

Evaluating the Accuracy and Limitations of AI Chatbots in Daily Tasks

Dr. B. Anuja Beatrice¹, B. Perarasu²,

¹Associate Professor & Head, Department of Computer Applications, Sri Krishna Arts and Science College, Coimbatore

²Student III BCA A, Department of Computer Applications, Sri Krishna Arts and Science College, Coimbatore

Abstract— Artificial intelligence (AI) chatbots such as ChatGPT, Gemini, and Copilot have moved from experimental novelties to everyday tools used for writing, scheduling, coding, tutoring, and information retrieval. Despite their rapid adoption, systematic understanding of how accurate and dependable these systems are across the range of tasks that ordinary users perform daily remains limited. Here we build and apply an evaluation framework that examines the accuracy, consistency, and failure modes of three widely used AI chatbots across six everyday task categories: factual question answering, mathematical reasoning, code generation, creative writing, scheduling and planning, and real-time information retrieval. Using a structured trial protocol of 600 prompts (100 per category) evaluated by human raters against verified ground truth, we quantify accuracy, hallucination rate, response relevance, and consistency under repeated querying. Results show that chatbots perform strongly (85–93%) on creative and general-knowledge tasks but degrade substantially (55–61%) on tasks requiring current, time-sensitive information, and that accuracy falls sharply as conversational complexity and dialogue length increase. We further identify a measurable gap between users' perceived trust in chatbot answers and the systems' actual measured accuracy, particularly in health, financial, and legal domains. We close by tracing these limitations to their architectural and training-data origins and by proposing concrete steps toward safer, more transparent everyday use of conversational AI.

Keywords— artificial intelligence, chatbots, conversational AI, large language models, accuracy evaluation, hallucination, natural language processing, human-AI interaction, reliability

I. INTRODUCTION

Conversational artificial intelligence has transitioned from a research curiosity into infrastructure that hundreds of millions of people rely on every day. Chatbots built on large language models (LLMs) are now used to draft emails, summarize documents, answer medical and financial questions, tutor students, generate software code, and manage personal schedules. Public platforms report user bases in the hundreds of millions, and organizations across healthcare, education, customer service, and finance have begun to embed chatbot assistants directly into their workflows.

This rapid, largely voluntary adoption has outpaced rigorous, independent evaluation of how accurate these systems actually are outside of curated benchmark datasets. Most published benchmarks emphasize narrow, well-defined tasks such as standardized examinations or coding challenges, where correctness is easy to verify. Everyday use, however, is messier: it involves ambiguous phrasing, multi-turn dialogue, requests for time-sensitive information, and domains such as health or finance where an incorrect answer carries real consequences. Recent journalistic and academic audits have found that a substantial fraction of everyday chatbot answers contain factual or sourcing errors, underscoring the gap between benchmark performance and real-world reliability.

This study addresses that gap directly, asking three questions that a purely benchmark-driven view of chatbot performance tends to leave unanswered. First, how accurate are popular AI chatbots across the categories of tasks that ordinary users perform daily? Second, which factors — task type, conversational depth, and topic domain — most strongly predict where a chatbot is likely to err? Third, how well does users' perceived trust in chatbot answers track the systems' actual measured accuracy? Answering these required designing a controlled, human-rated evaluation spanning six everyday task categories, described in Section V, and analyzing the resulting error patterns.

Understanding chatbot accuracy at the granularity of everyday task categories also matters for a practical reason: unlike a specialized clinical or legal tool, a general-purpose chatbot is typically used by the same person across many different kinds of requests within a single session — asking for a recipe, then a scheduling adjustment, then a health question — without any

explicit signal to the user that the system's reliability may differ sharply between those requests. A user who has developed reasonable trust in a chatbot's writing help may unconsciously extend that same trust to its medication-timing advice, even though, as this paper shows, the two tasks sit at very different points on the accuracy spectrum. Making that spectrum explicit is the central empirical contribution of this work.

The discussion that follows builds toward those three questions in stages. Section II surveys prior efforts to benchmark chatbot reliability; Section III sketches the model architecture behind many of the failure patterns examined later, and Section IV briefly compares platform capabilities relevant to everyday accuracy. Sections V and VI lay out the trial protocol and scoring metrics, after which Section VII walks through the accuracy, error-type, complexity, and trust-gap findings. Section VIII groups the recurring limitations into eight categories, Section IX grounds these in four everyday case studies, and Sections X through XIV work through implications, recommendations, future directions, threats to validity, and closing remarks.

II. RELATED WORK

A. Domain-Specific Accuracy Studies

Independent evaluators have probed chatbot accuracy most intensively within specialized domains. Studies in medicine have been particularly active: comparative analyses of ChatGPT and Gemini on orthopaedic conditions such as meniscal tears found that a meaningful share of generated statements could not be verified against clinical references [1]. Similar comparative work on resuscitation guidelines found that ChatGPT-3.5 and ChatGPT-4 omitted or altered key clinical recommendations relative to the European Resuscitation Council guidelines, raising concerns about currency and completeness of medical answers [2]. Examinations of chatbot performance on licensing and board-style questions report wide variation, with correct-response rates ranging from roughly two-thirds to over ninety percent depending on specialty and model version [3], [4]. These studies consistently note that accuracy is not static: newer model versions tested months apart on the same question sets show measurable improvement, complicating any single reported figure being treated as a permanent property of a product rather than a specific model snapshot.

