

Supplementary material

Supplementary methods

Datasets

KorMedMCQA used the doctor development (164) and test (435) files of the sean0042 release at revision 79efd6f. MedQA used the four-option development (1,272) and test (1,273) files of the GBaker mirror at revision 17af935, checked item by item against the BigBio four-option release at revision 484a6c0. Development questions were used only for pilot and request checks.

Answering prompts

The exact Korean instruction was as follows.

`exam.question`에 대한 정답을 선택하세요. 문항에서 요구하는 조건에 따라 다섯 선택지 중 가장 적절한 하나를 고르세요.

The exact English instruction was as follows.

Select the correct answer to `exam.question`. Following the conditions requested in the question, choose the single most appropriate of the four options.

Jev and Clef received the question as structured state, the instruction as the choice instruction and the option mapping as choice criteria.

```
{"state":{"exam":{"question":"<QUESTION>"}},
  "questions":{"answer":{"type":"choice","instructions":"<INSTRUCTION>",
    "criteria":{"A":"<OPTION A>","B":"<OPTION B>","C":"<OPTION C>","D":"<OPTION D>","E":"<OPTION E>"}}}
```

OpenAI Decisions received the same state serialized as its input text and one choice question with the same instruction and the options as value and description pairs in published order.

```
{"model":"gpt-6-luna","input":{"\"exam\":{\"question\":\": \"<QUESTION>\"}}\",
  "questions":[{"type":"choice","name":"answer","instructions":"<INSTRUCTION>",
    "choices":[{"value":"A","description":"<OPTION A>"}, ...]]}
```

For generative models, the single user message was the instruction, two newline characters, a compact question object, two newline characters and a compact option object, using UTF-8 text without ASCII escaping. No additional system or developer prompt was supplied.

<INSTRUCTION>

```
{"exam":{"question":"<QUESTION>"}}
```

```
{"A":"<OPTION A>","B":"<OPTION B>","C":"<OPTION C>","D":"<OPTION D>","E":"<OPTION E>"}
```

English requests omit option E. Generative models returned a JSON object containing only a required answer field restricted to the item's option labels. One structured final answer was scored per question and model.

Probability validation and calibration

A decision-model response was valid when it contained exactly the offered options with probabilities between zero and one that summed to one within rounding, and when the selected option had the highest probability. OpenAI Decisions rounds probabilities to two decimals and Clef to four. Probabilities were renormalized to sum to one. OpenAI Decisions occasionally returned a refusal without probabilities, which was invalid. Expected calibration error used probability bins with edges at 0.50, 0.70, 0.80, 0.90 and 0.95.

Letter-probability baseline

GPT-6 Luna without reasoning received the same request as its reference configuration, a single structured answer, with up to 20 token log probabilities. At the answer token, the probabilities of the valid option letters were renormalized, and the probability of the generated letter served as its confidence. Responses with less than half of the probability on valid letters were invalid.

Referral, oracle and cost

For each model, valid answers were sorted by increasing selected probability and the first k were referred, where k was 20% of all questions (87 Korean and 255 English). Answers tied at the cut contributed their average result, which is the expected value under random order. Invalid answers were never referred and remained incorrect. Random referral selected k valid answers at random, and the oracle referred first the answers that GPT-6.1 Sol corrects, which gives the highest accuracy achievable with k referrals to Sol. Bootstrap intervals rebuilt each ranking within every resample.

Costs used published list prices [1,3,S1-S4], with input priced as uncached and reasoning or thinking tokens as output. The decision endpoints bill input tokens only. Retries and deployment overhead were excluded.

Answer-order experiment

Each question received every cyclic rotation of its option texts across the fixed letters, including an unchanged-order request, plus one identical repeat of that request, giving 8,975 requests per model. Clef orders option keys alphabetically before encoding [S5], but a rotation still moves each answer text to a different letter and position. Cyclic rotation changes position and letter together and does not cover every permutation. Questions were included when every request returned a valid answer, and referral used each model's threshold for its least confident 20% of main-collection answers.

