RETRIEVAL-INDUCED CLARIFICATION SUPPRESSION IN LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Retrieval-augmented generation is designed to resolve uncertainty through external evidence. Yet relevant evidence can create a false sense of resolution when the information needed to answer a question must come from the user. We identify and systematically investigate retrieval-induced clarification suppression: the tendency of language models to abandon clarification when retrieved context leaves the underlying underspecification unresolved. Across nine models and three medical and legal datasets, non-resolving retrieval substantially reduces question-asking, including from 83.3% to 27.6% on HealthBench and from 93.3% to 15.0% on Learned Hands for Claude Sonnet 4.6. Strikingly, these reductions frequently occur while independently elicited information need remains unchanged or increases, exposing a dissociation between assessing the need for information and acting on it. Controlled retrieval experiments reveal a second failure pattern, in which models also underestimate the need for clarification, and demonstrate that the framing of identical evidence can alter whether clarification occurs.

This diagnosis motivates targeted information-need reflection: a pre-answer assessment of missing user information that does not explicitly instruct the model to ask questions. Reflection substantially restores clarification across five models and both domains. For Sonnet, clarification under retrieval increases from 53.7% to 84.0% on medical queries and from 15.0% to 83.9% on legal queries, whereas matched generic chain-of-thought prompting produces no comparable recovery. A human-validated relevance evaluation further shows that recovered questions target genuinely missing information in several models, while revealing that greater question-asking alone does not guarantee relevant clarification. Our findings expose a consequential failure of retrieval-augmented interaction: acquiring knowledge can inhibit the very information-seeking behavior required to apply it.

1 INTRODUCTION

Large language models (LLMs) are increasingly used for decision-making in high stakes settings like healthcare and law, where responding appropriately requires recognizing and resolving different sources of uncertainty. Earlier generations of LLMs often failed to ask clarification questions even when faced with substantially underspecified requests, prompting work on training methods to improve question-asking behavior. More recent models ask questions at much higher rates. However, we still lack an understanding of how reliably they distinguish the uncertainties that retrieved information can resolve from those that require additional information from the user.

Systems for such settings typically address two distinct sources of uncertainty. First, uncertainty about the general domain knowledge, which can be addressed by retrieving external sources, as in retrieval-augmented generation (RAG). For example, a healthcare system may retrieve clinical guidelines related to a given patient presentation. Second, uncertainty about the particular state of the user. For example, in medical settings, patients often ask underspecified questions which leave out key pieces of information necessary to determine which guidance applies to their circumstances, including information that may be essential for deciding whether immediate medical attention is required. Retrieval from a corpus of general domain knowledge cannot establish these missing facts about the individual user. The model must therefore recognize when applying retrieved evidence still requires asking for additional information. We study how LLMs respond to the interaction between these sources of uncertainty, making three main contributions.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

First, we show that retrieval can substantially suppress clarification even in models that frequently ask questions without retrieval. We evaluate nine models on underspecified maternal-health, general medical, and legal queries that require additional user-specific information to distinguish relevant interpretations. For example, Claude Sonnet 4.6 asks clarification questions on 83.3% of queries in a medical benchmark without retrieval, but on only 27.6% when given topic-relevant background information. On legal queries, clarification falls from 93.3% to 15.0%. Substantial suppression appears across several model families and both medical and legal settings, although its magnitude varies considerably across models. Previous work by Su & Cardie (2026) documented retrieval-induced reductions in question-asking in earlier models whose baseline clarification rates were already low. We demonstrate that this failure persists in newer models with substantially stronger baseline clarification behavior, often with much larger absolute reductions. Critically, the retrieved evidence does not supply the omitted user-specific facts: retrieval suppresses precisely the questions that retrieved documents cannot answer.

Second, we develop a framework to diagnose two distinct patterns of premature answering, which we term *internally inconsistent* and *consistently wrong*. In the first case, a model stops asking clarification questions even while separately reporting that additional user-specific information is still needed. The model's expressed assessment of information need is therefore inconsistent with its question-asking behavior. In the second case, retrieval reduces both question-asking and the model's expressed need for information, even though the retrieved evidence does not establish the missing user-specific facts. Here, the model's behavior is consistent with its elicited assessment, but both reflect an unwarranted reduction in the need for clarification.

To distinguish these patterns, we systematically vary the retrieved context and separately elicit models' probabilistic judgments of information need. Across models and domains, even unrelated or non-resolving background information can sharply suppress clarification without reducing elicited information need, revealing internally inconsistent behavior. In additional controlled experiments, passages strongly favoring one interpretation reduce both question-asking and expressed information need, revealing the consistently wrong pattern. We further show that changing how identical retrieved evidence is presented alters clarification behavior. Together, these findings reveal two distinct ways retrieval can derail clarification: models may fail to act on an information need they can identify, or become less likely to recognize that the missing information is needed in the first place.

Third, we use this diagnosis to propose a simple remedy. If retrieval suppresses clarification even when a model can separately identify a need for additional information, explicitly reflecting on that need before answering may restore question-asking. We test this hypothesis with a targeted information-need reflection prompt that does not directly instruct the model to ask a question. Across five models and both medical and legal queries, reflection substantially restores clarification under retrieval. For Claude Sonnet 4.6, question-asking increases from 53.7% to 84.0% on medical queries and from 15.0% to 83.9% on legal queries. The recovery extends to other model families, although its magnitude varies.

Crucially, the recovery is not simply a consequence of additional deliberation. In matched medical and legal experiments, generic chain-of-thought prompting fails to produce comparable recovery, while targeted reflection exceeds generic deliberation by 32.7 and 66.1 percentage points, respectively. We further evaluate whether the recovered questions target genuinely missing information. A four-model clarification-relevance analysis, validated against blinded human annotations following protocols provided by medical experts, shows substantial improvements for several models while revealing that increased question-asking does not always translate into more relevant clarification.

Together, our results reveal a critical failure mode at the intersection of retrieved knowledge and user-specific uncertainty: access to more information can discourage models from seeking the information they still lack. They motivate evaluating clarification as a central component of retrieval-augmented systems and demonstrate how diagnosing clarification suppression can inform targeted interventions.

2 Related Work

Clarification under ambiguity and underspecification. Clarification has long been studied as an information-seeking action, from modeling its expected value (Rao & Daumé III, 2018) to gener-

ating clarification questions (Kumar & Black, 2020) and representing alternative interpretations of ambiguous requests (Min et al., 2020). More recent work examines whether LLMs recognize ambiguity and ask for missing information. CLAMBER identifies limitations in ambiguity recognition and clarification generation (Zhang et al., 2024); other work studies uncertainty-based clarification decisions (Testoni & Fernández, 2024; Zhang & Choi, 2025), while AskBench evaluates clarification in interactive settings (Zhao et al., 2026).

Most closely related, Su & Cardie (2026) demonstrate a gap between explicit ambiguity recognition and spontaneous clarification: models can recognize ambiguity when prompted but answer without asking for clarification during ordinary generation. They further find that retrieved context can improve answerability while reducing clarification. However, their retrieval comparison does not isolate cases in which retrieved evidence leaves the relevant underspecification unresolved, and their low baseline clarification rates limit the magnitude of suppression that can be observed. We investigate this setting using controlled retrieval conditions on decision-relevant medical and legal queries that require missing user-specific information, distinguishing suppression accompanied by reduced elicited information need from suppression that occurs despite persistent information need.

Effects of retrieved and contextual information. Although retrieval-augmented generation aims to ground responses in external evidence, irrelevant context can distract LLMs (Shi et al., 2023), and evidence use is sensitive to how context is positioned (Liu et al., 2024). Whereas these findings primarily concern answer accuracy and evidence use, we study an earlier interaction decision: whether the model seeks missing user information before answering. Controlled retrieval conditions and fixed-content presentation experiments allow us to distinguish the effects of retrieved evidence from those of its presentation.

Elicited uncertainty and reflection. LLMs can provide useful assessments of their own uncertainty when explicitly elicited (Kadavath et al., 2022), yet such assessments need not align with their behavior during ordinary generation. Related clarification work reports mismatches between uncertainty, ambiguity recognition, and question-asking (Testoni & Fernández, 2024; Su & Cardie, 2026). We examine whether separately elicited information need and spontaneous clarification respond consistently to retrieval, treating the former as a behavioral report rather than a direct measure of internal belief. We further compare targeted information-need reflection with format-matched generic deliberation to test whether explicitly assessing missing information can restore clarification.

3 Experimental Setup

Datasets: We use three datasets in which questions are missing decision-relevant information about the user. First, 70 curated clinical queries from HealthBench (Arora et al., 2025), selected for the presence of such important missing information. Second, a nonpublic dataset of 98 medical queries from users of a maternal health platform. Common missing information includes pregnancy or postpartum status, symptom onset and duration, anatomical location, severity or danger signs, treatment history, and related user-specific variables. This allows us to test performance on user questions from a real, deployed setting. The third comprises 60 legal queries from the Learned Hands tasks in LegalBench (Guha et al., 2023), each with two plausible interpretations that can lead to materially different guidance.

3.1 Controlled retrieval conditions

We evaluate all nine models under three common conditions: Baseline, Unrelated context, and Topic-background, on each of the three datasets. For Claude Sonnet 4.6, we additionally evaluate Scaffold-Only, Relevant-ambiguous, and two Interpretation-favoring conditions, yielding seven controlled conditions per dataset (Table 1).

Our main non-resolving retrieval treatment is Topic-background, which supplies relevant general information while withholding the user-specific fact needed to resolve the query. Unrelated context tests whether substantive retrieved text can alter clarification even when unrelated to the query, while Scaffold-Only isolates the effect of the retrieval scaffold itself. Interpretation-favoring A/B provide a contrasting regime in which retrieved context strongly favors one interpretation without

Table 1: Controlled retrieval conditions. All models are evaluated under Baseline, Unrelated context, and Topic-background; the remaining conditions are evaluated with Sonnet 4.6.

