usability satisfaction,
203 symptom expression, question appropriateness, content clarity, usefulness as a health consultation
204 tool, and willingness to use the pathway again. Direct preference items assessed which pathway was
205 easier to use, more helpful for expressing symptoms, more trustworthy, and preferred for future use.
206 Participant demographic characteristics, including age group and sex, were used as descriptive and
207 exploratory variables.

208 The physician clinical review form was used to assess AI outputs in five domains: anamnesis, safety
209 net, clinical accuracy, comprehensiveness, and language. Physician reviewers also assigned gold
210 standard diagnoses and diagnosis groups for comparison with AI-generated outputs.

211 2.9 Data management

212 Data were collected automatically through the Ailix: GI Triage platform and stored in an encrypted
213 Google Spreadsheet with access limited to the research team. AI consultation transcripts were
214 recorded directly in the platform database. Each participant was assigned a unique code under a
215 pseudonymization system that replaced personal identifiers. Personal identifiers (name and telephone
216 number) were kept in a separate spreadsheet accessible only to the lead researcher for verification
217 purposes only. Data were not transferred to any third party outside the research team. Data handling
218 followed the principles of the Indonesian Personal Data Protection Law No. 27 of 2022, including a
219 lawful basis for processing, data minimization, encryption, and respect for the rights of data subjects.

220 2.10 Data analysis

221 Data analysis combined quantitative statistical analysis and qualitative thematic analysis, aligned
222 with the clinical validation, user experience, and system development aims of the study.

223 The quantitative analysis had three components. First, descriptive analysis was used to describe
224 participant characteristics and the clinical case profile. Categorical variables, including sex, age
225 group, red flag status, and physician-defined diagnosis groups, were presented as frequencies and
226 percentages. Reviewer scores and satisfaction scores were summarized using medians and
227 interquartile ranges because these variables were ordinal or non-normally distributed.

Second, clinical validation analysis was performed by comparing AI-generated outputs from each pathway with physician assessment as the gold standard. Red flag detection was evaluated using overall accuracy, confusion matrices, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and Gwet's AC1. Diagnostic performance was evaluated using Top-1 and Top-3 diagnostic match. Paired differences between the structured and conversational pathways for binary outcomes, including red flag accuracy, Top-1 diagnostic match, and Top-3 diagnostic match, were assessed using McNemar tests. Diagnosis-group agreement between AI output and physician assessment was assessed using Gwet's AC1. Sensitivity analyses were also conducted using exact Top-1 diagnosis-label agreement and exact Top-3 diagnosis-set agreement. Gwet's AC1 was selected because it is more robust than Cohen's Kappa when category distribution is imbalanced. Agreement was interpreted using the Landis and Koch scale: ≥ 0.81 = almost perfect, $0.61-0.80$ = substantial, $0.41-0.60$ = moderate, $0.21-0.40$ = fair, and ≤ 0.20 = slight or poor agreement.

Third, comparative and exploratory analyses were conducted to compare the structure and conversational pathways. McNemar tests were used for paired binary outcomes, including red flag accuracy, Top-1 diagnostic match, and Top-3 diagnostic match. Wilcoxon signed-rank tests were used to compare paired ordinal or non-normally distributed scores, including reviewer domain scores, total reviewer scores, and satisfaction scores. Effect sizes for Wilcoxon analyses were calculated as follows,

$$r = |Z| / \sqrt{N}$$

Direct user preference items were analyzed using binomial tests against an expected 50:50 distribution. Exploratory subgroup analyses assessed whether model performance varied by diagnosis group, sex, and age group, using descriptive summaries, chi-square tests, or exact tests when expected cell counts were small. A two-sided p-value of <0.05 was considered statistically significant.

The qualitative analysis was a thematic analysis of the participants' open-ended feedback, following the six-phase approach of Braun and Clarke familiarization with the data, generating initial codes, searching for themes, reviewing themes, defining themes, and reporting (16). This analysis aimed to identify the strengths, weaknesses, and development suggestions for the platform from the user perspective.

Statistical analyses were performed using IBM SPSS Statistics version 31.0.2.0 and R version 4.4.1. Gwet's AC1 and its confidence interval were calculated using the R package "irrCAC". Thematic analysis was supported by manual coding in Microsoft Excel.

2.11 Ethical considerations

This study was reviewed and approved by the Research Ethics Committee of Sekolah Tinggi Ilmu Kesehatan (STIKes) Mayapada, Jakarta, Indonesia (approval number 002/STIKES-M/KEP/I/SLE/IV/2026, dated 27 April 2026). The ethics review was conducted by an accredited, independent committee not affiliated with Sirka or the study sponsor, in order to avoid a conflict of interest. The study was guided by the three basic ethical principles (respect for persons, beneficence, and justice), the seven World Health Organization standards for health research, and the 2016 CIOMS international ethical guidelines.

Digital informed consent was obtained from every participant before any study activity. The information sheet explained the purpose, procedures, risks, benefits, data confidentiality, and the right to withdraw at any time without any effect on the Sirka services the participant received. Consent was given actively through a digital checkbox with a timestamp.

The study carried minimal risk, as it involved only the completion of online questionnaires with no medical procedure. Identified risks and their mitigation were: an explicit disclaimer that the AI output is not a physician's diagnosis; data encryption, pseudonymization, and compliance with the Personal Data Protection Law for data privacy; and an automatic alarm-symptom detection system with physician validation (human-in-the-loop) for AI error. Participants who completed the full study received compensation of IDR 35,000 in digital wallet (GoPay) balance, given after completion as recognition of the time spent (10–15 minutes), not as an inducement to participate.

3 Results

3.1 Participants and case characteristics

A total of 102 participants completed both consultation pathways and were included in the final analysis (Table 1). There were no missing data for the main demographic and outcome variables. Most participants were female, with 75 participants (73.5%), while 27 participants (26.5%) were male. The largest age group was 20-30 years, with 62 participants (60.8%), followed by 31-40 years with 25 participants (24.5%) and >40 years with 15 participants (14.7%).