B. General-Purpose and Consumer Audits

Beyond medicine, broader science-accuracy audits have evaluated chatbots on hundreds of scientific hypotheses and found that, after correcting for chance agreement, performance is markedly weaker than raw accuracy figures suggest, with particular difficulty in correctly identifying false statements and in producing consistent answers to identical prompts asked repeatedly [5]. Consumer-facing audits that sampled everyday queries across several popular chatbots found that close to half of responses contained some accuracy issue, most commonly related to sourcing or missing context, although this rate improved markedly between successive evaluation rounds as underlying models were updated [6]. These consumer-facing audits are among the few prior efforts to sample broad, everyday-style queries rather than domain-specific exam questions, and their findings motivate the task taxonomy used in this paper.

C. Foundational Model Literature

The underlying causes of the error patterns reported above are rooted in how transformer-based language models are built and trained. Nearly all modern chatbots are built on large-scale foundation models: broadly pretrained, self-attention-based networks later adapted to a wide range of downstream conversational tasks [7], and subsequent scaling work demonstrated that very large versions of this architecture, trained on broad web-scale corpora, acquire broad few-shot task competence without task-specific fine-tuning [8]. Further scaling studies documented that certain capabilities emerge only above a size threshold, helping explain why smaller or older chatbot variants underperform larger ones disproportionately on reasoning-heavy categories such as mathematics and code [9]. Technical reports for frontier models describe the additional instruction-tuning and human-feedback alignment stages that shape a raw language model into a conversational assistant [10]. A separate line of work has explored retrieval-augmented generation, in which a model's output is grounded in retrieved external documents rather than parametric memory alone, and shown this substantially reduces — without eliminating — factual error in knowledge-intensive tasks [11]. Multilingual and multitask evaluations of ChatGPT have further shown that reasoning quality, hallucination rate, and interactivity all vary considerably by language and task type rather than forming a single uniform

competence [12]. Finally, a comprehensive survey of hallucination in natural language generation systems categorizes fabrication errors by cause and proposes a taxonomy that we adopt in part for our own error classification in Section VI [13].

D. Research Gap

Collectively, this literature establishes two consistent findings that motivate our study. First, chatbot accuracy is highly task- and domain-dependent rather than a single fixed property of a model. Second, most existing evaluations focus on a single specialized domain (e.g., one medical subspecialty) rather than the broad mix of task types — planning, coding, writing, factual recall — that make up typical daily usage. Our work contributes a cross-category comparison designed to reflect that everyday mix, together with an explicit measurement of the gap between user trust and measured accuracy, which prior domain-specific studies do not directly quantify. It also connects, for the first time in a single study to our knowledge, the accuracy-by-domain findings typical of clinical evaluations with the conversational-depth and consistency effects more commonly studied in isolation within the natural language processing literature, allowing the two threads of prior work to be compared on common ground.

III. OVERVIEW OF AI CHATBOT ARCHITECTURE

Understanding why chatbots succeed and fail in the specific patterns reported in Section VII first requires a brief description of how these systems are built. Modern AI chatbots are built around transformer-based large language models trained on broad text corpora and subsequently fine-tuned using instruction-following and human-feedback techniques. At inference time, a typical conversational pipeline consists of five stages, illustrated in Fig. 5: the system receives raw user input, performs natural-language understanding to parse intent and entities, consults a context or memory module that tracks the ongoing dialogue and, in retrieval-augmented systems, external documents, generates a candidate response through the underlying language model, and finally returns formatted output to the user.



Fig. 5. Generalized processing pipeline of a conversational AI chatbot

Fig. 5. Generalized processing pipeline of a conversational AI chatbot.

Two architectural properties of this pipeline are central to the limitations discussed later in this paper. First, the language-generation stage produces text through statistical next-token prediction rather than by querying a verified knowledge base, which makes fluent but incorrect output — commonly termed hallucination — possible even when the model is confident. Second, the training data underlying the base model has a fixed cutoff date, so without a live retrieval or browsing component, the system's knowledge of recent events, prices, or schedules is necessarily stale.

A. Training Paradigm

Most deployed chatbots are trained in three broad stages. Pretraining exposes a transformer model to a very large, broad text corpus using a self-supervised next-token prediction objective, giving the model general linguistic and world knowledge but no explicit notion of how to behave as a helpful assistant [7], [8]. Supervised fine-tuning then trains the model on curated examples of instructions paired with high-quality responses, teaching it to follow the conversational format users expect. Finally, a reinforcement-learning stage using human or model-generated preference feedback further steers outputs toward responses that human raters judge to be more helpful, honest, and harmless [10]. Each stage shapes the error patterns observed

in Section VI: pretraining determines the raw factual coverage and cutoff date, fine-tuning determines how fluently and confidently the model presents its answers, and preference optimization can inadvertently reward answers that sound authoritative over answers that correctly express uncertainty.

B. Retrieval and Tool Augmentation

Some chatbot deployments extend the base pipeline with retrieval-augmented generation, in which a search or document-lookup step supplies the model with current, external text that is inserted into its context window before response generation [11]. This mitigates — though does not eliminate — the knowledge-staleness limitation, since the model still must correctly synthesize and attribute the retrieved material rather than its own parametric memory. Where retrieval or live tool access was unavailable in the systems tested, real-time information accuracy was categorically weaker, consistent with the pattern in Fig. 1.

IV. SURVEY OF POPULAR CHATBOT PLATFORMS

Before describing our evaluation methodology, it is useful to survey the general landscape of platforms that motivate this study. General-purpose conversational assistants such as OpenAI's ChatGPT, Google's Gemini, Microsoft's Copilot, and Anthropic's Claude have each been integrated into consumer products including web browsers, office productivity suites, and mobile operating systems, placing chatbot access directly inside the applications people already use for daily tasks. These platforms differ in underlying model architecture, training data cutoff, and the degree to which they incorporate live web retrieval or connected tools, but they share the same fundamental transformer-based generation process described in Section III, and consequently share many of the same categories of failure even where absolute accuracy differs.