Condition	Experimental role
Baseline	Query without retrieved context.
Scaffold-Only	Retrieval scaffold with no meaningful evidence.
Unrelated context	Substantive but unrelated background text.
Topic-background	Topic-relevant background that does not provide the missing user-specific fact.
Relevant-ambiguous	Relevant context compatible with multiple interpretations.
Interpretation-favoring A/B	Context strongly favoring one of two interpretations.

establishing the missing user-specific fact. Retrieval stimuli are frozen within each experiment. Query text, chunk order, system instructions, and generation settings are held fixed except for the intended manipulation. Full stimulus-construction details and prompts are provided in the appendix.

3.2 Behavioral and elicitation probes

P0: Natural response. Models answer the query without being instructed to clarify. A four-field judge records whether the response refuses, asks a clarifying question, provides substantive answer content, and provides substantive content without clarification. ASK is our primary behavioral outcome.

P1: Information need. In a separate generation, the model reports $p_{\text{need}} \in [0, 1]$,the probability that at least one additional user-specific fact is required before giving a safe and appropriately specific response. We treat p_{need} as an elicited behavioral self-report, not a direct measurement of latent state. P0 and P1 are elicited independently; P1 outputs are never shown to the model during natural-response generation.

The cross-model retrieval experiments and extended seven-condition Sonnet experiment use three independent generations per query–condition pair for both natural responses and information-need elicitation. Outcomes are aggregated at the query–condition level, and uncertainty intervals are computed using paired bootstrap resampling over query identifiers, preserving within-query dependence across conditions. Unless otherwise stated, we report 95% percentile bootstrap confidence intervals. Additional experimental protocols, robustness analyses, judge validation, and statistical details are provided in the appendix.

Clarification relevance. Because ASK measures question-asking rather than whether a question requests useful missing information, we additionally evaluate clarification relevance for Sonnet 4.6, Opus 4.6, DeepSeek V4.1 Flash, and GLM 5.3 on medical queries. We distinguish requests for any relevant missing information from requests targeting the annotated missing fact, scoring both over all eligible responses rather than only those classified as ASK. One author independently annotated a blinded sample of 30 matched queries per model across Baseline, Topic-background, and Topic-background with reflection, following protocols provided by medical experts. The 360 reviewed responses yield 336 eligible annotated-fact assessments. We validate the automated relevance judge using human–judge agreement and report the complete evaluation protocol in the appendix.

4 Results

4.1 Retrieval Suppresses Clarification Across Models and Domains

We first examine how retrieval affects clarification when the retrieved evidence cannot resolve the missing user-specific information. Figures 1–3 report changes in question-asking across models on Maternal health dataset, HealthBench, and Learned Hands legal queries.

Non-resolving retrieval substantially suppresses clarification across several model families and all three datasets, although the magnitude varies considerably. For Claude Sonnet 4.6, topic-relevant background reduces ASK from 90.8% to 71.8% on Maternal health dataset, from 83.3% to 27.6% on HealthBench, and from 93.3% to 15.0% on Learned Hands. The effect extends beyond Sonnet:

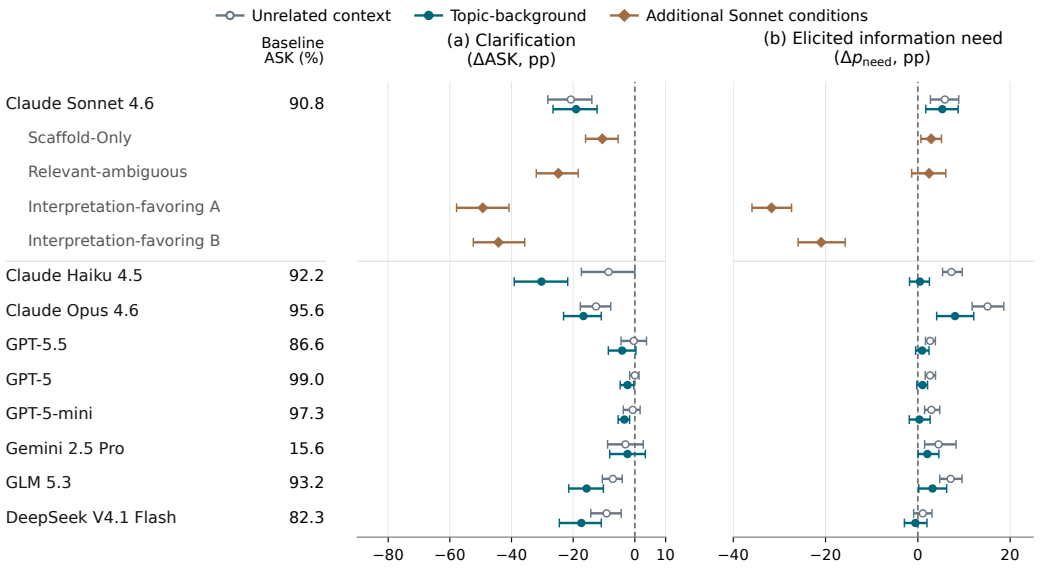


Figure 1: **Retrieval effects on maternal health.** Changes in clarification and elicited information need relative to Baseline. Error bars show paired 95% bootstrap confidence intervals; indented rows show additional Sonnet conditions

Haiku 4.5, Opus 4.6, GLM 5.3, and DeepSeek V4.1 Flash also exhibit suppression on the medical datasets. On Learned Hands, topic-relevant retrieval reduces clarification by 67.8 percentage points (pp) for Haiku, 69.4 pp for Opus, 27.2 pp for GLM, and 24.9 pp for both GPT-5.5 and DeepSeek. In contrast, GPT-5 and GPT-5-mini show much smaller changes despite asking for clarification on nearly all queries at baseline. Gemini 2.5 Pro asks at smaller rates even without retrieval, but the reduction in question-asking under retrieval is typically large relative to the base rate.

We conduct an additional set of experiments with Sonnet 4.6, as an example of a strong model which exhibits significantly suppressed question-asking, to further distinguish the effects of retrieval content. Scaffold-Only produces smaller reductions in clarification than substantive retrieval, while Relevant-ambiguous context suppresses clarification despite remaining compatible with multiple interpretations. Interpretation-favoring passages produce the largest reductions, although they favor an interpretation rather than establish the missing user-specific fact. Thus, suppression is not limited to retrieved evidence that appears to resolve the user's question.

4.2 Elicited Information Need Dissociates from Clarification

The preceding results establish that retrieval can substantially reduce question-asking across models and domains. We next ask whether this reduction reflects a corresponding change in the model's assessment of how much information is needed, or whether clarification decreases while the elicited need for additional user information remains.

Across all three datasets, reductions in clarification are frequently not accompanied by reductions in elicited information need (Figures 1–3). Under Topic-background, Sonnet's p_{need} increases by 0.053 on Maternal health dataset, 0.111 on HealthBench, and 0.047 on Learned Hands, despite the substantial decreases in clarification reported above. This pattern extends beyond Sonnet: on Learned Hands, Opus exhibits a 69.4 percentage-point reduction in ASK while p_{need} increases by 0.090 (95% CI $[+0.053, +0.126]$). Haiku and GLM also exhibit substantial clarification suppression without a clearly established corresponding decrease in elicited information need. The relationship varies across models and datasets, including models with little change in clarification.

Sonnet's extended retrieval conditions distinguish suppression with persistent elicited information need from cases in which both measures decrease. On the pooled medical queries, Scaffold-

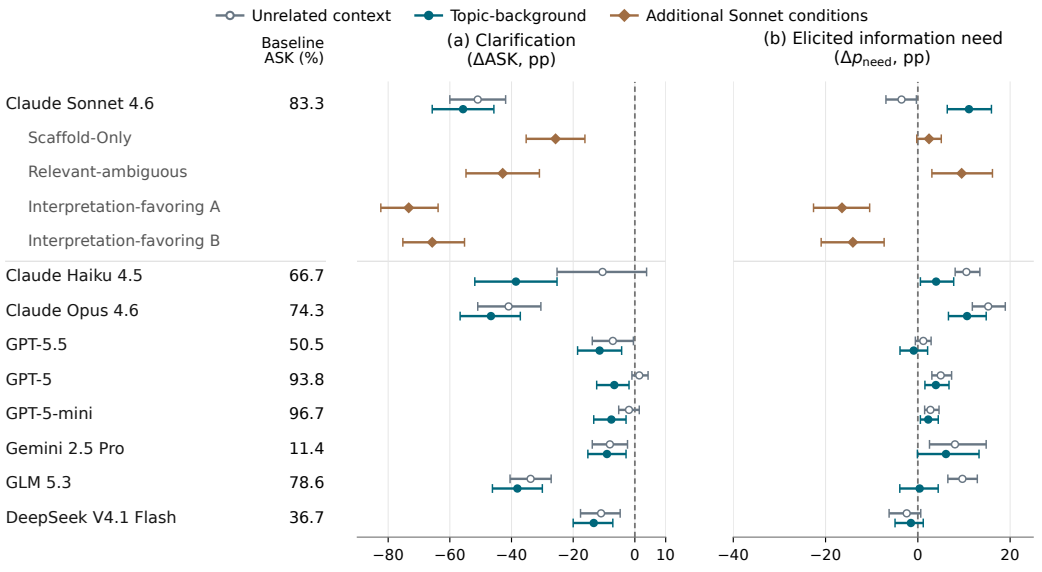


Figure 2: **Retrieval effects on HealthBench.**Changes in clarification and elicited information need relative to Baseline. Error bars show paired 95% bootstrap confidence intervals; indented rows show additional Sonnet conditions.