Error analysis

Korean content labels were assigned by GPT-6.1 Sol, instructed to characterize each question without answering it, and 71 of 435 questions were labeled as depending on Korean law or administration. The author reviewed a blinded sample of 113 questions, comprising every question flagged by an earlier model labeling and 40 random others, and agreed with the label on 95.6% of them (Cohen's kappa 0.90). In every disagreement the author judged the question to depend on Korean law and the model did not.

MedQA physician ratings [16] were matched to all 1,273 test questions by exact question text, options and key, with three or four raters per question. A question was flagged when a strict majority reported missing important information, no acceptable official answer or several acceptable answers, and a physician majority endorsed the model's answer when more than half of the raters listed it as acceptable.

Twenty-four English questions were missed by at least three of the four most accurate reference models. A physician majority had flagged 45.8% of them, against 14.5% of other questions, and flagged questions were missed this way 4.76 times as often (95% CI 2.17 to 10.48, Katz interval). A majority reported no acceptable official answer for 29.2% of them, against 6.6% of other questions (risk ratio 5.41, 95% CI 2.30 to 12.71). The four models chose the same wrong option on 19 of the 24, and at least one physician had selected that option on 13.

Cautious errors had a selected probability at or below the model's twentieth percentile of main-collection probabilities. Rotation instability was the share of nonzero rotations whose answer differed from the unchanged-order response. For Jev the answer-order experiment was a separate collection, so the measure describes how stable each question was for Jev rather than the specific error. Bootstrap resamples drew valid answers with replacement.

Supplementary references

- S1. OpenAI. API pricing. OpenAI API documentation. Accessed 30 September 2026.
<https://developers.openai.com/api/docs/pricing>
- S2. Anthropic. Pricing. Claude documentation. Accessed 30 September 2026.
<https://platform.claude.com/docs/en/about-claude/pricing>
- S3. Google. Gemini Developer API pricing. Accessed 30 September 2026.
<https://ai.google.dev/gemini-api/docs/pricing>
- S4. Cloudflare. Clef. Workers AI model documentation. Accessed 6 October 2026.
<https://developers.cloudflare.com/workers-ai/models/clef/>
- S5. Cloudflare. Clef model card and source code. Hugging Face. Revision 2f3de3dd85f379784083b0814d997ab627200f0c. Accessed 6 October 2026.
<https://huggingface.co/Cloudflare/clef/tree/2f3de3dd85f379784083b0814d997ab627200f0c>
- S6. Republic of Korea. Act on the Prevention of Acquired Immunodeficiency Syndrome, Article 8-2, version effective 12 September 2020. National Law Information Center. Accessed 7 October 2026.
<https://www.law.go.kr/LSW/lsSideInfoP.do?docCls=jo&joBrNo=02&joNo=0008&lsiSeq=220905&urlMode=lsScJoRltInfoR>
- S7. Korea Disease Control and Prevention Agency, Rural Development Administration. Tick-borne infections: scrub typhus and SFTS. Standard lecture materials for field education [in Korean]. Accessed 7 October 2026.
<https://www.kdca.go.kr/bbs/kdca/49/246779/download.do>
- S8. World Health Organization. Ottawa Charter for Health Promotion. 1986.
<https://www.who.int/publications/i/item/WH-1987>

Supplementary tables

Supplementary Table 1. Model configurations and versions. The decision models accept no reasoning, temperature or output-length setting and return no generated text. No service returned a dated snapshot identifier. Reasoning and thinking tokens count toward the output allowances, and GPT-6.1 Sol accepts no setting below low. Claude Sonnet 5.5 used between_tools without tools, which prevents up-front thinking. Claude Opus 5.5 answers that reached the allowance were regenerated once with a larger allowance. Two Opus answers remained incomplete, and one Opus and one Sonnet answer were refused. No request contained retrieval, tools, another model's answer, a key or a requested explanation.

