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.

A simulation-based framework for sizing paired benchmarks in the evaluation of clinical artificial intelligence, applied to 168 expert-level dyslipidaemia items

Mete Ucdal

meteucdal@hacettepe.edu.tr

Etimesgut Asker Hastanesi

Karya Yurtsever

Hacettepe University Hospital

Pınar Yıldız

Eskişehir Osmangazi University

Ayşen Akalın

Eskişehir Osmangazi University

Kadir Uğur Mert

Eskişehir Osmangazi University

Gülay Sain Güven

Hacettepe University Hospital

Research Article

Keywords: Sample size, Research design, Artificial intelligence, Benchmarking, Decision support systems, clinical, Statistical power, Hyperlipidemias

Posted Date: September 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10682358/v1

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

Additional Declarations: No competing interests reported.

Abstract

Background.

Comparative evaluations of clinical artificial intelligence increasingly use a paired design in which competing systems answer a shared bank of expert-authored items and are compared by McNemar's test. Sample-size methodology for that design is long established[1, 2] but is almost never applied here, and two features of contemporary benchmarks escape the usual normal-approximation formulae. The exact test is discrete, imposing a minimum discordance count below which no bank size can reach significance, and a bank on which strong systems frequently score perfectly cannot rank them at all. We derived a sizing framework that handles both, applied it prospectively, and report what it did and did not make estimable.

Methods.

Required bank size was obtained by Monte Carlo simulation over the trinomial distribution of paired outcomes, with 40,000 replications per candidate size, applying the exact McNemar test to each and searching ascending sizes for 80% power. Ceiling probability was accuracy raised to the power of bank size. The framework then built the instrument reported here. We authored 214 single-best-answer dyslipidaemia items, analysed the 168 carrying unanimous agreement across three independent reference keys, retained the 46 non-unanimous items as a pre-specified abstention probe, and stratified all items before scoring by three pre-specified quality criteria. Responders were three strictly nested configurations of a neurosymbolic multi-agent system, two comparator large language models, and six practising physicians.

Results.

The framework predicted correctly in both directions. The total architectural contrast was resolved at 8.9 percentage points and 97.7% realised power, whereas the orchestration-only contrast remained unresolved at 2.4 points and 13.3% power, requiring approximately 1051 items. Ceiling probability for a 97%-accurate system fell from 48.1% at 24 items to 0.6% at 168. Of thirteen reported contrasts, five were resolved, five underpowered, two unresolvable at feasible scale, and one is reported as a hypothesis.

Conclusions.

Bank size, not the system under test, determined which conclusions were reachable. Comparative evaluations of clinical artificial intelligence should justify bank size a priori and report realised power and required size for every contrast, including the null ones.

Trial registration.

Not applicable.

Background

Comparative evaluations of clinical artificial intelligence increasingly adopt a paired design. Competing systems, and often a panel of clinicians, answer the same bank of expert-authored items, and performance is compared with McNemar's test on the discordant pairs. Sample-size methodology for that design has existed for decades. Closed-form approximations for the power of the McNemar test are available and have been compared against exact power in detail[1, 2], and the choice among them is well characterised. What is missing is not the statistical theory but its application, since evaluations in this field are customarily built to whatever length expert respondents can complete, typically a few dozen items, and almost none states the item count its own contrasts would require.

Two features of contemporary artificial-intelligence benchmarks make that omission more consequential than it would be elsewhere, and neither is handled by the standard normal-approximation formulae. The first is discreteness. The exact conditional test evaluates the discordance count against a binomial distribution with success probability one half, so with no discordance in the opposing direction a two-sided P below 0.05 requires at least six one-directional discordant items. No increase in bank size relaxes that floor; bank size only changes the probability of reaching it. The second is a ceiling constraint that is specific to a literature in which the systems under comparison are all strong. A system with true accuracy p attains a perfect score with probability p raised to the power n, so a short bank on which a capable system frequently scores perfectly cannot rank systems at all, whatever the analysis. Both properties bind hardest exactly where this field now operates, since expert-level medical benchmarks increasingly report near-ceiling performance for frontier models[3, 4, 5].

The substantive claims at stake make the omission harder to excuse. Agentic and neurosymbolic architectures are advocated on the premise that decomposing a task across cooperating roles and grounding outputs in explicit checks improves reliability[6, 7], while other work reports that multi-agent pipelines can compound errors across calls[8, 9], and a distinctive rationale for such systems is safety rather than raw accuracy, in the form of rule checking and principled abstention on under-determined cases[10, 11]. Each of those claims is a component-level claim, and component-level contrasts are smaller than whole-system contrasts, so they are precisely the contrasts a short bank cannot resolve. A component that produces no measurable change on an instrument known in advance to lack the resolution to detect it has not been shown to be inert. It has not been tested, and the current literature does not distinguish the two situations.

We are able to demonstrate the point rather than assert it, because we made the error ourselves and can report both halves of the comparison. In a preceding, unpublished evaluation of the system described below, on a 24-item bank, the system and its own neural backbone each scored 24 of 24, every paired

comparison returned an exact McNemar P value of at least 0.25, the pre-specified mechanism-isolating ablation was not estimable because there were no discordant pairs to analyse, the symbolic verifier altered no output, and the abstention controller never activated. None of those were findings about the architecture. All were findings about the instrument, and all were predictable in advance from the two constraints above.

The objective of this study was therefore methodological. We set out to derive a framework that fixes the item count a paired evaluation requires before any item is authored, to apply it prospectively to build an adequately sized instrument, and to report honestly what the resulting instrument did and did not make estimable. Three substantive hypotheses were pre-specified for the application study. H1 held that the neurosymbolic multi-agent system for lipidology (NS-MML) is at least as accurate as frontier large language models and physicians. H2 held that symbolic verification, rather than multi-agent orchestration or model scale alone, accounts for any accuracy advantage, and was to be tested by a three-arm ablation. H3 held that the architectural benefit extends beyond accuracy on solvable items to appropriate abstention and avoidance of unsafe recommendations, and was to be tested on an item set deliberately enriched with under-determined problems. A framework of this kind earns credibility only if it predicts correctly in both directions, so we report the contrasts it anticipated would fail alongside those it anticipated would succeed. A second requirement follows from the first, since the resolution of a benchmark depends on the informativeness of its items, so item quality was treated as a measured, pre-specified stratification variable rather than as an assumption.

Dyslipidaemia was chosen as the application domain because contemporary lipid guidelines encode risk stratification, treatment targets, and therapeutic sequencing algorithmically[12, 13], which makes the correct answer often derivable from an explicit rule set and therefore makes the domain a fair test bed for an architecture whose stated mechanism is explicit rule checking. We had also previously reported that general-purpose large language models exceeded physicians on standardised dyslipidaemia cases in this domain[14], which supplies a comparison point.

Methods

Design and reporting

This was a comparative, cross-sectional benchmarking study with blinded scoring, designed and reported per the TRIPOD + AI statement,[15] the TRIPOD-LLM extension for studies using large language models,[16] and informed by PROBAST + AI.[17] Completed checklists are provided as supplementary files, with an element-by-element mapping in Additional file 1, Note S2. All scenarios were fictional and no patient-identifiable data were used.

The benchmark-sizing framework

Because the framework determined the design, it is described in full before the instrument it produced.

Notation and structure

For each item, the paired outcome of a system S and a comparator C falls into one of four cells. Let p_b denote the probability that S is correct and C is incorrect, p_c the probability that S is incorrect and C is correct, and p_a and p_d the probabilities of concordant correct and concordant incorrect responses. The exact McNemar test conditions on the total number of discordant pairs m equal to b plus c, and evaluates b against a binomial distribution with parameters m and one half. Power therefore depends only on p_b, p_c, and the bank size n. Marginal accuracy affects power only insofar as it constrains those two quantities. This is why the framework is expressed in discordance terms rather than in accuracy terms, and why two studies with identical accuracy differences can have radically different power.

Constraint 1, the discreteness floor

Because the conditional reference distribution is binomial with success probability one half, the two-sided exact P value with c equal to zero is 2 raised to the power one minus b. Significance at an alpha of 0.05 therefore requires b of at least 6, giving P equal to 0.0312, since b equal to 5 yields P equal to 0.0625. No increase in bank size relaxes this floor. Bank size only raises the probability of accumulating six or more one-directional discordant items. An evaluation whose expected discordance count is below six is incapable of significance regardless of the size of the observed effect.

Constraint 2, the power requirement

For a candidate bank size n, power was estimated by drawing 40,000 replications from the trinomial distribution with probabilities p_b, p_c, and one minus p_b minus p_c, applying the exact McNemar test to each replication, and recording the proportion of replications with two-sided P below 0.05. The required size is the smallest n at which estimated power reaches 80%. Simulation rather than a closed-form approximation was used precisely because the discreteness described under Constraint 1 makes normal-approximation formulae optimistic at the sizes relevant to clinician-led work.