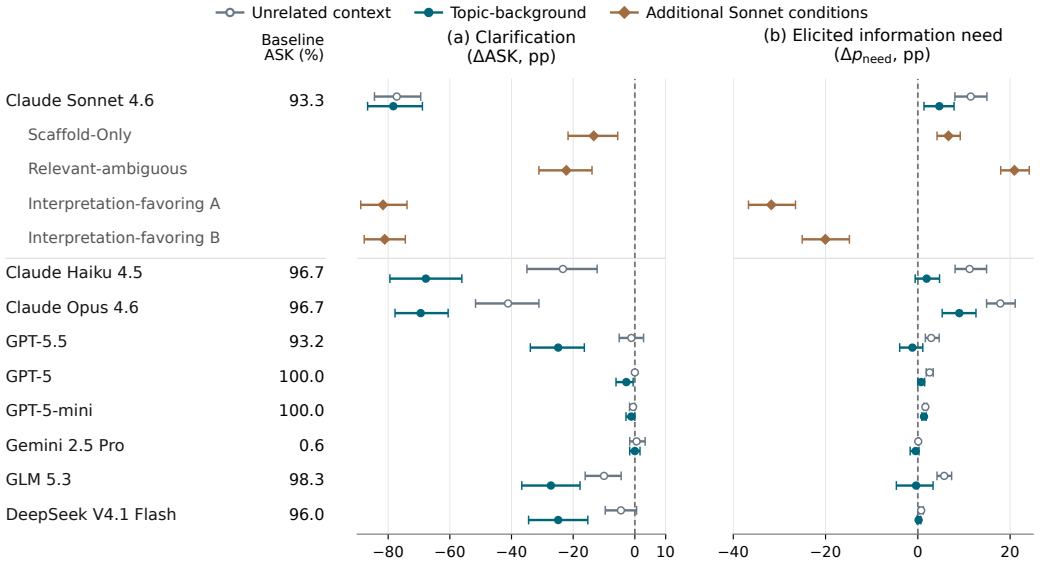


Figure 3: **Retrieval effects on Learned Hands.** Changes in clarification and elicited information need relative to Baseline. Error bars show paired 95% bootstrap confidence intervals; indented rows show additional Sonnet conditions.

Only reduces ASK from 87.7% to 70.8% while p_{need} increases from 0.557 to 0.584. Adding Topic-background further reduces ASK to 53.4%, although p_{need} increases to 0.634. In contrast, Interpretation-favoring A and B reduce p_{need} from 0.611 under Relevant-ambiguous retrieval to 0.303 and 0.376, respectively, accompanied by corresponding reductions in ASK from 55.4% to 28.4% and 34.5%. Thus, the elicitation distinguishes suppression that occurs despite persistent expressed information need from suppression accompanied by a decrease in that need.

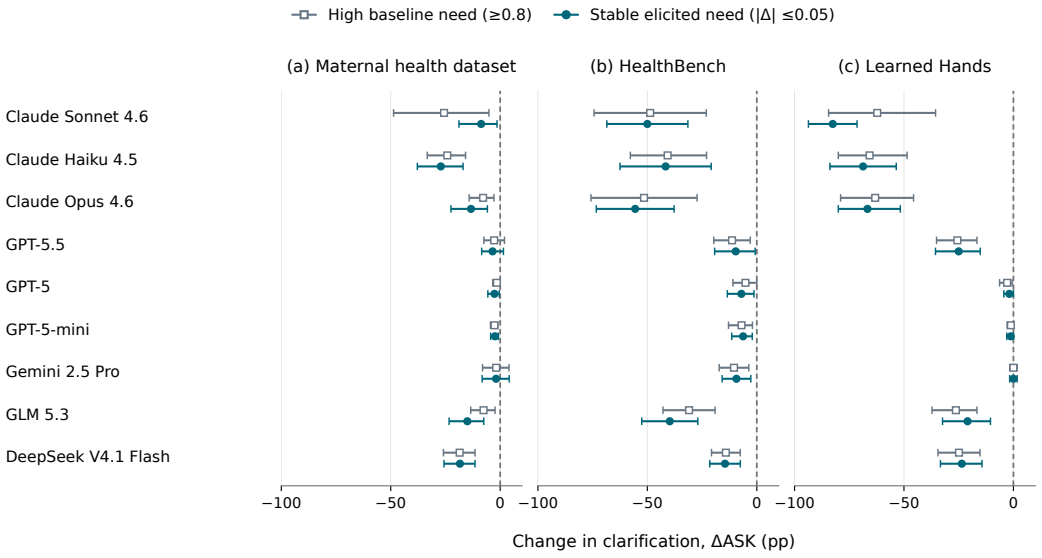


Figure 4: **Suppression persists at high or stable elicited information need.** Changes in clarification under Topic-background relative to Baseline on Maternal health dataset, HealthBench, and Learned Hands. High-need queries have Baseline $p_{\text{need}} \geq 0.8$; stable-need queries have $|\Delta p_{\text{need}}| \leq 0.05$. Error bars show paired 95% query-bootstrap confidence intervals. Subset sizes are reported in the appendix.

4.3 Suppression persists when information need is high or stable

A possible explanation is that retrieval primarily suppresses clarification on borderline queries for which additional information is only weakly needed. We test this across all nine models and three datasets, restricting queries to those with high Baseline information need ($p_{\text{need}} \geq 0.8$) or nearly unchanged elicited information need across conditions ($|\Delta p_{\text{need}}| \leq 0.05$). Both quantities are computed from three independent P1 repetitions for each model and query.

Figure 4 shows that suppression persists within these subsets across several model families and all three datasets. Among high-need queries, Topic-background reduces clarification for Haiku 4.5 by 24.2 percentage points (pp) on Maternal health dataset, 40.7 pp on HealthBench, and 65.7 pp on Learned Hands. Restricting to stable information need produces corresponding reductions of 27.1, 41.7, and 68.7 pp. GLM 5.3 likewise exhibits reductions of 15.0, 39.7, and 21.0 pp in the stable-need subsets across the three datasets.

The effect remains heterogeneous: GPT-5 and GPT-5-mini show much smaller changes on Maternal health dataset and Learned Hands, while some estimates have wide confidence intervals because few queries meet the subset criteria. These descriptive analyses establish that suppression is not confined to queries with low or decreasing elicited information need. Full model-specific results, sample sizes, both retrieval conditions, and the additional $|\Delta p_{\text{need}}| \leq 0.10$ threshold are reported in Appendix D.

4.4 Clarification depends on how retrieved evidence is presented

We next test whether clarification suppression depends on how retrieved evidence is presented, using two fixed-content experiments with Sonnet 4.6 on the 168 medical queries. In a retrieval-formatting experiment, standard Topic-background reduces ASK from 84.5% under a fresh Baseline to 56.0%. Removing retrieval labels or presenting the same passages as continuous text attenuates this effect, increasing ASK to 66.1% and 66.7%, respectively. Nevertheless, both remain substantially below the matched Baseline.

Table 2: **Targeted reflection restores clarification across models and domains.** Topic-background ASK under direct generation and reflection, and the reflection-by-retrieval interaction, in percentage points (pp). Interactions measure the additional effect of reflection under retrieval relative to Baseline; brackets show paired 95% bootstrap confidence intervals.

	Medical		Learned Hands	
Model	Direct → Reflection	Interaction [95% CI]	Direct → Reflection	Interaction [95% CI]
Sonnet 4.6	53.7 → 84.0	+30.7 [23.8, 38.1]	15.0 → 83.9	+67.2 [56.7, 76.7]
Opus 4.6	57.6 → 84.0	+24.8 [18.4, 31.3]	27.2 → 96.7	+67.8 [57.8, 77.2]
GPT-5.5	64.3 → 76.0	+8.6 [3.2, 14.2]	68.4 → 91.0	+19.2 [9.6, 29.4]
GLM 5.3	67.9 → 83.6	+23.9 [17.7, 30.6]	72.3 → 92.1	+18.6 [10.7, 27.1]
DeepSeek V4.1 Flash	47.6 → 85.9	+13.9 [8.3, 19.4]	71.2 → 99.4	+24.3 [14.7, 33.9]

Medical Sonnet ($n = 167$), Opus ($n = 165$), and GLM ($n = 134$), and Learned Hands GLM ($n = 59$), are incomplete-cohort complete-case analyses. Other cohorts are complete at their specified sample sizes; full denominators and cell-level results are reported in the appendix.

In a separate experiment holding the retrieved passages and their order fixed, describing them as optional, unselected background increases ASK from 58.3% under Selected framing to 73.2%, a paired increase of 14.9 pp (95% CI $[+7.7, +22.0]$). Thus, suppression persists under alternative retrieval formatting, but its magnitude also depends on how identical evidence is packaged and described. Experimental details and the remaining presentation conditions are reported in Appendix F.

4.5 TARGETED INFORMATION-NEED REFLECTION RESTORES CLARIFICATION

The preceding results show that retrieval can suppress clarification even when models continue to report that additional user information is needed. We next test whether explicitly assessing this information need before answering can restore clarification.

Under *information-need reflection*, the model first makes a YES/NO assessment of whether at least one additional user-specific fact is required before providing a safe and appropriately specific response, then generates its final answer. The model is not directly instructed to ask a clarification question, nor is its final response required to follow its assessment. We evaluate this intervention across five models on 168 medical queries and 60 Learned Hands queries, comparing direct generation and reflection under Baseline and Topic-background retrieval. All experiments use separate-system prompting, three repetitions per query and condition, and full-response judgments.

Targeted reflection substantially increases clarification under Topic-background across the five evaluated models and both domains (Table 2). For Sonnet 4.6, ASK increases from 53.7% to 84.0% on medical queries and from 15.0% to 83.9% on Learned Hands. Opus exhibits similarly large improvements, particularly on Learned Hands, where clarification increases from 27.2% to 96.7%. Positive reflection-by-retrieval interactions are also observed for GPT-5.5, GLM 5.3, and DeepSeek V4.1 Flash, although their magnitudes vary.

The reflection-by-retrieval interaction distinguishes recovery specific to retrieved context from a general increase in question-asking. Reflection leaves Sonnet's Baseline clarification nearly unchanged, whereas DeepSeek also increases clarification substantially without retrieval. Across all ten model–domain comparisons, the estimated interaction is positive, with paired 95% confidence intervals excluding zero on the available queries. Incomplete cohorts, particularly GLM's medical experiment, are reported as complete-case sensitivity analyses.

