# AI-Powered Personal Assistants: A Systematic Literature Review of Scheduling and Everyday Life Applications

**M Asif Chishti*, Mehwish Iqbal**

**Department of Computer Science, University of Engineering & Technology, Lahore, Pakistan.**

## ABSTRACT

AI-powered personal assistants are moving beyond simple question answering toward systems that can support scheduling, reminders, communication, task execution, personalized recommendations, and interaction with external services. However, conversational fluency alone does not determine whether such assistants are dependable in everyday use. This systematic literature review examines recent research on AI-powered personal and virtual assistants, with particular emphasis on scheduling, task management, and everyday productivity. Literature published from 2019 through 2026 was identified through IEEE Xplore, the ACM Digital Library, ScienceDirect, SpringerLink, and supplementary searches using Google Scholar and Semantic Scholar. The documented selection process identified 68 records, removed 14 duplicates, screened 54 records, assessed 30 full-text reports, and retained 23 focal studies for comparative synthesis. Because the included evidence comprised reviews, surveys, user studies, benchmarks, prototypes, simulations, and design research, a thematic narrative synthesis and review-specific quality/relevance appraisal were used rather than quantitative meta-analysis. The evidence shows that dependable everyday assistance requires more than accurate language understanding: assistants must maintain task state, interpret temporal and contextual information, select and execute the correct function, validate important parameters, recover from misunderstandings, and preserve meaningful user control. Personalization and proactive behavior can improve relevance, but their value is strongly conditioned by timing, transparency, privacy, suitability, and the user's ability to correct or override the system. The review therefore identifies operational reliability, contextual reasoning, controlled personalization, calibrated proactivity, and verifiable tool execution as interconnected requirements for future personal assistants. A review-derived conceptual framework is presented to organize these requirements and guide future design and evaluation.

**Keywords:** Intelligent personal assistants, virtual assistants, task scheduling, task execution, personalization, proactive assistance

## INTRODUCTION

Artificial intelligence has expanded digital assistants from relatively simple interfaces for information retrieval into interactive systems that can support communication, organization, decision-making, and action across digital services. AI-based digital assistants combine natural-language interaction with software functions and service access, allowing users to express goals conversationally instead of navigating each application manually [1]. Intelligent personal assistants form a broader class of such systems and commonly rely on natural-language processing, speech recognition, dialogue management, user modeling, and contextual information to interpret requests and provide assistance [2]. As these systems become more capable, the central question is no longer only whether an assistant can understand an utterance, but whether it can translate that understanding into an appropriate and dependable real-world action.

Scheduling and everyday task management expose this distinction particularly clearly. A request such as arranging a meeting, setting a reminder, postponing a task, sending a message, or coordinating several activities contains requirements that extend beyond language generation. The assistant may need to resolve temporal expressions, identify priorities and constraints, maintain task state across several conversational turns, interact with an external calendar or application, and determine whether confirmation is necessary before an action is committed. Research on time-management assistants has consequently emphasized that user models should remain understandable and inspectable rather than functioning as inaccessible representations of a person's priorities and routines [3]. Experimental work involving office tasks such as sending email, creating

timers or reminders, and retrieving contact information further demonstrates that assistant usefulness depends on the task and operating environment rather than on the mere availability of a voice interface [4].

The increasing use of personalization adds another layer to this problem. An assistant that knows a user's routines, preferred meeting times, communication habits, recurring commitments, or accessibility requirements can potentially reduce repetitive input and make recommendations more relevant. At the same time, incorrect personalization can repeatedly reproduce an erroneous assumption. A system that incorrectly infers that mornings are always preferred for meetings, for example, may create persistent scheduling friction rather than reducing it. Effective personalization therefore requires not only preference learning but also mechanisms that allow users to inspect, correct, and override what the assistant believes about them [3].

A related development is the movement from reactive toward proactive assistance. Reactive assistants generally wait for explicit requests, whereas proactive assistants may use context to decide that a reminder, suggestion, warning, or action could be useful before the user requests it. Proactivity is attractive for scheduling and time management because the value of assistance often depends on timing. However, evidence does not support the assumption that more proactive behavior is automatically better. A systematic review of proactive voice assistants found that the benefits of proactivity were context dependent, with clearer benefits in safety-critical or emergency settings and mixed findings in other scenarios [5]. The design challenge is therefore to determine not only what assistance might be useful, but when intervention is justified and how much autonomy the user is willing to delegate.

The shift from conversation to action also increases the importance of operational reliability. An assistant may correctly recognize that a user wants to change an appointment yet still select an incorrect function, modify the wrong argument, lose information introduced earlier in the dialogue, or report completion before an external service has actually accepted the change. Such failures are particularly important for scheduling and productivity because the consequences can persist outside the conversation itself. Recent function-calling research therefore evaluates assistants at the level of multi-turn action execution and parameter handling rather than relying only on the apparent quality of generated language.

Privacy, security, usability, and trust further constrain the amount of autonomy that an everyday assistant can safely exercise. Personalized and context-aware systems may process calendars, messages, contacts, interaction histories, routines, locations, or information from connected devices. Systematic evidence on conversational AI shows that privacy, security, and trust concerns overlap substantially when users evaluate these systems [6]. Trust is also related to perceived competence: users need reason to believe not merely that an assistant communicates naturally, but that it can perform the requested task dependably [7]. These human-centered requirements become more important, rather than less important, as assistants gain greater access to personal information and external tools.

Existing systematic reviews provide valuable broad accounts of intelligent assistants, including their architectures, interaction modes, application areas, and technological challenges [2], [8]. However, a broad taxonomy does not fully answer the narrower question faced in everyday productivity settings: what combination of capabilities is required for an assistant to convert a user's scheduling or task-management intention into a timely, correct, verifiable, and controllable outcome? The recent emergence of LLM-based personalization, multi-turn function calling, contextual notification timing, and increasingly proactive assistance makes this question especially relevant.

To address this gap, the present review asks three research questions.

RQ1: Which capabilities are most directly associated with effective scheduling, task management, and everyday productivity?

RQ2: Which technical and human-centered failure modes most consistently limit reliable task completion?

RQ3: How do personalization, contextual awareness, proactivity, external-tool integration, privacy, trust, and user control interact in the design of dependable personal assistants?

The comparative synthesis focuses on 23 focal studies published from 2019 through 2026 and represented by references [9]– [31]. References [1]– [8] are used to establish definitions, prior high-level evidence, and the broader conceptual context of the review. Within the focal evidence set, particular analytical weight is given to research that directly examines time management, notification timing, meeting scheduling, function execution, contextual communication, or everyday productivity. Studies centered on accessibility, education, healthcare, privacy, social interaction, or connected environments are retained where they provide enabling or boundary evidence relevant to the design of everyday assistants, but conclusions from these specialized domains are not treated as direct evidence of scheduling effectiveness.

The remainder of the paper is organized as follows. Section 2 synthesizes the literature thematically across task execution, personalization, proactivity, context, trust, privacy, usability, and social constraints. Section 3 describes the information sources, search strategy, eligibility criteria, study-selection process, data extraction, quality/relevance appraisal, and synthesis procedure. Section 4 presents findings organized around the three research questions. Section 5 discusses the cross-study implications and develops a review-derived conceptual framework for dependable personal assistance. Section 6 concludes the review, states its limitations, and identifies priorities for future research.

## LITERATURE REVIEW

Research on AI-powered personal assistants spans several overlapping areas, including conversational interaction, scheduling, personalized behavior, proactive intervention, external-tool execution, connected-device control, usability, privacy, and trust. Discussing these studies only in publication order obscures an important feature of the evidence: different studies address different parts of the end-to-end assistance problem. The literature is therefore synthesized here by analytical theme. Studies that directly address scheduling, notification timing, time management, task execution, or everyday productivity are distinguished from enabling evidence on personalization, context, function calling, and external-system integration and from boundary evidence concerning privacy, trust, usability, accessibility, social context, and specialized applications.

### Scheduling, Task Execution, and Operational Reliability

The studies most directly aligned with the review objective show that successful personal assistance is an end-to-end task rather than a language-generation task. Lingler et al. [11] examine notification timing by using contextual features to estimate whether an intervention is appropriate in an office setting. Their work shifts the scheduling problem from simply generating a reminder to deciding when that reminder should reach the user. Jeromela and Conlan [17] similarly approach time management as an interactive process in which a proactive assistant can make suggestions and encourage reflection while preserving user involvement. Akbar [21] extends this issue to a meeting-scheduling simulation, where proactive multi-agent behavior raises explicit questions about autonomy and human-in-the-loop control.