Constraint 3, the ceiling constraint

A system with true accuracy p attains a perfect score with probability p raised to the power n. Because a perfect score cannot be exceeded, a bank on which a strong system frequently attains a perfect score cannot rank systems, no matter how many discordant pairs the analysis would otherwise permit. Requiring the ceiling probability to fall below 5% for a system of 97% true accuracy implies at least 99 items, and requiring it to fall below 1% implies at least 152.

Prospective application to the present study

The pre-specified target was a net discordance gain of 7.5 percentage points, operationalised as p_b equal to 0.10 and p_c equal to 0.025. Simulation over ascending bank sizes returned 184 items for 80% power. The adjacent, more optimistic scenario of p_b equal to 0.10 and p_c equal to 0.02 returned 152 items, and the ceiling constraint required at least 152 items for the 1% criterion at 97% true accuracy. The operative recruitment floor was therefore the larger of the 152 items demanded by the moderate-gain scenario and the 152 items demanded by the ceiling constraint, and we record plainly that the more demanding 7.5-point target of 184 items was not reached. Allowing for exclusions at the reference-key stage, where our previous experience indicated a loss of approximately one fifth of drafted items, we set a recruitment target of at least 160 analysed items and drafted 214. The realised yield was 168 analysed items and 46 probe items, which is why the verification increment sits at exactly the 80% power boundary rather than above it.

The framework was also used prospectively to decide which contrasts were worth pre-specifying as primary. The orchestration-only contrast was expected to produce a small increment, and the framework indicated that resolving a 2-point net gain would require 1251 items. We therefore pre-specified it as a secondary contrast with an explicit acknowledgement that a null result would be uninformative, and we report its realised power rather than its P value alone. This is the practical value of the framework. It identifies unresolvable contrasts before resources are committed rather than after.

Two further sizing quantities were computed after the data were available and are labelled as such throughout, namely the number of items that would have been required to power each observed contrast and the number required to power the interaction between quality stratum and architectural benefit.

Benchmark instrument

We authored 214 single-best-answer items across five reasoning domains, namely drug-drug interactions and pharmacokinetics; statin-associated and immune-mediated myopathy; familial, monogenic, and severe dyslipidaemias; special populations, covering pregnancy, lactation, decompensated cirrhosis, advanced and dialysis-dependent kidney disease, HIV on antiretroviral therapy, and active malignancy; and novel therapeutics and biomarkers. Each item presented a full clinical vignette with five options and required integration of mechanism, trial evidence, and patient context rather than a single guideline threshold. Domains were balanced across the bank, with 32 to 34 analysed items each, and item topics are listed in Additional file 1, Table S1.

Forty-six draft items were excluded from the accuracy analysis because a single consensus reference answer could not be established among the source keys. Rather than being discarded, these items were retained as a pre-specified abstention probe, on the reasoning that a system whose stated contribution is principled abstention should decline an under-determined question whereas an unverified backbone should not. They are described in Additional file 1, Table S2. Because the items are newly authored, verbatim training-data contamination is unlikely, but contamination is nonetheless treated as a residual limitation, and each model’s version identifier and access date are recorded.

Item quality stratification

Item quality was measured before any scoring and used to stratify the bank. Every item was independently rated by three of the six physician panellists under a balanced incomplete block design in which each physician rated 84 of the 168 items, so that both strata were covered by three-rater assessment throughout. Raters scored methodological quality on a five-point scale and separately flagged perceived ambiguity or the existence of a defensible alternative answer, using the instrument and allocation reproduced in Additional file 1, Table S4. Rating preceded and was kept separate from answering, and raters did not see any system output.

Three criteria, all specified before scoring, defined the high-quality stratum, denoted Tier A. These were unanimous quality endorsement, meaning every assigned rater scored the item at least 4 of 5 and none flagged ambiguity; complete evidence provenance, meaning the reference answer was traceable to at least two independent guideline or trial sources; and unanimous first-pass agreement of the three reference keys with no key author requesting clarification. Items retained for analysis but failing at least one criterion formed the other-quality stratum, denoted Tier B. We deliberately used a conjunction of three criteria rather than the rating alone, so that stratum membership was not a restatement of the rating variable, and 23 Tier B items in fact attracted unanimous rater endorsement and were assigned to Tier B on provenance or first-pass key grounds.

We did not impose a numerical discrimination threshold as an inclusion criterion. Classical discrimination indices are bounded by the ability range of the reference panel, and on a bank where specialist respondents average above 80% correct such a threshold would exclude informative items for reasons unrelated to their quality. Realised discrimination is instead reported descriptively for both strata.

Reference standard

Three reference keys were derived independently. Key 1 was authored by the lead investigator from guideline sources during item construction and was locked on 20 June 2026. Key 2 was generated by Claude Opus 4.8 in a single deterministic run on 5 July 2026. Key 3 was produced on 8 July 2026 by a board-certified clinical pharmacologist not involved in item writing and not a member of the physician comparator panel. Only items on which all three keys agreed on the single best answer were retained, yielding the 168 analysed items, and the key set was locked on 9 July 2026, before any scoring and before the comparator evaluation runs. Unanimity was the retention rule, which removes adjudicator discretion, and disagreement was an exclusion criterion rather than something to be resolved.

This design protects the reference standard, which rests on two human-derived keys as well as the model-derived one. It does not protect the reported accuracy of Claude Opus 4.8. Because unanimity was the retention rule, every retained item is by construction an item on which that model’s answer matched the reference standard, and its score of 168 of 168 is guaranteed rather than observed. Claude

Opus 4.8 is accordingly reported for completeness and excluded from all inferential comparisons. The three system arms and Gemini 2.5 Pro contributed no keys, and their scores are unaffected.

The neurosymbolic multi-agent system

The NS-MML is a graph-orchestrated, multilayer neurosymbolic clinical-reasoning pipeline implemented in Python with LangGraph version 0.2.76 (Fig. 1). Its neural backbone was OpenAI GPT-5.5, model identifier gpt-5.5, accessed on 9 July 2026. The orchestration graph contained dedicated nodes for case normalisation, emergency triage, diagnostic reasoning, guideline grounding, management planning, contraindication and safety screening, symbolic verification, consensus synthesis, abstention control, final-answer generation, and audit logging. Each neural node emitted structured, auditable output rather than free-text chain-of-thought, and no chain-of-thought was stored or evaluated.

The symbolic component comprised a locked clinical rule base and lightweight ontology, namely NS-MML Rule Base v1.0 and NS-MML Clinical Ontology v1.0, encoding diagnostic invariants, mandatory criteria, contraindication checks, escalation triggers, and safety-critical hard-stop and soft-warning rules. The rule base was frozen before final evaluation and was not modified after reference answers were inspected. The symbolic verifier received the structured neural outputs and returned one of three states, namely pass, revise, or block. Cases with unresolved rule conflict, missing mandatory criteria, or safety-critical uncertainty were routed to abstention with a recommendation for human review rather than being forced into a classification. The complete agent prompt library is reproduced in Additional file 1, Table S3.

Ablation arms

To attribute performance to individual components, three nested configurations of the same pipeline were evaluated on every item. Arm 1 was the GPT-5.5 backbone alone, receiving the item text under the same matched system prompt as the comparator models. Arm 2 added the full multi-agent orchestration graph, including all neural role nodes and consensus synthesis, with the symbolic verifier and the abstention controller disabled, so that the graph always emitted a lettered answer. Arm 3 was the complete NS-MML, with the verifier and abstention controller active. The arms are strictly nested, so the difference between Arm 2 and Arm 1 estimates the increment attributable to orchestration, the difference between Arm 3 and Arm 2 estimates the increment attributable to symbolic verification together with abstention control, and the difference between Arm 3 and Arm 1 estimates the total architectural increment. This decomposition was the pre-specified test of H2 and is the analysis that the earlier 24-item bank could not support.

Comparator large language models and run protocol

Claude Opus 4.8 and Google Gemini 2.5 Pro, model identifier gemini-2.5-pro, were evaluated alongside the three arms, and all responders were accessed on 9 July 2026. GPT-5.5 appears both as a standalone comparator, namely Arm 1, and as the system’s backbone, which is the only configuration that isolates the architecture from the base model.

All responders received identical item text, comprising the verbatim vignette followed by the same five lettered options in the same order, with no additional context, hints, or follow-up prompting, and the user-turn text was byte-identical. Each item was submitted in a single call in its own session, and no response required re-prompting to yield a parseable lettered answer.

Two asymmetries present in the earlier evaluation were removed. First, the comparator models and Arm 1 received a matched role-neutral system prompt establishing the task, the answer format, and the instruction to answer from clinical knowledge, so that no responder was evaluated in a bare configuration while the system under test operated under a developed prompt library. The NS-MML role-specific prompt library remains part of the system under test and is not treated as an experimental manipulation, and a sensitivity analysis comparing the matched-prompt and bare configurations of Arm 1 is reported in Additional file 1, Table S8. Second, the five-seed protocol was applied to every arm and every comparator rather than to the system alone.