Based on physician-defined groups (Table 2) from the MQ pathway, the most common case groups were upper GI acid-peptic / foregut disorders and functional bowel / bowel habit disorders, each with 33 cases (32.4%). Infections / inflammatory bowel disorders and acute surgical / obstruction each accounted for 12 cases (11.8%). In the Convo pathway, the most common physician-defined diagnosis group was functional bowel / bowel habit disorder with 38 cases (37.3%), followed by upper GI acid-peptic / foregut disorders with 27 cases (26.5%). The non-diagnostic / non-GI / systemic category was more frequent in the Convo pathway than in the MQ pathway, with 20 cases (19.6%) vs 4 cases (3.9%).

Table 1. Participant characteristics

Characteristic	n (%)
Sex	
Male	27 (26.5)
Female	75 (73.5)
Age group	
20–30 years	62 (60.8)
31–40 years	25 (24.5)
>40 years	15 (14.7)
Total	102 (100.0)

298 **Table 2.** Physician-defined diagnosis groups by consultation pathway

Diagnosis group	MQ, n (%)	Convo, n (%)
Non-diagnostic / non-GI / systemic	4 (3.9)	20 (19.6)
Upper GI acid-peptic / foregut	33 (32.4)	27 (26.5)
Infectious / inflammatory bowel	12 (11.8)	14 (13.7)
Functional bowel / bowel habit disorder	33 (32.4)	38 (37.3)
Acute surgical abdomen / obstruction	12 (11.8)	0 (0.0)
Hepatobiliary / pancreatic	5 (4.9)	0 (0.0)
Anorectal / colorectal structural	0 (0.0)	1 (1.0)
Malabsorption / food intolerance / post-surgical	3 (2.9)	2 (2.0)
Total	102 (100.0)	102 (100.0)

299

300 3.2 Clinical validation of red flag detection

301 Overall red flag accuracy was 87.3% for MQ and 88.2% for Convo (Table 3). McNemar test showed
 302 no statistically significant difference in overall red flag accuracy between the two pathways ($p =$
 303 1.000). However, diagnostic performance analysis showed an important clinical difference between
 304 the two models. MQ identified 39 true-positive red flag cases and missed one red flag case, resulting
 305 in sensitivity of 97.5%, specificity of 80.6%, PPV of 76.5%, and NPV of 98.0%. In contrast, Convo
 306 classified all cases as non-red flag. This resulted in no true-positive cases, 12 false-negative cases,
 307 sensitivity of 0.0%, specificity of 100%, and NPV of 88.2%.

308 Agreement analysis using Gwet's AC1 showed substantial agreement of MQ red flag detection, with
 309 AC1 of 0.748 (95% CI 0.617-0.879). Convo showed almost perfect agreement, with AC1 of 0.868
 310 (95% CI 0.788-0.948). However, the higher AC1 in Convo should be interpreted with caution
 311 because it was driven by the predominance of non-red flag classifications and the absence of positive
 312 red flag outputs.

313 **Table 3.** Red flag detection performance compared with physician assessment

Model	TP	FP	TN	FN	Sensitivity	Specificity	PPV	NPV	Accuracy	Gwet's AC1
MQ	39	12	50	1	97.5%	80.6%	76.5%	98.0%	87.3%	0.748 (95% CI 0.617-0.879)
Convo	0	0	90	12	0.0%	100.0%	NA	88.2%	88.2%	0.868 (95% CI 0.788-0.948)

314 *TP = True Positive, FP = False Positive, TN = True Negative, FN = False Negative, PPV = Positive Predictive Value,*
 315 *NPV = Negative Predictive Value*

316

317 3.3 Diagnostic match and diagnosis-group agreement

318 Top-1 diagnosis match (Table 4) was achieved in 45 of 102 MQ cases (44.1%) and 40 of 102 Convo
 319 cases (39.2%). McNemar test showed no statistically significant difference between the two
 320 pathways for Top-1 diagnosis match ($\chi^2 = 0.291$; $p = 0.590$). Top-3 diagnosis match was achieved in
 321 26 of 102 MQ cases (25.5%) and 23 of 102 Convo cases (22.5%). This difference was also not
 322 statistically significant ($\chi^2 = 0.129$; $p = 0.719$).

When diagnosis agreement was assessed at the diagnosis-group level, MQ showed stronger agreement with physician-defined diagnosis groups than Convo. MQ had a percent agreement of 65.7% and Gwet's AC1 of 0.615 (95% CI 0.510–0.721), indicating substantial agreement. Convo had a percent agreement of 57.8% and Gwet's AC1 of 0.533 (95% CI 0.424–0.642), indicating moderate agreement.

Sensitivity analyses using exact diagnosis labels showed lower agreement than diagnosis-group analysis. For exact Top-1 diagnosis-label agreement, Gwet's AC1 was 0.421 for MQ and 0.368 for Convo. For exact Top-3 diagnosis-set agreement, Gwet's AC1 was 0.247 for MQ and 0.217 for Convo. These lower values were expected because exact diagnosis labels and Top-3 diagnosis sets were more granular and clinically overlapping than broader diagnosis groups.

Table 4. Diagnostic match and diagnosis-group agreement

Outcome	MQ	Convo	p-value	Interpretation
Primary diagnostic validation				
Top-1 diagnosis match	45/102 (44.1%)	40/102 (39.2%)	0.590	Not significantly different
Top-3 diagnosis match	26/102 (25.5%)	23/102 (22.5%)	0.719	Not significantly different
Diagnosis group Top-1 agreement	65.7%; AC1 = 0.615	57.8%; AC1 = 0.533	NA	MQ showed stronger agreement
Sensitivity analysis				
Exact Top-1 diagnosis-label agreement	44.1%; AC1 = 0.421	39.2%; AC1 = 0.368	NA	Lower agreement than diagnosis-group analysis
Exact Top-3 diagnosis-set agreement	25.5%; AC1 = 0.247	22.5%; AC1 = 0.217	NA	Lower agreement in both pathways

3.4 Clinical reviewer assessment

Clinical reviewer scores were compared across five domains: anamnesis, safety net, clinical accuracy, comprehensiveness, and language. Wilcoxon signed-rank test showed no significant difference between MQ and Convo for anamnesis, comprehensiveness, or language. Significant differences were found for safety net, clinical accuracy, and total reviewer score.

