Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Numerical Consistency and Accuracy of Large Language Models in Risk Communication for Cardiovascular Prevention in Older Adults: A Cross-Sectional Multi-Model Paired Study

Yanhui Jiao
Chinese Academy of Medical Sciences, Peking Union Medical College
Jiaxin Li
Cangzhou Central Hospital Yanshan Branch
Xiuzhen Xing
Cangzhou Central Hospital Yanshan Branch
Lei Cao
Cangzhou Central Hospital Yanshan Branch
Qing Li
189317576760163.com

Chinese Academy of Medical Sciences, Peking Union Medical College

Article

Keywords: large language model (LLM), artificial intelligence (AI), cardiovascular disease, primary prevention, risk communication, number needed to treat, reproducibility

Posted Date: November 3rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11277331/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Numerical Consistency and Accuracy of Large Language Models in Risk Communication for Cardiovascular Prevention in Older Adults: A Cross-Sectional Multi-Model Paired Study

Yanhui Jiao [a],Jiaxin Li [b], Xiuzhen Xing [b], Lei Cao[b] ,Qing Li [a*]

[a] Department of Health Policy and Management, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, 100730 , China

[b] First Department of Cardiology, Cangzhou Central Hospital Yanshan Branch, Cangzhou, 061300, China

*Corresponding author: Qing Li

School of Health Policy and Management, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing ,100730, China

*Corresponding author(s). E-mail(s): 18931757676@163.com;

Contributing authors: s2024026022@student.pumc.edu.cn;

These authors contributed equally to this work.

Abstract

Background: Cardiovascular disease remains the leading cause of morbidity and mortality worldwide, with its burden concentrated in older populations. In primary prevention, decisions to initiate pharmacological interventions such as statins depend on a shared understanding of benefits and harms. Large language models (LLMs) are increasingly used to convey such quantitative information, yet existing evaluations have focused almost exclusively on accuracy.

Objective: To evaluate the run-to-run consistency, accuracy, and cross-language consistency of mainstream LLMs when converting absolute benefit across different risk presentations in cardiovascular prevention scenarios for older adults.

Methods: In this cross-sectional multi-model paired study, we constructed 80 synthetic scenarios of cardiovascular primary prevention. Each scenario was presented in five risk formats (relative risk reduction, pre- and post-treatment event rates, natural frequencies, risk ratio, and a direct-read positive control). Three LLMs (deepseek-v4-flash, doubao-seed-2.1-turbo, and kimi-k2.6) answered each item independently three times in both Chinese and English, yielding 7,200 responses collected through official application programming interfaces on 3–4 September 2026. Two independently implemented automated scorers (Python and R) graded all responses against a locked gold standard using precision-aware tolerance rules, and the two implementations agreed on every item. The primary outcome was the run-to-run complete agreement rate, defined as identical values for all four numerical outputs across three repeated runs of the same item group (positive controls excluded). Item groups answered correctly but inconsistently were

analysed for the source of disagreement, and accuracy was reported as a secondary outcome. Between-model differences were compared using mixed-effects logistic regression with scenario as a random intercept.

Results: Complete agreement rates were 88.4% for deepseek, 98.4% for doubao, and 98.6% for kimi. Adjusted marginal agreement rates were 93.7% (95% CI: 90.1%–96.0%), 99.4% (98.6%–99.8%), and 99.5% (98.8%–99.8%), respectively. deepseek was significantly less consistent than both comparators (OR = 0.086 and 0.072; both $P < 0.001$), whereas doubao and kimi did not differ ($P = 0.946$). Consistency was higher in the English arm than in the Chinese arm (OR ≈ 4.2, $P < 0.001$). Of the 89 correct-but-inconsistent item groups, 88 differed only in NNT or NNH by one (rounding instability), and one additionally showed an ARR tail difference of 0.01 percentage points. Malignant inconsistency (materially different values, fluctuating correctness, or runs with no extractable answer) accounted for 0.16% of item groups (3/1,920). Cross-language numerical agreement ranged from 92.2% to 99.1%, and McNemar tests detected no asymmetry in accuracy between languages ($P = 0.48$–1.00). The four-value fully correct rate ranged from 99.8% to 100%, indicating a ceiling effect.

Conclusion: On this structured risk-communication task, the numerical conversion accuracy of current mainstream LLMs has approached the ceiling, and between-model differences lie primarily in run-to-run consistency. Correct-but-inconsistent responses arose almost entirely from rounding instability that does not alter clinical conclusions, and Chinese and English outputs were highly concordant. Evaluations of medical AI should incorporate run-to-run consistency, particularly the malignant inconsistency rate, alongside accuracy.

