

**Methods/Research · Educational evidence summary**

A protocol for external validation of Ask UnifiedACB responses

This methods article addresses a protocol for external validation of Ask UnifiedACB responses.

UnifiedACB · Published October 3, 2026 · Updated October 4, 2026

Read the current article

<https://unifiedacb.com/articles/099-a-protocol-for-external-validation-of-ask-unifiedacb-responses/>

Abstract

An answer interface can return the correct source value while still communicating uncertainty poorly or carrying the wrong medication into a follow-up. External evaluation must therefore examine more than isolated lookup accuracy. This article proposes a protocol for evaluating Ask UnifiedACB against frozen evidence records and realistic user tasks. It distinguishes existing internal regression testing from a future independent assessment. The proposed methods include independently assembled cases, predefined reference answers, multi-turn conversations, ambiguous product queries, source conflicts, missing entries, and questions beyond the evidence. Evaluation would separate factual errors from failures of qualification, context, usability, and clinical scope. No new external raters, user sample, accuracy estimate, or clinical outcome study are reported here. The intended result is a transparent account of what the interface retrieves, how it explains the evidence, and where users can misunderstand it. Passing such an evaluation would support a defined use case rather than validate a prescribing rule or universal clinical risk model.

Introduction

Ask UnifiedACB brings medication evidence into a conversational workflow. That workflow introduces questions that do not arise in a static table: which medicine a pronoun refers to, whether a comparison is still active, and whether a suggested follow-up preserves the limits of the preceding answer. A reliable evidence record does not by itself prove that every conversation communicates it correctly.

The objective of external evaluation is to test the delivered experience with independence from its implementation assumptions. Internal tests remain valuable, but an independently designed assessment can identify cases and misunderstandings that the development process did not anticipate. This article describes a proposed evaluation, not a completed certification.

Methods

Freeze the evaluated system and reference evidence

The study would record the interface release, evidence dataset version, content version, and evaluation date. Reference answers would be assembled from the frozen source observations and



relevant explanatory material. The reference set would preserve medication identity, source version, route and formulation qualifications, explicit zeroes, missing entries, and public-access restrictions.

Cases would be created independently of the implementation's benchmark where feasible. Development examples could be used for orientation, but the final assessment set should not be tuned to the same cases used to repair the system. Any overlap should be disclosed so that a successful result is not misrepresented as wholly independent testing.

Sample the tasks the product actually supports

The case set would include direct medication lookup, named-source questions, comparisons, explanatory questions, and article-context follow-ups. It would deliberately include ambiguous names, spelling variants, incomplete products, contradictory source values, and medicines absent from a selected source. Ordinary user wording should be represented alongside formal pharmacologic terms.

Multi-turn cases would test medication carryover, comparison changes, clarifying questions, and explicit topic switches. A correct first answer followed by an incorrect referent is a conversational failure. The protocol should also test whether starting a new conversation clears irrelevant prior context and whether an article entry point retains the article's evidence limits.

Suggested questions would be evaluated as part of the output. A suggestion can steer the user toward an unsupported premise even if the preceding answer is accurate. Reviewers would examine whether follow-ups remain relevant, understandable, and answerable within the available evidence, rather than rewarding novelty alone.

Define error categories before scoring

Factual fidelity would cover source values, identities, versions, and citations. Qualification fidelity would cover missingness, route/formulation scope, disagreements, and uncertainty. Context fidelity would cover the active medicine, comparison, and topic. Clinical scope would examine unsupported diagnosis, treatment instructions, or claims that a classification predicts an individual outcome.

The protocol would distinguish a wrong answer from an appropriate statement that information is unavailable. It would also distinguish an overly broad refusal from a justified limit: declining a prescribing decision should not prevent a source lookup that can be answered accurately. Evaluation should reward useful bounded responses, not simply short or cautious responses.

Independent reviewers would use a written rubric and record disagreements before adjudication. The number and expertise of reviewers, case sampling method, and final sample size would be specified in the complete study plan. No reviewer agreement statistic is claimed here. A future report should show the types and consequences of errors as well as aggregate performance.

Evaluate comprehension and usability separately

A user study could ask participants to interpret an answer, identify the source, explain a missing entry, or continue a comparison. Task completion, interpretation, and reported confidence would be measured separately. High confidence in a misunderstood answer would be a finding to investigate, not evidence of successful communication.