MQ had better safety net performance than Convo ($p < 0.001$; $r = 0.393$). This was the largest effect size among all reviewer domains. Convo had better clinical accuracy than MQ ($p = 0.004$; $r = 0.286$). Total reviewer score was significantly higher for MQ than Convo, with median 12.00 (11.00–14.00) vs 11.00 (10.00–13.00), respectively ($p = 0.003$; $r = 0.293$). These findings suggest that MQ performed better in safety-related aspects, while Convo performed better in clinical accuracy.

Table 5. Clinical reviewers' scores for MQ and Convo

Domain	MQ median (Q1–Q3)	Convo median (Q1–Q3)	p-value	Effect size r
Anamnesis	2.00 (2.00–2.00)	2.00 (2.00–2.00)	0.462	0.073
Safety net	2.00 (2.00–2.00)	2.00 (1.00–2.00)	<0.001*	0.393
Clinical accuracy	2.00 (1.00–2.00)	2.00 (2.00–3.00)	0.004*	0.286
Comprehensiveness	2.00 (2.00–2.00)	2.00 (2.00–3.00)	0.140	0.146
Language	3.00 (3.00–3.00)	3.00 (3.00–3.00)	0.096	0.165
Total reviewer score	12.00 (11.00–14.00)	11.00 (10.00–13.00)	0.003*	0.293

3.5 User satisfaction and preference

User satisfaction scores did not differ significantly between MQ and Convo (Table 6). The median satisfaction score was 42.50 (38.75-50.00) for MQ and 43.00 (39.00-50.00) for Convo. Wilcoxon signed-rank test showed no significant difference between the two pathways ($Z = -0.429$; $p = 0.668$; $r = 0.042$). Compared with MQ, Convo had higher satisfaction scores in 38 participants, equal satisfaction scores in 34 participants, and lower satisfaction scores in 30 participants.

Direct user preference showed a different pattern. Convo was selected more frequently than MQ for all four preference items. However, the difference was not significant for ease of use. Convo was selected as easier to use by 57 participants (55.9%), compared with 45 participants (44.1%) for MQ ($p = 0.276$). In contrast, Convo was selected significantly more often as more helpful in expressing symptoms, more trustworthy, and preferred for future use. Convo was selected as more helpful by 67 participants (63.7%; $p = 0.007$), and preferred for future use by 71 participants (69.6%, $p < 0.001$).

Table 6. User satisfaction and preference

Outcome	MQ	Convo	<i>p</i> -value
Satisfaction score, median (Q1–Q3)	42.50 (38.75–50.00)	43.00 (39.00–50.00)	0.668
Easier to use	45 (44.1%)	57 (55.9%)	0.276
More helpful	35 (34.3%)	67 (65.7%)	0.002*
More trustworthy	37 (36.3%)	65 (63.7%)	0.007
Preference for future use	31 (30.4%)	71 (69.6%)	<0.001*

Satisfaction score was analyzed using Wilcoxon signed-rank test. Preference items were analyzed using a binomial test against a 50:50 expected distribution.

3.6 Qualitative user and physician feedback

Thematic analysis of user feedback identified six themes (Table 7). Users valued the structured questionnaire because it was fast, simple, and easy to complete. However, fixed answer options sometimes limited their ability to describe symptoms accurately. In contrast, the conversational format allowed freer symptom expression and felt closer to a real consultation. Some users still found the chat format effortful because it required typing and clear symptom description. Users also suggested more adaptive questions, clearer diagnostic explanations, stronger safety disclaimers, and more human communication. Overall, user feedback supported a hybrid design that combines structured options with free-text elaboration.

Doctor reviewer notes showed six recurring clinical issues (Table 8). These included incomplete symptom characterization, missed or underemphasized red flags, mismatch between diagnosis and symptom pattern, insufficient assessment of contextual risk factors, generic education, and variable performance when user input was vague. These findings indicate that Ailix needs stronger structured history-taking, mandatory safety screening, clearer diagnostic prioritization, and more tailored patient education.

Table 7. Thematic analysis of user feedback on Ailix: GI Triage

Theme	Definition	Coding references	Unique users	Representative quote	Interpretation
Structured questionnaire is fast and easy, but restrictive	Users valued the structured questionnaire for its speed and simplicity, but several felt that fixed answer options could not fully represent their symptoms.	116	71	<i>“Lebih gampang digunakan tinggal klik-klik aja pilihannya, tapi ada pertanyaan yang opsinya kurang lengkap...”</i> (“It was easier to use because I only had to click the available options, but some questions did not provide enough answer choices....”)	The structured questionnaire improved efficiency and ease of use, but limited symptom expression when user experiences did not fit predefined choices.
Conversational chat allows freer symptom expression	Users valued the conversational format because it allowed freer, richer, and more personal symptom descriptions.	63	36	<i>“Jawaban bebas dan seperti berkonsultasi dengan dokter.”</i> (“Free-text answers made it feel as if actually consulting with a doctor....”)	The conversational model improved perceived naturalness and helped users describe symptoms in their own words.
Chat-based interaction needs more effort and user ability	Users noted that chat required more effort, including typing, thinking, and describing symptoms clearly.	44	27	<i>“Kadang bingung menjelaskan gejala yang dirasakan dengan spesifik.”</i> (“Sometimes, it was difficult to explain the symptoms specifically....”)	Conversational interaction was not universally easier because it depended on the user’s ability to articulate symptoms.
Questions and flow need to be more adaptive, specific, and clinically guided	Users suggested that questions should be more specific, adaptive, and guided by previous answers.	26	14	<i>“Pertanyaannya masih terlalu general kurang spesifik dan detail.”</i> (“The questions were still too general and lacked specificity and detail.”)	Both methods need more adaptive follow-up questions that respond to the user’s previous answers.
Users need clearer, safer, and more human diagnostic feedback	Users wanted clearer diagnostic explanations, safer wording, referral guidance, and a more human tone.	35	17	<i>“Bisa ditambahkan disclaimer seperti: Bukan diagnosis medis, segera konsultasi dengan dokter.”</i> (A disclaimer could be added, such as: “This is not a medical diagnosis. Please consult a doctor immediately.”)	Users needed clearer diagnostic hierarchy, safety communication, and reassurance to reduce misunderstanding and anxiety.