Keywords large language model (LLM); artificial intelligence (AI); cardiovascular disease; primary prevention; risk communication; number needed to treat; reproducibility

Introduction

Cardiovascular disease (CVD) remains the leading cause of morbidity and mortality worldwide, and its burden is concentrated in older populations[1]. In primary prevention, the decision to initiate pharmacological interventions such as statins depends on a shared understanding of treatment benefits and harms between clinicians and patients[2]. Because patients perceive no immediate change after taking medication, adherence depends largely on how they answer the question "how much does this drug benefit me"[2]. How benefits and risks are expressed therefore directly shapes patients' comprehension, decision-making, and adherence, and ultimately determines the real-world effectiveness of primary prevention[1].

The same treatment evidence can be expressed in several equivalent but numerically distinct formats, including relative risk reduction (RRR), rate differences (ARR/ARI), natural frequencies, risk ratios (RR), and the number needed to treat (NNT) and number needed to harm (NNH)[3]. The choice of format substantially influences the perceptions and choices of both patients and physicians, and converting between formats requires numeracy, a skill that is not universal even among clinicians[4]. The same evidence can thus lead to different interpretations and decisions simply because of how it is presented[5]. This is a long-standing problem in risk communication, and one that cannot be eliminated by improving numeracy alone[6].

Digital health tools are increasingly applied to CVD prevention to support risk assessment, patient education, and shared decision-making[7]. Artificial intelligence (AI), and LLMs in particular, now approach professional-level performance on medical question answering[8] and are increasingly used directly for health information queries. In high-frequency, standardised consultation scenarios such as cardiovascular prevention in older adults, these models have de facto assumed the role of "translating the numbers in the evidence for patients", making the reliability of their numerical conversion a new research question[5].

Existing evaluations, however, have focused almost exclusively on accuracy and have overlooked at least two dimensions. The first is whether repeated queries of the same question yield identical values (run-to-run consistency): even under theoretically deterministic decoding settings, inference services may return different results[9]. The second is whether answers remain self-consistent when the question is asked in different languages (cross-language consistency). For use cases in which model outputs are passed directly to patients, a model that is "almost always right but says something different each time" is not equivalent in usability to one that is equally accurate and stable[10].

Against this background, we used cardiovascular prevention scenarios in older adults as a testbed to evaluate three mainstream LLMs on the accuracy, run-to-run consistency, and cross-language consistency of absolute-benefit numerical conversion across five risk presentation formats. Inconsistent responses were graded by clinical severity. This study provides empirical evidence for the application and evaluation of medical AI in risk communication.

Results

2.1 Run-to-run consistency

A total of 7,200 valid responses were obtained, and 1,920 item groups entered the primary analysis. The overall L1 agreement rate was 95.2% (1,827/1,920). By model, the rates were 88.4% for deepseek (Chinese arm 83.1%, English arm 93.8%), 98.4% for doubao (97.2% and 99.7%, respectively), and 98.6% for kimi (97.5% and 99.7%, respectively). The mixed-effects model (Table 2) yielded adjusted marginal agreement rates of 93.7% (95% CI: 90.1%–96.0%) for deepseek, 99.4% (98.6%–99.8%) for doubao, and 99.5% (98.8%–99.8%) for kimi.Adjusted inter-run consistency by model and language arm is shown in Figure 1.

Deepseek was significantly less consistent than both doubao (OR = 0.086, $P < 0.001$) and

kimi (OR = 0.072, P < 0.001), whereas doubao and kimi did not differ (OR = 0.837, P = 0.946). The main effect of language was significant, with higher consistency in the English arm (OR ≈ 4.2, P < 0.001). The main effect of presentation format was not significant (adjusted marginal agreement rates 98.3%–99.1%; pairwise P = 0.60–1.00). Consistency by risk-communication format and language arm is shown in Figure 2. In format-stratified analyses, deepseek remained significantly less consistent than the other two models across all four computable formats (all P < 0.015). The GEE sensitivity analysis yielded concordant directions of effect.

2.2 Characteristics of correct-but-inconsistent responses

We analysed the source of disagreement for each of the 93 inconsistent item groups (Table 3, Figure 3). Of these, 89 (95.7%) were correct across all three runs but numerically non-identical ("correct-but-inconsistent": 72 groups for deepseek, 9 for doubao, and 8 for kimi). In 88 of these groups, the disagreement was confined to NNT or NNH differing by one—arising from oscillation between rounding up and rounding to the nearest integer—and one additional group (doubao Chinese arm) also showed an ARR tail difference of 0.01 percentage points; none altered any clinical conclusion. Under the pre-defined rules, one further group (deepseek Chinese arm), in which one run was judged incorrect but all numerical differences remained within the clinical-equivalence range (ARR by 0.01 percentage points, NNT by one), was graded benign, bringing the benign total to 90 groups (96.8%). Malignant inconsistency occurred in three groups (0.16%, 3/1,920): two involved materially different values with fluctuating correctness (one in the deepseek Chinese arm, where one run miscalculated the absolute risk increase and NNH; one in the kimi English arm,

where one run reported an ARR twice that of the other two runs), and one involved two of three runs producing content entirely unrelated to the stem with only one run answered correctly (doubao English arm). In other words, the between-model gap in agreement arose almost entirely from benign, rounding-level variation rather than from instability in clinical judgement. The severity classification of the 93 inconsistent item groups is summarised in Figure 3.

