

ACCURACY AND SAFETY OF THREE LARGE LANGUAGE MODELS ANSWERING TURKISH QUESTIONS ON ADULT IMMUNIZATION

Enes Zafer^{1✉}, Mehmet Yildiz²

Received on July 29, 2026

Presented by B. Petrunov, Member of BAS, on September 29, 2026

Abstract

Vaccine misinformation contributes to falling immunization coverage, and people increasingly bring vaccine questions to chatbots. We evaluated three large language models on 72 Turkish questions on adult immunization, in two arms: 40 patient-focused questions and 32 mapped to CDC/ACIP, WHO or Turkish Ministry of Health recommendations. The 216 responses were blind-coded and scored independently by two blinded specialists against a guideline-anchored answer key. Binary accuracy was 95.1% for Claude Opus 4.8, 93.8% for Gemini-3.1-Pro, and 86.1% for GPT-5.5. Patient-facing questions, including the classic vaccine myths, were answered well (93.8–96.2%), but accuracy on technical guideline content fell to 76.6% for GPT-5.5, which was also the most readable model and referred the reader to a clinician in only 53.1% of guideline-arm responses. Responses were more often incomplete than wrong.

Key words: adult immunization, artificial intelligence, large language models, patient education, vaccine misinformation

Introduction. Routine immunization coverage has stalled and in places declined; gains achieved before 2020 turned into losses that had not recovered by 2024 [1]. The World Health Organization identifies misinformation among the drivers, noting that people exposed to false claims often continue to rely on them even when accurate information is supplied [2]. Adult immunization is particularly

exposed, because uptake depends on individual decisions taken outside routine childhood schedules.

Increasingly those decisions are informed by chatbots. Large language models (LLMs) encode substantial clinical knowledge [3, 4] and are widely used for health questions [5, 6]. Their answers on vaccination have begun to be examined: ChatGPT responses to the WHO myth list were judged largely accurate and understandable [7]; a four-model comparison found correct-answer proportions of 0.86 to 0.98 but concluded that open-ended answers were too vague to be used alone [8]; evaluations of childhood [9] and human papillomavirus [10] vaccination reported accurate but incomplete content; and an English-Spanish comparison found similar accuracy but different readability [11].

Two gaps remain. Turkish-language performance on adult immunization has not been examined, and existing work treats vaccine questions as a single category, whereas what people actually ask ranges from a myth that needs rebutting to a technical protocol question that determines clinical action. These demand different things from a model. We therefore evaluated three leading LLMs in a blinded, dual-arm design that separates them.

Materials and methods. This was a cross-sectional, observer-blinded evaluation of LLM output, reported following the STROBE statement [12]. It involved no human participants, patient records or identifiable data and analysed only publicly available generated text; ethics committee approval was therefore not required.

The patient-focused arm comprised 40 questions across the eight subtopics listed in Table 1, including the classic myths that vaccines cause autism or infertility, that mRNA vaccines alter DNA, and that vaccines contain harmful mercury. The guideline arm comprised 32 questions, each derived from an explicit recommendation of CDC/ACIP, WHO or the Turkish Ministry of Health and mapped to that source item, covering indications, schedules, contraindications and precautions, intervals, co-administration, and risk-group regimens.

Table 1
Binary accuracy by subtopic and arm (two blinded specialists)

Subtopic	Guideline arm, %	Patient arm, %	<i>n</i>
COVID-19 vaccines	66.7	83.3	36
Misinformation and hesitancy	66.7	93.3	36
Influenza and respiratory vaccines	75.0	97.8	54
Adverse effects and safety	77.8	94.4	72
Special situations and risk groups	88.9	97.2	72
General rationale for immunization	94.4	100.0	42
Practical administration	97.2	100.0	66
Tetanus, zoster and other vaccines	100.0	96.7	54

n, number of expert-level ratings (two specialists per response), both arms combined

Guideline-anchored key points were prepared for all 72 questions, each naming the source and recording separately the points that guidance leaves to individualisation. For myth questions the key requires the mechanism, so that a bare denial was not scored as complete.