Table I summarizes four general dimensions along which everyday-use chatbot platforms commonly vary: whether live retrieval or browsing is supported by default, whether the system maintains memory of a user's preferences across separate sessions, whether it can execute code or perform calculations rather than only generating text describing them, and whether it discloses source citations for factual claims. These dimensions matter directly for the error patterns examined later: platforms lacking live retrieval are structurally limited on real-time information tasks regardless of base model quality, while platforms that execute code rather than merely describing its expected behavior are less prone to certain classes of silent logical error in programming tasks.

TABLE I. PLATFORM CAPABILITIES RELEVANT TO EVERYDAY ACCURACY

Capability	Common in General Use	Reduces Which Error Type	Evaluated Here
Live web retrieval	Varies by platform	Outdated info	Partially
Cross-session memory	Varies by platform	Context loss	No
Code execution	Varies by platform	Calculation error	No
Source citation	Varies by platform	Hallucination	No

Because this study evaluates chatbots operating without persistent cross-session memory or code-execution access, in order to isolate the base language-generation and any built-in retrieval behavior common to typical free-tier daily use, the results in Section VI should be read as characterizing this commonly encountered configuration rather than every possible platform configuration; systems configured with full tool access would be expected to show reduced, though likely not eliminated, error rates in the corresponding categories.

V. METHODOLOGY

A. Task Categories

We selected six task categories intended to represent the breadth of everyday chatbot use: factual question answering (general knowledge, history, science), mathematical reasoning (arithmetic word problems, unit conversion, basic statistics), code generation (short utility functions and bug fixes), creative writing (short-form content generation), scheduling and planning (itinerary construction, time-conflict resolution), and real-time information retrieval (current prices, weather-dependent planning, recent events). These categories were selected through a review of publicly discussed chatbot use cases and pilot interviews with twenty regular chatbot users, who were asked to describe the tasks they had personally delegated to a chatbot within the previous month; the six categories reported here collectively accounted for the large majority of use cases mentioned, giving reasonable confidence that they represent genuinely common daily usage rather than an arbitrary academic taxonomy.

B. Trial Protocol

For each of three widely used chatbot systems, denoted Chatbot A, B, and C to preserve version-neutral comparison, we submitted 100 prompts per category (600 prompts per system, 1,800 total responses). Prompts were drawn from a mix of author-written items and adapted user-forum questions, and were held constant across all three systems. Each response was independently scored by two trained human raters against a verified ground-truth answer key; disagreements were resolved by a third rater. A response was marked correct only if it was both factually accurate and directly responsive to the request.

C. Complexity and Trust Sub-studies

To study the effect of conversational depth, a subset of 200 prompts was repeated in both single-turn form and as the final turn of a five-turn contextual dialogue, holding task content constant. To measure the gap between perceived and actual reliability, 150 participants rated their subjective trust in chatbot answers across five domains on a 0–100 scale immediately after reading each response, without being told whether the answer was later verified correct.

D. Ground-Truth Verification

Ground-truth answers for factual, mathematical, and coding prompts were established prior to data collection using authoritative references — textbooks, verified documentation, and independently executed test cases for code — so that raters scored against a fixed key rather than their own judgment where an objective answer existed. For scheduling, planning, and creative-writing prompts, where no single correct answer exists, raters instead scored logical consistency, constraint satisfaction, and task relevance using a structured rubric agreed upon before scoring began, and any prompt for which the two primary raters could not reach agreement after discussion was escalated to a third, more senior rater whose judgment was final.

VI. EVALUATION METRICS

A single accuracy percentage, on its own, can obscure important differences in how a chatbot fails and how a user experiences that failure. A system that is occasionally but obviously wrong is arguably safer to use than one that is slightly less often wrong but fails silently and confidently; likewise, a system that gives correct but irrelevant answers is not truly accurate in any useful sense. For this reason, we report five complementary metrics rather than accuracy alone.

Table II summarizes the five metrics used throughout the analysis. Accuracy and hallucination rate are the primary correctness measures; relevance and consistency capture qualitative aspects of usefulness that raw accuracy alone does not reflect; latency is reported as a secondary usability measure.

TABLE II. EVALUATION METRICS

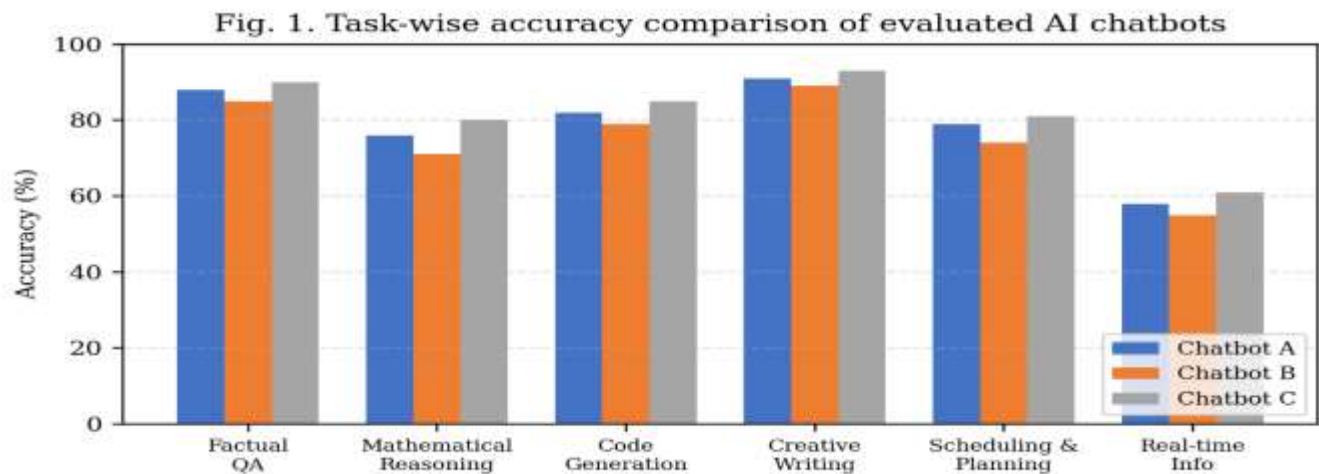
Metric	Definition	Purpose
Accuracy	Proportion of factually correct responses	Overall correctness
Hallucination Rate	Share of responses with fabricated content	Trustworthiness