These studies converge on the importance of context and timing, but they also reveal a limitation of the evidence base. Notification modeling [11], expert-oriented design work [17], and simulated meeting scheduling [21] represent different forms of evidence and cannot be treated as interchangeable demonstrations of real-world effectiveness. They collectively identify relevant requirements, but longitudinal field evidence showing how these mechanisms perform across months of actual calendar use remains limited.

Task execution introduces an additional source of difficulty. Wang et al. [14] developed HammerBench specifically to evaluate function calling in realistic multi-turn mobile-assistant scenarios. The benchmark includes imperfect instructions, changing intentions and arguments, and references to information introduced earlier in a conversation. Its findings show that a model can understand the broad purpose of a request and still fail at the level of function parameters. This distinction is central to everyday productivity: interpreting “move my meeting with Sara to Friday afternoon” is not sufficient if the assistant modifies the wrong appointment, supplies the wrong date parameter, or fails to verify that the calendar service accepted the change.

External-system studies reinforce this interpretation. TaskSense [15] uses large language models to translate higher-level natural-language requests into operations across heterogeneous sensor systems, while Ali et al. [19] demonstrate natural-language control of connected devices through an IoT-integrated personal assistant. These systems are not scheduling studies in a narrow sense, but they provide enabling evidence for a requirement that scheduling assistants increasingly share: natural-language intent must be converted into valid operations in another system. The relevant performance boundary therefore lies between understanding and verified execution.

Contextual communication provides another direct everyday example. Jain et al. [20] designed a proactive auto-response agent that used smartphone context to communicate a user's unavailability. Its two-week evaluation illustrates the potential value of context-aware automation while also exposing the need for transparency and personalization. In combination, [11], [14], [17], [20], and [21] indicate that dependable scheduling and everyday task support require at least four linked capabilities: maintaining the relevant task state, interpreting context and time, executing the appropriate action correctly, and preserving meaningful user authority over proactive or consequential behavior.

**Personalization, User Models, and Adaptation**

Personalization is frequently presented as a defining feature of next-generation assistants, but the literature distinguishes several different meanings of the term. Mok et al. [12] address personalization at the response level through HiCUPID, a benchmark designed to test whether LLM responses reflect information and preferences specific to an individual user. This is important because a response can sound considerate or helpful without actually being personalized. Benchmark-based evaluation therefore provides a mechanism for separating generic response quality from preference-sensitive behavior.

Long-term everyday use creates a different personalization problem. Wang et al. [13] examine how users themselves adapt after adopting intelligent personal assistants. Their survey-based analysis indicates that changes in established beliefs and routines are related to exploitative and exploratory IPA use, showing that adaptation is bidirectional: assistants learn about users, while users also change how they incorporate assistants into their routines. Consequently, an evaluation focused only on model-side adaptation can miss behavioral changes that influence sustained use.

Scrutability provides a necessary governance mechanism for such personalization. Jeromela [22] argues that representations maintained by an intelligent personal assistant should be open to user inspection and correction. This requirement is especially relevant to scheduling because an incorrect assumption about preferred working hours, recurring activities, travel time, or task priority can repeatedly influence subsequent actions. Scrutability therefore transforms personalization from an opaque prediction problem into an interaction in which stored assumptions can be challenged and revised.

Montalvo et al. [16] provide related evidence from user-centered social interaction design. Their work indicates that preferences of individual users can differ meaningfully from characteristics selected for an assumed target user. Although their emphasis is social interaction rather than scheduling accuracy, the broader implication is transferable: an assistant designed around population-level averages should not assume that a single style of prompting, proactivity, or communication will be acceptable to every user.

Evidence involving older adults further demonstrates why personalization cannot be separated from user characteristics. Pathirana et al. [10] synthesize 48 studies concerning older adults' interactions with AI-powered voice assistants and identify usage behavior, adoption barriers, learnability, user experience, cognitive and emotional effects, and design considerations as recurring themes. The review is valuable for inclusive design, but its population-specific nature also illustrates why findings from one user group should not automatically be generalized to all scheduling and productivity users.

Across these studies, personalization emerges as more than a mechanism for predicting what a user will prefer. Dependable personalization requires an evidence base for the preference, an ability to revise the preference

over time, visibility into consequential assumptions, and interaction designs that recognize meaningful differences among users.

**Proactivity, Context, and Interruption Timing**

Proactivity is particularly relevant to scheduling because many useful forms of assistance are time sensitive. An assistant that detects a likely scheduling conflict only after the missed meeting has little practical value. At the same time, an assistant that continuously interrupts users on the basis of uncertain predictions may reduce rather than improve productivity.

The direct studies illustrate different points along this autonomy continuum. Lingler et al. [11] focus on estimating notification timing from context-sensitive information. Jeromela and Conlan [17] explore proactive suggestions for time management while emphasizing reflection and continued user involvement. Jain et al. [20] evaluate an agent that communicates on a user's behalf when contextual information indicates unavailability. Akbar [21] examines proactive scheduling behavior in a simulated multi-agent environment and explicitly considers the role of human control. Rather than supporting a simple conclusion that proactive assistants are preferable to reactive assistants, these studies indicate that proactivity is conditional on context quality, confidence, timing, and the consequences of acting incorrectly.

Social context can further complicate proactive behavior. Luria et al. [26] show that personal agents operating in shared homes encounter boundaries involving privacy, control, conflicting interests, and expectations among multiple people. A scheduling assistant operating in a household or shared workplace may therefore need to reason not only about one user's preference but also about whose calendar, authority, or information is relevant to an action.

The literature consequently supports a calibrated model of proactivity. Low-risk suggestions may reasonably be delivered with comparatively little friction, while actions that modify commitments, communicate private context, spend resources, or affect other people require stronger confidence, explanation, confirmation, or reversibility. The relevant research question is therefore not how to maximize proactivity, but how to match the degree of autonomy to contextual confidence and potential consequence.

**Trust, Privacy, Usability, and Social Constraints**

Greater personalization and autonomy create corresponding governance requirements. Schadelbauer et al. [18], using an online survey of 367 participants, show that trust in intelligent virtual assistants is influenced by user-related characteristics and general technology-trust tendencies. This means that identical system behavior may not produce identical trust judgments across users. Trust should therefore not be treated as a simple consequence of adding more capabilities.

Privacy is similarly connected to assistant functionality. Hernández Acosta and Reinhardt [24] survey privacy issues and proposed protections for voice-controlled digital assistants. These concerns become directly relevant to scheduling systems because contextual assistance may depend on calendars, messages, contacts, routines, or information collected from devices and environments. Increasing contextual intelligence without specifying what information is collected, how it is used, and how users can limit it would create a design trade-off rather than an unqualified improvement.

Usability evidence also cautions against equating technical capability with practical effectiveness. Dutsinma et al. [23] organize voice-assistant usability research using the ISO 9241-11 perspective and highlight variation in the way effectiveness, efficiency, satisfaction, and context of use are measured. Earlier commercial-assistant studies similarly report differences in response correctness, naturalness, voice recognition, and contextual understanding [29], [30]. Murad and Munteanu [31] focus particularly on interaction breakdowns and the need for conversational interfaces to support recovery when understanding fails. These findings are directly relevant to scheduling because recovery behavior determines whether a misunderstanding becomes a minor correction or an incorrectly executed task.

Application-specific studies provide useful boundary tests of reliability. Al-Kaisi et al. [25] demonstrate the use of a voice assistant in foreign-language learning, whereas Nobles et al. [27] and Alagha and Helbing [28] evaluate assistant responses in health-related information and help-seeking situations. These studies do not establish scheduling performance and should not be interpreted as though they do. Their value to the present review lies elsewhere: they demonstrate that a fluent and accessible interface can still produce inconsistent outcomes and that reliability requirements increase when the consequence of misunderstanding becomes more serious.

The broad review by Todericiu [9] places many of these issues in a larger virtual-assistant taxonomy spanning interaction types, task specificity, deployment approaches, multimodality, IoT integration, and multiple application domains. Its breadth is useful for understanding the technological landscape, but it also reinforces the need for the present review to separate broad assistant capability from evidence directly related to scheduling and everyday task completion.

**Cross-Study Synthesis and Evidence Gap**

Taken together, the evidence reveals three relationships that are less visible when studies are discussed sequentially. First, personalization, proactivity, and task execution are not independent improvements. Better personalization can support more relevant proactive decisions, but a more personalized proactive assistant also creates greater consequences when its user model is wrong. Second, external-tool integration increases usefulness only when operational correctness can be established. The ability to call a calendar, messaging application, sensor service, or connected device is not equivalent to reliable task completion unless the selected function, arguments, external state, and execution result are validated. Third, trust and privacy operate as constraints on autonomy rather than as separate adoption variables added after technical design.