Each question was posed once, in Turkish, to GPT-5.5 (OpenAI), Gemini-3.1-Pro (Google) and Claude Opus 4.8 (Anthropic) on 5 June 2026. A fresh session was opened for every question with chat history, personalisation and memory disabled, and an identical Turkish instruction was used for all three providers. Responses were recorded verbatim, giving 216, then separated from their source, randomly coded (R001 to R216) and presented in mixed order.

Two blinded specialists, in internal medicine (8 years of experience) and infectious diseases (4 years), independently scored every response from separately randomised copies of the form: binary accuracy, four-point accuracy, five 1-to-5 Likert dimensions (accuracy, scientific-clinical quality, completeness, clarity, safety) and clinician referral. Safety was anchored on potential for harm, which here includes reinforcing a false belief or increasing hesitancy. Word count and Turkish readability were recorded automatically [13, 14].

Models were compared per question with the Friedman test and arms with the Mann–Whitney U test and Cohen r . Agreement was quantified with ICC(2,1) [15] and Fleiss kappa [16], with percent agreement and the prevalence-adjusted bias-adjusted kappa (PABAK) reported alongside because binary accuracy was unbalanced. A two-sided p below 0.05 was significant.

Results. Both specialists scored all 216 responses. In the four-point ratings 16.0% were completely accurate, 57.9% accurate but inadequate, 17.8% mixed and 8.3% completely inaccurate.

Binary accuracy was 95.1% for Claude Opus 4.8, 93.8% for Gemini-3.1-Pro and 86.1% for GPT-5.5. All three answered patient-focused questions well (93.8–96.2%) and separated on guideline content, where accuracy fell to 76.6% for GPT-5.5 against 90.6% and 93.8% for the other two (Table 2). On the Likert dimensions the two arms differed only on clarity ($p < 0.001$, $r = 0.195$) and safety ($p = 0.0001$, $r = 0.105$); accuracy ($p = 0.579$), scientific-clinical quality ($p = 0.847$) and completeness ($p = 0.591$) did not. The arm effect is therefore visible in the binary correct-or-not judgement rather than in graded quality.

T a b l e 2

Binary accuracy by model and arm (two blinded specialists)

Metric	GPT-5.5	Gemini-3.1-Pro	Claude Opus 4.8
Overall, %	86.1	93.8	95.1
Patient arm, %	93.8	96.2	96.2
Guideline arm, %	76.6	90.6	93.8
Words, mean	198.3	284.3	264.2

The weakest areas were all in the guideline arm: COVID-19 vaccines and misinformation (66.7% each), influenza and respiratory vaccines (75.0%) and adverse effects and safety (77.8%) (Table 1). The classic myths were rebutted correctly and with the underlying mechanism in the great majority of responses; errors concentrated instead in schedule and eligibility detail, in intervals and co-administration, and in the boundary between a true contraindication and a precaution.

Models differed on accuracy (Friedman $p = 0.0002$), scientific-clinical quality ($p = 0.0019$), completeness ($p = 0.0053$), and safety ($p = 0.0125$), but not on clarity ($p = 0.131$), where all three scored near the top of the scale (Table 3). GPT-5.5 produced the shortest answers and the most readable Turkish, yet had the lowest accuracy. Clinician referral diverged sharply: Claude Opus 4.8 signposted professional care in every response in both arms, Gemini-3.1-Pro in about 85%, and GPT-5.5 in every patient-arm response but in only 53.1% of guideline-arm responses (Fig. 1).