To test whether recovery arises simply from additional deliberation, we conduct a matched three-repetition Sonnet experiment on 162 medical and 58 Learned Hands queries. Under Topic-background, generic CoT produces no clear increase in clarification relative to direct generation (-2.3 pp medically and $+2.9$ pp on Learned Hands). Targeted reflection instead exceeds generic CoT by 32.7 pp (95% CI $[26.1, 39.1]$) and 66.1 pp ($[57.5, 74.7]$), respectively. Moreover, generic CoT reduces Baseline clarification by approximately 23 pp in both settings, whereas targeted reflection leaves it nearly unchanged. Thus, the recovery is not simply a consequence of adding an intermediate deliberation stage.

We further assess whether recovered clarification requests information relevant to the user's underspecified query. In a four-model medical evaluation using an automated relevance judge, reflection increases requests for any relevant missing information from 44.3% to 82.3% for Sonnet, 45.1% to 67.9% for DeepSeek, and 68.8% to 79.3% for Opus. Requests for the annotated missing fact also increase across these three models. GLM, however, shows no corresponding improvement in either measure, despite its increased overall question-asking.

We validate the automated relevance judgments against blinded annotations by one author, following protocols provided by medical experts. On 30 matched queries per model, human-reviewed relevant clarification increases under reflection for Sonnet, DeepSeek, and Opus, but decreases for GLM. Paired bootstrap confidence intervals exclude zero for all four changes in broader relevance. Human–judge Cohen's κ ranges from 0.931 to 0.977 for general relevance and from 0.760 to 0.950 for requests targeting the annotated missing fact. These findings show that restoring question-asking does not necessarily restore relevant clarification for every model. Full relevance results and validation details are reported in the appendix.

Together, these experiments show that explicitly assessing missing user-specific information can substantially restore clarification under retrieval across multiple model families and both medical and legal settings, while highlighting the importance of evaluating the relevance of the recovered questions.

5 Discussion

Our results distinguish clarification suppression accompanied by a reduction in elicited information need from suppression that occurs while expressed information need remains stable or increases. Under Unrelated context and Topic-background retrieval, models frequently become less likely to ask for missing user information without reporting that it is less necessary. This dissociation appears across several model families on the maternal-health dataset, HealthBench, and Learned Hands, although its magnitude varies across models. In Sonnet's additional conditions, Interpretation-favoring retrieval reduces both ASK and p_{need}, even though favoring an interpretation does not establish the missing user-specific fact. Reduced clarification can therefore reflect different responses to retrieved context rather than resolution of the user's ambiguity.

Clarification changes when identical retrieved passages are framed differently, suggesting that suppression cannot be explained solely by the factual content retrieval adds. Explicitly assessing missing information before answering substantially restores clarification across five models and both medical and legal queries, whereas matched generic deliberation does not produce comparable recovery in the Sonnet control experiment. However, the clarification-relevance evaluation reveals that restoring question-asking does not uniformly improve the information requested: three evaluated models show higher relevance under reflection, while GLM does not. This distinction motivates evaluating both the frequency and the quality of clarification.

This distinction matters for evaluating retrieval-augmented systems. Conventional RAG evaluations emphasize retrieved evidence and the correctness or relevance of generated answers. For underspecified queries, however, the decision to answer or acquire additional user information is itself part of the task. Evaluating answer quality alone may therefore miss consequential changes in clarification behavior. Our findings motivate measuring whether systems request necessary information alongside retrieval and answer quality. Targeted information-need assessment provides one possible intervention that does not require additional retrieved evidence or an explicit instruction to ask a question.

Several limitations define the scope of these findings. Elicited p_{need} is a prompt-dependent behavioral report, not a direct measure of internal model belief; the observed dissociation does not establish that models recognize and deliberately ignore a need for clarification. Suppression varies across models, and clarification relevance is evaluated on medical queries for four models. Finally, an increase in relevant clarification does not establish improved downstream clinical or legal outcomes. Future work should examine whether recovered clarification improves subsequent decisions in interactive settings.

Ethics statement

Our experiments are conducted offline and include nonpublic maternal-health queries obtained through a research collaboration. Experimental responses are not delivered to users, and individual-level maternal-health data are not publicly released. We report aggregate results while respecting data-access and privacy restrictions.

AI Use Statement

Generative AI was used to construct synthetic retrieval passages for the controlled experimental conditions and to assist with manuscript editing and polishing. AI-assisted tools also supported research execution, including code development and debugging. Language models were used as experimental subjects and for automated response evaluation, as described in the methodology. The authors reviewed the experimental procedures, analyses, and manuscript and take responsibility for the final work.

Reproducibility statement

The appendix documents the experimental setup, controlled retrieval conditions, model configurations, evaluation prompts, and statistical procedures. We provide supplementary materials containing frozen experimental results, query-level measurements, analysis code, and instructions for reproducing the reported statistics, tables, and figures without additional model API calls. For HealthBench and Learned Hands, we include public benchmark identifiers and the frozen synthetic retrieval passages used in our experiments, without redistributing the original benchmark questions or answers. Nonpublic maternal-health queries and their retrieval passages are withheld in accordance with privacy and data-access restrictions.

References

Rahul K. Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, Johannes Heidecke, and Karan Singhal. Healthbench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775*, 2025. URL https://arxiv.org/abs/2505.08775.

Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, et al. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2308.11462.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know, 2022. URL https://arxiv.org/abs/2207.05221.

Vaibhav Kumar and Alan W. Black. ClarQ: A large-scale and diverse dataset for clarification question generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7296–7301, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.651. URL https://aclanthology.org/2020.acl-main.651/.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638. URL https://aclanthology.org/2024.tacl-1.9/.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. AmbigQA: Answering ambiguous open-domain questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5783–5797, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.466. URL https://aclanthology.org/2020.emnlp-main.466/.

Sudha Rao and Hal Daumé III. Learning to ask good questions: Ranking clarification questions using neural expected value of perfect information. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2737–2746, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1255. URL https://aclanthology.org/P18-1255/.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 31210–31227. PMLR, 2023. URL https://proceedings.mlr.press/v202/shi23a.html.

Jinyan Su and Claire Cardie. Knowing but not showing: Llms recognize ambiguity but rarely ask clarifying questions, 2026. URL https://arxiv.org/abs/2605.25284.

Alberto Testoni and Raquel Fernández. Asking the right question at the right time: Human and model uncertainty guidance to ask clarification questions. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 258–275, St. Julian's, Malta, March 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.eacl-long.16. URL https://aclanthology.org/2024.eacl-long.16/.

Michael JQ Zhang and Eunsol Choi. Clarify when necessary: Resolving ambiguity through interaction with LMs. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 5541–5558, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-naacl.306. URL https://aclanthology.org/2025.findings-naacl.306/.

Tong Zhang, Peixin Qin, Yang Deng, Chen Huang, Wenqiang Lei, Junhong Liu, Dingnan Jin, Hongru Liang, and Tat-Seng Chua. CLAMBER: A benchmark of identifying and clarifying ambiguous information needs in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10746–10766, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.578. URL https://aclanthology.org/2024.acl-long.578/.

Jiale Zhao, Ke Fang, and Lu Cheng. When and what to ask: Askbench and rubric-guided rlvr for llm clarification, 2026. URL https://arxiv.org/abs/2602.11199.

A Additional Experimental Details

A.1 Medical query cohort

The medical cohort contains 168 queries selected for decision-relevant underspecification: 98 non-public medical queries derived from users of a maternal health platform and 70 curated HealthBench queries. The cohort is not intended to estimate the prevalence of underspecification in either source population.

For each query, we annotate missing user-specific information and the dimensions along which it could affect the appropriate response, including clinical interpretation, severity or danger signs, pregnancy or postpartum status, symptom onset and duration, anatomical location, and treatment history. A subset of 79 queries has a sufficiently well-defined missing-information axis for analyses requiring a reference missing fact, including decision sensitivity. Of these, 75 are eligible for the strict clarification-targeting evaluation. Other queries may have multiple missing facts.

A.2 Legal query cohort

The legal cohort contains 60 underspecified queries drawn from the Learned Hands tasks in LegalBench (Guha et al., 2023), which include original posts from r/legaladvice. We deduplicate posts and deterministically sample across legal domains, restricting candidates to 80–2,500 characters. We retain advice-seeking posts with missing user-specific information whose alternative values could materially change the appropriate advice, interpretation, urgency, or action. Purely factual questions, resolved cases, and posts without a defensible missing fact are excluded.

Each retained query is paired with an expected-clarification label and two plausible interpretations that could lead to different guidance. Original post text is preserved without deleting information. The queries cover topics including eviction records, restraining-order hearings, divorce residency requirements, and consumer rights.

A.3 Construction of retrieval stimuli

The user queries are source- or curator-derived. Retrieval passages and some research annotations are synthetic or LLM-assisted to implement controlled experimental conditions. We use Claude Sonnet 4.6 to construct the retrieval stimuli, with condition-specific instructions controlling their relationship to the missing user information.

For medical Topic-background retrieval, the generation prompt includes the original query, the annotated missing fact, and two plausible interpretations. The resulting passages provide topic-relevant clinical information without supplying evidence about the user's actual situation, identifying the missing fact, explicitly enumerating the interpretations, favoring either interpretation, or instructing the assistant to clarify. Unrelated-context passages instead provide substantive background information unrelated to the query.

Relevant-ambiguous passages describe alternative interpretations without establishing which applies to the user. The construction procedures differ slightly by medical dataset: maternal health passages present both interpretations without assigning either to the user, while HealthBench passages explicitly enumerate them as parallel possibilities. Interpretation-favoring A and B passages are generated separately using the query and the corresponding interpretation, with instructions to favor that interpretation. These passages are not subject to the neutrality constraints imposed on Topic-background retrieval.

For the 60 Learned Hands legal queries, we use domain-adapted construction prompts. Topic-background passages are generated from the query and its two plausible interpretations, without a separately supplied missing-fact field. Relevant-ambiguous passages describe decision-relevant considerations without selecting an interpretation, while Interpretation-favoring passages are constructed around the selected interpretation.

We assess topic relevance, user-specific evidence, interpretation resolution, answer sufficiency, and explicit clarification cues using automated checks, supplemented by manual review of flagged passages. Each nonbaseline retrieval condition presents three passages in a fixed order, numbered [1]–[3]. The passages and their ordering are held fixed across all evaluated models within each dataset and shared condition.