For this reason, the 23 focal studies are interpreted according to their relationship to the central review objective. Direct evidence addresses scheduling, time management, notification timing, contextual communication, function execution, or closely related everyday productivity. Enabling evidence addresses capabilities required for such tasks, including personalization, user modeling, external-system interaction, and recoverable conversation. Boundary evidence identifies conditions affecting transfer to real-world use, including privacy, trust, usability, accessibility, social context, and specialized-domain reliability. Conclusions about scheduling are anchored preferentially in direct evidence; while enabling and boundary studies are used to explain mechanisms and limitations.

The principal gap is therefore not a lack of research on personal assistants in general. It is a shortage of longitudinal, end-to-end evaluations in which a personalized assistant interprets evolving scheduling intentions, interacts with real services, performs verifiable actions, handles conflicts and corrections, and is evaluated over time under realistic privacy and autonomy settings. The comparative evidence matrix in Table 1 makes this distinction explicit.

Table 1. Comparative Evidence Matrix of the 23 Focal Studies

| Ref. | Objective / Focus | Method / Evidence Base | Principal Review-Relevant Finding | Scope Class |
|---|---|---|---|---|
| [9] | Organize developments and applications of virtual assistants | Broad literature review and taxonomy across interaction modes, task types, deployment, multimodality, and IoT | VA capability is increasingly distributed across multimodal and connected environments; breadth limits scheduling-specific inference | Enabling |
| [10] | Examine AI-powered voice assistants for older adults | Systematic review of 48 studies selected from 109 records | Adoption and usefulness depend on learnability, user experience, barriers, and population-specific design requirements | Boundary |
| [11] | Model appropriate | Context-sensitive LLM modeling/evaluation in an | Useful proactive support depends on when an intervention is | Direct |

| | | | | |
|---|---|---|---|---|
| | notification timing | office-notification setting | delivered, not merely whether it can be generated | |
| [12] | Evaluate personalized LLM assistants | HiCUPID conversational dataset, benchmark, and automated preference-sensitive evaluation | Personalization should be evaluated against individual user information rather than generic helpfulness | Enabling |
| [13] | Explain continued and ambidextrous IPA use | Survey, PLS-SEM analysis, post-hoc analysis, and qualitative validation | Users adapt beliefs and routines as they exploit familiar capabilities and explore new IPA functions | Enabling |
| [14] | Evaluate realistic mobile function calling | Multi-turn benchmark based on mobile functions and realistic interaction trajectories | Parameter-level and multi-turn execution errors can occur even when broad user intent is understood | Direct |
| [15] | Translate natural-language requests into heterogeneous sensor operations | LLM-enabled system/prototype evaluation across sensor tasks | Natural language can coordinate external systems, but task decomposition and correct execution remain necessary | Enabling |
| [16] | Compare target-user and individual-user social interaction design | User-centered design framework/comparative preference analysis | Individual interaction preferences can diverge from population-level design assumptions | Enabling |
| [17] | Explore proactive assistance for time management | Expert discussion and design-oriented analysis | Proactive suggestions should support reflection and preserve user involvement rather than simply automate planning | Direct |
| [18] | Examine personality-related factors associated with IVA trust | Online survey of 367 participants | Trust varies with user-related technology-trust characteristics and cannot be reduced to system capability alone | Boundary |
| [19] | Support connected-device control for people with disabilities | IoT/IPA prototype with NLP and machine-learning evaluation on structured and unstructured commands | Natural-language assistants can execute actions in connected environments but require robust command interpretation | Enabling |
| [20] | Automate contextual communication during user unavailability | Contextual auto-response system and two-week study with 12 participants | Context-aware proactive communication can be useful, but transparency and personalization are important | Direct |
| [21] | Examine proactive meeting-scheduling behavior | Multi-agent simulation centered on meeting scheduling | Increasing autonomy creates a corresponding need for human-in-the-loop control | Direct |
| [22] | Develop the concept of scrutable personal assistants | Conceptual/design research on inspectable user models | User models should permit inspection and correction when they influence personalized recommendations or actions | Enabling |
| [23] | Synthesize voice-assistant usability research | Systematic review structured through ISO 9241-11 | Effectiveness, efficiency, satisfaction, and context are measured inconsistently across the | Boundary |

| | | | | |
|---|---|---|---|---|
| | | | literature | |
| [24] | Examine privacy risks and protections in voice-controlled assistants | Privacy-focused survey/review | Access to personal and contextual information creates privacy requirements that increase with assistant integration | Boundary |
| [25] | Assess voice-assistant use in foreign-language learning | Education-focused empirical application using the Alice assistant | Assistants can support recurring learning activities, but the evidence is domain specific rather than scheduling specific | Boundary |
| [26] | Investigate social boundaries for personal agents in shared homes | HCI/design research involving household and interpersonal-agent scenarios | Shared environments introduce conflicting preferences, privacy boundaries, and questions of authority | Boundary |
| [27] | Evaluate responses to addiction help-seeking queries | Standardized comparative query audit of major commercial assistants | Conversational accessibility does not guarantee consistently appropriate intent recognition or response behavior | Boundary |
| [28] | Compare responses to consumer vaccine questions | Exploratory comparative audit of Alexa, Google Assistant, and Siri | Information quality and usefulness can vary across assistant platforms and queries | Boundary |
| [29] | Compare user experience across commercial personal assistants | Comparative user study with 92 participants | Response correctness and perceived naturalness differed among evaluated platforms | Boundary |
| [30] | Examine practical user experience with commercial assistants | Survey involving 100 users of Google Assistant, Siri, Cortana, and Alexa | Voice recognition and contextual understanding remained important practical limitations | Boundary |
| [31] | Examine conversational breakdown and recovery | HCI/design analysis of voice-user-interface failures | Dependable interaction requires mechanisms that communicate failure and help users repair misunderstandings | Enabling |

# METHODOLOGY

A systematic literature review was conducted to identify, assess, and synthesize research concerning AI-powered personal assistants with relevance to scheduling, task management, and everyday life applications. The review covered formally published English-language research from 2019 through 2026 and included intelligent personal assistants, virtual assistants, voice assistants, conversational assistants, and closely related AI-enabled systems. Because the topic combines technical and human-centered evidence, the review considered empirical studies, benchmarks, system evaluations, surveys, design research, and relevant scholarly reviews when they contributed to one or more of the research questions.

The reporting of the study-selection process was guided by PRISMA 2020 [32]. Search reporting was additionally informed by PRISMA-S [33]. These frameworks were used to improve transparency in describing information sources, eligibility criteria, record screening, full-text assessment, and final inclusion.

### Information Sources

The search covered scholarly databases and academic discovery platforms relevant to computer science, artificial intelligence, human-computer interaction, information systems, and intelligent systems. IEEE Xplore

and the ACM Digital Library were used for computing, intelligent-system, and human-computer-interaction research. ScienceDirect and SpringerLink were used for peer-reviewed journal and conference literature involving artificial intelligence, conversational systems, usability, and related applications. Google Scholar and Semantic Scholar were used as supplementary discovery tools for cross-publisher searching and identification of related scholarly publications.

Only formally published scholarly work was eligible for the focal evidence set. Preprint-only records without a corresponding formally published version were excluded. The role of each information source is summarized in Table 2.

Table 2. Information Sources Used in the Review

| **Information Source** | **Role in the Search** |
|---|---|
| IEEE Xplore | Computing, artificial intelligence, intelligent systems, engineering, and related assistant research |
| ACM Digital Library | Human-computer interaction, conversational interfaces, intelligent assistants, and computing research |
| ScienceDirect | Peer-reviewed AI, intelligent-systems, usability, ergonomics, and human-interaction research |
| SpringerLink | Journal and conference research involving AI systems, intelligent assistants, accessibility, and user interaction |
| Google Scholar | Supplementary cross-publisher discovery and citation-based identification of relevant literature |
| Semantic Scholar | Supplementary scholarly discovery and identification of related publications |

**Search Strategy**

The search strategy combined terms describing assistant technologies with terms representing the functional and human-centered dimensions of interest. Search syntax was adapted where necessary to the interface of each information source while preserving the same conceptual groups.