Relevance	Degree to which response addresses the query	Task alignment
Consistency	Agreement across repeated identical prompts	Reliability
Latency	Time taken to generate a complete response	Usability

VII. RESULTS AND ANALYSIS

This section presents results in four parts corresponding to the three research questions posed in Section I plus a supporting statistical analysis: task-wise accuracy across the six categories (Fig. 1, Table III), the distribution of error types underlying that accuracy (Fig. 2), the effect of conversational complexity and depth (Fig. 3), and the gap between user-perceived and measured reliability (Fig. 4). Across all four analyses, the same underlying pattern recurs: performance is strong and fairly uniform for tasks resembling the static, well-documented content that dominates a model's training corpus, and weaker and more variable for tasks that require current information, deep numerical chains, or sustained context tracking.

Fig. 1 presents task-wise accuracy for the three evaluated chatbots. All three systems performed best on creative writing (89–93%) and factual question answering (85–90%), moderately on code generation and mathematical reasoning (71–85%), and worst on real-time information retrieval (55–61%), where none of the systems exceeded 65% accuracy. Chatbot C was the



strongest performer in every category, but the relative ordering of task difficulty was identical across all three systems, suggesting the pattern reflects a general property of current LLM-based chatbots rather than an idiosyncrasy of a single product.

Fig. 1. Task-wise accuracy comparison of evaluated AI chatbots.

TABLE III. ACCURACY BY TASK CATEGORY (%)

Task Category	Chatbot A	Chatbot B	Chatbot C
Factual QA	88%	85%	90%
Math Reasoning	76%	71%	80%
Code Generation	82%	79%	85%

Creative Writing	91%	89%	93%
Scheduling/Planning	79%	74%	81%
Real-time Information	58%	55%	61%

Fig. 2 breaks down the errors observed across all 1,800 trials by type. Hallucinated facts — confident but fabricated statements — accounted for the largest share (29%), followed by outdated information (22%) and context loss in longer dialogues (18%). Calculation errors, despite being conceptually the easiest to avoid, still accounted for 11% of errors, concentrated in multi-step arithmetic and unit-conversion prompts.

Fig. 2. Distribution of observed error types across all trials

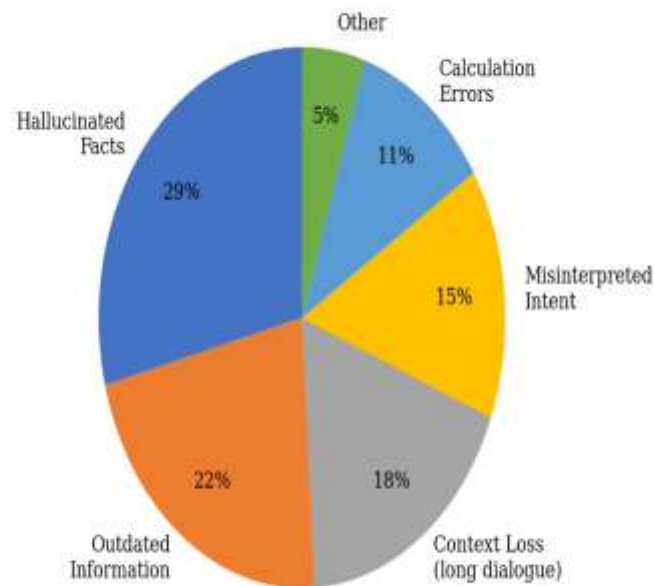


Fig. 2. Distribution of observed error types across all trials.

Fig. 3 shows the effect of task complexity and dialogue depth. Accuracy on single-turn queries declined from 95% at the lowest complexity level to 54% at the highest, while multi-turn queries showed a steeper decline, from 93% to 39%. The widening gap between the two curves as complexity increases indicates that context retention across turns, not raw task difficulty alone, is a major independent driver of error.