Responsive layouts, keyboard interaction, and supported browser behavior would be included as delivery conditions. Technical accessibility or rendering success would remain distinct from clinical comprehension. Any comparison between interface versions would need a prespecified design and consistent tasks rather than an informal impression of improvement.



Results

The published Validation & QA page reports a frozen 112-case Ask benchmark and 104,858 additional test queries with 335,687 assertions. These figures describe the existing internal release-testing record. They are not newly generated results from the proposed external evaluation, and they do not establish clinical validity. [UnifiedACB: Validation & QA](#)

No external participant study, independent rater assessment, or new conversational accuracy estimate was completed for this article. The proposed outputs are a case inventory, reference answers, rubric-based results, disagreement records, user interpretation findings, and a versioned list of limitations.

Future findings should identify the evaluated release explicitly. A subsequent interface or content change may affect behavior even if source values remain unchanged. Results should therefore be presented as evidence about a defined version and use case, with the scope of any retesting made clear.

Discussion

Aggregate accuracy can conceal consequential errors. Many easy lookups may overwhelm a smaller set of missing-as-zero failures or incorrect medication carryovers. Reporting performance by error type and task class makes the result more useful for repair and for judging the intended use. A single percentage cannot show whether the most important distinctions were preserved.

Externality also needs a precise meaning. Review by someone who did not write the interface is helpful, but independence can be weakened if the same person selected the cases, supplied reference answers, and adjudicated every disagreement. A report should describe these roles rather than use “externally validated” as an unexplained label.

The interface’s scientific source fidelity and a user’s clinical decision are different endpoints. A correct answer may improve access to evidence without improving prescribing outcomes. Demonstrating a clinical benefit would require an appropriate study of decisions and patient outcomes, including unintended effects. That claim should not be inferred from a successful lookup benchmark or usability session.

The protocol should preserve failures as reproducible cases. Once an error is repaired, the corrected case can enter internal regression testing while remaining identified as previously observed. A new independent assessment should include fresh cases, so the development process does not gradually turn the entire external set into familiar training material.

Limitations

This article does not provide a completed sampling frame, finalized scoring weights, recruited participants, or an external validation result. These are necessary components of a future study. The proposed dimensions may also need adaptation for different professional and public audiences.

Testing cannot exhaust every possible question or conversation. Results depend on the sampled tasks, reference evidence, users, and release. A successful evaluation would support confidence within that scope while leaving clinical judgment, evidence uncertainty, and individual treatment decisions outside the benchmark’s conclusions.



Conclusion

Evaluate Ask as both an evidence retrieval system and a conversation that people must understand. Preserve independence, test difficult context and missingness cases, and report errors by their meaning. Existing internal QA is a foundation for this work; it is not a substitute for the proposed external assessment.

References

1. Vennard O, Stewart C, Tolia M, Soiza RL, Myint PK. Anticholinergic Medication Burden Scales: A Systematic Review. J Am Geriatr Soc. 2026;74(6):1771-1784. [Source](#)
2. Chew ML, et al. Anticholinergic activity of 107 medications commonly used by older adults. J Am Geriatr Soc. 2008;56:1333-1341. [Source](#)
3. Carnahan RM, et al. The Anticholinergic Drug Scale as a measure of drug-related anticholinergic burden: associations with serum anticholinergic activity. J Clin Pharmacol. 2006;46:1481-1486. [Source](#)
4. Rudolph JL, et al. The anticholinergic risk scale and anticholinergic adverse effects in older persons. Arch Intern Med. 2008;168:508-513. [Source](#)
5. Lisibach A, Benelli V, Ceppi MG, et al. Quality of anticholinergic burden scales and their impact on clinical outcomes: a systematic review. Eur J Clin Pharmacol. 2021;77(2):147-162. [Source](#)
6. UnifiedACB. Validation & QA. Published internal benchmark scope and limits; RC12 release certification figures. [Source](#)

Explore the evidence

[Ask UnifiedACB](#) about a medication or source, browse the [Medication Library](#), or read the [methodology](#).

Disclosure

Funding: None. Conflict of interest: None declared for the UnifiedACB educational project.

Educational evidence summary. This content does not provide patient-specific medical advice and does not replace clinical judgment, medication reconciliation, or prescribing guidance.