The principal conceptual search expression was:

("intelligent personal assistant" OR "AI personal assistant" OR "virtual assistant" OR "voice assistant" OR "conversational assistant") AND ("scheduling" OR "time management" OR "task management" OR "everyday tasks" OR "daily activities" OR "personalization" OR "context awareness" OR "proactive assistance" OR "function calling" OR "tool use" OR "usability" OR "trust" OR "privacy")

The publication window covered work dated from January 2019 through 2026. Titles and abstracts were screened first. Records that appeared potentially relevant were assessed using the available full text. Publication information was checked to distinguish formally published work from preprint-only versions and to consolidate duplicate records referring to the same study.

**Eligibility Criteria**

A publication was eligible when it met the temporal, language, publication-status, system-relevance, and evidence-relevance requirements shown in Table 3. The scope was deliberately broad enough to identify capabilities and constraints that affect scheduling and everyday assistance, but direct conclusions concerning scheduling or productivity were not drawn solely from specialized application studies.

Table 3. Inclusion Criteria

| Criterion | Inclusion Requirement |
|---|---|
| Publication period | Published from 2019 through 2026 |
| Language | English |
| Publication status | Formally published scholarly journal article, conference paper, or peer-reviewed review |
| Research area | Computer science, artificial intelligence, human-computer interaction, information systems, or a directly relevant application field |
| System relevance | Examines an intelligent, virtual, voice, conversational, or AI-powered personal-assistant technology or a closely related assistant system |
| Functional relevance | Addresses scheduling, task management, everyday assistance, personalization, context awareness, proactive behavior, function/tool use, usability, trust, privacy, or closely related assistant capability |
| Evidence relevance | Provides empirical findings, benchmark evidence, system evaluation, design evidence, user evidence, or systematic synthesis relevant to at least one research question |
| Accessibility | Provides sufficient publication information and accessible content for meaningful assessment |

Studies were excluded according to the criteria in Table 4. In particular, a study was not retained simply because it addressed artificial intelligence generally. A clear relationship to intelligent personal assistance or an enabling/boundary issue affecting such systems was required.

Table 4. Exclusion Criteria

| Criterion | Exclusion Condition |
|---|---|
| Publication period | Published before 2019 |
| Language | Not available in English |
| Publication status | Preprint-only, non-scholarly, or lacking a formally published eligible version |
| Duplication | Duplicate or secondary record of a study already represented by its eligible publication |
| Accessibility | Insufficient publication information or content for meaningful assessment |
| Topic relevance | Does not examine an intelligent/AI-powered assistant or a closely related conversational-assistant system |
| Scope relevance | Does not contribute evidence concerning scheduling, everyday assistance, personalization, context, proactivity, task/tool execution, usability, trust, privacy, or related interaction requirements |

**Study Selection and Data Extraction**

The documented search and screening process identified 68 records. Fourteen duplicate records were removed, leaving 54 records for title-and-abstract screening. Twenty-four records were excluded at this stage. Thirty reports were assessed at full text, seven were excluded after eligibility assessment, and 23 focal studies were retained for the comparative synthesis. Figure 1 summarizes this process.

For every focal study, information was extracted on publication year, research objective, assistant or system type, methodological approach, sample/dataset/system setting where applicable, principal findings, reported limitations, and relationship to the research questions. These fields were used to construct the comparative evidence matrix in Table 1 and to support cross-study thematic synthesis.

References [9]–[31] constitute the 23 focal studies represented in the comparative matrix. References [1]–[8] are used for definitions, background, and prior high-level evidence and are not treated as additional rows in the final comparative matrix.

**Quality and Relevance Appraisal**

The included evidence was methodologically heterogeneous: it contained systematic reviews, surveys, controlled/user studies, computational benchmarks, system prototypes, simulations, design research, and conceptual work. Applying one design-specific risk-of-bias instrument uniformly across all of these study types would therefore have produced misleading comparisons. A structured review-specific appraisal was used instead.

Four appraisal domains were considered for each focal study: (1) clarity of the research objective and methodological design; (2) transparency of the participant sample, dataset, system, or evaluation setting, as applicable; (3) appropriateness and traceability of the analysis or evaluation relative to the conclusions drawn; and (4) explicit acknowledgement of limitations, contextual boundaries, or threats to generalizability. Evidence in each domain was considered clear, partial, or unclear. No composite numerical score was used, and a study was not excluded solely because one domain was limited. Instead, appraisal informed the strength and scope of the claims made during synthesis.

Relevance to the review objective was assessed separately using three categories. Direct evidence concerned scheduling, time management, notification timing, contextual communication, function execution, meeting scheduling, or closely related everyday productivity. Enabling evidence concerned capabilities required to perform these tasks, including personalization, scrutable user modeling, contextual reasoning, function/tool integration, connected-system interaction, and recoverable dialogue. Boundary evidence concerned factors that constrain real-world transfer, including privacy, trust, usability, accessibility, multi-user social context, or reliability demonstrated in a specialized application domain.

This distinction was used to reduce the risk of overgeneralization. Findings from specialized health, education, accessibility, or household studies were not interpreted as direct proof that a scheduling assistant would be effective. Instead, they were used to identify transferable reliability, usability, privacy, accessibility, and governance requirements.

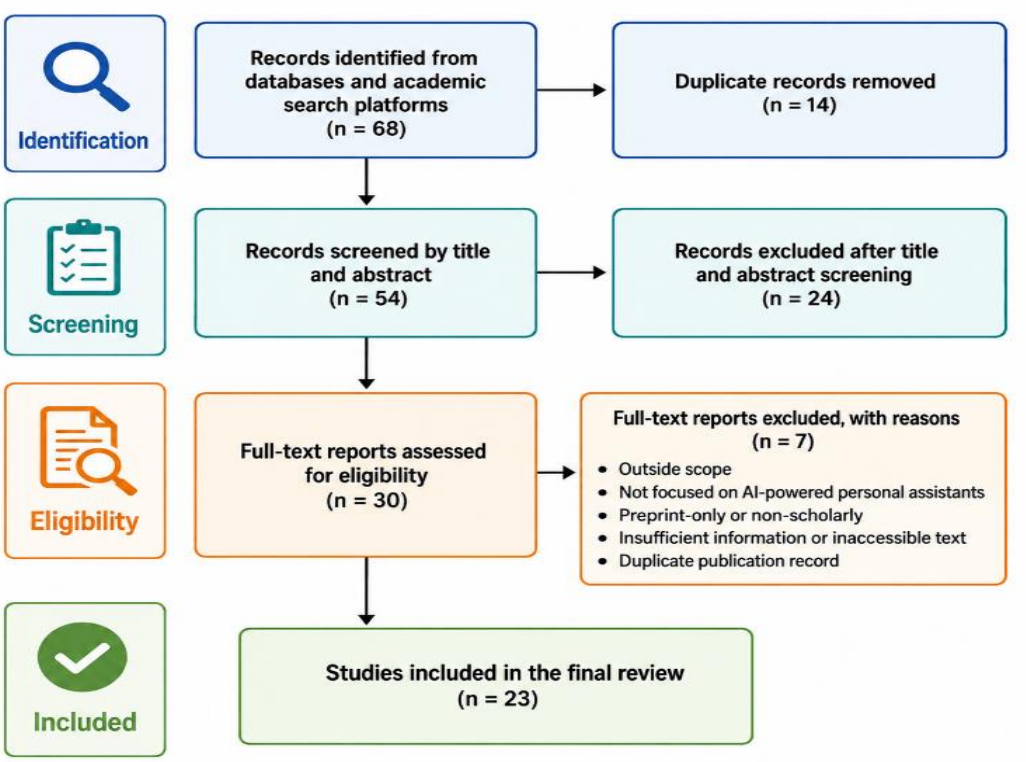
**Synthesis Procedure**

A quantitative meta-analysis was not performed because the included studies differed substantially in research design, participant population, system type, dataset, outcome measure, and application setting. A thematic narrative synthesis was therefore conducted.

Evidence was compared across five connected themes: (1) scheduling, time management, and task execution; (2) personalization, user models, and adaptation; (3) proactivity, context awareness, and intervention timing; (4) external-tool integration and operational reliability; and (5) trust, privacy, usability, accessibility, and social context. Particular attention was paid to convergence, contradiction, and contextual dependence across studies rather than counting every positive or negative result as equivalent.

The final synthesis was organized around the three research questions stated in Section 1. Claims about scheduling and everyday task performance were anchored preferentially in direct evidence. Enabling and boundary evidence was used to explain the technical mechanisms and human-centered conditions under which those capabilities may or may not transfer successfully to routine personal-assistant use. Fig 1 Adapted from PRISMA 2020 [32].

Figure 1. PRISMA 2020-guided flow of literature identification, screening, full-text eligibility assessment, and inclusion [32].