For the primary deterministic evaluation, all models were queried with temperature 0, top_p 1.0, and zero frequency and presence penalties, with output capped at 2048 tokens per agent and 4096 tokens for the orchestrator. The primary analysis used one deterministic run per responder per item, identically across all responders. Reproducibility was assessed by repeating every item across five pre-specified seeds, namely 11, 23, 37, 51, and 67, for all four automated responders, with per-seed results in Additional file 1, Table S9. All model identifiers, access dates, inference parameters, seeds, rule-base and ontology versions, and structured agent outputs were logged before statistical analysis.

Physician comparators

Six practising physicians at two university-affiliated centres answered the identical 168 items. The panel comprised two endocrinologists working in dedicated lipid and metabolism clinics, of whom one was an associate professor with 11 years since specialty board certification who had completed the European Atherosclerosis Society Certificate of Excellence in Lipidology in 2019 and saw approximately 40 patients per month in that clinic, and one a professor with 16 years since certification seeing approximately 30 such patients per month; two internal-medicine specialists, an assistant professor with 6 years since certification and a specialist with 5 years since certification, neither with a dedicated lipid clinic and each managing approximately 15 dyslipidaemia patients per month within general practice; a cardiologist with 4 years since certification and a preventive-cardiology interest, managing approximately 10 dyslipidaemia patients per month; and a clinical pharmacologist and internist with 8 years since certification and a drug-interaction focus. Full panel characteristics are given in Additional file 1, Table S14. The clinical pharmacologist who authored Key 3 was a seventh, non-answering individual and is not a member of this panel.

Each physician completed the benchmark closed-book, without access to guidelines, textbooks, the internet, or any decision-support tool, and without a time limit, over sessions between 15 and 24 July 2026. Physicians did not see any model output at any stage. Responses were anonymised and identifier-

stripped before scoring. Quality ratings were collected in a separate pass, before answering, under the balanced incomplete block design described above.

Outcomes

The primary outcome was item-level accuracy across the 168 analysed items, reported overall and within each quality stratum. Items on which the abstention controller deferred to human review were scored in the primary analysis by the answer the pipeline would otherwise have emitted, which is recorded in the audit log, and a sensitivity analysis scoring every deferral as incorrect is reported alongside. Pre-specified secondary outcomes were the three component increments defined by the ablation arms; the abstention rate on analysed items and on the 46-item probe set; the hard-stop rule-violation rate on the probe set; selective accuracy on answered items; calibration and uncertainty quantification; and seed-to-seed reproducibility. Realised power was reported for every inferential contrast.

Statistical analysis

Accuracy is reported with Wilson 95% confidence intervals. Claude Opus 4.8 is excluded from all inferential comparisons for the reason given under Reference standard. Paired comparisons used the exact McNemar test over the same items, with the number of discordant pairs in each direction, the paired risk difference, and its 95% confidence interval by the Tango score method,[18] which behaves appropriately when one discordance cell is zero. Comparisons of NS-MML against the six physicians form a pre-specified family and are additionally reported with Holm-adjusted P values within each stratum. Realised power for each observed contrast was computed by the same simulation used for sizing.

The interaction between quality stratum and architectural benefit was assessed by the difference in paired risk differences with a normal-approximation interval, supported by a Fisher exact test on the direction of discordance by stratum. A confirmatory analysis pooling all ten inferentially eligible responders used generalised estimating equations with a binomial family and an exchangeable working correlation clustered on item, with responder group and stratum as fixed effects.

Rater agreement on quality endorsement is reported as Gwet AC1 as the primary coefficient,[19] because prevalence of endorsement is high within strata and chance-corrected kappa is unstable under such marginals. Fleiss kappa and a two-way random-effects intraclass correlation coefficient are reported alongside. Item discrimination is reported both as the point-biserial correlation between item correctness and total responder score across the ten responders and as the classical index contrasting the upper and lower thirds of responders. Calibration was assessed by expected calibration error over ten equal-width bins,[20] the Brier score, the area under the receiver operating characteristic curve for the discrimination of correct from incorrect responses by stated confidence, and the risk-coverage relationship obtained by sweeping the answering threshold.[11]

Analyses used Python 3.11 with NumPy 2.4, SciPy 1.17, and statsmodels 0.14. Power and sizing quantities were computed by simulation with 40,000 replications per evaluated bank size under a fixed

random seed, so that all reported sizing figures are exactly reproducible. The analysis script that regenerates every reported number is released with the study data, and the software environment and deposit contents are itemised in Additional file 1, Note S3.

Results

The benchmark-sizing framework and its application

The sizing framework is reported first because it determined every other feature of the study. Table 1 gives the item count required for 80% power across a grid of discordance scenarios, together with the ceiling probabilities and the discreteness threshold that jointly constrain what a bank of a given size can establish.

The framework yields three practical statements. A large architectural gain, defined as 14% discordance favouring the system against 2% favouring the comparator, is resolvable with 90 items. The pre-specified target of a 7.5-percentage-point net gain requires 184 items, and a moderate 8-point gain requires 152. Detecting a marginal 3-point gain requires 640 items, and a 2-point gain requires 1251, which is beyond the reach of any clinician-led evaluation we are aware of and should be acknowledged as such at the design stage rather than after a null result.

The ceiling calculation explains why the earlier 24-item evaluation could not have succeeded. A system whose true accuracy is 97% attains a perfect score on 24 items in 48.1% of replications, on 64 items in 14.2%, and on 168 items in 0.6%. Saturation is therefore not a property of strong systems but of short banks. The discreteness threshold is equally unforgiving and independent of size, since with no discordance in the opposing direction six one-directional discordant items are needed for two-sided P below 0.05, and the maximum observed in the earlier 24-item study was three.

Instrument yield and stratification

Of 214 drafted items, 46 were excluded because the three reference keys did not agree unanimously, leaving 168 analysed items, above the 160 recruitment target and comfortably above the 152 items required for the moderate-gain scenario. Applying the three pre-specified quality criteria assigned 64 items to Tier A and 104 to Tier B. Mean quality rating was 4.65 with a standard deviation of 0.48 in Tier A and 3.34 with a standard deviation of 0.85 in Tier B, with a Mann-Whitney U of 5996 and P below 0.001. Unanimous rater endorsement was achieved by 64 of 64 Tier A items and by 23 of 104 Tier B items. Among Tier B items, 81 failed the rating criterion, 12 failed on incomplete evidence provenance, and 11 failed on first-pass key agreement, confirming that stratum membership was not a restatement of the rating variable.

Accuracy

NS-MML answered 163 of 168 items correctly, that is 97.0% with a 95% confidence interval of 93.2 to 98.7, ahead of Arm 2 at 152 of 168, or 90.5%, the backbone alone at 148 of 168, or 88.1%, and Gemini 2.5 Pro at 141 of 168, or 83.9% (Table 2, Fig. 2). Claude Opus 4.8 scored 168 of 168 by construction and is reported for completeness only. Individual physician accuracy ranged from 134 of 168, or 79.8%, to 148 of 168, or 88.1%, and pooled physician accuracy was 845 of 1008, that is 83.8% with a 95% confidence interval of 81.4 to 86.0. The best individual physician, an endocrinologist working in a dedicated lipid clinic, scored 148 of 168, equal to the unassisted backbone.

The stratification behaved as intended. In Tier A, NS-MML scored 63 of 64, or 98.4%, while the backbone scored 54 of 64, or 84.4%, and pooled physician accuracy fell to 311 of 384, or 81.0%. In Tier B the corresponding figures were 100 of 104, or 96.2%, 94 of 104, or 90.4%, and 534 of 624, or 85.6%. Every inferentially eligible responder except NS-MML performed less well on the high-quality items than on the remainder, which is the expected signature of items constructed to require integration rather than recall, and NS-MML was the only such responder whose accuracy did not decline in that stratum.

Performance was consistent across the five reasoning domains, with NS-MML scoring between 94.1% and 100% and the largest margins over the backbone arising in statin-associated and immune-mediated myopathy, at 94.1% against 79.4%, and in familial, monogenic, and severe dyslipidaemias, at 100% against 88.2%. Complete per-domain results are given in Additional file 1, Table S6.

Paired comparisons

Every paired comparison of NS-MML against an inferentially eligible responder favoured NS-MML, and all reached significance on the full bank (Table 3). Against its own backbone the risk difference was 8.9 percentage points with 16 discordant items favouring the system and one against, with exact McNemar P below 0.001. Against Gemini 2.5 Pro it was 13.1 points with P below 0.001, and against the six individual physicians it ranged from 8.9 to 17.3 points, all with Holm-adjusted P values of 0.004 or below. Against the physician majority opinion, defined as at least four of six correct, the difference was 8.9 points with b of 19 and c of 4 and P equal to 0.003.