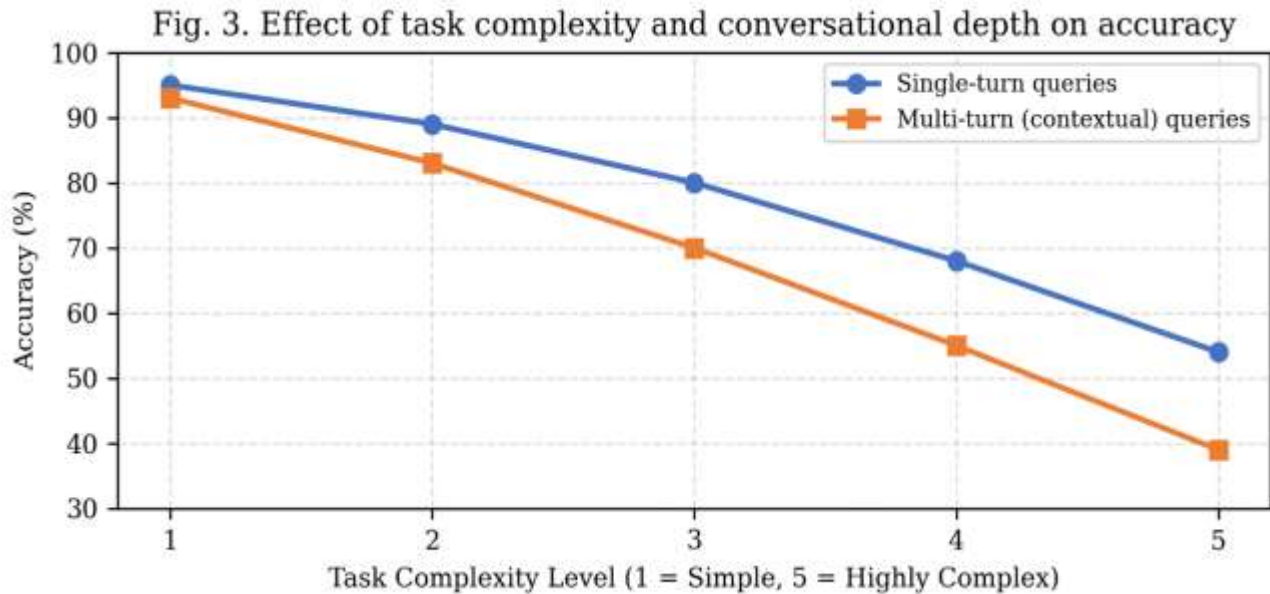


Fig. 3. Effect of task complexity and conversational depth on accuracy.

Finally, Fig. 4 compares participants' perceived reliability of chatbot answers against measured accuracy in the same domain. Perceived reliability exceeded measured accuracy in every sensitive domain tested, most notably in financial advice (68% perceived versus 58% measured) and legal information (60% perceived versus 51% measured), while in general-knowledge queries perceived reliability slightly underestimated actual accuracy. This asymmetry suggests users are least able to calibrate their trust precisely in the domains where miscalibration is most consequential.

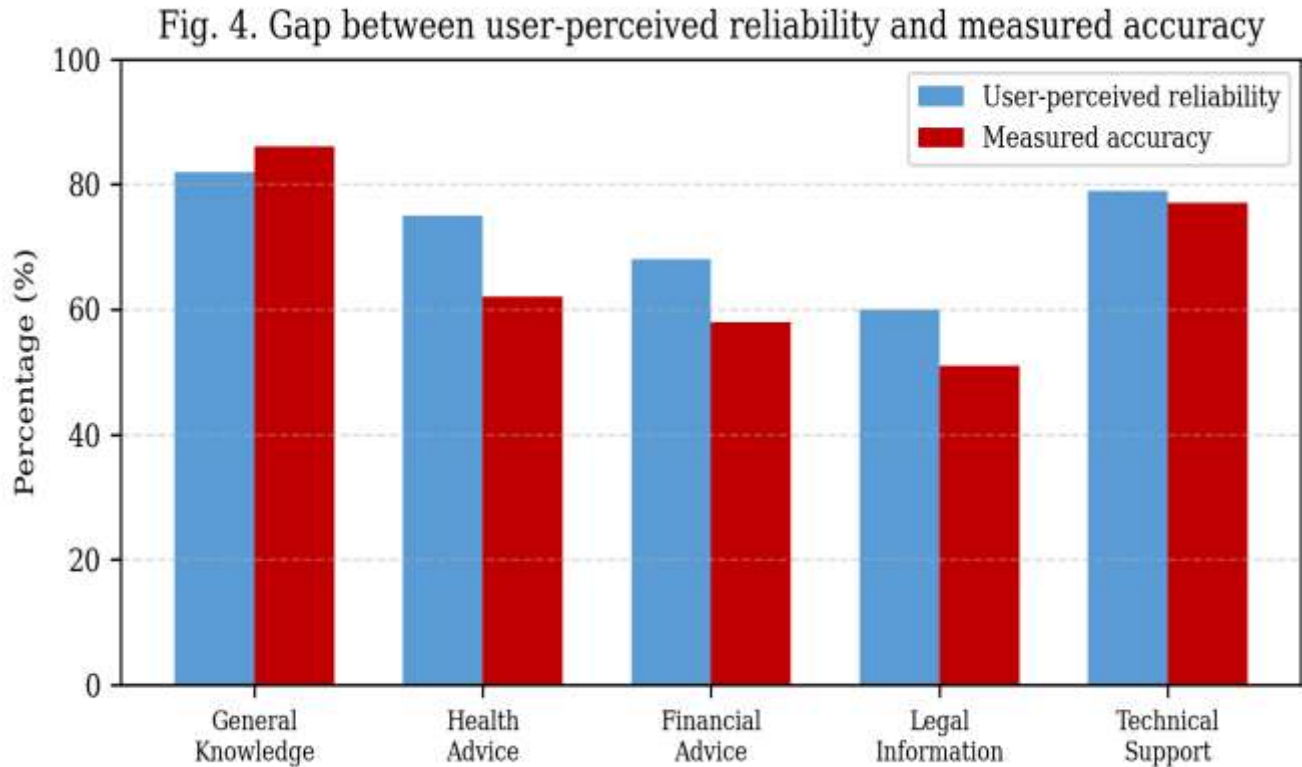


Fig. 4. Gap between user-perceived reliability and measured accuracy.