## FINDINGS

The thematic synthesis shows that the major challenges in AI-powered personal assistance occur at the intersections among language understanding, context, user modeling, action execution, and user governance. The reviewed evidence does not support treating these capabilities independently. An assistant may understand a request but execute it incorrectly; it may predict a useful action but intervene at an inappropriate time; or it may personalize accurately while using information in a way that the user considers opaque or intrusive. The findings are therefore organized around the three research questions.

### Capabilities Required for Scheduling and Everyday Task Support

RQ1 asks which capabilities are most directly associated with effective scheduling, task management, and everyday productivity. The strongest direct evidence comes from work on notification timing [11], mobile function execution [14], proactive time-management assistance [17], contextual communication [20], and meeting-scheduling autonomy [21]. Across these studies, effective assistance depends on a chain of capabilities rather than on a single model-level measure.

The first requirement is persistent interpretation of user intent and task state. Scheduling requests frequently evolve over several turns: a user may introduce a meeting, change the day, add a participant, reconsider the time, or refer indirectly to information mentioned earlier. HammerBench [14] is particularly informative because its evaluation includes imperfect instructions, changing intentions and arguments, and references to prior conversational information. The resulting parameter-level failures demonstrate that broad intent recognition is not equivalent to correct execution.

The second requirement is temporal and contextual reasoning. Lingler et al. [11] show that notification support can incorporate contextual features to estimate when an intervention is appropriate. Time-management research [17] likewise suggests that proactive assistance should support planning and reflection rather than simply generating more reminders. In practice, an assistant must distinguish between the existence of a potentially useful action and the suitability of performing or suggesting that action at the current moment.

The third requirement is dependable interaction with external systems. TaskSense [15] and the IoT-integrated assistant developed by Ali et al. [19] demonstrate that natural-language requests can be transformed into operations across sensors or connected devices. For everyday assistants, the same principle applies to calendars, messaging systems, task managers, maps, contacts, and other applications. However, successful integration requires more than producing a tool call. The selected operation, arguments, permissions, target object, and resulting external state must be consistent with the user's intention.

The fourth requirement is verifiable completion and recoverability. A personal assistant should not treat the generation of a plausible confirmation message as evidence that an external task has succeeded. For consequential actions, system feedback should distinguish among an intended action, an attempted action, and a verified completed action. When execution fails or ambiguity remains, the assistant should support correction rather than silently substituting an assumption. The interaction-breakdown analysis in [31] reinforces the importance of understandable recovery behavior.

The fifth requirement is controlled personalization. HiCUPID [12] demonstrates why personalization must be evaluated against information that is genuinely specific to an individual. Scrutable-user-model research [22] adds an important requirement for scheduling: stored assumptions about priorities, routines, preferences, and recurring behavior should remain inspectable and correctable. Together, these studies indicate that personalization should reduce repeated effort without converting previous behavior into an inflexible rule.

RQ1 is therefore answered by an end-to-end capability set: accurate interpretation of evolving intent, temporal and contextual reasoning, persistent task state, preference-sensitive personalization, correct function and parameter selection, verified external execution, understandable recovery, and user control over actions that materially affect commitments or other people.

**Recurrent Failure Modes and Human-Centered Constraints**

RQ2 asks which technical and human-centered failures most consistently limit dependable task completion. The literature identifies several recurring failure modes, but their significance varies according to context.

A central technical failure is the gap between semantic understanding and operational correctness. HammerBench [14] shows that parameter errors can occur even when an assistant has broadly understood what a user wants. This failure mode is particularly consequential for scheduling because a small parameter difference—such as the wrong participant, date, duration, reminder time, or calendar—can produce a formally successful but practically incorrect action.

Contextual errors form a second failure mode. A proactive assistant may possess sufficient information to generate a suggestion without possessing sufficient information to justify an intervention. The studies on notification timing [11], time-management assistance [17], contextual auto-response [20], and proactive scheduling [21] collectively show that context influences not only what an assistant should do but also whether it should act autonomously, ask for confirmation, make a suggestion, or remain silent.

A third limitation is opacity in personalization. As an assistant accumulates preferences and interaction history, an incorrect assumption can persist across later decisions. Scrutability [22] is therefore not simply a user-interface convenience; it is a mechanism for preventing personalization errors from becoming recurrent task errors. The relevance of individualized design is also supported by [16], which shows that preferences of individual users may differ from target-user averages.

Privacy and trust create a fourth set of constraints. Privacy-focused research [24] demonstrates that voice-controlled assistants can process information that users regard as sensitive, while the trust study in [18] indicates that trust judgments are influenced by characteristics of users as well as by the assistant. This becomes especially important when scheduling requires access to calendars, messages, contact relationships, routines, or shared household/workplace information. A technically accurate recommendation may still be unacceptable if the user does not understand why particular personal information was accessed or how it influenced the recommendation.

Usability and failure recovery form a fifth limitation. The systematic usability review [23] shows substantial diversity in how effectiveness, efficiency, satisfaction, and context are evaluated. Commercial-assistant studies [29], [30] similarly identify differences or shortcomings in recognition, contextual understanding, response quality, and naturalness. Murad and Munteanu [31] demonstrate why conversational breakdown should be treated as part of system design: misunderstanding is unavoidable, so dependable systems require mechanisms through which the user can recognize and repair it.

Application-specific studies [25], [27], [28] reinforce an important boundary condition. They show that conversational systems can be useful in specialized settings while also producing uneven outcomes. These studies should not be read as direct evaluations of scheduling, but they demonstrate a transferable principle: apparent conversational competence should not be used as a proxy for reliability in the underlying task or information domain.

RQ2 is therefore answered by a connected set of limitations: parameter and execution errors, incomplete or incorrect context, poorly timed proactivity, persistent personalization errors, opaque use of personal information, inconsistent usability, weak error recovery, and trust problems arising when system confidence exceeds demonstrated competence.

**Interaction among Personalization, Proactivity, Tool Use, and User Control**

RQ3 examines how personalization, contextual awareness, proactivity, external-tool integration, privacy, trust, and user control interact. The evidence indicates that increasing one form of assistant intelligence often increases the need for a corresponding governance mechanism.

Personalization can reduce user effort by allowing an assistant to reuse relevant preferences, but stronger personalization requires stronger mechanisms for inspection and correction. HiCUPID [12] provides a way to evaluate whether outputs actually reflect user-specific information, while scrutability research [22] emphasizes user access to the assumptions underlying personalized behavior. Research on user adaptation [13] also suggests that long-term assistant use changes the behavior of the user, so personalization should remain responsive rather than assuming that earlier routines represent permanent preferences.

Proactivity creates a similar trade-off. The potential value of proactive behavior lies in reducing the need for repeated explicit requests, particularly in time-sensitive settings. However, the direct studies [11], [17], [20], [21] indicate that proactivity should be calibrated according to contextual evidence, timing, consequence, and user preference. The appropriate response to uncertainty is not identical for all tasks. A low-risk suggestion can be easy to dismiss, whereas rescheduling a commitment, sending information to another person, or changing a connected device may justify confirmation.

External-tool integration increases both usefulness and consequence. TaskSense [15], HammerBench [14], and the IoT assistant in [19] demonstrate different mechanisms through which natural language can lead to operational actions. As this capability expands, verification becomes increasingly important. Tool access should therefore be accompanied by parameter validation, permission boundaries, execution feedback, action history, and—where technically possible—reversibility.

Trust emerges from this interaction rather than existing as a separate final layer. If an assistant is transparent but repeatedly executes incorrect actions, transparency alone will not make it trustworthy. Likewise, a highly accurate assistant may still be rejected if users feel that it accesses personal information or acts autonomously without understandable boundaries. Evidence concerning trust [18], privacy [24], and shared-home social boundaries [26] supports treating user control as a continuing property of the interaction rather than a one-time permission decision.

The resulting relationship can be summarized as follows: greater personalization increases the need for scrutability; greater proactivity increases the need for timing confidence and configurable autonomy; greater external-tool access increases the need for validation and verification; and greater contextual data use increases

the need for privacy controls and explanation. These relationships explain why technical capability and responsible user governance must develop together.

### Evidence Boundaries and Implications

The synthesis also reveals important limitations in the current evidence base. Only part of the focal literature directly evaluates scheduling, time management, notification timing, contextual communication, or action execution. Several studies instead provide enabling or boundary evidence from accessibility [10], [19], education [25], household interaction [26], healthcare-related queries [27], [28], general usability [23], or broad virtual-assistant research [9]. These studies remain useful, but their findings should not be generalized beyond the contexts in which they were obtained.