Model	Identifier	Version	Reasoning or thinking setting	Output token allowance	Collection date, 2026
Jev	jev-1.13.0	1.13.0	Not applicable	Not applicable	17 September
OpenAI Decisions	gpt-6-luna	No dated snapshot	Not applicable	Not applicable	7 October
Clef	clef	Public weights revision 2f3de3d [S5]	Not applicable	Not applicable	7 October
GPT-6 Luna letter probabilities	gpt-6-luna	No dated snapshot	none, with token log probabilities	128	6 October
GPT-6.1 Sol	gpt-6.1-sol	No dated snapshot	low	4096	30 September
GPT-6 Luna none	gpt-6-luna	No dated snapshot	none	128	30 September
GPT-6 Luna max	gpt-6-luna	No dated snapshot	max	128000	30 September
Claude Sonnet 5.5	claude-sonnet-5-5	No dated snapshot	between_tools, effort low	128	30 September
Claude Opus 5.5	claude-opus-5-5	No dated snapshot	adaptive, effort low	16000, increased to 64000 for incomplete outputs	30 September
Gemini 3.8 Flash	gemini-3.8-flash	No dated snapshot	LOW	8192	17 September
GPT-6.1 Sol, Korean content labels	gpt-6.1-sol	No dated snapshot	service default, resolved as medium	16000	6 October

Supplementary Table 2. Standalone counts. Invalid, refused or incomplete answers count as incorrect. Every invalid OpenAI Decisions answer was a refusal, and refusals recurred on identical repeat requests.

Cohort	Model	Correct / planned	Valid responses
Korean	Jev 1.13.0	378/435	435
	OpenAI Decisions	400/435	434
	Clef	363/435	435
	GPT-6 Luna letter probabilities	408/435	435
	GPT-6.1 Sol low	429/435	435
	Claude Opus 5.5 low	430/435	435
	Gemini 3.8 Flash low	428/435	435
	GPT-6 Luna max	427/435	435
	Claude Sonnet 5.5 low	412/435	435
	GPT-6 Luna none	405/435	435
English	Jev 1.13.0	1117/1273	1272
	OpenAI Decisions	1111/1273	1267
	Clef	1079/1273	1273
	GPT-6 Luna letter probabilities	1128/1273	1270
	GPT-6.1 Sol low	1232/1273	1273
	Claude Opus 5.5 low	1230/1273	1270
	Gemini 3.8 Flash low	1233/1273	1273
	GPT-6 Luna max	1217/1273	1273
	Claude Sonnet 5.5 low	1203/1273	1272
	GPT-6 Luna none	1138/1273	1273

Supplementary Table 3. Selected-option probability and the separately reported confidence field. Both measures were computed on the same valid test answers with the definitions of the main analysis. TypeSafe AI describes the field as derived from the probabilities [24]. Its derivation for OpenAI Decisions and the hosted Clef service is not documented, and the public Clef code returns the selected probability itself. ECE, expected calibration error.

Cohort	Model	Mean difference, confidence minus probability	Error AUROC, probability / confidence	ECE, probability / confidence	Errors caught at 20%, probability / confidence, %
Korean	Jev 1.13.0	-0.046	0.861 / 0.861	0.043 / 0.088	56.1 / 56.1
	OpenAI Decisions	-0.030	0.935 / 0.935	0.044 / 0.074	91.2 / 91.2
	Clef	-0.174	0.856 / 0.854	0.061 / 0.203	59.7 / 61.1
English	Jev 1.13.0	-0.039	0.899 / 0.901	0.009 / 0.031	71.6 / 72.0
	OpenAI Decisions	-0.033	0.903 / 0.903	0.028 / 0.020	71.4 / 71.4
	Clef	-0.145	0.864 / 0.864	0.020 / 0.135	62.4 / 63.4

Supplementary Table 4. Paired differences in error AUROC. Each bootstrap resample drew source questions with replacement and computed each model's AUROC on its own valid answers. Because the models make different errors, the differences describe how well each model ranks its own errors.