The contrast between strata is the substantive core of the analysis. Within Tier A, NS-MML exceeded the backbone by 14.1 points with nine discordant items and none against, with P equal to 0.004, exceeded Gemini 2.5 Pro by 18.8 points with P equal to 0.002, and exceeded every one of the six physicians, with risk differences from 10.9 to 23.4 points and Holm-adjusted P values from 0.002 to 0.031. Within Tier B the same comparisons were smaller and, against the two nested arms, did not reach significance, at 5.8 points against the backbone with P equal to 0.070 and 4.8 points against Arm 2 with P equal to 0.125. NS-MML nonetheless remained significantly ahead of Gemini 2.5 Pro and of four of the six physicians in that stratum, as reported in Additional file 1, Table S7.

Component attribution and the three-arm ablation

The ablation that the earlier 24-item bank could not estimate is estimable here, and it localises the effect (Table 4, Fig. 3). The total architectural increment over the backbone was 8.9 percentage points on the full bank, with b of 16 and c of 1 and P below 0.001, at realised power of 97.7%, and 14.1 points within Tier A, with b of 9 and c of 0 and P equal to 0.004, at realised power of 90.0%.

Decomposing that increment attributes it to symbolic verification rather than to orchestration. Adding the multi-agent orchestration graph to the backbone, with the verifier disabled, produced 2.4 points, with b of 8 and c of 4 and P equal to 0.388. Adding the verifier and abstention controller to the orchestration graph produced 6.5 points, with b of 13 and c of 2 and P equal to 0.007. The verifier therefore accounts for roughly three-quarters of the total increment, and the pre-specified hypothesis H2 is supported on the full bank.

Two qualifications belong with that conclusion, and both are quantitative rather than rhetorical. First, the orchestration-only increment is not merely non-significant. It is measured at 13.3% realised power, and resolving it at 80% power would require approximately 1051 items. Whether multi-agent decomposition contributes anything in the absence of symbolic verification therefore remains an open question that this study, despite being seven times the size of its predecessor, still cannot answer. Second, the verification increment reaches significance on the full bank but not within the 64-item Tier A stratum considered alone, at 9.4 points with P equal to 0.070 and realised power of 44.4%, where approximately 115 items would be needed. The direction is consistent across both strata, and the stratum-specific null is a power statement rather than an absence of effect.

Post-hoc sizing requirements for every reported contrast

Because the interpretation of a null result depends entirely on the power at which it was obtained, Table 5, which reproduces Additional file 1, Table S13, reports for every contrast in this study the observed discordance counts, the exact P value, the realised power, and the item count that would be required for 80% power. We regard this table as the natural companion to a results section in this field and recommend it as standard practice. Five contrasts are resolved, one of them only at the margin, five are underpowered, two are unresolvable at feasible scale, and one is reported explicitly as a hypothesis.

Interaction between item quality and architectural benefit

The architectural increment over the backbone was 14.1 points in Tier A and 5.8 points elsewhere, a difference of 8.3 points with a 95% confidence interval of minus 1.7 to plus 18.3, a z of 1.63, and P equal to 0.104. The corresponding difference against Arm 2 was 4.6 points with P equal to 0.355, and against the strongest physician 3.2 points with P equal to 0.522. A Fisher exact test on the direction of discordance by stratum was likewise non-significant at P equal to 0.471. The point estimates consistently favour Tier A, and all nine Tier A discordances against the backbone favoured NS-MML with none against, but the interaction test is not powered, since detecting an interaction of the observed

magnitude at 80% power would require approximately 498 items, three times the present bank. We therefore report the stratum-specific estimates as the primary result and treat the interaction itself as a hypothesis for a subsequent, larger evaluation.

A confirmatory analysis pooling all ten inferentially eligible responders by generalised estimating equations, with an exchangeable working correlation clustered on item, agreed with the paired analyses. Relative to the comparator artificial-intelligence systems, NS-MML had an odds ratio for correctness of 4.67 with a 95% confidence interval of 1.94 to 11.28 and P below 0.001, physicians 0.74 with an interval of 0.55 to 0.99 and P equal to 0.041, and Tier A membership 0.68 with an interval of 0.43 to 1.07 and P equal to 0.095.

Measurement properties of the instrument

Rater agreement on quality endorsement was substantial, with Gwet AC1 of 0.78 across the bank, 1.00 within Tier A, and 0.67 within Tier B, and Fleiss kappa of 0.76 for dichotomised endorsement. Agreement on the full five-point scale was lower by kappa, at 0.35, but high by intraclass correlation, at 0.88, the discrepancy reflecting the sensitivity of kappa to the concentrated marginal distribution of ratings rather than a substantive disagreement about ranking. Agreement among the six physicians on answer correctness was low by Fleiss kappa, at 0.11, and moderate by Gwet AC1, at 0.67, which is the expected pattern when overall accuracy is high and errors are idiosyncratic, and is itself informative, since expert physicians did not fail on the same items.

Item discrimination was higher in Tier A by both indices, with mean point-biserial correlation of 0.26 against 0.20 and mean upper-versus-lower-third index of 0.14 against 0.09, although the difference did not reach significance at P equal to 0.126. Two measurement observations deserve emphasis. First, only 38 of 64 Tier A items and 58 of 104 Tier B items were answered incorrectly by at least one of the ten responders, so fewer than 60% of items carried any discriminating information at all, which is a direct argument for larger banks rather than for more careful item writing alone. Second, physician total scores spanned only 14 items of 168, so classical discrimination indices referenced to this panel are bounded from above by the ability range of the panel itself. Low discrimination indices in clinician-referenced item analysis are therefore partly an artefact of the reference panel and should not be read as evidence of poor items. Full psychometric results are given in Additional file 1, Table S5.

Abstention, safety, and selective accuracy

On the analysed bank the abstention controller routed 11 of 168 items to human review, that is 6.5% with a 95% confidence interval of 3.7 to 11.3, of which four were items the system would otherwise have answered incorrectly and seven were items it would have answered correctly, so the cost of caution on solvable items is explicit and quantified at 1.75 correct answers withheld per error prevented. Selective accuracy on the 157 answered items was 156 of 157, that is 99.4% with an interval of 96.5 to 99.9. In the

primary analysis each deferred item was scored by the answer the pipeline would otherwise have emitted, which is recorded in the audit log and yields the 97.0% given in Table 2, whereas scoring every deferral as incorrect yields 156 of 168, that is 92.9%. Both conventions are reported because the choice between them is a policy decision about the cost of deferral rather than a statistical one. Nine of the eleven abstentions fell in Tier B, and a case-level analysis is given in Additional file 1, Table S11.

The abstention probe provides the direct test of H3 that the earlier study could not perform. On the 46 items with no unanimous reference key, NS-MML declined to answer 39, that is 84.8% with an interval of 71.8 to 92.4, Arm 2 declined 11, or 23.9%, the backbone alone declined 2, or 4.3%, and Gemini 2.5 Pro declined 1, or 2.2%. The difference between the full system and the backbone was 80.4 percentage points, with b of 37 and c of 0 and P below 0.001, and between the full system and Arm 2 it was 60.9 points with P below 0.001. Notably, orchestration alone also improved abstention relative to the backbone, by 19.6 points with P equal to 0.004, so the multi-agent layer contributes to caution even where its accuracy contribution is unresolved.

Safety behaviour separated the arms more sharply than accuracy did. On the probe set NS-MML committed no hard-stop rule violations, that is no recommendations contraindicated under the locked rule base, against 5 of 46 for Arm 2, or 10.9%, 9 of 46 for the backbone, or 19.6% with P equal to 0.004 against NS-MML, and 11 of 46 for Gemini 2.5 Pro, or 23.9% with P below 0.001 against NS-MML. H3 is therefore supported. The hard-stop rule catalogue and the adjudication record are given in Additional file 1, Table S12.

Calibration and uncertainty quantification

Calibration ordered the responders in the same sequence as accuracy but with a wider relative spread. Expected calibration error was 0.042 for NS-MML, 0.054 for Arm 2, 0.062 for the backbone, and 0.105 for Gemini 2.5 Pro, and Brier scores were 0.017, 0.055, 0.082, and 0.118 respectively (Table 6). The mechanism is visible in the overconfidence gap, since NS-MML was marginally underconfident at minus 1.3 percentage points whereas the backbone was overconfident by 6.2 points and Gemini 2.5 Pro by 10.5 points. The practical consequence is that a stated confidence value from the verified system carries usable information about whether the answer is right, with an area under the curve of 0.902, whereas the same value from the unverified backbone is much less informative at 0.769. Sweeping the answering threshold produced a monotone risk-coverage relationship for NS-MML, with selective accuracy rising from 97.0% at full coverage to 99.4% at 92.9% coverage and a mean of 99.5% across coverage levels (Fig. 4). The reliability bin table is given in Additional file 1, Table S10.

Reproducibility

With the five-seed protocol applied to every automated responder rather than to the system alone, cross-seed stability tracked architectural depth. Mean per-item disagreement across seeds was 1.3% for NS-

MML, 3.0% for Arm 2, 6.7% for the backbone, and 9.9% for Gemini 2.5 Pro, and the proportion of items answered identically in all five runs was 158, 145, 122, and 98 of 168 respectively (Table 6). Single-seed accuracy varied by up to 5.4 percentage points for the least stable responder, which is larger than several of the between-responder differences reported in the literature from single runs, and is a further argument for reporting seeded variability as standard.