2.3 Cross-language consistency

Pairing the Chinese and English responses of the same model, item, and run (1,200 pairs per model), the proportion with all four values identical was 99.1% for kimi, 95.2% for doubao, and 92.2% for deepseek (Table 4). McNemar tests detected no significant asymmetry in four-value correctness between languages (0–2 discordant pairs; P = 0.48–1.00). McNemar tests detected no significant asymmetry in four-language correctness between languages (0–2 discordant pairs; P = 0.48–1.00). The four-value fully correct rate ranged from 99.8% to 100% across models and language arms (Table 5), indicating a ceiling effect for accuracy on this task.

Tables

Table 1. Models and runtime configurations

Model	Model ID	Temperature	Thinking mode	Provider
deepseek	deepseek-v4-flash	0	Light thinking (platform default)	DeepSeek
doubao	doubao-seed-2-1-turbo-260628	0	Disabled	Volcano Engine
kimi	kimi-k2.6	1 (locked by platform)	Enabled (platform default)	Moonshot AI

Note: Between-model comparisons reflect differences between the deployed services as a whole rather than differences in decoding parameters alone.

Table 2. Run-to-run complete agreement rates and mixed-effects model results

Model / comparison	L1 agreement, % (Chinese / English)	Adjusted agreement, % (95% CI)	OR (P value)
deepseek	83.1 / 93.8	93.7 (90.1–96.0)	Reference
doubao	97.2 / 99.7	99.4 (98.6–99.8)	vs deepseek: 0.086 (<0.001)
kimi	97.5 / 99.7	99.5 (98.8–99.8)	vs deepseek: 0.072 (<0.001)
doubao vs kimi	—	—	0.837 (0.946)

Note: Mixed-effects logistic regression, L1 ~ model × presentation format + language + (1 | scenario), n = 1,920 item groups; OR < 1 indicates lower agreement for the former model.

Table 3. Severity grading of inconsistent item groups (number of groups)

Model	Inconsistent groups	Benign (within clinical-equivalence range)	Malignant
deepseek	74	73	1
doubao	10	9	1
kimi	9	8	1
Total	93	90	3

Table 4. Cross-language numerical agreement and McNemar tests

Model	Cross-language numerical agreement, %	Correct in both languages (pairs)	Correct in one language only (Chinese / English)	McNemar P value
deepseek	92.2	1,198	0 / 2	0.480
doubao	95.2	1,198	2 / 0	0.480
kimi	99.1	1,199	1 / 0	1.000

Table 5. Four-value fully correct rates by model and language (%)

Model	Language	Four-value fully correct	ARR correct	NNT correct	Positive control fully correct
deepseek	Chinese	99.8	100.0	99.9	100
deepseek	English	100.0	100.0	100.0	100
doubao	Chinese	100.0	100.0	100.0	100
doubao	English	99.8	99.8	99.8	100
kimi	Chinese	100.0	100.0	100.0	100
kimi	English	99.9	99.9	99.9	100

Note: Primary-outcome items only (F5 excluded); n = 960 per cell. ARI and NNH accuracy likewise ranged from 99.8% to 100% and are omitted.

Figures

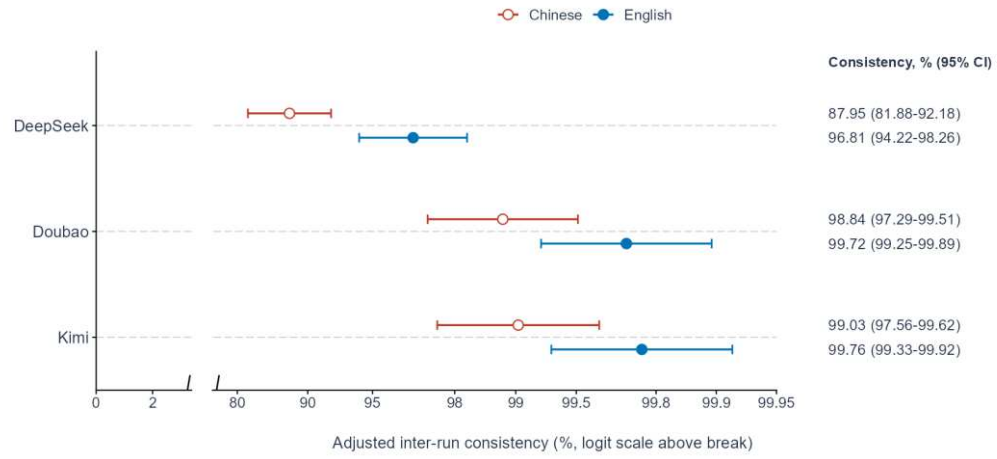


Figure 1. Adjusted inter-run consistency by model and language arm.
Estimates are marginal means (95% CI) from the mixed-effects logistic regression (L1 ~ model × presentation format + language + (1 | scenario), n = 1,920 item groups). The x-axis is on a logit scale above the axis break. Open circles, Chinese arm; filled circles, English arm. Numerical labels show adjusted consistency, % (95% CI).