The evidence is also heterogeneous in methodology. A benchmark can expose execution errors under controlled conditions but cannot establish long-term user acceptance. A survey can identify trust or adoption relationships but does not directly measure whether a calendar operation was correct. A design study can identify important interaction principles without demonstrating their frequency or effect in field deployment. A systematic review can synthesize a broad area while inheriting the variability of its underlying studies. For these reasons, the present review emphasizes convergence among different forms of evidence while preserving their methodological boundaries.

A particularly important gap concerns longitudinal end-to-end evaluation. The literature contains evidence on individual components—personalization, notification timing, function calling, proactive behavior, privacy, trust, and usability—but fewer studies evaluate these capabilities together during sustained use. Future assistants will need to perform chains of behavior in which context is interpreted, preferences are applied, an action is selected, an external tool is called, the result is verified, and the user can inspect or correct what happened. Evaluation should therefore measure the complete assistance loop rather than only one conversational turn or one isolated model capability.

Overall, the findings indicate that the transition from reactive to adaptive personal assistance should not be measured by increasing autonomy alone. The more meaningful criterion is whether an assistant can provide the right degree of assistance for the task while maintaining operational correctness, contextual appropriateness, transparency, privacy, and user authority.

## DISCUSSION AND CONCEPTUAL FRAMEWORK

The synthesis indicates that future AI-powered personal assistants require a balanced combination of contextual intelligence, dependable execution, personalization, and user governance. The framework presented in this section is review-derived: it integrates recurring requirements identified across the evidence base and is intended as a conceptual design and evaluation model rather than an empirically validated implementation architecture. This distinction is important because the included studies vary substantially in design, setting, and outcome measure.

### Interpreting the Evidence as an End-to-End Assistance Process

The principal implication of the review is that everyday assistance should be modeled as an end-to-end process. A scheduling interaction does not finish when an assistant generates an appropriate sentence. It begins with understanding what the user wants and may continue through context interpretation, preference retrieval, decision-making, external-tool selection, parameter construction, execution, verification, and possible correction.

This interpretation connects findings that are often studied separately. Personalization research [12], [22] addresses how user-specific information should influence decisions. Notification and proactive-assistance studies [11], [17], [20], [21] address whether and when the system should intervene. Function-calling and external-system work [14], [15], [19] addresses the conversion of natural-language intention into operational

action. Trust, privacy, usability, and social-context research [18], [23], [24], [26], [31] addresses the conditions under which users can reasonably permit such behavior.

These findings suggest that failure in one stage can invalidate success in another. Accurate personalization cannot compensate for an incorrectly executed calendar function. Correct function selection cannot compensate for an intervention made at an unacceptable moment. Strong contextual reasoning cannot justify an action if the user has not delegated the relevant authority. Dependability is therefore a property of the complete interaction chain.

**Review-Derived Capability Framework**

The framework contains five interconnected capability layers: user understanding, context management, intelligent reasoning and decision-making, controlled action execution, and feedback/trust/continuous improvement. Privacy, security, transparency, and user control operate across these layers rather than being treated as final additions.

The User Understanding Layer represents preferences, routines, priorities, accessibility requirements, and relevant interaction history. Personalization should be based on information sufficiently specific to improve the user's task while allowing consequential assumptions to be inspected or corrected [12], [22]. The purpose of this layer is not to construct the largest possible user profile; it is to maintain only the information required to make assistance meaningfully more relevant.

The Context Management Layer represents temporal information, availability, current activity, environmental conditions where relevant, conversation history, and task state. Scheduling depends strongly on this layer because the same suggestion may be appropriate at one moment and disruptive at another. Evidence on notification timing and proactive assistance indicates that context should influence both the content of an intervention and the decision to intervene at all [11], [17].

The Intelligent Reasoning and Decision Layer interprets the user's objective, resolves ambiguity, identifies constraints, selects an appropriate level of autonomy, and determines whether additional information or confirmation is necessary. This layer should explicitly represent uncertainty rather than converting every ambiguous request into a confident action. In a scheduling context, uncertainty may concern the intended event, participant, time zone, duration, priority, or permission to modify an existing commitment.

The Controlled Action Execution Layer converts decisions into interactions with calendars, messaging applications, task managers, connected services, sensors, or other external systems. Evidence from function-calling and connected-system studies shows why this layer requires validation [14], [15], [19]. Before execution, important parameters should be checked against the interpreted intention. After execution, the assistant should use available external feedback to determine whether the requested state was actually achieved.

The Feedback, Trust, and Continuous Improvement Layer support correction, explanation, action review, preference adjustment, and recovery from failure. A user should be able to indicate that an action was wrong, that a preference has changed, or that a particular form of proactivity is unwanted. Feedback should update future assistance without making one correction an unjustified permanent rule.

**User Control, Privacy, and Action Risk**

A key implication of the synthesis is that user control should vary with action consequence rather than being implemented as a uniform confirmation prompt for every operation. Excessive confirmation would eliminate much of the efficiency that makes personal assistants useful; insufficient confirmation could allow uncertain interpretations to produce consequential actions.

A practical design principle is therefore risk-calibrated autonomy. Low-consequence, easily reversible assistance can operate with relatively more autonomy when contextual confidence is high. More consequential

actions should require stronger evidence, clearer disclosure, or explicit confirmation. Relevant factors include whether an action changes another person's expectations, discloses contextual information, modifies an important commitment, controls a physical device, accesses sensitive information, or is difficult to reverse.

Privacy should follow the same principle of proportionality. Contextual and personalized assistance should not assume that every technically available source of information should automatically be used. Users should be able to understand which categories of information support important behaviors and to limit access without making the complete system unusable. Privacy-oriented work [24] and research on social boundaries in shared environments [26] show why personal information and authority can become especially complex when multiple users or shared devices are involved.

### Operational Verification and Recovery

The evidence on tool use requires a stronger definition of task success. A useful personal assistant should distinguish among four states: the user expressed an intention; the assistant interpreted an intended action; the assistant attempted the action; and the external system verified the result. Conflating these states can create false confirmations.

For example, generating the statement "Your meeting has been moved to 3:00 PM" should not itself count as successful scheduling. Where the integrated calendar service provides confirmation, success should depend on the returned state of the relevant event. If execution fails, the assistant should state that the action was not completed and provide an appropriate recovery path. If the target event is ambiguous, the assistant should resolve that ambiguity before making an irreversible or socially consequential change.

HammerBench [14] shows why parameter-level validation is necessary in multi-turn function calling. Interaction-breakdown research [31] similarly shows why the user needs a clear path to repair misunderstanding. Together, these findings suggest that execution feedback and recovery behavior should be treated as core assistant capabilities rather than exceptional error handling.

### Relationship to Existing Assistant Research

The framework differs from broad virtual-assistant taxonomies because its organizing principle is dependable everyday task completion rather than classification of assistant technologies. Broad reviews [9] provide useful categories involving modality, task specificity, deployment, and application area, but scheduling effectiveness cuts across these categories. A voice-based and a text-based assistant may face the same parameter-validation problem; a mobile and smart-home assistant may face the same privacy problem; and an LLM-based and conventional assistant may both need mechanisms for user correction.

The framework also integrates human-centered evidence with technical execution evidence. This is important because the literature suggests no simple trade in which better underlying AI automatically yields better everyday assistance. Usability [23], trust [18], privacy [24], individual preferences [16], and social boundaries [26] remain relevant even when language models become more capable. At the same time, human-centered acceptance cannot substitute for operational correctness when the assistant is expected to perform external actions.

## LIMITATIONS AND FUTURE RESEARCH

This review has several limitations. First, it is restricted to English-language formally published work dated from 2019 through 2026. Relevant preprint-only research and work in other languages were therefore outside the focal evidence set. Second, the 23 studies are heterogeneous in design and include reviews, surveys, benchmarks, prototypes, simulations, user studies, and conceptual/design research. This heterogeneity prevents meaningful quantitative meta-analysis and limits direct comparison of effect sizes or performance measures.

Third, only a subset of the evidence directly evaluates scheduling, time management, contextual communication, notification timing, or task execution. Studies involving older adults, education, accessibility,

healthcare-related questions, smart homes, and privacy were retained principally because they identify enabling capabilities or real-world boundary conditions. Their findings should not be interpreted as direct estimates of scheduling effectiveness.

Fourth, the review-specific quality and relevance appraisal improves transparency across a mixed evidence base but does not replace design-specific risk-of-bias instruments that would be appropriate for a future review restricted to a narrower and methodologically homogeneous set of studies. Finally, the available manuscript search record supports aggregate PRISMA flow reporting but does not provide a complete database-specific historical retrieval log for quantitative comparison of search yields across every information source.