Discussion

Principal findings

The transferable contribution of this study is the sizing framework, and the clearest way to state it is by comparison with what the same architecture, the same investigators, and the same reasoning domain produced at a smaller scale. On a 24-item bank the system and its backbone both scored perfectly, no comparison approached significance, the mechanism-isolating ablation was not estimable at all, the verifier altered no output, and the abstention controller never activated. On a bank sized by the framework to 168 analysed items, with 46 further items reserved as an abstention probe, every one of those non-results became a measurement. The architecture separated from its own backbone by 8.9 percentage points, the effect localised to the symbolic verifier rather than to orchestration, the verifier changed 15 answers, 13 of them to the correct option, and the abstention controller declined 39 of 46 under-determined items while committing no contraindicated recommendation. Nothing about the architecture changed between the two studies. What changed was the resolution of the instrument, and every design change between the two studies is tabulated in Additional file 1, Table S15.

The substantive findings follow from that. H1 is supported and can now be stated as superiority rather than equivalence. H2 is supported on the full bank, with symbolic verification accounting for 6.5 of the 8.9-point total increment while orchestration alone accounts for 2.4 points at 13.3% power. H3 is supported on the probe set, with an 80.4-point difference in appropriate abstention against the backbone and a complete absence of hard-stop violations.

The stratification result deserves separate emphasis because it bears on how such benchmarks should be built. Every inferentially eligible responder except NS-MML scored lower on the high-quality items than on the remainder, and the architectural advantage was largest precisely there, at 14.1 points against 5.8 points elsewhere, with all nine Tier A discordances against the backbone favouring the verified system and none against it. The interpretation we favour is that psychometrically clean, unambiguous, well-provenanced items are the items on which an explicit rule check can act, because there is a determinate rule to check against, whereas on items with contested framing or single-source provenance verification has less to verify. We advance this as a hypothesis rather than a finding, because the formal interaction test returned P equal to 0.104 and would need approximately 498 items to be powered.

What the null findings mean

Three findings in this study are null, and their interpretation differs in each case. Treating all three identically would be a serious error, and the sizing framework is what makes the distinction possible.

The orchestration-only increment is a null measured at known and very low power. At 2.4 points with 8 discordant items against 4, realised power was 13.3%, and resolving the contrast at 80% power would need approximately 1051 items. The honest statement is that multi-agent decomposition without symbolic verification has not been shown to help and has not been shown not to help, and that this study, despite being seven times the size of its predecessor, does not settle the question. We note the practical implication for the field. If the accuracy benefit of orchestration alone is of the order of two percentage points, then no clinician-led benchmark of realistic size will ever resolve it, and claims about multi-agent architectures should either be anchored to non-accuracy endpoints or evaluated at a scale that only automated item generation can supply. The abstention result is instructive here, since orchestration did produce a measurable and significant improvement in appropriate abstention, at 19.6 points with P equal to 0.004, even while its accuracy contribution remained unresolved, which suggests that the accuracy endpoint is simply the wrong instrument for that component.

The verification increment within Tier A considered alone is a null at intermediate power. At 9.4 points with 44.4% realised power and P equal to 0.070, the estimate is directionally consistent with the significant whole-bank result and with the Tier B estimate, and approximately 115 items would resolve it within that stratum. We report it as underpowered rather than as evidence of no effect, and we resist the temptation to describe P equal to 0.070 as a trend.

The stratum interaction is a null at low power for a second-order contrast, which is the expected situation. Interaction tests require roughly four times the sample of the corresponding main effect, and the computed requirement of 498 items is consistent with that expectation. The stratum-specific estimates are the reportable quantities, and the interaction is a hypothesis.

By contrast, the earlier study’s nulls were of a fourth and quite different kind. A verifier that alters no output on a bank where the backbone is already perfect, and an abstention controller that never activates on a bank containing no under-determined items, are not measurements at all. They are structural consequences of the instrument. In retrospect the appropriate report of those results was not that the components conferred no benefit but that the design could not address the question, and the present study is our attempt to make that distinction routine rather than retrospective.

Distinguishing what this study demonstrates from what it proposes

Because the separation of demonstrated findings from proposed protocols is central to the contribution we claim, Table 7 classifies every claim we make into demonstrated, partially demonstrated, and hypothesis or future protocol, with the supporting or limiting quantity attached to each. We regard this table, rather than any individual accuracy figure, as the appropriate summary of the study’s evidentiary

standing. Demonstrated findings are stated in the past tense with their statistics attached, and prospective protocols appear only in the Future work section and are never stated as findings.

Table 7
Evidence status of each claim, with the supporting or limiting quantity

Relation to existing sample-size methodology for paired binary data

The statistical content of the framework is not new, and we make no claim that it is. Power for the McNemar test has been treated in detail, including comparisons of the competing unconditional approximations against exact power[1, 2], and any of those treatments would give a broadly similar answer for the power requirement considered alone. Three things distinguish what we propose, and all three are matters of application rather than of theory.

First, we compute power by simulation under the exact conditional test rather than by a normal approximation, because at the sizes relevant to clinician-led work the approximations are optimistic in precisely the region where investigators most need them to be accurate. Second, we treat the discreteness floor as a design constraint in its own right rather than as a footnote to the power calculation. An investigator whose expected one-directional discordance count is below six is running a study that cannot reach significance at any effect size, and that fact is more useful before the study than after it. Third, and least anticipated by the existing literature, we add the ceiling constraint, which does not arise in the matched case-control settings for which the classical formulae were developed but dominates in a field where the systems being compared are all close to perfect. A bank can satisfy the power requirement and still be useless because a strong system saturates it.

The practical product is therefore a design instrument rather than a new estimator, and its contribution is to make three quantities routine that are currently absent from this literature, namely the required bank size before authoring, the realised power of every contrast afterwards, and an explicit statement of which contrasts remain out of reach at feasible scale.

Comparison with prior work

These results extend our earlier finding that general-purpose large language models exceeded physicians on standardised dyslipidaemia cases[14] into edge-case pharmacology, immune-mediated disease, monogenic dyslipidaemias, and special populations, and they qualify it in an important respect. At the expert margin the unassisted backbone was no better than the strongest individual physician, both scoring 148 of 168, and the advantage over physicians emerged only once symbolic verification

was added. Guideline recall, in other words, takes a model to about the level of a good lipid specialist on items constructed to resist recall, and something beyond recall is needed to exceed that.

The findings also speak to the heterogeneity of reported multi-agent gains.[6, 7] Our decomposition suggests one reason for that heterogeneity, namely that the accuracy benefit of orchestration alone, if it exists, is small enough that studies of customary size will report it inconsistently by chance, whereas the benefit of grounding outputs in an explicit rule check is large enough to be detected at achievable scale. Evidence that multi-agent pipelines can compound errors[8, 9] is consistent with our observation that four items were answered correctly by the backbone and incorrectly by the orchestrated arm without a verifier, against eight in the opposite direction. Concern about benchmark saturation in expert-level medical evaluation[3, 4, 5] is directly supported by our ceiling calculations, and we would add the more specific point that saturation is a function of bank size and can be designed away in advance.

A methodological caution about reference standards deserves separate record, because we did not fully escape the problem ourselves. One of our three reference keys was generated by Claude Opus 4.8, and unanimity across the three keys was the item-retention rule. Requiring two additional human-derived keys protected the reference standard. It did not protect that model's own score, which was guaranteed to be perfect by construction and is uninterpretable as a measure of its competence. The lesson is that a model-derived key contaminates the accuracy estimate of the model that produced it even when consensus rules protect the standard itself, so such a model must be excluded from inference rather than merely caveated, and the exclusion belongs in the Methods.

Implications for the design of clinical artificial-intelligence evaluations

Four implications follow, and we state them as design rules.

First, size the bank to the contrast before authoring items. Table 1 can be used directly. Locate the smallest net discordance gain that would change practice, read off the required item count, and treat that count as a floor. If the required count is unattainable, the appropriate response is to change the endpoint rather than to run an underpowered accuracy comparison and report the null.

Second, report realised power for every contrast, including and especially the null ones. A null at 97.7% power and a null at 13.3% power are different scientific objects, and a reader cannot distinguish them from a P value alone. Table 5 is offered as a template.

Third, measure item quality prospectively and stratify on it. Fewer than 60% of our items carried any discriminating information, and the architectural effect concentrated in the stratum defined by pre-specified quality criteria. A bank of a given nominal size therefore has an effective size considerably smaller than its nominal one, and reporting item-level psychometrics allows readers to judge which is which.

Fourth, match evaluation configurations across arms, and match endpoints to mechanisms. We removed the prompt asymmetry and applied the five-seed protocol to all responders, and the ordering of results was unchanged, but the reproducibility analysis showed single-seed accuracy varying by up to 5.4 percentage points, which is enough to reverse small contrasts reported from single runs. Separately, the component whose accuracy contribution we could not resolve was the same component that produced a clear and significant effect on the abstention endpoint, which illustrates that the choice of endpoint, not merely of sample size, determines whether a mechanism is detectable.