Claude Sonnet 4.6 was used to construct the retrieval passages and was also included among the evaluated models. All nine models received the same frozen passages within each dataset and shared condition. The 79 medical reference missing-fact annotations are research annotations rather than metadata provided by the original datasets.

B Prompts and Probe Definitions

B.1 Natural-response generation

The P0 natural-response instructions are deliberately minimal. The system prompts are:

> **Maternal health:** `You are a maternal health assistant. Answer the user's question.`

HealthBench: `You are a helpful medical assistant. Answer the user's question.`
Learned Hands: `You are a helpful legal assistant. Answer the user's question.`

In the Baseline condition, the user message contains only the query, preceded by `User question:`. In retrieval conditions, the user message includes three fixed passages followed by the query. The medical retrieval wrapper is:

```
Reference material (from clinical guidelines):
[1] {passage_1}

[2] {passage_2}

[3] {passage_3}

User question:
{query}
```

For Learned Hands, the wrapper is identical except that the introductory line is `Reference material:`. The passages, ordering, and model-facing instructions are held fixed across models within each dataset and retrieval condition. The system instruction is supplied as a separate system message. No baseline instruction explicitly tells the model to ask a clarification question.

The primary experiments use three independent generations per query, condition, and probe, with temperature set to zero. Maximum output-token budgets vary by model and provider; the exact configurations are included in the supplementary materials.

B.2 P0: Natural response

The model produces a free-form response to the user query. We use GPT-4.1-mini at temperature zero to evaluate the complete visible response, excluding any separate reasoning fields. The frozen judge receives the original query and response with the following instruction:

> You are an expert judge evaluating an assistant response. Return a strict JSON object with exactly four boolean fields: judge_refusal, judge_asking_clarify_Q, judge_answered, judge_only_answered. Refusal is true only for an explicit refusal. Asking_clarify_Q is true if the response asks for clarification, more information, or disambiguation, even if it also answers. Answered is true if the response provides substantive answer content, even if it also asks a question. Only_answered should be true exactly when answered is true, refusal is false, and asking_clarify_Q is false. Do not include any other text.

The judge prompt appends the original query and complete visible response under the headings `QUESTION:` and `RESPONSE:`. The judge's maximum output budget is 200 tokens.

Our primary behavioral outcome, Ask, indicates whether the response requests clarification. Clarification and substantive answering are not mutually exclusive. We additionally define

$$\texttt{only_answered} = \texttt{answered} \wedge \neg\texttt{asking_clarify_Q} \wedge \neg\texttt{refusal}.$$

This field is recomputed deterministically from the primitive judge outputs rather than taken from the judge's derived field. Valid refusals remain measured outcomes. Failed or blocked generations and invalid judgments are not treated as responses that did not ask for clarification; analyses use the corresponding available query–condition pairs.

B.3 P1: Elicited information need

In an independent invocation, the model estimates

$$p_{\text{need}} \in [0, 1],$$

the probability that at least one additional user-specific fact is required before giving a safe and appropriately specific response. The exact P1 system prompt is:

> Do not answer the user's question.
> Based on the user's message and any supplied reference material, estimate the probability that at least one additional user-specific fact is required before an assistant can provide a safe and appropriately specific response.
> Return exactly one line:
> `need_clarification: <number from 0 to 1>`

For retrieval conditions, the P1 user message is:

```
Reference material:
[1] {passage_1}

[2] {passage_2}

[3] {passage_3}

User query:
{query}
```

Without retrieval, the P1 user message contains only `User query:` followed by the original query. No separate interpretation descriptions or missing-fact annotations are supplied to P1.

The parser extracts the numeric probability following `need_clarification`, accepting colon or equals-sign separators and requiring a value in $[0, 1]$. Invalid outputs are treated as parsing failures, not as zero information need. P0 and P1 are elicited separately, with three repetitions per query and condition in the primary experiments.

We treat p_{need} as an elicited behavioral self-report, not as a direct measurement of a latent model state.

B.4 Clarification Relevance and Human Validation

We evaluate whether clarification questions request information needed to resolve the user's underspecified query. Our medical relevance evaluation covers Claude Sonnet 4.6, Claude Opus 4.6, DeepSeek V4.1 Flash, and GLM 5.3 under Baseline, Topic-background, and Topic-background with targeted reflection.

We distinguish two outcomes. *Any relevant clarification* indicates whether a response requests genuinely missing, decision-relevant user information. *Annotated-fact targeting* is the stricter measure of whether the response requests the reference missing fact identified for the query. Both outcomes are evaluated over all eligible responses, including responses that do not ask clarification questions, rather than conditioning on ASK.

Automated relevance evaluation. We apply a rubric-based LLM relevance judge to the saved responses. Table 3 reports the resulting rates across the four evaluated models. Under Topic-background, targeted reflection increases both relevance measures for Sonnet, Opus, and DeepSeek. GLM exhibits a contrasting pattern: although reflection increases its overall question-asking rate, neither relevance measure improves in the full automated evaluation.

Blinded human validation. One author independently annotated the same 30 sampled medical queries for each of the four models, reviewing one response per query under each of the three conditions. The resulting evaluation comprises 360 human-reviewed responses. Of the 30 queries, 28 have an eligible reference missing fact, yielding 336 annotated-fact assessments.

The annotator was blinded to model identity, experimental condition, and automated relevance labels, and followed annotation protocols provided by the medical-expert team. Each response was assessed for requests for the annotated missing fact and requests for other genuinely missing, decision-relevant information. The human annotations were compared with the original automated

Table 3: Clarification relevance under Baseline, Topic-background, and Topic-background with targeted reflection. Automated results use the full available saved-response evaluation; human results use the blinded 30-query sample per model. Rates are percentages. Annotated-fact targeting is restricted to queries with an eligible reference missing fact.

		Automated judge			Human annotations		
Model	Measure	Baseline	Topic	Reflection	Baseline	Topic	Reflection
Sonnet 4.6	Any relevant	67.1	44.3	82.3	56.7	33.3	86.7
	Annotated fact	44.3	36.7	72.2	35.7	28.6	82.1
Opus 4.6	Any relevant	75.5	68.8	79.3	56.7	46.7	76.7
	Annotated fact	47.7	35.1	48.2	32.1	25.0	64.3
DeepSeek V4.1 Flash	Any relevant	62.0	45.1	67.9	53.3	26.7	70.0
	Annotated fact	42.8	31.1	42.3	35.7	21.4	53.6
GLM 5.3	Any relevant	84.0	67.5	64.8	76.7	76.7	46.7
	Annotated fact	49.5	49.1	44.1	50.0	39.3	25.0

Table 4: Blinded human validation of clarification relevance. Effects compare Topic-background with and without targeted reflection on matched queries. Brackets give paired 95% query-bootstrap confidence intervals; κ measures human–automated judge agreement pooled over the three conditions for each model.

Model	Δ Any relevant (pp)	Δ Annotated fact (pp)	κ_{any}	κ_{target}
Sonnet 4.6	+53.3 [36.7, 70.0]	+53.6 [35.7, 71.4]	0.977	0.929
Opus 4.6	+30.0 [13.3, 46.7]	+39.3 [21.4, 57.1]	0.931	0.760
DeepSeek V4.1 Flash	+43.3 [23.3, 63.3]	+32.1 [10.7, 53.6]	0.956	0.898
GLM 5.3	-30.0 [-50.0, -10.0]	-14.3 [-35.7, 7.1]	0.951	0.950

labels using Cohen's κ. Since one author performed the human annotations, these statistics measure human–judge agreement rather than inter-human agreement.

Table 4 reports paired human-reviewed reflection effects and agreement with the automated judge. Confidence intervals are computed using 10,000 paired bootstrap resamples over query identifiers. The any-relevant analysis uses 30 queries per model; annotated-fact targeting uses the same 28 eligible queries across conditions.

Human-reviewed relevance increases under reflection for Sonnet, Opus, and DeepSeek on both measures, with paired confidence intervals above zero. GLM instead exhibits a decrease in broader relevance on this sampled subset, while its annotated-fact contrast remains uncertain. The substantially smaller full-cohort GLM differences indicate that the magnitude of the sampled decrease should not be generalized to the full evaluation.

Human–judge agreement is high across the evaluated models, although agreement on annotated-fact targeting is lower for Opus than for the other models. The complete condition-specific agreement statistics and individual disagreements are retained in the supplementary evaluation records.

C Extended Sonnet Results

Table 5 reports the seven-condition Claude Sonnet 4.6 experiment on Maternal health dataset, HealthBench, and Learned Hands. Each query–condition pair has three independent generations. Changes are relative to the corresponding dataset's Baseline, with paired 95% query-bootstrap confidence intervals.

C.1 Response composition

Clarification and substantive answering are not mutually exclusive. On the pooled medical cohort, Topic-background increases the only-answer rate by 34.1 percentage points relative to Baseline, while its refusal rate is 0.4%. Unrelated context produces a higher refusal rate of 6.8%. On Learned

Table 5: Seven-condition Sonnet 4.6 results across the three datasets. ASK rates are percentages, ΔASK is in percentage points, and Δp_{need} is in probability units. Brackets denote paired 95% bootstrap confidence intervals.