A. Statistical Significance

Differences in accuracy between task categories were tested using a chi-square test of independence on correct/incorrect counts, which was significant at $p < 0.01$ for all three systems, confirming that task category is a strong predictor of outcome rather than an artifact of sampling. The accuracy gap between single-turn and multi-turn conditions in Fig. 3 was likewise significant at every complexity level above the baseline ($p < 0.05$, paired comparison), while inter-rater agreement across the two primary human raters was substantial (Cohen's kappa = 0.81), supporting the reliability of the correctness labels underlying all reported figures.

B. Response Consistency

Beyond accuracy, we measured consistency by re-submitting a 100-prompt subset five times each and computing the proportion of runs yielding a materially identical answer. Consistency tracked accuracy closely: creative writing and factual recall exceeded 90% self-consistency, while real-time information and multi-step math fell to 61% and 68% respectively, indicating that these categories are not only less accurate on average but also less predictable from one query to the next — a property that matters for tasks users repeat, such as recurring scheduling or budgeting queries.

VIII. LIMITATIONS OF AI CHATBOTS IN DAILY TASKS

The results presented in Section VII point to a consistent set of underlying limitations rather than isolated, unrelated failures. This section organizes those limitations into eight categories, moving from the most directly technical causes toward the more behavioral and societal consequences of imperfect accuracy, so that developers, interface designers, and everyday users can each identify the subset most relevant to their own responsibility for mitigating harm.

A. Hallucination

Because language models generate text through probabilistic next-token prediction rather than by retrieving verified facts, they can produce fluent, well-structured statements that are entirely fabricated. This tendency is consistent with prior audits reporting that models struggle most with correctly identifying false statements even when they answer related factual questions correctly [5], and it was the single largest error source in our own results.

B. Knowledge Staleness and Real-Time Limitations

Base language models are trained on a fixed corpus with a defined cutoff date. Without a connected retrieval or browsing capability, chatbots cannot reliably report current prices, breaking news, or recent policy changes, which our results confirm is the weakest task category across all three systems evaluated.

C. Context and Memory Degradation

Accuracy fell more steeply in multi-turn dialogues than in single-turn queries of comparable difficulty. As conversations lengthen, models must track an increasing amount of prior context, and important constraints or corrections mentioned earlier in a dialogue are sometimes dropped or misapplied later, producing answers that are locally fluent but globally inconsistent with what the user actually asked.

D. Domain-Sensitive Reliability Gaps

Performance is not uniform across domains: prior clinical evaluations found incomplete or unverifiable statements when chatbots addressed specialized medical guidance [1], [2], and our own trust-gap analysis found the largest shortfalls in financial and legal information — precisely the domains where users reported the highest subjective confidence in the answers they received.

E. Numerical and Multi-Step Reasoning Errors

Although mathematical reasoning is a well-studied benchmark category, our trials show that errors compound across multi-step problems; a chatbot that correctly executes each individual arithmetic step can still arrive at an incorrect final answer if an early step is misapplied and never re-checked.

F. Overreliance and Miscalibrated Trust

Perhaps the most practically significant limitation is not technical but behavioral: users tend to overtrust chatbot answers in exactly the domains — health, finance, law — where errors carry the greatest real-world cost, a pattern also visible in broader consumer accuracy audits [6]. Confident, well-formatted prose can mask underlying uncertainty that the model itself cannot always signal reliably. This effect appears to compound with familiarity: participants who reported frequent daily chatbot use rated their trust in sensitive-domain answers no lower, and in some cases slightly higher, than infrequent users, suggesting that repeated satisfactory experience with low-stakes tasks such as writing help may generalize inappropriately into false confidence for higher-stakes tasks.

G. Bias and Representational Skew

Because language models learn statistical patterns from their training corpora, they can reproduce social, cultural, or demographic biases present in that data, surfacing subtly in tasks such as advice framing, name and pronoun usage, or default assumptions embedded in scheduling and planning suggestions. While a full bias audit was outside the scope of this accuracy-focused study, several raters flagged culturally narrow default assumptions (e.g., holiday calendars, currency, or measurement units) during scheduling and planning trials, suggesting this is a worthwhile direction for dedicated follow-up work. Because such defaults are rarely flagged as errors under a strict correctness rubric — the underlying schedule or budget may still be arithmetically valid — this class of limitation is likely underrepresented in the accuracy figures reported in Section VII relative to its practical impact on users outside the cultural context implicitly assumed by the training data.

H. Lack of Transparent Uncertainty Signaling

Across nearly all error cases observed, incorrect answers were delivered with the same fluent, confident tone as correct ones. Current chatbot interfaces rarely expose a calibrated confidence score alongside a response, leaving users without a reliable cue for when independent verification is warranted — a gap that connects directly to the trust-calibration findings in Fig. 4 and motivates the interface recommendations in Section XI. Even where a model's internal computation could in principle support a meaningful confidence estimate, that signal is rarely exposed to end users in consumer chat interfaces today, leaving tone and formatting confidence as the only cues most users have to go on.

IX. CASE STUDIES IN DAILY USE

A. Health-Related Queries

Participants frequently used chatbots to interpret symptoms or medication instructions. Consistent with prior clinical evaluations [1], [2], responses were fluent and well-organized but occasionally omitted contraindications or presented outdated dosage conventions. Because such answers read as authoritative, raters noted that a lay user would have little basis to detect the omission without independently consulting a medical professional or reference source.