Cohort	Comparison	Difference (95% CI)
Korean	OpenAI Decisions minus GPT-6 Luna letter probabilities	0.091 (0.007 to 0.180)
	Jev minus OpenAI Decisions	-0.074 (-0.118 to -0.029)
	Jev minus Clef	0.005 (-0.045 to 0.058)
	OpenAI Decisions minus Clef	0.079 (0.033 to 0.126)
	Jev minus GPT-6 Luna letter probabilities	0.017 (-0.062 to 0.102)
English	OpenAI Decisions minus GPT-6 Luna letter probabilities	0.104 (0.062 to 0.147)
	Jev minus OpenAI Decisions	-0.004 (-0.031 to 0.022)
	Jev minus Clef	0.035 (0.007 to 0.065)
	OpenAI Decisions minus Clef	0.039 (0.011 to 0.069)
	Jev minus GPT-6 Luna letter probabilities	0.100 (0.054 to 0.146)

Supplementary Table 5. Errors caught and accuracy by share of answers referred. Shares are of all planned questions. The oracle refers errors first and, for accuracy, prioritizes errors that GPT-6.1 Sol corrects. Accuracy replaces the referred answers with GPT-6.1 Sol answers over the full planned denominator.

Cohort	Model	Referred, %	Errors caught, %	Oracle errors caught, %	Accuracy after referral, %	Random referral, %	Oracle, %
Korean	Jev 1.13.0	10	41.4	77.2	92.1	88.1	97.0
		20	56.1	100.0	94.0	89.2	99.1
		30	78.2	100.0	96.9	90.4	99.1
		50	96.5	100.0	98.4	92.8	99.1
	OpenAI Decisions	10	60.3	100.0	96.4	92.6	98.6
		20	91.2	100.0	98.2	93.2	98.6
		30	91.2	100.0	98.2	93.9	98.6
		50	100.0	100.0	98.4	95.2	98.6
	Clef	10	27.8	61.1	87.6	85.0	93.6
		20	59.7	100.0	92.9	86.5	99.5
		30	79.2	100.0	95.9	88.0	99.5
		50	91.7	100.0	97.9	91.1	99.5
	GPT-6 Luna letter probabilities	10	70.4	100.0	97.7	94.3	99.1
		20	76.6	100.0	98.0	94.8	99.1
		30	79.5	100.0	98.1	95.2	99.1
		50	85.4	100.0	98.2	96.2	99.1
English	Jev 1.13.0	10	44.8	81.9	92.3	88.6	97.6
		20	71.6	100.0	94.7	89.5	97.6
		30	87.5	100.0	95.9	90.4	97.6
		50	97.7	100.0	96.7	92.2	97.6
	OpenAI Decisions	10	46.6	81.4	92.0	88.2	97.3
		20	71.4	100.0	94.1	89.1	97.3
		30	88.6	100.0	95.5	90.0	97.3
		50	98.7	100.0	96.3	91.8	97.3
	Clef	10	39.2	65.5	90.3	86.0	94.7
		20	62.4	100.0	93.0	87.2	98.0
		30	79.4	100.0	95.0	88.4	98.0
		50	93.8	100.0	96.2	90.8	98.0
	GPT-6 Luna letter probabilities	10	55.6	89.4	93.7	89.4	97.5
		20	68.0	100.0	94.7	90.2	97.5
		30	72.0	100.0	94.9	91.0	97.5
		50	80.0	100.0	95.4	92.6	97.5

Supplementary Table 6. Answer and referral changes under option rotation and identical repetition. Each nonzero rotation and the identical repeat were compared with a contemporary unchanged-order response. Korean questions had four nonzero rotations and English questions three. Questions were included when all their requests gave valid answers, which excluded questions with an OpenAI Decisions refusal or a Jev output whose selected option lacked the maximum probability. Referral thresholds were 0.628, 0.736 and 0.661 in Korean and 0.770, 0.800 and 0.739 in English for Jev, OpenAI Decisions and Clef. The unchanged-order requests of OpenAI Decisions and Clef reproduced every valid main-collection answer and probability.