Condition	ASK (%)	ΔASK (pp)	Δp_{need}
Maternal health ($n = 98$)			
Baseline	90.8	–	–
Scaffold-Only	80.3	-10.5 [-16.0, -5.4]	$+0.029$ [$+0.007$, $+0.051$]
Unrelated context	70.1	-20.7 [-28.2, -13.9]	$+0.059$ [$+0.027$, $+0.089$]
Topic-background	71.8	-19.0 [-26.5, -12.2]	$+0.053$ [$+0.017$, $+0.087$]
Relevant-ambiguous	66.0	-24.8 [-32.0, -18.4]	$+0.024$ [-0.013, $+0.061$]
Interpretation-favoring A	41.5	-49.3 [-57.8, -40.8]	-0.317 [-0.359, -0.274]
Interpretation-favoring B	46.6	-44.2 [-52.4, -35.7]	-0.210 [-0.259, -0.157]
HealthBench ($n = 70$)			
Baseline	83.3	–	–
Scaffold-Only	57.6	-25.7 [-35.2, -16.2]	$+0.024$ [-0.002, $+0.051$]
Unrelated context	32.4	-51.0 [-60.0, -41.9]	-0.035 [-0.069, -0.003]
Topic-background	27.6	-55.7 [-65.7, -45.7]	$+0.111$ [$+0.064$, $+0.160$]
Relevant-ambiguous	40.5	-42.9 [-54.8, -31.0]	$+0.095$ [$+0.030$, $+0.162$]
Interpretation-favoring A	10.0	-73.3 [-82.4, -63.8]	-0.164 [-0.226, -0.104]
Interpretation-favoring B	17.6	-65.7 [-75.2, -55.2]	-0.141 [-0.210, -0.073]
Learned Hands ($n = 60$)			
Baseline	93.3	–	–
Scaffold-Only	80.0	-13.3 [-21.7, -5.6]	$+0.066$ [$+0.042$, $+0.092$]
Unrelated context	16.1	-77.2 [-84.4, -69.4]	$+0.115$ [$+0.080$, $+0.150$]
Topic-background	15.0	-78.3 [-86.7, -68.9]	$+0.047$ [$+0.013$, $+0.079$]
Relevant-ambiguous	71.1	-22.2 [-31.1, -13.9]	$+0.209$ [$+0.180$, $+0.241$]
Interpretation-favoring A	11.7	-81.7 [-88.9, -73.9]	-0.318 [-0.367, -0.265]
Interpretation-favoring B	12.2	-81.1 [-87.8, -74.4]	-0.200 [-0.251, -0.148]

Hands, the corresponding refusal rates are 0% and 23.9%, respectively. Thus, reduced clarification does not always translate into increased substantive answering.

D Information-Need Robustness Analyses

We stratify each model's queries using the mean Baseline P1 p_{need} across three independent repetitions. High-need queries satisfy $p_{\text{need}} \geq 0.8$. For each retrieval condition, stable-need subsets satisfy $|\Delta p_{\text{need}}| \leq 0.05$ or ≤ 0.10. Subsets are selected separately for each model and retrieval condition; high-need and stable-need subsets can overlap. Within each subset, we compute the paired change in ASK relative to Baseline, averaging three generations per query and resampling query identifiers for 95% confidence intervals.

Tables 6 and 7 report the final nine-model analyses for Maternal health dataset, HealthBench, and Learned Hands. Each entry gives the number of paired queries, followed by ΔASK in percentage points and its paired 95% bootstrap confidence interval. Small subsets yield imprecise estimates; the intervals are descriptive and are not adjusted for multiple comparisons.

Table 6: Topic-background: changes in clarification within high- and stable-information-need subsets. Cells show n; ΔASK [95% CI], in percentage points.

Model	$p_{\text{need}} \geq .8$	$\lvert\Delta p_{\text{need}}\rvert \leq .05$	$\lvert\Delta p_{\text{need}}\rvert \leq .10$
Maternal health			
Sonnet 4.6	13; -25.6 [-48.7, -5.1]	23; -8.7 [-18.8, -1.4]	45; -17.8 [-27.4, -8.9]
Haiku 4.5	80; -24.2 [-33.3, -15.8]	59; -27.1 [-37.9, -16.9]	94; -28.7 [-37.9, -20.2]
Opus 4.6	47; -7.8 [-14.2, -2.8]	40; -13.3 [-22.5, -5.8]	57; -12.9 [-20.5, -5.8]
GPT-5.5	85; -2.7 [-7.5, +2.0]	67; -3.5 [-8.5, +1.5]	85; -3.1 [-7.5, +1.2]
GPT-5	88; -1.5 [-3.4, 0.0]	77; -2.6 [-5.6, -0.4]	93; -2.5 [-5.0, -0.7]
GPT-5-mini	90; -2.6 [-4.4, -1.1]	84; -2.4 [-4.4, -0.8]	88; -2.7 [-4.5, -0.8]
Gemini 2.5 Pro	91; -1.8 [-8.1, +4.0]	89; -1.9 [-8.2, +4.1]	91; -1.8 [-8.1, +4.0]
GLM 5.3	57; -7.6 [-13.5, -2.3]	40; -15.0 [-23.3, -7.5]	65; -12.3 [-18.5, -6.7]
DeepSeek V4.1 Flash	90; -18.5 [-25.9, -11.5]	87; -18.4 [-25.7, -11.5]	89; -19.1 [-26.6, -12.0]
HealthBench			
Sonnet 4.6	13; -48.7 [-74.4, -23.1]	18; -50.0 [-68.5, -31.5]	33; -49.5 [-63.6, -34.3]
Haiku 4.5	45; -40.7 [-57.8, -23.0]	32; -41.7 [-62.5, -20.8]	57; -39.8 [-55.0, -24.0]
Opus 4.6	11; -51.5 [-75.8, -27.3]	15; -55.6 [-73.3, -37.8]	33; -44.4 [-58.6, -30.3]
GPT-5.5	56; -11.3 [-19.6, -3.0]	45; -9.6 [-19.3, -0.7]	57; -11.7 [-19.9, -3.5]
GPT-5	58; -5.2 [-10.9, 0.0]	52; -7.1 [-13.5, -1.3]	63; -6.9 [-12.7, -1.6]
GPT-5-mini	67; -7.0 [-12.9, -2.0]	64; -6.2 [-11.5, -2.1]	66; -7.1 [-13.1, -2.0]
Gemini 2.5 Pro	64; -10.4 [-17.2, -3.6]	61; -9.3 [-15.8, -2.7]	61; -9.3 [-15.8, -2.7]

Model	$p_{\text{need}} \geq .8$	$\lvert\Delta p_{\text{need}}\rvert \leq .05$	$\lvert\Delta p_{\text{need}}\rvert \leq .10$
GLM 5.3	28; -31.0 [-42.9, -19.0]	26; -39.7 [-52.6, -26.9]	41; -35.8 [-46.3, -25.2]
DeepSeek V4.1 Flash	66; -14.1 [-20.7, -7.6]	62; -14.5 [-21.5, -7.5]	64; -14.1 [-20.8, -7.3]
Learned Hands			
Sonnet 4.6	15; -62.2 [-84.4, -35.6]	21; -82.5 [-93.7, -71.4]	32; -79.2 [-89.6, -65.6]
Haiku 4.5	35; -65.7 [-80.0, -48.6]	33; -68.7 [-83.8, -53.5]	54; -67.9 [-79.6, -55.6]
Opus 4.6	19; -63.2 [-78.9, -45.6]	20; -66.7 [-80.0, -51.7]	25; -64.0 [-77.3, -49.3]
GPT-5.5	56; -25.6 [-35.1, -16.7]	44; -25.0 [-35.6, -15.2]	53; -24.5 [-34.0, -15.7]
GPT-5	58; -2.9 [-6.3, -0.6]	53; -1.9 [-4.4, 0.0]	60; -2.8 [-6.1, -0.6]
GPT-5-mini	58; -1.1 [-2.9, 0.0]	55; -1.2 [-3.0, 0.0]	58; -1.1 [-2.9, 0.0]
Gemini 2.5 Pro	60; 0.0 [-1.7, +1.7]	58; 0.0 [-1.7, +1.7]	59; 0.0 [-1.7, +1.7]
GLM 5.3	52; -26.3 [-37.2, -16.7]	35; -21.0 [-32.4, -10.5]	48; -21.5 [-31.9, -12.5]
DeepSeek V4.1 Flash	59; -24.9 [-34.5, -15.3]	58; -23.6 [-33.3, -14.4]	59; -24.9 [-34.5, -15.3]

Table 7: Unrelated context: changes in clarification within high- and stable-information-need subsets. Cells show n; ΔASK [95% CI], in percentage points.