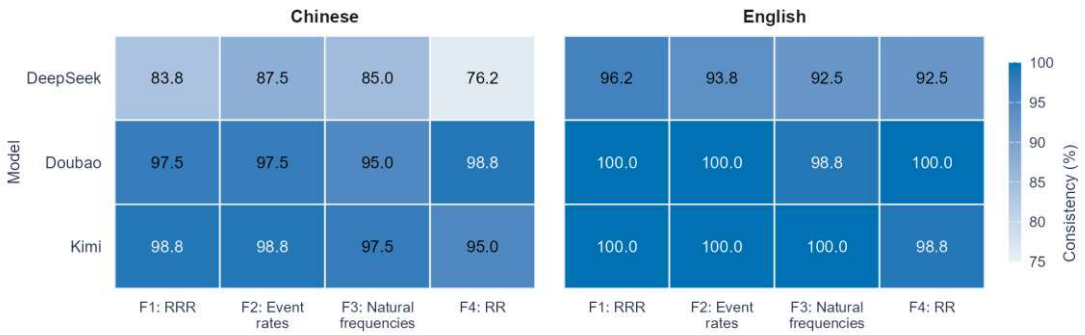


Figure 2. Run-to-run consistency by risk-communication format and language arm.

Each cell shows the percentage of item groups with complete numerical agreement across three runs; every cell is based on n = 80 scenarios. F1, relative risk reduction format; F2, pre- and post-treatment event rate format; F3, natural frequency format; F4, risk ratio format.

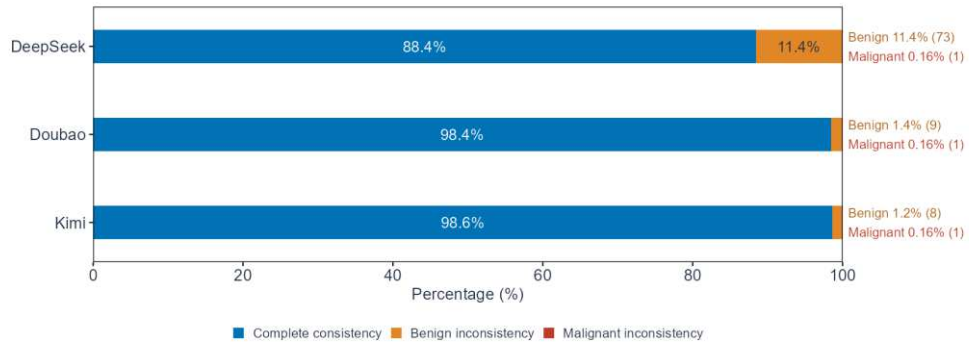


Figure 3. Severity classification of the 93 inconsistent item groups.

Complete consistency (all three runs numerically identical), benign inconsistency (correct in all runs with differences confined to the clinical-equivalence range: ARR/ARI ≤ 0.01 percentage points, NNT/NNH ≤ 1), and malignant inconsistency (materially different values, fluctuating correctness, or runs with no extractable answer). Numbers in parentheses are group counts; denominators are 640 item groups per model.

Methods

Abbreviations: AI, artificial intelligence; API, application programming interface; ARI, absolute risk increase; ARR, absolute risk reduction; CI, confidence interval; CVD, cardiovascular disease; GEE, generalised estimating equations; LLM, large language model; NNH, number needed to harm; NNT, number needed to treat; OR, odds ratio; RR, risk ratio; RRR, relative risk reduction.

3.1 Study design

This was an evaluation of numerical computation and risk-presentation performance by LLMs based on synthetic scenarios, covering 80 cardiovascular primary prevention scenarios in older adults and five risk presentation formats. The study was conducted between 3 and 4 September 2026 using a two-stage design. First, the research team generated 80 standardised numerical scenarios by script according to pre-specified parameter distributions (baseline