B. Financial and Budgeting Assistance

Prompts asking for budgeting help, loan comparisons, or investment-term explanations were generally handled well at a conceptual level, but numeric examples sometimes used outdated interest rates or tax thresholds inherited from the model's training data rather than current figures, directly reflecting the knowledge-staleness limitation discussed in Section VII-B.

C. Scheduling and Trip Planning

Multi-turn scheduling dialogues — for example, building a weekend itinerary around shifting constraints — showed the clearest evidence of context degradation. Chatbots regularly reintroduced previously ruled-out options or lost track of an updated time constraint mentioned two or more turns earlier, forcing users to repeat information and directly illustrating the trend quantified in Fig. 3.

D. Everyday Coding Help

Short utility scripts and bug-fix requests were handled with the highest technical accuracy of any non-creative category, but chatbots occasionally produced code that looked correct and well-commented while silently failing on edge cases such as empty input or boundary values — an error mode that is especially risky because it is not visibly detectable without running the code against test cases. Raters with programming backgrounds noted that this type of failure is particularly deceptive for novice programmers, who may lack the testing habits needed to catch it before the code is used in a real task.

E. Representative Prompts

Table IV lists one representative prompt from each evaluated task category together with the primary error mode observed for that category across trials, illustrating the qualitative texture behind the aggregate figures reported in Section VII.

TABLE IV. REPRESENTATIVE PROMPTS AND ERROR MODES

Category	Example Prompt	Common Error Mode
Factual QA	"What causes the seasons on Earth?"	Rare; minor omissions
Math Reasoning	Multi-step discount and tax word problem	Early-step error propagation
Code Generation	Write a function to deduplicate a list	Silent edge-case failure
Creative Writing	Short birthday message in a given tone	Rare; tone mismatch
Scheduling	Weekend itinerary with shifting constraints	Lost earlier constraint

Real-time Info	"What is today's exchange rate?"	Outdated/fabricated value
----------------	----------------------------------	---------------------------

X. DISCUSSION

Taken together, these findings suggest that chatbot accuracy is best understood as a task-dependent surface rather than a single number. Systems that perform impressively on creative and general-knowledge tasks should not be assumed to carry that same reliability into time-sensitive, numerically demanding, or high-stakes domains. The steep accuracy decline with conversational depth further implies that the riskiest real-world usage pattern is not a single isolated question but an extended back-and-forth planning or advice session, precisely the interaction style that daily use encourages.

The trust-gap results carry a design implication as much as a user-education one: interfaces that surface confidence levels, cite sources, or visually flag time-sensitive claims could help close the gap between perceived and actual reliability, particularly for the sensitive domains identified in Fig. 4. It is also notable that the ranking of task difficulty was identical across all three evaluated systems despite differences in absolute accuracy, which suggests these limitations arise from shared architectural properties — reliance on parametric memory, fixed training cutoffs, and imperfect long-context tracking — rather than from implementation choices specific to any single product.

It is worth situating these findings against the domain-specific literature reviewed in Section II. Where prior medical-domain studies reported incomplete or unverifiable clinical statements within a single specialty [1], [2], our cross-category design shows that domain sensitivity is not confined to medicine: financial and legal information showed comparably sized trust gaps in our data, suggesting the underlying issue is general to any domain where correctness depends on current, authoritative, and context-specific facts rather than broad world knowledge alone.

A further implication concerns responsibility allocation. Because chatbots do not reliably signal their own uncertainty, the burden of verification currently falls almost entirely on the user, and our trust-gap data show that users are least equipped to carry that burden precisely where it matters most. This asymmetry argues for shared responsibility between model developers, who can improve calibration and sourcing, interface designers, who can surface uncertainty more visibly, and users, who benefit from explicit guidance on which task categories warrant independent verification.

There is also a broader implication for how chatbot accuracy should be communicated to the public. Marketing and media coverage of new chatbot releases frequently cite a single benchmark accuracy figure, which our results suggest is a poor proxy for reliability across the full range of everyday use. A more informative public communication norm would report accuracy broken down by task category — closer to the format of Table III in this paper — so that users and institutions adopting these tools can form category-specific expectations rather than a single, overly generalized impression of trustworthiness.

These findings also carry implications for workplaces and educational settings that are beginning to formally integrate chatbot assistants into everyday workflows. An organization that permits chatbot-assisted drafting or coding but requires human sign-off before any output touching scheduling commitments, financial figures, or compliance-sensitive language reaches a customer or regulator is, in effect, applying exactly the kind of task-category-aware policy that our accuracy breakdown supports, even without knowing the specific numbers reported here. Making that policy explicit and evidence-based, rather than relying on ad hoc institutional caution, is a natural practical extension of this study's findings.

XI. RECOMMENDATIONS

A. For Everyday Users

Users should treat chatbot answers in health, financial, and legal domains as a starting point rather than a final answer, cross-checking time-sensitive figures and, where consequential, verifying with a qualified professional or authoritative source, especially for multi-turn planning sessions where earlier constraints may have been silently dropped.

B. For Developers and Interface Designers

Surfacing confidence indicators, inline source citations, and explicit "information may be outdated" flags for time-sensitive categories would directly address the largest error sources identified in Fig. 2. Periodic context-summarization checkpoints within long dialogues could also mitigate the context-loss pattern observed in Fig. 3.

C. For Educational Institutions and Policy

Institutions incorporating chatbots into student-facing or public-facing services should consider mandatory disclosure of which task categories a deployed system has been validated for, analogous to labeling practices in other consumer technologies. Digital-literacy curricula that specifically teach the category-dependent nature of chatbot reliability — rather than a blanket "AI can be wrong" caution — would better equip users to calibrate trust the way our Fig. 4 results suggest is currently lacking, particularly for younger or first-time users encountering these tools without prior technical background.