Hybrid design is preferred for future development	Users repeatedly suggested combining structured choices with free-text or chat elements.	153	73	“Tetap wawancara via chat, tapi sediakan opsi juga.” (“Keep the chat-based interview, but also provide answer option.”)	The strongest user-driven recommendation was a hybrid flow that combines structured choices for clarity and safety with free-text input for flexibility.
---	--	-----	----	---	--

Table 8. Thematic analysis of doctor reviewer notes

Theme	Definition	Coding references	Unique cases	Clinical implication
Incomplete symptom characterization and chronology	Doctor notes indicated incomplete exploration of core GI symptom details, including location, onset, duration, severity, stool pattern, bowel habit, pain character, triggers, relieving factors, and symptom progression.	518	102	The AI needs more systematic GI history-taking prompts before generating diagnostic or educational output.
Missed or underemphasized red flags and safety escalation	Doctor notes mentioned alarm symptoms, red flags, or the need for clearer escalation, including bleeding, melena, weight loss, anemia, obstipation, persistent vomiting, dysphagia, severe pain, pregnancy-related caution, or urgent evaluation.	293	96	Safety-net logic should be mandatory, explicit, and rule-based, especially for alarm symptoms.
Diagnosis or recommendation mismatch with available clinical pattern	Doctor notes identified diagnostic outputs or recommendations that were too broad, premature, unsupported, or less consistent with the symptom pattern.	121	81	The AI needs better diagnostic prioritization, uncertainty framing, and alignment between history and output.
Insufficient assessment of contextual risk factors and secondary causes	Doctor notes identified missing contextual history, such as medication use, NSAID use, antibiotics, alcohol, H. pylori risk, pregnancy, thyroid disease, family history, surgery, diet, infection exposure, or psychosocial factors.	184	94	Clinical prompts should include context-sensitive risk factor checks based on the presenting GI complaint.
Generic or insufficiently tailored patient education	Doctor notes described education as too general, too broad, unfocused, incomplete, or not tailored to the most likely diagnosis and risk profile.	62	51	Patient education should be individualized, structured by likely diagnosis, and linked to clear next steps.

Variable system performance and need for robustness to patient input Doctor notes showed that the AI sometimes captured relevant context well, but performance varied depending on vague, contradictory, or incomplete user responses. 170 86 The platform should clarify contradictory answers, retain prior context, and combine conversational flexibility with structured clinical checks.

3.7 Exploratory analysis of factors associated with model performance

Exploratory subgroup analysis was conducted to assess whether model performance varied by diagnosis group, age, or sex (Table 9 and Table 10). Based on diagnosis group, diagnostic match varied across case categories. In the MQ pathway, Top-1 match was highest in hepatobiliary / pancreatic cases (80%) and infectious/inflammatory bowel cases (66.7%), while it was lower in acute surgical abdomen/obstruction cases (16.7%). In the Convo pathway, Top-1 match was highest in functional bowel/bowel habit disorder cases (57.9%) and upper GI acid-peptic / foregut cases (44.4%). These findings were interpreted descriptively because several groups had small case numbers.

Demographic exploratory analysis showed no significant association between age or sex and red flag accuracy in either pathway. MQ-Top1 match was also not associated with sex or age. For Convo, Top-1 match was higher among female participants, with 34 of 75 females (45.3%) and 6 of 27 males (22.2%) achieving Top-1 match (Pearson $p = 0.035$; Fisher exact $p = 0.04$). This finding was interpreted as exploratory because the male subgroup was smaller and the analysis was not powered for definitive subgroup inference.

Table 9. Exploratory diagnostic match by diagnostic group

Pathway	Diagnosis group	Total cases	Top-1 match, n (%)	Top-3 match, n (%)
MQ	Non-diagnostic / non-GI / systemic	4	0 (0.0)	2 (50.0)
MQ	Upper GI acid-peptic / foregut	33	15 (45.5)	7 (21.2)
MQ	Infectious / inflammatory bowel	12	8 (66.7)	2 (16.7)
MQ	Functional bowel / bowel habit disorder	33	14 (42.4)	7 (21.2)
MQ	Acute surgical abdomen / obstruction	12	2 (16.7)	4 (33.3)
MQ	Hepatobiliary / pancreatic	5	4 (80.0)	4 (80.0)
MQ	Malabsorption / food intolerance / post-surgical	3	2 (66.7)	0 (0.0)
Convo	Non-diagnostic / non-GI / systemic	20	0 (0.0)	0 (0.0)
Convo	Upper GI acid-peptic / foregut	27	12 (44.4)	15 (55.6)
Convo	Infectious / inflammatory bowel	14	4 (28.6)	1 (7.1)
Convo	Functional bowel / bowel habit disorder	38	22 (57.9)	6 (15.8)
Convo	Anorectal / colorectal structural	1	1 (100.0)	1 (100.0)
Convo	Malabsorption / food intolerance / post-surgical	2	1 (50.0)	0 (0.0)

Table 10. Exploratory demographic analysis of red flag accuracy and Top-1 match

Factor	Outcome	Main finding	<i>p</i> -value
Sex	MQ RF accuracy	Male 85.2%; Female 88.0%	Fisher <i>p</i> = 0.741
Sex	Convo RF accuracy	Male 88.9%; Female 88.0%	Fisher <i>p</i> = 1.000
Age	MQ RF accuracy	20–30: 88.7%; 31–40: 88.0%; >40: 80.0%	Exact <i>p</i> = 0.631
Age	Convo RF accuracy	20–30: 88.7%; 31–40: 88.0%; >40: 86.7%	Exact <i>p</i> = 1.000
Sex	MQ Top-1 match	Male 44.4%; Female 44.0%	Pearson <i>p</i> = 0.968
Sex	Convo Top-1 match	Male 22.2%; Female 45.3%	Pearson <i>p</i> = 0.035; Fisher <i>p</i> = 0.040
Age	MQ Top-1 match	20–30: 46.8%; 31–40: 44.0%; >40: 33.3%	Exact <i>p</i> = 0.690
Age	Convo Top-1 match	20–30: 40.3%; 31–40: 28.0%; >40: 53.3%	Exact <i>p</i> = 0.283

4 Discussion

4.1 Principal Findings

This study evaluated two Indonesian-language AI pathways for gastrointestinal triage in a within-subject design with 102 participants. The structured rule-based pathway (MQ) and the conversational pathway (Convo) showed different performance profiles across clinical and user outcomes. MQ performed better in the safety net domain and in the total reviewer score, while Convo performed better in clinical accuracy and was preferred by users for helpfulness, trustworthiness, and future use. Top-1 and Top-3 diagnostic match rates were not significantly different between the two pathways. Overall red flag accuracy appeared similar between MQ and Convo, but sensitivity values showed a large clinical difference that is more meaningful for safety. These findings show that no single pathway was superior across all dimensions, and that clinical performance and user preference did not always point in the same direction.