event risk 5%–35%, including both 5-year and 10-year time horizons; relative risk reduction 15%–35%; absolute increase in harm risk 0.2%–3.0%; the clinical plausibility of these ranges was confirmed by cardiovascular specialists). The gold standard, comprising the exact values and acceptable integer sets for absolute risk reduction (ARR), absolute risk increase (ARI), NNT, and NNH, was computed in parallel using pre-locked formulas, independently verified by a statistician, and fully locked before any model output was obtained. Each scenario was rendered into four arithmetically equivalent but differently worded presentations (relative risk reduction format, pre- and post-treatment event rate format, natural frequency format, and risk ratio format), plus one direct-read positive control item designed to distinguish "cannot compute" from "failed to read a given value". Each item was administered in two language arms (Chinese and English) with identical underlying values and equivalent semantics. Second, three LLMs (deepseek-v4-flash, doubao-seed-2.1-turbo-260628, and kimi-k2.6) answered each item through their official application programming interfaces (APIs) using a frozen prompt template. The prompt instructed the models to use only the values given in the item stem and to report the four quantities in fixed fields with a brief explanation. Each item was presented in an independent session to avoid contextual carry-over and was repeated independently three times, yielding 7,200 responses. All raw outputs (including model ID, temperature, timestamps, and token usage) were retained for audit. Responses were scored by an automated scorer that extracted the four values from the fixed fields, compared them against the gold standard under precision-aware rules, and assigned pre-specified error labels (E0–E5). The scorer was implemented independently in Python and R, and the two implementations agreed on every response. Refusals and parsing failures were retained in the denominator and never excluded. Model versions, temperatures, endpoints, and access dates were frozen, documented, and archived together with the raw data. The study workflow is illustrated in Additional file 1: Figure S1.

3.2 Scenario bank and numerical gold standard

Eighty standardised synthetic clinical scenarios were generated by script according to pre-specified parameter distributions, representing common decision contexts in cardiovascular primary prevention among older adults: age 65–80 years; varying sex and risk-factor combinations (hypertension, dyslipidaemia, smoking, obesity, family history of

premature CVD, and others); 5-year or 10-year baseline event risk of 5%–35%; treatment relative risk reduction of 15%–35%; and absolute increase in harm risk of 0.2%–3.0%. The clinical plausibility of each parameter range was confirmed item by item by cardiovascular specialists. An illustrative example: a 77-year-old woman with dyslipidaemia, hypertension, and a family history of premature CVD has a 10-year baseline risk of cardiovascular events of 26.9%; statin therapy confers a relative risk reduction of 16%, and the risk of harm events increases from 1.2% to 1.4%. The corresponding gold standard is an ARR of 4.3 percentage points, an NNT of 24 (exact value 23.26), an ARI of 0.2 percentage points, and an NNH of 500.

For each scenario, the gold standard was computed in parallel using pre-locked formulas: post-treatment risk p1 = p0 × (1 − relative risk reduction); ARR = p0 − p1 (percentage points); ARI = harm risk under treatment − harm risk under control (this subtraction had to be performed by the model and was not given in the stem); NNT = 1/ARR and NNH = 1/ARI, with integer answers accepted when rounded up or rounded to the nearest integer. The gold standard was independently verified by a statistician and fully locked before any model output was obtained. Each scenario was rendered into four arithmetically equivalent but differently worded presentations: (1) relative risk reduction format; (2) pre- and post-treatment event rate format; (3) natural frequency format (per 1,000 persons); and (4) risk ratio format (the stem gives RR to two decimal places only). One additional direct-read positive control item, in which the stem states ARR and ARI directly, was included to distinguish "cannot compute" from "failed to read a given value". Each item was administered in Chinese and English arms with identical underlying values and equivalent semantics.

These scenarios were deliberately constructed to cover clinical heterogeneity in baseline risk, magnitude of benefit, time horizon, and risk factors, but they do not encompass the full complexity encountered in routine clinical practice.

3.3 Models and prompting

All interactions were conducted through the official APIs of the three models: deepseek-v4-flash (DeepSeek, temperature 0), doubao-seed-2.1-turbo-260628 (Volcano

Engine, temperature 0, thinking mode disabled), and kimi-k2.6 (Moonshot AI, temperature locked at 1 by the platform). Data were collected on 3–4 September 2026 under a "platform-default or best-available configuration" principle. Model versions, temperatures, endpoints, and access dates were documented throughout and archived with the data (Table 1).

All items used a standardised, frozen prompt template to ensure methodological consistency and reproducibility. The model had access to all numerical information given in the stem (baseline risk, relative benefit or an equivalent presentation, harm risks in both groups, and the time horizon). The core instruction read: "Using only the values given in the stem, calculate and report the absolute risk reduction (ARR), absolute risk increase (ARI), number needed to treat (NNT), and number needed to harm (NNH) over 10 years (or the time horizon given in the stem); output the four values in the fixed fields with a brief explanation; do not use any information beyond the stem."

Each item was administered in an independent session, executed by a single script to ensure consistency, and repeated independently three times. Model outputs were evaluated exactly as produced, with no editing, modification, regeneration, or post-processing before scoring. All raw outputs (including model ID, temperature, and timestamps) were retained for audit.

3.4 Evaluation procedure

An automated scorer extracted the four values (ARR, ARI, NNT, NNH) from the fixed fields of each model output and compared them with the gold standard under pre-defined rules. Precision-aware adjudication was used: whether a response constituted the correct rounding of the gold standard was judged at the precision at which the model answered, so that legitimate rounding was not misclassified as error. For the information loss inherent in the risk ratio format (RR given to two decimal places only), both the true value and the value derived from the presented RR were accepted as correct (dual gold standard). Hazardous errors were flagged with a pre-defined label system: E0, no value given; E1, relative quantity reported as absolute; E2, unit misuse causing relative–absolute confusion; E3, reciprocal of a relative risk; E4, time-horizon drift; E5, direction error. The scorer was implemented independently in

Python and R, and the two implementations agreed on all 7,200 responses. Refusals and parsing failures were retained in the denominator and never excluded. Positive control items were analysed separately and did not contribute to the primary outcome.