Cohort	Model	Outcome	Questions	Rotation, %	Repeat, %	Excess, percentage points (95% CI)
Korean	Jev 1.13.0	Answer	434	6.45	2.30	4.15 (2.25 to 6.11)
		Referral decision	434	8.81	1.84	6.97 (5.07 to 8.99)
	OpenAI Decisions	Answer	430	5.58	0.00	5.58 (3.90 to 7.33)
		Referral decision	430	10.23	0.00	10.23 (8.08 to 12.50)
	Clef	Answer	435	10.06	0.00	10.06 (7.93 to 12.18)
		Referral decision	435	16.44	0.00	16.44 (13.85 to 19.20)
English	Jev 1.13.0	Answer	1270	4.80	1.73	3.07 (2.10 to 4.07)
		Referral decision	1270	7.14	2.13	5.01 (3.81 to 6.22)
	OpenAI Decisions	Answer	1257	6.79	0.00	6.79 (5.67 to 7.98)
		Referral decision	1257	10.85	0.00	10.85 (9.49 to 12.22)
	Clef	Answer	1273	9.22	0.00	9.22 (7.93 to 10.50)
		Referral decision	1273	12.67	0.00	12.67 (11.21 to 14.19)

Supplementary Table 7. Cautious and confident errors in Korean. Definitions follow Table 4, with thresholds of 0.628, 0.736 and 0.661. The difference in rotation change between cautious and confident errors was 43.8 percentage points (95% CI 30.3 to 57.2) for Jev, 30.8 (9.2 to 51.9) for OpenAI Decisions and 5.1 (-11.7 to 21.6) for Clef. With three confident errors, the OpenAI Decisions percentages are imprecise.

Measure	Jev, cautious	Jev, confident	OpenAI Decisions, cautious	OpenAI Decisions, confident	Clef, cautious	Clef, confident
Errors, n	32	25	31	3	43	29
Answer changed under rotation, %	46.8	3.0	43.3	12.5	41.3	36.2
Corrected by GPT-6.1 Sol, %	96.9	88.0	90.3	33.3	97.7	96.6
Mean reference accuracy, %	88.5	85.3	77.4	44.4	91.9	92.0
GPT-6 Luna none chose the same wrong option, %	18.8	20.0	54.8	66.7	14.0	17.2
Law and policy content, %	46.9	32.0	61.3	0.0	39.5	24.1

Supplementary Table 8. Accuracy on Korean law and policy questions and on physician-flagged English questions. Law and policy questions (71 of 435) are those whose answer depends on Korean legal or administrative rules, as labeled by GPT-6.1 Sol without keys or model answers. Physician-flagged questions (192 of 1273) are those for which a majority of three or four US physicians reported missing information, no acceptable official answer or several acceptable answers [16]. Invalid answers count as incorrect.

Model	Korean law and policy, % correct	Other Korean, % correct	Law and policy share of errors, %	English physician-flagged, % correct	Other English, % correct	Flagged share of errors, %
Jev 1.13.0	67.6	90.7	40.4	65.6	91.7	42.3
OpenAI Decisions	73.2	95.6	54.3	70.3	90.3	35.2
Clef	66.2	86.8	33.3	67.7	87.8	32.0
GPT-6 Luna letter probabilities	78.9	96.7	55.6	76.0	90.8	31.7
GPT-6.1 Sol low	95.8	99.2	50.0	92.2	97.6	36.6
Claude Opus 5.5 low	100.0	98.6	0.0	93.2	97.2	30.2
Gemini 3.8 Flash low	93.0	99.5	71.4	91.7	97.8	40.0
GPT-6 Luna max	95.8	98.6	37.5	87.5	97.0	42.9
Claude Sonnet 5.5 low	87.3	96.2	39.1	85.9	96.0	38.6
GPT-6 Luna none	73.2	97.0	63.3	76.6	91.7	33.3

Supplementary Table 9. Selected errors that persisted across models. Cases illustrate recurring error sources among questions missed by at least two of the four most accurate reference models (GPT-6.1 Sol, Claude Opus 5.5, Gemini 3.8 Flash and GPT-6 Luna max). The nine configurations are the three decision models and the six reference models. Korean legal and public health comments cite the primary source where an exact provision or guidance document was identified, and English comments come from the published physician ratings [16]. Comments are descriptive and were not clinician adjudicated, and official keys were used in every analysis.