The observed agreement values did not reach the initial 95% benchmark. This indicates that the current version of Ailix: GI Triage should be interpreted as an early-stage AI triage support tool, not as a stand-alone diagnostic system. The findings support further refinement of the system, particularly in diagnosis mapping and mandatory red flag escalation.

Taken together, the dissociation between safety performance, diagnostic performance, user preference, and clinical reviewer judgement is the most informative finding of this study. The four outcome dimensions did not align, and each told a different story. Safety performance favoured MQ, diagnostic accuracy as judged by reviewers favoured Convo, user preference favoured Convo, and diagnosis-group agreement also favoured MQ. A simple summary statement that one pathway is "better" would therefore misrepresent the data. The practical interpretation is that no single end-to-end LLM pathway, structured or conversational, was simultaneously safe, accurate, and acceptable for unsupervised GI triage in Indonesian, and any deployable system must be designed against all four dimensions at once rather than optimised on one.

4.2 Safety-critical performance in red flag detection

The McNemar test result of *p* = 1.000 for overall red flag accuracy should not be read as evidence that the two pathways are equivalent in safety. MQ achieved a sensitivity of 97.5% and captured almost all true red flag cases. Convo, in contrast, classified all cases as non-red flag, with a sensitivity of 0.0%, and produced 12 false-negative cases. The apparent equivalence in accuracy was

432 driven by the high prevalence of non-red flag cases, which masked the difference in clinically critical
433 performance.

434 This pattern is a known limitation of accuracy as a single performance metric when class distribution
435 is imbalanced (13,14). The higher Gwet's AC1 observed for Convo (0.868) compared with MQ
436 (0.748) should also be interpreted with caution. Gwet's AC1 is more robust to the prevalence paradox
437 than Cohen's Kappa (13,14), but it can still reflect agreement that is dominated by the majority class.
438 In this study, the Convo AC1 value reflected agreement on non-red flag cases rather than agreement
439 on the detection of true red flags.

440 In GI triage, the clinical cost of false negatives is substantially higher than the cost of false positives,
441 because missing a serious presentation can delay urgent care. Previous work on LLMs in clinical
442 decision support has highlighted that hallucination and incomplete safety responses remain important
443 risks for patient-facing systems (18,19). Our results add a concrete demonstration of this risk in a real
444 triage workflow: an AI pathway can appear safe by overall accuracy while failing to detect any red
445 flag case in practice.

446 Our finding is consistent with the broader online symptom checker literature. Sutaria et al. evaluated
447 four commercial online symptom checkers and reported that they sought red flag symptoms in only a
448 small proportion of relevant cases, despite achieving acceptable overall triage accuracy on emergency
449 vignettes (22). Their conclusion that diagnostic accuracy alone is not an adequate measure of safety
450 for triage tools applies directly to our results. The Convo pathway in our study achieved 88.2%
451 overall accuracy and an AC1 of 0.868 on red flag detection, yet identified zero of the twelve true red
452 flag cases. Without disaggregating sensitivity and specificity, the system would appear acceptably
453 safe; with disaggregation, it is clear that the system would have missed every patient who needed
454 urgent escalation.

455 The 95% benchmark used in the sample-size calculation should also be reinterpreted in light of these
456 results. The observed AC1 of 0.748 for MQ red flag detection and 0.615 for MQ diagnosis-group
457 agreement both fell well below the benchmark, even though MQ was the better-performing pathway
458 across most clinical outcomes. This indicates that the benchmark was aspirational rather than
459 achieved, and that the current Ailix: GI Triage system, in either configuration, is not yet at the level
460 of agreement that would justify autonomous triage. The honest interpretation is that this study
461 establishes a baseline for further iteration, not evidence of clinical readiness.

462 4.3 Diagnostic agreement and clinical reviewer assessment

463 Top-1 and Top-3 diagnostic match rates were not significantly different between the two pathways,
464 although MQ showed slightly higher diagnosis-group agreement (Gwet's AC1 0.615 vs 0.533).
465 Reviewer scores showed a more nuanced pattern. MQ was rated higher on safety net and on total
466 reviewer score, while Convo was rated higher on clinical accuracy. A possible explanation for the
467 higher clinical accuracy of Convo is that a free-conversation pathway can capture richer context
468 about a patient's complaint, including subtle wording and additional symptom information that a
469 fixed questionnaire flow may not request. Similar effects of conversational design on information
470 capture have been described in other clinical chatbot studies (20).

471 However, higher clinical accuracy alone is not sufficient for a triage tool. Without a robust safety
472 layer, an accurate diagnostic suggestion that misses a red flag can still result in inappropriate self-
473 management. This is consistent with broader observations from LLM benchmarking work in clinical

474 decision support, which has shown that diagnostic performance can vary independently of safety-
475 related performance (6,8,9).

476 The sensitivity analysis using exact diagnosis labels gave a more conservative picture of diagnostic
477 performance. Exact Top-1 diagnosis-label agreement fell to AC1 0.421 for MQ and 0.368 for Convo,
478 and exact Top-3 diagnosis-set agreement was lower still, at AC1 0.247 and 0.217, respectively. The
479 drop from diagnosis-group agreement to exact-label agreement is large and clinically meaningful. It
480 indicates that while the system was reasonably able to place a case in the correct broad category (for
481 example, identifying a presentation as functional bowel disorder rather than upper GI disease), it was
482 much less reliable at producing the specific diagnostic label that a physician would assign. For a
483 triage tool, broad-category agreement may be sufficient if the next step is a referral decision rather
484 than treatment, but the gap between broad and specific agreement should be presented transparently
485 to users so that they do not over-interpret AI-generated diagnostic labels.