3.5 Outcome measures and statistical analysis

The primary outcome was the run-to-run complete agreement rate (L1): within an item group (model × language × scenario × presentation format; n = 1,920 after exclusion of positive controls), all four numerical values were identical across the three repeated runs. Secondary outcomes included the four-value fully correct rate (all four values correct in a single response with no hazardous error); the definitional sensitivity metrics L2 (agreement in correctness labels) and L3 (clinically equivalent agreement: ARR/ARI differing by no more than 0.01 percentage points and NNT/NNH by no more than one); clinical severity grading of inconsistencies (benign: all three runs answered in the required format, with numerical differences within the clinical-equivalence range, including NNT/NNH rounding instability of ±1 and ARR/ARI tail differences of ±0.01 percentage points; malignant: materially different values, fluctuating correctness, or containing a run with no extractable answer—runs that failed to produce extractable values were counted as erroneous runs); and the cross-language numerical agreement rate together with paired comparison of Chinese–English correctness on the same item.

A mixed-effects logistic regression was fitted with L1 agreement of the item group as the dependent variable (fixed effects: model × presentation format + language; random intercept: scenario). Adjusted marginal agreement rates (95% CI) and pairwise odds ratios are reported.

Generalised estimating equations (exchangeable correlation) served as a sensitivity analysis. Cross-language correctness was compared with McNemar tests. Two-sided $\alpha = 0.05$. Analyses were performed in R 4.4 (packages lme4, emmeans, and geepack).

3.6 Ethics

Because the study involved only hypothetical, non-identifiable synthetic clinical scenarios and included no human participants or patient data, ethical approval was not required under institutional policy.

Discussion

This simulation study evaluated three LLMs on absolute-benefit numerical conversion across five risk presentation formats in cardiovascular prevention scenarios for older adults, with run-to-run consistency as the primary outcome. The assessed models stably produced absolute-benefit quantities consistent with clinical interpretation in structured numerical scenarios. Run-to-run inconsistencies were almost entirely confined to rounding-level variation that does not alter conclusions, and accuracy, as a secondary outcome, had approached the ceiling.

4.1 Between-model differences in run-to-run consistency

One of the most salient findings of this study is that run-to-run consistency differed substantially and robustly between models. The complete agreement rate of deepseek (88.4%) was significantly lower than that of doubao (98.4%) and kimi (98.6%), with adjusted odds of inconsistency 11–14 times higher, and this difference held across all four computable formats and both languages. Most notably, 11.6% of item groups still received different values from deepseek at temperature 0, a setting theoretically corresponding to deterministic decoding. This suggests that the widely held assumption that "temperature 0 guarantees reproducibility" does not hold for production inference services [11]. The point matters particularly in risk communication, where the trust of patients and physicians in a model rests on the expectation that the same question receives the same answer [12], and numerical fluctuation may erode

that trust [13]. Output variation under deterministic settings indicates that factors beyond decoding parameters, such as implementation details of the inference service, also affect reproducibility. This is consistent with prior warnings that non-determinism should not be neglected in the evaluation of LLMs [12]. It should be noted that we could not confirm the specific mechanism from external observation; the above interpretation is a reasonable inference only.

4.2 Correct-but-inconsistent responses are dominated by rounding-level variation

Item-level analysis of the 93 inconsistent groups showed that 96.8% were benign: 95.7% were correct-but-inconsistent (all three runs answered correctly), with disagreements almost entirely confined to NNT or NNH differing by one (only one group additionally showed an ARR tail difference of 0.01 percentage points), and one further group contained an incorrect run while its numerical differences remained within the clinical-equivalence range. Truly conclusion-altering, malignant inconsistency (materially different values, fluctuating correctness, or runs with no extractable answer) accounted for only 0.16%.This observation suggests that consistency and accuracy are two independent evaluation dimensions [14]: a model can be almost always correct yet give slightly different numbers each time; conversely, minor numerical fluctuation does not necessarily imply unstable clinical judgement. Clinically, an NNT differing by one would not change any primary prevention decision[15], but receiving two different NNTs for the same question asked three times may still be perceived as unreliable in real communication [16]. We therefore propose the malignant inconsistency rate, rather than an undifferentiated agreement rate, as the clinical yardstick for consistency evaluation.