Question	Topic	Official key	Model answers	Comment
KorMedMCQA doctor-2022-1-18	Disclosure of HIV infection of a soldier	C, patient and unit commander	A, patient only, by Jev, OpenAI Decisions, Clef, Sol, Luna none, Gemini and Sonnet	The examination-era AIDS Prevention Act added notification of the institution head in communal military settings, an exception to the general confidentiality rule [S6]. The statute names the institution head and the key names the unit commander.
KorMedMCQA doctor-2023-1-17	Approval of emergency transfer when consent cannot be obtained	A, mayor, county head or district head	C, public health centre head, by Jev, OpenAI Decisions and Sol. B by Luna max	The key assigns this approval to the local government head. The applicable examination-era provision was not independently verified.
KorMedMCQA doctor-2024-1-74	Preventive focus of farmer education on severe fever with thrombocytopenia syndrome	A, personal hygiene	D, vector control, by all nine configurations	National field-education materials emphasize protective clothing, laundering and washing after outdoor work, which supports personal protection [S7]. The boundary between the two category labels is not settled.
KorMedMCQA doctor-2022-1-75	Ottawa Charter action area for assessing health effects of a dam project	D, building healthy public policy	A, creating supportive environments, by OpenAI Decisions, Sol, Opus and Luna max. The other configurations chose the key	The Charter places systematic assessment of the health impact of a changing environment under creating supportive environments [S8], whereas the key treats the dam project as a matter of public policy.
KorMedMCQA doctor-2024-3-70	Acute impaired consciousness after fever, vaccination, antihistamine and corticosteroid	A, delirium	C, functional neurological symptom disorder, by Jev, Luna none, Sol, Sonnet and Opus	Normal EEG, MRI and cerebrospinal fluid findings accompany fever, leukocytosis, raised C-reactive protein and recent sedating medication.
MedQA test-00160	Splenic zone linked to infection risk after splenectomy	C, zones 1 and 2	A by OpenAI Decisions, Luna none, Luna max, Gemini, Sonnet and Opus. B by Jev and Sol	The question refers to numbered zones on a slide that was not provided. All three physicians reported missing information.
MedQA test-00564	Sexual dysfunction during citalopram treatment	A, lower the dose	B, add bupropion, by all nine configurations	Two of three physicians selected bupropion and reported no acceptable official answer.
MedQA test-00112	Mechanism of retrosternal burning with eating	B, esophageal fibrosis	A, decreased lower esophageal sphincter tone, by Jev, Clef, Sol, Gemini and Luna max	Two of three physicians selected decreased sphincter tone and reported no acceptable official answer.
MedQA test-00874	Next step for cough two weeks after a febrile illness	D, supportive care	A, chest radiograph, by all configurations except Opus	Two of three physicians selected chest radiograph and reported no acceptable official answer.

Supplementary Table 10. MedQA results without physician-flagged questions. The main measures were repeated on the 1081 English questions without a physician flag, referring the least confident 20% of them. GPT-6.1 Sol alone answered 97.6% of these questions correctly.

Model	Accuracy, %	Error AUROC	Errors caught at 20%, %	Accuracy after referral, %	Random referral, %	Oracle, %
Jev 1.13.0	91.7	0.917	81.9	96.9	92.9	98.6
OpenAI Decisions	90.3	0.916	83.7	96.2	91.7	98.1
Clef	87.8	0.879	71.2	95.2	89.7	98.7
GPT-6 Luna letter probabilities	90.8	0.818	71.3	96.1	92.2	98.4