Future research should therefore prioritize longitudinal evaluations in realistic scheduling and productivity environments. Such studies should measure end-to-end task success rather than conversational response quality alone. Useful measures include correct event or task identification, parameter accuracy, conflict detection, execution success, external-state verification, correction cost, interruption timing, false proactive interventions, user override frequency, and recovery after misunderstanding.

Future studies should also evaluate personalization over time. Research is needed to determine how rapidly assistants should update preferences, how users can inspect inferred routines, how conflicting preferences are resolved, and how systems distinguish stable preferences from temporary behavior. Multilingual and cross-cultural studies are needed because language, communication practices, privacy expectations, and scheduling norms can influence interaction.

Finally, evaluation should consider configurable autonomy rather than a binary distinction between reactive and autonomous assistants. The most useful system may not be the system that performs the largest number of actions without asking. It may instead be the system that learns which actions a particular user is comfortable delegating, identifies situations in which uncertainty warrants confirmation, and makes completed actions transparent and recoverable. Figure 2 summarizes this review-derived framework.

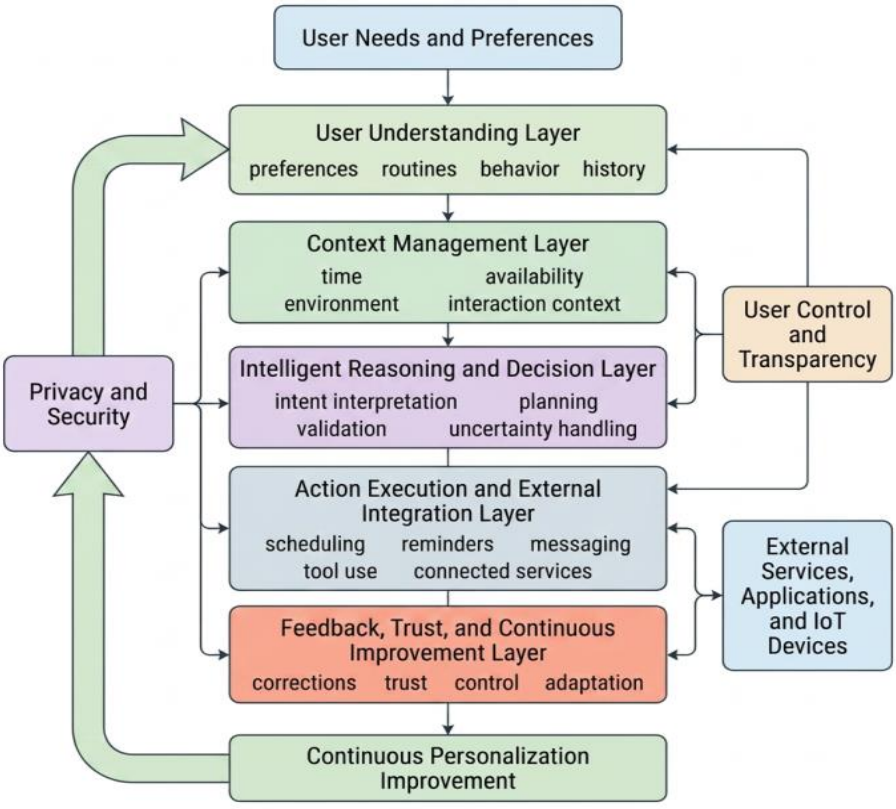



Figure 2. Review-derived conceptual framework for adaptive and dependable AI-powered personal assistance.

## CONCLUSION

This systematic literature review synthesized 23 focal studies published from 2019 through 2026 on AI-powered personal assistants, with particular attention to scheduling, time management, task execution, and everyday life support. By reorganizing the evidence thematically and distinguishing direct, enabling, and boundary evidence, the review shows that the central challenge is no longer simply whether an assistant can produce an appropriate conversational response. Dependable assistance requires the system to convert an evolving user intention into an appropriate, timely, verifiable, and controllable outcome.

With respect to RQ1, effective scheduling and everyday task support depend on a linked set of capabilities: accurate interpretation of user intent, maintenance of task state across conversational turns, temporal and contextual reasoning, meaningful personalization, correct function and parameter selection, reliable external-system interaction, execution verification, and understandable recovery from uncertainty or failure. These capabilities should be evaluated as an end-to-end process because success at one stage cannot compensate for failure at a later stage.

With respect to RQ2, the most persistent limitations are contextual and temporal errors, parameter-level execution failures, weak recovery from misunderstanding, poorly calibrated proactive intervention, opaque or outdated user models, privacy exposure, inconsistent usability, and trust problems created when assistant behavior exceeds demonstrated competence. The reviewed evidence also shows that these limitations are interdependent. Increasing access to context can support better predictions while simultaneously increasing privacy requirements; increasing autonomy can reduce interaction effort while increasing the consequence of incorrect inference.

With respect to RQ3, personalization, proactivity, and external-tool integration produce the greatest value when they are accompanied by appropriate governance mechanisms. Personalization increases the need for scrutability and correction. Proactivity increases the need for timing confidence and user-configurable autonomy. Tool execution increases the need for parameter validation, permission boundaries, execution feedback, action history, and reversibility where possible. Privacy, transparency, trust, and user control should therefore be treated as design requirements that shape how assistant intelligence is deployed rather than as acceptance variables considered only after a system has been built.

The review also identifies limitations in the current evidence base. The included studies use heterogeneous research designs and outcome measures, preventing quantitative meta-analysis. Only part of the literature directly evaluates scheduling and everyday productivity, while studies in accessibility, education, healthcare, IoT, and shared-home contexts mainly provide enabling or boundary evidence. The review is also limited to English-language formally published literature from 2019 through 2026. The review-specific quality and relevance appraisal supports interpretation across heterogeneous evidence but does not replace design-specific risk-of-bias assessment for a future review restricted to one empirical methodology.

Future research should prioritize longitudinal and ecologically realistic evaluations of assistants that operate across real calendars, task managers, messaging systems, and related services. Evaluation should report end-to-end task success, parameter accuracy, conflict handling, execution verification, interruption timing, recovery cost, privacy expectations, user overrides, and configurable autonomy. Research should also examine multilingual and cross-cultural interaction, accessibility requirements, shared and multi-user environments, cross-platform task execution, and the long-term stability and correctability of personalized user models.

Overall, progress toward effective AI-powered personal assistants depends on integrating increasingly capable language and reasoning models with dependable execution and accountable interaction design. The most useful future assistant will not necessarily be the one that acts most autonomously. It will be the one that understands enough context to provide relevant support, knows when uncertainty requires user involvement, performs authorized actions correctly, verifies what actually occurred, and allows the user to remain informed and in control.

# DECLARATIONS

### Competing Interests

The authors declare that they have no competing interests.

### Funding Information

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

### Author Contributions

***M Asif Chishti:*** Conceptualized the study, designed the research framework, conducted the literature review, contributed to literature screening and thematic analysis, synthesized the findings, and drafted and revised the manuscript.

***Mehwish Iqbal:*** Conceptualized the study, contributed to the research framework, conducted the literature review, contributed to literature screening and thematic analysis, synthesized the findings, and drafted and revised the manuscript.

**Both authors:** Reviewed and approved the final version of the manuscript.

### Data Availability Statement

No new primary participant-level data were generated in this study. The review analyzes publicly available scholarly publications cited in the manuscript. The study-selection counts and principal extracted characteristics used for the synthesis are reported in the PRISMA flow diagram and comparative evidence table.