4.3 Accuracy approaches the ceiling

These findings accord with the overall performance of LLMs on structured medical assessment tasks: models excel on tasks with explicit rules, accuracy approaches the ceiling, and accuracy metrics lose discriminative power on such tasks [17]. Setting run-to-run consistency as the primary outcome responds directly to this shift. Once "computing correctly" no longer differentiates models, "answering stably" becomes the next critical dimension of evaluation. During the pre-analysis phase, accuracy on the risk ratio format appeared lower; verification confirmed that this arose from information loss in the stem rather

than from model error, and the difference disappeared after the dual-gold-standard treatment. Evaluators must therefore distinguish "errors of the model" from "pseudo-errors created by measurement" [18].

4.4 Cross-language consistency

The cross-language results support the above conclusions from another angle. Numerical agreement between Chinese and English ranged from 92.2% to 99.1% across the three models, and McNemar tests detected no language asymmetry in correctness, indicating that numerical conversion ability on this task does not depend on the language of the query. For application scenarios serving predominantly Chinese-speaking older populations in China, this is an encouraging signal. Nevertheless, the significant main effect favouring the English arm (driven mainly by deepseek) indicates that independent evaluation for non-English settings remains necessary.

4.5 Implications for practice and evaluation

These findings are consistent with the broader view that AI should augment, not replace, clinical expertise [17]. Within this framework, the appropriate role of LLMs in risk communication is to perform the structured, verifiable intermediate step of "translating the numbers in the evidence", freeing clinicians for higher-level tasks such as individualised decision-making and patient communication, while model outputs remain under professional verification.

From an application and regulatory perspective, our results have two implications. First, in high-frequency, standardised consultation scenarios such as cardiovascular prevention in older adults, current mainstream models can already convey absolute-benefit values at a level suitable for assisted use, for example as tools for patient education and consultation support. Second, for product evaluation, "passing a single test" is insufficient evidence of reliability [19]: products with identical accuracy may differ vastly in repeatability [20]. Run-to-run consistency, particularly the malignant inconsistency rate, should be incorporated into the evaluation and acceptance framework for medical AI products. Our dual-platform cross-scoring pipeline and locked gold standard provide a replicable paradigm for such evaluation.

4.6 Limitations

Several limitations must be acknowledged. First, the study was based on hypothetical synthetic scenarios and cannot fully capture the ambiguity, missing information, and comorbidity complexity of real patient consultations. Although the scenarios covered clinical heterogeneity in baseline risk, magnitude of benefit, time horizon, and risk factors, they did not span the full range of routine practice, and the conclusions should not be extrapolated to real clinical conversations. Second, all evaluations were conducted under simulated conditions, and no conclusions can be drawn about clinical effectiveness or patient outcomes. Third, only three models and a single disease area were included; the models represent a snapshot as of September 2026, and version updates may alter the conclusions. The three configurations followed platform defaults or best-available settings rather than full alignment (kimi-k2.6 temperature locked at 1; thinking mode disabled for doubao), and between-model comparisons should therefore be read as comparisons of the deployed services as a whole. Fourth, two deviations from the original study protocol were documented in the deviation log: the primary outcome was changed from the four-value fully correct rate to run-to-run consistency, in view of the ceiling effect observed in both the pre-analysis and the formal scoring, and the planned human double-rating dimension of "presentational fidelity" was not included following the shift in analytic focus. Fifth, run-to-run consistency was based on only three repetitions; finer-grained characterisation of variability requires more runs. Finally, this study evaluated the accuracy and consistency of computation and conversion in fixed numerical scenarios, which is not equivalent to evaluating the overall clinical usability of the models. LLMs can still produce confident outputs that lack supporting evidence [21], and professional oversight is indispensable for any practical application of such tools.

4.7 Future directions

These limitations point to directions for future work: validating the findings in real consultation corpora and more complex scenarios; expanding coverage of models and disease areas; increasing the number of repetitions sufficiently to characterise output distributions; and directly comparing model outputs with conversions performed by clinical experts. To our knowledge, this is among the first studies to systematically evaluate the computational accuracy and run-to-run consistency of LLMs in absolute benefit and risk presentation against

a fixed numerical gold standard, addressing an evaluation gap in the current literature that emphasises accuracy over consistency [22].

Conclusion

In conclusion, on a structured absolute-benefit conversion task in cardiovascular prevention for older adults, the three mainstream LLMs achieved near-ceiling accuracy, and between-model differences lay primarily in run-to-run consistency rather than correctness. Run-to-run inconsistencies were almost entirely benign, rounding-level variation that does not alter clinical conclusions, whereas malignant inconsistency was rare (0.16%) but clinically relevant. Run-to-run consistency—particularly the malignant inconsistency rate assessed against clinical-equivalence criteria—should complement accuracy in the evaluation framework for medical AI, and model outputs should remain under professional verification before being used to communicate benefits and harms to patients.

Data availability： The datasets used and analyzed during the current study are available from the corresponding author upon reasonable request.

Supplementary Materials: Supplementary Material 1.

Funding: Not applicable.

Acknowledgments: Not required.

Declaration of competing interest: The authors declare no conflicts of interest.