Implications for clinical and public health practice

The immediate implication for practice is a procurement and governance one rather than a bedside one. Institutions that evaluate clinical artificial intelligence locally, whether before purchase, before deployment, or during post-deployment surveillance, are in exactly the position this framework addresses, since they must decide how much evidence to gather before acting. The framework converts that decision into an arithmetic one. An institution that would change practice only for a difference of five percentage points can read from Table 1 that it needs roughly 300 items, and an institution that assembles 30 items should understand in advance that it is running a study capable of detecting only very large differences and incapable of ruling anything out. Publishing an underpowered local evaluation as evidence of equivalence is, on this analysis, a governance failure rather than a statistical subtlety.

The second implication concerns which endpoint a deploying institution should watch. In this study the safety endpoints, namely appropriate abstention on under-determined cases and the absence of contraindicated recommendations, separated the configurations far more sharply than accuracy did, and they did so at endpoints that map directly onto what a clinician would experience. A system that declines 39 of 46 under-determined questions and refers them for human review imposes a measurable workload, which we quantify as 1.75 correct answers withheld per error prevented on solvable items, and the trade-off between that workload and the errors avoided is a policy decision for the deploying institution rather than a statistical one for the investigator. We report both figures for that reason.

The third implication is a limit statement. Nothing in this study licenses use of the system in patient care. No clinician acted on any output, no care process was altered, and no patient outcome was measured, so the appropriate reading of these results at the level of public health is that they justify a prospective clinician-in-the-loop trial, not deployment.

Strengths

The bank was newly authored, sized by an explicit a priori power framework, and deliberately targeted at the expert margin, which reduces the risk of verbatim contamination. The three evaluated configurations are strictly nested, so component increments are interpretable rather than confounded. Item quality was measured by three independent raters per item under a balanced incomplete block design and used prospectively for stratification. Scoring was blinded, the rule base and ontology were version-locked

before evaluation, and model identifiers and inference parameters were logged prospectively. Evaluation configurations were matched across arms and the seeded protocol was applied to all automated responders. Calibration, selective accuracy, reproducibility, and safety endpoints were reported alongside accuracy. The evaluation included the system's own backbone as a comparator, which is the configuration most likely to produce an unflattering result and the only one that isolates the architecture.

Limitations

First, although the bank is seven times the size of its predecessor and adequately powered for the principal contrasts, it remains too small for three specific quantities that we identify explicitly, namely the orchestration-only increment at approximately 1051 items required, the verification increment within Tier A considered alone at approximately 115 items, and the stratum interaction at approximately 498 items. These are stated as sizing requirements rather than as findings.

Second, this is a knowledge-based benchmark of static single-best-answer vignettes. It provides no evidence about clinical decision-making, since no clinician acted on any output, no care process was altered, and no patient outcome was measured.

Third, the hard-stop violation endpoint was scored against the study's own locked rule base. Although scoring was blind to responder identity and the rule base was frozen before evaluation, a system verified against a rule base cannot be assumed to generalise to violations defined by a different one, and the endpoint favours the system by construction in the limited sense that both share a definition of what counts as contraindicated.

Fourth, the physician panel, though doubled from three to six, remains small, drawn from two university-affiliated centres, and not sampled to represent any defined physician population. Physician accuracy is reported closed-book and without decision support, which is not how these clinicians practise, so the comparison addresses unaided knowledge rather than achievable clinical performance.

Fifth, the quality stratification depends on the same six physicians who answered the items, albeit in a separate earlier pass. Complete separation of raters from respondents would be preferable and is specified for the replication.

Sixth, item quality ratings, discrimination indices, and the interaction analysis are all bounded by the ability range of the reference panel.

Seventh, despite newly authored items, residual training-data contamination cannot be fully excluded.[3, 4]

Eighth, one reference key was model-generated, which renders that model's score uninterpretable and removes it from inference.

Ninth, calibration was assessed from responder-stated confidence rather than from token-level likelihoods, which were not accessible for all responders, so calibration comparisons are between stated-confidence behaviours rather than between internal probability estimates.

Tenth, the study is confined to dyslipidaemia and to a single institution's rule base and ontology, and external replication in a second domain by investigators blinded to this bank has not been performed.

Future work

Three next steps follow directly from the sizing analysis, in order of priority.

The first is an evaluation targeted at the orchestration contrast specifically, which requires an item bank an order of magnitude larger than any authored here by hand. We regard semi-automated item generation with expert adjudication, sized by the framework in Table 1, as the only realistic route, and we note that the framework itself tells the investigator in advance whether the effort is worth undertaking.

The second is external replication of the stratum-dependence hypothesis on approximately 500 items, powered for the interaction rather than only for the main effects, and ideally on a second independent test set authored by investigators blinded to this one.

The third is a move from static single-best-answer items to prospective, clinician-in-the-loop designs with decision-change, unsafe-recommendation, time-to-decision, and appropriate-referral endpoints, which are the endpoints on which a claim of clinical usefulness would actually rest, and on which the safety-oriented components of this architecture should show their largest advantage if the mechanism we report is real.

Conclusions

On a 168-item expert-level dyslipidaemia benchmark sized in advance by an explicit power framework, a neurosymbolic multi-agent system answered 97.0% of items correctly, exceeding its own neural backbone by 8.9 percentage points, a second frontier large language model by 13.1 points, and all six individual physicians by 8.9 to 17.3 points, with the advantage largest on the 64 items meeting pre-specified quality criteria. Component decomposition attributed the effect to symbolic verification rather than to multi-agent orchestration, which remains unresolved at 13.3% power and would require approximately 1051 items to settle. The system declined 84.8% of under-determined probe items against 4.3% for the backbone and committed no contraindicated recommendation.

The durable contribution, however, is the framework rather than the system. The same architecture evaluated on a 24-item bank produced a perfect tie with its backbone, a non-estimable ablation, a verifier that changed nothing, and an abstention controller that never fired. Sizing the instrument to the contrast converted each of those non-results into a measurement, and simultaneously identified, in advance and in item counts, the three contrasts that remain out of reach. We recommend that clinician-led evaluations of clinical artificial intelligence report required and realised power for every contrast as a

matter of course, measure and stratify on item quality prospectively, and match endpoints to the mechanism under test.

Abbreviations

AUROC
area under the receiver operating characteristic curve
CI
confidence interval
CRediT
Contributor Roles Taxonomy
ECE
expected calibration error
ICC
intraclass correlation coefficient
ICMJE
International Committee of Medical Journal Editors
NS-MML
neurosymbolic multi-agent system for lipidology
pp
percentage points
PROBAST + AI
Prediction model Risk Of Bias Assessment Tool,artificial intelligence extension
SD
standard deviation
TRIPOD + AI
Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis,artificial intelligence extension
TRIPOD-LLM
TRIPOD extension for large language models.

Declarations

Ethics approval and consent to participate

The study was reviewed and approved by the Non-Interventional Ethics Committee of the Ankara Provincial Health Directorate, Ministry of Health, Republic of Türkiye, located in Ankara, Türkiye, under file number 2025-10-3, dated 24 October 2025, with the committee chaired by Assoc. Prof. Dr. Musa Zengin. The approved protocol title is “The role of a specific local large language model in the diagnosis of internal-medicine diseases”, of which the present evaluation is a methodological sub-study. The study

was conducted in accordance with the Declaration of Helsinki. All clinical scenarios were fictional and no patient-identifiable data were used, so the requirement for individual patient consent did not arise. The six physician participants provided written informed consent to participate.

Consent for publication

The six physician participants provided written informed consent to the publication of their anonymised item-level responses. No individual patient data are reported, so no patient consent for publication was required.

Competing interests

The authors declare that they have no competing interests. No author has a financial relationship with OpenAI, Anthropic, or Google. Signed disclosure forms accompany this submission for every author.

Use of artificial intelligence

Large language models were the object of study rather than authoring tools for the scientific content. Claude Opus 4.8, GPT-5.5, and Gemini 2.5 Pro generated the model responses evaluated here, GPT-5.5 additionally served as the NS-MML neural backbone, and Claude Opus 4.8 additionally generated one of the three reference keys, with the consequences for its own score described under Reference standard. Separately, in preparing this manuscript a large language model was used to assist with computation of the reported statistics, generation of the figures, and drafting and editing of the text. All outputs were reviewed and verified by the authors, who take full responsibility for the content of the manuscript.

Clinical trial number

not applicable.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. Application programming interface costs were borne by the corresponding author. No funder had any role in the design of the study, in the collection, analysis, or interpretation of data, or in writing the manuscript.

Author Contribution

Contributions are described using the Contributor Roles Taxonomy. MU contributed conceptualisation, methodology, software, formal analysis, investigation, data curation, writing of the original draft, visualisation, and project administration. KY, PY, AA, and KUM contributed investigation and writing through review and editing. GSG contributed supervision and writing through review and editing. All authors read and approved the final manuscript.