486 The exploratory diagnostic match by diagnosis group also deserves direct interpretation. In the MQ
487 pathway, Top-1 match was very high for hepatobiliary and pancreatic cases (80.0%, 4 of 5) and
488 infectious or inflammatory bowel cases (66.7%, 8 of 12), but low for acute surgical abdomen or
489 obstruction cases (16.7%, 2 of 12). In the Convo pathway, Top-1 match was highest for functional
490 bowel disorders (57.9%) and lowest for the non-diagnostic or non-GI category (0%, 0 of 20).
491 Although the small case counts in some groups require caution, the pattern suggests that performance
492 is uneven across the diagnostic spectrum: the platform was better at conditions with stereotyped
493 presentations and weaker at acute surgical presentations, which are precisely the conditions where a
494 missed red flag carries the greatest clinical cost. Any deployable version of Ailix should therefore
495 prioritise robust performance in the acute and serious end of the GI spectrum, not in the cases where
496 the system is already strong.

497 4.4 User satisfaction, preference, and qualitative feedback

498 Total user satisfaction scores were not significantly different between MQ and Convo, but the direct
499 preference items leaned clearly toward Convo. Convo was preferred for helpfulness in expressing
500 symptoms, trustworthiness, and future use, while ease of use did not differ significantly between the
501 two pathways. This pattern suggests that an aggregated satisfaction score and explicit preference
502 items can capture different aspects of user experience, and that a pathway preferred by users is not
503 necessarily the safest pathway.

504 The qualitative feedback from both users and clinicians supported this interpretation. Users described
505 Convo as more natural and easier to talk to, which is consistent with prior findings that
506 conversational design supports engagement in health chatbots (20,21). Clinicians, however, noted
507 that MQ provided more consistent safety-related questions, while Convo would need a mandatory red
508 flag layer to be acceptable for triage use. The combination of stronger user preference for the
509 conversational pathway and stronger safety performance of the structured pathway is the central
510 tension that the system design must resolve.

511 The qualitative findings explain the trade-off between user experience and clinical safety. The
512 structured questionnaire was efficient and consistent, but less flexible. The conversational model was
513 more natural and helped users explain symptoms, but it also required more effort and produced less
514 consistent clinical safety signals.

515 Doctor reviewer notes reinforce this concern. GI triage depends on detailed history-taking and careful
516 red flag detection. If symptom chronology, stool pattern, medication exposure, and alarm symptoms

517 are missed, diagnostic output may become too broad or premature. Therefore, conversational AI
518 should not function as the only safety layer.

519 The main implication is that Ailix should use a hybrid model. Conversational input can improve user
520 comfort and symptom expression, while structured rule-based questions should secure red flag
521 screening and key GI history elements. The final output should clearly separate likely diagnosis,
522 alternative diagnoses, red flag status, self-care advice, and when to seek medical care. This approach
523 preserves the user advantage of Convo while improving the safety and consistency of MQ.

524 The volume and consistency of the qualitative findings strengthen this interpretation. Doctor
525 reviewer notes recorded 518 coding references on incomplete symptom characterization across all
526 102 cases, and 293 references on missed or underemphasized red flags across 96 of 102 cases. These
527 were the two most frequent clinical issues in the entire qualitative dataset. The pattern is not isolated
528 to a few cases; it is systemic. At the same time, user feedback returned 153 coding references from
529 73 users requesting a hybrid design that combines structured and free-text elements. The agreement
530 between independent clinician concerns and user preferences on the same design direction is itself a
531 strong signal: both groups identified the same gap, from different vantage points, and pointed to the
532 same solution.

533 An exploratory finding deserves careful, non-overclaimed comment. In the Convo pathway, Top-1
534 diagnostic match was higher among female participants (45.3%, 34 of 75) than male participants
535 (22.2%, 6 of 27), with Fisher's exact $p = 0.040$. No comparable difference was seen in the MQ
536 pathway or for red flag accuracy. Given the small male subgroup ($n = 27$) and the exploratory nature
537 of the analysis, this should not be read as a confirmed sex-based performance difference. However,
538 the broader clinical AI literature has documented that diagnostic AI tools can show sex-stratified
539 performance gaps when training data are imbalanced or when symptom presentations differ by sex
540 (23). The conversational pathway in particular relies on natural-language interpretation of free-text
541 symptom descriptions, which is a setting in which subtle differences in expression style could affect
542 performance. Prospective evaluation with a balanced sex distribution will be needed before any sex-
543 related performance pattern can be confirmed or excluded.

544 4.5 Implications for hybrid AI triage design

545 The Introduction outlined an expectation that a structured model would be more consistent but less
546 flexible, while a conversational model would be more natural but more prone to inaccurate or unsafe
547 output. The present results support this expectation. MQ is safer and more consistent; Convo is more
548 flexible and more acceptable to users. Each pathway has a role, but neither pathway is sufficient on
549 its own.

550 For Ailix: GI Triage, the conversational pathway should not be deployed as a stand-alone
551 consultation route. Red flag detection should be implemented as a mandatory, rule-based, non-
552 skippable layer that runs before any AI-generated recommendation is returned to the user. The
553 conversational pathway can then be used for free-text symptom elaboration, patient education, and
554 engagement. Final output should include a clear escalation pathway to a human clinician, and the
555 safety layer should be active before any diagnosis or recommendation is shown. This design pattern
556 is consistent with the human-in-the-loop and safety-layered approaches that recent reviews of LLMs
557 in patient-facing care have recommended (18,19,21).

558 Specifically, the design priorities identified by the qualitative analysis can be translated into concrete
559 system requirements. The five most clinically consequential ones are: (i) mandatory and explicit red

flag screening before any diagnostic output, ideally implemented as a non-skippable rule-based layer with structured questions on alarm symptoms; (ii) systematic GI history-taking prompts covering symptom location, onset, duration, severity, stool pattern, pain character, triggers, and progression, which were inconsistently captured in both pathways; (iii) context-sensitive risk factor checks, including medication use, NSAID exposure, antibiotics, alcohol, pregnancy, and family history, which appeared as a recurring gap across 94 of 102 cases; (iv) clearer diagnostic prioritization with uncertainty framing, explicit disclaimers that the output is not a physician diagnosis, and a structured presentation of likely diagnosis, alternatives, red flag status, self-care advice, and when to seek care; and (v) handling of contradictory or vague user input, with the system clarifying inconsistencies before generating output. These five requirements are derived directly from the doctor reviewer themes in Table 8 and should serve as the development specification for the next iteration of Ailix.