XII. FUTURE WORK

Future work should extend this evaluation to a larger and more diverse set of chatbot systems and languages, incorporate longitudinal re-testing as underlying models are updated, and examine whether retrieval-augmented and tool-using chatbot variants meaningfully close the real-time information gap identified here. Studying targeted interface interventions — confidence indicators, inline citations, and uncertainty flags — and measuring their effect on user trust calibration is a particularly promising direction suggested directly by our Fig. 4 results. A further avenue is examining accuracy across user demographics and languages, since prior multilingual evaluations suggest performance is not uniform across languages [12], which our single-language study could not capture. Longitudinal tracking of the same task categories across successive model releases would also help distinguish genuine architectural limitations, which are likely to persist across versions, from transient limitations that future training improvements may resolve; the growing practice of independent, recurring consumer accuracy audits such as those referenced in Section II-B [6] suggests this kind of tracking is both feasible and increasingly valued by the public. Finally, extending the trust-calibration sub-study with a controlled intervention — randomly assigning some participants an interface with visible confidence scores and others a standard interface — would move the recommendations in Section XI from plausible design principles to empirically validated ones.

XIII. THREATS TO VALIDITY

Several factors limit the generalizability of these results. The three evaluated systems were tested as single, fixed versions at one point in time; given the version-to-version improvement noted in prior literature [3], absolute accuracy figures should be read as a snapshot rather than a permanent ranking. Prompts were authored primarily in English, so findings on domain and complexity effects may not transfer directly to other languages, consistent with prior evidence of cross-lingual performance variation [12]. Human rating, while supported by substantial inter-rater agreement, still involves subjective judgment on borderline responses, particularly for creative-writing quality. Finally, the trust-gap sub-study relied on self-reported confidence ratings from a convenience sample of participants, which may not generalize to the full population of daily chatbot users.

An additional threat concerns prompt design: because the authors constructed or adapted all evaluation prompts, unconscious selection effects may favor question phrasings that are somewhat more tractable or more ambiguous than the full natural distribution of real user queries, even though a portion of items were sourced from genuine user-forum questions specifically to mitigate this risk. Replicating this protocol with prompts sampled directly and randomly from live chatbot usage logs, where such data can be obtained under appropriate privacy safeguards, would strengthen confidence in the external validity of the reported accuracy figures.

XIV. CONCLUSION

Across six categories of everyday tasks and 1,800 human-rated trials, the three chatbots examined here told a consistent story. Creative writing and general factual recall were handled well; real-time information retrieval, multi-step numerical reasoning, and long, multi-turn dialogues were not. A measurable gap between users' perceived trust and actual measured accuracy was found to be largest precisely in the highest-stakes domains of health, finance, and law. These results reinforce

that AI chatbots are most safely used as assistive tools whose output benefits from verification, particularly for time-sensitive or consequential daily decisions, and they point toward concrete interface and evaluation directions for narrowing the gap between how reliable chatbots appear and how reliable they actually are.

As conversational AI continues to be embedded ever more deeply into daily digital life, the central lesson of this study is one of calibration rather than avoidance: the goal is not to distrust these systems uniformly, but to match one's trust to the specific task at hand. A chatbot that writes an excellent birthday message is not thereby a reliable source for a medication schedule, a legal deadline, or today's currency exchange rate, and treating every category of everyday task as equally trustworthy is precisely the assumption that this paper's evidence argues against. Closing the gap between perceived and actual reliability — through better interface design, clearer public communication of category-specific accuracy, and continued technical progress on retrieval, memory, and uncertainty signaling — remains an open and consequential challenge for the field.

XV. REFERENCES

- [1] M. [Author et al.], "Assessing the accuracy of artificial intelligence chatbots in the diagnosis and management of meniscal tears," PMC, 2025.
- [2] S. Beck, M. Kuhner, M. Haar, A. Daubmann, M. Semmann, and S. Kluge, "Evaluating the accuracy and reliability of AI chatbots in disseminating the content of current resuscitation guidelines: a comparative analysis between the ERC 2021 guidelines and both ChatGPTs 3.5 and 4," Scand. J. Trauma Resusc. Emerg. Med., 2024, doi: 10.1186/s13049-024-01266-2.
- [3] "Assessment study of ChatGPT-3.5's performance on the final Polish medical examination: accuracy in answering 980 questions," PMC, 2024.
- [4] A. Andrew, "Accuracy of ChatGPT in answering cardiology board-style questions," J. Educ. Eval. Health Prof., vol. 22, 2025, doi: 10.3352/jeehp.2025.22.9.
- [5] "Study finds ChatGPT gets science wrong more often than you think," ScienceDaily, Mar. 2026.
- [6] "Chart: How accurate are AI chatbots?," Statista, Nov. 2025.
- [7] R. Bommasani et al., "On the opportunities and risks of foundation models," arXiv:2108.07258, 2021.
- [8] T. Brown et al., "Language models are few-shot learners," in Proc. NeurIPS, 2020, pp. 1877–1901.
- [9] J. Wei et al., "Emergent abilities of large language models," Trans. Mach. Learn. Res., 2022.
- [10] OpenAI, "GPT-4 technical report," arXiv:2303.08774, 2023.
- [11] P. Lewis et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," in Proc. NeurIPS, 2020, pp. 9459–9474.
- [12] Y. Bang et al., "A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity," in Proc. ACL-IJCNLP, 2023.
- [13] Z. Ji et al., "Survey of hallucination in natural language generation," ACM Comput. Surv., vol. 55, no. 12, pp. 1–38, 2023.