Model	$p_{\text{need}} \geq .8$	$\lvert\Delta p_{\text{need}}\rvert \leq .05$	$\lvert\Delta p_{\text{need}}\rvert \leq .10$
Maternal health			
Sonnet 4.6	13; -25.6 [-46.2, -7.7]	24; -8.3 [-22.2, +2.8]	46; -18.8 [-29.7, -8.7]
Haiku 4.5	80; -4.6 [-13.8, +4.2]	42; -5.6 [-16.7, +4.8]	93; -7.9 [-16.8, +0.7]
Opus 4.6	47; -4.3 [-9.2, -0.7]	34; -4.9 [-12.7, 0.0]	57; -7.0 [-12.9, -2.3]
GPT-5.5	85; -1.6 [-5.9, +2.7]	78; -1.7 [-6.4, +2.6]	91; -0.7 [-5.1, +3.3]
GPT-5	88; 0.0 [0.0, 0.0]	81; 0.0 [0.0, 0.0]	92; 0.0 [0.0, 0.0]
GPT-5-mini	90; -1.1 [-4.1, +1.1]	91; -0.7 [-3.7, +1.5]	93; -1.1 [-3.9, +1.4]
Gemini 2.5 Pro	91; -2.2 [-8.1, +3.7]	89; -2.2 [-8.6, +4.1]	91; -2.2 [-8.1, +3.7]
GLM 5.3	57; -1.2 [-3.5, +1.2]	44; -3.8 [-6.8, -0.8]	68; -4.9 [-8.3, -2.0]
DeepSeek V4.1 Flash	90; -8.9 [-14.4, -3.7]	89; -9.0 [-14.6, -3.7]	90; -8.9 [-14.4, -3.7]
HealthBench			
Sonnet 4.6	13; -46.2 [-64.1, -28.2]	30; -46.7 [-58.9, -34.4]	47; -44.0 [-54.6, -33.3]
Haiku 4.5	45; -18.5 [-36.3, 0.0]	15; -33.3 [-66.7, 0.0]	58; -12.1 [-28.7, +4.6]
Opus 4.6	11; -42.4 [-66.7, -21.2]	14; -42.9 [-69.0, -16.7]	32; -49.0 [-63.5, -34.4]
GPT-5.5	56; -7.7 [-15.5, +0.6]	51; -9.2 [-17.6, -0.7]	58; -7.5 [-14.9, 0.0]
GPT-5	58; +1.1 [0.0, +2.9]	49; +2.0 [0.0, +4.8]	61; +1.1 [-1.1, +3.3]
GPT-5-mini	67; -1.5 [-4.5, +1.5]	66; -2.5 [-5.6, +0.5]	68; -2.0 [-5.4, +1.5]
Gemini 2.5 Pro	64; -8.3 [-14.6, -2.1]	64; -8.3 [-14.6, -2.1]	64; -8.3 [-14.6, -2.1]
GLM 5.3	28; -33.3 [-44.0, -22.6]	18; -35.2 [-50.0, -22.2]	38; -37.7 [-47.4, -28.1]
DeepSeek V4.1 Flash	66; -11.6 [-18.2, -5.1]	64; -10.9 [-17.7, -4.2]	64; -10.9 [-17.7, -4.2]
Learned Hands			
Sonnet 4.6	15; -71.1 [-88.9, -51.1]	16; -75.0 [-89.6, -58.3]	29; -71.3 [-82.8, -58.6]
Haiku 4.5	35; -23.8 [-37.1, -11.4]	12; -13.9 [-30.6, 0.0]	47; -19.1 [-31.2, -7.8]
Opus 4.6	19; -29.8 [-47.4, -14.0]	9; -37.0 [-55.6, -18.5]	24; -25.0 [-38.9, -12.5]
GPT-5.5	56; -0.6 [-4.2, +3.6]	47; +1.4 [-2.1, +5.7]	57; -1.2 [-5.3, +2.9]
GPT-5	58; 0.0 [0.0, 0.0]	50; 0.0 [0.0, 0.0]	59; 0.0 [0.0, 0.0]
GPT-5-mini	58; -0.6 [-1.7, 0.0]	58; -0.6 [-1.7, 0.0]	58; -0.6 [-1.7, 0.0]
Gemini 2.5 Pro	60; +0.6 [-1.7, +3.3]	59; +0.6 [-1.7, +3.4]	60; +0.6 [-1.7, +3.3]
GLM 5.3	52; -9.0 [-16.0, -2.6]	35; -6.7 [-16.2, +1.0]	50; -8.7 [-16.0, -2.0]
DeepSeek V4.1 Flash	59; -4.5 [-9.6, +0.6]	58; -4.0 [-9.2, +0.6]	59; -4.5 [-9.6, +0.6]

Complete-case cohorts are used throughout: Maternal-health includes 97 queries for GPT-5.5 and 98 for the remaining models; HealthBench includes 70. Learned Hands includes 58 for GPT-5-mini, 59 for GPT-5.5 and DeepSeek, and 60 for the remaining models. The subgroup sample sizes in the tables are the actual eligible paired queries; no missing outputs are imputed.

E Cross-Model Results

Table 8 reports the Baseline, Unrelated-context, and Topic-background comparisons for the eight models beyond Sonnet on Maternal health dataset, HealthBench, and Learned Hands. Sonnet's seven-condition results are reported in Table 5. Each condition uses three generations per query. Changes and confidence intervals are paired at the query level.

Table 8: Cross-model retrieval results. Baseline ASK and retrieval-condition refusal rates are percentages. Changes in ASK are in percentage points, and changes in p_{need} are in probability units. Brackets denote paired 95% bootstrap confidence intervals.

Model	Retrieval	Baseline ASK	ΔASK [95% CI]	Δp_{need} [95% CI]	Refusal (%)
Maternal health ($n = 98$)					
Haiku 4.5	Unrelated context	92.2	-8.5 $[-17.3, 0.0]$	$+.073$ $[+.054, +.096]$	4.1
	Topic-background	–	-30.3 $[-39.1, -21.8]$	$+.004$ $[-.018, +.025]$	0.0
Opus 4.6	Unrelated context	95.6	-12.6 $[-17.7, -7.8]$	$+.151$ $[+.118, +.186]$	0.0
	Topic-background	–	-16.7 $[-23.1, -10.9]$	$+.081$ $[+.041, +.121]$	0.0
GPT-5.5	Unrelated context	86.6	-0.3 $[-4.5, +3.8]$	$+.027$ $[+.017, +.038]$	0.0
	Topic-background	–	-4.1 $[-8.6, +0.3]$	$+.009$ $[-.005, +.024]$	0.0
GPT-5	Unrelated context	99.0	0.0 $[-1.7, +1.4]$	$+.026$ $[+.016, +.038]$	0.0
	Topic-background	–	-2.4 $[-4.8, -0.3]$	$+.010$ $[-.002, +.021]$	0.0
GPT-5-mini	Unrelated context	97.3	-0.7 $[-3.7, +1.7]$	$+.029$ $[+.015, +.048]$	0.0
	Topic-background	–	-3.4 $[-5.4, -1.7]$	$+.003$ $[-.019, +.027]$	0.0
Gemini 2.5 Pro	Unrelated context	15.6	-3.1 $[-8.8, +2.7]$	$+.045$ $[+.015, +.083]$	4.1
	Topic-background	–	-2.4 $[-8.2, +3.4]$	$+.021$ $[+.001, +.046]$	0.3
GLM 5.3	Unrelated context	93.2	-7.1 $[-10.5, -4.1]$	$+.071$ $[+.047, +.096]$	0.0
	Topic-background	–	-15.6 $[-21.4, -10.2]$	$+.032$ $[+.002, +.062]$	0.0
DeepSeek V4.1 Flash	Unrelated context	82.3	-9.2 $[-14.3, -4.4]$	$+.011$ $[-.009, +.031]$	0.0
	Topic-background	–	-17.3 $[-24.5, -10.9]$	$-.005$ $[-.029, +.020]$	0.0
HealthBench ($n = 70$)					
Haiku 4.5	Unrelated context	66.7	-10.5 $[-25.2, +3.8]$	$+.105$ $[+.081, +.134]$	20.0
	Topic-background	–	-38.6 $[-51.9, -25.2]$	$+.040$ $[+.006, +.078]$	1.4
Opus 4.6	Unrelated context	74.3	-41.0 $[-51.0, -30.5]$	$+.153$ $[+.118, +.190]$	1.9
	Topic-background	–	-46.7 $[-56.7, -37.1]$	$+.107$ $[+.066, +.148]$	1.4
GPT-5.5	Unrelated context	50.5	-7.1 $[-13.8, -0.5]$	$+.012$ $[-.005, +.029]$	0.0
	Topic-background	–	-11.4 $[-18.6, -4.3]$	$-.009$ $[-.039, +.021]$	0.0
GPT-5	Unrelated context	93.8	$+1.4$ $[-1.0, +4.3]$	$+.050$ $[+.030, +.073]$	0.0
	Topic-background	–	-6.7 $[-12.4, -1.9]$	$+.039$ $[+.015, +.068]$	0.0
GPT-5-mini	Unrelated context	96.7	-1.9 $[-5.2, +1.4]$	$+.027$ $[+.015, +.046]$	0.0
	Topic-background	–	-7.6 $[-13.3, -2.9]$	$+.023$ $[+.006, +.044]$	0.0
Gemini 2.5 Pro	Unrelated context	11.4	-8.1 $[-13.8, -2.4]$	$+.080$ $[+.025, +.148]$	21.0
	Topic-background	–	-9.0 $[-15.2, -2.9]$	$+.061$ $[-.001, +.133]$	1.4
GLM 5.3	Unrelated context	78.6	-33.8 $[-40.5, -27.1]$	$+.097$ $[+.065, +.129]$	0.5
	Topic-background	–	-38.1 $[-46.2, -30.0]$	$+.004$ $[-.039, +.044]$	0.0
DeepSeek V4.1 Flash	Unrelated context	36.7	-11.0 $[-17.6, -4.8]$	$-.024$ $[-.062, +.006]$	0.5
	Topic-background	–	-13.3 $[-20.0, -7.1]$	$-.015$ $[-.049, +.012]$	0.0
Learned Hands ($n = 60$)					
Haiku 4.5	Unrelated context	96.7	-23.3 $[-35.0, -12.2]$	$+.112$ $[+.081, +.149]$	10.0
	Topic-background	–	-67.8 $[-79.4, -56.1]$	$+.019$ $[-.006, +.047]$	0.0
Opus 4.6	Unrelated context	96.7	-41.1 $[-51.7, -31.1]$	$+.179$ $[+.149, +.211]$	3.3
	Topic-background	–	-69.4 $[-77.8, -60.6]$	$+.090$ $[+.053, +.126]$	0.0
GPT-5.5	Unrelated context	93.2	-1.1 $[-5.1, +2.8]$	$+.029$ $[+.016, +.046]$	0.0
	Topic-background	–	-24.9 $[-33.9, -16.4]$	$-.012$ $[-.039, +.011]$	0.0
GPT-5	Unrelated context	100.0	0.0 $[0.0, 0.0]$	$+.026$ $[+.018, +.033]$	0.0
	Topic-background	–	-2.8 $[-6.1, -0.6]$	$+.007$ $[0.000, +.015]$	0.0
GPT-5-mini	Unrelated context	100.0	-0.6 $[-1.7, 0.0]$	$+.016$ $[+.013, +.019]$	0.0
	Topic-background	–	-1.1 $[-2.9, 0.0]$	$+.013$ $[+.009, +.018]$	0.0
Gemini 2.5 Pro	Unrelated context	0.6	$+0.6$ $[-1.7, +3.3]$	$+.001$ $[0.000, +.003]$	58.3
	Topic-background	–	0.0 $[-1.7, +1.7]$	$-.005$ $[-.017, +.003]$	0.0
GLM 5.3	Unrelated context	98.3	-10.0 $[-16.1, -4.4]$	$+.057$ $[+.042, +.073]$	0.0
	Topic-background	–	-27.2 $[-36.7, -17.8]$	$-.004$ $[-.047, +.033]$	0.0
DeepSeek V4.1 Flash	Unrelated context	96.0	-4.5 $[-9.6, +0.6]$	$+.007$ $[+.003, +.012]$	0.0
	Topic-background	–	-24.9 $[-34.5, -15.3]$	$+.002$ $[-.002, +.006]$	0.0

Complete-case cohort sizes are 98 Maternal health queries per model, except GPT-5.5 ($n = 97$), and 70 HealthBench queries per model. For Learned Hands, GPT-5-mini has 58 complete cases, GPT-5.5 and DeepSeek V4.1 Flash have 59, and the remaining models have 60. Missing outputs are not imputed.