4.6 Relevance for Indonesian digital health and GI consultation

This study provides early evidence for an Indonesian-language AI GI triage system. The relevance of such a system is in supporting initial access to consultation rather than replacing physician care, especially in the context of the high prevalence of GI disorders and the limited and uneven distribution of specialist care across Indonesia (1–4). Indonesia also has substantial variation in language use, digital literacy, and access to clinicians. Any AI triage system intended for use in this context must be designed to be clear, safe, and easy to understand, and must communicate the limits of its outputs explicitly. The World Health Organization's global strategy on digital health explicitly emphasizes the importance of fitting digital tools to local language and cultural contexts (10), and our findings are consistent with that direction.

4.7 Strengths and limitations

Strengths

This study is, to the best of our knowledge, the first validation of an Indonesian-language AI assistant for GI triage. It used a within-subject design, so that each participant served as their own comparison between the two pathways. Physician assessment was used as the gold standard. The study combined clinical performance metrics, structured reviewer scoring, user preference items, and qualitative feedback, which allowed safety, accuracy, and acceptance to be evaluated together rather than separately. The use of Gwet's AC1 also addressed a known limitation of Cohen's Kappa under imbalanced category distributions (13,14).

Limitations

Several limitations should be considered. First, the sample was drawn from active clients of a single digital health service rather than from the general Indonesian population, and the majority of participants were female and in the 20–30-year age group. Second, data collection was fully online, which may introduce selection bias related to digital literacy. Third, the physician gold standard was based on consultation transcripts only, without physical examination, laboratory tests, imaging, or endoscopy, and the clinical review was performed by a single physician. Fourth, several diagnosis groups had small case counts, and the subgroup analyses were exploratory and were not powered for definitive inference. Fifth, the absence of red flag detection in the Convo pathway should be interpreted as a system design and configuration issue rather than as an inherent limitation of conversational AI; the same underlying model could be implemented with a mandatory safety layer. Finally, the study did not assess actual downstream clinical outcomes, such as whether participants who received an escalation message subsequently sought medical care.

603 4.8 Future research

604 Future studies should evaluate a hybrid configuration that combines rule-based red flag detection
605 with conversational symptom exploration in the same workflow, and should test this configuration
606 against the current MQ and Convo pathways. External validation in broader populations, including
607 more balanced sex and age distributions and broader geographic and socioeconomic coverage, will
608 be important. Future work should also aim for a more balanced diagnostic case mix, include multiple
609 physician reviewers to allow inter-reviewer agreement analysis, and measure actual user referral
610 behavior after an AI consultation. Prospective safety evaluation, including monitoring of false-
611 negative red flag cases over time, would strengthen the evidence base for clinical deployment.
612 Finally, evaluation in local Indonesian language varieties and everyday colloquial speech would help
613 to identify gaps that may affect users outside the original development context.

614 4.9 Conclusion

615 In conclusion, this study provides early evidence that an Indonesian-language AI GI triage system can
616 support initial symptom assessment, but its safety depends on system design. MQ offered stronger
617 safety-net performance, while Convo provided better user acceptance and higher clinical accuracy
618 scores. The safest and most usable direction is a hybrid model that combines rule-based red flag
619 detection with conversational symptom exploration.

620 5 Conflict of Interest

621 AKT, SNF, and IA are employed by PT. Sirka Integra Medika, Jakarta, Indonesia, which developed
622 and operates the Ailix: GI Triage platform evaluated in this study. RA is the owner of both Sirka Ltd,
623 Singapore, the contracting entity that received the research grant from MENINBLOX Inc., and PT.
624 Sirka Integra Medika, the operating entity where the research was conducted, and is also a co-author
625 of this manuscript. ESK, HO, and JP are employed by MENINBLOX Inc., Gwangju, Republic of
626 Korea, the funder of this study, which has a collaborative commercial relationship with Sirka Ltd.
627 AKT, SNF, and IA report no other financial interests in either entity beyond their employment.

628 To mitigate these potential conflicts, ethical review of the study was conducted by an independent
629 research ethics committee not affiliated with either Sirka entity or with MENINBLOX Inc. (STIKes
630 Mayapada Research Ethics Committee, Jakarta, Indonesia; approval number 002/STIKES-
631 M/KEP/I/SLE/IV/2026). The analysis plan was specified in the study protocol prior to data
632 collection. The funder had no role in the design of the study, data collection, statistical analysis, or
633 interpretation of results beyond the contributions of the MENINBLOX-affiliated co-authors as listed
634 in the Author Contributions, and the decision to publish, including the reporting of safety findings
635 unfavorable to the conversational pathway, was made by the research team independently.

636 The authors declare no other financial or non-financial competing interests.

637 6 Author Contributions

638 AKT: Writing – original draft, Writing – review & editing, Conceptualization, Investigation,
639 Methodology. SNF: Writing – original draft, Writing – review & editing, Conceptualization, Data
640 curation, Formal analysis, Validation. RA: Writing – review & editing, Resources, Funding
641 Acquisition, Supervision. IA: Writing – review & editing, Project administration, Resources,
642 Supervision. HO: Writing – review & editing, Software, Conceptualization. ESK: Writing – review

643 & editing, Resources, Project Administration. JP: Writing – review & editing, Funding Acquisition,
644 Supervision.

645 **7 Funding**

646 This study was funded by MENINBLOX Inc., based in Gwangju, Republic of Korea, through a
647 research grant awarded under a memorandum of understanding with Sirka Science Ltd, Singapore.
648 The research was carried out by PT. Sirka Integra Medika, Jakarta, Indonesia, the operating affiliate
649 of Sirka Science Ltd. The funder had no role in study design, data collection, data analysis,
650 interpretation of results, or the decision to publish.