# REFERENCES

1. A. Maedche, C. Legner, A. Benlian, B. Berger, H. Gimpel, T. Hess, O. Hinz, S. Morana, and M. Söllner, “AI-based digital assistants: Opportunities, threats, and research perspectives,” *Business & Information Systems Engineering*, vol. 61, no. 4, pp. 535–544, 2019. https://doi.org/10.1007/s12599-019-00600-8
2. A. de Barcelos Silva et al., “Intelligent personal assistants: A systematic literature review,” *Expert Systems with Applications*, vol. 147, Art. no. 113193, 2020. https://doi.org/10.1016/j.eswa.2020.113193
3. J. Jeromela and O. Conlan, “Devising Scrutable User Models for Time Management Assistants,” in *Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization*, pp. 250–255, 2024. https://doi.org/10.1145/3631700.3665182
4. P. Haghighat, T. Nguyen, M. Valizadeh, M. Arvan, N. Parde, M. Kim, and H. Jeong, “Effects of an intelligent virtual assistant on office task performance and workload in a noisy environment,” *Applied Ergonomics*, vol. 109, Art. no. 103969, 2023. https://doi.org/10.1016/j.apergo.2023.103969
5. C. Bérubé et al., “Proactive behavior in voice assistants: A systematic review and conceptual model,” *Computers in Human Behavior Reports*, vol. 14, Art. no. 100411, 2024. https://doi.org/10.1016/j.chbr.2024.100411
6. A. Leschanowsky, S. Rech, B. Popp, and T. Bäckström, “Evaluating privacy, security, and trust perceptions in conversational AI: A systematic review,” *Computers in Human Behavior*, vol. 159, Art. no. 108344, 2024. https://doi.org/10.1016/j.chb.2024.108344
7. R. Huang, M. Kim, and S. J. Lennon, “Voice-based personal assistant (VPA) trust: Investigating competence and integrity,” *Telematics and Informatics Reports*, vol. 14, Art. no. 100140, 2024. https://doi.org/10.1016/j.teler.2024.100140

8. E. Islas-Cota, J. O. Gutierrez-Garcia, C. O. Acosta, and L.-F. Rodríguez, "A systematic review of intelligent assistants," *Future Generation Computer Systems*, vol. 128, pp. 45–62, 2022. https://doi.org/10.1016/j.future.2021.09.035
9. I. A. Todericiu, "Virtual Assistants: A Review of the Next Frontier in AI Interaction," *Acta Universitatis Sapientiae, Informatica*, vol. 17, Art. no. 1, 2025. https://doi.org/10.1007/s44427-025-00002-7
10. N. Pathirana, A. Htait, and E. Wanner, "AI-powered voice assistants for older adults: A literature review of insights, research practices, and future directions," *Universal Access in the Information Society*, vol. 25, Art. no. 95, 2026. https://doi.org/10.1007/s10209-026-01352-5
11. A. Lingler, H. A. Frijns, L. Boess, N. Murziakova, and P. Wintersberger, "Using LLMs to Model Notification Timing from Context-Sensitive Features," in *Proceedings of the Augmented Humans International Conference* 2026, pp. 406–419, 2026. https://doi.org/10.1145/3795011.3795067
12. J. Mok, I.-h. Kim, S. Park, and S. Yoon, "Exploring the Potential of LLMs as Personalized Assistants: Dataset, Evaluation, and Analysis," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10212–10239, 2025. https://doi.org/10.18653/v1/2025.acl-long.504
13. M. Wang, Y. Lu, and W. Li, "Ambidextrous use of intelligent personal assistants: The role of individual unlearning," Internet Research, pp. 1–20, 2025. https://doi.org/10.1108/INTR-07-2024-1099
14. J. Wang et al., "HammerBench: Fine-Grained Function-Calling Evaluation in Real Mobile Assistant Scenarios," in *Findings of the Association for Computational Linguistics*: ACL 2025, pp. 3350–3376, 2025. https://doi.org/10.18653/v1/2025.findings-acl.175
15. K. Liu et al., "TaskSense: A Translation-like Approach for Tasking Heterogeneous Sensor Systems with LLMs," in *Proceedings of the 23rd ACM Conference on Embedded Networked Sensor Systems (SenSys '25)*, pp. 213–225, 2025. https://doi.org/10.1145/3715014.3722070
16. F. Montalvo, P. Thai, P. Stephens, S. Hinkle, and J. Sasser, "User-centered Social Interaction Design in Intelligent Personal Assistants and Social Robots," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 68, no. 1, pp. 1871–1876, 2024. https://doi.org/10.1177/10711813241274655
17. J. Jeromela and O. Conlan, "Voicing Suggestions and Enabling Reflection: Results of an Expert Discussion on Proactive Assistants for Time Management," in *Proceedings of the 5th International Conference on Conversational User Interfaces*, Art. no. 48, pp. 1–6, 2023. https://doi.org/10.1145/3571884.3604317
18. L. Schadelbauer, S. Schlögl, and A. Groth, "Linking Personality and Trust in Intelligent Virtual Assistants," *Multimodal Technologies and Interaction*, vol. 7, no. 6, Art. no. 54, 2023. https://doi.org/10.3390/mti7060054
19. A.-E. A. Ali, M. Mashhour, A. S. Salama, R. Shoitan, and H. Shaban, "Development of an Intelligent Personal Assistant System Based on IoT for People with Disabilities," *Sustainability*, vol. 15, no. 6, Art. no. 5166, 2023. https://doi.org/10.3390/su15065166
20. P. Jain, R. Farzan, and A. J. Lee, "Laila is in a Meeting: Design and Evaluation of a Contextual Auto-Response Messaging Agent," in *Proceedings of the 2022 ACM Designing Interactive Systems Conference*, pp. 1457–1471, 2022. https://doi.org/10.1145/3532106.3533493
21. A. Akbar, "Proactivity in Intelligent Personal Assistants: A Simulation-Based Approach," *in Multi-Agent Systems*, Lecture Notes in Computer Science, vol. 13442, pp. 423–426, 2022. https://doi.org/10.1007/978-3-031-20614-6_24
22. J. Jeromela, "Scrutability of Intelligent Personal Assistants," in *Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization*, pp. 335–340, 2022. https://doi.org/10.1145/3503252.3534355
23. F. L. I. Dutsinma, D. Pal, S. Funilkul, and J. H. Chan, "A Systematic Review of Voice Assistant Usability: An ISO 9241–11 Approach," *SN Computer Science*, vol. 3, no. 4, Art. no. 267, 2022. https://doi.org/10.1007/s42979-022-01172-3
24. L. Hernández Acosta and D. Reinhardt, "A survey on privacy issues and solutions for Voice-controlled Digital Assistants," *Pervasive and Mobile Computing*, vol. 80, Art. no. 101523, 2022. https://doi.org/10.1016/j.pmcj.2021.101523

25. A. N. Al-Kaisi, A. L. Arkhangelskaya, and O. I. Rudenko-Morgun, “The didactic potential of the voice assistant ‘Alice’ for students of a foreign language at a university,” *Education and Information Technologies*, vol. 26, pp. 715–732, 2021. https://doi.org/10.1007/s10639-020-10277-2
26. M. Luria, R. Zheng, B. Huffman, S. Huang, J. Zimmerman, and J. Forlizzi, “Social Boundaries for Personal Agents in the Interpersonal Space of the Home,” in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–12, 2020. https://doi.org/10.1145/3313831.3376311
27. A. L. Nobles, E. C. Leas, T. L. Caputi, S.-H. Zhu, S. A. Strathdee, and J. W. Ayers, “Responses to addiction help-seeking from Alexa, Siri, Google Assistant, Cortana, and Bixby intelligent virtual assistants,” *npj Digital Medicine*, vol. 3, Art. no. 11, 2020. https://doi.org/10.1038/s41746-019-0215-9
28. E. C. Alagha and R. R. Helbing, “Evaluating the quality of voice assistants’ responses to consumer health questions about vaccines: An exploratory comparison of Alexa, Google Assistant and Siri,” *BMJ Health & Care Informatics*, vol. 26, no. 1, Art. no. e100075, 2019. https://doi.org/10.1136/bmjhci-2019-100075
29. A. Berdasco, G. López, I. Diaz, L. Quesada, and L. A. Guerrero, “User Experience Comparison of Intelligent Personal Assistants: Alexa, Google Assistant, Siri and Cortana,” *Proceedings*, vol. 31, no. 1, Art. no. 51, 2019. https://doi.org/10.3390/proceedings2019031051
30. A. S. Tulshan and S. N. Dhage, “Survey on Virtual Assistant: Google Assistant, Siri, Cortana, Alexa,” in *Advances in Signal Processing and Intelligent Recognition Systems*, *Communications in Computer and Information Science*, vol. 968, pp. 190–201, 2019. https://doi.org/10.1007/978-981-13-5758-9_17
31. C. Murad and C. Munteanu, “‘I Don’t Know What You’re Talking About, HALexa’: The Case for Voice User Interface Guidelines,” in *Proceedings of the 1st International Conference on Conversational User Interfaces*, Art. no. 9, pp. 1–3, 2019. https://doi.org/10.1145/3342775.3342795
32. M. J. Page et al., “The PRISMA 2020 statement: An updated guideline for reporting systematic reviews,” *BMJ*, vol. 372, Art. no. n71, 2021. https://doi.org/10.1136/bmj.n71
33. M. L. Rethlefsen, S. Kirtley, S. Waffenschmidt, A. P. Ayala, D. Moher, M. J. Page, J. B. Koffel, and the PRISMA-S Group, “PRISMA-S: An extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews,” *Systematic Reviews*, vol. 10, Art. no. 39, 2021. https://doi.org/10.1186/s13643-020-01542-z