T a b l e 3

Five-dimension quality by model, mean (SD), with Friedman test across models

Dimension (1-5)	GPT-5.5	Gemini-3.1-Pro	Claude Opus 4.8	p
Accuracy	3.74 (1.11)	3.63 (0.70)	3.99 (0.76)	0.0002
Scientific-clinical quality	3.68 (1.19)	3.69 (0.87)	3.97 (0.84)	0.0019
Completeness	3.85 (1.17)	3.92 (0.76)	4.22 (0.77)	0.0053
Clarity	4.49 (1.15)	4.72 (0.48)	4.81 (0.41)	0.131
Safety	4.49 (1.17)	4.75 (0.79)	4.85 (0.74)	0.0125

SD, standard deviation

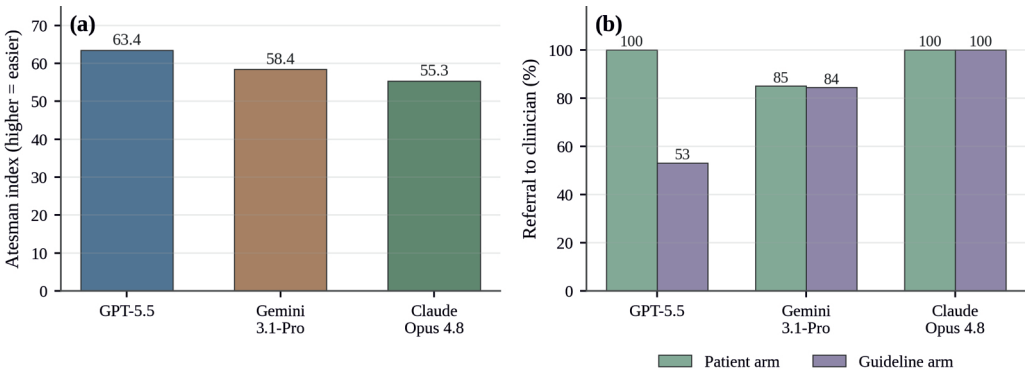


Fig. 1. (a) Turkish readability by model (Atesman index, higher indicates easier text). (b) Clinician-referral rate by arm as scored by the two specialists; GPT-5.5 refers in every patient-arm response but in only 53.1% of guideline-arm responses

Agreement between the two specialists was good to excellent: ICC(2,1) was 0.955 for safety, 0.737 for accuracy, 0.700 for clarity, 0.618 for scientific-clinical

quality, and 0.528 for completeness. For binary accuracy, percent agreement was 96.3%, Fleiss kappa 0.758, and PABAK 0.926.

Discussion. The shape of performance is more informative than the headline figure. Patient-facing questions were handled well by all three models, while the models separated on technical guideline content. These systems are therefore more dependable as a counter to misinformation than as a source of protocol detail. Asked whether a vaccine causes autism or infertility, or whether mRNA changes DNA, all three rejected the claim and supplied the mechanism, which is what distinguishes a persuasive rebuttal from a bare denial and is the element the WHO identifies as necessary when false beliefs persist despite correction [2]; this is consistent with earlier work on the WHO myth list [7]. Asked which pneumococcal schedule applies to a given risk group, or how long to wait between two vaccines, reliability drops.

The drop was largest for the model producing the shortest and most readable text, which also referred the reader to a clinician in only about half of its guideline-arm responses. Readability is therefore not a proxy for reliability, and the model that reads most easily may be the one most in need of a safety net.

Responses were more often incomplete than wrong: only 16.0% of ratings were completely accurate against 57.9% accurate but inadequate. In immunization this matters in a specific way, because an answer that correctly states that a vaccine is recommended for older adults but omits the interval, the risk-group qualification or the contraindication leaves the reader with a belief but not a decision. Completeness was also the dimension on which the two specialists agreed least, which suggests that what counts as sufficient here is itself contested.

Several limitations apply. Each response was collected once and LLMs vary between sessions. The answer key reflects guidance at a single date, and immunization recommendations are revised frequently, which plausibly contributes to the errors in schedule detail. Models were evaluated in a single-turn format, evaluation used two specialists, and multiple-dimension comparisons were exploratory. Findings apply to the reported model versions and access date.