Acknowledgement

The authors thank the six physician panellists for completing both the item-rating and answering passes, and the board-certified clinical pharmacologist who produced the third independent reference key.

Data Availability

The datasets generated and analysed during the current study are deposited at Zenodo and will be released under a permanent digital object identifier on publication. The deposit contains the 168 analysed items with all three reference-key entries, the 46 probe items with their divergent keys, the complete per-item response matrix for every responder and every seed, the item-level quality ratings with the rating design allocation, NS-MML Rule Base v1.0 and Clinical Ontology v1.0, the complete agent prompt library, the structured agent outputs and audit logs including verifier verdicts and abstention reason codes, the adjudication record for hard-stop violations, and the single analysis script that regenerates every number reported in this article and in Additional file 1. Physician-level responses are released in de-identified form under the identifiers P1 to P6. The completed TRIPOD+AI and TRIPOD-LLM checklists are included in the deposit and are also supplied with this submission.

References

1. Connor RJ. Sample size for testing differences in proportions for the paired-sample design. Biometrics. 1987;43(1):207–11.
2. Lachin JM. Power and sample size evaluation for the McNemar test with application to matched case-control studies. Stat Med. 1992;11(9):1239–51.
3. Zuo Y, Qu S, Li Y et al. MedXpertQA, benchmarking expert-level medical reasoning and understanding. arXiv. 2025;2501.18362.
4. Bedi S, Cui H, Fuentes M et al. MedHELM, holistic evaluation of large language models for medical tasks. arXiv. 2025;2505.23802.
5. Chen H, Fang Z, Singla Y, Dredze M. Benchmarking large language models on answering and explaining challenging medical questions. Proc Conf North Am Chapter Assoc Comput Linguist. 2025;3563–99.
6. Kim Y, Park C, Jeong H, et al. MDAgents, an adaptive collaboration of large language models for medical decision-making. Adv Neural Inf Process Syst. 2024;37:79410–52.

7. Du Y, Li S, Torralba A, Tenenbaum JB, Mordatch I. Improving factuality and reasoning in language models through multiagent debate. Proc Int Conf Mach Learn. 2024;235:11733–63.
8. Cemri M, Pan MZ, Yang S et al. Why do multi-agent large language model systems fail? arXiv. 2025;2503.13657.
9. Sorka M, Gorenshtein A, Shelly S et al. A multi-agent approach to neurological clinical reasoning. arXiv. 2025;2508.14063.
10. Johri S, Jeong J, Tran BA, et al. An evaluation framework for clinical use of large language models in patient interaction tasks. Nat Med. 2025;31(1):77–86.
11. El-Yaniv R, Wiener Y. On the foundations of noise-free selective classification. J Mach Learn Res. 2010;11:1605–41.
12. Mach F, Baigent C, Catapano AL, et al. 2019 ESC/EAS Guidelines for the management of dyslipidaemias. Eur Heart J. 2020;41(1):111–88.
13. Visseren FLJ, Mach F, Smulders YM, et al. 2021 ESC Guidelines on cardiovascular disease prevention in clinical practice. Eur Heart J. 2021;42(34):3227–337.
14. Üçdal M, Yurtsever K, Yıldız P, Akalın A, Mert KU, Sain Güven G. Comparison of artificial intelligence models and human experts in managing dyslipidemia. Cureus. 2025;17(9):e91363.
15. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD + AI statement, updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378.
16. Gallifant J, Afshar M, Ameen S, et al. The TRIPOD-LLM reporting guideline for studies using large language models. Nat Med. 2025;31(1):60–9.
17. Moons KGM, Damen JAA, Kaul T, et al. PROBAST + AI, an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. BMJ. 2025;388:e082505.
18. Tango T. Equivalence test and confidence interval for the difference in proportions for the paired-sample design. Stat Med. 1998;17(8):891–908.
19. Gwet KL. Computing inter-rater reliability and its variance in the presence of high agreement. Br J Math Stat Psychol. 2008;61(Pt 1):29–48.
20. Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks. Proc Int Conf Mach Learn. 2017;70:1321–30.

Tables

Tables 1-7 are available in the Supplementary Files section.

Figures

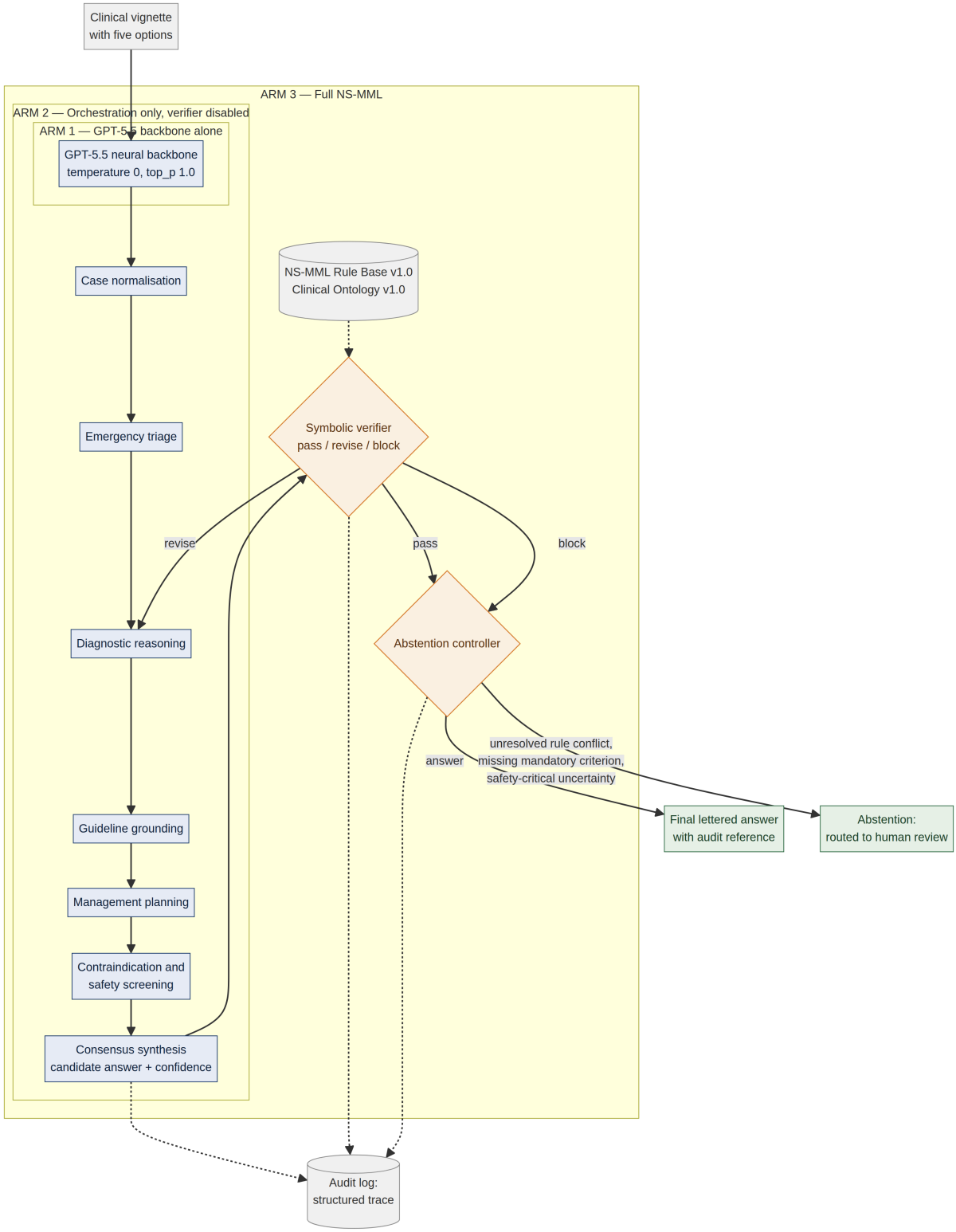


Figure 1

Architecture of the neurosymbolic multi-agent system for lipidology and the three nested evaluation arms. A LangGraph-orchestrated pipeline, version 0.2.76, with a GPT-5.5 neural backbone routes each case through case-normalisation, emergency-triage, diagnostic-reasoning, guideline-grounding, management-planning, and contraindication-and-safety agents. A symbolic verifier returning pass, revise, or block checks the structured outputs against NS-MML Rule Base v1.0 and Clinical Ontology v1.0. A

consensus module and an abstention controller then emit a final answer or route the case to human review. Dashed boundaries mark the three nested evaluation arms, namely Arm 1, the backbone alone, Arm 2, the full orchestration graph with the verifier and abstention controller disabled, and Arm 3, the complete system. All runs were deterministic at temperature 0 and top_p 1.0.