651 **8 Acknowledgments**

652 This research was supported by the Ministry of SMEs and Startups (MSS), Korea Institute for
653 Advancement of Technology (KIAT) through the Innovation Development (R&D) for Global
654 Regulation-Free Special Zone.

655 **9 Reference styles**

- 656 1. Kementerian Kesehatan RI. Profil Kesehatan Indonesia Tahun 2017. Jakarta: Kementerian
657 Kesehatan Republik Indonesia (2018).
- 658 2. Syam AF, Aulia C, Renaldi K, Simadibrata M, Abdullah M, Tedjasaputra TR. Revisi Konsensus
659 Nasional Penatalaksanaan Penyakit Refluks Gastroesofageal (GERD) di Indonesia. Jakarta:
660 Perhimpunan Gastroenterologi Indonesia (2016).
- 661 3. Kementerian Kesehatan RI. Data Registrasi dan Distribusi Tenaga Medis Indonesia. Jakarta:
662 Direktorat Jenderal Tenaga Kesehatan (2024). Available at: [https://eplanning-](https://eplanning-ditjennakes.kemkes.go.id)
663 [ditjennakes.kemkes.go.id](https://eplanning-ditjennakes.kemkes.go.id)
- 664 4. Konsil Kedokteran Indonesia. Data Registrasi Dokter Spesialis per 24 April 2024. Jakarta: KKI
665 (2024).
- 666 5. Ong JCL, Jin L, Elangovan K, Lim GYS, Lim DYZ, Sng GGR, et al. Large language model as
667 clinical decision support system augments medication safety in 16 clinical specialties. *Cell Rep Med.*
668 (2025) 6(10):102323. doi: 10.1016/j.xcrm.2025.102323
- 669 6. Gaber F, Shaik M, Allega F, Bilecz AJ, Busch F, Goon K, et al. Evaluating large language model
670 workflows in clinical decision support for triage and referral and diagnosis. *npj Digit Med.* (2025)
671 8:263. doi: 10.1038/s41746-025-01684-1
- 672 7. Gong EJ, Bang CS, Lee JJ, Park J, Kim E, Kim S, et al. Large language models in
673 gastroenterology: systematic review. *J Med Internet Res.* (2024) 26:e66648. doi: 10.2196/66648
- 674 8. Tao S, Guo Y, Chen W, Lin L, Wu H, Ke M, et al. Multiple large language models versus
675 experienced physicians in diagnosing challenging cases with gastrointestinal symptoms. *npj Digit*
676 *Med.* (2025) 8:75. doi: 10.1038/s41746-025-01462-z

677 9. Huo B, Boyle A, Marfo N, McKechnie T, Lee J, Mayol J, et al. Large language models for chatbot
678 health advice studies: a systematic review. *JAMA Netw Open*. (2025) 8(2):e2457879. doi:
679 10.1001/jamanetworkopen.2024.57879

680 10. World Health Organization. *Global Strategy on Digital Health 2020–2025*. Geneva: World
681 Health Organization (2021).

682 11. Brooke J. SUS: a “quick and dirty” usability scale. In: Jordan PW, Thomas B, Weerdmeester BA,
683 McClelland AL, editors. *Usability Evaluation in Industry*. London: Taylor & Francis (1996). p. 189–
684 194.

685 12. Zhou L, Bao J, Setiawan IMA, Saptono A, Parmanto B. The mHealth App Usability
686 Questionnaire (MAUQ): development and validation study. *JMIR Mhealth Uhealth*. (2019)
687 7(4):e11500. doi: 10.2196/11500

688 13. Gwet KL. Computing inter-rater reliability and its variance in the presence of high agreement. *Br*
689 *J Math Stat Psychol*. (2008) 61(1):29–48. doi: 10.1348/000711006X126600

690 14. Wongpakaran N, Wongpakaran T, Wedding D, Gwet KL. A comparison of Cohen's Kappa and
691 Gwet's AC1 when calculating inter-rater reliability coefficients: a study conducted with personality
692 disorder samples. *BMC Med Res Methodol*. (2013) 13:61. doi: 10.1186/1471-2288-13-61

693 15. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*.
694 (1977) 33(1):159–174. doi: 10.2307/2529310

695 16. Braun V, Clarke V. Using thematic analysis in psychology. *Qual Res Psychol*. (2006) 3(2):77–
696 101. doi: 10.1191/1478088706qp063oa

697 17. Council for International Organizations of Medical Sciences (CIOMS). *International Ethical*
698 *Guidelines for Health-related Research Involving Humans*. 4th ed. Geneva: CIOMS (2016).

699 18. Omar M, Sorin V, Collins JD, Reich D, Freeman R, Gavin N, et al. Large language models are
700 highly vulnerable to adversarial hallucination attacks in clinical decision support: a multi-model
701 assurance analysis. *medRxiv [Preprint]* (2025). doi: 10.1101/2025.03.18.25324184

702 19. Asgari E, Montaña-Brown N, Dubois M, et al. A framework to assess clinical safety and
703 hallucination rates of LLMs for medical text summarisation. *npj Digit Med*. (2025) 8:274. doi:
704 10.1038/s41746-025-01670-7

705 20. Nadarzynski T, Miles O, Cowie A, Ridge D. Acceptability of artificial intelligence (AI)-led
706 chatbot services in healthcare: a mixed-methods study. *Digit Health*. (2019) 5:2055207619871808.
707 doi: 10.1177/2055207619871808

708 21. Aggarwal A, Tam CC, Wu D, Li X, Qiao S. Artificial intelligence–based chatbots for promoting
709 health behavioral changes: systematic review. *J Med Internet Res*. (2023) 25:e40789. doi:
710 10.2196/40789

711 22. Sutaria S, Devakumar D, Mallinson P, Kinra S, Malak TT, Meczner A. Evaluating the use of red
712 flags by online symptom checkers. *BMC Health Serv Res*. (2025) 25:1263. doi: 10.1186/s12913-025-
713 13353-w

714 23. Cirillo D, Catuara-Solarz S, Morey C, Guney E, Subirats L, Mellino S, et al. Sex and gender
715 differences and biases in artificial intelligence for biomedicine and healthcare. npj Digit Med. (2020)
716 3:81. doi: 10.1038/s41746-020-0288-5

In review