In conclusion, these models are dependable at patient-facing myth correction but weaker on the technical protocol detail that determines who should receive which vaccine and when. Decisions on schedules, contraindications and risk-group eligibility should remain with a clinician.

REFERENCES

- [1] GBD 2023 Vaccine Coverage Collaborators (2025) Global, regional, and national trends in routine childhood vaccination coverage from 1980 to 2023 with forecasts to 2030, *Lancet*, **406**(10500), 235–260.

- [2] World Health Organization (2025) Message by the Director of the Department of Immunization, Vaccines and Biologicals at WHO – November/December 2025, Geneva, WHO.
- [3] SINGHAL K., S. AZIZI, T. TU et al. (2023) Large language models encode clinical knowledge, *Nature*, **620**(7972), 172–180.
- [4] SINGHAL K., T. TU, J. GOTTWEIS et al. (2025) Toward expert-level medical question answering with large language models, *Nat. Med.*, **31**(3), 943–950.
- [5] HALTAUFDERHEIDE J., R. RANISCH (2024) The ethics of ChatGPT in medicine and healthcare: a systematic review on large language models, *NPJ Digit. Med.*, **7**(1), 183.
- [6] AYRE J., E. CVEJIC, K. J. MCCAFFERY (2025) Who is asking ChatGPT health questions? Analysis of a nationally representative Australian community sample, *Med. J. Aust.*, **222**(4), 210–212.
- [7] DEIANA G., M. DETTORI, A. ARGHITTU et al. (2023) Artificial intelligence and public health: evaluating ChatGPT responses to vaccination myths and misconceptions, *Vaccines*, **11**(7), 1217.
- [8] SIEBIELEC J., F. RACIBORSKI (2026) Assessment of vaccine information accuracy across large language models, *Front. Public Health*, **14**, 1811462.
- [9] ERGUN OZDEL Z. G., E. CIFTCI, M. HACIMUSTAFAOGLU (2026) Are AI chatbots reliable sources for parents to ask about their concerns about childhood vaccination?, *Trends Pediatr.*, **7**(1), 63–70.
- [10] LAILY A., L. M. SCHWAB-REESE, M. DAVISH et al. (2026) Examining artificial intelligence chatbots' responses in providing human papillomavirus vaccine information for young adults, *JMIR Public Health Surveill.*, **12**, e79720.
- [11] JOSHI S., E. HA, A. AMAYA et al. (2024) Ensuring accuracy and equity in vaccination information from ChatGPT and CDC: mixed-methods cross-language evaluation, *JMIR Form. Res.*, **8**, e60939.
- [12] VON ELM E., D. G. ALTMAN, M. EGGER et al. (2007) The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies, *Lancet*, **370**(9596), 1453–1457.
- [13] ATESMAN E. (1997) Measuring readability in Turkish, *Dil Dergisi*, **58**, 71–74, (in Turkish).
- [14] BEZIRCI B., A. E. YILMAZ (2010) A software library for measuring text readability and a new readability metric for Turkish, *Dokuz Eylul Univ. Fac. of Eng., J. Sci. Eng.*, **12**(3), 49–62, (in Turkish).
- [15] KOO T. K., M. Y. LI (2016) A guideline of selecting and reporting intraclass correlation coefficients for reliability research, *J. Chiropr. Med.*, **15**(2), 155–163.
- [16] FLEISS J. L. (1971) Measuring nominal scale agreement among many raters, *Psychol. Bull.*, **76**(5), 378–382.

¹*Department of Internal Medicine, Sakarya Training and Research Hospital, Adnan Menderes Cad. Saglik Sok. No. 195, 54100 Adapazari, Sakarya, Türkiye*
e-mail: ezafer@sakarya.edu.tr (<https://orcid.org/0009-0001-4547-0255>)

²*Giresun Provincial Health Directorate, 28100 Giresun, Türkiye*
e-mail: drmehmetyildiz54@gmail.com (<https://orcid.org/0000-0003-0706-0906>)