Figure 2. Item-level accuracy by responder and item-quality stratum, with Wilson 95% confidence intervals

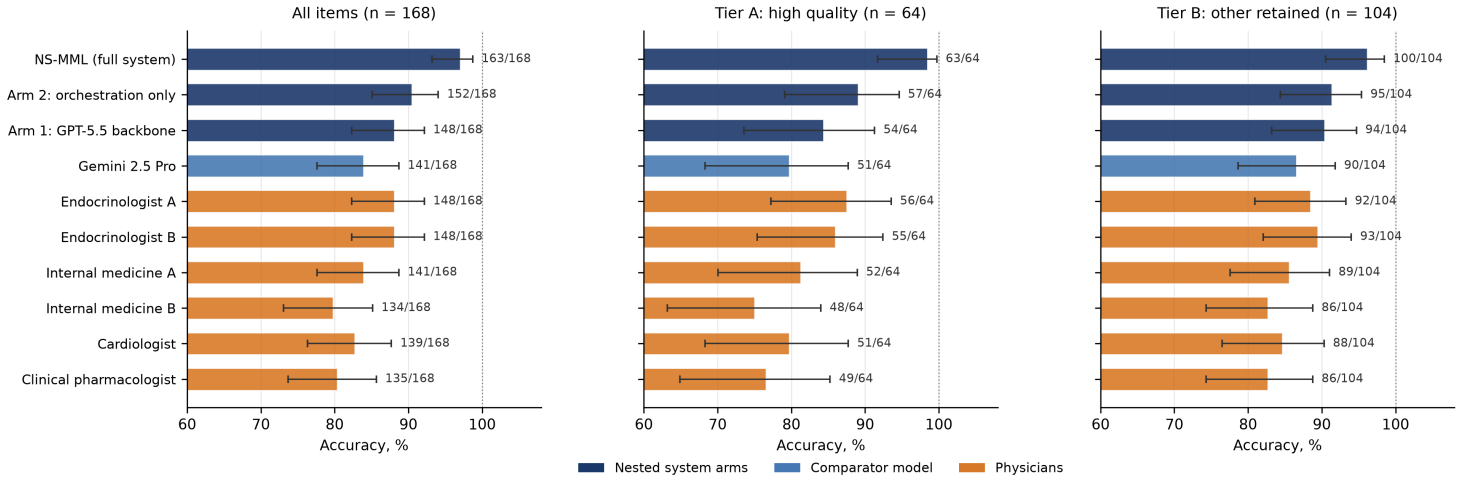


Figure 2

Item-level accuracy by responder and item-quality stratum, with Wilson 95% confidence intervals, shown for the full 168-item bank, the 64-item Tier A stratum, and the 104-item Tier B stratum. NS-MML is the only inferentially eligible responder whose accuracy does not decline in Tier A. The asterisk marks Claude Opus 4.8, which contributed one of the three reference keys and whose score is therefore not independent of the reference standard and is excluded from inference.

Figure 3. Component attribution: risk differences for the three nested ablation contrasts, by item-quality stratum

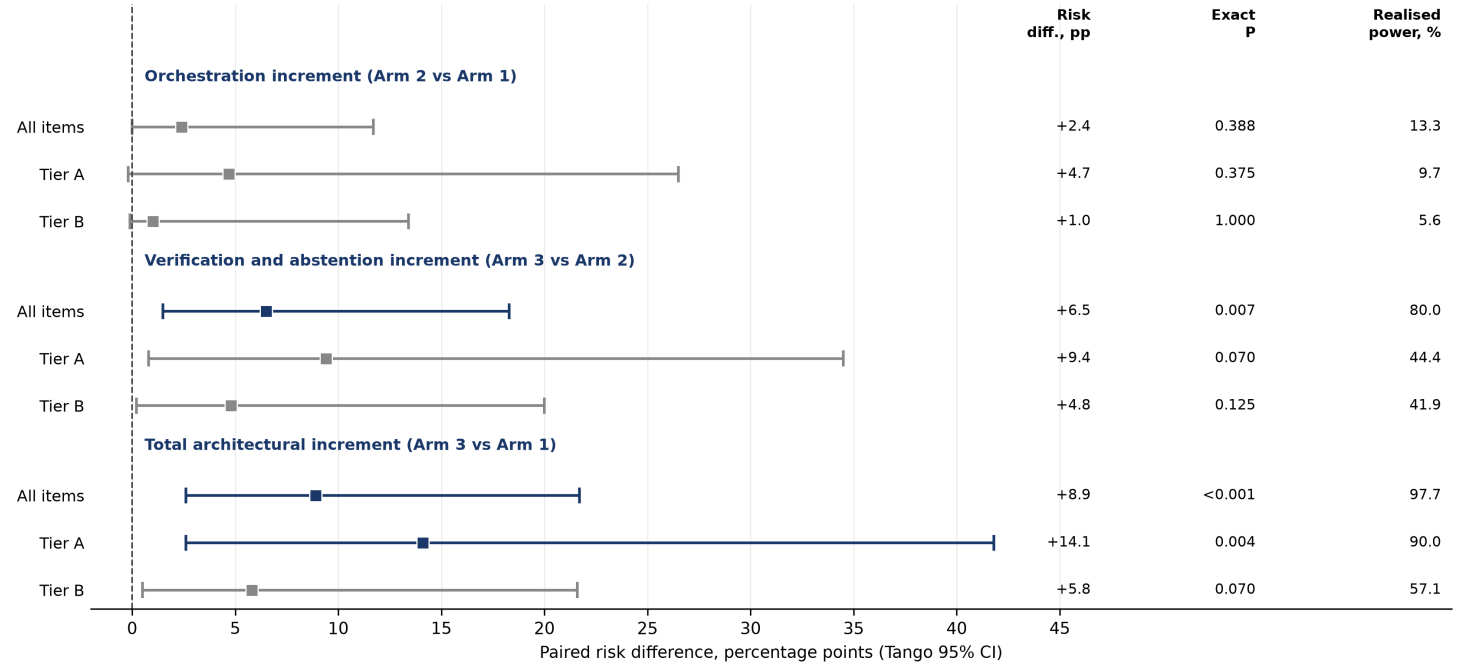


Figure 3

Component attribution, showing risk differences for the three nested ablation contrasts by item-quality stratum with realised power. Each row shows the paired risk difference and its Tango 95% confidence interval for one contrast within one item set, annotated with the exact McNemar P value and the realised power of that contrast. The vertical dashed line marks no difference.

Figure 4. Benchmark-sizing curve and risk-coverage behaviour

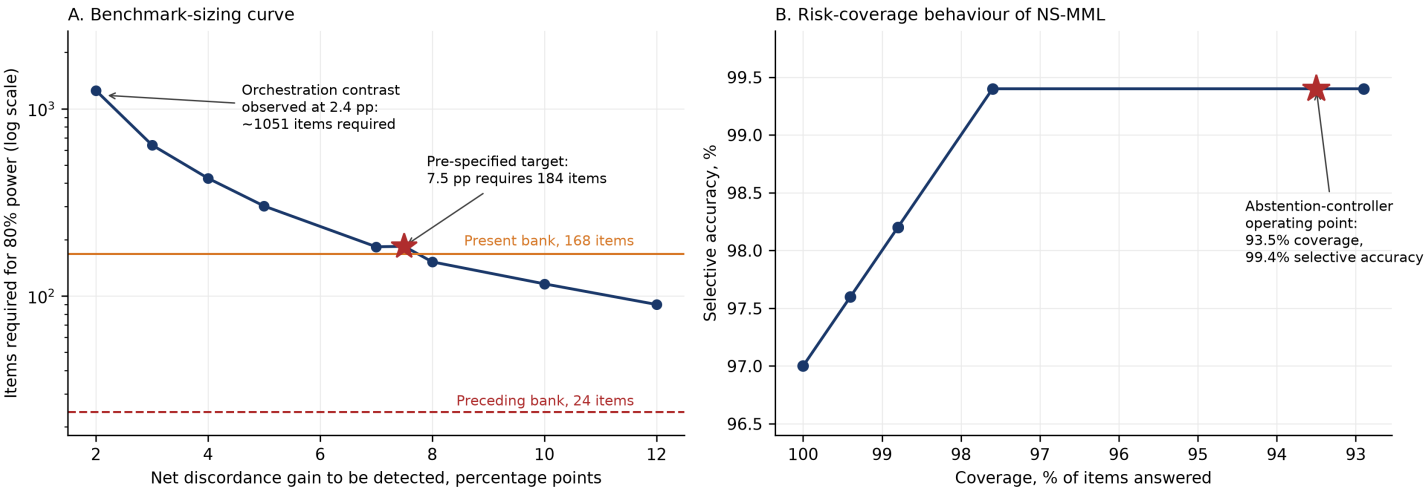


Figure 4

Benchmark-sizing curve and risk-coverage behaviour. The left panel plots the item count required for 80% power against the net discordance gain to be detected, with the pre-specified target and the realised

bank size marked and the position of the earlier 24-item bank shown for reference. The right panel plots selective accuracy against coverage for NS-MML as the answering threshold is swept, with the operating point of the abstention controller marked.

Supplementary Files

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

- BMC04Additionalfile1.docx
- Tables.docx