Author contributions: Y.J.: Conceptualization, Methodology, Formal analysis, Data curation, Writing – Original Draft Preparation. J.L., X.X., and L.C.: Validation (review of the expert confirmation forms). Q.L.: Supervision, Project administration, Funding acquisition, Writing – Review & Editing (corresponding author). All authors read and approved the final manuscript.

References:

1. Perk J, De Backer G, Gohlke H, Graham I, Reiner Ž, Verschuren WMM, et al. European Guidelines on Cardiovascular Disease Prevention in Clinical Practice (Version 2012). International Journal of Behavioral Medicine. 2012 2012/12/01;19(4):403–88.
2. Brown MT, Bussell JK. Medication adherence: WHO cares? Mayo Clin Proc. 2011 Apr;86(4):304–14.
3. McDonagh TA, Metra M, Adamo M, Gardner RS, Baumbach A, Böhm M, et al. 2021 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure. Eur Heart J. 2021 Sep 21;42(36):3599–726.
4. Dowell D, Haegerich TM, Chou R. CDC Guideline for Prescribing Opioids for Chronic Pain--United States, 2016. Jama. 2016 Apr 19;315(15):1624–45.
5. Trevena LJ, Zikmund-Fisher BJ, Edwards A, Gaissmaier W, Galesic M, Han PKJ, et al. Presenting quantitative information about decision outcomes: a risk communication primer for patient decision aid developers. BMC Medical Informatics and Decision Making. 2013 2013/11/29;13(2):S7.

6. Nelson W, Reyna VF, Fagerlin A, Lipkus I, Peters E. Clinical implications of numeracy: theory and practice. Ann Behav Med. 2008 Jun;35(3):261–74.
7. Visseren FLJ, Mach F, Smulders YM, Carballo D, Koskinas KC, Bäck M, et al. 2021 ESC Guidelines on cardiovascular disease prevention in clinical practice: Developed by the Task Force for cardiovascular disease prevention in clinical practice with representatives of the European Society of Cardiology and 12 medical societies With the special contribution of the European Association of Preventive Cardiology (EAPC). European Heart Journal. 2021;42(34):3227–337.
8. Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, Amin M, et al. Toward expert-level medical question answering with large language models. Nature Medicine. 2025 2025/03/01;31(3):943–50.
9. Hager P, Jungmann F, Holland R, Bhagat K, Hubrecht I, Knauer M, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. Nature Medicine. 2024 2024/09/01;30(9):2613–22.
10. Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. npj Digital Medicine. 2023 2023/07/06;6(1):120.
11. Ouyang S, Zhang J, Harman M, Wang M. An Empirical Study of the Non-Determinism of ChatGPT in Code Generation. ACM Transactions on Software Engineering and Methodology. 2025 01/22;34:1–28.
12. Araujo T, Helberger N, Kruikemeier S, de Vreese CH. In AI we trust? Perceptions about automated decision-making by artificial intelligence. AI & SOCIETY. 2020 2020/09/01;35(3):611–23.
13. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Medicine. 2019 2019/10/29;17(1):195.
14. Zhao WX, Zhou K, Li J, Tang T, Dong Z, Hou Y, et al. A Survey of Large Language Models. Frontiers of Computer Science. 2026 2026/05/09;20(12):2012627.
15. Cook RJ, Sackett DL. The Number Needed to Treat: A Clinically Useful Measure of Treatment Effect. BMJ. 1995 1995/02/18;310(6977):452–4.
16. Díaz-Rodríguez N, Del Ser J, Coeckelbergh M, Prado M, Herrera-Viedma E, Herrera F. Connecting the Dots in Trustworthy Artificial Intelligence: From AI Principles, Ethics, and Key Requirements to Responsible AI Systems and Regulation2023.
17. Dwivedi YK, Kshetri N, Hughes L, Slade EL, Jeyaraj A, Kar AK, et al. Opinion Paper: "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. International Journal of Information Management. 2023 2023/08/01/;71:102642.
18. Park SH, Han K. Methodologic Guide for Evaluating Clinical Performance and Effect of Artificial Intelligence Technology for Medical Diagnosis and Prediction. Radiology. 2018 2018/03/01;286(3):800–9.
19. Panayides AS, Amini A, Filipovic ND, Sharma A, Tsaftaris SA, Young A, et al. AI in Medical Imaging Informatics: Current Challenges and Future Directions. IEEE Journal of Biomedical and Health Informatics. 2020;24(7):1837–57.

20. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics. 2020 2020/01/02;21(1):6.
21. Zhang Y, Li Y, Cui L, Cai D, Liu L, Fu T, et al. Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models. Computational Linguistics. 2025 12/01;51:1373–418.
22. Raiaan MAK, Mukta MSH, Fatema K, Fahad NM, Sakib S, Mim MMJ, et al. A Review on Large Language Models: Architectures, Applications, Taxonomies, Open Issues and Challenges. IEEE Access. 2024;12:26839–74.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementarymaterial.docx