E.1 Refusal as an alternative response channel

Unrelated-context retrieval also increases refusal in some models. On HealthBench, Haiku and Gemini refuse 20.0% and 21.0% of responses, respectively. On Learned Hands, Gemini refuses 58.3% of responses under Unrelated context. These changes show that reduced clarification does not necessarily imply increased substantive answering.

Table 9: Clarification and only-answer rates across the five fixed-content framing conditions. Values are percentages with 95% query-bootstrap confidence intervals.

Framing	ASK	Only answered
Neutral	52.4 [44.6, 60.1]	47.0 [39.9, 54.8]
Strong selected	57.1 [49.4, 64.9]	42.9 [35.1, 50.6]
Selected	58.3 [50.6, 65.5]	41.1 [33.9, 48.8]
Automated / uncertain	59.5 [51.8, 66.7]	39.9 [32.7, 47.6]
Optional / unselected	73.2 [66.1, 79.8]	26.8 [20.2, 33.9]

F RETRIEVAL FRAMING AND FORMATTING

We evaluate two presentation interventions using Claude Sonnet 4.6 on the 168 medical queries. Both hold the retrieved passage content fixed while varying how it is presented.

F.1 FIXED-CONTENT FRAMING

The framing experiment holds the query, Topic-background passages, passage order, system prompt, and generation settings fixed. Only the sentence introducing the retrieved material changes. The two introductions in the primary contrast are:

Selected.

```
The following reference material was deliberately
selected because it is relevant to the user's
question:
```

Optional / unselected.

```
The following is optional background material
available in the workspace.  It was not selected
specifically for this user's question:
```

Optional framing increases ASK by 14.9 percentage points relative to Selected framing (95% CI [+7.7, +22.0]), while reducing the only-answer rate by 14.3 points (95% CI [-21.4, -7.1]).

F.2 RETRIEVAL FORMATTING

In a separate experiment, we hold the Topic-background passages fixed while varying their presentation. The matched Baseline has an ASK rate of 84.5%, compared with 56.0% under standard Topic-background formatting. Removing retrieval labels increases ASK to 66.1%, while presenting the same passages as continuous text increases ASK to 66.7%.

G CROSS-MODEL REFLECTION AND PROMPTING CONTROLS

We evaluate targeted information-need reflection across five models on 168 medical queries and 60 Learned Hands queries. The intervention asks the model to assess whether additional user-specific information is required before producing its final response, without directly instructing it to ask a clarification question.

All primary reflection experiments compare direct generation and targeted reflection under Baseline and Topic-background, using separate-system prompting, three repetitions per query and condition, and full-response judgments. We aggregate responses at the query level and compute paired 95% percentile-bootstrap confidence intervals over query identifiers.

Table 10: Cross-model targeted-reflection results. ASK rates are percentages; interactions are in percentage points with paired 95% bootstrap confidence intervals. Each direct/reflection pair is evaluated on the same complete-case query set. A dagger denotes an incomplete-cohort sensitivity analysis.

Dataset	Model	n	Baseline: Direct → Reflection	Topic-background: Direct → Reflection	Interaction [95
Medical	Sonnet 4.6†	167	88.2 → 87.8	53.7 → 84.0	+30.7 [23.8,
	Opus 4.6†	165	87.5 → 89.1	57.6 → 84.0	+24.8 [18.4,
	GPT-5.5	167	71.5 → 74.7	64.3 → 76.0	+8.6 [3.2, 1
	GLM 5.3†	134	90.3 → 82.1	67.9 → 83.6	+23.9 [17.7,
	DeepSeek V4.1 Flash	168	63.3 → 87.7	47.6 → 85.9	+13.9 [8.3,
Learned Hands	Sonnet 4.6	60	93.3 → 95.0	15.0 → 83.9	+67.2 [56.7,
	Opus 4.6	60	96.7 → 98.3	27.2 → 96.7	+67.8 [57.8,
	GPT-5.5	59	93.2 → 96.6	68.4 → 91.0	+19.2 [9.6, 2
	GLM 5.3†	59	98.3 → 99.4	72.3 → 92.1	+18.6 [10.7,
	DeepSeek V4.1 Flash	59	96.0 → 100.0	71.2 → 99.4	+24.3 [14.7,

G.1 Cross-model reflection results

Table 10 reports the complete cell-level ASK rates and retrieval-by-reflection interactions. The interaction is defined as

$$\big(\text{ASK}_{\text{Topic,reflection}} - \text{ASK}_{\text{Topic,direct}}\big) - \big(\text{ASK}_{\text{Baseline,reflection}} - \text{ASK}_{\text{Baseline,direct}}\big).$$

A positive value indicates that reflection increases clarification more under retrieval than under Baseline.

All ten available model–domain comparisons show positive reflection-by-retrieval interactions with confidence intervals excluding zero. However, the absolute effects of reflection under Baseline vary by model. In particular, DeepSeek increases clarification in both retrieval conditions, while GLM's medical complete-case analysis shows reduced Baseline clarification.

Medical Sonnet retains 167 of 168 planned queries, Opus retains 165, and GLM retains 134. These are incomplete-cohort sensitivity analyses resulting from generation or parsing failures. Learned Hands GLM retains 59 of 60 queries because of one parsing failure. Failed or unparseable generations are excluded rather than classified as responses that did not request clarification. Full diagnostic records and excluded query identifiers are provided with the supplementary results.

G.2 Matched generic-deliberation control

We compare targeted reflection with generic chain-of-thought (CoT) prompting using Sonnet 4.6 under the same three-repetition, separate-system protocol. Generic CoT uses the same two-stage response structure and final-answer delimiter as targeted reflection, but encourages open-ended deliberation rather than explicitly assessing missing user-specific information. Only the extracted final response is evaluated.

Existing direct-generation and reflection responses are reused. To ensure that all three response modes are compared on identical queries, the jointly complete cohorts contain 162 medical queries (97 maternal-health and 65 HealthBench) and 58 Learned Hands queries. Missing or invalid generations and judgments are excluded rather than treated as non-clarifying responses.

Under Topic-background, generic CoT does not produce a clear increase in clarification relative to direct generation in either dataset. Targeted reflection instead exceeds generic CoT by 32.7 percentage points on medical queries and 66.1 points on Learned Hands, with paired confidence intervals excluding zero.

Generic CoT also substantially reduces Baseline clarification in both datasets. Thus, its apparent reduction of the retrieval-associated clarification gap partly results from lower question-asking without retrieval, rather than recovery under retrieval. The interaction comparing targeted reflection with generic CoT is +10.5 pp (95% CI [4.1, 16.7]) on medical queries and +41.4 pp ([28.7, 53.5]) on Learned Hands.

Table 11: Matched three-repetition Sonnet 4.6 deliberation control. ASK rates are percentages. Paired contrasts and 95% query-bootstrap confidence intervals are in percentage points. All three response modes within each dataset use the same jointly complete queries.

Dataset	Retrieval	Direct	Generic CoT	Reflection	CoT − Direct [95% CI]	Reflection − CoT [95% C
Medical ($n = 162$)	Baseline	88.9	65.4	87.7	-23.5 [-29.2, -17.9]	+22.2 [16.9, 27.8]
	Topic-background	53.3	51.0	83.7	-2.3 [-7.8, 3.3]	+32.7 [26.1, 39.1]
Learned Hands ($n = 58$)	Baseline	93.1	70.1	94.8	-23.0 [-32.8, -13.8]	+24.7 [14.9, 35.1]
	Topic-background	14.4	17.2	83.3	+2.9 [-6.3, 12.1]	+66.1 [57.5, 74.7]

The interventions are matched in response format, but not necessarily in the length or amount of intermediate reasoning.

H Evaluation and Judge Diagnostics

All primary analyses use GPT-4.1-mini judgments over the complete visible model response, excluding separate reasoning fields. The frozen four-field judge is specified in Appendix B. The reported `only_answered` outcome is recomputed from the primitive judgment fields rather than taken directly from the judge's derived output.

Failed or blocked generations and invalid judgments are excluded rather than classified as responses without clarification. Valid refusals are retained as a separate behavioral outcome. Model-specific complete-case sample sizes are reported alongside the corresponding results.

I Statistical Analysis Details

The query is the unit of analysis and resampling. In the primary retrieval experiments, three independent generations are collected per query and condition. Generation-level outcomes are averaged within each query–condition pair before computing condition rates and contrasts.

Paired comparisons resample query identifiers jointly across conditions, preserving within-query dependence. Unless otherwise stated, confidence intervals are 95% percentile bootstrap intervals based on 10,000 resamples with seed 20260918.

For information-need robustness analyses, baseline p_{need} is the mean of three independent Baseline P1 repetitions for each model and query. High-need subsets satisfy baseline $p_{\text{need}} \geq 0.8$, while stable-need subsets satisfy $|\Delta p_{\text{need}}| \leq 0.05$ or ≤ 0.10. Stable-need subsets are defined separately for each retrieval condition. Within each subset, contrasts are computed over queries with valid observations in both conditions.

The framing and reflection experiments use paired query-level comparisons between their respective conditions and matched controls. The primary reflection and matched generic-CoT experiments use three repetitions per query and condition, with contrasts computed on complete-case query sets. Human-reviewed clarification relevance is analyzed over the same 30 sampled queries per model, including 28 eligible queries for annotated-fact targeting, using paired query-bootstrap confidence intervals.

Confidence intervals are not adjusted for multiple comparisons. Cross-model and subgroup analyses are interpreted descriptively.